LÖSUNG 2C

a.

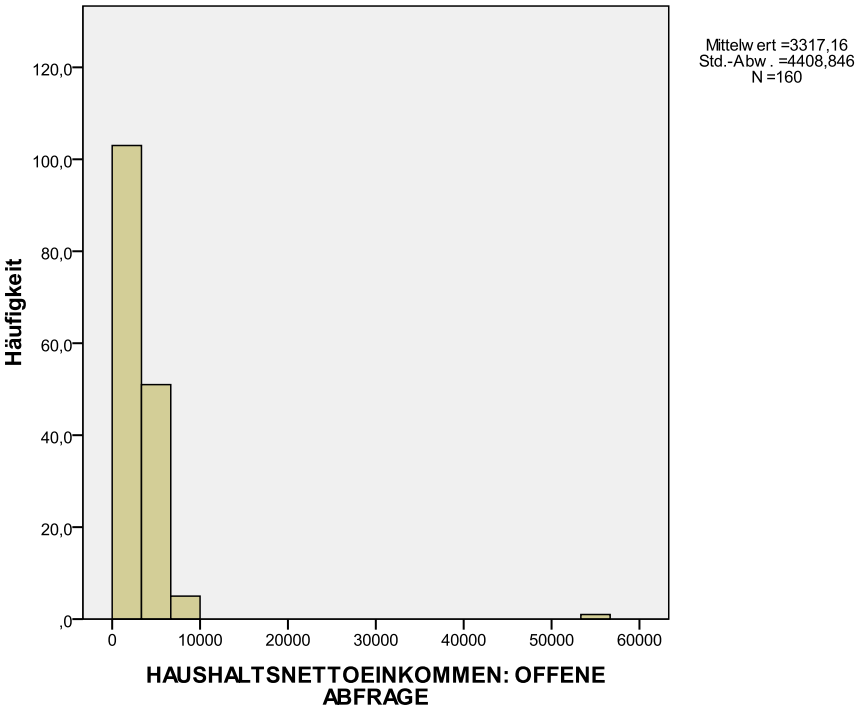
- Bei HHEINK handelt es sich um eine metrische Variable.

- Bei den Analysen sollen Extremwerte ausgeschlossen werden. Man sollte sich darüber im Klaren sein, dass Extremwerte statistische Maßzahlen zur Beschreibung einer Verteilung stark beeinflussen und verzerren können. Auch Untersuchungen zu Zusammenhängen zwischen Variablen können durch Extremwerte "gestört" werden: Möglich wäre, dass ein Zusammenhang gar nicht sichtbar wird oder auch, dass ein Zusammenhang gemessen wird, obwohl er gar nicht besteht.

Extremwerte feststellen:

- Es gibt verschiedene Möglichkeiten, Extremwerte zu finden und deren Fälle zu identifizieren.

- Ein Histogramm zeigt, dass mindestens ein Wert, der zwischen 50000 und 60000 DM liegt, aus der Verteilung der Einkommen heraus fällt („Grafik“, „Diagrammerstellung“, „Histogramm“ wählen, durch Doppelklicken auf das Symbol für ein „Einfaches Histogramm“ dieses in die Diagrammvorschau übertragen, HHEINK auf „X-Achse?“ ziehen).



- Eine bessere Möglichkeit bietet sich im Menü „Analysieren“, "Deskriptive Statistiken", "Explorative Datenanalyse". Hier kann man anfordern, dass 5 größte und 5 kleinste Werte mit der Fallnummer und/oder einer Identifikationsvariable ausgegeben werden.

  Dazu geht man wie folgt vor: In der Dialogbox "Explorative Datenanalyse" übertragen Sie die Variable HHEINK das Eingabefeld "Abhängige Variablen" und Variable NR in das Eingabefeld "Fallbeschriftung". Klicken Sie auf die Schaltfläche "Statistik" und wählen Sie in der sich öffnenden Dialogbox die Option "Ausreißer“.

  Aus der resultierenden Tabelle kann man entnehmen, dass die SPSS-Fallnummer 91 mit der ALLBUS-Identifikationsnummer 3479 einen sehr hohen Wert in Höhe von 55.000 DM hat. Möglicherweise handelt es sich um einen Datenfehler.

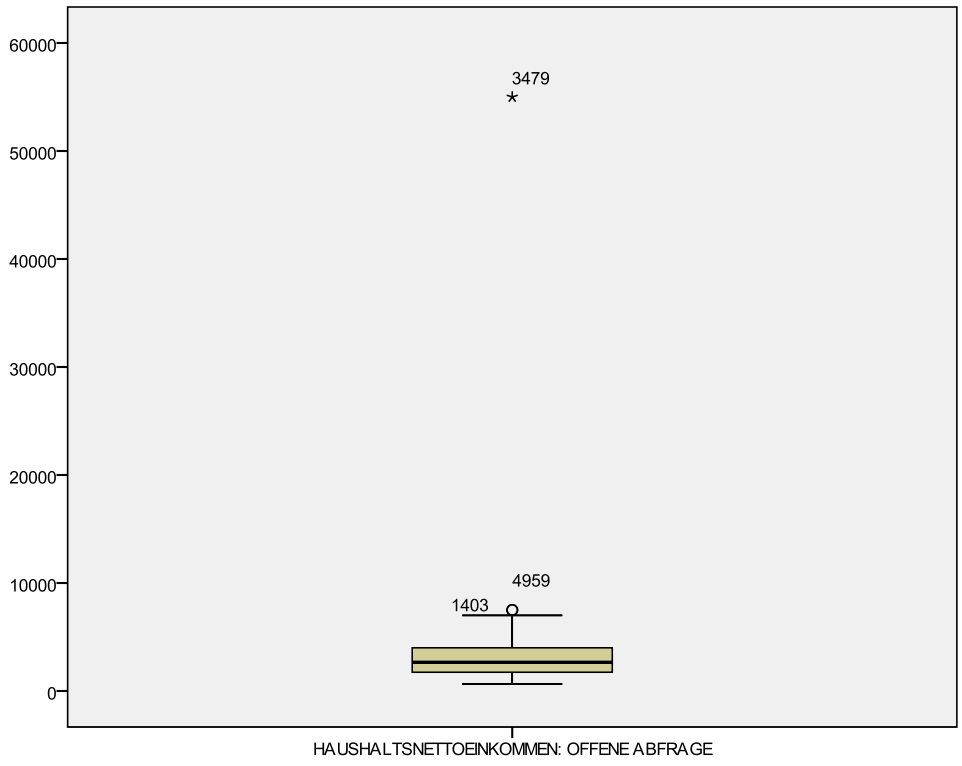
Man findet den Fall im Daten-Editor, in dem man dort in der Datenansicht die Variable HHEINK markiert, "Bearbeiten", "Suchen" klickt und den Wert 55000 eingibt. Natürlich kann man auch den Fall 91 einfach durch Scrollen im Dateneditor aufsuchen.

Extremwerte

			Fallnummer	nr	Wert
hheink	Größte Werte	1	91	3479	55000
		2	137	4959	7500
		3	177	1403	7500
		4	106	3964	7000
		5	283	4583	7000
	Kleinste Werte	1	294	4911	650
		2	1	39	680
		3	130	4704	700
		4	129	4612	800
		5	192	2006	850[a]

a. Nur eine partielle Liste von Fällen mit dem Wert 850 wird in der Tabelle der unteren Extremwerte angezeigt.

Mit "Explorative Datenanalyse" wird auch ein Boxplot-Diagramm ausgegeben. Auch in diesem Diagramm kann man den Fall 3479 als einen Extremwert erkennen. (Näheres zum Boxplot-Diagramm s. Lösung zu Aufgabe 2d und Kap. 32.13).



Einkommensklassen bilden:

- Aus obiger Tabelle mit den Extremwerten ist ersichtlich, dass (bei Ausschluss des Extremwertes in Höhe von 55.000) 7.500 der höchste Wert für HHEINK ist. Eine Klassenbildung bis unter 8.000 erfasst alle Daten bis auf den Ausreißer.

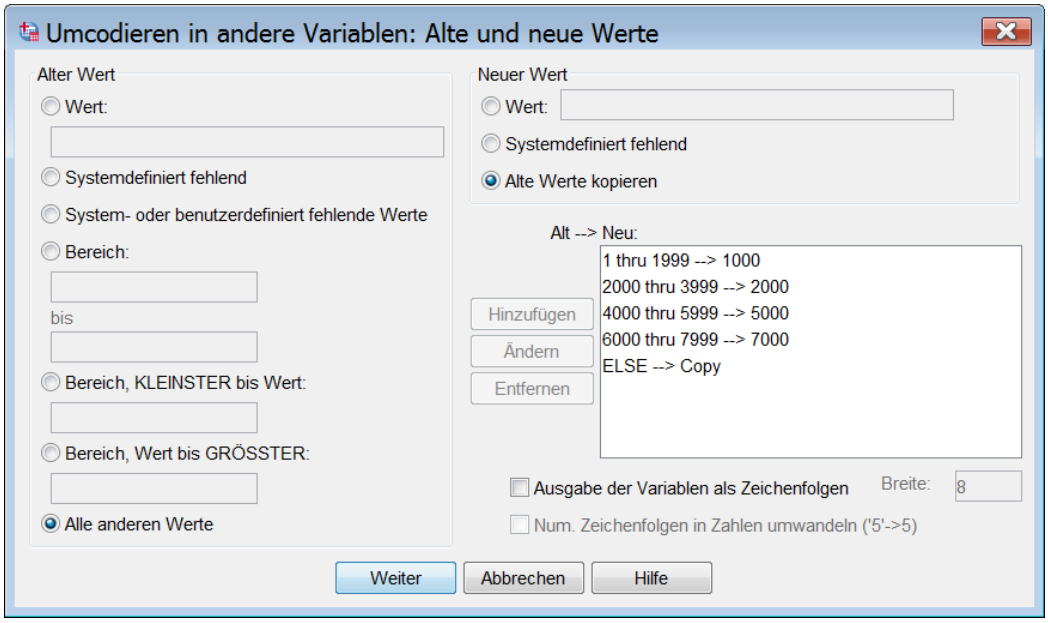
- Die Klassenbildung für HHEINK wird im Menü "Transformieren", "Umkodieren in andere Variablen…" vorgenommen. Man sollte nicht in die gleiche Variable umkodieren, da dann die Ursprungswerte von HHEINK verloren gehen. Als neue Variable kann man z.B. den Variablennamen HHEINK3 nehmen, eine neue Beschriftung für den Variablennamen (z.B. Haushalteinkommen, klassifiziert) sowie Beschriftungen für die Variablenwerte vergeben.

  Man sollte als kodierte Werte für HHEINK3 die Klassenmitte der Wertebereiche von HHEINK nehmen (also: Wert 1.000 für HHEINK3 für den Wertebereich 1 bis 1.999 von HHEINK, 3000 für den Wertebereich 2000 bis 3999 usw.). Dieses ist unbedingt erforderlich, wenn man auf der Basis der klassifizierten Daten statistische Maßzahlen berechnen möchte.

  Zu überlegen ist, wie man bei der Rekodierung mit den als nutzerdefinierten fehlenden Werten sowie dem Extremwert in Höhe von 55.000 DM umgehen möchte. HHEINK hat 0 und 97000 bis 99000 als fehlende Werte.

  Lösung 1: Nachdem man die Rekodierung für die letzte Klasse (6.000 bis 7.999) übertragen hat, werden "alle anderen Werte" in "alte Werte kopieren" übertragen. Für die Werte 0, 55000, 99997, 99998 und 99999 werden in der Variablenansicht des Daten-Editors die entsprechenden Label vergeben. Anschließend definiert man 55.000 bis 99.999 sowie 0 als fehlende Werte. Damit ist in der neuen Variable HHEINK3 sichtbar, dass man auch den Extremwert als fehlend definiert hat. Auch die Gründe für fehlende Werte bleiben sichtbar (Transparenz).

  *Lösung 2: W*ie bei Lösung 1, nur mit dem Unterschied, dass nun "alle anderen Werte" in "Systemdefiniert fehlend" übertragen werden. Man kann jetzt den Grund für fehlende Werte nicht mehr erkennen. Daher sollte man wohl die erste Lösung bevorzugen.



(Eine alternative Möglichkeit der Klassifizierung metrischer Variablen bietet das Menü „Transformieren“. „Visuelles Klassieren“. Kapitel 5.4)

Häufigkeitsauszählung:

- Auszählen mit „Analysieren“ „Deskriptive Statistiken“, „Häufigkeiten…“.

hheink3 Haushaltseinkommen (klassifiziert)

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	1000 (- u. 2000)	47	15,6	29,6	29,6
	3000 (2000 - u. 4000)	68	22,6	42,8	72,3
	5000 (4000 - u. 6000)	32	10,6	20,1	92,5
	7000 (6000 - u. 8000)	12	4,0	7,5	100,0
	Gesamt	159	52,8	100,0	
Fehlend	55000 Extremwert	1	,3		
	99997 verweigert	126	41,9		
	99998 weiss nicht	1	,3		
	99999 keine Angabe	14	4,7		
	Gesamt	142	47,2		
Gesamt		301	100,0		

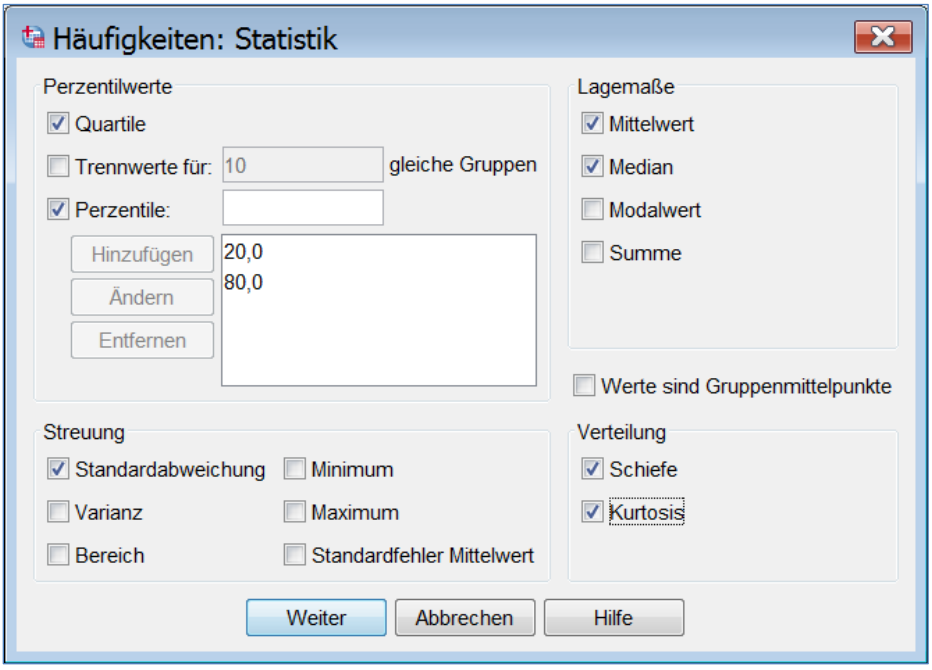
Statistische Maßzahlen:

- *Für unklassifizierte Daten.*

  Zur Berechnung statistischer Maßzahlen sollte man immer wenn es möglich ist von den Ursprungswerten, also von HHEINKOM, ausgehen. Geht man von den klassifizierten Daten aus, so kommt es zu Ungenauigkeiten.

  Um den Extremwert 55.000 auszuschließen, wählt man "Daten", "Fälle auswählen", "Falls Bedingung zutrifft" und kann als Bedingung HHEINK ~= 55.000 eingeben (alternativ: HHEINK < 55.000), ("~" ist das Zeichen für das logische "nicht"). (Man kann aber auch 55000 als fehlenden Wert deklarieren.)

  Das Berechnen geschieht im Menü "Deskriptive Statistiken", "Häufigkeiten": Variable HHEINK übertragen, Schaltfläche "Statistik" klicken und die Optionsschalter für die gewünschten Maßzahlen wählen, Perzentile eingeben und mit "Hinzufügen" übertragen.



Statistiken

hheink HAUSHALTSNETTOEINKOMMEN: OFFENE ABFRAGE

N	Gültig	159
	Fehlend	0
Mittelwert		2992,11
Median		2600,00
Standardabweichung		1596,503
Schiefe		,802
Standardfehler der Schiefe		,192
Kurtosis		,033
Standardfehler der Kurtosis		,383
Perzentile	20	1600,00
	25	1725,00
	50	2600,00
	75	4000,00
	80	4400,00

Es handelt sich um eine linkssteile Verteilung (Median < Mittelwert). Auch das Schiefemaß weist mit einem positiven Wert auf das Vorliegen einer linkssteilen Verteilung hin. Die Wölbung der Verteilung entspricht einer Normalverteilung (Kurtosis = 0,033).

Wegen der Linkssteilheit der Verteilung ist der Mittelwert (das arithmetische Mittel) kein guter Repräsentant für das mittlere Haushaltseinkommen. Der Median eignet sich dafür besser.

Die Streuung ist, gemessen an der Standardabweichung, im Vergleich zum mittleren Wert sehr hoch. Es gibt also große Unterschiede in den Haushaltseinkommen.

Die mittleren 50 % der Haushaltseinkommen liegen zwischen 1725 und 4000 DM, die mittleren 60 % zwischen 1600 und 4400 DM.

- *Für klassifizierte Daten.*

  Bei der Berechnung der Maßzahlen auf der Basis der klassifizierten Daten (HHEINK3) muss man in der Dialogbox "Häufigkeiten: Statistik" "Werte sind Gruppenmittelpunkte" anklicken.

  Man sieht bei einem Vergleich der Ergebnisse, dass diese z.T. sehr stark voneinander abweichen. Das liegt z.T. daran, dass wir nur wenige (vier) Einkommensklassen gebildet haben. Man sollte also besser mehr als vier Klassen bilden.

  Generell sollte man statistische Maßzahlen einer Verteilung nach Möglichkeit immer aus den Ursprungsdaten berechnen.

Statistiken

hheink3

N	Gültig	159
	Fehlend	142
Mittelwert		3113,21
Median		2947,83[a]
Standardabweichung		1789,495
Schiefe		,532
Standardfehler der Schiefe		,192
Kurtosis		-,441
Standardfehler der Kurtosis		,383
Perzentile	20	1288,70[b]
	25	1565,22
	50	2947,83
	75	4530,00
	80	4848,00

a. Aus gruppierten Daten berechnet

b. Perzentile werden aus gruppierten Daten berechnet.

b.

- Das arithmetische Mittel für die Ursprungswerte und für die klassifizierten Daten unterscheidet sich nicht, wenn das arithmetische Mittel der Werte innerhalb einer jeden Klasse gleich der jeweiligen Klassenmitte ist. Diese Bedingung ist praktisch nie erfüllt.

  Um die Bedingung für die zweite Einkommensklasse zu prüfen, berechnen wir für diese Klasse das arithmetische Mittel.

  Zunächst wird mit „Daten“, „Fälle auswählen“, „Falls Bedingung zutrifft“ und der Bedingung hheink >= 2000 & hheink < 4000 die Auswahl der Fälle der Einkommensklasse durchgeführt (s.o.).

  Anschließend wird mit „Analysieren“, „Deskriptive Statistiken, „Deskriptive Statistik…“ die Dialogbox „Deskriptive Statistik“ geöffnet. Dort übertragen sie Auswertungsvariable HHEINK in das Feld „Variable(n)“ und klicken auf „Optionen…“. In der sich öffnenden Dialogbox wählen Sie „Mittelwert“. Mit „Weiter“ und „OK“ wird dieser für die zweite Einkommensklasse berechnet. Der Mittelwert ist mit 2747,62 viel kleiner als die Klassenmitte 3000.

Deskriptive Statistik

	N	Mittelwert
hheink	68	2747,62
Gültige Werte (Listenweise)	68	

c.

Die meisten wählbaren Maßzahlen in den Menüs stimmen überein: Mittelwert, Standardabweichung etc. Aber im Menü Häufigkeiten lassen sich auch Median, Quartile, Perzentile, der Modal-

wert und Trennwerte für gleiche Gruppen berechnen sowie die Option „Werte sind Gruppenmittelpunkte“ auswählen.

d.

- Robuste Lageparameter fordert man im Menü „Analysieren“, „Deskriptive Statistiken“, "Explorative Datenanalyse", Untermenü "Statistiken" durch Wählen der Option "M-Schätzer" an. Bevor diese für die Variable HHEINK berechnet werden, muss mit „Daten“, „Fälle auswählen“ der Extremwert 55000 ausgeschlossen werden (s. o.)[1].

  Die robusten Lageparameter sind als mittlere Werte erwartungsgemäß kleiner als das arithmetische Mittel und liegen somit näher am Median. Das liegt daran, dass vom mittleren Wert abweichende Werte der Verteilung mit je kleinerem Gewicht in die Berechnung einfließen, desto weiter der Wert vom mittleren Wert abweicht.

  Die verschiedenen M-Schätzer unterscheiden sich durch die unterschiedliche Gewichtung von hohen und kleinen Werten.

M-Schätzer

	M-Schätzer nach Huber[a]	Tukey-Biweight[b]	M-Schätzer nach Hampel[c]	Andrews-Welle[d]
hheink	2752,15	2645,40	2784,37	2642,63

a. Die Gewichtungskonstante ist 1,339.

b. Die Gewichtungskonstante ist 4,685.

c. Die Gewichtungskonstanten sind 1,700, 3,400 und 8,500

d. Die Gewichtungskonstante ist 1,340*pi.

[1] Bei den robusten Lageparametern extreme Werte schon weniger berücksichtigt, so dass Sie zu keiner so großen Verzerrung führen, darin liegt gerade ihr Sinn. Man muss also Ausreißer nicht unbedingt entfernen. Wenn allerdings solch extreme Abweichungen vorliegen, wie bei unserem Fall mit einem Einkommen von 55000, sollte man diesen doch auch hier schon vorher bereinigen.