Univariate Verteilungen

(1) Analyse: "deskriptive Statistiken"

„Analysieren“ -> „deskriptive Statistiken“
-> „deskriptive Statistik “

Keine tabellarische Darstellung der Häufigkeitsverteilung

(2) Analyse: "Häufigkeitsverteilung"

„Analysieren“ -> „deskriptive Statistiken“
-> „Häufigkeitsverteilung“

Tabellarische Häufigkeitsverteilungen

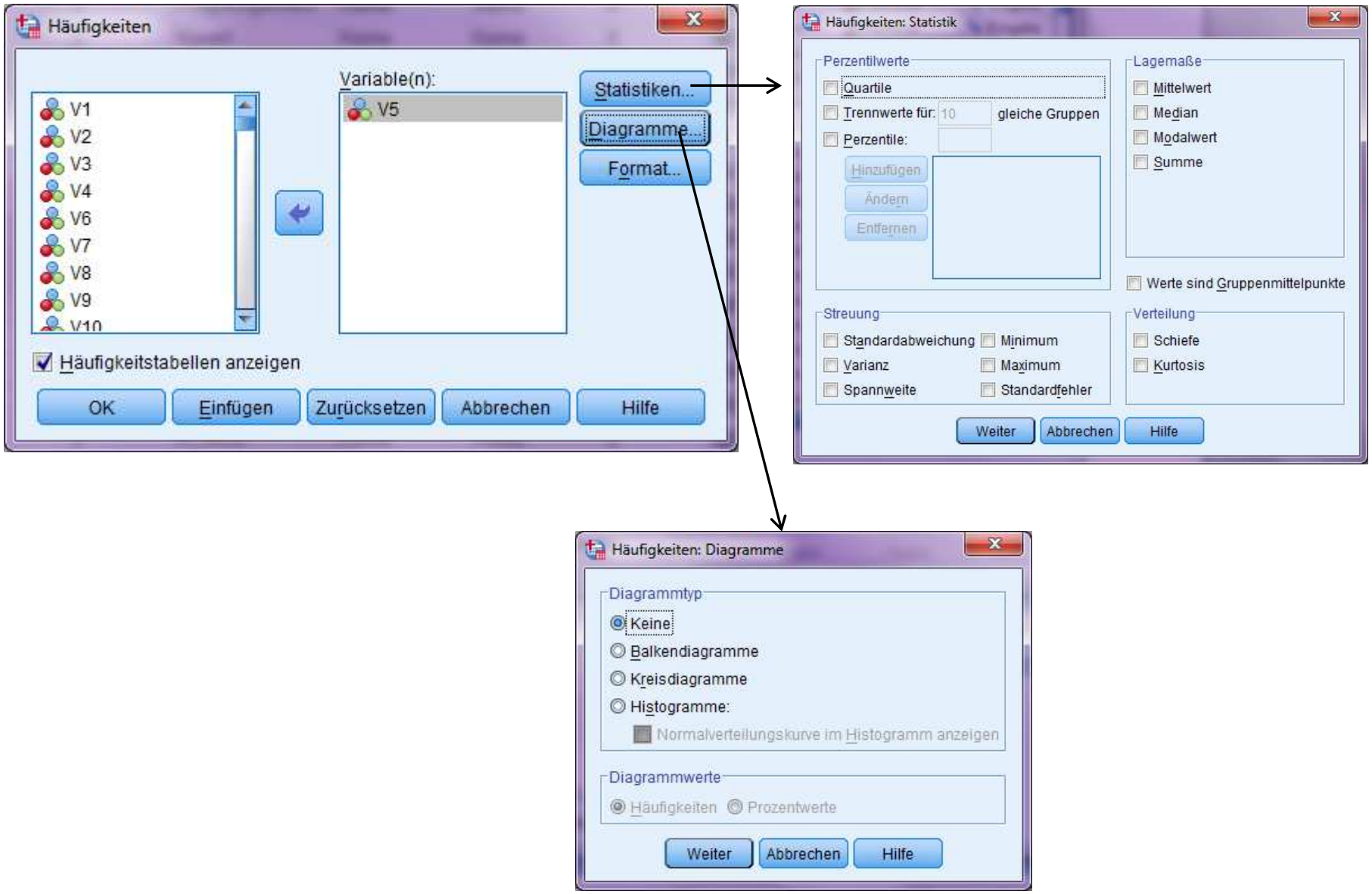
Quartile, Perzentile, Median
Modalwert,
Mittelwert

Varianz, Standardabweichung
Spannweite,

Standardfehler

Balkendiagramm,
Histogramm

Optionen "Häufigkeitsverteilung"



Zugehöriger Syntaxbefehl: Frequencies

Beispiel:

frequencies v5.

oder:

frequ v5.

Liefert lediglich tabellarische Häufigkeitsverteilung.
Für weitere Optionen siehe Syntaxreferenz oder wähle im Menü „einfügen“.

1 2

V18 Unterstützungspersonen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00	1	,4	,4	,4
	2,00	4	1,6	1,6	2,0
	3,00	12	4,9	4,9	6,9
	4,00	24	9,8	9,8	16,7
	5,00	16	6,5	6,5	23,3
	6,00	18	7,3	7,3	30,6
	7,00	15	6,1	6,1	36,7
	8,00	16	6,5	6,5	43,3
	9,00	6	2,4	2,4	45,7
	10,00	44	17,9	18,0	63,7
	11,00	9	3,7	3,7	67,3
	12,00	20	8,1	8,2	75,5
	13,00	1	,4	,4	75,9
	14,00	8	3,3	3,3	79,2
	15,00	17	6,9	6,9	86,1
	16,00	2	,8	,8	86,9
	17,00	2	,8	,8	87,8
	18,00	3	1,2	1,2	89,0
	19,00	2	,8	,8	89,8
	20,00	13	5,3	5,3	95,1
	21,00	1	,4	,4	95,5
	23,00	1	,4	,4	95,9
	25,00	2	,8	,8	96,7
	26,00	1	,4	,4	97,1
	30,00	3	1,2	1,2	98,4
	35,00	2	,8	,8	99,2
	42,00	1	,4	,4	99,6
	50,00	1	,4	,4	100,0
	Gesamt	245	99,6	100,0	
Fehlend	System	1	,4		
Gesamt		246	100,0		

4 5 3

Häufigkeitsauszählung für V18 („Unterstützungspersonen“)

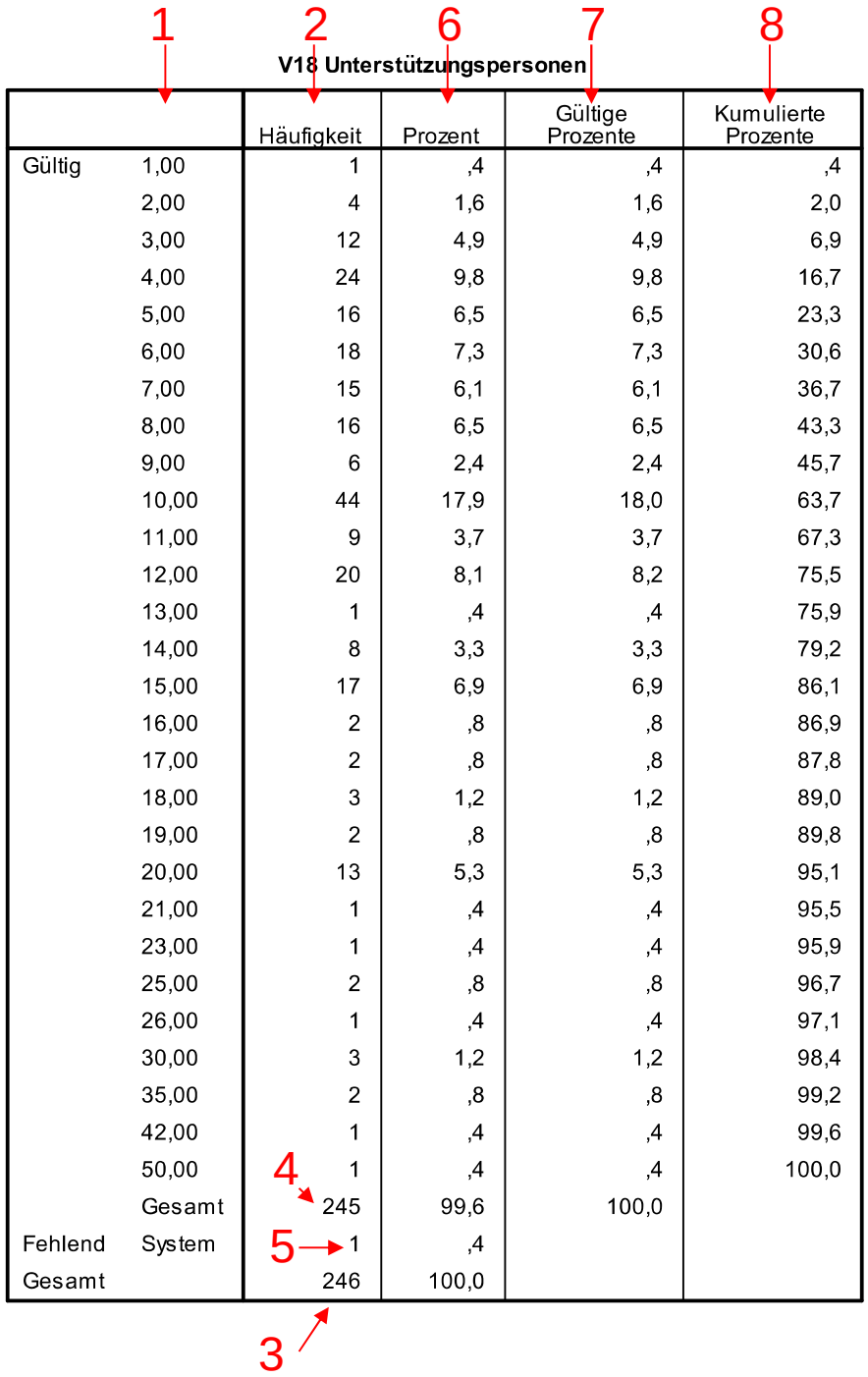
1: Aufgetretene Ausprägungen

2: Absolute Häufigkeiten, mit die einzelnen Ausprägungen aufgetreten sind

3: Gesamt-Fallzahl (Anzahl der Datenvektoren im Datensatz)

4: Valide Fallzahl: Anzahl an Datenvektoren mit gültiger Ausprägung von V18

5: Missing-Fälle: Anzahl an Fällen mit nicht definierten (fehlenden) Werten bei V18.



Häufigkeitsauszählung für V18 („Unterstützungspersonen“)

6: Relative Häufigkeiten inklusive der Missing-Fälle. (Prozentuierungsbasis ist 246)

7: Valide relative Häufigkeiten ohne Missing-Fälle. (Prozentuierungsbasis ist 245)

8: Kumulierte relative Häufigkeiten.
z.B. Rund drei Viertel aller Befragten (75,5%) nannten bis zu 12 Unterstützungspersonen.

Beispiel: Unterschiedliche Arten von Missing-Values in Häufigkeitsverteilungen:

v140 WICHTIGKEIT VON RUHE UND ORDNUNG

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1 AM WICHTIGSTEN	1088	31,8	33,0	33,0
	2 AM ZWEITWICHTIGSTEN	806	23,5	24,5	57,5
	3 AM DRITTWICHTIGSTEN	860	25,1	26,1	83,6
	4 AM VIERTWICHTIGSTEN	541	15,8	16,4	100,0
	Gesamt	3295	96,2	100,0	
Fehlend	8 WEISS NICHT	54	1,6		
	9 KEINE ANGABE	72	2,1		
	System	4	,1		
	Gesamt	130	3,8		
Gesamt		3425	100,0		

Ausprägungen 8 und 9 wurden als benutzerdefiniert fehlend codiert.
Zusätzlich sind System-fehlende Werte vorhanden.

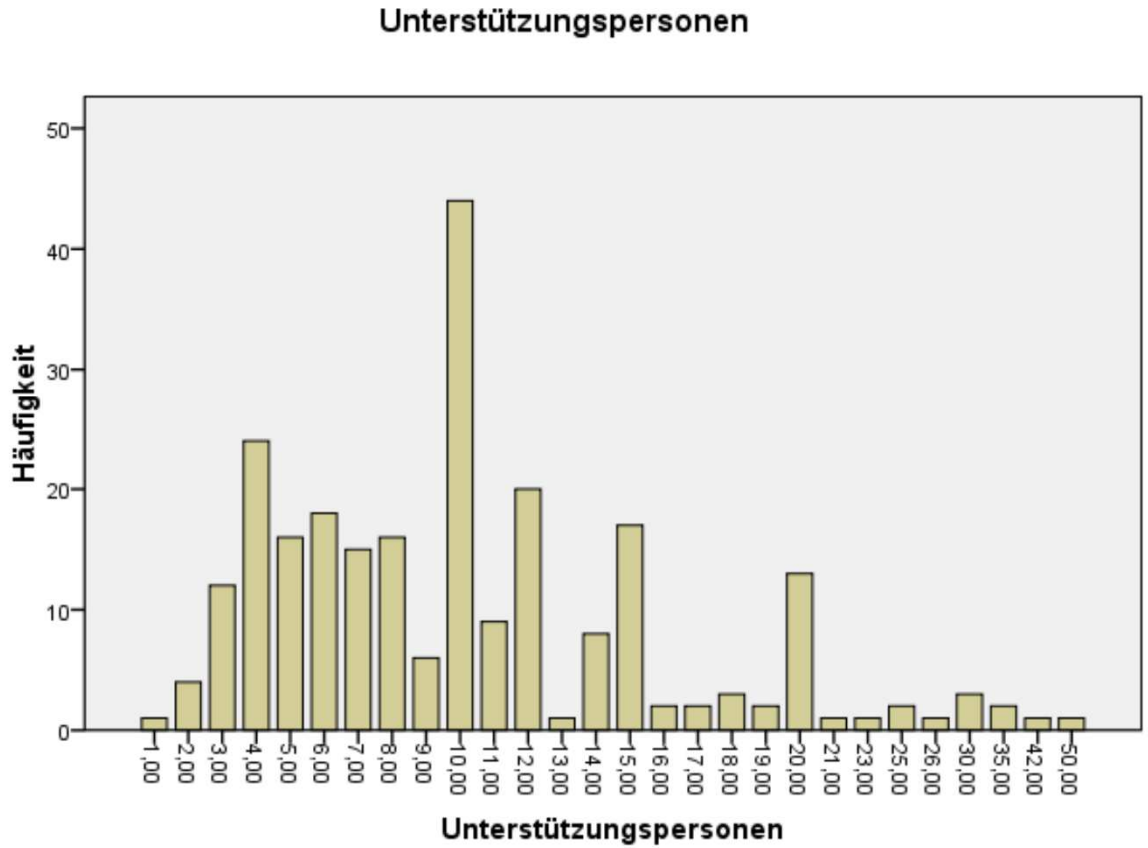
Alle drei Arten fehlender Werte bleiben bei der Berechnung der validen relativen Häufigkeiten unberücksichtigt.

Grafische Veranschaulichung der Verteilung

Balkendiagramm

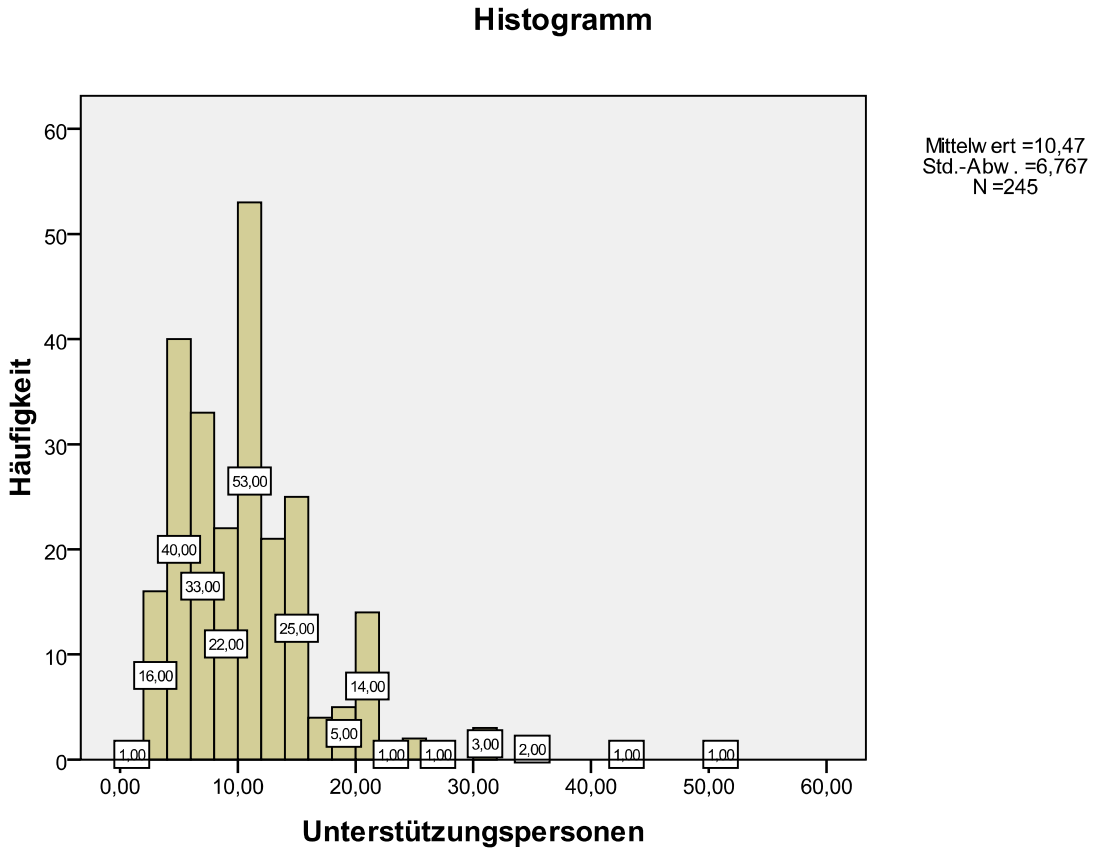
Bei vielen Ausprägungen ungünstig

Nicht besetzte Ausprägungen werden nicht dargestellt. Daher ist die Skalierung der x-Achse nicht konstant!



Grafische Veranschaulichung der Verteilung

Histogramm: Vorteil: Die Skalierung der x-Achse ist konstant.
Nicht besetzte Ausprägungsbereichte werden dargestellt.
Automatische Klasseneinteilung (in diesem Fall in 2-er Schritten)
Die Flächen entsprechen der Häufigkeit.



Lagemaße

Modalwert
ab nominalem Messniveau

Median, Quartile, Perzentile
ab ordinalem Messniveau

Arithmetisches Mittel
ab intervallskalen-Niveau

Streuungsmaße

Varianz
ab intervallskalen-Niveau

Standardabweichung
ab intervallskalen-Niveau

Quartilsabstand
ab ordinalem Messniveau
(als Abstand von Rangplätzen)

Range
ab Intervallskalen Niveau

Lagemaße

Modus: die am häufigsten vorkommende Ausprägung.

Der Modus muss nicht eindeutig sein (z.B. Bimodale Verteilung)

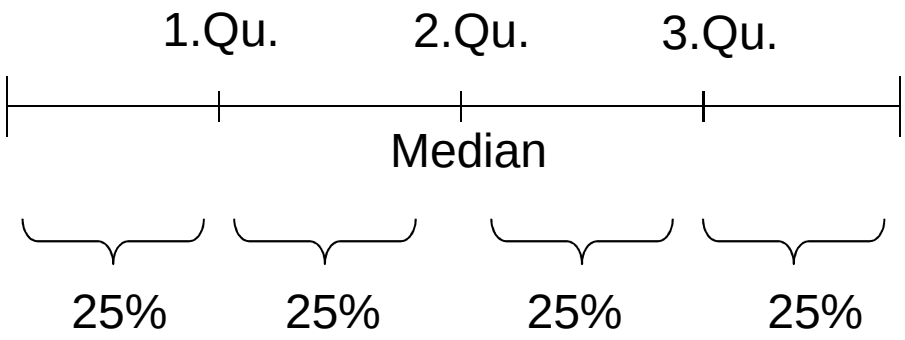
Median: Merkmalswert, der in der Mitte einer geordneten Messreihe liegt.

Über dem Median liegen genau so viele Fälle (50%) wie darunter.

Der Median teilt die Verteilung in zwei Hälften.

Quartile:

Eine geordnete Messreihe wird in vier gleiche Abschnitte unterteilt.
1.Quartil: 25%, 2.Quartil: Median, 3.Quartil: 75%



Perzentile: Verallgemeinerung für beliebige prozentuelle Abschnitte.
Das 25%-Perzentil entspricht dem 1. Quartil

V18 Unterstützungspersonen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00	1	,4	,4	,4
	2,00	4	1,6	1,6	2,0
	3,00	12	4,9	4,9	6,9
	4,00	24	9,8	9,8	16,7
	5,00	16	6,5	6,5	23,3
	6,00	18	7,3	7,3	30,6
	7,00	15	6,1	6,1	36,7
	8,00	16	6,5	6,5	43,3
	9,00	6	2,4	2,4	45,7
	10,00	44	17,9	18,0	63,7
	11,00	9	3,7	3,7	67,3
	12,00	20	8,1	8,2	75,5
	13,00	1	,4	,4	75,9
	14,00	8	3,3	3,3	79,2
	15,00	17	6,9	6,9	86,1
	16,00	2	,8	,8	86,9
	17,00	2	,8	,8	87,8
	18,00	3	1,2	1,2	89,0
	19,00	2	,8	,8	89,8
	20,00	13	5,3	5,3	95,1
	21,00	1	,4	,4	95,5
	23,00	1	,4	,4	95,9
	25,00	2	,8	,8	96,7
	26,00	1	,4	,4	97,1
	30,00	3	1,2	1,2	98,4
	35,00	2	,8	,8	99,2
	42,00	1	,4	,4	99,6
	50,00	1	,4	,4	100,0
	Gesamt	245	99,6	100,0	
Fehlend	System	1	,4		
Gesamt		246	100,0		

Statistiken

V18 Unterstützungspersonen

N	Gültig	245
	Fehlend	1
Median		10,0000
Modus		10,00
Perzentile	25	6,0000
	50	10,0000
	75	12,0000

Arithmetisches Mittel:

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

Das Arithmetische Mittel ist der Schwerpunkt einer Verteilung, von dem die Summe der Abweichungen der einzelnen Werte gleich Null ist:

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$$

Eine Lineartransformation in x führt zur selben Lineartransformation des Mittelwertes:

$$x_i' = a + b * x_i \Rightarrow \bar{x}' = a + b * \bar{x}$$

Wird der Mittelwert mit n multipliziert, dann erhält man die Summe der Merkmalsausprägungen in der Stichprobe.

$$\bar{x} * n = \sum_{i=1}^{n} x_i$$

zu beachten: Das Arithmetische Mittel ist empfindlich gegenüber Ausreißern!

V18 Unterstützungspersonen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00	1	,4	,4	,4
	2,00	4	1,6	1,6	2,0
	3,00	12	4,9	4,9	6,9
	4,00	24	9,8	9,8	16,7
	5,00	16	6,5	6,5	23,3
	6,00	18	7,3	7,3	30,6
	7,00	15	6,1	6,1	36,7
	8,00	16	6,5	6,5	43,3
	9,00	6	2,4	2,4	45,7
	10,00	44	17,9	18,0	63,7
	11,00	9	3,7	3,7	67,3
	12,00	20	8,1	8,2	75,5
	13,00	1	,4	,4	75,9
	14,00	8	3,3	3,3	79,2
	15,00	17	6,9	6,9	86,1
	16,00	2	,8	,8	86,9
	17,00	2	,8	,8	87,8
	18,00	3	1,2	1,2	89,0
	19,00	2	,8	,8	89,8
	20,00	13	5,3	5,3	95,1
	21,00	1	,4	,4	95,5
	23,00	1	,4	,4	95,9
	25,00	2	,8	,8	96,7
	26,00	1	,4	,4	97,1
	30,00	3	1,2	1,2	98,4
	35,00	2	,8	,8	99,2
	42,00	1	,4	,4	99,6
	50,00	1	,4	,4	100,0
	Gesamt	245	99,6	100,0	
Fehlend	System	1	,4		
Gesamt		246	100,0		

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
V18 Unterstützungspersonen	245	1,00	50,00	10,4694	6,76691	45,791
Gültige Werte (Listenweise)	245					

Streuungsmaße / Dispersionsmaße

Lagemaße sind bei manchen Verteilungen bedingt informativ.

Z.B. Durchschnittsnote bei einer Klausur = 3.
Es könnten beispielsweise alle ein Befriedigend haben oder aber 50% sehr gut und 50% nicht genügend.

Daher ist zusätzlich zu einem Lagemaß ein Streuungsmaß angebracht.

Variationsbreite (Range): Spannweite bzw. Differenz zwischen größtem und kleinstem Meßwert.

Der Range kann von Extremwerten stark beeinflusst werden. Daher wird der Range häufig als Differenz von Perzentilwerten berechnet (z.B. Differenz zwischen 5%- und 95%-Perzentil

Quartilsabstand: Abstand zwischen 1. und 3. Quartil

V18 Unterstützungspersonen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00	1	,4	,4	,4
	2,00	4	1,6	1,6	2,0
	3,00	12	4,9	4,9	6,9
	4,00	24	9,8	9,8	16,7
	5,00	16	6,5	6,5	23,3
	6,00	18	7,3	7,3	30,6
	7,00	15	6,1	6,1	36,7
	8,00	16	6,5	6,5	43,3
	9,00	6	2,4	2,4	45,7
	10,00	44	17,9	18,0	63,7
	11,00	9	3,7	3,7	67,3
	12,00	20	8,1	8,2	75,5
	13,00	1	,4	,4	75,9
	14,00	8	3,3	3,3	79,2
	15,00	17	6,9	6,9	86,1
	16,00	2	,8	,8	86,9
	17,00	2	,8	,8	87,8
	18,00	3	1,2	1,2	89,0
	19,00	2	,8	,8	89,8
	20,00	13	5,3	5,3	95,1
	21,00	1	,4	,4	95,5
	23,00	1	,4	,4	95,9
	25,00	2	,8	,8	96,7
	26,00	1	,4	,4	97,1
	30,00	3	1,2	1,2	98,4
	35,00	2	,8	,8	99,2
	42,00	1	,4	,4	99,6
	50,00	1	,4	,4	100,0
	Gesamt	245	99,6	100,0	
Fehlend	System	1	,4		
Gesamt		246	100,0		

Statistiken

V18 Unterstützungspersonen

N	Gültig	245
	Fehlend	1
Spannweite		49,00
Perzentile	5	3,0000
	25	6,0000
	50	10,0000
	75	12,0000
	95	20,7000

Range=50-1=49

Range basierend auf dem Abstand zwischen dem 5. und dem 95. Perzentil= 20.7-3=17,7

Quartilsabstand=12-6=6, d.h., die mittleren 50% der Verteilung gaben zwischen 6 und 12 Unterstützungspersonen an. Sie unterscheiden sich daher in einem Ausmaß von 6 angegebenen Personen.

Exkurs: Perzentilberechnung in SPSS:

Die Ermittlung erfolgt durch einen gewichteten Durchschnitt.

z.B. Beim 95% Perzentil wird zunächst errechnet, dass bei n=245 der Rangplatz, bei dem die ersten 95% der Stichprobe überschritten werden bei 233,7 liegt.
Das Perzentil wäre sozusagen die Ausprägung des 233,7. Falles der geordneten Messreihe.

Da es nur ganzzahlige Rangplätze gibt, wird ein Durchschnitt zwischen der Ausprägung der 233. und der 234. Person als Perzentilwert berechnet.

In den Durchschnitt geht die Ausprägung des 233. Falles mit einem Gewicht von 0.3 und die Ausprägung des 234. Falles mit 0.7 ein.

In diesem Beispiel: 233. Fall besitzt Ausprägung 20, 234.Fall besitzt Ausprägung 21. Daher wird als 95%-Perzentil (0.3*20+0.7*21)/2=20.7 ermittelt.

Varianz:

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

Mittlere quadratische Abweichung vom Mittelwert. Da sich die Einheit der Varianz auf die quadrierte Skalierung von X bezieht, ist diese schwer interpretierbar.

Daher wird die Wurzel der Varianz in Form der **Standardabweichung** berechnet.

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

(„Mittlere Abweichung vom Mittelwert“)

Beispiel: Mittelwert / Standardabweichung

Noten:

3 3 3 3 3

$$\bar{x} = \frac{15}{5} = 3$$

$$s = \sqrt{s^2} = \sqrt{\frac{(3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2 + (3-3)^2}{5}} = \sqrt{\frac{0}{5}} = 0$$

5 5 3 1 1

$$\bar{x} = \frac{15}{5} = 3$$

$$s = \sqrt{s^2} = \sqrt{\frac{(5-3)^2 + (5-3)^2 + (3-3)^2 + (1-3)^2 + (1-3)^2}{5}} = \sqrt{\frac{16}{5}} = 1.7$$

1 2 3 4 5

$$\bar{x} = \frac{15}{5} = 3$$

$$s = \sqrt{s^2} = \sqrt{\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}} = \sqrt{\frac{10}{5}} = 1.4$$

Das quadrieren der Abweichungen vom Mittelwert bei der Varianzberechnung lässt einerseits das Vorzeichen der Abweichung unberücksichtigt (für die Streuung einer Variable ist es unerheblich ob Abweichung vom Mittelwert „überdurchschnittlich“ oder „unterdurchschnittlich“ ist).

Weiterhin bekommen dadurch aber auch größere Abweichungen stärkeres Gewicht.

Wird nicht quadriert sonder der Absolutwert verwendet, dann würde die AD-Streuung (average deviation) resultieren:

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

Beispiel (z.B. erreichte Punkte bei einem Test):

Gruppe 1 (6 Personen): 2, 2, 2, 6, 6, 6. Mittelwert: 24 / 6 = 4
Gruppe 2 (6 Personen): 0, 3, 3, 5, 5, 8. Mittelwert: 24 / 6 = 4

Gruppe 1:

x_i	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$
2	$\lvert 2-4\rvert=2$	$(2-4)^2=4$
2	$\lvert 2-4\rvert=2$	$(2-4)^2=4$
2	$\lvert 2-4\rvert=2$	$(2-4)^2=4$
6	$\lvert 6-4\rvert=2$	$(6-4)^2=4$
6	$\lvert 6-4\rvert=2$	$(6-4)^2=4$
6	$\lvert 6-4\rvert=2$	$(6-4)^2=4$
Σ	12	24
	$AD=\frac{1}{n}\sum \lvert x_i - \bar{x}\rvert = \frac{12}{6} = 2$	$s^2=\frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{24}{6} = 4$ $s=\sqrt{4}=2$

Gruppe 2:

x_i	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$
0	$\lvert 0-4\rvert=4$	$(0-4)^2=16$
3	$\lvert 3-4\rvert=1$	$(3-4)^2=1$
3	$\lvert 3-4\rvert=1$	$(3-4)^2=1$
5	$\lvert 5-4\rvert=1$	$(5-4)^2=1$
5	$\lvert 5-4\rvert=1$	$(5-4)^2=1$
8	$\lvert 8-4\rvert=4$	$(8-4)^2=16$
Σ	12	36
	$AD=\frac{1}{n}\sum \lvert x_i - \bar{x}\rvert = \frac{12}{6} = 2$	$s^2=\frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{36}{6} = 6$ $s=\sqrt{6}=2{,}45$

SPSS berechnet die Varianz nicht nach der dargestellten Formel, sondern die korrigierte Stichprobenvarianz. Dabei wird nicht durch n, sondern durch n-1 dividiert (ein relevanter Unterschied ergibt sich nur bei kleinen Stichproben).

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

Deskriptive Statistik

	N	Minimum	Maximum	Mittelwert	Standardabweichung	Varianz
Gruppe1	6	2,00	6,00	4,0000	2,19089	4,800
Gruppe2	6	,00	8,00	4,0000	2,68328	7,200
Gültige Werte (Listenweise)	6					

Umrechnung auf die zuvor dargestellte Varianz durch Multiplikation mit $\frac{n-1}{n}$

Im Beispiel:

Gruppe 1: korrigierte Varianz=4,8. Varianz=4,8 * 5/6 = 4
Gruppe 2: korrigierte Varianz=7,2. Varianz=7,2 * 5/6 = 6

Exkurs: Korrigierte Stichprobenvarianz

Die korrigierte Stichprobenvarianz ist dann nötig, wenn auf Basis der Varianz in einer Stichprobe auf die Varianz in der Grundgesamtheit geschlossen werden soll.

Es kann gezeigt werden, dass die Mittelwerte von Varianzen aus unendlich vielen Stichproben des Umfangs n der Grundgesamtheit die Varianz in der Grundgesamtheit um den Faktor (n-1)/n unterschätzt.

(-> die Stichprobenvarianz ist kein „erwartungstreuer Schätzer" des Populationsparameters).

Wenn somit auf Basis der Stichprobenvarianz auf die Varianz in der Grundgesamtheit geschlossen werden soll, dann ist die Varianz nicht mit n sondern mit (n-1) im Nenner zu berechnen.

V18 Unterstützungspersonen

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1,00	1	,4	,4	,4
	2,00	4	1,6	1,6	2,0
	3,00	12	4,9	4,9	6,9
	4,00	24	9,8	9,8	16,7
	5,00	16	6,5	6,5	23,3
	6,00	18	7,3	7,3	30,6
	7,00	15	6,1	6,1	36,7
	8,00	16	6,5	6,5	43,3
	9,00	6	2,4	2,4	45,7
	10,00	44	17,9	18,0	63,7
	11,00	9	3,7	3,7	67,3
	12,00	20	8,1	8,2	75,5
	13,00	1	,4	,4	75,9
	14,00	8	3,3	3,3	79,2
	15,00	17	6,9	6,9	86,1
	16,00	2	,8	,8	86,9
	17,00	2	,8	,8	87,8
	18,00	3	1,2	1,2	89,0
	19,00	2	,8	,8	89,8
	20,00	13	5,3	5,3	95,1
	21,00	1	,4	,4	95,5
	23,00	1	,4	,4	95,9
	25,00	2	,8	,8	96,7
	26,00	1	,4	,4	97,1
	30,00	3	1,2	1,2	98,4
	35,00	2	,8	,8	99,2
	42,00	1	,4	,4	99,6
	50,00	1	,4	,4	100,0
	Gesamt	245	99,6	100,0	
Fehlend	System	1	,4		
Gesamt		246	100,0		

Statistiken

V18 Unterstützungspersonen

N	Gültig	245
	Fehlend	1
Mittelwert		10,4694
Standardabweichung		6,76691
Varianz		45,791

Die Befragten nannten durchschnittlich 10,5 Unterstützungspersonen.

Die mittlere Abweichung von diesem Durchschnitt beträgt 6,8 Unterstützungspersonen

Ergänzende Literaturempfehlung:

Benninghaus, Hans (2001). Einführung in die Sozialwissenschaftliche Datenanalyse. München, Oldenbourg, 6.Auflage oder höher.