Vorbemerkung zu den SPSS-Aufgaben: Die Daten können mit folgendem Befehl geöffnet werden

```
get file="arbeit98s.sav".
```

Fehlende Werte (Missing Values, z.B. „keine Antwort“, „weiss nicht“) sind in den Daten mit negativen Zahlen codiert und können mit dem folgenden Befehl von den Analysen ausgeschlossen werden:

```
missing values all (-9 thru -1).
```

Lösungen 1: Variablen und Skalenniveaus

1-1. a) diskret: endliche bzw. abzählbar unendliche Anzahl Ausprägungen, z.B. Geschlecht, Zähldaten; stetig (kontinuierlich): unendlich viele Ausprägungen innerhalb eines Intervalles möglich, z.B. Körpergrösse

b) qualitativ (kategorial): Ausprägungen sind Qualitäten/Kategorien, z.B. Berufe; quantitativ: Ausprägungen entsprechen Intensität/Ausmass, z.B. Einkommen

1-2. a) nominal, diskret; b) ordinal, diskret; c) absolut, diskret; d) ordinal, diskret (normalerweise); e) intervall, diskret (Kalenderzeit ist aber an sich stetig); f) ratio, stetig; g) nominal, diskret; h) nominal, diskret

1-3. a) intervall und tiefer; b) ordinal; c) ordinal und höher; d) ratio und höher; e) intervall und tiefer; f) ratio und tiefer (Konstante muss grösser als 0 sein)

Lösungen 2: Summenzeichen

2-1. a) $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$

b) $\sum_{i=1}^{3} x_i^2 y_i = x_1^2 y_1 + x_2^2 y_2 + x_3^2 y_3$

c) $\sum_{i=1}^{2} (x_i + y_i)^2 = (x_1 + y_1)^2 + (x_2 + y_2)^2 = (x_1^2 + 2x_1 y_1 + y_1^2) + (x_2^2 + 2x_2 y_2 + y_2^2) = x_1^2 + x_2^2 + 2x_1 y_1 + 2x_2 y_2 + y_1^2 + y_2^2$

d) $\left[\sum_{i=4}^{5} y_i^2\right]^{1/2} = [y_4^2 + y_5^2]^{1/2}$

e) $\sum_{i=4}^{6} \sum_{j=1}^{4} x_i y_j = (x_4 + x_5 + x_6)(y_1 + y_2 + y_3 + y_4) = x_4 y_1 + x_4 y_2 + x_4 y_3 + x_4 y_4 + x_5 y_1 + x_5 y_2 + x_5 y_3 + x_5 y_4 + x_6 y_1 + x_6 y_2 + x_6 y_3 + x_6 y_4$

f) $\sum_{i=1}^{2} (x_i^2 + y_i^2) = x_1^2 + y_1^2 + x_2^2 + y_2^2$

g) $\sum_{i=1}^{4} (x_i - 5) = x_1 + x_2 + x_3 + x_4 - 4 \cdot 5$

2-2. a) $= 3 + 4 + 4 + 0 + 6 = 17$

b) $= 3^2 \cdot 2 + 4^2 \cdot 5 + 4^2 \cdot 1 = 114$

c) $= 3^2 + 4^2 + 2 \cdot 3 \cdot 2 + 2 \cdot 4 \cdot 5 + 2^2 + 5^2 = 106$

d) $= [3^2 + 4^2]^{1/2} = 5$

e) $= (0 + 6 + 2)(2 + 5 + 1 + 3) = 88$

f) $= 3^2 + 2^2 + 4^2 + 5^2 = 54$

g) $= 3 + 4 + 4 + 0 - 4 \cdot 5 = -9$

2-3. a) $\sum_{i=1}^{4} x_{i1} = 2 + 3 + 1 + 2 = 8$

b) $\sum_{j=1}^{3} x_{3j} = 1 + 8 + 5 = 14$

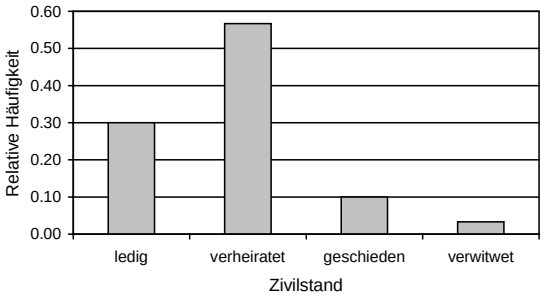
c) $\sum_{i=3}^{4} \sum_{j=1}^{3} x_{ij} = 1 + 8 + 5 + 2 + 7 + 7 = 30$

d) $\sum_{i=1}^{4} \sum_{j=1}^{3} x_{i2}x_{3j} = (9 + 5 + 8 + 7)(1 + 8 + 5) = 406$

Lösungen 3: Häufigkeitsverteilungen und grafische Darstellung

3-1.

Zivilstand	h_j	f_j	Winkel
ledig	900	0.300	108°
verheiratet	1700	0.567	204°
geschieden	300	0.100	36°
verwitwet	100	0.033	12°
	3000	1.000	360°



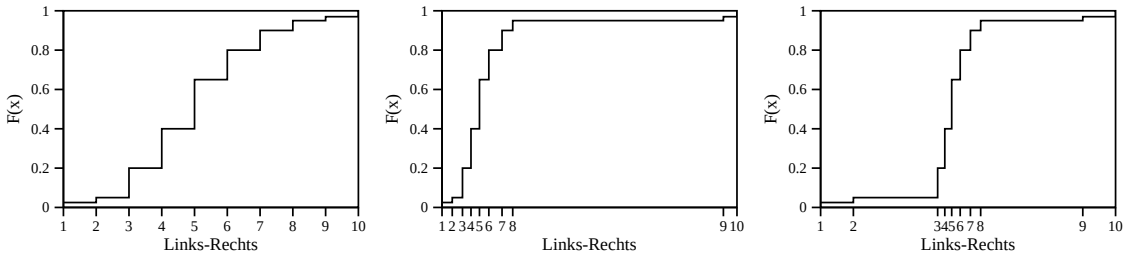
3-2. 810 Fälle entsprechen einem relativen Anteil von $1 - 0.45 - 0.28 = 0.27$. Die Anzahl Fälle im Total ergibt sich somit als $810/0.27 = 3000$.

Politikinteresse	h_j	f_j
1. viel	810	0.27
2. mittel	1350	0.45
3. wenig	840	0.28
Total	3000	1.00

3-3. Sinnvoll bei (quasi-)stetigen Merkmalen; die Fläche einer Säule entspricht der relativen Häufigkeit der Klasse.

3-4. Vgl. Jann (2002: S. 28).

3-5. Das Verfahren ist problematisch, weil es sich um eine Ordinalskala handelt. Die Kurve kann durch die Wahl unterschiedlicher Abstände zwischen den Kategorien manipuliert werden. Alle nachfolgenden Darstellungen sind im Prinzip zulässig (wobei man normalerweise wohl die erste Darstellung verwenden würde).



*3-6. Frauenanteil: 49.7%. Syntax:

```
frequencies sex.
```

*3-7. Anteil älter 30: 80.6%. Syntax:

```
compute alter=98-a5j.
variable label alter "Alter in Jahren".
frequencies alter.
```

*3-8. Anteil Frauen mit Tertiärausbildung: 7.0% + 6.4% = 13.4%; Anteil Männer mit Tertiärausbildung: 21.5% + 9.8% = 31.3% ⇒ Das Bildungsniveau der Männer in der Stichprobe ist deutlich höher als das der Frauen. Syntax:

```
sort cases by sex.
split file by sex.
frequencies g1.
split file off.
```

*3-9. Die Partizipationsquote war 61.8%. Zu beachten ist aber, dass sich viele Befragte nicht erinnern konnten, ob sie die an den Abstimmungen teilgenommen haben. Die Annahme liegt nahe, dass diese Personen eher *nicht* teilgenommen haben (sonst würden sie sich wohl erinnern). Syntax:

```
frequencies i17.
```

*3-10. Partizipation der 18–30-Jährigen: 45.8%; Partizipation der über 30-Jährigen: 65.7% ⇒ Die Teilnahme an den Abstimmungen hängt deutlich mit dem Alter zusammen. Syntax:

```
temporary.
select if alter<=30.
frequencies i17.

temporary.
select if alter>30.
frequencies i17.
```

Oder:

```
compute alt01=alter>30.
value label alt01 0"30 oder jünger" 1"über 30".
sort cases by alt01.
split file by alt01.
frequencies i17.
split file off.
```

*3-11. Symmetrisch-unimodale Verteilung. Syntax:

```
graph /bar=b29.
```

*3-12. Asymmetrisch-multimodale, leicht rechtsschiefe Verteilung. Syntax:

```
graph /histogram=alter.
```

Oder:

```
frequencies alter /format=notable /histogram.
```

Lösungen 4: Univariate Kennzahlen

4-1. Modus: $M = 35$

Median: $\tilde{x} = (x_{(16/2)} + x_{(16/2+1)})/2 = (x_{(8)} + x_{(9)})/2 = (37 + 39)/2 = 38$

arith. Mittel: $\bar{x} = 665/16 = 41.56$

Quartile (np ist in beiden Fällen ganzzahlig):
$Q_1 = x_{0.25} \in [x_{(16\cdot 0.25)} = x_{(4)} = 35, x_{(16\cdot 0.25+1)} = x_{(5)} = 35] \Rightarrow Q_1 = 35$
$Q_3 = x_{0.75} \in [x_{(16\cdot 0.75)} = x_{(12)} = 45, x_{(16\cdot 0.75+1)} = x_{(13)} = 48] \Rightarrow Q_1 \in [45, 48]$

Lineare Interpolation für Q_3:
$Q_3 = x_{([0.75(16+1)]_G)} + (0.75(16+1) - [0.75(16+1)]_G)(x_{([0.75(16+1)]_G+1)} - x_{([0.75(16+1)]_G)})$
$= x_{(12)} + (12.75 - 12)(x_{(13)} - x_{(12)}) = 45 + 0.75(48 - 45) = 47.25$

Varianz: $s^2 = 28977/16 - \bar{x}^2 = 83.62$

Standardabweichung: $s = \sqrt{83.62} = 9.14$

Stichprobenvarianz/-standardabweichung (Nenner: $n - 1$): $s^2 = 89.20, \quad s = 9.44$

4-2. a) $M = 300$, $\tilde{x} = 300$, $\bar{x} = 7660/20 = 383$, $s^2 = 4667400/20 - 383^2 = 86681$, $s = 294.4$

b) Klassenmitten (auch Alternativen denkbar): 100, 300, 500, 700, 1100.5
M = Klasse 200-399
$\tilde{x} = 200 + 200(0.5 - 0.15)/0.55 = 327.27$
$\bar{x} = 8000.5/20 = 400.025$
$s^2 = 4441100.3/20 - 400.025^2 = 62035.014$, $s = 249.07$

Die Lagemasse werden mit den klassierten Daten relativ gut wiedergegeben (nur geringe Abweichungen zu den mit den Originaldaten berechneten Lagemassen). Allerdings ist die Varianz der klassierten Daten deutlich kleiner (da die Varianz innerhalb der Klassen eliminiert wird).

4-3. Median

4-4. $\bar{y} = 1.3 - 0.7\bar{x} = 0.11$

$s_Y^2 = (-0.7)^2 s^2 = 0.1225$

$s_Y = |-0.7|s = 0.35$

4-5. $\bar{z} = 0$, $s_Z = 1$

4-6. a) $\tilde{x} = 100$, $\bar{x} = 101.3$

b) Der Median verändert sich nicht, da die unmittelbar um den Median liegenden Werte die gleichen bleiben (98 und 100). Wird ein Messwert um einen Betrag d verändert, so wirkt sich dies auf das aritmetische Mittel mit d/n aus, wobei n die Fallzahl sei. Da 3 Werte um je 3 Einheiten verringert wurden (Fallzahl: 10), ist das korrigierte arithmertische Mittel gegeben als: $\bar{x}^{\text{korr}} = \bar{x} - 3 \cdot 3/10 = \bar{x} - 0.9 = 100.4$

4-7. $\bar{x} = 9$

$R = 12 - 7 = 5$

$IQR = 10 - 8 = 2$

$AD = (7|7 - 9| + 14|8 - 9| + \ldots)/60 = 0.93$

$s^2 = (7 \cdot 7^2 + 14 \cdot 8^2 + \ldots)/60 - 9^2 = 1.53$

$v = \sqrt{1.53}/9 = 0.137$

4-8. $\bar{x}_B = 461.2$, $s_B = 75.05$, $\bar{x}_W = 3647$, $s_W = 453.9$

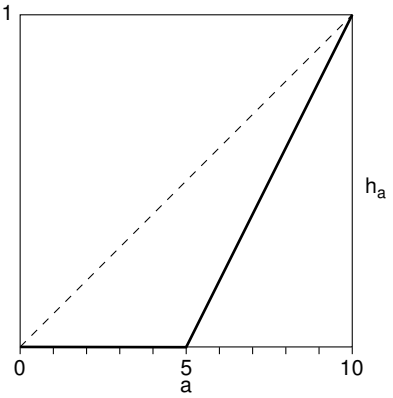
$v_B = 75.05/461.2 = 0.163$

$v_W = 453.9/3647 = 0.124$

Zwar weist die Datenreihe für Wien eine höhere Varianz auf, bei Betrachtung der Variationskoeffizienten wird aber ersichtlich, dass die Mietpreise in Bern stärker variieren (d.h. die relativen Unterschiede sind in Bern grösser).

4-9. a) $G = 0.5$

Erklärung (Geometrie): Die Gesamtfläche des Dreiecks unterhalb der Diagonale ist gleich $\frac{1}{2}ah_a$. Lorenzkurve halbiert a, die Fläche des Dreiecks unterhalb der Lorenzkurve beträgt also $\frac{1}{2}(\frac{1}{2}ah_a)$. Der Gini-Koeffizient setzt die beiden Flächen ins Verhältnis, also $[\frac{1}{2}(\frac{1}{2}ah_a)]/(\frac{1}{2}ah_a) = \frac{1}{2}$.



b) $G = \frac{2(1\cdot0+2\cdot0+...+6\cdot1+...+10\cdot1)}{10\cdot5} - \frac{10+1}{10} = \frac{2\cdot40}{50} - \frac{11}{10} = 1.6 - 1.1 = 0.5$

*4-10. Arith. Mittel: 5351.64; Median: 5000.00; Std. Abw.: 2754.88; Rechtsschiefe Verteilung da arith. Mittel > Median. Syntax:

```
missing value netinco (lo thru 0) /*Ausschluss von Pers. ohne Einkommen*/.

temporary.
select if c1=1.
descriptives netinco.
```

Besser, da auch Median:

```
temporary.
select if c1=1.
frequencies netinco /format=notable /statistics=all.
```

*4-11. Die Einkommen der Männer variieren stärker: $v_F = 1665.343/4168.25 = 0.400$, $v_M = 2959.543/5833.52 = 0.507$. Syntax:

```
temporary.
select if c1=1.
means netinco by sex.
```

Box-Plot:

```
temporary.
select if c1=1.
examine variables=netinco by sex /plot=boxplot
 /statistics=none /nototal.
```

Lösungen 5: Bivariate Kennzahlen I

5-1. $h_{11} = 1395.2$, $h_{12} = 214.8$, $h_{21} = 1228.8$, $h_{22} = 189.2$

5-2. $\chi^2 = \frac{100(0\cdot0-40\cdot60)^2}{60\cdot40\cdot60\cdot40} = 100, \quad \phi = \frac{0\cdot0-40\cdot60}{\sqrt{60\cdot40\cdot60\cdot40}} = -1$

oder:
$\chi^2 = \frac{(0-\frac{40\cdot60}{100})^2}{\frac{40\cdot60}{100}} + \frac{(40-\frac{40\cdot40}{100})^2}{\frac{40\cdot40}{100}} + \frac{(60-\frac{60\cdot60}{100})^2}{\frac{60\cdot60}{100}} + \frac{(0-\frac{60\cdot40}{100})^2}{\frac{60\cdot40}{100}} = 100$

$\phi = \sqrt{\frac{100}{100}} = 1 \quad \Rightarrow$ perfekter Zusammenhang

	0	1	
0	0	40	40
1	60	0	60
	60	40	100

5-3. $d\% = \left(\frac{60}{120} - \frac{32}{80}\right) \cdot 100 = (0.5 - 0.4) \cdot 100 = 10$

$OR = \frac{60\cdot48}{32\cdot60} = 1.5$

$\phi = \frac{60\cdot48-32\cdot60}{\sqrt{92\cdot108\cdot120\cdot80}} = 0.098$

	m	*w*	
R	60	32	92
NR	60	48	108
	120	80	200

5-4. a)

		X		
		Land	Kleinst.	Grossst.	Total
	gesch.	0.091	0.143	0.429	0.283
Y	n.gesch.	0.909	0.857	0.571	0.717
	Total	1.000	1.000	1.000	1.000

b) $\chi^2 = \frac{(5-\frac{75\cdot55}{265})^2}{\frac{75\cdot55}{265}} + \frac{(10-\frac{75\cdot70}{265})^2}{\frac{75\cdot70}{265}} + \frac{(60-\frac{75\cdot140}{265})^2}{\frac{75\cdot140}{265}} + \frac{(50-\frac{190\cdot55}{265})^2}{\frac{190\cdot55}{265}} + \frac{(60-\frac{190\cdot70}{265})^2}{\frac{190\cdot70}{265}} + \frac{(80-\frac{190\cdot140}{265})^2}{\frac{190\cdot140}{265}} = 31.4$

$V = \sqrt{\frac{31.4}{265(2-1)}} = 0.344, \quad K = \sqrt{\frac{31.4}{265+31.4}} = 0.325, \quad K^* = \frac{0.325}{\sqrt{(2-1)/2}} = 0.460$

$\lambda_X = \frac{(50+60+80)-190}{265-190} = 0, \quad \lambda_Y = \frac{(60+80)-140}{265-140} = 0$

$\tau_X = \frac{\frac{1}{265}\left(\frac{5^2+50^2}{55} + \frac{10^2+60^2}{70} + \frac{60^2+80^2}{140}\right) - \frac{75^2+190^2}{265^2}}{1-\frac{75^2+190^2}{265^2}} = 0.118$

$\tau_Y = \frac{\frac{1}{265}\left(\frac{5^2+10^2+60^2}{75} + \frac{50^2+60^2+80^2}{190}\right) - \frac{55^2+70^2+140^2}{265^2}}{1-\frac{55^2+70^2+140^2}{265^2}} = 0.072$

c)

		X		
		Land	klein/gross	Total
	gesch.	5	70	75
Y	n.gesch.	50	140	190
	Total	55	210	265

$\chi^2 = \frac{265(5\cdot140-70\cdot50)^2}{75\cdot190\cdot55\cdot210} = 12.6$

$V = \sqrt{\frac{12.6}{265(2-1)}} = 0.218, \quad K = \sqrt{\frac{12.6}{265+12.6}} = 0.213, \quad K^* = \frac{0.213}{\sqrt{(2-1)/2}} = 0.301$

$\tau_X = \frac{\frac{1}{265}\left(\frac{5^2+50^2}{55} + \frac{70^2+140^2}{210}\right) - \frac{75^2+190^2}{265^2}}{1-\frac{75^2+190^2}{265^2}} = 0.048, \quad \tau_Y = \frac{\frac{1}{265}\left(\frac{5^2+70^2}{75} + \frac{50^2+140^2}{190}\right) - \frac{55^2+210^2}{265^2}}{1-\frac{55^2+210^2}{265^2}} = 0.048$

Vergleich: Die Koeffizienten nehmen ab, weil sich das Scheidungsrisiko in Klein- und Grossstadt unterscheidet. Durch die Zusammenfassung fallen diese Unterschiede „unter den Tisch“.

5-5. $C = 5(6+7+3+8) + 2(7+8) + 2(3+8) + 6 \cdot 8 = 220$

$D = 2(2+1) + 7(3+1) + 6 \cdot 1 = 40$

$T_X = 5 \cdot 2 + 2(6+7) + 1 \cdot (3+8) + 6 \cdot 7 + 3 \cdot 8 = 113$

$T_Y = 5(2+1) + 2(6+3) + 2 \cdot 1 + 6 \cdot 3 + 7 \cdot 8 = 109$

$\tau_b = \frac{220-40}{\sqrt{(220+40+113)(220+40+109)}} = 0.485, \quad \gamma = \frac{220-40}{220+40} = 0.692$

$\tau_b = \gamma$, wenn lediglich Diagonalfelder besetzt (keine Bindungen)

5-6.

$S \backslash F$	0	1
1	2	1
2	1	2
3	0	1

$C = 2(2+1) + 1 \cdot 1 = 7, \quad D = 1 \cdot 1 = 1, \quad T_X = 2 \cdot 1 + 1 \cdot 2 = 4$

$T_Y = 2 \cdot 1 + 1(2+1) + 2 \cdot 1 = 7$

$\tau_b = \frac{7-1}{\sqrt{(7+1+4)(7+1+7)}} = 0.447, \quad \gamma = \frac{7-1}{7+1} = 0.750$

*5-7. Prozentsatzdifferenz Erwerbstätige: −11.1; Prozentsatzdifferenz Vollzeiterwerbstätige: −6.0. Erwerbstätige Männer sind häufiger Gewerkschaftsmitglied als erwerbstätige Frauen. Die Unterschiede verringern sich aber, wenn nur vollzeitwerwerbstätige Personen in Betracht gezogen werden. Syntax:

```
temporary.
select if c1<>4.
crosstabs b24 by sex /cells=count column.

temporary.
select if c1=1.
crosstabs b24 by sex /cells=count column.
```

*5-8. Syntax:

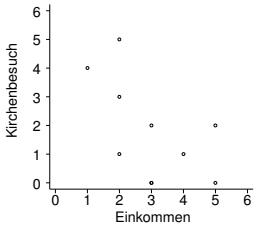
```
recode alter (lo thru 30=1) (31 thru 50=2) (51 thru hi=3) into altkat.
variable label altkat "Alterskategorien".
value label altkat 1"18-30" 2"31-50" 3"51-71".
frequencies altkat.

compute erwerb=range(c1,1,3).
value label erwerb 1"erwerbstätig" 0"nicht erwerbstätig".
frequencies erwerb.

crosstabs erwerb by altkat /cells=count column
 /statistics=chi cc phi lambda uc.
```

Lösungen 6: Bivariate Kennzahlen II

6-1. $\sum x_i = 30, \quad \sum y_i = 18, \quad \sum x_i^2 = 106$

$\sum y_i^2 = 60, \quad \sum x_i y_i = 42$

$r = \frac{10 \cdot 42 - 30 \cdot 18}{\sqrt{(10 \cdot 106 - 30^2)(10 \cdot 60 - 18^2)}} = -0.571$



6-2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
I	3	10	4	2	5	6	8	14	12	7	1	9	13	11
II	1	9	8	5	4	6	13	10	14	2	3	7	11	12

$\overline{rg} = 7.5, \quad \sum rg(x_i)^2 = \sum rg(y_i)^2 = 1015, \quad \sum rg(x_i) rg(y_i) = 958$

$r_S = \frac{958 - 14 \cdot 7.5 \cdot 7.5}{\sqrt{(1015 - 14 \cdot 7.5^2)(1015 - 14 \cdot 7.5^2)}} = 0.749$

(relativ hohe Übereinstimmung)

6-3. $\eta^2 = \frac{51(3.15-4.30)^2 + 311(4.45-4.30)^2 + 42(4.57-4.30)^2}{2476.1} = 0.031, \eta = 0.177$

3.1% der Einkommensvariation werden erklärt.

6-4. a) Punkt-biseriale Korrelation:

$\bar{y}_0 = 4, \quad \bar{y}_1 = 5.167, \quad s_Y = 2.759, \quad r_{pb} = \frac{5.167-4}{2.759} \sqrt{\frac{6 \cdot 4}{10^2}} = 0.207$

b) Rang-biseriale Korrelation:

$\overline{rg}_0 = \frac{19.5}{4} = 4.875, \quad \overline{rg}_1 = \frac{35.5}{6} = 5.917, \quad r_{rb} = \frac{2}{10}(5.917 - 4.875) = 0.208$

Positive Korrelation: Frauen sind stärker "links" eingestellt als Männer (bzw. Männer stärker "rechts" als Frauen).

*6-5. Es besteht praktisch kein monotoner Zusammenhang: $\gamma = -0.003$, $\tau_b = -0.002$, $r_S = -0.003$. Syntax:

```
crosstabs b29 by b30 /format=notables /statistics=gamma.
nonpar corr b29 b30 /print=both.
```

Es gibt aber Hinweise auf einen nicht-monotonen Zusammenhang. Wird anstatt der Links-Rechts-Skala die Abweichung von der politischen Mitte verwendet, so lauten die Masse: $\gamma = -0.080$, $\tau_b = -0.056$, $r_S = -0.064$. Das heisst, Personen, die politisch deutlich links oder rechts eingestellt sind, bekundeten ein höheres politisches Interesse als Personen in der Mitte. Syntax:

```
compute b30b=abs(5.5-b30).
fre b30b.
crosstabs b29 by b30b /format=notables /statistics=gamma.
nonpar corr b29 b30b /print=both.
```

*6-6. Zwischen Alter und Einkommen sowie Bildung und Einkommen bestehen deutliche positive Zusammenhänge. Der Zusammenhang zwischen Bildung und Alter ist hingegen nur sehr gering. Korrelationstabelle:

	Einkommen	Alter	Bildung
Einkommen	1.000	0.344	0.425
Alter	0.344	1.000	0.068
Bildung	0.425	0.068	1.000

$n = 488$

Syntax:

```
recode g1 (2=9) (3=10) (4=11) (5=12) (6=13) (7=15) (8=18) into bildung.
variable label bildung "Bildungsjahre".
frequencies bildung.

compute vz=c1=1.
filter by vz.
correlations variables=netinco alter bildung.
correlations variables=netinco alter bildung /missing=listwise.
filter off.
```

Lösungen 7: Inferenzstatistik

7-1. 1. Wahl: $(1 - P(A)) \cdot (1 - P(B)) = 0.9 \cdot 0.8 = 0.72$

2. Wahl: $P(A) \cdot (1 - P(B)) + (1 - P(A)) \cdot P(B) = 0.1 \cdot 0.8 + 0.9 \cdot 0.2 = 0.08 + 0.18 = 0.26$

Ausschuss: $P(A) \cdot P(B) = 0.1 \cdot 0.2 = 0.02$

7-2. a) Intuitiv: Bei 8 der 10 Bomben geht ein anonymer Hinweis ein (80%). Die Whs., dass eine Bombe an Bord ist UND ein Hinweis eingeht, ist also 8 zu 100 000. Dies bedeutet gleichzeitig, dass nur 8 der 400 anonymen Hinweise tatsächlich auf eine existierende Bombe verweisen. Die Whs., dass sich ein Hinweis als richtig herausstellt, beträgt somit 8 zu 400 (0.02 bzw. 2%).

b) Formal:

Ausgangslage:
$P(A) = 10/100\,000 = 0.0001$, $P(B) = 400/100\,000 = 0.004$, $P(B|A) = 0.80$, $P(A|B) = ?$

Zwischenschritt (Whs., dass A und B gleichzeitig auftreten):
$P(A \cap B) = P(B|A) \cdot P(A) = 0.8 \cdot 0.0001 = 0.00008$

Lösung (Whs., dass A eintritt, wenn B eingetreten ist):
$P(A|B) = P(A \cap B)/P(B) = 0.00008/0.004 = 0.02$

7-3.

	$P = 0.5$	$P = 0.95$
$t_P(10)$	−1.8125	1.8125
$t_P(17)$	−1.7396	1.7396
$\chi^2_P(15)$	7.2609	24.996
$\chi^2_P(28)$	16.928	41.337
$F_P(3,7)$	0.113	4.347
$F_P(7,3)$	0.230	8.887

Hinweise:
$t_P(n) = -t_{1-P}(n)$
$F_P(m,n) = \frac{1}{F_{1-P}(n,m)}$

7-4. a) $P(X \leq 180) = \Phi(\frac{x-\mu}{\sigma}) = \Phi(\frac{180-175}{10}) = \Phi(0.5) = 0.6915$

b) $P(X \geq 160) = 1 - \Phi(\frac{160-175}{10}) = 1 - \Phi(-1.5) = \Phi(1.5) = 0.9332$

c) $P(175 \geq X \geq 190) = \Phi(\frac{190-175}{10}) - \Phi(\frac{175-175}{10}) = \Phi(1.5) - \Phi(0.0) = 0.9332 - 0.5000 = 0.4332$

7-5. $p = \frac{2304}{6400} = 0.36$

approx. 95%-Konfidenzintervall:

$p \pm z_{1-\alpha/2}\hat{\sigma}_p = p \pm z_{0.975}\sqrt{\frac{p(1-p)}{n}} = 0.36 \pm 1.96\sqrt{\frac{0.36(1-0.36)}{6400}} = 0.36 \pm 0.01176$

$\Rightarrow 0.34824 \leq \pi \leq 0.37176$

7-6. Hypothesen: $H_0: \mu - \mu_0 = 0; \quad H_1: \mu - \mu_0 \neq 0$

Teststatistik: $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}, Z \sim N(0,1)$

$Z_1 = \frac{10.014-10}{0.1/\sqrt{5}} = 0.313 \quad Z_2 = \frac{10.052-10}{0.1/\sqrt{5}} = 1.163 \quad Z_3 = \frac{9.87-10}{0.1/\sqrt{5}} = -2.907$

In der Stichprobe 3 wird die Nullhypothese für $\alpha = 0.01$ abgelehnt, da $|Z_3| \geq z_{1-\alpha/2} = 2.58$. (In Stichprobe 3 kann mit mindestens 99%-iger Whs. davon ausgegangen werden, dass es sich nicht um zufällige Abweichungen handelt und die Produktionsmaschine neu eingestellt werden muss.) Stichproben 1 und 2 liegen innerhalb der zulässigen Grenzen.

7-7. $H_0: \mu_1 - \mu_2 \leq 0; \quad H_1: \mu_1 - \mu_2 > 0$

$T = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{s_1^2/n_1+s_2^2/n_2}} = \frac{13.34-12.73}{\sqrt{3.11^2/2502+3.84^2/361}} = 2.885$

(T ist für grosse Fallzahlen approx. standardnormalverteilt)

Schweizer weisen ein signifikant höheres Bildungsniveau auf, da $|T| \geq z_{1-\alpha} = 1.64$.

7-8. approx. McNemar-Test: $\chi^2 = \frac{(h_{12}-h_{21})^2}{h_{12}+h_{21}} \overset{a}{\sim} \chi^2(1)$

$\chi^2 = \frac{(6-14)^2}{6+14} = 3.2 < \chi^2_{0.95}(1) = 3.84$ (kein signifikanter Zusammenhang)

7-9. χ^2-Test: $\chi^2 = \frac{(8-15\cdot 51/124)^2}{15\cdot 51/124} + \cdots + \frac{(21-72\cdot 29/124)^2}{72\cdot 29/124} = 18.3 > \chi^2_{0.95}(4) = 9.49$

Likelihood-Ratio-Test:

$LR = 2(8\ln[8/(15\cdot 51/124)] + \cdots + 21\ln[21/(72\cdot 29/124)]) = 21.8 > \chi^2_{0.95}(4) = 9.49$

(die Merkmale sind mit grosser Wahrscheinlichkeit nicht unabhängig)

*7-10. Die Mittelwertsdifferenz von CHF 1665.27 ist hochsignifikant: $|T| = 8.02$ (ungleiche Varianzen). Syntax:

```
temporary.
select if c1=1.
t-test /groups=sex(0 1) /variables=netinco.
```

*7-11. Der Zusammenhang ist hochsignifikant: $\chi^2 = 193.06$ bei 32 Freiheitsgraden. Syntax:

```
crosstabs b29 by g1 /cells=column /statistics=chisq.
```

Wird die Kategorie „andere Ausbildung“ augeschlossen, kann die Bildungsvariable als ordinalskaliert betrachtet werden. Die ordinalen Zusammenhangsmasse τ_b und γ betragen dann -0.305 und -0.424 (ebenfalls hochsignifikant). Das heisst, höhere Ausbildung geht mit stärkerem politischen Interesse einher (man beachte, dass das politische Interesse „verkehrt“ skaliert ist: hohe Werte auf der Skala bedeuten tiefes Interesse). Syntax:

```
temporary.
sel if g1<9.
crosstabs b29 by g1 /format=notable /statistics=chisq btau gamma.
```

*7-12. Der Zusammenhang zwischen Gewerkschaftsmitgliedschaft und Geschlecht ist signifikant: $\chi^2 = 11.86$, $p = 0.001$; Fisher's Exact Test: $p = 0.001$. Syntax:

```
temporary.
select if c1<>4.
crosstabs b24 by sex /cells=count column
 /statistics=chisq.
```

Werden jedoch nur vollzeiterwerbstätige Personen berücksichtigt, sind die Unterschiede nicht mehr signifikant: $\chi^2 = 1.95$, $p = 0.163$; Fisher's Exact Test: $p = 0.174$. Das heisst, der Zusammenhang hat offenbar zumindest teilweise auch mit dem Beschäftigungsgrad zu tun. Syntax:

```
temporary.
select if c1=1.
crosstabs b24 by sex /cells=count column
 /statistics=chisq.
```

*7-13. Der Zusammenhang ist Signifikant. Median-Test: $p = 0.000$; Rangsummen-Test: $p = 0.000$. Syntax:

```
npar tests /median=b29 by sex(0 1).
npar tests /m-w=b29 by sex(0 1).
```

Werden jedoch nur Personen mit Tertiärausbildung betrachtet, ist der Zusammenhang nicht mehr signifikant. Median-Test: $p = 0.712$; Rangsummen-Test: $p = 0.058$. Dieses Resultat ist dadurch zu erklären, dass Bildung einen Einfluss auf des politische Interesse hat und Frauen im Durchschnitt ein geringeres Bildungsniveau aufweisen. Syntax:

```
temporary.
select if range(g1,7,8).
npar tests /median=b29 by sex(0 1).
temporary.
select if range(g1,7,8).
npar tests /m-w=b29 by sex(0 1).
```

Lösungen 8: Lineare Regression

8-1. a) Einkommen: $\bar{x} = \frac{1}{n}\sum_i x_i = \frac{30}{10} = 3$

$s_X^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 = \frac{106}{10} - 9 = 1.6$

Kirche: $\bar{y} = \frac{1}{n}\sum_i y_i = \frac{18}{10} = 1.8$

$s_Y^2 = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 = \frac{60}{10} - 3.24 = 2.76$

Kovarianz:
$s_{XY} = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y} = \frac{42}{10} - 5.4 = -1.2$

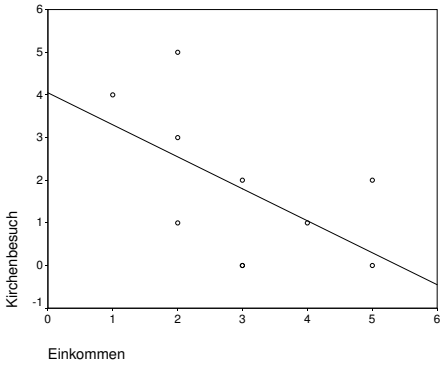
Regressionsgleichung:

$\hat{\beta} = \frac{S_{XY}}{s_X^2} = \frac{-1.2}{1.6} = -0.75$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 1.8 + 0.75 \cdot 3 = 4.05$

$\hat{Y} = \hat{\alpha} + \hat{\beta}X = 4.05 - 0.75X$

Streudiagramm



b) $\hat{Y}(X = 3) = 4.05 - 0.75 \cdot 3 = 1.8$

c) Nicht-erklärte Varianz: $\frac{SQR}{SQT} = 1 - R^2 = 1 - \frac{s_{XY}^2}{s_X^2 s_Y^2} = 1 - \frac{-1.2^2}{1.6 \cdot 2.76} = 0.674$
(67% der Varianz sind nicht erklärt bzw. 33% sind erklärt)

d) $r = \frac{s_{XY}}{s_X s_Y} = \frac{-1.2}{\sqrt{1.6}\sqrt{2.76}} = -0.57$

$T = \frac{-0.57}{\sqrt{1-(-0.57)^2}}\sqrt{10-2} = -1.96 \quad \Rightarrow \quad |T| < t_{0.975}(8) = 2.31$ (nicht signifikant)

8-2. a) $\hat{\beta} = \frac{s_{XY}}{s_X^2} = \frac{0.353}{3.449} = 0.102$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 4.169 - 0.102 \cdot 3.730 = 3.789$

Regressionsgleichung: $\hat{Y} = \hat{\alpha} + \hat{\beta}X = 3.789 + 0.102X$

b) $\hat{\beta}^z = \hat{\beta}\frac{s_X}{s_Y} = 0.102\frac{\sqrt{3.449}}{\sqrt{1.824}} = 0.140$

Std. Regressionsgleichung: $\hat{Y}^z = \hat{\beta}^z X^z = 0.140X^z$

Bei der std. Regression werden alle Variablen vor der Berechnung der Gerade z-transformiert. Es werden so die Einheiten der Variablen eliminiert. Bei der Einfachregression (nur eine unabh. Var.) entspricht der std. Koeff. gerade dem Korrelationskoeffizienten zwischen X und Y. Im multivariaten Fall können die std. Koeffizienten als relative Stärken der Effekte interpretiert werden. (Veränderung von Y in Standardabweichungen bei Erhöhung von X um eine Standardabweichung)

c) $R^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2} = \frac{0.353^2}{3.449 \cdot 1.824} = 0.020$

d) Hypothesen: $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0, \quad H_1: \beta_j \neq 0$ für mindestens ein j

Prüfgrösse: $F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} = \frac{0.020}{1-0.020}\frac{89-1-1}{1} = 1.776, \quad F \sim F(p, n-p-1)$

Kritischer Wert: $F_{1-\alpha}(p, n-p-1) = F_{0.95}(1, 87)$, für hohe Freiheitsgrade sind die Quantile der F-Verteilung meistens nur lückenhaft tabelliert, der genaue Wert kann aber durch Interpolation geschätzt werden, z.B.: $F_{0.95}(1, 87) \approx F_{0.95}(1, 60) + (F_{0.95}(1, 120) - F_{0.95}(1, 60))\frac{87-60}{120-60} = 4.001 + (3.920 - 4.001)\frac{27}{60} = 3.965$ (der genaue Wert wäre 3.951)

Entscheid: Da $F < F_{0.95}(1, 87)$ kann die Nullhypothese nicht abgelehnt werden. Das Modell liefert also keinen signifikanten Erklärungsbeitrag.

8-3. a) $H_0: \beta_j = 0, \quad H_1: \beta_j \neq 0, \quad T_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$

$n = 1\,287$	$\hat{\beta}$	$\hat{\sigma}$	T
Konstante)	1.631	.440	3.707
Umweltbetroffenheit (0–10)	$-2.51 \cdot 10^{-2}$	.016	-1.569
Umweltwissen (0–10)	.115	.023	5.000
Umweltbewusstsein (0–10)	.191	.028	6.821
Geschlecht (1 'weiblich')	.586	.096	6.104
Alter in Jahren	$1.03 \cdot 10^{-2}$	.003	3.433

Bei der vorliegenden Anzahl Freiheitsgrade $n - p - 1 = 1281$ kann die Standardnormalverteilung als Prüfverteilung verwendet werden: $z_{1-\alpha/2} = 1.960$. Somit sind mit Ausnahme des Koeffizienten für "Umweltbetroffenheit" alle Effekte signifikant von null verschieden.

b) $\hat{Y} = 1.631 - 0.0251 \cdot 10 + 0.115 \cdot 10 + 0.191 \cdot 10 + 0.0103 \cdot 35 = 4.80$

*8-4. Modell a): Ein zusätzliches Altersjahr führt zu einem Einkommenszuwachs vom durchschnittlich 78 Franken. Modell b): Ein zusätzliches Bildungsjahr erhöht das Einkommen um durchschnittlich CHF 483. Modell c): Männer verdienen im Durchschnitt CHF 1665 mehr als Frauen. Modell d): Vor allem der Geschlechtseffekt verringert sich deutlich. Das heisst, die Einkommensdifferenz zwischen Frauen und Männern ist teilweise auf Unterschiede in der Bildung zurückzuführen (oder dem Alter; allerdings wird das Durschnittsalter für Männer und Frauen etwa gleich sein). Anteil erklärter Varianz in Modell d): 33%. Syntax:

```
filter by vz /* Variable "vz" aus Aufgabe 6-6 */.
regression /dependent=netinco /method=enter alter.
regression /dependent=netinco /method=enter bildung.
regression /dependent=netinco /method=enter sex.
regression /dependent=netinco /method=enter alter bildung sex.
filter off.
```

*8-5. Syntax:

```
filter by vz.
plot /format=regression /plot=netinco with alter.
plot /format=regression /plot=netinco with bildung.
filter off.
```

*8-6. Da die abhängige Variable logarithmiert wurde, können die Effekte näherungsweise als Prozenteffekte interpretiert werden. Eine Erhöhung der Bildung um ein Jahr führt bei den Frauen zu einem Einkommenszuwachs von gut 6%, bei den Männern gut 9%. Genauer: $\exp(0.06390) - 1 = 6.60\%$ und $\exp(0.09303) - 1 = 9.75\%$. Bei der Berufserfahrung ist der Effekt etwas schieriger zu interpretieren, da zwei Parameter berücksichtigt werden müssen. Aufgrund der Vorzeichen lässt sich festhalten, dass der Effekt eines zusätzlichen Jahres Berufserfahrung zunächst positiv ist, mit zunehmender Berufserfahrung aber immer kleiner und schliesslich negativ wird. Der Wendepunkt liegt bei den Frauen bei $0.0457/(2 \cdot 0.000805) = 28.4$ Jahren Berufserfahrung, bei den Männern bei $0.06308/(2 \cdot 0.000989) = 31.9$ Jahren. Eine Veränderung von beispielsweise 0 auf 1 Jahr Berufserfahrung führt bei den Frauen zu eine Einkommenszuwachs von $\exp(0.0457 - 0.000805) - 1 = 4.6\%$, eine Veränderung von 40 auf 41 Jahre zu einem Einkommensrückgang: $\exp(0.0457 - 0.000805 \cdot (41^2 - 40^2)) - 1 = -1.9\%$. Syntax:

```
if netinco>0 lny=ln(netinco).
rename variables bildung=educ.
compute exp=alter-educ-6.5.
if exp<0 exp=0.
compute exp2=exp**2.
if e14>0 lnh=ln(e14).
descriptives lny educ exp exp2 lnh.

sort cases by sex.
split file by sex.
regression /dependent=lny /method=enter educ exp exp2 lnh.
split file off.
```

*8-7. Hinweis: „S17C“ und „S17G“ sind umgekehrt gepolt. Der Zusammenhang ist negativ, d.h. je mehr „rechts“, desto tiefer das Umweltbewusstsein ($r = -0.286$). Eine Erhöhung auf der Skala der politischen Orientierung um einen Punkt führt zu einem Rückgang um 0.271 Punkte auf der Skala des Umweltbewusstseins. (Problematisch: die Skalen werden hier als metrisch interpretiert.) Syntax:

```
compute ub=((5-s17a)+(5-s17b)+(s17c-1)+(5-s17d)
           +(5-s17e)+(5-s17f)+(s17g-1)+(5-s17h))*5/16.
variable label ub "Umweltbewusstsein".
value label ub 0"tief" 10"hoch".
descriptives ub.

correlate ub b30.
nonpar corr ub b30.
regression /dependent=ub /method=enter b30.
```