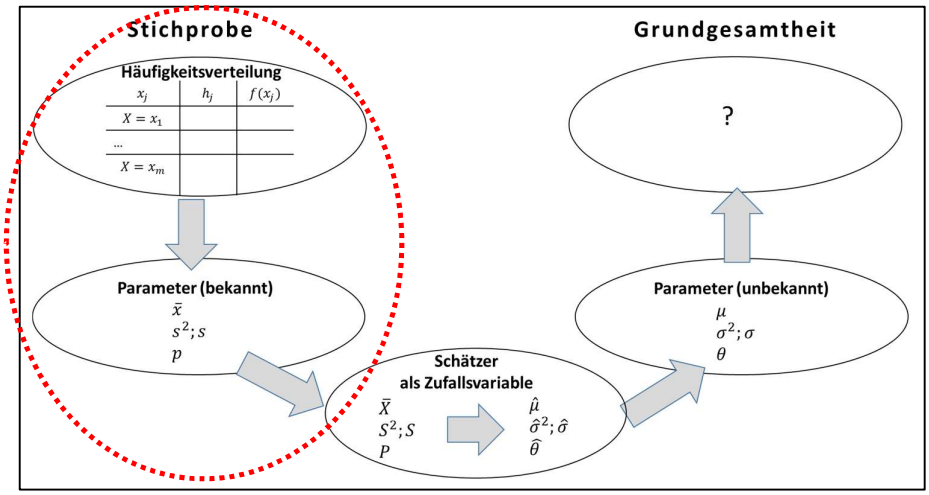
B. BESCHREIBENDE STATISTIK

Das Gebiet der beschreibenden Statistik wird hier in die univariate und die multivariate Statistik differenziert. Bei der **univariaten Statistik** wird nur ein einzelnes Merkmal betrachtet. Das könnte z.B. das Gehalt der Probanden bei einer Befragung sein. Für das Merkmal Gehalt werden dann die Parameter wie Mittelwert und Streuung berechnet. Wird mehr als ein Merkmal betrachtet - zum Beispiel der Zusammenhang von Gehalt, Geschlecht und Alter - spricht man von **multivariater Statistik**. Häufig wird in der Literatur noch die **bivariaten Statistik**, bei der nur zwei Merkmale betrachtet werden, als Sonderfall der multivariaten Statistik behandelt. Dieses könnten zum Beispiel das Gehalt und das Geschlecht sein. Die bivariate Statistik ist bereits ein erster Schritt hin zur Veranschaulichung und Analyse von Zusammenhängen, die erst in dem Kapitel „Schließende Statistik“ dargestellt werden.

Den Zusammenhang der in der beschreibenden Statistik auftretenden Größen veranschaulicht der linke, rot markierte Teil der nachstehenden Grafik. Nur dieser ist in diesem Kapitel von Relevanz.

- Das wesentliche Ergebnis der beschreibenden Statistik ist erst einmal die **Häufigkeitsverteilung**, die einen tabellarischen Überblick über das Auftreten bestimmter Werte in der Stichprobe gibt.
- Die **Parameter** Mittelwert usw. sind Maßzahlen zur Beurteilung der Stichprobe.



Das Kapitel zur univariaten Statistik wird des Weiteren in **nicht klassierte** und **klassierte Verteilungen** differenziert. Bei der Klassierung geht es um die Zusammenfassung von Merkmalsrealisation zu Gruppen bzw. Klassen. So könnte z.B. das Einkommen von befragten Probanden in Gruppen

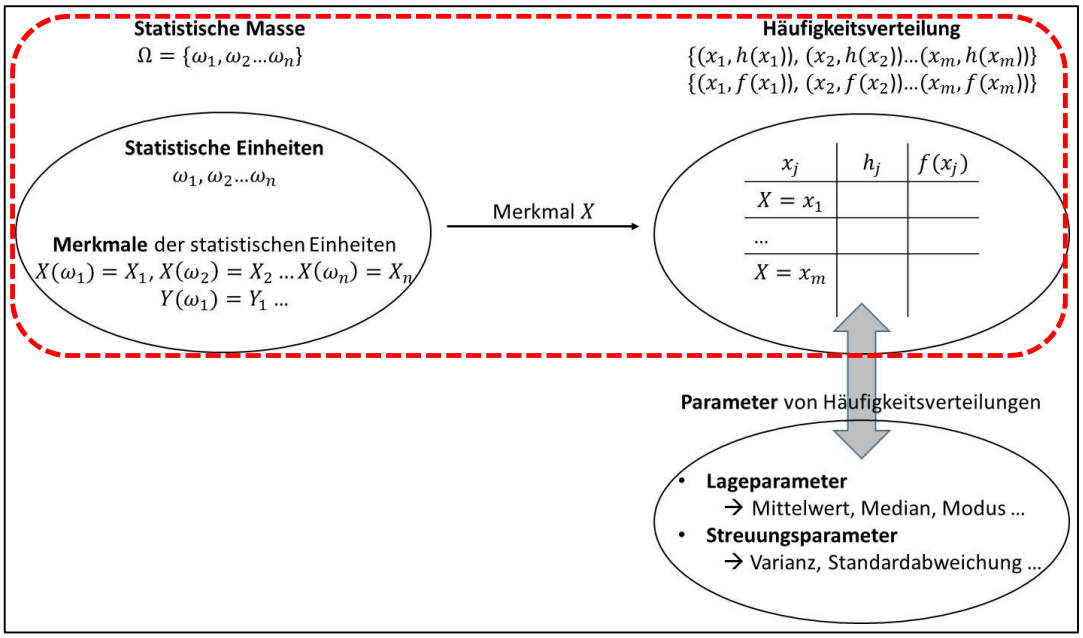
von 10.000 € /p.a. bis 15.000 €/p.a.,
von 15.001 €/p.a. bis 20.000 €/p.a.
usw.

zusammengefasst werden. Diese Thematik ist heute aufgrund der technischen Entwicklung aber nicht mehr von großer Bedeutung.

2 Univariate Statistik

2.1 Nichtklassierte Häufigkeitsverteilung

Wie grundsätzlich in unseren Skripten folgt nach einer kurzen Einführung immer eine kompakte Darstellung der Definitionen und Sätze für den kundigen Leser und danach dann die Veranschaulichung oder auch Herleitung. Dieser Kapitel betrifft nur die oberen, gestrichelt eingerahmten Definitionen der Grafik.



Das nachstehende Beispiel, bei dem 20 Tabletten einer Packung in Gramm gewogen wurden, konkretisiert die obige Skizze noch einmal. Die statistische Masse in Form der Urliste ist links abgebildet, die entsprechende Häufigkeitsverteilung rechts. **Die verwendeten Größen gilt es allgemein zu definieren**.

Urliste (Statistische Masse)

lfd. Nr.	Gramm
1	0,65
2	0,65
3	0,64
4	0,65
5	0,65
6	0,62
7	0,63
8	0,61
9	0,65
10	0,64
11	0,65
12	0,68
13	0,66
14	0,67
15	0,62
16	0,61
17	0,62
18	0,68
19	0,68
20	0,62

Häufigkeitsverteilung

Gewicht (g)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	,61	2	10,0	10,0	10,0
	,62	4	20,0	20,0	30,0
	,63	1	5,0	5,0	35,0
	,64	2	10,0	10,0	45,0
	,65	6	30,0	30,0	75,0
	,66	1	5,0	5,0	80,0
	,67	1	5,0	5,0	85,0
	,68	3	15,0	15,0	100,0
	Gesamt	20	100,0	100,0	

2.1.1 Mathematische Definitionen

Nachstehende Definitionen fassen die mathematischen Grundlagen dieses Kapitels für den kundigen Leser zusammen. Weniger kundige Leser finden Erklärungen dazu in dem Abschnitt „Herleitung & Berechnungen“.

Statistische Einheit ω

Die statistische Einheit ω ist das kleinste Element in der Statistik. Die statistische Einheit ist Träger von Informationen, die für eine statische Untersuchung von Interesse sind

Anmerkungen

- Jede Einheit muss zeitlich, räumlich und sachlich abgegrenzt sein
- Synonyme: Element, Beobachtungseinheit, Merkmalsträger

Statistische Gesamtheit Ω

Eine Menge wohlunterschiedener statistischer Einheiten $\omega_1, \omega_2, \omega_3, \dots, \omega_n$ mit übereinstimmenden Identifikationsmerkmalen heißt statistische Gesamtheit Ω mit

$$\Omega(\omega_1, \omega_2, \omega_3, \dots, \omega_n) = \Omega(\omega_\mathrm{i}) \text{ mit } i = 1, \dots, n$$

Merkmal X

Eine Eigenschaft einer statistischen Einheit, die charakteristisch für eine bestimmte Fragestellung ist, heißt Merkmal X.

Anmerkungen

- Merkmale werden mit lateinischen Großbuchstaben X, Y, Z bezeichnet.
- Statistische Einheiten Merkmals

Merkmalswert $\mathbf{X(\omega_i) = X_i}$

Den Wert, den ein Merkmal X einer statischen Einheit ω_i mit $i = 1, \dots, n$ annimmt, heißt Merkmalswert $X(\omega_i) = X_i$

Merkmalsrealisation

Für ein- oder mehrfach auftretende Merkmalswerte steht ***eine*** Merkmalsrealisation x. Die Menge aller m Merkmalsrealisationen einer Gesamtheit werden wie folgt bezeichnet:

$$x_1, x_2, x_3, \dots, x_m = x_j \text{ mit } j = 1, \dots, m$$

Anmerkungen

- Synonym: Merkmalsausprägung
- Beispiele

Merkmale	*Merkmalsausprägung ; Merkmalsrealisation*
Alter	1,2,3,4,...
Geschlecht	männlich, weiblich

Absolute und relative Häufigkeit

Es sei X ein beliebig skaliertes Merkmal, das mit $m \leq n$ voneinander verschiedene Merkmalsrealisationen x_j für eine Gesamtheit von n Einheiten statistisch erhoben wurde.

Dann heißt für alle $j = 1, 2, \dots, m$ die Anzahl

$$h(X = x_j) = h(x_j) = h_j$$

der Einheiten mit der Merkmalsrealisation x_j **absolute Häufigkeit** der Merkmalsrealisation x_j.

Die Anteilszahl

$$f(X = x_j) = f(x_j) = f_j = \frac{h_j}{\sum_{j=1}^{m} h_j} = \frac{h_j}{n}$$

heißt **relative Häufigkeit** der Merkmalsrealisation x_j.

Häufigkeitsverteilung

Es sei X ein beliebig skaliertes Merkmal. Dann heißt das m-Tupel

$$\{(x_1, h_1), \dots (x_j, h_j), \dots (x_m, h_m)\}$$

absolute Häufigkeitsverteilung des Merkmals X.

Das m –Tupel

$$\{(x_1, f_1), \dots (x_j, f_j), \dots (x_m, f_m)\}$$

relative Häufigkeitsverteilung des Merkmals X.

Summenhäufigkeit

Die Kumulation der absoluten bzw. relativen Häufigkeiten derjenigen Merkmalsrealisationen eines mindestens ordinal skalierten Merkmals X, die die Merkmalsrealisation x_j nicht überschreiten, heißt

absolute Summenhäufigkeit mit

$$H(x_j) = H_j = h(X \leq x_j) = \sum_{r=1}^{j} h_r$$

bzw. **relative Summenhäufigkeit** mit

$$F(x_j) = F_j = f(X \leq x_j) = \sum_{r=1}^{j} f_r$$

Skalen

Die Merkmalsrealisationen eines Merkmals werden für die Analyse in irgendeiner Form an einander gereiht. Das kann z.B. bei dem Familienstand die Reihung „ledig – verheiratet – verwitwet – geschieden“ oder bei den Ergebnissen einer Klausur die Reihung „1 – 2 – 3 – 4 – 5“ oder auch umgekehrt sein. Diese Darstellung wird als Skala bezeichnet.

Von großer Relevanz für die mathematischen Optionen ist die Art die Skala, das sogenannte Skalenniveau. Bei nominal skalierten Merkmalen wie beispielsweise dem Familienstand lässt sich z.B. kein Mittelwert berechnen; bei den erreichten Punkten von Klausuren hingegen ist dieses hingegen sinnvoll.

Definition Skala

Eine Zeichen- bzw. Zahlenmenge zur relationstreuen Abbildung von Merkmalsausprägungen heißt Skala.

Differenzierung von Skalen nach Merkmalsarten

- **… nach dem Skalenniveau (Messniveau)**

Skalenniveau			
nicht-metrisch		metrisch	
nominal	ordinal	kardinal	
• Verschiedenartigkeit	• Verschiedenartigkeit • Rangreihenfolge	• Verschiedenartigkeit • Rangreihenfolge • zählbare Unterschiede (Abstand; Vielfaches mittels reeller Zahlen	
		Skala besitzt ***keinen*** natürlichen Nullpunkt	Skala besitzt ***einen*** natürlichen Nullpunkt
Nominalskala	***Ordinalskala***	***Intervallskala***	***Verhältnisskala***
Familienstand Beruf	Windstärken Schulnoten	Temperatur (Celsius/Fahren.) Intelligenz IQ	Längenmaße Gewichtsmaße

- **... nach der Art der Messbarkeit eines (kardinalen) Merkmals**
  - **Diskrete Merkmale**

    Ein kardinal skaliertes Merkmal, das in einem endlichen Intervall nur einzelne (abzählbar endlich viele) Merkmalswerte annehmen kann, heißt diskretes (diskontinuierlich) Merkmal.
  - **Stetige Merkmale**

    Ein kardinal skaliertes Merkmal, das in einem endlichen Intervall jeden beliebigen Merkmalswert annehmen kann, heißt stetiges (kontinuierliches) Merkmal.

2.1.2 Herleitungen & Berechnungen

Bei den nachfolgenden Erläuterungen ist es sinnvoll sich parallel immer die Definitionen aus dem vorherigen Abschnitt zu vergegenwärtigen, um die Erklärungen mit der Definition abzugleichen.

2.1.2.1 Urliste, Statistische Masse, Statistische Einheit, Merkmal, Merkmalswert

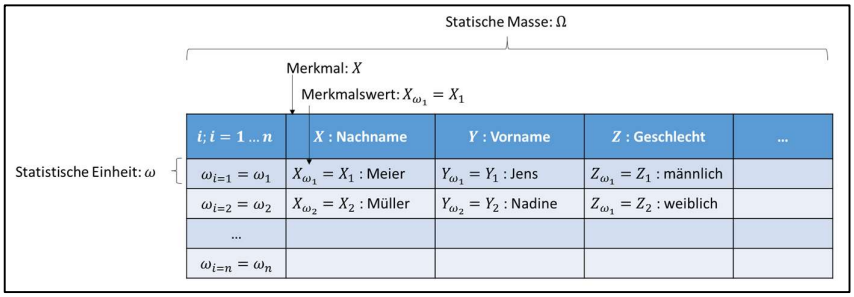
Bei den nachstehenden Erläuterungen sind immer zwei Perspektiven auf das Datenmaterial

- die Urliste und
- die Häufigkeitsverteilung

zu unterscheiden. Die Urliste stellt, wie der Name bereits sagt, die ursprünglichen Daten in Listenform dar. Sie enthält alle Daten der Erhebung in tabellarische Form.

lfd. Nummer	Nachname	Vorname	Geschlecht
1	Meier	Jens	männlich
2	Müller	Nadine	weiblich
…			

Die folgende Urliste ist inhaltlich identisch mit der vorstehenden Tabelle. Sie ist allerdings schon um die Variablenbezeichnungen, die noch zu erklären sind, ergänzt.

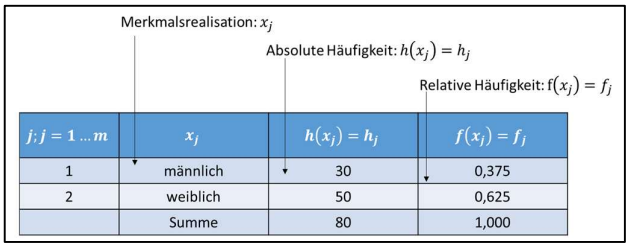


Die Urliste ist der Ausgangspunkt jeder statistischen Untersuchung. Die Gesamtheit der Daten bezeichnet man als **statistische Masse Ω** (Omega groß); jeden Datensatz – hier jede Zeile – als **statistische Einheit $\boldsymbol{\omega}$** (Omega klein). Die **Anzahl der Elemente $\boldsymbol{n}$** bezeichnet die Anzahl der Befragten also die Anzahl der Zeilen. Um eine Menge von Elementen – zum Beispiel der Urliste – allgemein zu beschreiben bedient man sich einer Indexvariable. Im Zusammenhang mit der Urliste wird die **Indexvariable $\boldsymbol{i}$** verwendet, wobei i Werte von 1 bis n annehmen kann. Hat man beispielsweise 80 Probanden befragt, ist $n = 80$ und der Index läuft von 1 bis 80. Da die Schreibweise $\omega_1, \omega_2, \omega_3, \ldots \omega_{80}$ zur Beschreibung der Urliste etwas unbequem ist verwendet man die Schreibweise $\omega_i \; mit \; i = 1 \; bis \; 80$.

Bei jedem Probanden (Befragten) werden bestimmte Informationen erfragt, wie z.B. der Nachname, der Vorname, das Geschlecht usw., die als **Merkmale $\boldsymbol{X, Y, Z}$** ... mit einem großen lateinischen Buchstaben beschrieben werden. X ist folglich das Merkmal „Nachname“, Y das Merkmal „Vorname“ usw. Der konkrete Wert bei dem ersten Probanden bei dem Merkmal „Nachname“ ist „Meier“, was als **Merkmalswert $\boldsymbol{X_{\omega_1}}$** oder in kurzer Schreibeweise einfach **$\boldsymbol{X_1}$** bezeichnet wird. **Es ist folglich zwischen dem abstrakten Merkmal $\boldsymbol{X}$ und dem konkreten Merkmalswert $\boldsymbol{X_{\omega_1} oder \; X_1}$ zu differenzieren.**

2.1.2.2 Häufigkeitsverteilung, Merkmalsrealisation

Aus Gründen der Übersichtlichkeit werden die Daten der Urliste zum Beispiel in Häufigkeitsverteilungen überführt. Eine Häufigkeitsverteilung bezieht sich bei der univariaten Statistik immer nur auf ein Merkmal, in diesem Beispiel $X = Geschlecht$.



$j; j = 1 \dots m$	x_j	$h(x_j) = h_j$	$f(x_j) = f_j$
1	männlich	30	0,375
2	weiblich	50	0,625
	Summe	80	1,000

Bei einer Häufigkeitsverteilung werden alle möglichen Werte, die ein Merkmal annehmen kann, einmal in der Verteilung aufgeführt. Im obigen Beispiel sind das die Optionen „männlich“ und „weiblich“. Diese werden als **Merkmalsrealisationen x_j** (kleiner lateinischer Buchstabe) bezeichnet, wobei die **Anzahl der Merkmalsrealisationen mit m** und der **Laufindex** mit j bezeichnet wird. Damit gilt für das obige Beispiel mit den beiden Merkmalsrealisation „männlich“ und „weiblich“ also $m = 2$ die folgende allgemeine Schreibweise: $x_j \; mit \; j = 1 \; bis \; 2$.

Für die weiteren Betrachtungen ist folglich wichtig, zwischen

- dem Merkmal X,
- dem Merkmalswert X_i (lateinischer Großbuchstabe)
- und der Merkmalsrealisation x_j (lateinischer Kleinbuchstabe)

zu unterscheiden. Das Merkmal ist die abstrakte Beschreibung, der Merkmalswert der konkrete Wert eines Merkmals einer statistischen Einheit ω_i und die Merkmalsrealisation der Wert in der Häufigkeitsverteilung. Identische Merkmalswerte können in der Urliste natürlich mehrfach auftreten, wenn zum Beispiel mehrere Probanden den Merkmalswert „männlich“ aufweisen. Identische Merkmalsrealisationen hingegen sind nicht möglich, da in einer Häufigkeitsverteilung ja gerade die Merkmalsrealisation die Anzahl der Datensätze mit einem identischen Merkmalswert angeben soll. Oben ist beispielsweise $x_1 = 30$ und bedeutet, dass 30 der befragten Probanden als Geschlecht „männlich“ angegeben haben. Ferner ist bei der Indexierung zu beachten:

- Merkmalswerte $X_i \; mit \; i = 1 \; bis \; n$
- Merkmalsrealisationen $x_j \; mit \; j = 1 \; bis \; m$

2.1.2.3 Häufigkeiten

An dem nachstehenden Beispiel sollen die unterschiedlichen Definitionen zu Häufigkeiten erläutert werden. Es handelt sich dabei um die Analyse eines Medikaments, bei dem 20 Tabletten auf unterschiedliche Merkmale hin untersucht wurden. In diesem Beispiel ist aber nur das Merkmal $X = Gewicht$ relevant.

$i = 1 \dots 20$	X : Gewicht	Y : Durchmesser	…
ω_1	$X(\omega_1) = 0{,}65$		
ω_2	$X(\omega_2) = 0{,}65$		
ω_3	$X(\omega_3) = 0{,}64$		
ω_4	$X(\omega_4) = 0{,}65$		
ω_5	$X(\omega_5) = 0{,}65$		
ω_6	$X(\omega_6) = 0{,}62$		
ω_7	$X(\omega_7) = 0{,}63$		
ω_8	$X(\omega_8) = 0{,}61$		
ω_9	$X(\omega_9) = 0{,}65$		
ω_{10}	$X(\omega_{10}) = 0{,}64$		
ω_{11}	$X(\omega_{11}) = 0{,}65$		
ω_{12}	$X(\omega_{12}) = 0{,}68$		
ω_{13}	$X(\omega_{13}) = 0{,}66$		
ω_{14}	$X(\omega_{14}) = 0{,}67$		
ω_{15}	$X(\omega_{15}) = 0{,}62$		
ω_{16}	$X(\omega_{16}) = 0{,}61$		
ω_{17}	$X(\omega_{17}) = 062$		
ω_{18}	$X(\omega_{18}) = 0{,}68$		
ω_{19}	$X(\omega_{19}) = 0{,}68$		
ω_{20}	$X(\omega_{20}) = 0{,}62$		

Die statistische Gesamtheit Ω ist in diesem Beispiel die Menge der 20 Tabletten. Die statistischen Einheiten ω_i mit $i = 1\ bis\ 20$ entsprechen dann jeder einzelnen Tablette mit allen gemessenen Merkmalswerten für $X, Y,$ ...(zum Beispiel X=Gewicht, Y= Durchmesser). Wichtig ist die Unterscheidung zwischen den Merkmalswerten $X(\omega_i) = X_i$ einer Tablette und der Merkmalsrealisation x_j. Aus der obigen Urliste ergibt sich die nachstehende Häufigkeitsverteilung. Aus der Urliste ergibt sich bei der Merkmalsrealisation $x_1 = 0{,}61$, dass 2 Tabletten ein Gewicht von 0,61 Gramm aufweisen, 4 Tabletten ein Gewicht von 0,62 Gramm aufweisen.

Häufigkeitsverteilung in SPSS

Gewicht (g)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	,61	2	10,0	10,0	10,0
	,62	4	20,0	20,0	30,0
	,63	1	5,0	5,0	35,0
	,64	2	10,0	10,0	45,0
	,65	6	30,0	30,0	75,0
	,66	1	5,0	5,0	80,0
	,67	1	5,0	5,0	85,0
	,68	3	15,0	15,0	100,0
	Gesamt	20	100,0	100,0	

Eigene Häufigkeitsverteilung (Excel)

x_j	$h(x_j)=h_j$	$f(x_j)$	$F(x_j)$
0,61	2	0,10	0,10
0,62	4	0,20	0,30
0,63	1	0,05	0,35
0,64	2	0,10	0,45
0,65	6	0,30	0,75
0,66	1	0,05	0,80
0,67	1	0,05	0,85
0,68	3	0,15	1,00
Summe	20	1,00	

Die absolute Häufigkeit ist die Anzahl der Fälle, bei denen der Merkmalswert X_i oder allgemeine das Merkmal X mit der Merkmalsrealisation x_j übereinstimmt, also gilt $X_i = x_j$ oder einfach $X = x_j$

$$h(X = x_j) = h(x_j) = h_j\ .$$

Umgangssprachlich zählt man einfach die Fälle mit einer bestimmten Merkmalsrealisation. Des Weiteren ist die relative Häufigkeit

$$f(X = x_j) = f(x_j) = f_j = \frac{h_j}{\sum_{j=1}^{m} h_j} = \frac{h_j}{n}$$

von Bedeutung, die den prozentualen Anteil der absoluten Häufigkeit an der Gesamtzahl wiedergibt. Die absolute Häufigkeit h_j wird durch die Gesamtanzahl der Fälle n dividiert. Für x_1 bedeutet das

$$f(x_1) = \frac{2}{20} = 0{,}1$$

oder 10% (wenn der Quotient mit 100 multipliziert wird). Die letzte Spalte kumuliert nur die relativen Häufigkeiten einer Realisation zur nächsten. Die Spalten der obigen Tabelle noch einmal in Kürze zusammengefasst:

- x_j Merkmalsrealisation
- $h(x_j)$ absolute Häufigkeit
- $f(x_j)$ relative Häufigkeit
- $F(x_j)$ relative Summenhäufigkeit

Die vorstehende Häufigkeitsverteilung stellt man in der Mathematik auch gerne in Mengen-Schreibweise dar (für die praktische Anwendung nicht zwingend wichtig). Eine Zeile der Häufigkeitsverteilung besteht dann mit der Merkmalsrealisation und der relativen Häufigkeit aus dem Tupel $(x_1, \mathrm{f}(x_1))$. Die gesamte Häufigkeitsverteilung entspricht dann der Menge aller Tupel:

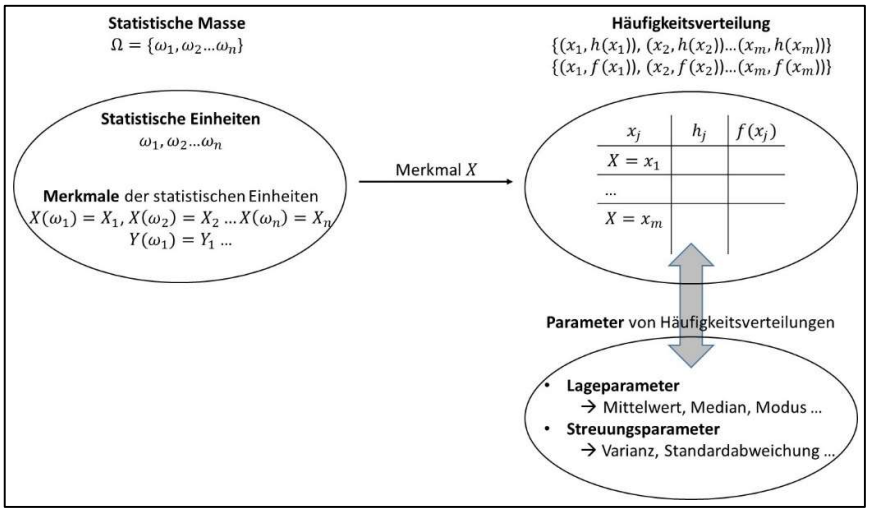
$$\{(x_1, \mathrm{f}(x_1)), \dots\ (x_j, \mathrm{f}(x_j)), \dots\ (x_m, \mathrm{f}(x_m))\}$$

bzw. im konkreten Beispiel

$$\{(0{,}61; 0{,}10); (0{,}92; 0{,}20); \dots; (0{,}68; 0{,}15)\}$$

2.1.2.4 Zusammenfassung

Die bereits bekannte Grafik veranschaulicht noch einmal den Weg von der Urliste zur Häufigkeitsverteilung einschließlich aller Bezeichnungen.



2.1.2.5 Skalen

Der letzte Abschnitt der Definitionen befasst sich mit den Skalen. Grundsätzlich ist eine Skala nichts weiter als eine Menge von Zahlen oder Zeichen, die die Merkmalswerte darstellt. Das können zum Beispiel bei dem Merkmal Gehalt einfach Ziffern sein; beim Merkmal Geschlecht kann es entweder der Text „männlich; weiblich“, die Buchstaben „m; w“ oder auch Ziffern „1; 2“ sein.

Skalenniveau

Eine Möglichkeit der Differenzierung von Skalen ist die nach dem Skalenniveau. Man unterscheidet dabei die folgenden Optionen.

- Nominalskala

  Eine Nominalskala differenziert ausschließlich nach der Bezeichnung der Merkmalsausprägung. Ein typisches Beispiel ist der Familienstand der als Text (ledig, verheiratet, verwitwet, geschieden) oder auch als Schlüssel (0; 1; 2; 3) angegeben werden kann. Auch wenn ein Schlüssel numerischer Natur ist, kann man hier keine Reihenfolge oder Hierarchie bilden. Auch die Berechnung eines Mittelwertes wäre unsinnig. Man kann bei diesem Merkmal ausschließlich die Verschiedenheit der Merkmalswerte feststellen.

- Ordinalskala

  Demgegenüber kann bei einer Originalskala eine Rangreihenfolge gebildet werden. Typische Beispiele sind „Windstärken“ und „Hotel-Klassifizierungen mittels Sternen“. Es lässt sich hier festhalten, dass ein 3-Sterne-Hotel besser ist als ein 2-Sterne-Hotel. Ob allerdings die erste Kategorie doppelt so gut ist wie die zweite lässt sich nicht sagen.

- Intervallskala

  Der bei der Ordinalskala angesprochene Mangel weist die Intervallskala nicht mehr auf. Ein typisches Beispiel hierfür ist die Temperatur, bei der sich eindeutig sagen lässt, dass 20 Grad doppelt so warm ist wie 10 Grad. Allerdings lässt sich die Temperatur sowohl in Celsius als auch in Fahrenheit messen, d.h. es existiert bei beiden Skalen kein eindeutig definierter Nullpunkt.

- Verhältnisskala

  Der Mangel des nicht eindeutig definierten Nullpunktes entfällt bei der Verhältnisskala. Typische Beispiele sind Längen – oder Gewichtsmaße, bei denen der Wert 0 immer eindeutig keine Länge bzw. kein Gewicht beschreibt.

Die Intervall- und Verhältnisskalen werden auch als metrische oder kardinale Skalen bezeichnet.

Eine weitere Differenzierung, die allerdings nur kardinale Skalen betrifft, ist die in stetige und diskrete Merkmale.

- Diskrete Merkmale

  Diskrete Merkmale zeichnen sich dadurch aus, dass nur bestimmte Merkmalswerte auftreten können; Zwischenwerte sind ausgeschlossen. Ein typisches Beispiel hierfür ist die Anzahl von Kindern in einer Familie, die mit 0 beginnend nur ganzzahlige Werte annehmen kann.

- Stetige Merkmale

  Demgegenüber können stetige Merkmale – zumindest in einem Intervall – jeden beliebigen Wert annehmen. Die Körpergröße ist hierfür ein Beispiel, die bei exakter Messung mit mehreren Nachkommastellen jeden Wert innerhalb eines Bereiches annehmen kann.

Anmerkung zur Praxis

- Da die Daten sehr häufig in Form von Schlüsseln gespeichert werden, lassen sich mit statistischer Software bei fast allen Merkmalen Parameter berechnen, auch wenn dieses wie oben gezeigt unsinnig ist (s. Mittelwert beim Familienstand). Aus der Perspektive der Berechenbarkeit von Parametern sind metrischen Skalen wünschenswert, weil sich mit nominal und original skalierten Daten nur wenige statistische Berechnungen durchführen lassen. Der Anwender muss sich daher vor jeder Berechnung über das zugrunde liegende Skalenniveau Gedanken machen.
- Ein weiteres Problem mit dem Skalenniveau wird hingegen in Praxis häufig zu Gunsten der Berechenbarkeit gelöst.

2. Wie stehen Sie zu folgenden Aussagen Ihre Gemeinde/ nähere Umgebung betreffend?	stimme voll und ganz zu	stimme eher zu	Unentschieden	stimme eher nicht zu	stimme überhaupt nicht zu
Meine Gemeinde ist kinderfreundlich	❑	❑	❑	❑	❑
Ich fühle mich über Angebote für Familien gut informiert	❑	❑	❑	❑	❑

Der hier abgebildete Auszug aus einem Fragebogen ermöglicht den Probanden eine Einschätzung auf einer Skala von „sehr schlecht“ bis „sehr gut“ abzugeben, wobei die Verschlüsselung mit den Ziffern 1 bis 5 erfolgt. Damit handelt es sich um ein ordinales Skalenniveau, da zwar eine Reihenfolge angegeben werden kann, aber nicht geklärt ist, ob der Abstand zwischen den Merkmalswerten immer identisch also äquidistant ist. Unterstellt man aber diese Äquidistanz liegt ein metrischen Skalenniveau vor und es lassen sich die weiteren statistischen Berechnungen ohne Probleme durchführen. Damit folglich weiterführende statistische Berechnung durchgeführt werden können, werden in der Praxis sehr häufig ordinale Skalen als metrische Skalen interpretiert.

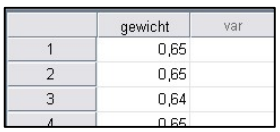
2.1.3 SPSS

Soll die Aufgabe mit den Gewicht von Tabletten mittels SPSS gelöst werden, so ist wie folgt vorzugehen:

- Einrichtung der Variablen, die der Erfassung des Merkmals *X* dienen
- Eingabe der Daten
- Auswertung

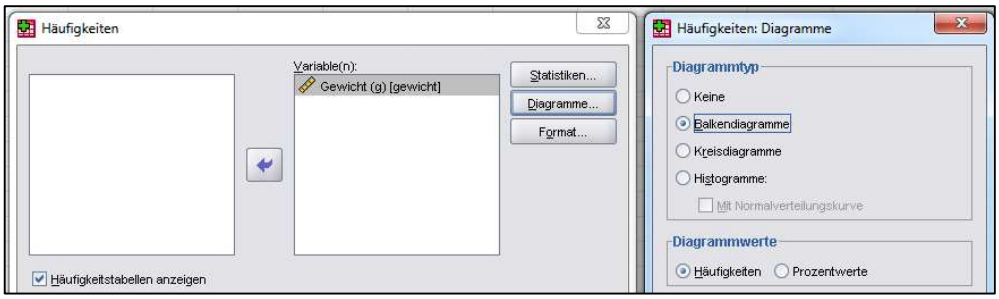
Einrichtung der SPSS-Datei

Datenerfassung



Menü „Analysieren“ -> „Deskriptive Statistik“ -> „Häufigkeiten“

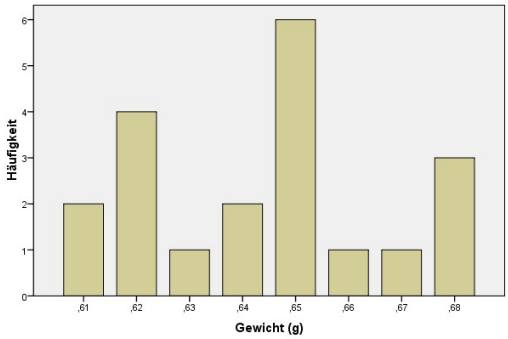
- Auswahl der Variablen & Button „Diagramme“



- Ausgabe

Gewicht (g)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	,61	2	10,0	10,0	10,0
	,62	4	20,0	20,0	30,0
	,63	1	5,0	5,0	35,0
	,64	2	10,0	10,0	45,0
	,65	6	30,0	30,0	75,0
	,66	1	5,0	5,0	80,0
	,67	1	5,0	5,0	85,0
	,68	3	15,0	15,0	100,0
	Gesamt	20	100,0	100,0	



2.2 Klassierte Häufigkeitsverteilungen

Die Klassierung von Daten hatte in der Vergangenheit im Wesentlichen zwei Gründe.

- Zum einen war es der deutlich **reduzierte Rechenaufwand** bei klassierten Daten und
- zum anderen eine **verbesserte Aussagefähigkeit** bei einer geringeren Anzahl von Merkmalsrealisationen.

Das erste Argument hat heute aufgrund der softwaregestützten Auswertung nicht mehr die Relevanz wie bei der früher üblichen manuellen Auswertung. Es bleibt nur das Argument der Übersichtlichkeit, wenn eine große Anzahl von Merkmalsrealisationen auf eine geringere Anzahl von Merkmalsrealisations-Klassen reduziert wird. Grundsätzlich kann dieses Argument in den folgenden Situationen von zum Tragen kommen:

- Bei stetigen Merkmalen ist die Klassierung für eine übersichtliche Auswertung fast immer erforderlich, da bei stetigen Merkmalen eben durch die Stetigkeit eine Vielzahl von Merkmalsrealisationen möglich sind.
- Bei diskreten Merkmalen wird die Klassierung dann notwendig, wenn eine hohe Anzahl von Merkmalsrealisation zur Unübersichtlichkeit führt. Bei welcher die Anzahl von Realisation dieses der Fall ist, ist – wie immer bei solchen Fragestellungen – nicht eindeutig definiert. Die Auffassung des Autors hierzu ist: mehr als acht Merkmalsrealisation sind nicht mehr übersichtlich.
- Ein Nachteil der Klassierung besteht in der Verfälschung der Berechnungen. Bei dem Beispiel „Körpergrößen“ könnte man beispielsweise Klassen von 5 cm Breite einführen; d.h. von 152,5 cm bis 157,5 cm usw. Die Berechnungen werden jetzt statt mit den exakten Werten mit den Klassenmitten durchgeführt; also 155 cm. Es ist sofort ersichtlich, dass sich Verfälschungen z.B. bei der Berechnung des Mittelwertes ergeben.

Da eine Klassierung häufig für die grafischen Auswertung erforderlich ist, geht man in der Praxis den Weg der Umkodierung in eine neue Variable. Dabei behält man die alten Werte bei und codiert sie z.B. mit dem Klassenmittel um. Die Grafiken können dann mit der klassierten Werten und die Berechnungen mit der exakten Werten durchgeführt werden.

Bei der nachstehenden Häufigkeitsverteilung wurde die Körpergröße von 100 Probanden ermittelt. Eine Häufigkeitstabelle mit jeder auftretenden Merkmalsrealisation wäre sehr unübersichtlich. Daher wurden Klassen mit einer Klassenbreite von 5 cm gebildet:

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	153,00	1	1,0	1,0	1,0
	154,00	1	1,0	1,0	2,0
	155,00	2	2,0	2,0	4,0
	156,00	3	3,0	3,0	7,0
	157,00	3	3,0	3,0	10,0
	158,00	5	5,0	5,0	15,0
	159,00	6	6,0	6,0	21,0
	160,00	4	4,0	4,0	25,0
	161,00	5	5,0	5,0	30,0
	162,00	7	7,0	7,0	37,0
	163,00	5	5,0	5,0	42,0
	164,00	5	5,0	5,0	47,0
	165,00	6	6,0	6,0	53,0
	166,00	7	7,0	7,0	60,0
	167,00	5	5,0	5,0	65,0
	168,00	4	4,0	4,0	69,0
	169,00	5	5,0	5,0	74,0
	170,00	5	5,0	5,0	79,0
	171,00	6	6,0	6,0	85,0
	172,00	4	4,0	4,0	89,0
	173,00	3	3,0	3,0	92,0
	174,00	2	2,0	2,0	94,0
	175,00	3	3,0	3,0	97,0
	176,00	1	1,0	1,0	98,0
	177,00	1	1,0	1,0	99,0
	178,00	1	1,0	1,0	100,0
	Gesamt	100	100,0	100,0	

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	155,00	10	10,0	10,0	10,0
	160,00	27	27,0	27,0	37,0
	165,00	28	28,0	28,0	65,0
	170,00	24	24,0	24,0	89,0
	175,00	10	10,0	10,0	99,0
	180,00	1	1,0	1,0	100,0
	Gesamt	100	100,0	100,0	

2.2.1 Mathematische Definitionen

Klassierung

Die Zusammenfassung von vielen voneinander unterschiedlichen Merkmalswerten eines diskreten bzw. stetigen Merkmals heißt Klassierung.

Anmerkungen

- Bei einer großen Zahl von Merkmalsausprägungen wird eine Häufigkeitsverteilung unübersichtlich.
  Beispiel: ALTER
- Bei stetigen Merkmalen ist eine Häufigkeitsverteilung nicht mehr möglich, da die absolute Häufigkeit je jede Merkmalsausprägung nur sehr gering ist.
  Beispiel: GEWICHT

Klasse

Es sei X ein kardinalskaliertes Merkmal. Dann heißt für alle $k = 1, 2, \dots, m$ das geordnete Merkmalswerte-Intervall $x_k^u \leq X \leq x_k^o$ Klasse.

Anmerkungen

- Der Index k bestimmt die Klasse. D.h. die erste Klasse wird als Klasse x_1 usw. bezeichnet
- Die Indices $u; o$ stehen für die Unter- bzw. Obergrenze der Klasse.

Klassenbreite

Die Differenz $\Delta x_k = x_k^o - x_k^u$ von oberer und unterer Klassengrenze heißt Klassenbreite.

Anmerkungen

- Um eine Vergleichbarkeit bei der statistischen Analyse von Daten zu ermöglichen, ist es sinnvoll, die Klassenbreite für alle Klasse konstant zu halten.
- Grundsätzlich muss die Breite aller Klassen nicht gleich sein; ist dieses aber der Fall wird von äquidistanten Klassen gesprochen.

Klassenmittel

Das einfache arithmetische Mittel

$$\bar{x}_k = \frac{\sum_{i=1}^{h_k} x_{ik}}{h_k}$$

der Merkmalswerte, die zu einer Klasse gehören, heißt Klassenmittel.

Anmerkungen

- h_k ist als die Anzahl der Elemente einer Klasse definiert.
- Es werden demzufolge alle Merkmalswerte einer Klasse addiert und durch die Anzahl dividiert (klassisches arithmetisches Mittel).
- Es handelt sich hierbei um das arithmetische Mittel einer Klasse.

Klassenmitte

Das einfache arithmetische Mittel

$$x_k^* = \frac{x_k^o + x_k^u}{2}$$

aus der oberen und der unteren Klassengrenze heißt Klassenmitte.

Anmerkungen

- Die Klassenmitte wird häufig als Repräsentant für die Klasse verwendet.
- Die Klassenmitte nimmt keine Rücksicht auf die Verteilung der Werte innerhalb der Klasse, wie z.B. das Klassenmittel.

Klassenhäufigkeit

Die Anzahl der Elemente. Deren Merkmalswerte in eine bestimmte Klasse fallen, heißt absolute Häufigkeit.

$$h(x_k^*) = h_k = h(x_k^u \leq X \leq x_k^o)$$

Analog gilt für die relative Häufigkeit

$$f(x_k^*) = f_k = f(x_k^u \leq X \leq x_k^o)$$

Häufigkeitsdichte

Der Quotient aus der absoluten bzw. der relativen Häufigkeit und der Breite einer Klasse

$$h^D(x_k^*) = h_k^D = \frac{h_k}{\Delta x_k} \text{ bzw. } f_k^D = \frac{f_k}{\Delta x_k}$$

heißt Häufigkeitsdichte.

2.2.2 Herleitungen & Berechnungen

Die Erläuterungen dieses Kapitels basieren auf einem Beispiel, bei dem in einer empirischen Erhebung u. a. die Körpergrößen von 100 Probanden ermittelt wurden. Die Urliste hat somit die folgende Struktur:

$i = 1 \dots 100$	X : Größe	Y : Gewicht	…
ω_1	$X(\omega_1) = 161$		
ω_2	$X(\omega_2) = 162$		
ω_3	$X(\omega_3) = 166$		
ω_4	$X(\omega_4) = 161$		
ω_5	$X(\omega_5) = 171$		
…			

Da hier nur das Merkmal „Körpergröße“ von Bedeutung ist, sind die Größen gemessen in cm in platzsparender Form nachstehend zusammengefasst.

161	162	166	161	171	159	160	174	165	163
161	178	157	156	160	172	167	162	164	156
177	162	167	168	157	164	176	166	171	169
171	155	170	158	171	167	161	172	169	161
160	164	162	170	168	165	173	159	173	166
170	154	165	162	174	158	156	165	160	165
172	167	173	166	164	168	175	158	163	169
171	166	159	162	159	171	163	158	167	168
163	153	172	170	158	164	162	175	165	169
170	155	169	159	163	159	166	157	166	175

Die unten stehende Häufigkeitsverteilung macht das Problem der Unübersichtlichkeit aufgrund der Vielzahl von Merkmalsrealisationen deutlich. Ein Problem, das bei stetigen Merkmalen typisch ist.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	153,00	1	1,0	1,0	1,0
	154,00	1	1,0	1,0	2,0
	155,00	2	2,0	2,0	4,0
	156,00	3	3,0	3,0	7,0
	157,00	3	3,0	3,0	10,0
	158,00	5	5,0	5,0	15,0
	159,00	6	6,0	6,0	21,0
	160,00	4	4,0	4,0	25,0
	161,00	5	5,0	5,0	30,0
	162,00	7	7,0	7,0	37,0
	163,00	5	5,0	5,0	42,0
	164,00	5	5,0	5,0	47,0
	165,00	6	6,0	6,0	53,0
	166,00	7	7,0	7,0	60,0
	167,00	5	5,0	5,0	65,0
	168,00	4	4,0	4,0	69,0
	169,00	5	5,0	5,0	74,0
	170,00	5	5,0	5,0	79,0
	171,00	6	6,0	6,0	85,0
	172,00	4	4,0	4,0	89,0
	173,00	3	3,0	3,0	92,0
	174,00	2	2,0	2,0	94,0
	175,00	3	3,0	3,0	97,0
	176,00	1	1,0	1,0	98,0
	177,00	1	1,0	1,0	99,0
	178,00	1	1,0	1,0	100,0
	Gesamt	100	100,0	100,0	

Daten und Merkmale

Als statistische Einheit ω ist hier die befragte Person anzusehen, die statistische Gesamtheit Ω ist dann die Menge aller 100 Probanden; somit gilt $n = 100$. Das für diese Untersuchung relevante Merkmal X ist die Körpergröße. Die Merkmalswerte $X(\omega_i) = X_i$, was für ω_1 bedeutet $X(\omega_1) = X_1 = 161$. Bezogen auf die Gesamtheit gilt für den Index i: $i = 1\ bis\ 100$.

Klassierung

Beim vorliegenden Beispiel wurde Klassenbreite von 5 cm angenommen und die Klassen wie folgt gebildet:

- Klasse: 152,5 bis 157,5 Klassenmitte: 155,0
- Klasse: 157,5 bis 162,5 Klassenmitte: 160,0
- …

Die Klassen müssen wie üblich indiziert werden und man verwendet zur Differenzierung von den anderen Indices hier die Variable k.

Die Klassenbreite von 5 cm wurde gewählt, da hierdurch eine übersichtliche Anzahl von Klassen bei der Häufigkeitsverteilung entsteht. Da alle Intervalle gleichlang sind, gilt für alle k, dass die Differenz zwischen Ober- und Untergrenze immer identisch ist:

$$\Delta x_k = x_k^O - x_k^U = 5$$

Dieses ist nicht immer der Fall, da gerne bei der ersten und der letzten Merkmalsrealisation offene Grenzen gewählt werden. So wäre hier bei der ersten Realisation auch ein Intervall von „bis 157,5 cm“ und bei der letzten Realisation ein Intervall von „mehr als 177,5 cm“ denkbar. Bei dieser Art der Intervallbildung ist darauf zu achten, dass die Randintervalle nicht einen zu großen Wertebereich umfassen, um Fehler bei der Interpretation zu vermeiden.

Die Intervalle selbst sind so gelegt, dass die Grenzen immer zwischen zwei Merkmalswerten liegen und so Grenzfälle vermieden werden. Dieses wäre der Fall, wenn ein Merkmalswert direkt auf eine Intervallgrenze fällt. Tritt ein solcher Fall auf, wird der Wert je zur Hälfte in beiden Klassen berücksichtigt. Bei manuellen Berechnungen führt dieses aber zu einem erheblichen Rechenaufwand und sollte daher vermieden werden.

Klassenmittel

Das Klassenmitte ist mathematisch nichts anderes als das bekannte arithmetische Mittel, jedoch bezogen auf die Werte der jeweiligen Klasse. Von daher wird die Variable um den Index k die jeweilige Klasse erweitert.

$$\bar{x}_k = \frac{\sum_{i=1}^{h_k} x_{ik}}{h_k}$$

Für Interpretationen spielt dieser Parameter aber nur eine untergeordnete Rolle und soll daher nicht weiter vertieft werden.

Klassenmitte

Die Klassenmitte ist gegenüber dem Klassenmittel aber relevant für die Häufigkeitsverteilung. Hierbei handelt es sich um den Mittelwert zwischen der Ober- und der Untergrenze. Die Klassenmitte wird zur Differenzierung mit dem Akzent $*$ und dem Index k für die entsprechende Klasse versehen.

$$x_k^* = \frac{x_k^o + x_k^u}{2}$$

Für die erste Klasse bedeutet dieses:

$$x_1^* = \frac{157{,}5 + 152{,}5}{2} = 155$$

Klassenhäufigkeit

Die Definitionen für die absolute und die relative Häufigkeit gelten analog zu den nicht klassierten Verteilungen; der Unterschied besteht mit x_k^* statt x_j im Argument. Der Ausdruck $h(x_k^u \leq X \leq x_k^o)$ beschreibt das Zählen aller Merkmalswerte X_i innerhalb der Intervallgrenzen des $k-ten$ Intervalls.

$$h(x_k^*) = h_k = h(x_k^u \leq X \leq x_k^o)$$

$$f(x_k^*) = f_k = f(x_k^u \leq X \leq x_k^o)$$

Häufigkeitsverteilung

In diesem Beispiel gilt für den Index k mit $k = 1\ bis\ 6$. In der Tabelle stellen die Spalten folgendes dar:

- $\Delta x_k = x_k^O - x_k^U$: Klassenbreite mit Obergrenze – Untergrenze
- x_k^* : Klassenmitte
- $h(x_k^*)$: absolute Häufigkeit
- $f(x_k^*)$: relative Häufigkeit
- $F(x_k^*)$: relative Summenhäufigkeit

$\Delta x_k = x_k^O - x_k^U = 5$	x_k^*	$h(x_k^*)$	$f(x_k^*)$	$F(x_k^*)$
152,5 - 157,5	155,00	10	10,0	10,0
157,5 - 162,5	160,00	27	27,0	27,0
162,5 - 167,5	165,00	28	28,0	28,0
167,5 - 172,5	170,00	24	24,0	24,0
172,5 - 177,5	175,00	10	10,0	10,0
177,5 - 182.5	180,00	1	1,0	1,0
	Gesamt	100	100,0	100,0

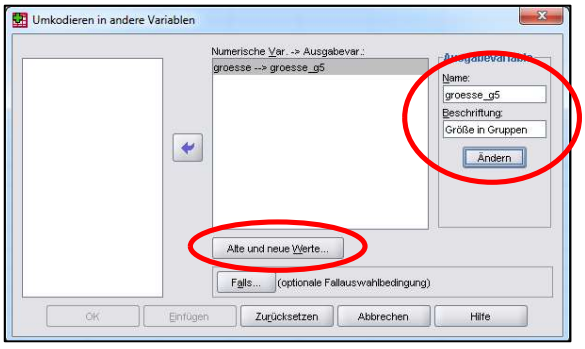
2.2.3 SPSS - Umkodieren von Variablen

Um in SPSS Klassen zu bilden, müssen die Variablen umkodiert werden. D.h. die alten Werte werden neuen Werten nach einer bestimmten Vorschrift zugeordnet. Von den beiden zur Verfügung stehenden Varianten, „Umkodieren in dieselbe Variable“ oder „Umkodierung in eine andere Variable“, sollte immer die Variante „in eine andere Variable“ gewählt werden, da sonst die Ursprungswerte verloren gehen.

In diesem Fall ist es sinnvoll, die Werte den jeweiligen Klassenmitten x_k^* zuzuordnen. Ferner sind hier die umkodierten Werte einer neuen Variablen zuzuweisen, damit die Ausgangswerte erhalten bleiben.

- **Zielvariable definieren**

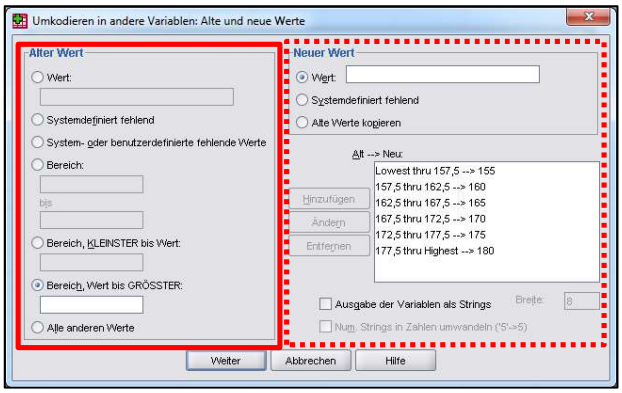
  Der Namen und die Beschriftung der Zielvariablen wird in den entsprechenden TextBoxen eingegeben und anschließend die über den Butten „Ändern“ bestätigt.



  Mittels des Button „Alte und neue Werte“ gelangt zu den Regeln der Umkodierung.

- **Regeln für Umkodierung festlegen**

  Das Fenster zum Umkodieren ist in zwei zentrale Frames unterteilt: den „Alten Werten“ auf der linken Seite (durchgezogen Rot) und den „Neuen Werten“ (gestrichelt Rot) auf der rechten Seite. Alle eingegebenen Regeln werden auf der rechten Seite angezeigt. Man kann diese Regeln mittels der entsprechenden Button hinzufügen, ändern oder löschen.



  Mittels des Button „Weiter“ werden die dann Regeln angewendet.

- **Ergebnis im Dateneditor**

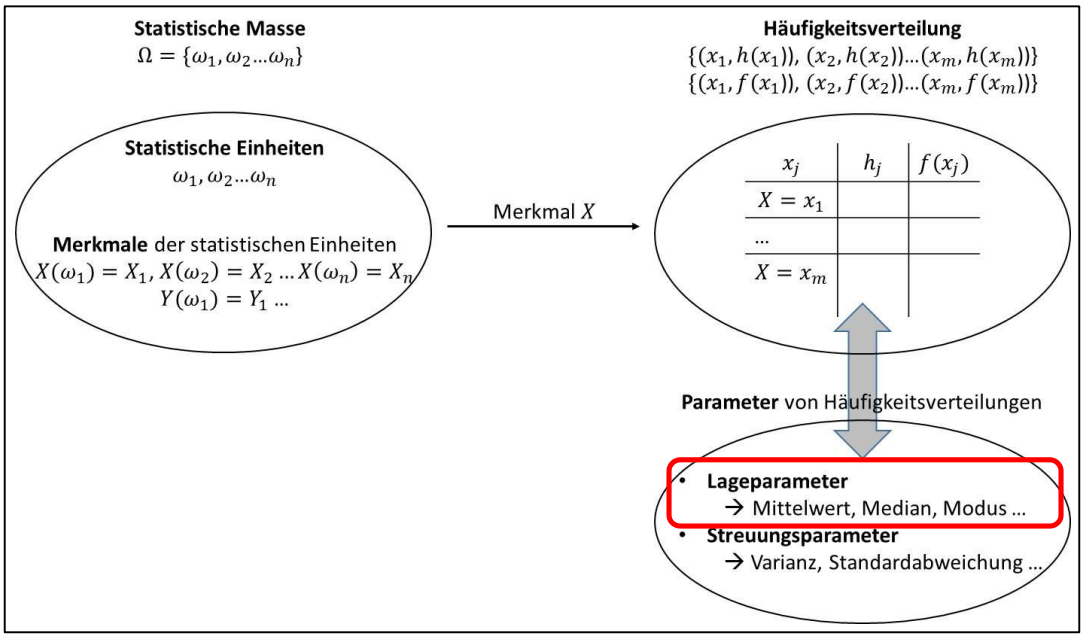
	groesse	groesse_g5
1	161,00	160,00
2	161,00	160,00
3	177,00	175,00
4	171,00	170,00
5	160,00	160,00
6	170,00	170,00
7	172,00	170,00
8	171,00	170,00
9	163,00	165,00
10	170,00	170,00
11	162,00	160,00

2.3 Lageparameter

Häufigkeitsverteilungen sind zwar übersichtlicher als Urlisten, lassen sich aber trotzdem schlecht interpretieren und ggf. vergleichen. Von daher benötigt man Kennzahlen zur Charakterisierung der Verteilung.

Lageparameter sind eine Klasse von Kennzahlen, die die Struktur der Häufigkeitsverteilung, d.h. die Lage der Verteilung - ähnlich der Kurvendiskussion in der Analysis - beschreiben. Durch die Einordnung in die bekannte Grafik wird noch einmal deutlich, dass die Lageparameter eine komprimierte Beschreibung der Häufigkeitsverteilung liefern. Ein typisches Beispiel für einen Lageparameter ist der Mittelwert. Er beschreibt beispielsweise bei einer Klausur wie viel Punkte die Studierenden im Mittel erreicht haben.

Der Nachteil des Mittelwertes ist aber, dass er nichts über sein Zustandekommen aussagt. Ein Mittelwert von 3,0 kann sich dadurch ergeben, dass alle Studierenden eine 3,0 geschrieben haben oder aber dadurch, dass die eine Hälfte eine 1,0 und die andere Hälfte eine 5,0 geschrieben hat. An diesem Beispiel wird deutlich, dass zu den Lageparametern weitere Kennzahlen, man nennt sie Streuungsparameter, hinzukommen müssen.



Nachstehend das Beispiel, das im Folgenden detailliert besprochen wird.

0,65		x_j	$h(x_j) = h_j$	$f(x_j)$	$F(x_j)$
0,65		0,60	0	0,00	0%
0,64		0,61	2	0,10	10%
0,65		0,62	4	0,20	30%
0,65		0,63	1	0,05	35%
0,62		0,64	2	0,10	45%
0,63		0,65	6	0,30	75%
0,61		0,66	1	0,05	80%
0,65		0,67	1	0,05	85%
0,64		0,68	3	0,15	100%
0,65		0,69	0	0,00	100%
0,68		0,70	0	0,00	100%
0,66			20	1,00	
0,67					
0,62		***Lageparameter***			
0,61		Mittelwert		0,644	
0,62		Median		0,650	
0,68		Modus		0,650	
0,68					
0,62					

2.3.1 Mathematische Definitionen

Quantile

Seien $X^*(\omega_i) = X_i^*$ die auf- oder absteigend sortierten Merkmalswerte $X(\omega_i) = X_i$. Sei ferner X ein mindestens ordinalskaliertes Merkmal. Dann heißt der Merkmalswert x_q, die die aufsteigend geordnete Folge aller beobachteten Merkmalswerte X_i derart zweiteilt, dass q Anteile unterhalb und $1 - q$ Anteile oberhalb der Merkmalsausprägung x_q liegen, wobei $0 \leq q \leq 1$ gilt, Quantil.

Anmerkungen

- Quantile teilen die geordneten Elemente der Gesamtheit in gleich große Teile auf. Die nachstehende Tabelle zeigt die in der Statistik üblichen Quantile.

q	*Benennung*
0,50	Median
0,25	1. Quartil
0,50	2. Quartil
0,75	3. Quartil
0,10	1. Dezil
0,20	2. Dezil
…	…
0,90	9. Dezil
0,01	1. Perzentil
0,02	2. Perzentil
…	…
0,99	99. Perzentil

Median (Zentralwert)

Es sei X ein mindestens ordinalskaliertes Merkmal. Dann heißt die Merkmalswert x_q, mit $q = 0{,}5$, die die aufsteigend geordnete Folge aller beobachteten Merkmalsausprägungen derart zweiteilt, dass 50% unterhalb und 50% oberhalb der Merkmalsausprägung, liegen Median:

$$\bar{x}_Z = x_{0,5}$$

und ist definiert durch

$$\bar{x}_Z = \begin{cases} X_{(\frac{n+1}{2})} \text{ für n ungerade} \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}) \text{ für n gerade} \end{cases}$$

Modus (Häufigster Wert)

Der Merkmalswert eines beliebig skalierten Merkmals X, die am häufigsten auftritt, heiß Modus $\bar{x}_M$.

- Für nichtklassierte Daten gilt
  $\bar{x}_M$: dasjenige x_j, für das $h(x_j)$ maximal ist
- Bei klassierten Daten gilt:
  $\bar{x}_M$: dasjenige x_k^* , für das $h(x_k^*)$ maximal ist

Anmerkungen

- Der Modus ist nur dann definiert, wenn genau ein häufigster Wert existiert.

Arithmetisches Mittel

Es sei X ein kardinalskaliertes Merkmal. Dann heißt der Wert, der sich ergibt, wenn für alle $i = 1, 2, 3, \dots, n$ die Summe der beobachteten Merkmalswerte X_i gleichmäßig auf alle Merkmalsträger verteilt wird, arithmetisches Mittel.

$$\bar{x} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{j=1}^{m} h_j * x_j \; = \sum_{j=1}^{m} f_j * x_j$$

Anmerkungen

- Das arithmetische Mittel lässt sich bei großen Datenmengen vereinfacht mittels Häufigkeitsverteilung berechnen.
- Da bei klassierten Daten nicht mit den exakten Werten x_k sondern stellvertretend mit der Klassenmitte x_k^* gerechnet wird, können sich bei der Berechnung des arithmetischen Mittels je nach Berechnungsverfahren unterschiedliche Werte ergeben. Haben alle Klassen die gleiche Breiten $\Delta x = \Delta x_1 = \Delta x_2 \dots = \Delta x_m$ liegt die maximale Abweichungen der beiden Werte bei

$$\left| \frac{1}{n}\sum_{j=1}^{m} h_j * x_j \; - \frac{1}{n}\sum_{j=1}^{m} h_j * x_j^* \right| \leq \frac{\Delta x}{2}$$

Geometrisches Mittel

Es sei X ein verhältnisskaliertes Merkmal. Dann heißt der Wert, der sich aus der Wurzel des Produktes aller Merkmalswerte geometrisches Mittel.

$$\bar{\mathrm{x}}_{\mathrm{G}} = \sqrt[\mathrm{n}]{\mathrm{X}_1 * \mathrm{X}_2 * \mathrm{X}_3 * \dots * \mathrm{X}_\mathrm{n}} = \sqrt[n]{x_1^{h_1} * x_2^{h_2} * x_3^{h_3} * \dots * x_m^{h_m}}$$

Anmerkungen

- Die Bestimmung des geometrischen Mittels ist nur dann sinnvoll, wenn die Merkmalswerte geometrisch mit einander verknüpft sind. Dieses ist z.B. bei der Entwicklung eines Kapitals gegeben, bei dem sich dieses um einen Faktor $q = 1 + \frac{p}{100}$ vermehrt.

Gewichtetes (arithmetisches) Mittel

Es sei X ein kardinalskaliertes Merkmal. Dann heißt die Summe für alle $i = 1, 2, 3, \dots, n$ der beobachteten Merkmalswerte X_i gewichtet mit dem Gewicht g_i gewichtetes arithmetisches Mittel $\bar{x}_{gew}$:

$\bar{x}_{gew} = \sum_{i=1}^{n} X_i * g_i$ mit $g_i > 0$ für alle i und $\sum_{i=1}^{n} g_i = 1$

oder

$\bar{x}_{gew} = \frac{\sum_{i=1}^{n} X_i * g_i}{\sum_{i=1}^{n} g_i}$ mit $g_i > 0$ für alle i.

Anmerkungen

- Da bei dieser Vorgehensweise jedes Element X_i der Gesamtheit mit einem anderen Gewicht bewertet wird, ist eine Berechnung auf Basis der Merkmalsrealisation x_j der Häufigkeitsverteilung nicht möglich.

2.3.2 Herleitungen & Berechnungen

Die Erläuterungen erfolgen anhand des bekannten Beispiels mit dem Gewicht der Tabletten, die noch einmal als Urliste und als Häufigkeitsverteilung aufgeführt sind.

Urliste

X_i	Gewicht
X_1	0,65
X_2	0,65
X_3	0,64
X_4	0,65
X_5	0,65
X_6	0,62
X_7	0,63
X_8	0,61
X_9	0,65
X_{10}	0,64
X_{11}	0,65
X_{12}	0,68
X_{13}	0,66
X_{14}	0,67
X_{15}	0,62
X_{16}	0,61
X_{17}	0,62
X_{18}	0,68
X_{19}	0,68
X_{20}	0,62

Häufigkeitsverteilung

x_j	$h(x_j) = h_j$	$f(x_j)$	$F(x_j)$
0,60	0	0,00	0%
0,61	2	0,10	10%
0,62	4	0,20	30%
0,63	1	0,05	35%
0,64	2	0,10	45%
0,65	6	0,30	75%
0,66	1	0,05	80%
0,67	1	0,05	85%
0,68	3	0,15	100%
0,69	0	0,00	100%
0,70	0	0,00	100%
Summe	20	1,00	

Berechnung des arithmetischen Mittels

Die Parameter einer Verteilung können immer auf verschiedenen Wegen - anhand der Uhr Liste, anhand der absoluten Häufigkeiten und anhand der relativen Häufigkeiten - berechnet werden.

$$\bar{x} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{j=1}^{m} h_j * x_j \; = \sum_{j=1}^{m} f_j * x_j$$

Der erste Teil des Ausdrucks

$$\bar{x} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{\mathrm{n}}$$

stellt die Berechnung mit den Werten der Urliste dar. Dieser Quotient beschreibt etwas umständlich im Zähler die Summation alle Werte der Urliste, die dann durch die Anzahl der Werte dividiert wird.

Der zweite Teil des Ausdrucks

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

ist die Kurzschreibweise des Zählers in Form des Summenzeichens und ist für einige Leser eventuell etwas ungewohnt. Das Summenzeichen $\sum_{i=1}^{n} X_i$ ist eine bequeme Schreibweise für die Summation $X_1 + X_2 + X_3 + \cdots X_n$ und beschreibt, dass die Summe aller X_i beginnend mit $i = 1$, also X_1 und endend mit $i = n$, also X_n gebildet werden soll.

Der dritte Ausdruck $\frac{1}{n}\sum_{j=1}^{m} h_j * x_j$ verwendet die absoluten Häufigkeiten h_j der Häufigkeitsverteilung, die mit der jeweiligen Merkmalsrealisation multipliziert und für alle Realisation summiert wird. Für $x_2 = 0{,}61$ mit $h(x_2) = 2$ ergibt sich damit der Wert 1,22, was der Summe der beiden Merkmalswerte $X_8 = 0{,}61$ und $X_{16} = 0{,}61$ aus der Urliste entspricht. Vereinfacht: ob man $0{,}61 + 0{,}61$ oder

$0{,}61 * 2$ rechnet ist für das Ergebnis unerheblich. Es ist aber meistens ein deutlich geringer Rechenaufwand, wenn anstatt der Werte der Urliste die Werte der Häufigkeitsverteilung verwendet werden, da die Häufigkeitsverteilung weniger Werte enthält.

Bei dem vierten Ausdruck wird der Quotient $\frac{1}{n}$ in das Summenzeichen gezogen, wodurch sich aus $\frac{1}{n} * h_j = \frac{h_j}{n}$ die relative Häufigkeit f_j ergibt.

$$\frac{1}{n}\sum_{j=1}^{m} h_j * x_j = \sum_{j=1}^{m} \frac{1}{n} * h_j * x_j = \sum_{j=1}^{m} f_j * x_j$$

Alle mathematischen Ansätze führen zu dem gleichen Ergebnis; die Berechnung anhand der Urliste wird bei großen n aber schnell umständlich, was dann für das Arbeiten mit der Häufigkeitsverteilung spricht.

- … anhand der Urliste

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{20} * (0{,}65 + 0{,}65 + 0{,}64 + +065 + \cdots + 0{,}62) = 0{,}644$$

- … anhand der absoluten Häufigkeiten der Häufigkeitsverteilung

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} h_j * x_j = \frac{1}{20} * ((0{,}61 * 2) + (0{,}62 * 4) + \cdots + (0{,}68 * 3) = 0{,}644$$

- … anhand der relativen Häufigkeiten der Häufigkeitsverteilung

$$\bar{x} = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{x_j} * f(x_j) = (0{,}61 * 0{,}1) + (0{,}62 * 0{,}2) + \cdots (0{,}68 * 0{,}15) = 0{,}644$$

Bestimmung des Medians

Bei der Bestimmung des Medians sind die beiden Fälle n gerade bzw. ungerade zu unterscheiden.

$$\bar{x}_Z = \begin{cases} X_{(\frac{n+1}{2})} \text{ für n ungerade} \\ \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}) \text{ für n gerade} \end{cases}$$

Ziel des Medians ist die Bestimmung des mittleren Wertes (nicht des Mittelwertes), der die aufsteigend (oder absteigend) sortierten Werte in die oberen und die unteren 50 % aufteilt. Dieses soll wieder an dem Beispiel „Gewicht von Tabletten“ erläutert werden. Die linke Tabelle zeigt die Urliste in der bekannten Reihenfolge; in der rechten Tabelle sind die Werte aufsteigend sortiert. Die Sortierung der Messwerte ist eine zwingende Voraussetzung für die Bestimmung des Medians.

In diesem Beispiel ist n=20 gerade und damit muss der Median zwischen dem 10. und dem 11. Wert liegen. Die formal korrekte Bestimmung des Medians erfolgt entsprechend der obigen Formel durch

$$\bar{x}_Z = \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}) = \frac{1}{2}(X_{\frac{20}{2}} + X_{\frac{20}{2}+1}) = \frac{1}{2}(X_{10} + X_{11}) = \frac{1}{2}(0{,}65 + 0{,}65) = 0{,}65$$

X_i	Gewicht
X_1	0,65
X_2	0,65
X_3	0,64
X_4	0,65
X_5	0,65
X_6	0,62
X_7	0,63
X_8	0,61
X_9	0,65
X_{10}	0,64
X_{11}	0,65
X_{12}	0,68
X_{13}	0,66
X_{14}	0,67
X_{15}	0,62
X_{16}	0,61
X_{17}	0,62
X_{18}	0,68
X_{19}	0,68
X_{20}	0,62

	X_i	Gewicht
1	X_8	0,61
2	X_{16}	0,61
3	X_6	0,62
4	X_{15}	0,62
5	X_{17}	0,62
6	X_{20}	0,62
7	X_7	0,63
8	X_3	0,64
9	X_{10}	0,64
10	X_1	0,65
11	X_2	0,65
12	X_4	0,65
13	X_5	0,65
14	X_9	0,65
15	X_{11}	0,65
16	X_{13}	0,66
17	X_{14}	0,67
18	X_{12}	0,68
19	X_{18}	0,68
20	X_{19}	0,68

Wäre die Anzahl n ungerade –z.B. $n = 21$, würde die Mittelung zwischen den beiden Werten entfallen und der Median könnte direkt einer Position zuordnet werden:

$$\bar{x}_Z = X_{(\frac{n+1}{2})} = X_{(\frac{21+1}{2})} = X_{11}$$

Für die Interpretation ist festzuhalten, dass sowohl unterhalb als auch oberhalb des Medians $\bar{x}_Z = 0{,}65$ jeweils 50% der Messwerte liegen.

Bestimmung des Modus‘

Für die Bestimmung des Modus‘ ist die Spalte $h(x_j)$ oder $f(x_j)$ relevant. Gesucht ist das x_j mit der größten absoluten oder relativen Häufigkeit. In diesem Fall gilt:

$$\bar{x}_M = 0{,}65$$

Bestimmung des geometrischen Mittels

Diese Berechnung ist bei den hier vorliegenden Daten **nicht sinnvoll**. Die ausführliche Begründung erfolgt im Abschnitt Interpretation (s. Abschnitt 0).

2.3.3 Interpretationen

Mittelwert ⇔ Median ⇔ Modus

Die Parameter Mittelwert, Median und Modus lassen im Zusammenhang betrachtet einige Rückschlüsse auf die Lage der Verteilung zu:

- Symmetrische Verteilung, wenn gilt:
  $\bar{x} = \bar{x}_Z$; also Mittelwert = Median.
- Symmetrische und eingipflige Verteilung, wenn gilt:
  $\bar{x} = \bar{x}_Z = \bar{x}_M$; also Mittelwert = Median = Modus.

Mittelwert ⇔ Median

Im vorherigen Abschnitt wurde die Berechnung des Medians und des Mittelwertes dargestellt. Es stellt sich die Frage, wo der Unterschied zwischen diesen beiden Parametern liegt und wann welcher Parameter sinnvoll ist. Die linke Tabelle zeigt die bisherigen Daten des Beispiels „Gewicht von Tabletten“, zu den bekannten Werten $\bar{x} = 0{,}6440\ und\ \bar{x}_Z = 0{,}6500$ führt. Bei der rechten Tabelle wurde der Messwert $X_{19} = 1.000$ von 0,68 g auf 1.000 g verändert. Hierdurch verändert sich der Mittelwert

drastisch auf $\bar{x} = 50{,}61$; der Median hingegen bleibt unverändert, was zur folgenden Schlussfolgerung führt:

- Beim Mittelwert gehen alle Werte proportional in die Berechnung ein. Bei starker Abweichung eines Messwertes – man bezeichnet diesen als Ausreißer - verleitet der Mittelwert zu einer falschen Interpretation.
- Der Median teilt die aufsteigend sortierten Werte in die unteren und die oberen 50%. Hierbei kommen Ausreißer nicht zur Geltung.

	X_i	Gewicht
1	X_8	0,61
2	X_{16}	0,61
3	X_6	0,62
4	X_{15}	0,62
5	X_{17}	0,62
6	X_{20}	0,62
7	X_7	0,63
8	X_3	0,64
9	X_{10}	0,64
10	X_1	0,65
11	X_2	0,65
12	X_4	0,65
13	X_5	0,65
14	X_9	0,65
15	X_{11}	0,65
16	X_{13}	0,66
17	X_{14}	0,67
18	X_{12}	0,68
19	X_{18}	0,68
20	X_{19}	0,68

Mittelwert	0,6440
Median	0,6500

	X_i	Gewicht
1	X_8	0,61
2	X_{16}	0,61
3	X_6	0,62
4	X_{15}	0,62
5	X_{17}	0,62
6	X_{20}	0,62
7	X_7	0,63
8	X_3	0,64
9	X_{10}	0,64
10	X_1	0,65
11	X_2	0,65
12	X_4	0,65
13	X_5	0,65
14	X_9	0,65
15	X_{11}	0,65
16	X_{13}	0,66
17	X_{14}	0,67
18	X_{12}	0,68
19	X_{18}	0,68
20	X_{19}	1000,00

Mittelwert	50,6100
Median	0,6500

Für die Praxis ist aus diesen Erkenntnissen folgendes zu schließen:

- Sowohl der Mittelwert als auch der Median haben beide ihre Berechtigung. Der Mittelwert ist gut interpretierbar und auch für weitere statistische Berechnungen zwingend erforderlich. Er hat aber den Nachteil, empfindlich auf Ausreißer zu reagieren. Der Median hat den Vorteil der Unempfindlichkeit gegenüber Ausreißern, ist aber für weitere statistische Berechnungen nicht geeignet.
- Der Median ist zum Beispiel im Rahmen von Einkommensanalysen von großer Bedeutung, da er einen anschaulicheren Wert als ein durch sehr hohe Einkommen verzerrten Mittelwert liefert.
- Bei statistischen Untersuchungen sind immer beide Parameter zu berechnen. Weichen die Werte der beiden Parameter erheblich voneinander ab, so ist dieses ein Indiz für Ausreißer. In einem solchen Fall sollten die Werte der Urliste noch einmal genauer analysiert werden.
- Wird der Mittelwert, den man für weitere Berechnung benötigt, durch Ausreißer stark verfälscht, müssen diese ggf. aus dem Datensatz eliminiert werden. Hierdurch wird der Mittelwert realistisch und die weiteren, darauf basierenden Berechnungen aussagefähiger. **Ein solches Vorgehen ist aber auf jeden Fall offen zu legen und im Auswertungsbericht zu dokumentieren.**

Geometrisches Mittel ⇔ Arithmetisches Mittel

Das geometrische Mittel wurde im bisherigen Beispiel vernachlässigt, da immer nur entweder das arithmetische oder das geometrische Mittel berechnet werden kann; beide bei den gleichen Daten zu berechnen ist sinnlos. Welcher der beiden Mittelwerte bei einer Analyse sinnvoll ist, ist abhängig von

der Daten-Struktur. Beim arithmetischen Mittel sind die aufeinander folgenden Daten additiv und bei beim geometrischen Mittel multiplikativ miteinander verknüpft.

Das folgende Beispiel zeigt mit dem Wachstum einer Population eine geometrischen Datenstruktur (in diesem Fall das negative Wachstum). Hier soll aus den Daten die durchschnittliche Wachstumsrate über 10 Jahre ermittelt werden.

BevölkerungEntwicklung (geometrisches Mittel)

Population 85.000.000

Jahr	*Wachstum*	*q*	*Bevölkerung*
1. Jahr	-2,0	0,980	83.300.000,00
2. Jahr	-2,0	0,980	81.634.000,00
3. Jahr	-3,0	0,970	79.184.980,00
4. Jahr	-4,0	0,960	76.017.580,80
5. Jahr	-3,0	0,970	73.737.053,38
6. Jahr	-2,0	0,980	72.262.312,31
7. Jahr	-2,0	0,980	70.817.066,06
8. Jahr	-1,0	0,990	70.108.895,40
9. Jahr	-1,0	0,990	69.407.806,45
10. Jahr	-0,5	0,995	69.060.767,42
arithmetischs Mittel	-2,050	0,979500	69.097.658,22
geomtrisches Mittel		0,979448	69.060.767,42
Wachstum in %		-2,055231	

Der Vergleich macht den Unterschied deutlich. Das arithmetische Mittel der Prozentsätze in der zweiten Spalte ergibt $\bar{x} = -2{,}05$; also ein durchschnittliches Wachstum von -2,05%. Das korrekte Ergebnis, das geometrische Mittel in der dritten Spalte, hingegen führt zu $\bar{x}_G = -2{,}055$; also durchschnittlich -2,055% Wachstum. Dieser Unterschied mag gering erscheinen, führt aber bei dem arithmetischen Mittel zu einem Bevölkerungsbestand von 69.097.658 und bei dem geometrischen Mittel zu 69.060.767.

Voraussetzung für den geometrischen Mittelwert ist eine geometrische Struktur der Daten. Als Exkurs sei hier auf die geometrische Folge verwiesen. Etwas einfacher formuliert liegt eine geometrische Struktur dann vor, wenn man von einem Folgenglied zum nächsten durch die Multiplikation mit einem Faktor, der unterschiedlich sein kann, kommt. Wichtig ist hier die mathematische Operation „Multiplikation“, die von einem Element zum nächsten Element führt, und nicht die der „Addition“. Schaut man sich einmal die untere Tabelle an, so gelangt man bei einem Wachstum von -2% durch den Faktor 0,98 von 85.000.000 zu 83.300.000. Es gilt also

$$85.000.000 * 0{,}98 = 83.300.00$$

Der Faktor 0,98 ergibt sich aus

$$q = 1 - \frac{p}{100} = 1 - \frac{2}{100} = 0{,}98.$$

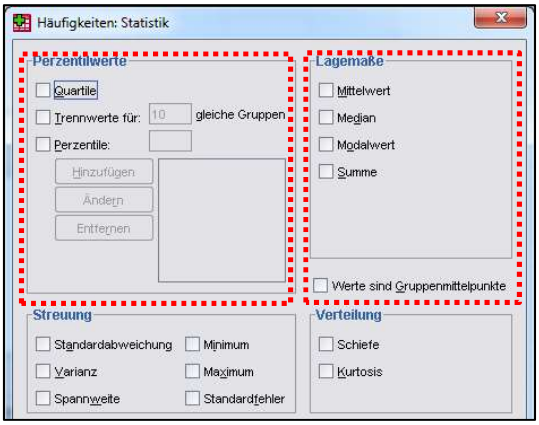
Diese Struktur findet man sehr häufig bei Wachstums- oder auch Verzinsungsprozessen in der Ökonomie.

Liegt den Daten eine geometrische Struktur zugrunde, so muss anstatt des arithmetischen Mittels das geometrische Mittel verwendet werden, bei dem die Merkmalswerte multipliziert und dann aus dem Produkt die n-te Wurzel gezogen wird. Hingegen werden beim arithmetischen Mittel die Werte addiert und dann durch n dividiert.

Die nachstehenden Werte zeigen die, in diesem Fall nur geringen Unterschied von 0,000052 der beiden Berechnungen. Bei größeren Wachstumsraten ist der Unterschied dann entsprechend größer.

2.3.4 SPSS

Die Berechnung der Lageparameter mittels SPSS ist einfach; man gelangt über das Item „Häufigkeiten“ und dort über den Button „Statistik“ zu dem nachstehenden Fenster. Die Perzentile und Lagemaße sind die beiden oberen Groupboxen und können dort nach Bedarf selektiert werden. Eine Option für die Berechnung des geometrischen Mittels existiert nicht.



Gewicht (g)

N	Gültig	20
	Fehlend	0
Mittelwert		,6440
Median		,6500
Modus		,65

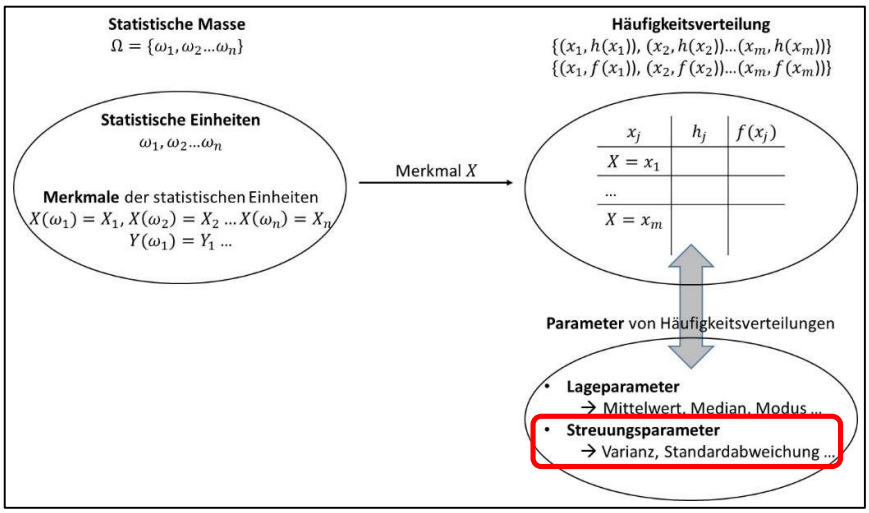
2.4 Streuungsparameter

Die Aussagekraft von Mittelwerten ist eingeschränkt. Bei einer Klausur kann ein Mittelwert von 3,0 dadurch zustande kommen,

- dass alle Schüler eine 3,0 geschrieben haben oder
- dadurch, dass 50% eine 1,0 und 50% eine 5,0 geschrieben haben.

In beiden Fällen ist der Mittelwert identisch. Um diesen Parameter zusätzlich bewerten zu können, benötigt man Streuungsparameter. Die ***Standardabweichung*** als ein Beispiel für einen Streuungsparameter beschreibt ***die mittlere Abweichung der Messwerte vom Mittelwert*** und könnte damit zu einer besseren Beurteilung der obigen Situation führen.

Streuungsparameter sind ebenso wie Lageparameter Kennzahlen einer Häufigkeitsverteilung, die, wie die Bezeichnung bereits sagt, im Gegensatz zur Lage jetzt die Streuung der Parameter beschreibt.



2.4.1 Mathematische Definitionen

Varianz

Sei X ein kardinal skaliertes Merkmal. Dann heißt das arithmetische Mittel der quadrierten Abweichungen der Merkmalswerte X_i von ihrem arithmetischen Mittel $\bar{x}$ Varianz S^2.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{m} h_j * (x_i - \bar{x})^2 = \sum_{j=1}^{m} f_j * (x_j - \bar{x})^2$$

Anmerkungen

- Die Varianz wird im Gegensatz zur absoluten mittleren Abweichung d immer als Abweichung vom arithmetischen Mittel $\bar{x}$ berechnet. Die Varianz hat eine quadratische Dimension.
- Bei klassierten Daten erfolgt einer Korrektur der Varianz mit Hilfe der Sheppardschen Korrektur

$$s^2_{korr} = s^2 - \frac{b^2}{12}$$

Varianz unter Berücksichtigung von Freiheitsgraden

Sei X ein kardinal skaliertes Merkmal. Dann heißt das arithmetische Mittel der quadrierten Abweichungen der Merkmalswerte X_i von ihrem arithmetischen Mittel $\bar{x}$ Varianz S^2.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n-1}\sum_{j=1}^{m} h_j * (x_i - \bar{x})^2$$

Varianz unter Berücksichtigung des Verschiebungssatzes

Sei X ein kardinal skaliertes Merkmal. Dann gilt für die Varianz s^2 die folgende Beziehung:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{x}^2$$

Anmerkungen

- Mit Hilfe dieses Verschiebungssatzes lässt sich die Varianz vereinfacht ermitteln, da nicht erst die Differenzen $X_i - \bar{x}$ berechnet werden müssen.

Alternative Berechnungen der Varianz (ohne Freiheitsgrade)

Urliste	$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$
Häufigkeitsverteilung	$s^2 = \frac{1}{n}\sum_{j=1}^{m} h_j * (x_i - \bar{x})^2 = \sum_{j=1}^{m} f_j * (x_j - \bar{x})^2$
Verschiebungssatz	$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{x}^2$

Bei der Berücksichtigung der Freiheitsgrade verändert sich der Quotient von $\frac{1}{n}$ in $\frac{1}{n-1}$.

Standardabweichung

Sei X ein kardinal skaliertes Merkmal. Dann heißt die positive Wurzel aus der Varianz die Standardabweichung S.

$$s = +\sqrt{s^2}$$

Anmerkungen

- Die Standardabweichung ist nicht mehr von quadratischer Dimension. Der Wert ist daher direkt am Beispiel interpretierbar.

Variationskoeffizient

Sei X ein kardinal skaliertes Merkmal. Der Quotient aus Standardabweichung und arithmetischem Mittel heißt Variationskoeffizient v.

$$v = \frac{s}{\bar{x}}$$

Anmerkungen

- Der Variationskoeffizient stellt eine relative Größe dar. Es wird die Schwankung in Relation zum Mittelwert betrachtet, wodurch auch unterschiedliche Merkmalwerte in Beziehung gesetzt werden können.

Mittlere absolute Abweichung

Sei X ein kardinal skaliertes Merkmal. Dann heißt das arithmetische Mittel der Abweichungen der Merkmalswerte X_i vom Median $\bar{x}_z$ mittlere absolute Abweichung d.

$$d = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{x}_z| = \frac{1}{n}\sum_{j=1}^{m} h_j * |x_j - \bar{x}_z| = \sum_{j=1}^{m} f_j * |x_j - \bar{x}_z|$$

Anmerkungen

- Die mittlere absolute Häufigkeit wird üblicher Weise als Abweichung vom Zentralwert oder Median berechnet.
- Es sind aber auch andere Bezugsgrößen denkbar. Bei der Verwendung des Medians $\bar{x}_z$ wird d minimal.

Mittlere absolute Abweichung unter Berücksichtigung von Freiheitsgraden

Sei X ein kardinal skaliertes Merkmal. Dann heißt das arithmetische Mittel der Abweichungen der Merkmalswerte X_i vom Median $\bar{x}_z$ mittlere absolute Abweichung d.

$$d = \frac{1}{n-1}\sum_{i=1}^{n}|X_i - \bar{x}_z| = \frac{1}{n-1}\sum_{j=1}^{m} h_j * |x_j - \bar{x}_z|$$

Spannweite (Range)

Sei X ein kardinal skaliertes Merkmal. Dann heißt die Differenz aus dem größtem und dem kleinsten Merkmalswert Spannweite w.

$$w = \max_{i=1,\dots,n} X_i - \min_{i=1,\dots,n} X_i$$

Quantile

Seien $X^*(\omega_i) = X_i^*$ die auf- oder absteigend sortierten Merkmalswerte $X(\omega_i) = X_i$. Sei ferner X ein mindestens ordinalskaliertes Merkmal. Dann heißt der Merkmalswert x_q, die die aufsteigend geordnete Folge aller beobachteten Merkmalswerte X_i derart zweiteilt, dass q Anteile unterhalb und $1 - q$ Anteile oberhalb der Merkmalsausprägung x_q liegen, wobei $0 \leq q \leq 1$ gilt, Quantil.

Anmerkungen

- Median
  ist der Merkmalswert $x_{0,5} = \bar{x}_z$, der die aufsteigend sortierten Merkmalswerte X_i^* in 2 gleich große Teilgesamtheiten zerlegt.
- Quartile
  sind die Merkmalswerte $x_{0,25}; x_{0,5}; x_{0,75}$, die die aufsteigend sortierten Merkmalswerte X_i^* in 4 gleichgroße Teilgesamtheiten zerlegen.
- Perzentil

sind die Merkmalswerte $x_{0,01}; x_{0,02}; x_{0,03} \dots x_{0,99}$, die die aufsteigend sortierten Merkmalswerte X_i^* in 100 gleichgroße Teilgesamtheiten zerlegen.

2.4.2 Herleitungen

2.4.2.1 Grundgedanke der Varianz & Standardabweichung

Die Varianz und die Wurzel aus der Varianz - die Standardabweichung - sind zwei zentrale Parameter der Statistik. Die Varianz misst die quadratische Streuung der Merkmalswerte um den Mittelwert. Bei genauer Betrachtung ist die Varianz auch wieder ein Mittelwert. In diesem Fall allerdings der Mittelwert der (quadratischen) Abweichungen der Merkmalswerte vom Mittelwert. Diese Idee führt im ersten Schritt zur Berechnung der Abweichungen der Merkmalswerte vom Mittelwert:

$$X_i - \bar{x}$$

Um den Mittelwert dieser Differenzen zu bestimmen, müssen diese Differenzen summiert und dann durch die Anzahl dividiert werden:

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \bar{x}$$

Der Nachteil dieses Ansatzes ist, dass die Messwerte nach oben und nach unten vom Mittelwert abweichen, sie sich dadurch gegenseitig aufheben und die Streuung somit nicht korrekt wiedergegeben wird. Eine mögliche Lösung dieses Problems besteht in der Quadratur der Differenzen:

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

Der Nachteil dieses Ansatzes ist die quadratische Dimension der Varianz, den man dann durch das Ziehen der Wurzel wieder behebt. Diese Größe, die Standardabweichung, besitzt dann die Dimension der Messwerte:

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

Vergleicht man den Mittelwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

und die Varianz

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

so könnte man die Varianz auch als den Mittelwert von $(X_i - \bar{x})^2$ sehen. Die folgende Schreibweise ist in der beschreibenden Statistik nicht üblich, würde diesen Sachverhalt aber gut verdeutlichen:

Mittelwert : $\bar{x}_{X_i}$

Varianz : $\bar{x}_{(X_i - \bar{x})^2}$

2.4.2.2 Verschiebungssatz

Der Verschiebungssatz ermöglicht eine einfachere Berechnung der Varianz, da nicht erst die Differenzen $X_i - \bar{x}$ berechnet werden müssen. Dieser Ansatz wird in späteren Kapiteln sehr häufig verwendet werden. Von daher erfolgt an dieser Stelle der Beweis.

Um den Beweis verstehen zu können, soll noch einmal die Formel für den Mittelwert und eine kleine Umformung durch die Multiplikation mit „n“ in Erinnerung gebracht werden:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \mid *n$$

$$\Leftrightarrow n\bar{x} = \sum_{i=1}^{n} X_i$$

Der eigentliche Beweis sieht danach wie folgt aus:

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

$$\Leftrightarrow s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{x} + \bar{x}^2) \mid Binomische\ Formel$$

$$\Leftrightarrow s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} 2X_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right) \mid Auflösung\ der\ Summe$$

$$\Leftrightarrow s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{x}\sum_{i=1}^{n} X_i + n\bar{x}^2\right)$$

$$\Leftrightarrow s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2\right) \mid s.Vorbemerkung$$

$$\Leftrightarrow s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right)$$

$$\Leftrightarrow s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{x}^2\right)$$

Diese Variante kann für die Berechnungen noch weiter vereinfacht werden:

$$\Leftrightarrow s^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}n\bar{x}^2$$

$$\Leftrightarrow s^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{x}^2$$

Der Term $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ entspricht auch wieder einem Mittelwert, nämlich dem der quadrierten Messwerte X_i, also X_i^2 und könnte daher auch als $\overline{x^2}$ beschrieben werden. Daraus würde sich dann ergeben

$$\Leftrightarrow s^2 = \overline{x^2} - \bar{x}^2$$

Zusammenfassend lässt sich festhalten: die Formel $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$ ist sehr gut für die Herleitung der Varianz aber nicht für deren Berechnung geeignet. Die Formel nach dem Verschiebungssatz $s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{x}^2\right)$ hingegen ist für die Herleitung unbrauchbar, ermöglicht aber eine einfachere Berechnung. Im weiteren Verlauf wird daher überwiegend mit der Varianz nach dem Verschiebungssatz gearbeitet.

2.4.2.3 Freiheitsgrade & Schließende Statistik

Die obigen Berechnungen zeigen unterschiedliche Ergebnisse bei der Varianz und der Standardabweichung.

0,65		x_j	$h(x_j) = h_j$	$f(x_j)$	$F(x_j)$		
0,65		0,60	0	0,00	0%		
0,64		0,61	2	0,10	10%		
0,65		0,62	4	0,20	30%		
0,65		0,63	1	0,05	35%		
0,62		0,64	2	0,10	45%		
0,63		0,65	6	0,30	75%		
0,61		0,66	1	0,05	80%		
0,65		0,67	1	0,05	85%		
0,64		0,68	3	0,15	100%		
0,65		0,69	0	0,00	100%		
0,68		0,70	0	0,00	100%		
0,66			20	100,00%			
0,67							
0,62		***Lageparameter***					
0,61		Mittelwert		0,644			
0,62		Median		0,650			
0,68		Modus		0,650			
0,68		***Streuungsparameter***					
0,62		Varianz		0,00052			
		Standarabweichung		0,02280	=>	[0,621 ; 0,667]	
		Variationskoeffizient		0,03541			
		Variationskoeffizient in %		3,54%			
		mittlere abs. Abweichung		0,0189	=>	[0,631 ; 0,669]	

Bei der manuellen Berechnung ergibt sich bei der Division durch die Anzahl $n = 20$ eine Varianz von $s^2 = 0{,}000494$. Bei der Berechnung durch SPSS (auch Excel) ergibt sich bei der Division durch die um 1 reduzierte Anzahl also $n - 1 = 20 - 1 = 19$ eine Varianz von $s^2 = 0{,}00052$.

manuelle Berechnung	***SPSS***
$s^2 = \frac{1}{20} * 0{,}00988 = 0{,}0004940$	$s^2 = \frac{1}{19} * 0{,}00988 = 0{,}000520$

Dieser Abweichung liegt eine unterschiedliche Betrachtungsweise zugrunde. Letztlich kann diese erst im Rahmen der schließenden Statistik geklärt werden. An dieser Stelle soll als Exkurs eine anschauliche Herleitung stehen.

Eingangs wurde der Unterschied zwischen der beschreibenden und der schließenden Statistik deutlich gemacht. Statistische Software wie SPSS wird in der empirischen Forschung eingesetzt, bei der die Ergebnisse einer Stichprobe auf eine Grundgesamtheit übertragen werden. Es handelt sich damit um das Gebiet der schließenden Statistik. Bei der Übertragung von Erkenntnissen von einer Stichprobe auf die Grundgesamtheit treten naturgemäß Fehler auf, denn eine Stichprobe ist eben nur ein Teil der Grundgesamtheit. Ggf. kann, ohne dass man es merkt, die Stichprobe „unglücklich“ zusammengesetzt sein; so können z.B. bei einer Befragung von Studierenden zu deren Monatseinkommen zufällig nur die „Großverdiener“ ausgewählt worden sein, was dann natürlich bei den Schlüssen auf die Grundgesamtheit zu Fehlern führen würde. Grundsätzlich kann man aber festhalten, dass mit zunehmender Größe der Stichprobe dieser mögliche Fehler immer geringer wird.

Wirkungsweise der Freiheitsgrade

Das später noch zu erläuternde Konzept der Freiheitsgrade trägt diesen Überlegungen Rechnung, in dem nicht mit n, sondern einem korrigierten Wert gerechnet wird. In Fall der Varianz ist der korrigierte Wert $n - 1$. Vergleicht man die beiden Formeln

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 \text{ und } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

so ergibt sich ein Unterschied in den Quotienten $\frac{1}{n}$ und $\frac{1}{n-1}$ wobei für große n gilt:

$$\frac{1}{n} \approx \frac{1}{n-1}$$

D.h., dass mit größer werdender Stichprobe der Unterschied zwischen beiden Ergebnissen immer geringer wird. Dieses ist sinnvoll, da mit größer werdender Stichprobe diese der Grundgesamtheit immer

näherkommt und damit der Schluss von der Stichprobe auf die Grundgesamtheit mit einem immer geringeren Risiko behaftet ist.

Allgemeines Konzept der Freiheitsgrade

Im speziellen Fall der Varianz bestimmen sich die Freiheitsgrade durch $n-1$; allgemein betrachtet ergeben sich die Freiheitsgrade durch

$$Freiheitsgrade = n - k$$

wobei k die Anzahl der in die Berechnung einfließenden Variablen ist.

Veranschaulichung der Freiheitsgrade

Anschaulich lassen sich die Freiheitsgrade als die maximale Anzahl von frei wählbaren Werten (daher der Begriff **Freiheit**sgrade) bei einer Berechnung bezeichnen, die frei veränderlich sind, ohne dass das Ergebnis verändert wird. Als Beispiel stelle man sich die Berechnung eines Mittelwertes von 10 Zahlen vor, der den Wert 0 ergibt. Soll dieser Mittelwert 0 weiterhin bestehen bleiben, können max. 9 Werte beliebig geändert werden. Der 10. Wert ergibt sich dann aus den vorherigen 9 Werten, da der Mittelwert wieder 0 werden soll. Die 9 frei wählbaren Elemente stellen die Freiheitsgrade dieser Berechnung dar. Sie bestimmen sich aus der Anzahl der untersuchten Fälle $n = 10$ und der Anzahl der Variablen $k = 1$ und somit $Freiheitsgrade = 10 - 1 = 9$.

Betrachtet man die Berechnung eines Korrelationskoeffizienten von **zwei** Variablen bei der 30 Werte zur Verfügung stehen, so berechnen sich die Freiheitsgrade wie folgt

$$Freiheitsgrade = 30 - 2 = 28,$$

da in diesem Fall zwei Variablen in die Berechnung einfließen.

2.4.3 Berechnungen

Die Optionen zur Berechnung der Varianz sind vielfältig. Die Varianz kann mit der Urliste aber auch mit der Häufigkeitsverteilung berechnet werden. Eine weitere Vorgehensweise bietet der Verschiebungssatz, wofür die Spalte X_i^2 erforderlich ist. Alle Verfahren werden nachstehend erläutert.

i	X_i	X_i^2
1	0,65	0,42
2	0,65	0,42
3	0,64	0,41
4	0,65	0,42
5	0,65	0,42
6	0,62	0,38
7	0,63	0,40
8	0,61	0,37
9	0,65	0,42
10	0,64	0,41
11	0,65	0,42
12	0,68	0,46
13	0,66	0,44
14	0,67	0,45
15	0,62	0,38
16	0,61	0,37
17	0,62	0,38
18	0,68	0,46
19	0,68	0,46
20	0,62	0,38
Summe	12,8800	8,3046
Anzahl	20	
Mittelwert	0,6440	

x_j	$h(x_j) = h_j$	$f(x_j)$	$F(x_j)$
0,60	0	0,00	0%
0,61	2	0,10	10%
0,62	4	0,20	30%
0,63	1	0,05	35%
0,64	2	0,10	45%
0,65	6	0,30	75%
0,66	1	0,05	80%
0,67	1	0,05	85%
0,68	3	0,15	100%
0,69	0	0,00	100%
0,70	0	0,00	100%
	20	100,00%	

Varianz anhand der Urliste

Bei dieser Vorgehensweise ist für jeden Merkmalswert die Abweichung zum Mittelwert zu berechnen und diese Differenz dann zu Quadrieren. Da, wie bereits erläutert, die Varianz eigentlich auch ein Mittelwert ist, sind diese Quantitäten Differenzen zu summieren und durch die Anzahl zu dividieren. Die

Varianz ist aufgrund ihrer quadratischen Dimension schlecht zu veranschaulichen. Dieses gelingt aber mit der Standardabweichung.

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

$$s^2 = \frac{1}{20} * ((0{,}65 - 0{,}644)^2 + (0{,}65 - 0{,}644)^2 \dots)$$

$$s^2 = \frac{1}{20} * 0{,}00988 = 0{,}0004940$$

Varianz anhand der relativen Häufigkeiten

Die Berechnung mit den relativen Häufigkeiten erschließt sich unmittelbar und führt natürlich zum gleichen Ergebnis. Der Vorteil dieser Vorgehensweise liegt in dem geringeren Rechenaufwand gegenüber der Berechnung mit der Urliste.

$$\mathrm{s}^2 = \sum_{\mathrm{j}=1}^{\mathrm{m}} f(x_j) * (\mathrm{x_j} - \bar{x})^2$$

$$\mathrm{s}^2 = (0{,}61 - 0{,}644)^2 * 0{,}1 \dots (0{,}68 - 0{,}644)^2 * 0{,}15 = 0{,}000494$$

Varianz anhand des Verschiebungssatzes

Der Beweis für den Verschiebungssatz ist bereits bekannt; hier geht es nur um die Anwendung. Der Verschiebungssatz ist bei vielen Berechnungen in der Statistik von Bedeutung; man arbeitet zwar bei diesem Ansatz wieder mit der Urliste, kommt aber ohne eine aufwendige Berechnung der Abweichungsquadrate aus. Bei dem ersten Term $\sum_{i=1}^{n} X_i^2$ sind nur die Merkmalswerte zu quadrieren und dann zu summieren. Einen zweiten Teil des Terms $n\bar{x}^2$ ist der Mittelwert zu Quadrieren und mit der Anzahl n zu multiplizieren. Für die nachstehende Berechnung wurde die Summe $\sum_{i=1}^{n} X_i^2 = 8{,}3046$ der obigen Tabelle entnommen:

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{x}^2\right) = \frac{1}{20} * (8{,}3046 - 20 * 0{,}644^2) = 0{,}000494$$

Standardabweichung

Da die Standardabweichung nicht mehr quadratisch ist, lässt sie sich deutlich besser als die Varianz interpretieren.

$$s = +\sqrt{s^2} = +\sqrt{0{,}000494} = 0{,}0222261$$

Die obigen 20 Messwerte weichen damit im Mittel um $\pm 0{,}02223$ vom Mittelwert von $\bar{x} = 0{,}644$ ab. Damit ergibt sich das folgende Intervall für die mittlere Abweichung vom Mittelwert:

$$[0{,}644 - 0{,}022; 0{,}644 + 0{,}022] = [0{,}622; 0{,}666]$$

Variationskoeffizient

Beim Variationskoeffizienten wird die Standardabweichung in Bezug zum Mittelwert gesetzt, wodurch entsprechend der Prozent- oder Dreisatzrechnung die Standardabweichung relativ zum Mittelwert ausgedrückt wird. Eine Multiplikation mit 100 ergibt dann den Prozentwert. Relative Wert lassen sich im Gegensatz zu absoluten Werten besser vergleichen. Ein typisches Beispiel hierfür ist der Vergleich von Kursschwankungen bei Aktien; eine Standardabweichung von 10 € ist bei einem Aktienwert von 50 € anders zu beurteilen als bei einem Aktienwert von 500 €. In diesem Beispiel entspricht die Standardabweichung $s = 0{,}222261$ einer Abweichung vom Mittelwert von 3,45%.

$$v = \frac{s}{\bar{x}} = \frac{0{,}0222261}{0{,}644} = 0{,}0345126$$

Mittlere absolute Abweichung

Die mittlere absolute Abweichung wird üblicherweise mit dem Median (Zentralwert) berechnet. Berechnet man die Abweichungen zwischen einem Merkmalswert und dem Mittelwert, so besteht immer das Problem, dass sich positive und negative Abweichungen gegenseitig eliminieren. Bei der Varianz wird dieses Problem durch das Quadrieren der Abweichungen, bei der mittleren absoluten Abweichung hingegen durch den Betrag gelöst. Bei der Betragsoperation wird das Ergebnis der Differenz vorzeichenlos betrachtet, wodurch auch das Aufrechnen positiver und negativer Abweichungen verhindert wird.

Bei der nachstehenden Formel ist der erste Term für die Berechnungen mit der Urliste, der zweite für die absoluten Häufigkeiten und der dritte für die relativen Häufigkeiten.

$$d = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{x}_z| = \frac{1}{n}\sum_{j=1}^{m} h_j * |x_j - \bar{x}_z| = \sum_{j=1}^{m} f_j * |x_j - \bar{x}_z|$$

$$d = \frac{1}{20} * (|0{,}65 - 0{,}65| + |0{,}65 - 0{,}65| + |0{,}64 - 0{,}65| \ldots) = 0{,}018$$

Da diese Größe nicht quadratischer Natur ist, kann sie direkt interpretiert werden. Die mittlere absolute Abweichung vom Median beträgt $0{,}018$. Auch dieser Sachverhalt lässt sich als Intervall schreiben:

$$[0{,}65 - 0{,}018; 0{,}65 + 0{,}018] = [0{,}632; 0{,}668]$$

2.4.4 Interpretationen

Standardabweichung

Die Varianz oder Standardabweichung ist erst einmal einfach nur eine rechnerische Größe. Ob dieser Wert **gut** oder **schlecht** ist, hängt von der Problemstellung ab. Untersucht man beispielsweise aus produktionstechnischen Gründen die Stärke von Stahlblechen, so will man natürlich immer möglichst gleich starke Platten produzieren und damit eine geringe Standardabweichung der Messungen haben. Betrachtet man hingegen eine Klausur, so wären ein Ergebnis, bei dem alle Studierenden eine 3,0 haben, testtheoretisch nicht gut, da eine Klausur eben auch die guten von den schlechten Leistungen differenzieren soll. Hier wäre folglich eine große Standardabweichung wünschenswert.

Mittlere absolute Abweichung ⇔ Standardabweichung

Die Definitionen im ersten Teil sind so gewählt, dass die **Standardabweichung** immer im Zusammenhang mit dem **arithmetischen Mittel**, die **mittlere absolute Abweichung** wird immer im Zusammenhang mit dem **Median** (Zentralwert) verwendet wird.

Dieses muss aber nicht so sein. Grundsätzlich können anstelle des arithmetischen Mittels bei der Standardabweichung und des Medians bei der mittleren absoluten Abweichung beliebige Werte verwendet werden. Der berechnete Parameter sagt dann aus, inwieweit die Merkmalswerte dann um diesen beliebigen Wert streuen. Allerdings führt bei der Standardabweichung jeder andere Wert als das arithmetische Mittel und bei der mittleren absoluten Abweichung jeder andere Wert als der Median zu einer größeren Streuung. Man spricht hier von der Minimum-Eigenschaft des arithmetischen Mittels bzw. des Medians. Diese Erkenntnis führt dann auch zu den entsprechenden Definitionen.

Lage- und Streuungsparameter

Die nachstehenden Beispiele verdeutlichen noch einmal an unterschiedlichem Zahlenmaterial die Notwendigkeit, den Mittelwert immer im Zusammenhang mit der Varianz oder Standardabweichung zu interpretieren. Beide Verteilungen weisen einen Mittelwert von 3,0 auf. Erst die Varianz bzw. die Standardabweichung macht deutlich, dass die Einzelwerte beider Verteilungen doch sehr unterschiedlich um den jeweiligen Mittelwert schwanken.

Die Standardabweichung von $1{,}384$ ist so zu interpretieren, dass die einzelnen Messwerte im Mittel um $\pm 1{,}384$ um den Mittelwert von $3{,}0$ schwanken. Damit ergibt sich das folgende Intervall für die mittlere Abweichung vom Mittelwert $[3{,}0 - 1{,}384; 3{,}0 + 1{,}384] = [1{,}616; 4{,}384]$.

- **Beispiel 1 für Häufigkeitsverteilung mit Mittelwert 3,0**

Note	abs. Häufigkeit
1	5
2	3
3	8
4	3
5	5

Streuungsparameter	
Mittelwert	3,000
Varianz	1,917
Standardabweichung	1,384
Variationskoeffizient	0,461

- **Beispiel 2 für Häufigkeitsverteilung mit Mittelwert 3,0**

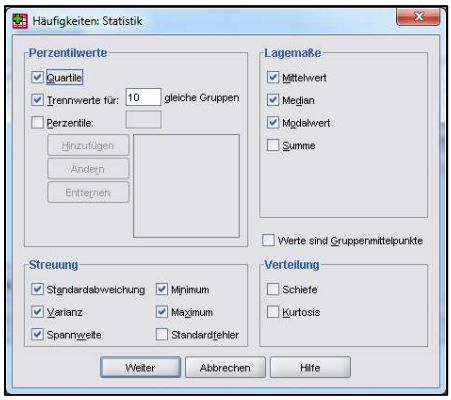
Note	abs. Häufigkeit
1	0
2	0
3	24
4	0
5	0

Streuungsparameter	
Mittelwert	3,000
Varianz	0,000
Standardabweichung	0,000
Variationskoeffizient	0,000

2.4.5 SPSS

Die Berechnung der Lage- und Streuungsparameter erfolgt in SPSS über den Menüpunkt „Häufigkeiten“ und dort den Button „Statistik“.

Einstellung des Fensters „Statistik“



Ausgabe

Die Berechnungen von SPSS liefern die nachstehenden Werte für die Streuungsparameter.

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	,61	2	10,0	10,0	10,0
	,62	4	20,0	20,0	30,0
	,63	1	5,0	5,0	35,0
	,64	2	10,0	10,0	45,0
	,65	6	30,0	30,0	75,0
	,66	1	5,0	5,0	80,0
	,67	1	5,0	5,0	85,0
	,68	3	15,0	15,0	100,0
	Gesamt	20	100,0	100,0	

N	Gültig	20
	Fehlend	0
Mittelwert		,6440
Median		,6500
Modus		,65
Standardabweichung		,02280
Varianz		,000520
Spannweite		,07
Minimum		,61
Maximum		,68
Perzentile	25	,6200
	50	,6500
	75	,6575

3 Bivariate Statistik

Bei der bisherigen Betrachtungsweise stand immer nur ein Merkmal X im Fokus der Analyse. Werden zwei oder mehr Merkmale $X, Y, \ldots$ betrachtet, steht i.d.R. nicht die Verteilung des einzelnen Merkmals im Vordergrund sondern auch der ***Zusammenhang zwischen den Merkmalen***. Letzteres ist nicht zwangsläufig, tritt aber sehr häufig auf.

Bei der Betrachtung von zwei Merkmalen spricht man von bivariater Statistik, bei mehr Merkmalen von multivariater Statistik. Grundsätzlich könnte eine beliebige Anzahl von Merkmalen analysiert und in einer Häufigkeitsverteilung dargestellt werden. Jedoch treten erhebliche Einschränkungen bei der Anschauung bei mehr als zwei Variablen auf.

Zur ersten Veranschaulichung dient hier das aus der Ökonomie bekannte Beispiel des Zusammenhanges zwischen Einkommen ⇔ Konsum. In den Zeilen stehen die Konsumklassen, in die Spalten die Einkommensklassen. Der markierte Wert 7 steht für sieben Probanden, die bei einem Einkommen von 20 Geldeinheiten einen Konsum von 15 Geldeinheiten aufweisen.

- Gut erkennbar ist an dieser Tabelle, dass mit steigendem Einkommen auch der Konsum steigt, da bei höherem Einkommen die geringen Konsumklassen weniger Probanden aufweisen.
- Des Weiteren ist – wenn auch etwas schlechter - erkennbar, dass bei steigendem Einkommen die Sparquote steigt, d.h. mit zunehmendem Einkommen immer mehr vom Einkommen gespart wird. Erkennbar ist dieses daran, da der Anteil der oberhalb des blauen Pfeils liegenden Fälle, also der Fälle, die nur einen Teil des Einkommens konsumieren, immer größer wird.

Häufigkeitsverteilung mit absoluten Häufigkeiten

		Einkommen				
		10,00	15,00	20,00	25,00	Gesamt
Konsum	5,00	2	0	0	0	2
	10,00	8	5	1	1	15
	15,00	0	5	7	8	20
	20,00	0	0	2	8	10
	25,00	0	0	0	3	3
Gesamt		10	10	10	20	50

Häufigkeitsverteilung mit bedingten relativen Häufigkeiten als Spaltenprozente

Die Argumentation bezüglich des Zusammenhanges zwischen der Sparquote und dem Einkommen lässt sich deutlich besser anhand der bedingten relativen Spalten-Häufigkeit führen. Betrachtet man die Prozente **unterhalb** des blauen Pfeils, so wird deutlich, dass der Anteil deren, die ihr gesamtes Einkommen konsumieren, von 80% bei 10 GE Einkommen auf 15% bei 25 GE Einkommen sinkt.

			Einkommen				
			10,00	15,00	20,00	25,00	Gesamt
Konsum	5,00	Anzahl	2	0	0	0	2
		% innerhalb von Einkommen	20,0%	,0%	,0%	,0%	4,0%
	10,00	Anzahl	8	5	1	1	15
		% innerhalb von Einkommen	80,0%	50,0%	10,0%	5,0%	30,0%
	15,00	Anzahl	0	5	7	8	20
		% innerhalb von Einkommen	,0%	50,0%	70,0%	40,0%	40,0%
	20,00	Anzahl	0	0	2	8	10
		% innerhalb von Einkommen	,0%	,0%	20,0%	40,0%	20,0%
	25,00	Anzahl	0	0	0	3	3
		% innerhalb von Einkommen	,0%	,0%	,0%	15,0%	6,0%
Gesamt		Anzahl	10	10	10	20	50
		% innerhalb von Einkommen	100,0%	100,0%	100,0%	100,0%	100,0%

3.1 Mathematische Definitionen

Zweidimensionale Urliste

Grundlage für die weiteren Definitionen ist die nachstehende Struktur einer zweidimensionalen Urliste:

i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
…	…	…
n	X_n	Y_n

Merkmalswert $\mathbf{X(\omega_i) = X_i}$ bzw $\mathbf{Y(\omega_i) = Y_i}$

Den Wert, den ein Merkmal X oder Y einer statischen Einheit ω_i mit $i = 1, \dots, n$ annimmt, heißt Merkmalswert $X(\omega_i) = X_i$ bzw. $Y(\omega_i) = Y_i$

Merkmalsrealisationen

Für ein- oder mehrfach auftretende Merkmalswerte steht ***eine*** Merkmalsrealisation x für das Merkmal X und y für das Merkmal Y. Die Menge aller m_x bzw. m_y Merkmalsrealisationen einer Gesamtheit werden wie folgt bezeichnet:

$x_1, x_2, x_3, \dots, x_{m_x} = x_k$ mit $k = 1, \dots, m_x$

$y_1, y_2, y_3, \dots, y_{m_y} = y_l$ mit $l = 1, \dots, m_y$

Zweidimensionale absolute Häufigkeitsverteilung

Gegeben sei eine statistische Gesamtheit Ω mit n Einheiten $\omega_1, \omega_2, \dots, \omega_n$, an denen die Merkmale X und Y mit den m_x und m_y sich voneinander unterscheidenden Merkmalsrealisationen x_k und y_l beobachtet wurden.

Dann heißt die Anzahl, mit der das Ausprägungspaar $(x_k; y_l)$ beobachtet wurde, **absolute Häufigkeit $\boldsymbol{h(x_k; y_l)}$**.

Die nachstehende Tabelle der absoluten Häufigkeiten ist definiert als **Kreuztabelle der Merkmale $\boldsymbol{X}$ und $\boldsymbol{Y}$**.

X \ Y	y_1	y_2	…	y_l	…	y_{m_Y}	h
x_1	$h_{1,1}$	$h_{1,2}$	…	$h_{1,l}$	…	h_{1,m_Y}	$h(x_1)$
x_2	$h_{2,1}$	$h_{2,2}$	…	$h_{2,l}$	…	h_{1,m_Y}	$h(x_2)$
⋮	⋮	⋮		⋮		⋮	⋮
x_k	$h_{k,1}$	$h_{k,2}$	…	$h_{k,l}$	…	h_{k,m_Y}	$h(x_k)$
⋮	⋮	⋮		⋮		⋮	⋮
x_{m_X}	$h_{m_X,1}$	$h_{m_X,2}$	…	$h_{m_X,l}$	…	h_{m_X,m_Y}	$h(x_{m_X})$
h	$h(y_1)$	$h(y_2)$	…	$h(y_l)$	…	$h(y_{m_Y})$	n

Anmerkung

Weitere Bezeichnungen für eine Kreuztabelle

Nominalskalierte Merkmale	Kontingenztabelle
Ordinal- oder kardinalskalierte Merkmale	Korrelationstabelle
Tabelle mit $m_x = m_y$	quadratische Kreuztabelle
Tabelle für zwei dichotome Merkmale	Vierfeldermatrix

Relative Häufigkeit

Gegeben sei eine Häufigkeitstabelle mit den absoluten Häufigkeiten $h(x_k; y_l)$. Dann ist die **relative Häufigkeit** definiert durch

$$f(x_k; y_l) = \frac{h(x_k; y_l)}{n}$$

Die nachstehende Tabelle ist als **zweidimensionale relative Häufigkeitstabelle** definiert.

X \ Y	y_1	y_2	$\cdots$	y_l	$\cdots$	y_{m_Y}	f
x_1	$f_{1,1}$	$f_{1,2}$	$\cdots$	$f_{1,l}$	$\cdots$	f_{1,m_Y}	$f(x_1)$
x_2	$f_{2,1}$	$f_{2,2}$	$\cdots$	$f_{2,l}$	$\cdots$	f_{1,m_Y}	$f(x_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	$f_{k,1}$	$f_{k,2}$	$\cdots$	$f_{k,l}$	$\cdots$	f_{k,m_Y}	$f(x_k)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_{m_X}	$f_{m_X,1}$	$f_{m_X,2}$	$\cdots$	$f_{m_X,l}$	$\cdots$	f_{m_X,m_Y}	$f(x_{m_X})$
f	$f(y_1)$	$f(y_2)$	$\cdots$	$f(y_l)$	$\cdots$	$f(y_{m_Y})$	1

Anmerkung

- Die Bestimmung der relativen Häufigkeit bezieht sich immer auf das n der gesamten Tabelle.

Bedingte relative Häufigkeit

Gegeben sei eine Häufigkeitstabelle mit den absoluten Häufigkeiten $h(x_k; y_l)$. Die Häufigkeit, mit der das Merkmal X die Merkmalsrealisation x_k annimmt unter der Bedingung, dass das Merkmal Y die Ausprägung y_l besitzt, heißt **bedingte relative Häufigkeit**

$$f(x_k|y_l) = \frac{h(x_k; y_l)}{h(y_l)}$$

Die folgende Tabelle ist als **zweidimensionale, bedingt relative Häufigkeitstabelle** definiert.

X \ Y	y_1	y_2	$\cdots$	y_l	$\cdots$	y_{m_Y}
x_1	$f(x_1 \| y_1)$	$f(x_1 \| y_2)$	$\cdots$	$f(x_1 \| y_l)$	$\cdots$	$f(x_1 \| y_{m_Y})$
x_2	$f(x_2 \| y_1)$	$f(x_2 \| y_2)$	$\cdots$	$f(x_2 \| y_l)$	$\cdots$	$f(x_2 \| y_{m_Y})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_k	$f(x_k \| y_1)$	$f(x_k \| y_2)$	$\cdots$	$f(x_k \| y_l)$	$\cdots$	$f(x_k \| y_{m_Y})$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_{m_X}	$f(x_{m_X} \| y_1)$	$f(x_{m_X} \| y_2)$	$\cdots$	$f(x_{m_X} \| y_l)$	$\cdots$	$f(x_{m_X} \| y_{m_Y})$
Summe	1	1	…	1	…	1

Anmerkung

- Diese Berechnung ist identisch mit den **Spaltenprozenten**.
- Analog gilt für die Berechnung der **Zeilenprozente**

$$f(y_l|x_k) = \frac{h(x_k; y_l)}{h(x_k)}$$

- Die Bestimmung der bedingten relativen Häufigkeit bezieht sich immer auf das n der Spalte oder Zeile der Tabelle.

Kovarianz (empirische Kovarianz)

Seien X und Y zwei kardinalskalierte Merkmale. Dann ist das arithmetische Mittel aus dem Produkt der jeweiligen Abweichungen vom Mittelwert als empirische Kovarianz definiert

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})$$

Anmerkung

- Aus den bisherigen Analysen eines Merkmals ist die Varianz als Güte für die quadratische Streuung der Messwerte um den Mittelwert bekannt:
  $$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(X_i - \bar{x})$$
  Die Kovarianz greift diese Definition der Streuung auf, verschränkt aber die beiden Merkmale durch das Produkt der jeweiligen Abweichung vom Mittelwert. Die Kovarianz ist damit ein Parameter für die gemeinsame Streuung der beiden Merkmale.
- Die nachstehende Formel stellt den Ansatz für die Berechnung mit den Daten aus der Häufigkeitsverteilung dar.
  $$s_{x,y} = \sum_{k=1}^{m_x}\sum_{l=1}^{m_y} h_{k,l}(x_k - \bar{x})(y_l - \bar{y})$$

Kovarianz unter Berücksichtigung der Freiheitsgrade (Streuungsparameter)

Seien X und Y zwei kardinalskalierte Merkmale. Dann ist das arithmetische Mittel aus dem Produkt der jeweiligen Abweichungen vom Mittelwert als Kovarianz definiert

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})$$

Kovarianz unter Berücksichtigung des Verschiebungssatzes

Seien X und Y zwei kardinalskalierte Merkmale. Dann gilt für die empirische Kovarianz die folgende Beziehung:

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) = \frac{1}{n}(\sum_{i=1}^{n} X_i Y_i - n\bar{x}\bar{y})$$

Richtung des Zusammenhanges

- Ist das Vorzeichen der Kovarianz positiv $Cov > 0$ so liegt eine gleichläufige Variation vor. D.h., dass die Veränderung eines Merkmals zu einer gleichgerichteten Veränderung des anderen Merkmals führt.
- Ist das Vorzeichen der Kovarianz negativ $Cov < 0$ so liegt eine gegenläufige Variation vor. D.h., dass die Veränderung eines Merkmals führt zu einer gegenläufigen Veränderung des anderen Merkmals.

Anmerkung

- Die vergleichende Interpretation der Kovarianz, d.h. ob die Streuung hoch oder niedrig ist, ist derzeit noch nicht möglich, da die Kovarianz maßstabsabhängig ist.

3.2 Herleitungen & Berechnungen

3.2.1 Aufbau & Interpretation von Kreuztabellen

Urliste => Kreuztabelle

Nachstehend sind zwei Formen einer Urliste (nur Ausschnitte: Excel und SPSS) mit zwei Merkmalen abgebildet.

Urliste in Excel

	Einkommen	Konsum
i	X_i	Y_i
1	10	5
2	10	5
3	10	10
4	10	10
5	10	10
6	10	10
7	10	10
8	10	10
9	10	10
10	10	10
11	15	10
12	15	10

Urliste in SPSS

	Einkommen	Konsum
1	10,00	5,00
2	10,00	5,00
3	10,00	10,00
4	10,00	10,00
5	10,00	10,00
6	10,00	10,00
7	10,00	10,00
8	10,00	10,00
9	10,00	10,00
10	10,00	10,00
11	15,00	10,00
12	15,00	10,00

Eine Häufigkeitsverteilung für zwei oder mehr Merkmale wird auch als Kreuztabelle bezeichnet. Soll diese manuell anhand der obigen Daten erzeugt werden, so geht man sukzessiv entsprechend der nachstehenden Kreuztabelle vor. Die erste Zelle beinhaltet alle Fälle, die ein Einkommen von 10 GE und einen Konsum von 5 GE aufweisen. Die Anzahl dieser Fälle ermittelt man anhand der Urliste, was in diesem Fall 2 Fälle sind. Bei den anderen Zellen ist analog vorzugehen, was dann zu der folgenden kompletten Kreuztabelle führt.

		Einkommen			
		10	**15**	**20**	**25**
Konsum	**5**	2			
	10	8	5	1	1
	15		5	7	8
	20			2	8
	25				3

In SPSS ist dieser Prozess wesentlich schneller zu realisieren. Auch in Excel lassen sich sehr einfach Kreuztabellen über das Menü „Pivot-Tabellen“ erzeugen.

		Einkommen				
		10,00	15,00	20,00	25,00	Gesamt
Konsum	5,00	2	0	0	0	2
	10,00	8	5	1	1	15
	15,00	0	5	7	8	20
	20,00	0	0	2	8	10
	25,00	0	0	0	3	3
Gesamt		10	10	10	20	50

Kreuztabelle mit absoluten Häufigkeiten

Nachstehend wird durch die Pfeile der Zusammenhang zwischen den Definitionen und der Kreuztabelle verdeutlicht.

X \ Y	y_1	y_2	$\cdots$	y_l	$\cdots$	y_{m_Y}	h
x_1	$h_{1,1}$	$h_{1,2}$	$\cdots$	$h_{1,l}$	$\cdots$	h_{1,m_Y}	$h(x_1)$
x_2	$h_{2,1}$	$h_{2,2}$	$\cdots$	$h_{2,l}$	$\cdots$	h_{1,m_Y}	$h(x_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	$h_{k,1}$	$h_{k,2}$	$\cdots$	$h_{k,l}$	$\cdots$	h_{k,m_Y}	$h(x_k)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_{m_X}	$h_{m_X,1}$	$h_{m_X,2}$	$\cdots$	$h_{m_X,l}$	$\cdots$	h_{m_X,m_Y}	$h(x_{m_X})$
h	$h(y_1)$	$h(y_2)$	$\cdots$	$h(y_l)$	$\cdots$	$h(y_{m_Y})$	n

		Einkommen				
		10,00	15,00	20,00	25,00	Gesamt
Konsum	5,00	2	0	0	0	2
	10,00	8	5	1	1	15
	15,00	0	5	7	8	20
	20,00	0	0	2	8	10
	25,00	0	0	0	3	3
Gesamt		10	10	10	20	50

Randverteilungen

Die bisher bekannten eindimensionalen Verteilungen entsprechenden Randverteilung, d.h. der Summenspalte bzw. Summenzeile.

Einkommen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	10,00	10	20,0	20,0	20,0
	15,00	10	20,0	20,0	40,0
	20,00	10	20,0	20,0	60,0
	25,00	20	40,0	40,0	100,0
	Gesamt	50	100,0	100,0	

Konsum

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	5,00	2	4,0	4,0	4,0
	10,00	15	30,0	30,0	34,0
	15,00	20	40,0	40,0	74,0
	20,00	10	20,0	20,0	94,0
	25,00	3	6,0	6,0	100,0
	Gesamt	50	100,0	100,0	

Relative Häufigkeiten

Kreuztabellen mit ausschließlich absoluten Häufigkeiten sind schwer zu interpretieren. Durch relative Häufigkeiten wird dieses deutlich verbessert. Bei Kreuztabellen werden drei Arten von relativen Häufigkeiten unterschieden:

- Relative Häufigkeit auf die Gesamtsumme
- Bedingte relative Häufigkeit bezüglich der Spalten
- Bedingte relative Häufigkeit bezüglich der Zeile

- **Relative Häufigkeit auf die Gesamtsumme**

  Bei der relativen Häufigkeit auf die Gesamtsumme wird der prozentuale Anteil der Zelle in Relation zur Gesamtanzahl der Fälle bestimmt. Die erste Zelle beinhaltet zwei Fälle, wodurch sich der relative Anteil von $\frac{2}{50} = 0{,}04$ oder 4% ergibt.

			Einkommen				Gesamt
			10,00	15,00	20,00	25,00	
Konsum	5,00	Anzahl	2	0	0	0	2
		% der Gesamtzahl	4,0%	0,0%	0,0%	0,0%	4,0%
	10,00	Anzahl	8	5	1	1	15
		% der Gesamtzahl	16,0%	10,0%	2,0%	2,0%	30,0%
	15,00	Anzahl	0	5	7	8	20
		% der Gesamtzahl	0,0%	10,0%	14,0%	16,0%	40,0%
	20,00	Anzahl	0	0	2	8	10
		% der Gesamtzahl	0,0%	0,0%	4,0%	16,0%	20,0%
	25,00	Anzahl	0	0	0	3	3
		% der Gesamtzahl	0,0%	0,0%	0,0%	6,0%	6,0%
Gesamt		Anzahl	10	10	10	20	50
		% der Gesamtzahl	20,0%	20,0%	20,0%	40,0%	100,0%

  Diese relativen Häufigkeiten sind für Interpretationen meist auch nicht brauchbar. Hierfür sind die bedingten relativen Häufigkeiten besser geeignet.

- **Bedingte relative Häufigkeiten für Spalten (Einkommen)**

  Die nachstehende Definition für die bedingte relative Häufigkeit einer Spalte soll an der ersten Spalte erläutert werden.

$$f(x_k|y_l) = \frac{h(x_k; y_l)}{h(y_l)}$$

  Man spricht hier von einer bedingten Häufigkeit, da in der ersten Spalte prozentualen Werte unter der Bedingung, dass das Einkommen 10 GE entspricht, berechnet werden. x_k ist die absolute Häufigkeit der k-ten Zelle unter der Bedingung y_l.

			Einkommen				Gesamt
			10,00	15,00	20,00	25,00	
Konsum	5,00	Anzahl	2	0	0	0	2
		% innerhalb von Einkommen	20,0%	0,0%	0,0%	0,0%	4,0%
	10,00	Anzahl	8	5	1	1	15
		% innerhalb von Einkommen	80,0%	50,0%	10,0%	5,0%	30,0%
	15,00	Anzahl	0	5	7	8	20
		% innerhalb von Einkommen	0,0%	50,0%	70,0%	40,0%	40,0%
	20,00	Anzahl	0	0	2	8	10
		% innerhalb von Einkommen	0,0%	0,0%	20,0%	40,0%	20,0%
	25,00	Anzahl	0	0	0	3	3
		% innerhalb von Einkommen	0,0%	0,0%	0,0%	15,0%	6,0%
Gesamt		Anzahl	10	10	10	20	50
		% innerhalb von Einkommen	100,0%	100,0%	100,0%	100,0%	100,0%

- **Bedingte relative Häufigkeiten für Zeilen (Konsum)**

  Bei den bedingten relativen Häufigkeiten werden die prozentualen Werte für die jeweilige Zeile bestimmt und die Argumentation gilt analog.

  $$f(y_l|x_k) = \frac{h(x_k; y_l)}{h(x_k)}$$

  y_l ist die absolute Häufigkeit der l-ten Zelle unter der Bedingung x_k.

			Einkommen				Gesamt
			10,00	15,00	20,00	25,00	
Konsum	5,00	Anzahl	2	0	0	0	2
		% innerhalb von Konsum	100,0%	0,0%	0,0%	0,0%	100,0%
	10,00	Anzahl	8	5	1	1	15
		% innerhalb von Konsum	53,3%	33,3%	6,7%	6,7%	100,0%
	15,00	Anzahl	0	5	7	8	20
		% innerhalb von Konsum	0,0%	25,0%	35,0%	40,0%	100,0%
	20,00	Anzahl	0	0	2	8	10
		% innerhalb von Konsum	0,0%	0,0%	20,0%	80,0%	100,0%
	25,00	Anzahl	0	0	0	3	3
		% innerhalb von Konsum	0,0%	0,0%	0,0%	100,0%	100,0%
Gesamt		Anzahl	10	10	10	20	50
		% innerhalb von Konsum	20,0%	20,0%	20,0%	40,0%	100,0%

Interpretation

Bei der Tabelle mit den Spaltenprozenten steht das Einkommen in den Spalten. Es ist festzustellen, dass mit höherem Einkommen zwar der Konsum steigt, jedoch nicht mehr das gesamte Einkommen konsumiert wird. Volkswirte sprechen in diesem Zusammenhang mit einer zunehmenden Sparneigung bzw. von einer abnehmenden Konsumneigung, die auch plausibel ist. Die Tabelle mit den Zeilenprozenten lässt sich diesbezüglich deutlichschlechter interpretieren.

Empfehlungen für den Tabellenaufbau

- Die **relativen** Häufigkeiten auf die Gesamtanzahl sind erfahrungsgemäß seltener für die Interpretation notwendig.
- Die **bedingten relativen** Zeilen- oder Spaltenhäufigkeiten sind hingegen häufig erforderlich.
  - Grundsätzlich ist es vorteilhaft, die bedingten relativen Häufigkeiten aus Sicht der unabhängigen Größe und **nicht** aus Sicht der abhängigen Größe berechnen.
  - Bei dem obigen Beispiel wird die Auswirkung des Einkommens auf den Konsum untersucht. Das Einkommen ist hier die **unabhängige** und der Konsum die **abhängige** Größe.
  - Da die unabhängige Variable das Einkommen ist und in den Spalten steht, ist mit den Spaltenprozenten zu arbeiten.
  - Grundsätzlich käme man zu dem gleichen, gut interpretierbaren Ergebnis, wenn das Einkommen in den Zeilen angeordnet wäre und die Zeilenprozente ausgewiesen würden. Erfahrungsgemäß ist aber die Anordnung der unabhängigen Größe in den Spalten für den Betrachter leichter zu erfassen.
- Des Weiteren sollte Folgendes beachtet werden:
  - Tabellen sollten zur leichteren Interpretation möglichst einfach aufgebaut werden. Dabei sind mehr als zwei Dimensionen, die grundsätzlich denkbar sind, zu vermeiden.
  - Zellen sollten neben den absoluten Häufigkeiten **mindestens eine** aber auf jeden Fall auch **nur eine bedingte Häufigkeit** enthalten.
  - Während einer Präsentation sollte nicht zwischen den Interpretationsweisen gewechselt werden.

3.2.2 Kovarianz

3.2.2.1 Herleitung

Die Berechnung der Varianz bei einem Merkmal ist bekannt. Das Quadrieren der Differenz $X_i - \bar{x}$ lässt sich auch in Form eines Produktes schreiben:

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

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x}) * (Y_i - \bar{y})$$

Diese Schreibweise zeigt sehr deutlich die Analogie von Varianz und Kovarianz. Der Unterschied besteht darin, dass bei der Kovarianz das Produkt aus den Differenzen beider Merkmale besteht und auf diese Weise beide Merkmale miteinander verschränkt werden.

Auch bei der Kovarianz kann mit Hilfe des Verschiebungssatzes die Berechnung vereinfacht werden. Vor dem Beweis sei noch einmal auf die folgenden Vorüberlegungen verwiesen

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \Leftrightarrow n\bar{x} = \sum_{i=1}^{n} X_i \ ,$$

die natürlich für beide Merkmale X und Y gilt.

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})$$

$$\Leftrightarrow s_{x,y} = \frac{1}{n}(\sum_{i=1}^{n} X_iY_i - X_i\bar{y} - Y_i\bar{x} + \bar{x}\bar{y})$$

$$\Leftrightarrow s_{x,y} = \frac{1}{n}(\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\bar{y} - \sum_{i=1}^{n} Y_i\bar{x} + \sum_{i=1}^{n} \bar{x}\bar{y})$$

$$\Leftrightarrow s_{x,y} = \frac{1}{n}(\sum_{i=1}^{n} X_iY_i - \bar{y}\sum_{i=1}^{n} X_i - \bar{x}\sum_{i=1}^{n} Y_i + \mathrm{n}\bar{x}\bar{y})$$

$$\Leftrightarrow s_{x,y} = \frac{1}{n}(\sum_{i=1}^{n} X_iY_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + \mathrm{n}\bar{x}\bar{y})$$

$$\Leftrightarrow s_{x,y} = \frac{1}{n}(\sum_{i=1}^{n} X_iY_i - n\bar{y}\bar{x} - n\bar{x}\bar{y} + \mathrm{n}\bar{x}\bar{y})$$

$$\Leftrightarrow s_{x,y} = \frac{1}{n}(\sum_{i=1}^{n} X_iY_i - n\bar{y}\bar{x}\ )$$

Auch hier könnte eine weitere Vereinfachung erfolgen:

$$\Leftrightarrow s_{x,y} = \frac{1}{n}\sum_{i=1}^{n} X_iY_i\ - \bar{y}\bar{x}$$

Da der Term $\frac{1}{n}\sum_{i=1}^{n} X_iY_i$ dem Mittelwert von X_iY_i entspricht ist auch die folgende Schreibweise möglich:

$$\Leftrightarrow s_{x,y} = \overline{xy} - \bar{y}\bar{x}$$

3.2.2.2 Berechnungen

Die Kovarianz ist die entscheidende und in diesem Zusammenhang neue Größe bei der Betrachtung zweier Merkmale. Die Berechnung der Varianz bei einem Merkmal ist bekannt. Nun lässt sich die Quadratur der Differenz $X_i - \bar{x}$ auch in Form eines Produktes schreiben:

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

Diese Schreibweise zeigt sehr deutlich die Analogie von Varianz und Kovarianz. Der Unterschied besteht darin, dass bei der Kovarianz das Produkt aus den Differenzen beider Merkmale besteht und auf diese Weise beide Merkmale miteinander verschränkt werden.

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})$$

Auch bei der Kovarianz gibt es verschiedene Möglichkeiten der Berechnung. Die gebräuchlichsten sind in diesem Fall aber die mittels der Urliste und die mittels des Verschiebungssatzes. Der Ansatz über die Kreuztabelle wird in diesem Fall weniger verwendet.

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) = \frac{1}{n}(\sum_{i=1}^{n} X_i Y_i - n\bar{x}\bar{y})$$

Bei den nachstehenden Beispielen werden exemplarisch die Kovarianzen berechnet und diskutiert. Die nachstehenden Daten stellen die Gewinn und DV-Kosten zweier Unternehmen über mehrere Jahre dar. Der Zusammenhang ist in beiden Fällen mittels der Kovarianz zu untersuchen.

	Unternehmen 1			***Unternehmen 2***		
Jahr	**Gewinn**	**DV-Kosten**	**X_i*Yi**	**Gewinn**	**DV-Kosten**	**X_i*Yi**
1	10	30	300	100	300	30.000
2	15	30	450	150	300	45.000
3	15	100	1.500	150	1.000	150.000
4	20	50	1.000	200	500	100.000
5	20	100	2.000	200	1.000	200.000
6	25	80	2.000	250	800	200.000
7	30	50	1.500	300	500	150.000
8	30	100	3.000	300	1.000	300.000
9	30	250	7.500	300	2.500	750.000
10	35	180	6.300	350	1.800	630.000
11	35	330	11.550	350	3.300	1.155.000
12	40	200	8.000	400	2.000	800.000
13	45	400	18.000	450	4.000	1.800.000
14	50	500	25.000	500	5.000	2.500.000
15	50	600	30.000	500	6.000	3.000.000
Mittelwert	30	200	118.100	300	2.000	11.810.000
$s_{x,y}$ (n)		1.873,3			187.333,3	
$s_{x,y}$ (n-1)		2.007,1			200.714,3	

Kovarianz Unternehmen 1

- **Urliste**

$$s_{x,y} = \frac{1}{15-1} * ((10-30) * (30-200) + (15-30 * (30-200) \dots)$$

$$s_{x,y} = \frac{1}{14} * 28.100 = 2.007{,}143$$

- **Verschiebungssatz**

$$s_{x,y} = \frac{1}{15-1} * (118.100 - 15 * 30 * 200) = 2.007{,}143$$

Kovarianz Unternehmen 2

- **Urliste**

$$s_{x,y} = \frac{1}{15-1} * ((100-300) * (300-2.000) + (150-300) * (300-2.000) \dots)$$

$$s_{x,y} = \frac{1}{14} * 2.810.000 = 200.714{,}3$$

- **Verschiebungssatz**

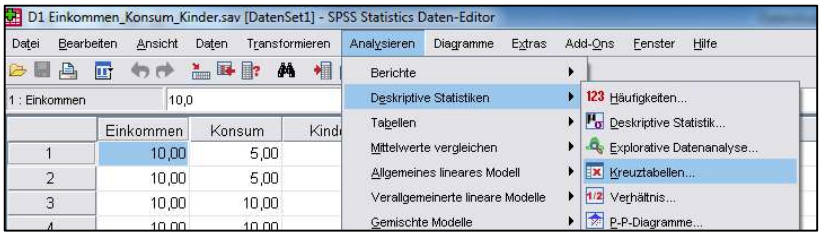
$$s_{x,y} = \frac{1}{15-1} * (11.810.000 - 15 * 300 * 2.000) = 200.714{,}3$$

Interpretation

- In beiden Fällen ist die Kovarianz positiv, d.h. der Zusammenhang ist **gleichgerichtet** und somit führt die Steigerung der unabhängigen Größe zu einer Steigerung der abhängigen Größe.
- Obwohl beide Grafiken bis auf den Maßstab vergleichbar sind und man eine identische Kovarianz vermuten würde, ist diese mit $s_{x,y} = 2.007{,}43$ bzw. $s_{x,y} = 200.714{,}3$ unterschiedlich. Beide Kovarianzen unterscheiden sich um den Faktor $10 * 10 = 100$, der sich aus den mit 10 multiplizierten Merkmalswerten ergibt.
- **Die Kovarianz ist somit maßstabsabhängig und in dieser Form nur wenig aussagefähig. Dieser Mangel wird aber bei späteren Berechnungen korrigiert.**

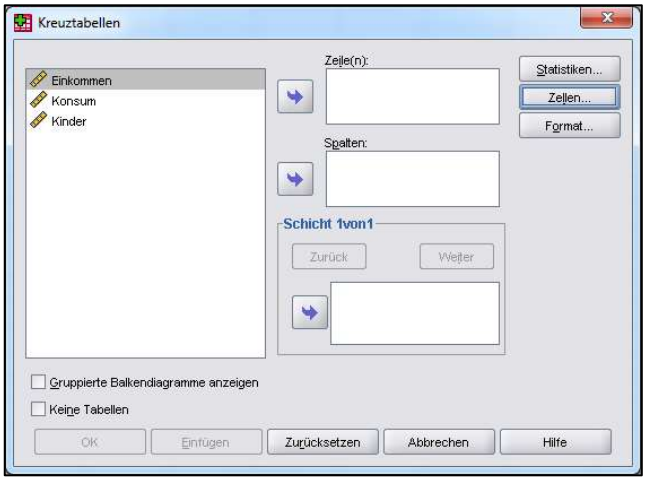
3.3 SPSS

Auswahl des Menü-Items „Kreuztabellen“



Grundstruktur der Tabelle

Fenster für die grundlegende Struktur der Tabelle, d.h. die zu analysierenden Variablen und deren Anordnung



Erläuterung

- Die Felder „Zeilen“ und „Spalten“ enthalten die Variablen, die in diesen Bereichen angezeigt werden sollen.
- Stehen in diesen Felder mehrere Variablen werden mehrere Kreuztabellen erzeugt.
- Über das Feld „Schicht 1 von 1“ besteht die Option einer weiteren Differenzierung der Zeilenwerte, womit im Grunde eine dreidimensionale Häufigkeitsverteilung generiert wird. Dabei wird zuerst nach der in „Schicht 1 von 1“ stehenden Variablen und dann nach der in „Zeilen“ stehen Variablen differenziert.

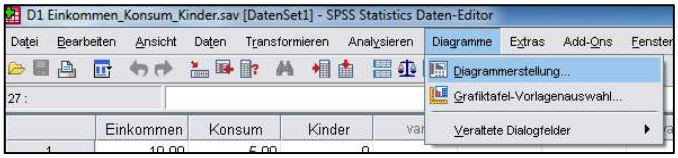
Zellen-Inhalte definieren

Über den Button „Zellen …“ gelangt man zum nachstehenden Fenster, in dem die anzuzeigenden Werte in den Zellen festgelegt werden. Neben den „Beobachteten Häufigkeiten“ lassen sich dort auch die bedingten relativen Häufigkeiten selektieren.

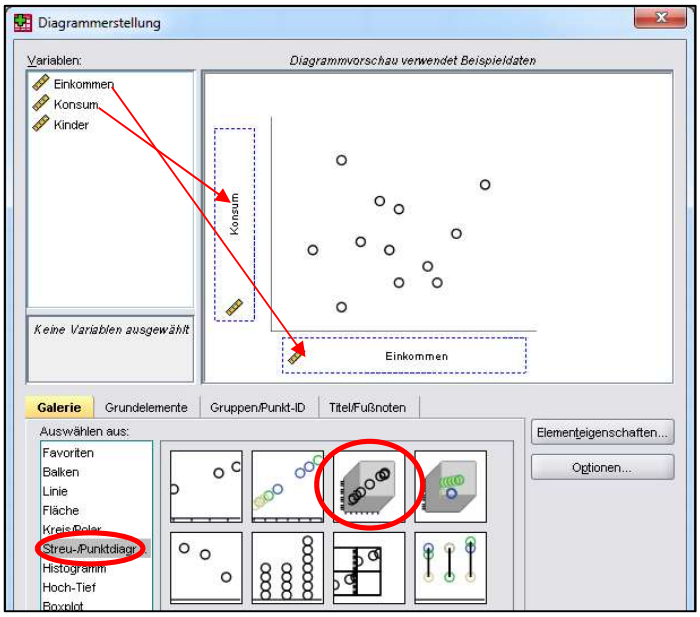
Graphiken

Über das Item „Diagramme“ -> „Diagrammdarstellung“ wird der Dialog für graphische Darstellungen geöffnet.

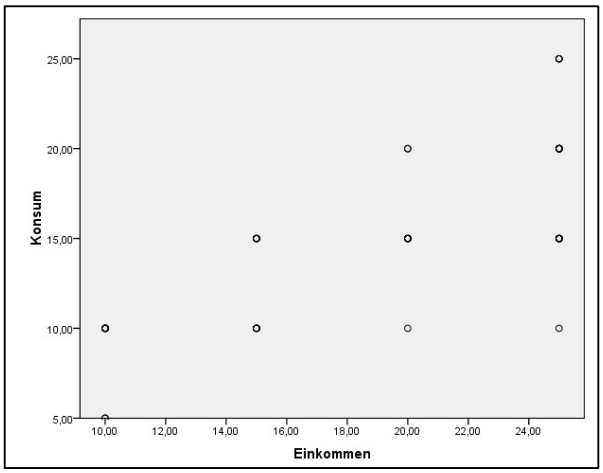
- Menü-Item „Diagrammerstellung“



- Optionsfenster für Diagrammerstellung
  - In diesem Fall ist die Diagrammform „Streu- / Punktdiagramm“ sinnvoll.
  - Von den möglichen Optionen für Streudiagramme ist die eingekreiste Option zu wählen.
  - Die Variable „Einkommen“ ist dann als die unabhängige Variable auch die X-Achse zu ziehen und die Variable „Konsum“ als abhängige Variable auf die Y-Achse.
  - Die letzte Grafik zeigt dann das unter diesen Bedingungen erzeugt Streudiagramm.



- Einfaches Streudiagramm



4 Zusammenfassung zur deskriptiven Statistik

4.1 Univariate Statistik

Nicht klassierte Häufigkeitsverteilungen

statistische Größe	*Benennung*
Statistische Einheit	ω
Anzahl der Elemente der Gesamtheit	n
Bezeichnung der Elemente der Gesamtheit	$\omega_1; \omega_2; \omega_3; \dots\ \omega_n$ ω_i für $i = 1\ bis\ n$
Statistische Gesamtheit	$\Omega(\omega_1, \omega_2, \omega_3, \dots, \omega_n)$ $\Omega(\omega_i)$ mit $i = 1, \dots, n$
Merkmal	X
Merkmalswert	$X(\omega_i) = X_i$
Anzahl der möglichen Merkmalsrealisationen	m
Bezeichnung der Merkmalsrealisationen	$x_1;\ x_2;\ x_3; \dots;\ x_m$ x_j für $j = 1\ bis\ m$
absolute Häufigkeit bezogen auf eine bestimmte Ausprägung x_j	$h(x_j) = h_j$
relative Häufigkeit bezogen auf eine bestimmte Ausprägung x_j	$f(x_j) = f_j = \frac{h(x_j)}{n}$
Relative Summenhäufigkeit	$F(x_j) = f(X \le x_j) = \sum_{r=1}^{j} f_r$

Häufigkeitsfunktion (relative Häufigkeiten)	$f(x) = \begin{cases} f(x_j); \text{für } x = x_j \\ 0; \text{für } x \neq x_j \end{cases}$
Verteilungsfunktion (relative Häufigkeiten)	$F(a) = \begin{cases} 0 \text{ für alle } a < x_i \\ F(x_j) \text{ für alle } x_j \le a \le x_{j+1}; j = 1; 2 \dots m-1 \\ 1 \text{ für alle } a \ge x_m \end{cases}$

Klassierte Häufigkeitsverteilungen

statistische Größe	*Benennung*
Anzahl der möglichen Ausprägungen (Anzahl der Klassen oder Intervalle)	m
Bezeichnung der Ausprägungen anhand der Klassenmitte (daher x_k^* statt x_j als Mittelwert der Klasse)	$x_1^*, x_2^*, x_3^*, \dots, x_m^*$ x_k^* für $k = 1\ bis\ m$
absolute Häufigkeit bezogen auf eine bestimmte Ausprägung $\bar{x}_j$	$h(x_k^*) = h_k$
relative Häufigkeit bezogen auf eine bestimmte Ausprägung $\bar{x}_j$	$f(x_k^*) = f_k$
Häufigkeitsdichte	$h^D(x_k^*) = h_k^D = \frac{h_k}{\Delta x_k}$

Lageparameter

Median (Zentralwert)

$$\bar{x}_Z = \begin{cases} X^*_{(\frac{n+1}{2})} \text{ für n ungerade} \\ \frac{1}{2}(X^*_{\frac{n}{2}} + X^*_{\frac{n}{2}+1} \text{ für n gerade} \end{cases}$$

Modus (Häufigster Wert)

Für nichtklassierte Daten gilt
$\bar{x}_M$: dasjenige x_j, für das $h(x_j)$ maximal ist

Bei klassierten Daten gilt:
$\bar{x}_M$: dasjenige x_k^* , für das $h(x_k^*)$ maximal ist

Arithmetisches Mittel

$$\bar{x} = \frac{X_1 + X_2 + X_3 + \cdots X_n}{\mathrm{n}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{j=1}^{m} h_j * x_j \; = \sum_{j=1}^{m} f_j * x_j$$

Geometrisches Mittel

$$\bar{\mathrm{x}}_{\mathrm{G}} = \sqrt[\mathrm{n}]{\mathrm{X}_1 * \mathrm{X}_2 * \mathrm{X}_3 * \ldots * \mathrm{X}_{\mathrm{n}}} = \sqrt[n]{x_1^{h_1} * x_2^{h_2} * x_3^{h_3} * \ldots * x_m^{h_m}}$$

Gewichtetes (arithmetisches) Mittel

$\bar{x}_{gew} = \sum_{i=1}^{n} X_i * g_i$ mit $g_i > 0$ für alle i und $\sum_{i=1}^{n} g_i = 1$
oder
$\bar{x}_{gew} = \frac{\sum_{i=1}^{n} X_i * g_i}{\sum_{i=1}^{n} g_i}$ mit $g_i > 0$ für alle i.

Streuungsparameter

Spannweite (Range)

$$w = \max_{i=1,\ldots,n} X_i - \min_{i=1,\ldots,n} X_i$$

Mittlere absolute Abweichung

$$d = \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{x}_z| = \frac{1}{n}\sum_{j=1}^{m} h_j * |x_j - \bar{x}_z| = \sum_{j=1}^{m} f_j * |x_j - \bar{x}_z|$$

$$d = \frac{1}{n-1}\sum_{i=1}^{n} |X_i - \bar{x}_z|$$

Varianz

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{m} h_j * (x_j - \bar{x})^2 = \sum_{j=1}^{m} f_j * (x_j - \bar{x})^2$$

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

Standardabweichung

$$s = +\sqrt{s^2}$$

Variationskoeffizient

$$v = \frac{s}{\bar{x}}$$

Zulässigen Parameter in Abhängigkeit vom Skalenniveau.

Kollektivmaß	Skalenniveau			
	Nominal-skala	Ordinal-skala	Intervall-skala	Verhältnis-skala
Lageparameter				
Modus	Ja	Ja	Ja	Ja
Median		Ja	Ja	Ja
Arithmetisches Mittel			Ja	Ja
Geometrisches Mittel				Ja
Streuungsparameter				
Spannweite		Ja	Ja	Ja
Quartile / Quartilsabstand		Ja	Ja	Ja
Mittlere absolute Abweichung			Ja	Ja
Varianz / Standardabweichung			Ja	Ja
Variationskoeffizient				Ja

4.2 Bivariate Statistik

statistische Größe	***Benennung***
Statistische Einheit	ω
Anzahl der Elemente der Gesamtheit	n
Bezeichnung der Elemente der Gesamtheit	$\omega_1; \omega_2; \omega_3; \dots\ \omega_n$ ω_i für $i = 1\ bis\ n$
Statistische Gesamtheit	$\Omega(\omega_1, \omega_2, \omega_3, \dots, \omega_n)$ $\Omega(\omega_i)$ mit $i = 1, \dots, n$
Merkmal	X, Y
Merkmalswert	$X(\omega_i) = X_i$ $Y(\omega_i) = Y_i$

Anzahl der möglichen Merkmalsrealisationen	m_x bzw. m_y
Bezeichnung der Merkmalsrealisationen	$x_1;\ x_2;\ x_3; \dots;\ x_{m_x}$ x_k für $k = 1\ bis\ m_x$ $y_1;\ y_2;\ y_3; \dots;\ y_{m_y}$ y_l für $l = 1\ bis\ m_y$
absolute Häufigkeit für $(x_k; y_l)$	$h(x_k; y_l)$
relative Häufigkeit für $(x_k; y_l)$	$f(x_k; y_l) = \frac{h(x_k; y_l)}{n}$
bedingte relative Häufigkeit für Spalten $(x_k; y_l)$	$f(x_k \mid y_l) = \frac{h(x_k; y_l)}{h(y_l)}$
bedingte relative Häufigkeit für Zeilen $(x_k; y_l)$	$f(x_k \mid y_l) = \frac{h(x_k \mid y_l)}{h(x_k)}$
Randhäufigkeit bezüglich der Ausprägung x_k	$h(x_k) = \sum_{l=1}^{m_y} h(x_k \mid y_l)$
Randhäufigkeit bezüglich der Ausprägung y_l	$h(y_l) = \sum_{k=1}^{m_x} h(x_k \mid y_l)$

Kovarianz

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) = \frac{1}{n}(\sum_{i=1}^{n} X_i Y_i - n\bar{y}\bar{x}\,)$$