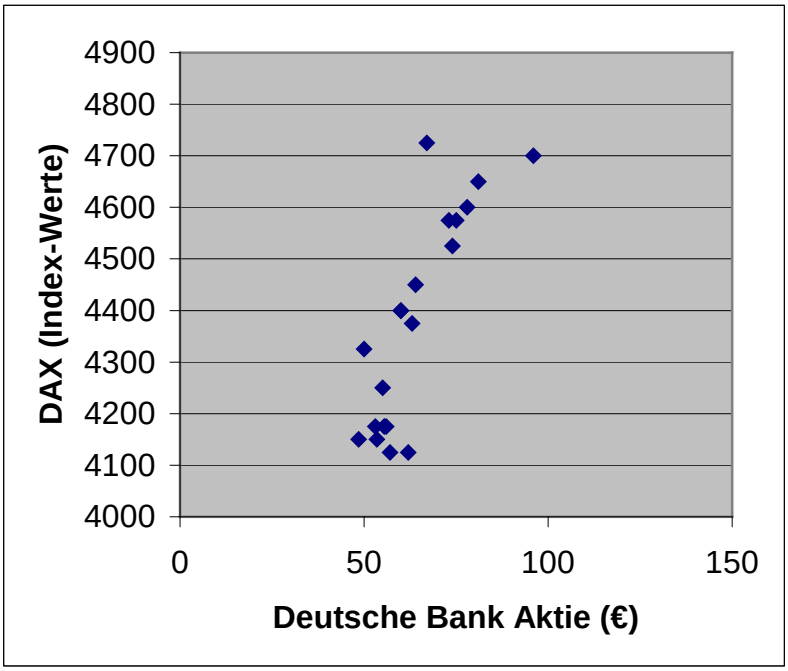
Linearer Zusammenhang von Datenreihen

Vielen Problemen liegen (möglicherweise) lineare Zusammenhänge zugrunde:

- Mein Internetanbieter verlangt eine Grundgebühr und rechnet minutenweise ab
- Ich bestelle ein Taxi und bezahle die Anfahrt und darüber hinaus pro gefahrenen Kilometer einen festen Betrag

In den Sozial-, Natur- oder Wirtschaftswissenschaften gewinnt man Daten jedoch häufig durch Befragungen, Messungen, Entwicklungen, ... die nicht sofort einen Rückschluss auf einen möglicherweise vorhandenen linearen Zusammenhang zulassen. Die folgende Tabelle und Grafik zeigen zu verschiedenen Zeitpunkten (fiktive) Werte der Deutschen Bank Aktie und des Deutschen Aktien Index (DAX):

Datum	Deutsche Bank-Aktie (€)	DAX (Index-Werte)
03.04.2007	63	4375
04.04.2007	74	4525
05.04.2007	81	4650
06.04.2007	78	4600
07.04.2007	55	4250
10.04.2007	62	4125
11.04.2007	96	4700
12.04.2007	75	4575
13.04.2007	64	4450
14.04.2007	67	4725
17.04.2007	48,5	4150
18.04.2007	73	4575
19.04.2007	79	4925
20.04.2007	55,5	4175
21.04.2007	56	4175
24.04.2007	53,5	4150
25.04.2007	60	4400
26.04.2007	60	4400
27.04.2007	50	4325
28.04.2007	53	4175
03.05.2007	57	4125



Betrachtet man die Graphik, so könnte man vermuten, dass die Entwicklung des Wertes der Deutschen Bank Aktie ein guter Indikator für die Entwicklung des DAX ist. Wäre dies der Fall, so könnte man die Deutsche Bank Aktie nutzen, um (mehr oder weniger sichere) Aussagen über die Entwicklung des DAX treffen zu können.

a) Würden Sie aufgrund der Daten einen (linearen) Zusammenhang vermuten?
b) Angenommen, man könnte solch einen Zusammenhang auf irgendeine Art mehr oder wenig sicher nachweisen, wozu könnte das nützlich sein?

Man kann sich also fragen, ob und wenn ja inwiefern man solch einen Zusammenhang auch rechnerisch beschreiben kann.

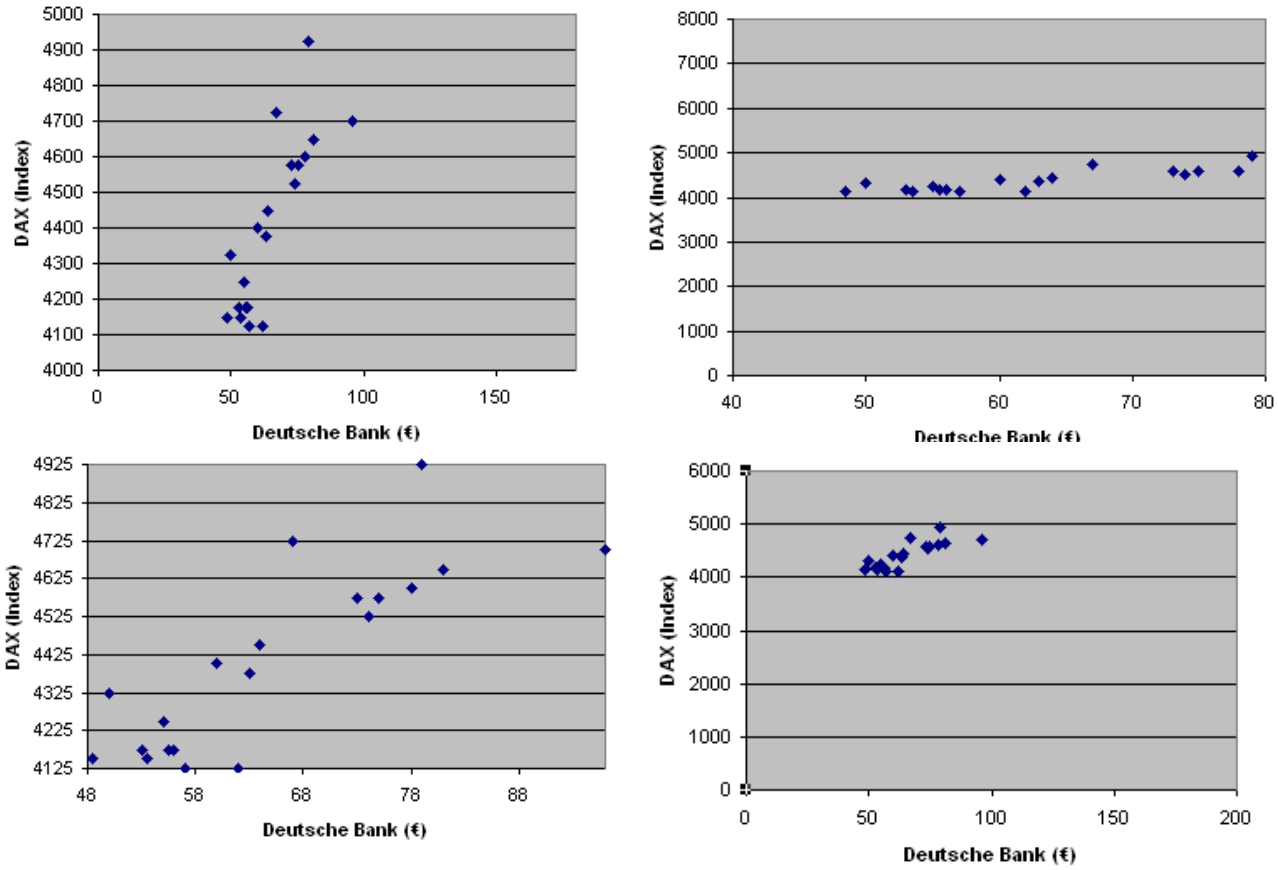
Das bekannteste Verfahren in der Statistik, mit dem „die Stärke" eines linearen Zusammenhanges zweier Datenreihen gemessen wird, ist die Bestimmung des Korrelationskoeffizienten. Die Idee des Korrelationskoeffizienten geht auf Francis Galton (1822-1911) zurück. Aber erst sein Schüler Karl Pearson (1857-1936) arbeitete 1897 das Konzept zu Galtons Idee aus. Da auch Auguste Bravais (1811-1863) daran beteiligt gewesen war, benennt man den Korrelations-koeffizienten auch Bravais-Pearson-Korrelationskoeffizient.

Karl Pearson

Die Idee des Korrelationskoeffizienten

Ob ein linearer Zusammenhang vorliegt, sollte **nicht** anhand der Datendarstellung beantwortet werden. Durch Manipulation der Datendarstellung kann ein linearer Zusammenhang hervorgehoben aber auch „vertuscht" werden. Durch verschiedene Achsenskalierungen kann diese Manipulation erzeugt werden.

c) Welche Darstellungen legen einen linearen Zusammenhang nahe, welche weniger?



„Teure und billige" Aktien

Ob Aktien teuer oder billig sind macht ohne Bezugspunkt streng genommen keinen Sinn, denn ab wann ist eine Aktie z.B. teuer: Ab 50€, ab 60€ oder erst ab 70€? Die Begriffe teuer und billig können aber in **Bezug auf einen Durchschnitt** Sinn machen: Z.B. teurer oder billiger als der Durchschnitt. Nicht die absoluten Werte, sondern die Abweichung der Werte von ihrem Mittelwert geben den Begriffen teuer und billig eine sinnvolle (und messbare) Bedeutung.

Der 1. Schritt:

Pearsons erste Idee war es daher, von allen Daten einer Datenreihe den jeweiligen Mittelwert abzuziehen. Hierdurch streuen nun die Daten mehr oder weniger symmetrisch um 0 herum (siehe Abbildung).

Mit welcher Einheit soll man rechnen?

Auch wenn die Begriffe teuer und billig nun einen Sinn machen, so könnte man die unterschiedlichen Skalierungen der Achsen immer noch manipulieren (z.B. durch Strecken und Stauchen), um einen linearen Zusammenhang zu verstärken oder zu vertuschen. **Dies ist aber nur deshalb möglich, weil beide Koordinatenachsen unterschiedlich skaliert sind** (d.h. sie haben verschiedene Maßeinheiten und -zahlen). Die Skalen müssten also derart „vereinheitlicht" werden, dass eine Einheit auf der einen Achse auch einer Einheit auf der anderen Achse entspricht.

Der 2. Schritt:

Pearsons zweite Idee war es daher, alle Daten einer Datenreihe nach dem 1. Schritt durch deren jeweilige Standardabweichung zu teilen. Hierdurch werden die Daten stärker „beieinandergerückt", als sie vorher noch lagen.

Die folgende Tabelle zeigt zu dem 1. und 2. Schritt ein Zahlenbeispiel:

		1.Schritt		2.Schritt	
Kurswert (KW)	Dax	KW-MW	DAX-MW	(KW-MW)/SA	(DAX-MW)/SA
20	4300	-6	0	**-1,12**	**0,00**
25	4325	-1	25	**-0,19**	**1,22**
33	4275	7	-25	**1,31**	**-1,22**
Mittelwert (MW)					
26	4300				
Standardabweichung (SA)					
5,35	20,41				

d) Vervollständigen Sie:

		1.Schritt		2.Schritt	
Kurswert (KW)	Dax	KW-MW	DAX-MW	(KW-MW)/SA	(DAX-MW)/SA
24	4210				
31	4260				
44	4250				
Mittelwert (MW)					
Standardabweichung (SA)					

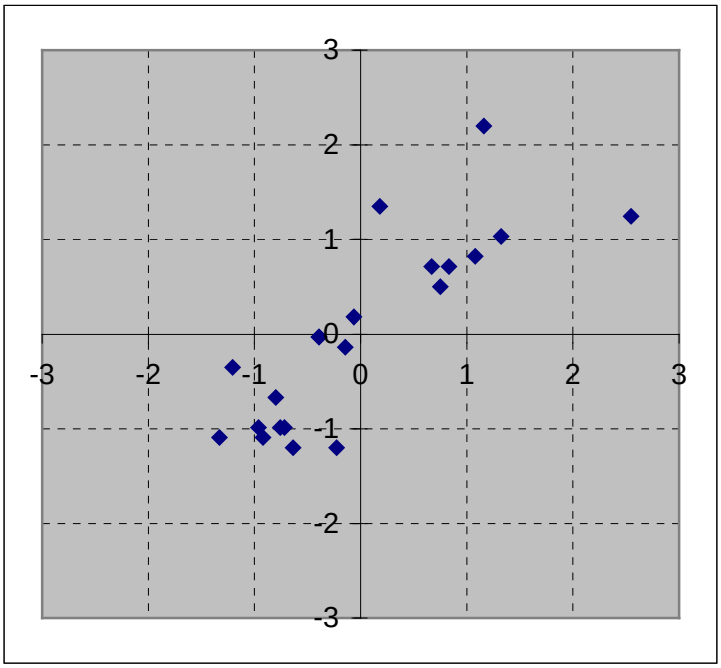
e) Vervollständigen sie

		1.Schritt		2.Schritt	
Kurswert (KW)	Dax	KW-MW	DAX-MW	(KW-MW)/SA	(DAX-MW)/SA
56	5240				
62	5100				
71	5122				
77	5199				
Mittelwert (MW)					
Standardabweichung (SA)					

f) Berechnen sie zur ersten Tabelle, d) und e) jeweils den Mittelwert und die Standardabweichung der 3 bzw. 4 neu gewonnenen Daten zu (KW-MW)/SA und (DAX-MW)/SA. Was fällt auf? Wie lässt es sich erklären?

Was haben diese neuen Daten mit linearem Zusammenhang zu tun?

Liegen mit der Idee aus dem 1. und 2. Schritt viele Datenpunkte im rechten oberen und linken unteren Quadranten, so deutet dies schon einmal näherungsweise auf einen linearen Zusammenhang hin. Würden diese Punkte nun auch noch alle auf einer Geraden liegen, so wäre der lineare Zusammenhang bereits optimal zu sehen. Pearson suchte aber nach einer **Zahl**, die diesen Zusammenhang misst (ohne dafür eine Grafik zu erstellen – was verständlich ist, wenn man sich überlegt, man müsste tausende von Daten erst zeichnen bevor man sie auswerten könnte). Rechts abgebildet sind z.B. die Deutsche Bank und DAX Daten vom Anfang nachdem sie durch Schritt 1 und 2 „transformiert" wurden.



Der 3. Schritt:

Pearson schlug vor bei allen Punkten die x- mit der y-Koordinate zu multiplizieren und den Mittelwert all dieser Produkte zu bilden. Diesen Mittelwert nennt man **Korrelationskoeffizient**.

Wieso misst der Korrelationskoeffizient den linearen Zusammenhang?

Die folgende Tabelle zeigt den 3. Schritt noch einmal rechnerisch:

2.Schritt

Kurswert (KW)	Dax	(KW-MW)/SA	(DAX-MW)/SA	Produkte
20	4300	**-1,12**	**0,00**	0,00
25	4325	**-0,19**	**1,22**	-0,23
33	4275	**1,31**	**-1,22**	-1,60
				Mittelwert
			Korrelationskoeffizient: r =	-0,61

Um das Berechnen des Korrelationskoeffizienten r zu üben, hier ein paar einfache Wertetabellen:

1.	x	y	2.	x	y	3.	x	y	4.	x	y	5.	x	y	6.	x	y	7.	x	y
	1	1		1	2		1	3		1	4		1	5		1	6		1	7
	2	3		2	4		2	5		2	1		2	3		2	4		2	5
	3	5		3	3		3	3		3	4		3	5		3	2		3	3
	4	7		4	6		4	5		4	3		4	3		4	3		4	1

g) Berechnen Sie jeweils r. Stellen sie die Daten auch graphisch dar. Für welches r würden Sie

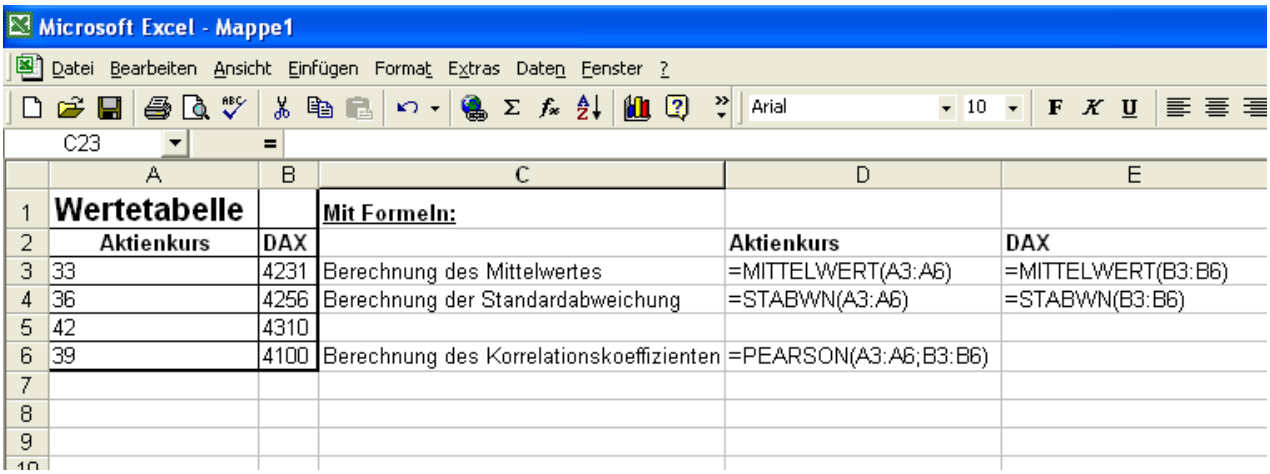
- keinen
- einen schwachen
- einen mittleren
- einen starken und
- einen perfekten

linearen Zusammenhang vermuten?

h) Erstellen sie (sinnvollerweise mit Hilfe von Excel) Wertetabellen, so dass $r = \pm 1$; $r \approx \pm 0{,}75$; $r \approx \pm 0{,}5$; $r \approx \pm 0{,}25$; $r \approx 0$; $r \approx \pm 2$. Was fällt auf?

i) Zurück zu Seite 1: Berechnen sie r für die Deutsche Bank Aktie und den DAX. Lässt sich näherungsweise ein linearer Zusammenhang vermuten?

Hier noch ein paar Hilfen für den Einsatz von Excel zur Berechnung der Lösungen (denn alle Aufgaben handschriftlich zu berechnen ist langweilig, wenn man die Idee des Korrelationskoeffizienten verstanden hat):

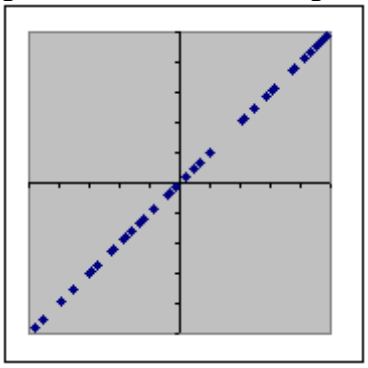


	A	B	C	D	E
1	**Wertetabelle**		**Mit Formeln:**		
2	**Aktienkurs**	**DAX**		**Aktienkurs**	**DAX**
3	33	4231	Berechnung des Mittelwertes	=MITTELWERT(A3:A6)	=MITTELWERT(B3:B6)
4	36	4256	Berechnung der Standardabweichung	=STABWN(A3:A6)	=STABWN(B3:B6)
5	42	4310			
6	39	4100	Berechnung des Korrelationskoeffizienten	=PEARSON(A3:A6;B3:B6)	
7					
8					
9					

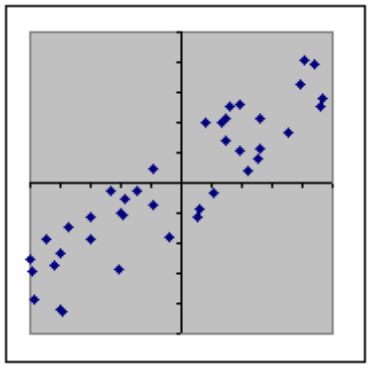
Die Berechnung des Korrelationskoeffizienten durch die Formel „=Pearson(...)" würdigt auf diese Weise sicher noch einmal die Leistung von Karl Pearson. Denn welcher weitere Nachname hat es schon als „Formel" in die Formelsammlung von Excel geschafft?

Korrelationen sehen!

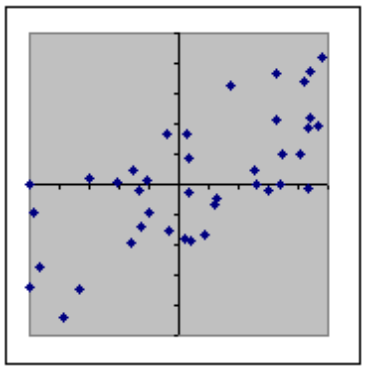
Der Korrelationskoeffizient ist nun nichts anderes, als eine Zahl, die beschreibt, wie sehr die „Punktwolke" (oder auch eine Gerade, die man sich passend durch die Punktwolke gezeichnet denkt) steigt oder fällt:



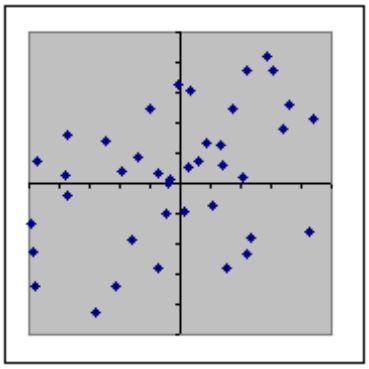
Korrelationskoeffizient =



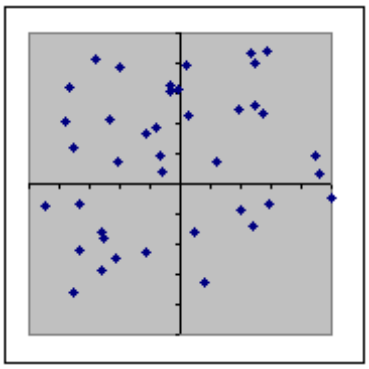
Korrelationskoeffizient =



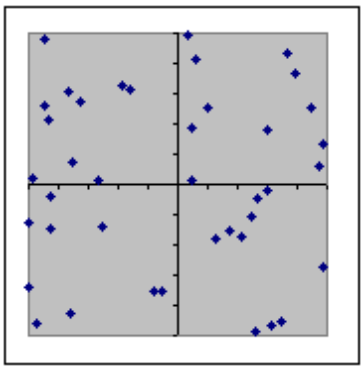
Korrelationskoeffizient =



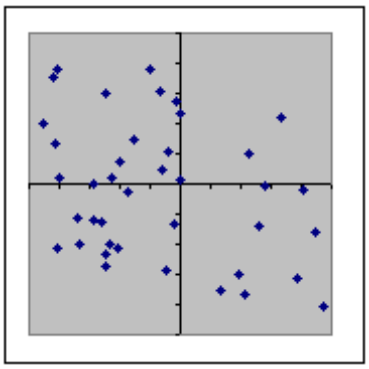
Korrelationskoeffizient =



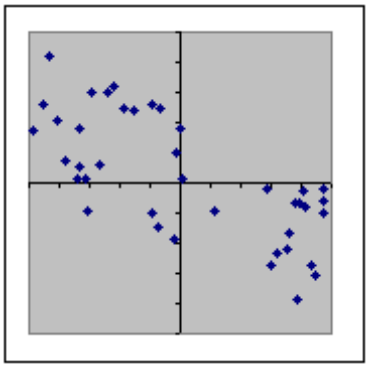
Korrelationskoeffizient =



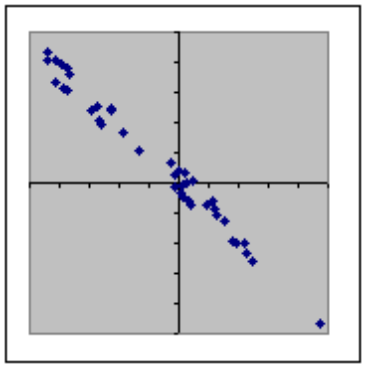
Korrelationskoeffizient =



Korrelationskoeffizient =



Korrelationskoeffizient =



Korrelationskoeffizient =

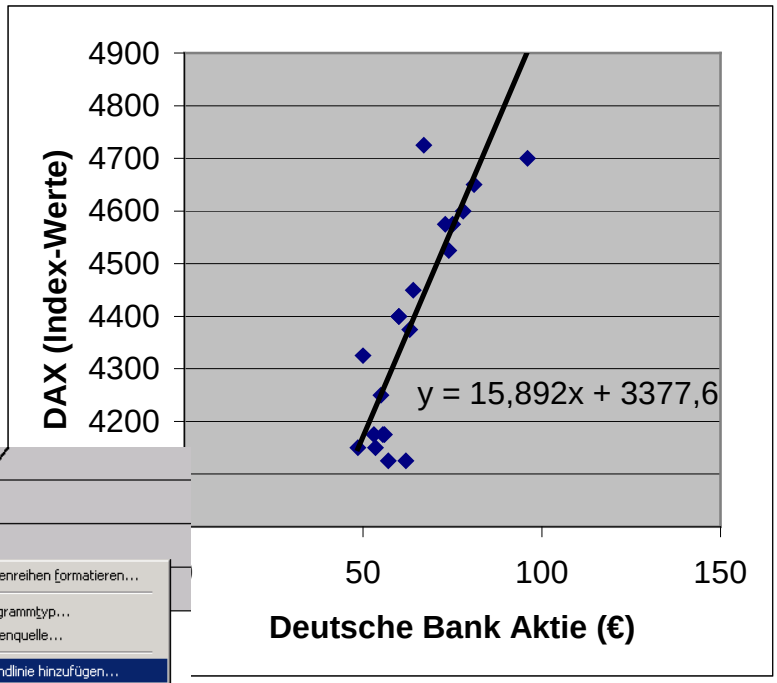
j) Welche Korrelationskoeffizienten schätzen Sie?

Wenn man nachvollzogen hat, wie ein Korrelationskoeffizient zustande kommt, fällt es leichter zu verstehen, was er bedeutet: Der Korrelationskoeffizient misst, wie stark zwei Merkmale **linear zusammenhängen**. Ein solcher linearer Zusammenhang lässt sich grafisch am besten durch eine Gerade darstellen. Ein Korrelationskoeffizient von 0.8 bedeutet also:

> *„Wenn die transformierten Daten zum Gewicht um eine bestimmten Wert größer werden, dann werden die z-transformierten Daten zur Größe **voraussichtlich** um das 0,8-fache dieses Wertes größer. Dieser ‚bestimmte Wert' wird also in dem oben beschriebenen quadratischen Koordinatenkreuz gemessen, z. B. muss man, wenn man 2cm nach rechts geht, 80% ·2 cm, also 1,6 cm nach oben gehen."*

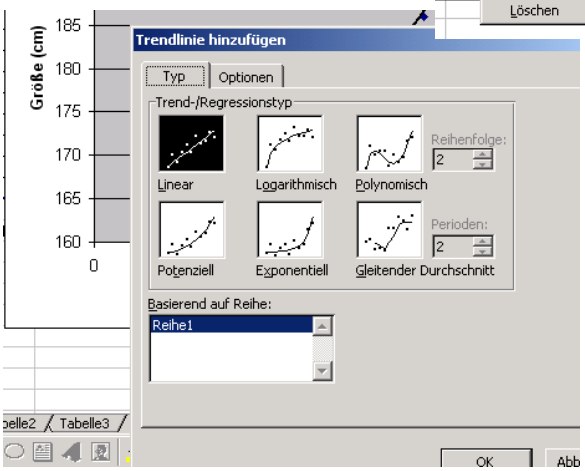
Ermittlung einer Trendgerade

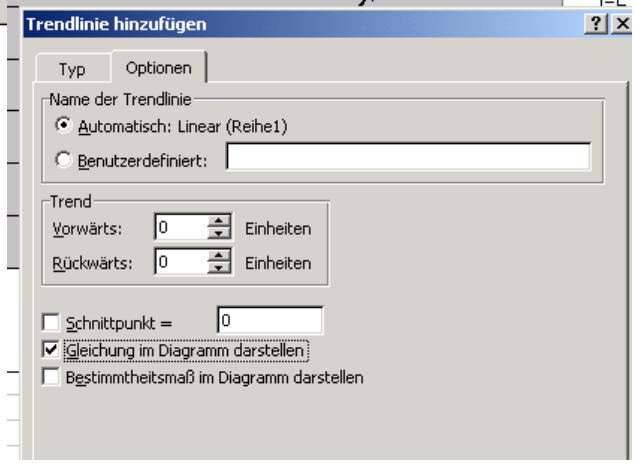
Mit Hilfe des Korrelationskoeffizienten überprüft man zunächst, ob ein linearer Zusammenhang vorliegt. Ist dies näherungsweise der fall, so stellt sich natürlich die Frage, wie dieser lineare Zusammenhang in Form einer Gerade(ngleichung) aussehen kann.



Wenn man die Daten mit Excel graphisch darstellt und einen Datenpunkt mit der rechten Maustaste anklickt, so erscheint das rechts abgebildete Fenster.

Klickt man „Trendlinie hinzufügen, so erscheint:





Hier wählt man den Trend-/Regressionstyp „Linear" aus (neben den vielen anderen, die es sonst noch gibt) und anschließend auf den Reiter „Optionen". Dort klickt man „Gleichung im Diagramm darstellen" an und bestätigt mit Ok.

Doch wie kommt die Gleichung der Trendgeraden zustande? Die Idee lässt sich folgendermaßen beschreiben:

1. Idee: Man sucht einen Datenpunkt der möglichst in der Mitte der Punktwolke liegt. Hierfür gibt es praktisch keinen besseren als $\left(\overline{x} \,/\, \overline{y}\right)$.

2. Idee: Neben einem Punkt braucht man für die Gerade noch eine Steigung. Diese ermittelt man mit Hilfe des Korrelationskoeffizienten:

Der Korrelationskoeffizient r gibt die „Steigung" der transformierten Daten an. Wenn man nun wüsste, um das wie vielfache die original y-Werte (DAX) stärker oder schwächer streuen, als die original x-Werte (Deutsche Bank), so könnte man r mit diesem Faktor vergrößern bzw. verkleinern und würde so die „Steigung" der original Daten erhalten.

Aufgaben:

k) Berechnen Sie die Steigung der Trendgerade zu den DAX-Deutsche Bank-Daten (Vergleichen Sie Ihr Ergebnis mit der Steigung 15,892 der Geradengleichung im obigen Diagramm).

l) Erklären Sie, woher die +3377,6 der Geradengleichung im obigen Diagramm kommen.