Datenanalyse mit Excel

Wintersemester 2013/14

KORRELATIONRECHNUNG

Korrelationsrechnung

- Ziel der Korrelationsrechnung besteht im bivariaten Fall darin, die **Stärke** des **Zusammenhangs** zwischen zwei interessierenden statistischen Variablen **aufzudecken** und zu **quantifizieren**.
- Beispiele: Tabakpreis und Nikotinkonsum, Fertilität und Frauenerwerbsquote, Berufserfahrung und Gehalt…
- Fragen: Besteht eine Beziehung zwischen den zwei interessierenden Variablen? Wie sieht diese Beziehung aus? Wie kann die Beziehung quantifiziert werden?

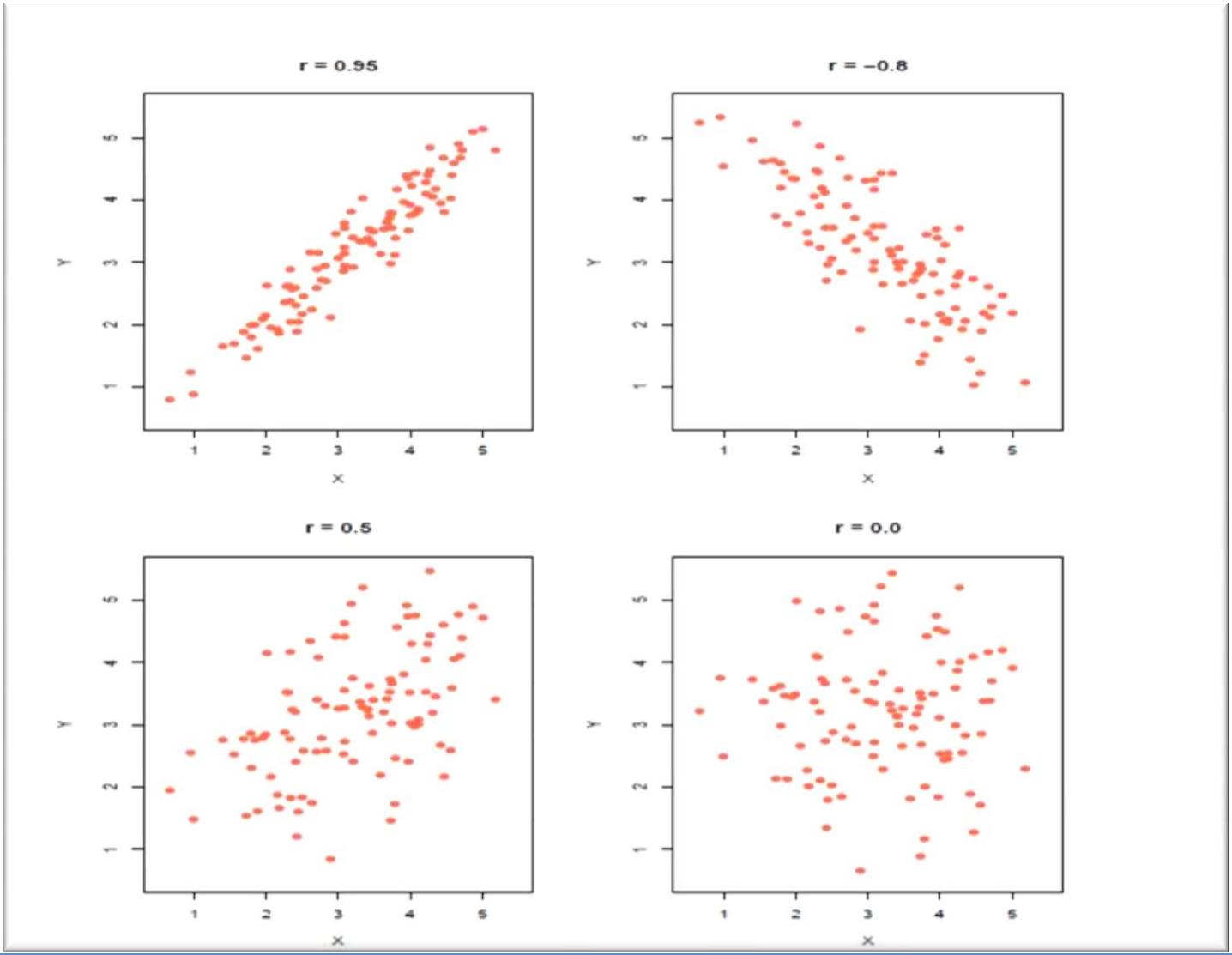
- Je nachdem, von welcher **Skalenqualität** die jeweiligen Untersuchungsvariablen sind, werden unterschiedliche Berechnungsmethoden zu verwenden.

Streudiagramm

- In einem ersten Schritt kann man die beobachteten Wertepaare (x_1 , y_1) (x_2 , y_2) (x_3 , y_3) … (x_n , y_n) in einem **Streudiagramm** visualisieren.
- Ein Streudiagramm ermöglicht die grafische Darstellung eines Zusammenhangs von zwei (zumeist metrischen) Variablen.
- Der Zusammenhang wird als Punktewolke in einem Koordinatensystem dargestellt.
- Die Ausprägung einer Variablen A wird auf der x-Achse abgetragen und die Ausprägung einer Variablen B auf der y-Achse.
- Das Streudiagramm gibt Hinweise auf die Beziehung der Wertepaare. Grundsätzlich kann die Beziehung klassifiziert werden in:
  - Keine Beziehung: undefinierte Punktewolke.
  - Lineare Beziehung: Punkte formen (grob) eine Linie.
  - Nichtlineare Beziehung: Punkte formen (grob) eine Kurve.

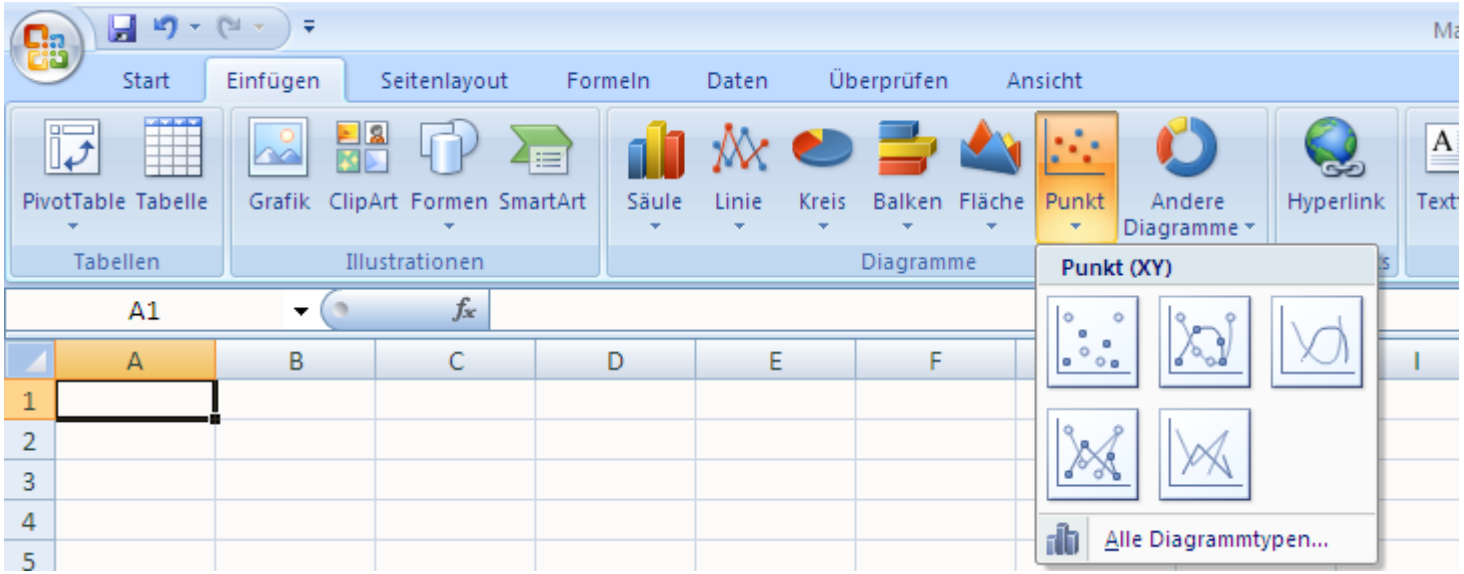
- Das Streudiagramm erlaubt zudem, Ausreißer zu identifizieren.

Streudiagramm

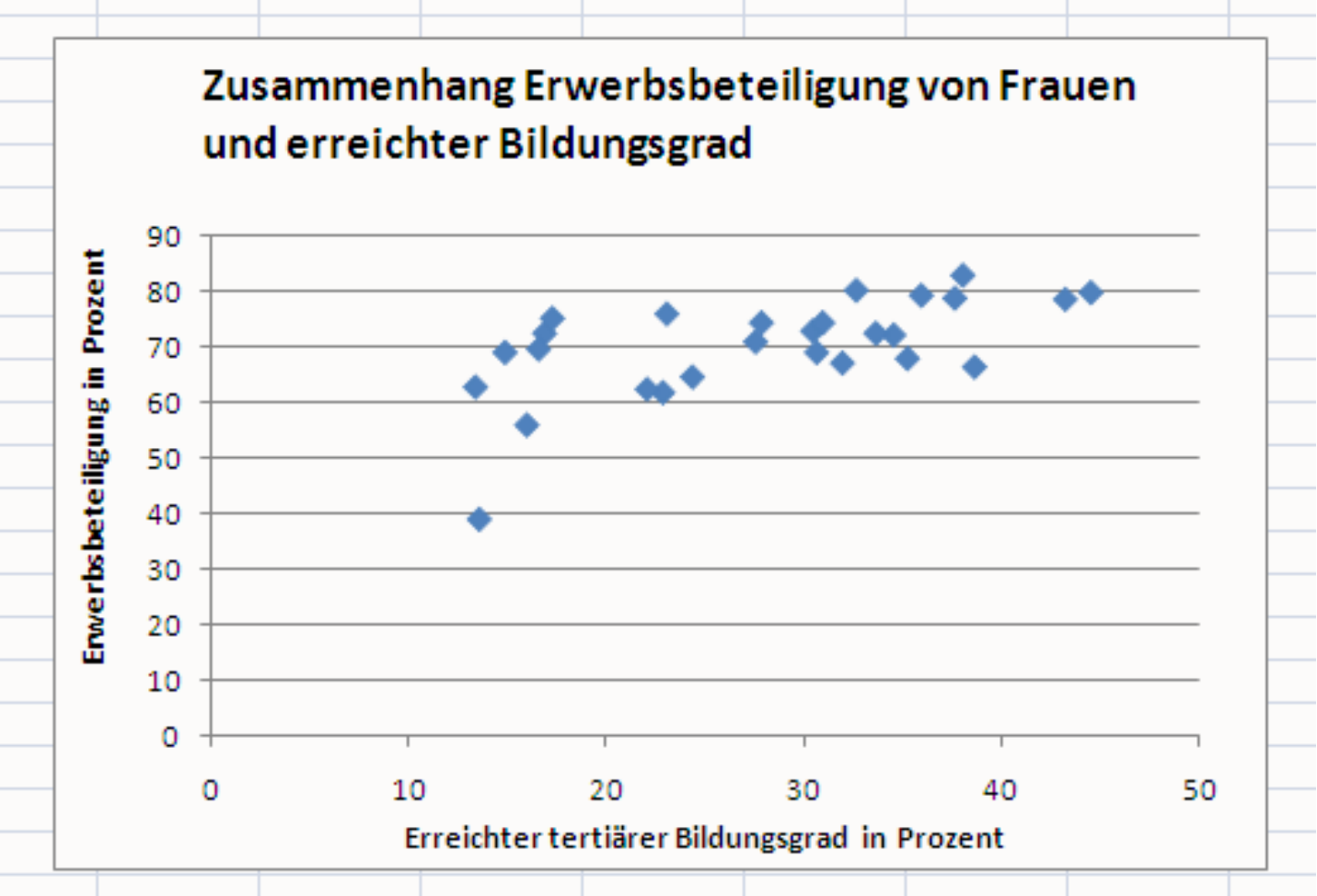


Streudiagramm

Öffnen Sie die Datei Erwerbsquote_Frauen_Bildungsgrad.xlsx und erstellen Sie ein Streudiagramm.



Streudiagramm



Korrelationsrechnung bei metrisch skalierten Variablen

Zusammenhangsrechnung bei metrisch skalierten Variablen:

- Geeignet: Berechnung mit dem Korrelationskoeffizient von Bravais/Pearson:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} * \sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

- Excel Funktion **Korrel(.)**
- Die Korrelation nimmt Werte zwischen -1 und 1 an.
- Bei r = +1 liegt ein maximal starker, gleichgerichteter Zusammenhang vor. Dies bedeutet, dass die Punkte (x_n , y_n) alle auf einer Geraden liegen.
- Bei r= -1 liegt ein maximal starker, gegenläufiger Zusammenhang vor.
- Bei r = 0 liegt kein Zusammenhang vor.
- Zwischenwerte können entsprechend interpretiert werden.

Korrelationsrechnung bei metrisch skalierten Variablen

- Das Quadrat des Korrelationskoeffizienten von Bravais/Pearson ist der **Determinationskoeffizient.**
- Er zeigt den Anteil der Varianz der abhängigen Variablen Y, der durch die Variabilität der unabhängigen Variablen X, unter Nutzung der Annahme des linearen Zusammenhangs zwischen beiden Variablen, statistisch erklärt wird.

- Achtung: der Korrelationskoeffizient von Bravais/Pearson bemisst die Stärke eines bivariaten Zusammenhangs, von dem implizit vorausgesetzt wird, dass er **linear** ist!

Korrelationsrechnung bei metrisch skalierten Variablen

	A	B	C
0	Tschechische Republik	69,0	14,9
1	Dänemark	78,7	37,7
2	Deutschland	75,9	23,1
3	Estland	79,7	44,6
4	Irland	66,4	38,7
5	Griechenland	61,8	22,9
6	Spanien	69,0	30,7
7	Frankreich	72,8	30,5
8	Italien	56,0	16,0
9	Zypern	72,1	34,6
0	Lettland	80,1	32,7
1	Litauen	79,2	36,0
2	Luxemburg	67,1	32,0
3	Ungarn	62,4	22,1
4	Malta	39,1	13,6
5	Niederlande	74,3	31,0
6	Österreich	72,4	16,9
7	Polen	64,6	24,4
8	Portugal	75,1	17,3
9	Rumänien	62,8	13,4
0	Slowenien	74,3	27,9
1	Slowakei	69,6	16,6
2	Finnland	78,5	43,3
3	Schweden	82,8	38,1
4	Vereinigtes Königreich	72,4	33,7
5			
6	**Korrelation**		**0,61**
7	**Determinationskoeffzient**		**0,37**

Faustregeln für die Interpretation des Zusammenhangs:

Grob:

- Schwacher Zusammenhang: 0 - 0.30
- Mittlerer Zusammenhang: 0.30 - 0.70
- Starker Zusammenhang: 0.70 – 1

Detaillierter

- Sehr geringer Zusammenhang: 0 - 0.10
- Geringer Zusammenhang: 0.10 - 0.30
- Mittlere Zusammenhang: 0.40 - 0.70
- Hoher Zusammenhang: 0.70 - 0.90
- Sehr hoher Zusammenhang: über 0.90

Achtung: keine universell geltenden Regeln!!!

=KORREL(B8:B34;C8:C34)
=C36^2

Korrelationsrechnung

Fehlinterpretation:

- Gefundene Korrelationen müssen mit Vorsicht hinterfragt werden.
- Eine Scheinkorrelation kann auftreten, weil beide Variable hoch mit einer Dritten korrelieren. Bekanntestes Beispiel: Geburten und Störche.
- Eine verdeckte Korrelation kann vorkommen, wenn sich die Korrelationen von Subgruppen der Stichprobe gegenseitig neutralisieren.

Korrelation und Kausalität:

- Aus der Kennzahl selbst kann nicht abgelesen werden, was Ursache und was Wirkung ist.
- Kennzahlen können nur messen, ob die Daten einen statistischen Zusammenhang aufweisen, nicht, ob es auch tatsächlich einen kausalen Zusammenhang gibt!

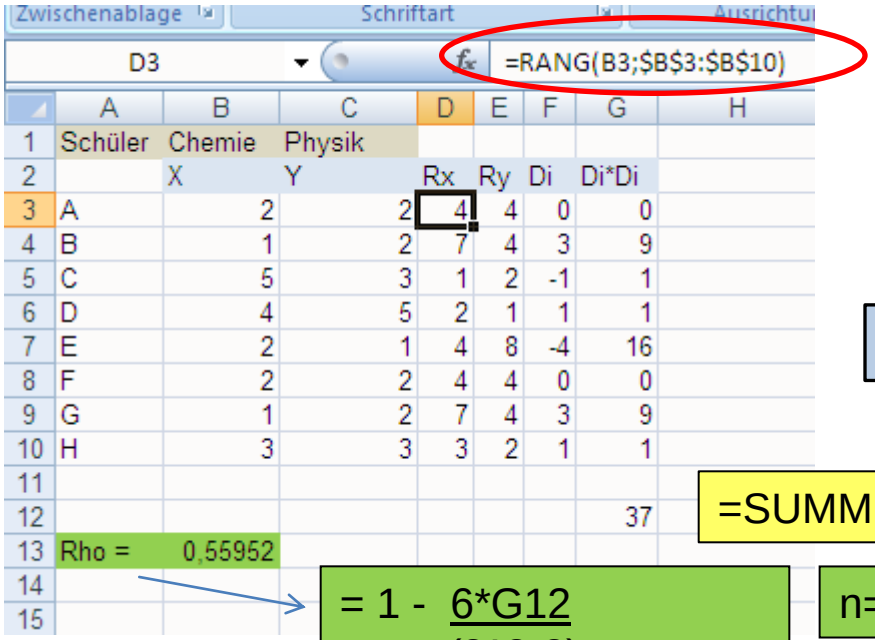
Korrelationsrechnung bei ordinal skalierten Variablen

- Geeignet: Der Rangkorrelationskoeffizient von Spearman:

$$r_{SP} = 1 - \frac{6 * \sum D_i^2}{n^3 - n}$$

- Berechnung in Excel ‚per Hand‘.
- Interpretation des Rangkorrelationskoeffizienten wie schon bei dem Korrelationskoeffizienten von Bravais/Pearson.

Korrelationsrechnung bei ordinal skalierten Variablen



	A	B	C	D	E	F	G	H
1	Schüler	Chemie	Physik					
2		X	Y	Rx	Ry	Di	Di*Di	
3	A	2	2	4	4	0	0	
4	B	1	2	7	4	3	9	
5	C	5	3	1	2	-1	1	
6	D	4	5	2	1	1	1	
7	E	2	1	4	8	-4	16	
8	F	2	2	4	4	0	0	
9	G	1	2	7	4	3	9	
10	H	3	3	3	2	1	1	
11								
12							37	
13	Rho =	0,55952						
14								
15								

Die Rangplätze können über die Excel Funktion **Rang(.)** ermittelt werden.

Di=Rx-Ry

=SUMME(G3:G11)

= 1 - $\frac{6*G12}{(8^3-8)}$

n= 8 Schüler

Korrelationsrechnung bei nominal skalierten, dichotomen Variablen

- Für die Zusammenhangsrechnung von zwei nominalen Variablen mit jeweils zwei Ausprägungen ist der Vierfelder-Koeffizient geeignet:

$$\phi = \frac{a * d - b * c}{\sqrt{S_1 * S_2 * S_3 * S_4}}$$

X	x1	x2	Summe
Y			
y1	a	b	S1
y2	c	d	S2
Summe	S3	S4	n

B7 | =(B3*C4-C3*B4)/WURZEL(D3*D4*B5*C5)

	A	B	C	D	E
1	Geschlecht	männlich	weiblich	Summe	
2	Zustimmung zur Einführung des Rachverbots				
3	ja	2	4	6	
4	nein	5	1	6	
5	Summe	7	5	12	
6					
7	Phi =	-0,50709			
8					

Korrelationsrechnung bei nominal skalierten, dichotomen Variablen

- **Alternativ**: Verwendung des Korrelationskoeffizienten von Bravais/Pearson. Achtung: nur bei **zwei dichotomen** Variablen.

18	Geschlecht		x männlich	x weiblich	Summe
19			1	0	
20	Zustimmung zur Einführung des Rachverbots				
21	y ja	1	2	4	6
22	y nein	0	5	1	6
23	Summe		7	5	12

x	y
1	1
1	1
1	0
1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1
0	0
Korrelation	-0,50709255

Die Kombination (1,1) kommt insgesamt 2 Mal vor. Die Kombination (0,1) kommt insgesamt 4 Mal vor …

Korrelationsrechnung bei nominal skalierten, nicht-dichotomen Variablen

- Für die Zusammenhangsrechnung von zwei nominalen, nicht-dichotomen Variablen eignet sich der Kontingenzkoeffizient von Pearson.
- Für die Berechnung des Kontingenzkoeffizienten werden die Daten zunächst in einer Kreuztabelle (Pivottabelle) dargestellt (kein notwendiger Schritt).

- Unabhängig von der Berechnung des Kontingenzkoeffizienten eignet sich diese tabellarische Darstellungsform für große Datenmengen.

Exkurs: Erstellen einer Pivot Tabelle

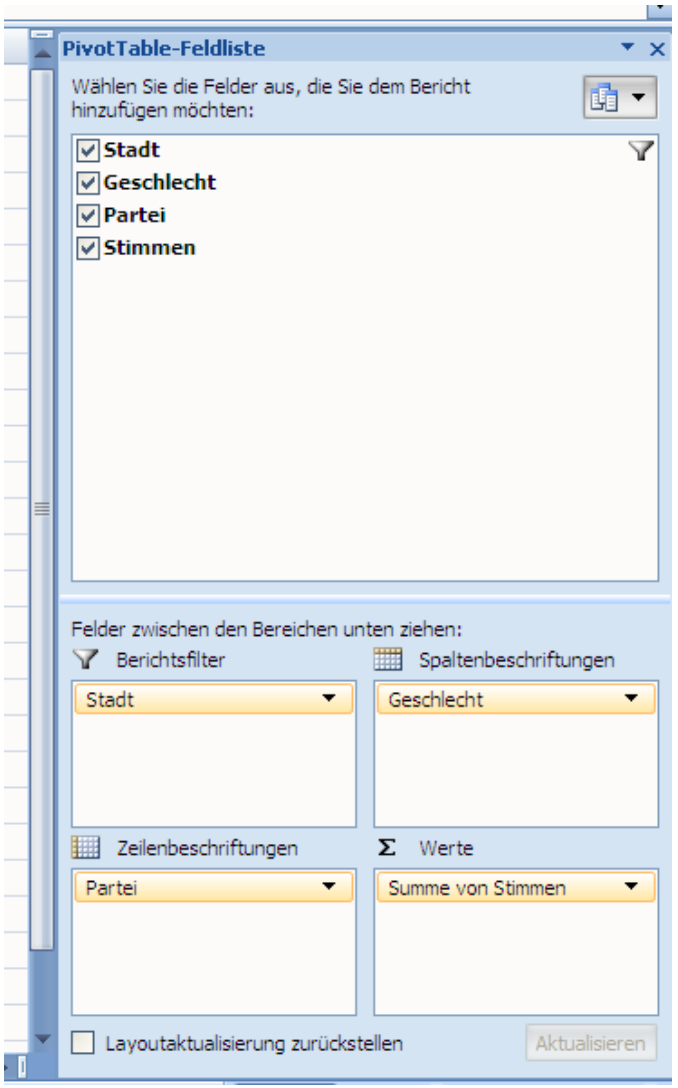
- Eine Pivot-Tabelle ist eine Auswertungstabelle. Dieses Instrument bietet die Möglichkeit, große Datensätze anzuordnen, zusammenzufassen und zu analysieren. Mit Hilfe von Drop-Down Listen können Sie die Tabelle auf die interessierenden Merkmalsausprägungen reduzieren.

Öffnen Sie die Datei ‚Wahlen.xlsx'.
Markieren Sie den Tabellenbereich und gehen Sie auf ‚Einfügen' → ‚Pivot Tabelle' → ‚Pivot Table'.
Wählen Sie für die Ablage der Pivot Tabelle die Option ‚Neues Arbeitsblatt'.
Drücken Sie auf OK.
Ein neues Tabellenblatt mit der Bezeichnung ‚Pivottabelle' öffnet sich.

Stadt	Geschlecht	Partei	Stimmen
Stadt A	männlich	CDU/CSU	388
Stadt A	männlich	SPD	325
Stadt A	männlich	FDP	54
Stadt A	männlich	Die Grünen	88
Stadt A	männlich	Sonstige	28
Stadt A	weiblich	CDU/CSU	419
Stadt A	weiblich	SPD	311
Stadt A	weiblich	FDP	62
Stadt A	weiblich	Die Grünen	98
Stadt A	weiblich	Sonstige	25
Stadt B	männlich	CDU/CSU	400
Stadt B	männlich	SPD	420
Stadt B	männlich	FDP	40
Stadt B	männlich	Die Grünen	70
Stadt B	männlich	Sonstige	25
Stadt B	weiblich	CDU/CSU	350
Stadt B	weiblich	SPD	400
Stadt B	weiblich	FDP	50
Stadt B	weiblich	Die Grünen	60
Stadt B	weiblich	Sonstige	30

Exkurs: Erstellen einer Pivot Tabelle

Ziehen Sie die Variable ‚Stadt' in das Feld ‚Bereichsfilter',
die Variable ‚Geschlecht' in das Feld ‚Spaltenbeschriftung', die Variable ‚Partei' in das Feld ‚Zeilenbeschriftung' und die Variable ‚Stimmen' in das Feld ‚Σ Werte'.



Exkurs: Erstellen einer Pivot Tabelle

Durch die Drop-Down Funktion können Sie relevante Merkmals-ausprägungen auswählen.

Stadt	(Alle)		
Summe von Stimmen	Geschlecht		
Partei	männlich	weiblich	Gesamtergebnis
CDU/CSU	788	769	1557
Die Grünen	158	158	316
FDP	94	112	206
Sonstige	53	55	108
SPD	745	711	1456
Gesamtergebnis	1838	1805	3643

Korrelationsrechnung bei nominal skalierten, nicht-dichotomen Variablen

- Wählen Sie für die Berechnung des Kontingenzkoeffizienten Stadt A aus. Kopieren Sie die Tabelle in ein neues Tabellenblatt und wählen Sie beim Einfügen die Option ‚Werte und Zahlenformate'.
- Zur Berechnung des Kontingenzkoeffizienten wird zunächst die Frage beantwortet, wie viele männliche (weibliche) CDU/CSU- (SPD-, FDP-, die Grünen-, Sonstige) Wähler zu erwarten wären, wenn Unabhängigkeit bestünde.
- Bspw. Berechnung des Erwartungswert der männlichen CDU/CSU-Wähler:

  (807*883)/1798=396, 32.
- Die restlichen Erwartungswerte werden entsprechend berechnet (Berechnung mit absoluten Zellbezügen!).

Korrelationsrechnung bei nominal skalierten, nicht-dichotomen Variablen

Zwischenablage | Schriftart | Ausrichtung | Zahl | Formatvorlag

G5 | =$B$10*D5/$D$10

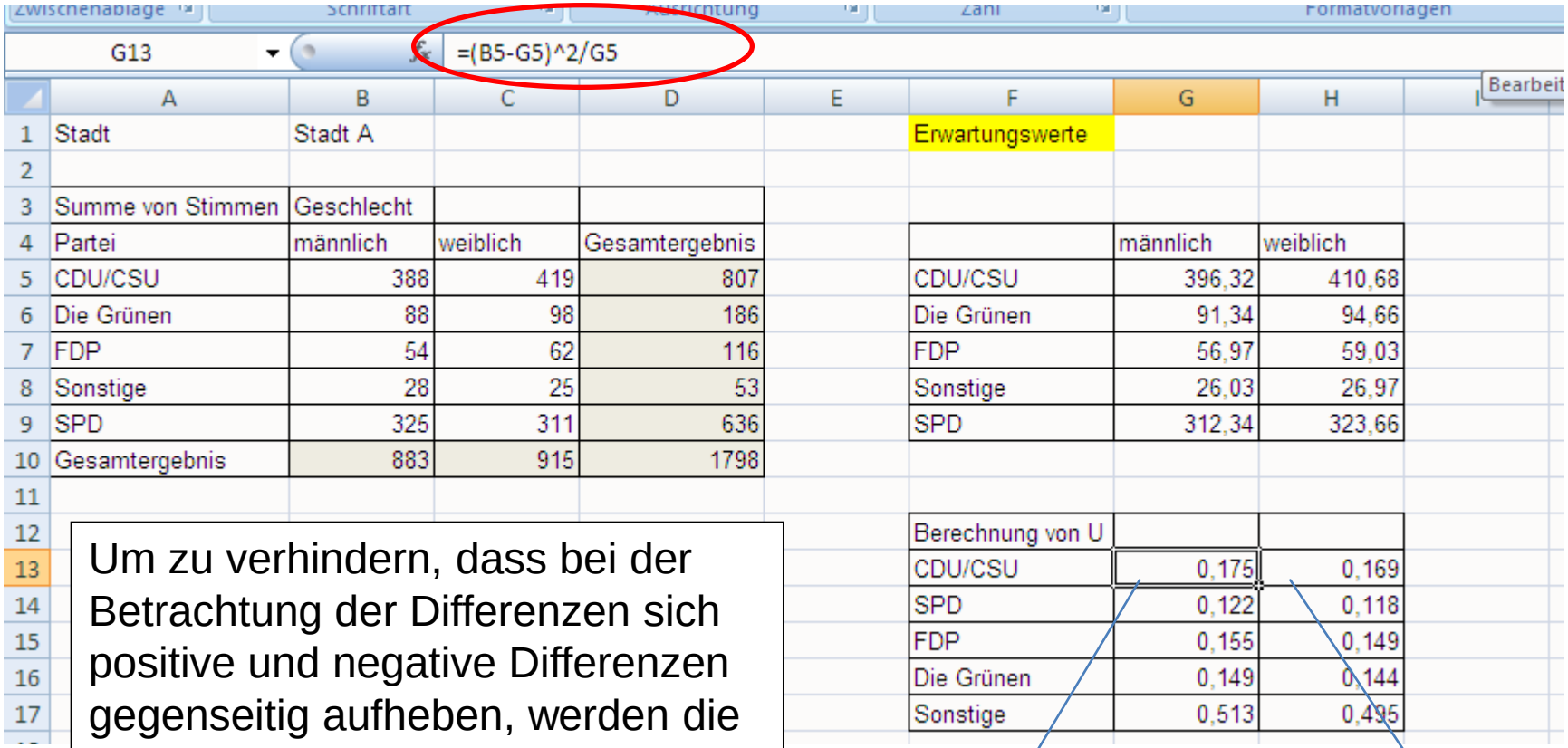
	A	B	C	D	E	F	G	H
1	Stadt	Stadt A				Erwartungswerte		
2								
3	Summe von Stimmen	Geschlecht						
4	Partei	männlich	weiblich	Gesamtergebnis			männlich	weiblich
5	CDU/CSU	388	419	807		CDU/CSU	396,32	410,68
6	Die Grünen	88	98	186		Die Grünen	91,34	94,66
7	FDP	54	62	116		FDP	56,97	59,03
8	Sonstige	28	25	53		Sonstige	26,03	26,97
9	SPD	325	311	636		SPD	312,34	323,66
10	Gesamtergebnis	883	915	1798				

= D5*$B$10/$D$10

= D5*$C$10/$D$10

Folgende Argumentation: Je weiter die bei Unabhängigkeit zu erwartenden Werte von denen abweichen, die tatsächlich beobachtet werden, desto weiter ist der Befund von der Unabhängigkeit entfernt bzw. desto stärker hängen die betrachteten Variablen voneinander ab. Das heißt, je größer die Differenzen sind, desto stärker ist der Zusammenhang.

Korrelationsrechnung bei nominal skalierten, nicht-dichotomen Variablen



Zwischenablage | Schriftart | Ausrichtung | Zahl | Formatvorlagen

G13 | =(B5-G5)^2/G5

	A	B	C	D	E	F	G	H
1	Stadt	Stadt A				Erwartungswerte		
2								
3	Summe von Stimmen	Geschlecht						
4	Partei	männlich	weiblich	Gesamtergebnis			männlich	weiblich
5	CDU/CSU	388	419	807		CDU/CSU	396,32	410,68
6	Die Grünen	88	98	186		Die Grünen	91,34	94,66
7	FDP	54	62	116		FDP	56,97	59,03
8	Sonstige	28	25	53		Sonstige	26,03	26,97
9	SPD	325	311	636		SPD	312,34	323,66
10	Gesamtergebnis	883	915	1798				
11								
12						Berechnung von U		
13						CDU/CSU	0,175	0,169
14						SPD	0,122	0,118
15						FDP	0,155	0,149
16						Die Grünen	0,149	0,144
17						Sonstige	0,513	0,495

Um zu verhindern, dass bei der Betrachtung der Differenzen sich positive und negative Differenzen gegenseitig aufheben, werden die Differenzen quadriert. Zusätzlich werden sie durch den jeweiligen Erwartungswert relativiert.

=(B5-G5)^2/G5

=(B5-G5)^2/G5

Korrelationsrechnung bei nominal skalierten, nicht-dichotomen Variablen

I19 | =SUMME(G19:H19)

	A	B	C	D	E	F	G	H	I
1	Stadt	Stadt A				Erwartungswerte			
2									
3	Summe von Stimmen	Geschlecht							
4	Partei	männlich	weiblich	Gesamtergebnis			männlich	weiblich	
5	CDU/CSU	388	419	807		CDU/CSU	396,32	410,68	
6	Die Grünen	88	98	186		Die Grünen	91,34	94,66	
7	FDP	54	62	116		FDP	56,97	59,03	
8	Sonstige	28	25	53		Sonstige	26,03	26,97	
9	SPD	325	311	636		SPD	312,34	323,66	
10	Gesamtergebnis	883	915	1798					
11									
12						Berechnung von U			
13						CDU/CSU	0,175	0,169	
14						SPD	0,122	0,118	
15						FDP	0,155	0,149	
16						Die Grünen	0,149	0,144	
17						Sonstige	0,513	0,495	
18									
19						Summe	1,114	1,075	2,189

Die quadrierten, relativen Abweichungen werden aufaddiert und zusammengefasst (hier 2,189). Bei Unabhängigkeit ergibt dieser Wert 0. Der Wert ist umso größer, je weiter die beobachteten von den erwarteten Häufigkeiten entfernt sind.

Korrelationsrechnung bei nominal skalierten, nicht-dichotomen Variablen

I19 | fx | =SUMME(G19:H19)

	A	B	C	D	E	F	G	H	I
1	Stadt	Stadt A				Erwartungswerte			
2									
3	Summe von Stimmen	Geschlecht							
4	Partei	männlich	weiblich	Gesamtergebnis			männlich	weiblich	
5	CDU/CSU	388	419	807		CDU/CSU	396,32	410,68	
6	Die Grünen	88	98	186		Die Grünen	91,34	94,66	
7	FDP	54	62	116		FDP	56,97	59,03	
8	Sonstige	28	25	53		Sonstige	26,03	26,97	
9	SPD	325	311	636		SPD	312,34	323,66	
10	Gesamtergebnis	883	915	1798					
11									
12						Berechnung von U			
13						CDU/CSU	0,175	0,169	
14						SPD	0,122	0,118	
15						FDP	0,155	0,149	
16						Die Grünen	0,149	0,144	
17						Sonstige	0,513	0,495	
18									
19						Summe	1,114	1,075	2,189
20								C =	0,03487372

Zur Berechnung des Kontingenzkoeffizienten nach Pearson wird die folgende Formel verwendet:

$$C = \sqrt{\frac{U}{U+n}}$$

Mit 0,035 besteht hier nur ein sehr geringer Zusammenhang zwischen Geschlecht und bevorzugter Partei.

=WURZEL(I17/(I17+1798))

Übung 8

1. Suchen Sie die Daten für die Erwerbstätigkeit von Frauen sowie Daten für die Fertilitätsrate in den EU-Ländern. Berechnen Sie Korrelation. Erstellen Sie eine geeignete Grafik und beschreiben Sie die Grafik in wenigen Sätzen.

Viel Erfolg

Bei Fragen:

j.bossert@wiwipa.uni-saarland.de