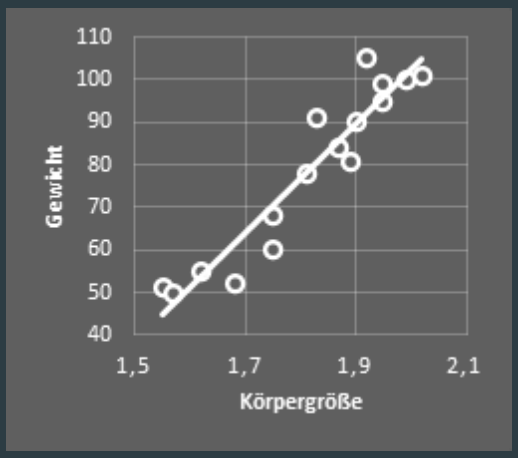


Lineare Regression

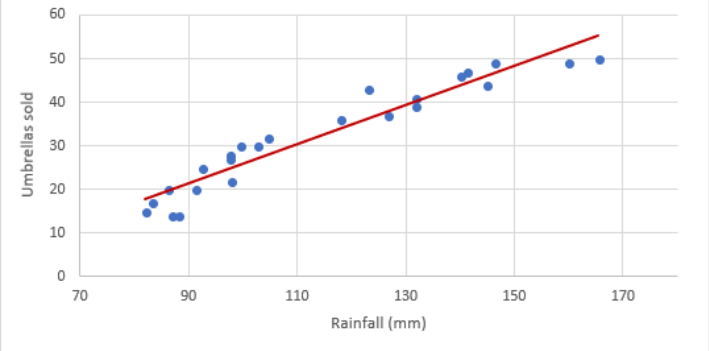
Uhlmann Rudolf

Inhalt

Lineare Regression

Einführungsbeispiel

Wenn es regnet geht das Geschäft mit den Regenschirmen gut. Nehmen wir als Beispiel die Verkaufszahlen für Regenschirme der letzten 24 Monate und betrachten den durchschnittlichen monatlichen Niederschlag für denselben Zeitraum.

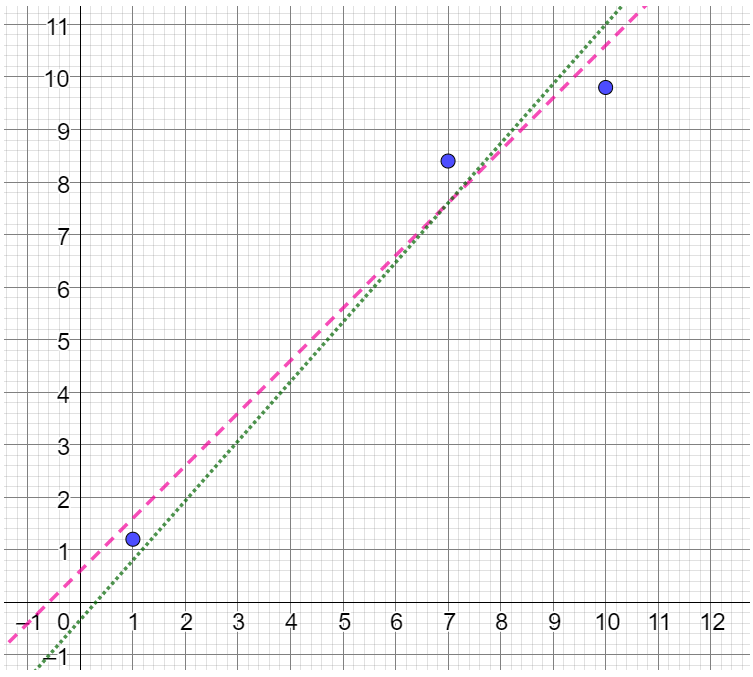


https://www.ablebits.com/office-addins-blog/2018/08/01/linear-regression-analysis-excel/

Das Diagramm zeigt eine annähernd lineare Beziehung zwischen den Niederschlagsmengen und den Regenschirm-Verkaufszahlen. Die rote Linie ist die sogenannte Trendlinie oder Ausgleichsgerade, die den linearen Verlauf im Mittel anzeigt.
Die „lineare Regression“ ist ein statistisches Verfahren, um in so einem Fall die bestmögliche lineare Funktion zu finden.

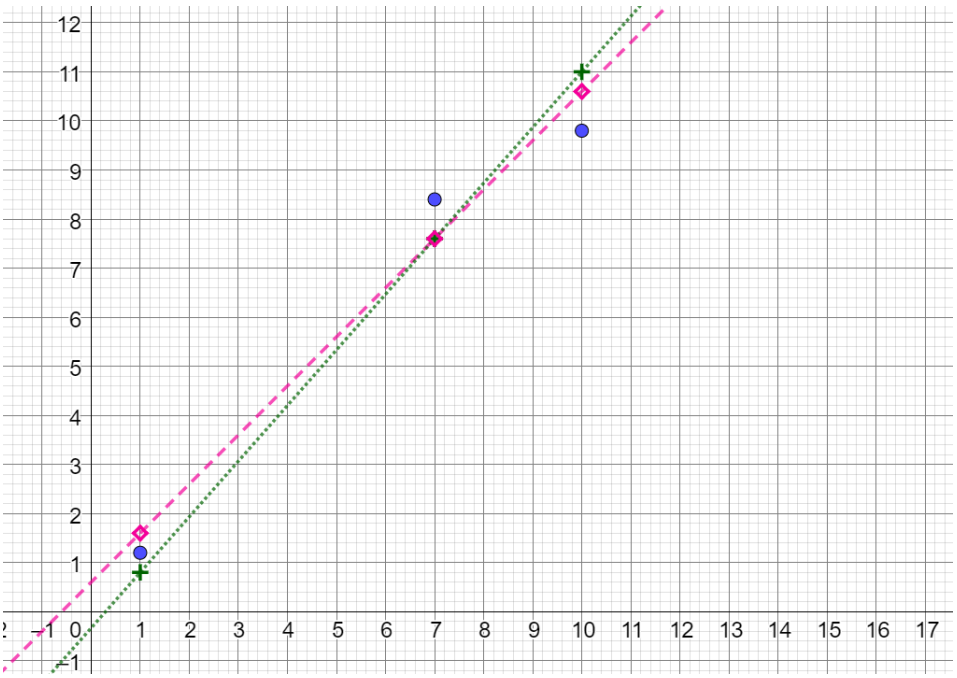
Beispiel: Betrachten wir ein einfaches Beispiel mit drei Datenpunkten:

x	y
1	1,2
7	8,4
10	9,9



Welche der beiden Ausgleichsgeraden repräsentiert den linearen Trend besser?

Betrachten wir die Abweichungen an den gegebenen Stellen $x_1 = 1$, $x_2 = 7$ und $x_3 = 10$ und dazu die sogenannten Residuen $r_i = y_i - \hat{y}_i$.



Die punktierte Linie:

x	y	$\hat{y}$	$y - \hat{y}$
1	1,2	0,8	0,4
7	8,4	7,6	0,8
10	9,9	11,0	-1,1

Die strichlierte Linie:

x	y	$\hat{y}$	$y - \hat{y}$
1	1,2	1,6	-0,4
7	8,4	7,6	0,8
10	9,9	10,6	-0,7

Die hier angewandte „METHODE DER KLEINSTEN QUADRATE“ stammt von C.F. Gauß. Die Methode liefert die Parameter k und d einer passenden Ausgleichsgeraden $y = k \cdot x + d$, indem die Summe der Abstandsquadrate minimiert wird:

$$Q(k,d) = \sum_i (y_i - \hat{y}_i)^2 \to min$$

Die punktierte Linie:

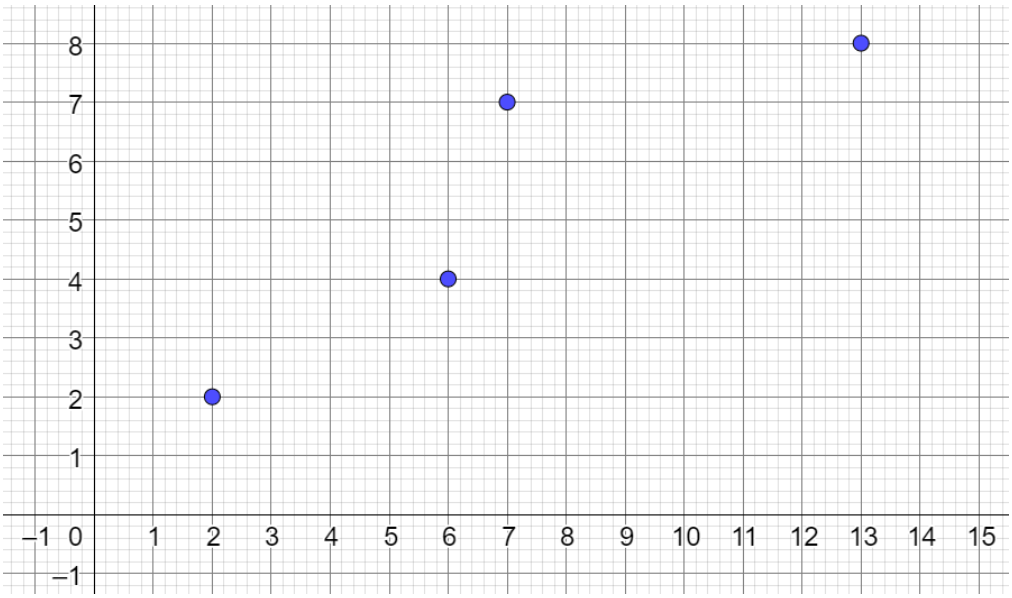
$y - \hat{y}$	$(y - \hat{y})^2$	
0,4	0,16	$\sum_i = 2{,}01$
0,8	0,64	
-1,1	1,21	

Die strichlierte Linie:

$y - \hat{y}$	$(y - \hat{y})^2$	
-0,4	0,16	$\sum_i = 1{,}29$
0,8	0,64	
-0,7	0,49	

Somit ist die strichlierte Linie eine besser passende Trendlinie (Ausgleichsgerade)! Sie ist aber trotzdem nicht die optimale.

Aufgabe: Zeichne mit Augenmaß eine passende Ausgleichsgerade ein. Berechne dann die Summe der Abstandsquadrate und vergleiche mit deinen Mitschülern!



Die Regressionsgerade

Die „Methode der kleinsten Quadrate“ bestimmt jene Geradengleichung

$$\hat{y} = k \cdot x + d$$

mit

$$Q(k,d) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - kx_i - d)^2 \rightarrow min$$

Dies gelingt, indem $Q(k,d)$ sowohl bezüglich k als auch d optimiert wird. Wir verwenden dazu die Differenzialrechnung mit den beiden Bedingungen $\frac{\partial Q}{\partial k} = 0$ und $\frac{\partial Q}{\partial d} = 0$!

Der optimale Wert für d

$$\frac{\partial Q}{\partial d} = \frac{\partial}{\partial d}\sum_i (y_i - kx_i - d)^2 = [Kettenregel] = \sum 2(y_i - kx_i - d)(-1) = \sum 2(kx_i + d - y_i) =$$

$$= \sum 2(kx_i - y_i) + \sum 2d = 2\sum kx_i - 2\sum y_i + n2d = 0$$

Auflösen nach d:

$$\rightarrow d_{opt} = \frac{-\sum kx_i + \sum y_i}{n} = \frac{-k\sum x_i + \sum y_i}{n} = -k\frac{\sum x_i}{n} + \frac{\sum y_i}{n} = -k\bar{x} + \bar{y}$$

wobei $\bar{x} = \frac{\sum x_i}{n}$ der Mittelwert aller x-Werte und $\bar{y} = \frac{\sum y_i}{n}$ der Mittelwert aller y-Werte ist. Somit ist der optimale Wert von d

$$d_{opt} = \frac{\sum y_i}{n} - k\frac{\sum x_i}{n}$$

Der optimale Wert für k

$$\frac{\partial Q}{\partial k} = \frac{\partial}{\partial k}\sum_i (y_i - kx_i - d)^2 = \sum 2(y_i - kx_i - d)(-x_i) = \sum 2(-y_i x_i - k{x_i}^2 - dx_i)$$

$$= 2k\sum {x_i}^2 + 2d\sum x_i - 2\sum y_i x_i = 0$$

Auflösen nach k:

$$2k\sum {x_i}^2 + 2(\bar{y} - k\bar{x})\sum x_i - 2\sum y_i x_i = 0$$

$$k\sum {x_i}^2 - k\bar{x}\sum x_i + \bar{y}\sum x_i = \sum y_i x_i$$

$$kn\sum {x_i}^2 - kn\bar{x}\sum x_i + n\bar{y}\sum x_i = n\sum y_i x_i$$

$$kn\sum {x_i}^2 - k\sum x_i \sum x_i + \sum y_i \sum x_i = n\sum y_i x_i$$

$$k_{opt} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - (\sum x_i)^2} = \frac{n\sum x_i y_i - n^2 \frac{\sum x_i}{n}\frac{\sum y_i}{n}}{n\sum {x_i}^2 - n^2\left(\frac{\sum x_i}{n}\right)^2} = \frac{\sum x_i y_i - n\frac{\sum x_i}{n}\frac{\sum y_i}{n}}{\sum {x_i}^2 - n\left(\frac{\sum x_i}{n}\right)^2}$$

$$k_{opt} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum {x_i}^2 - n\bar{x}^2}$$

Bei Vorgabe $\mathrm{d} = 0$***:***

Ist eine Ausgleichsgerade gewünscht, die definitiv durch den Ursprung verläuft, so ergibt sich für k_{opt}

$$\frac{\partial Q}{\partial k} = 2k\sum {x_i}^2 + 2 \cdot d(=0) \cdot \sum x_i - 2\sum y_i x_i = 0$$

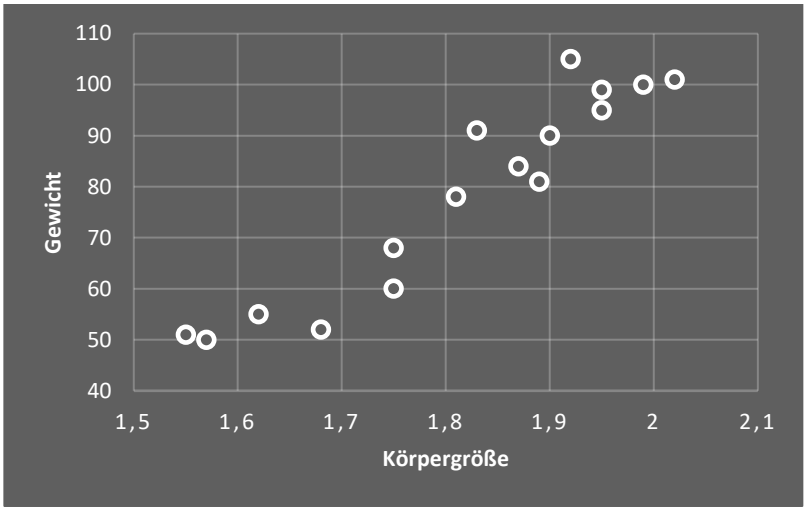
$$2k\sum {x_i}^2 - 2\sum y_i x_i = 0$$

$$k_{opt} = \frac{\sum x_i y_i}{\sum {x_i}^2}$$

Lineare Regression in Excel

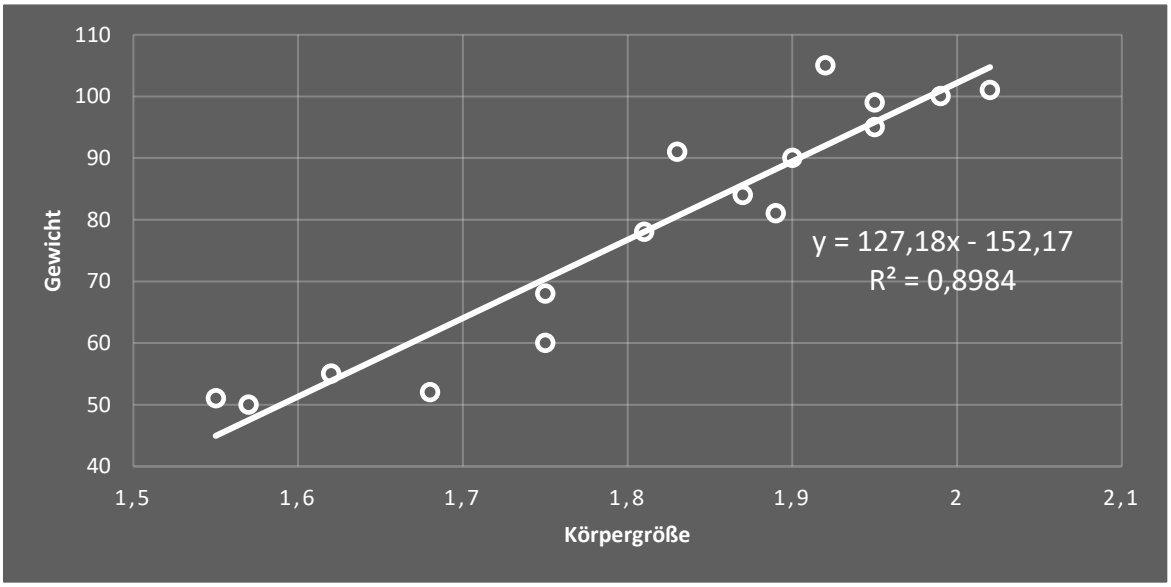
Beispiel: In einer Gruppe von 16 Personen wurde jeweils die Körpergröße und dass Gewicht gemessen.

Körpergrösse in m	Gewicht in kg
1,55	51
1,57	50
1,62	55
1,68	52
1,75	60
1,75	68
1,81	78
1,83	91
1,87	84
1,89	81
1,90	90
1,92	105
1,95	95
1,95	99
1,99	100
2,02	101



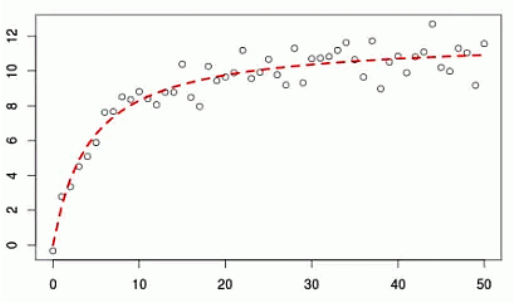
Die Tabelle in Excel übertragen und als Punktdiagramm dargestellt. Beachte, der Nullpunkt ist nicht im Diagramm abgebildet.

Mit einem Rechtsklick auf einen Datenpunkt wählt man aus dem Kontextmenü den Punkt „Trendlinie“. Es wird automatisch die lineare Trendlinie berechnet. Zusätzlich lassen sich die Gleichung und das Bestimmtheitsmaß R^2 anzeigen. Bei perfekter Übereinstimmung, wenn also alle Datenpunkte ohnehin auf einer geraden Linie liegen, ist $R^2 = 1$.



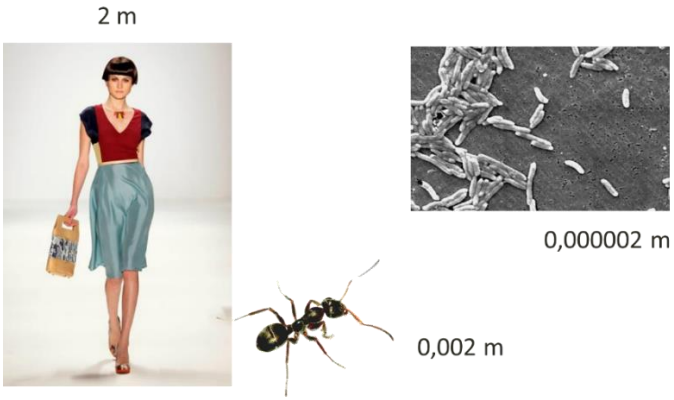
Regression nichtlinearer Zusammenhänge

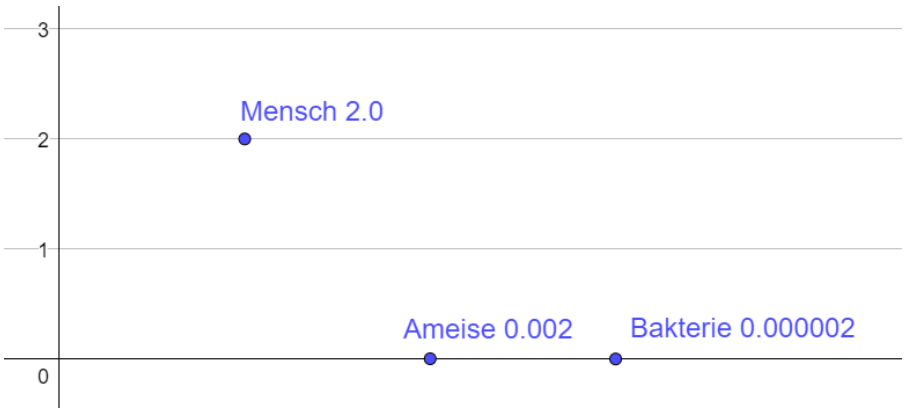
Um die lineare Regression auch auf andere Funktionstypen anwenden zu können, wird die Technik der „Linearisierung“ benötigt.



Linearisierung mittels logarithmischer Darstellung

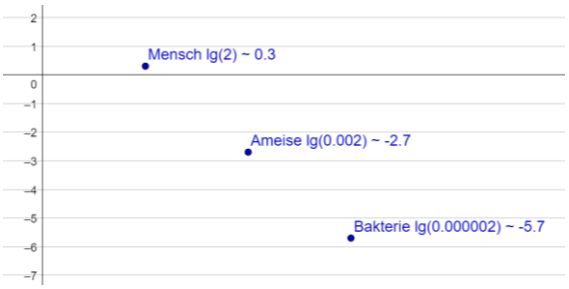
Die logarithmische Darstellung von Daten wird meist dann angewandt, wenn sehr kleine und gleichzeitig sehr große Zahlen dargestellt werden sollen.

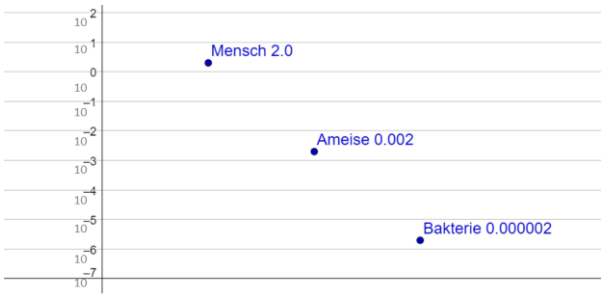




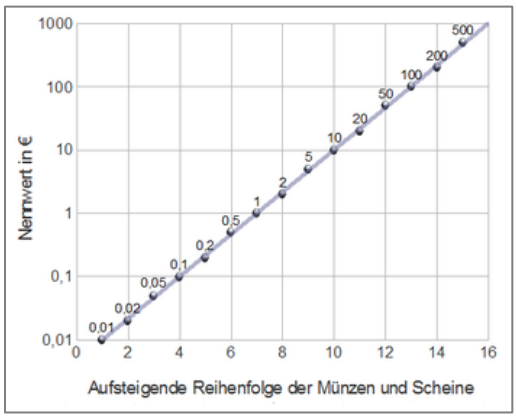
Man hat das Problem, dass entweder der kleine Bereich nicht mehr aufgelöst werden kann oder die großen Werte durch die Decke gehen. Der Trick besteht jetzt darin, dass man anstatt der darzustellenden Werte die logarithmierten Werte darstellt (normalerweise der dekadische Logarithmus). 1000 ist dann nur 3 und $0{,}000001$ ist -6, was betragsmäßig nicht wenig ist. Große Werte werden dadurch also betragsmäßig kleiner und kleine Werte größer.

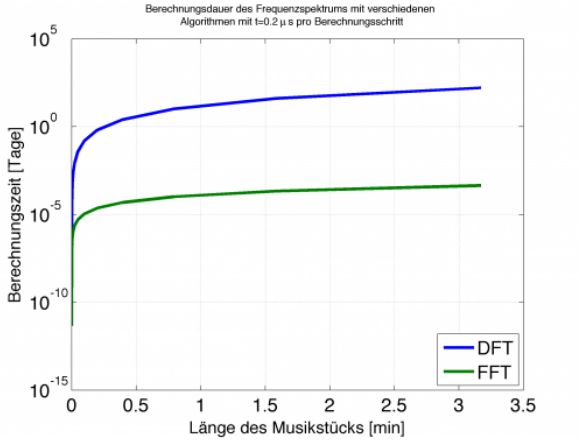
Man sieht auch schon, dass eine gewisse Linearisierung stattgefunden hat.





Im linken Diagramm sind die logarithmierten Werte pur dargestellt, im rechten Diagramm sind die dargestellten Logarithmen mit den originalen Werten beschriftet.
Hier noch weitere Beispiele logarithmischer Darstellungen:

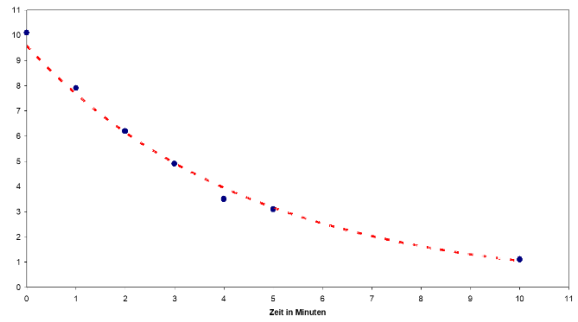




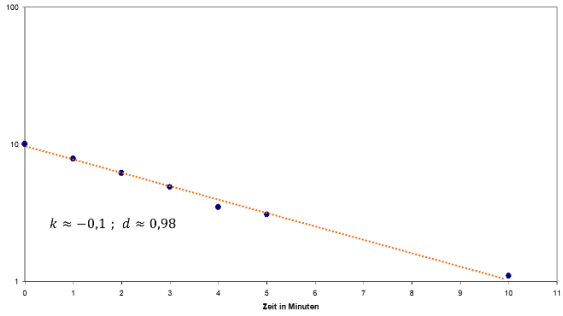
Die Exponentialfunktion in einfach logarithmischer Darstellung

In den Grafiken wird der Zerfall von Bierschaum untersucht. Links das Ergebnis in üblicher Darstellung, rechts in logarithmischer. Die logarithmische Darstellung hat zu einer Linearisierung geführt.

Schaumhöhe in cm



Schaumhöhe in cm



Satz: Der Graph einer Exponentialfunktion wird in der (einfach) logarithmischen Darstellung linear.

Beweis: Sei $y(x) = c \cdot a^x$. Das Logarithmieren beider Seiten führt zu

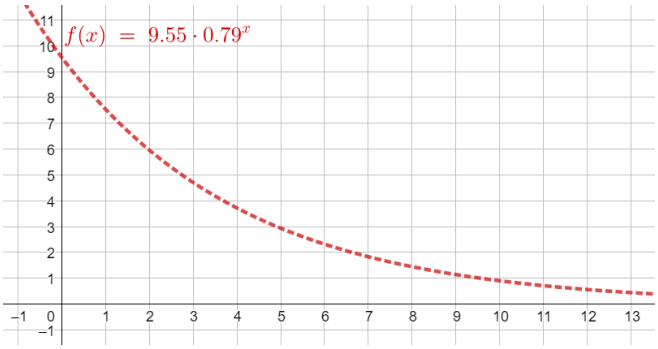
$$\lg\big(y(x)\big) = \lg(c \cdot a^x)$$

$$\lg\big(y(x)\big) = \lg(c) + \lg(a^x)$$

$$\lg\big(y(x)\big) = \lg(a) \cdot x + \lg(c) \ldots \text{ linear zu } x \ \blacksquare$$

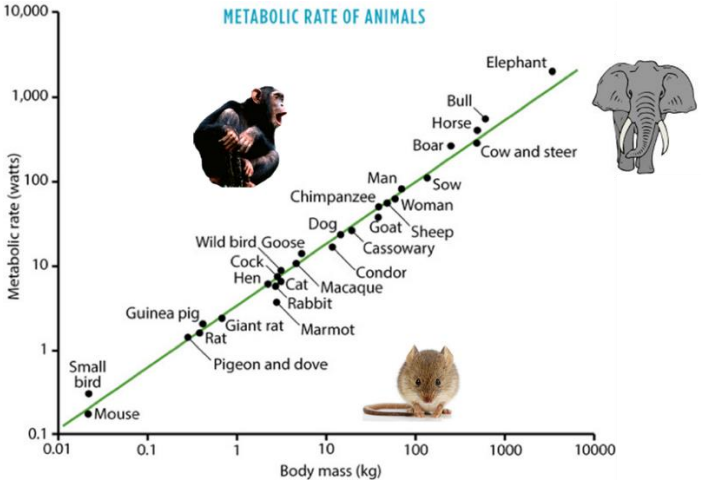
Die in der Grafik dargestellten Werte $k \approx -0{,}1$ und $d \approx 0{,}98$ beziehen sich auf die logarithmierten Werte. $d = 0{,}98$ bedeutet $c = 10^{0{,}98} = 9{,}55$. Der Schnittpunkt mit der Ordinate ist original der Wert c. Der Anstieg $k = -0{,}1$ bedeutet $a = 10^{-0{,}1} \approx 0{,}79$.

Die Funktionsgleichung der Trendlinie (Ausgleichskurve) lautet somit $y(x) \approx 9{,}55 \cdot 0{,}79^x$



Die doppelt logarithmische Darstellung

Von einer doppelt logarithmischen Darstellung spricht man, wenn sowohl x- als auch y-Achse logarithmisch skaliert sind.



Wird eine Korrelation in der doppelt logarithmischen Darstellung linear, so handelt es sich original um eine Potenzfunktion.

Satz: Der Graph einer Potenzfunktion wird in der doppelt logarithmischen Darstellung linear.

Beweis: Sei $y(x) = a \cdot x^r$. Das Logarithmieren beider Seiten führt zu

$$\lg\big(y(x)\big) = \lg(a \cdot x^r)$$

$$\lg\big(y(x)\big) = \lg(a) + r \cdot \lg(x)$$

$$\lg\big(y(x)\big) = r \cdot \lg(x) + \lg(a) \ldots \text{ linear zu } \lg(x) \quad \blacksquare$$

Zusammenfassung:

Der Graph einer Exponentialfunktion wird in der einfach logarithmischen Darstellung linear!
Der Graph einer Potenzfunktion wird in der doppelt logarithmischen Darstellung linear!

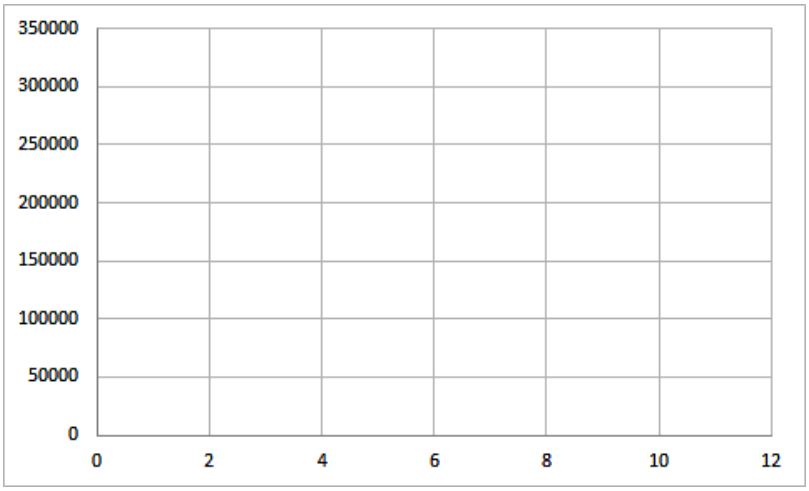
Übungen

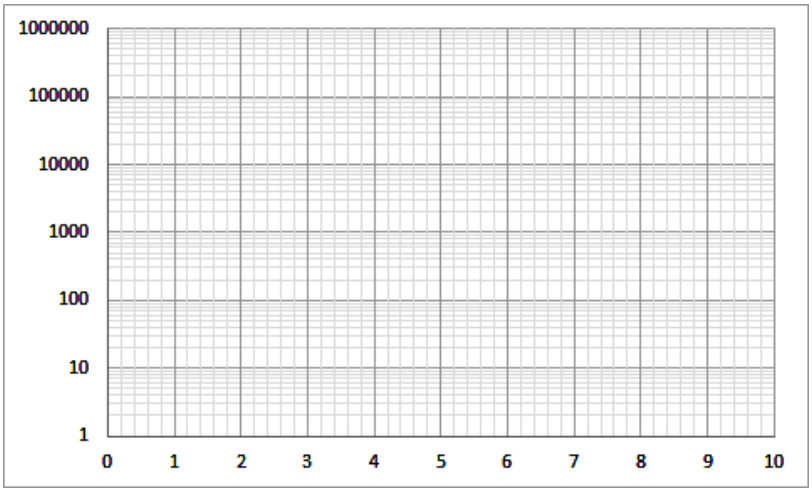
Aufgabe 1

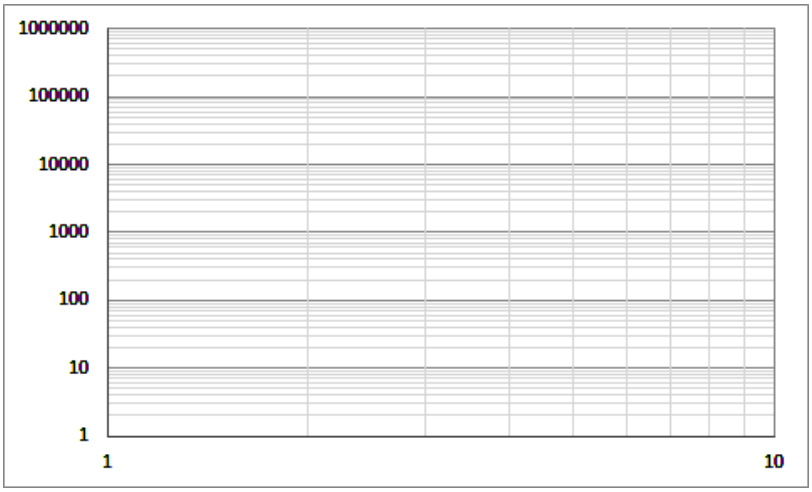
Angabe: Gegeben sind zehn Daten (x, y). Zeichne diese so genau es geht aber freihändig in die Diagramme ein!

Aufgabenstellung: Entscheide auf Grund der graphischen Ergebnisse, um welchen Funktionstyp es sich hierbei handelt.

x	$y(x)$
0	5
1	15
2	45
3	135
4	405
5	1215
6	3645
7	10935
8	32805
9	98415
10	295245





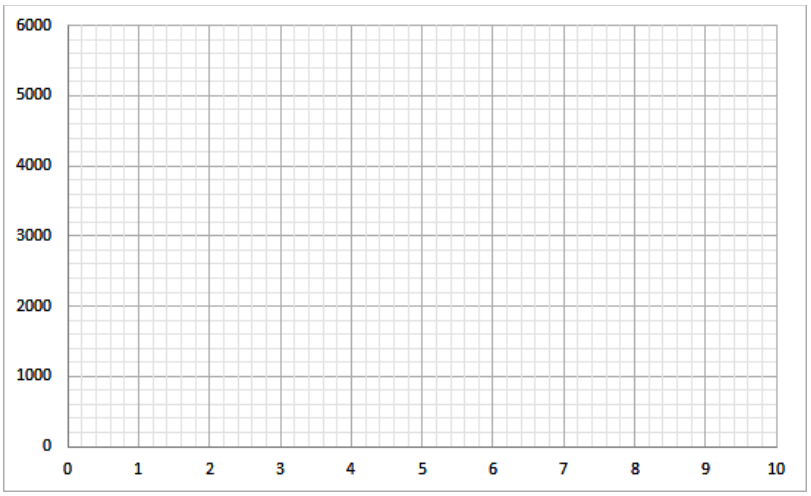


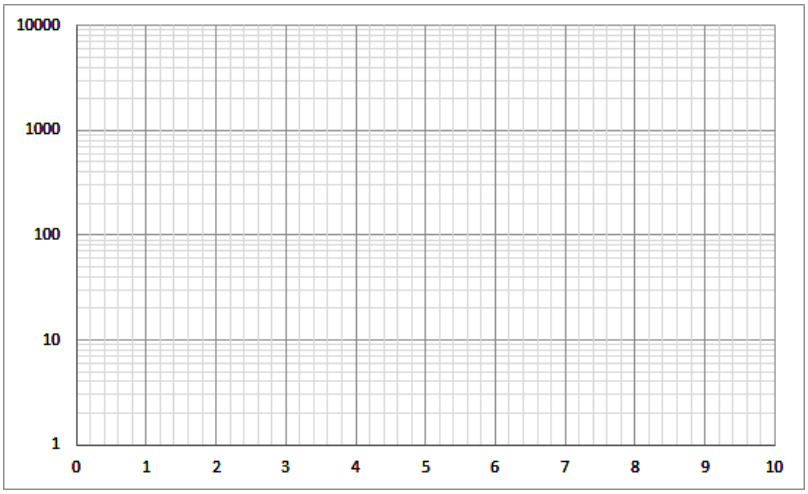
Aufgabe 2

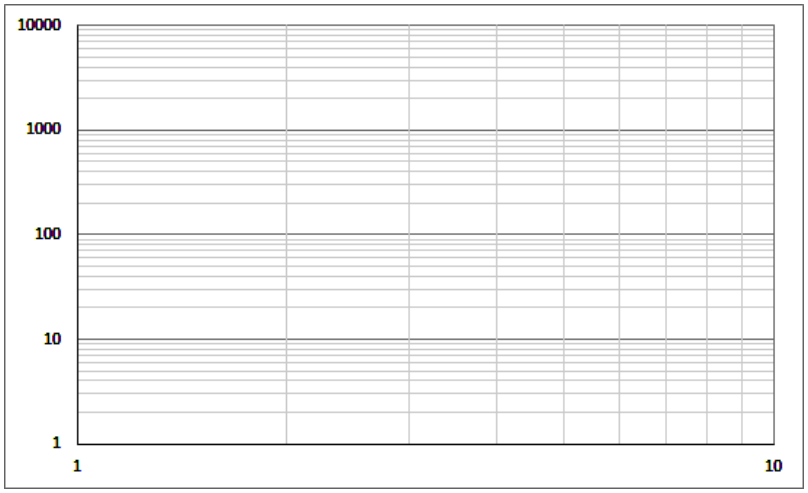
Angabe: Gegeben sind zehn Daten (x, y). Zeichne diese so genau es geht aber freihändig in die Diagramme ein!

Aufgabenstellung: Entscheide auf Grund der graphischen Ergebnisse, um welchen Funktionstyp es sich hierbei handelt.

x	$y(x)$
0	0
1	5
2	40
3	135
4	320
5	625
6	1080
7	1715
8	2560
9	3645
10	5000



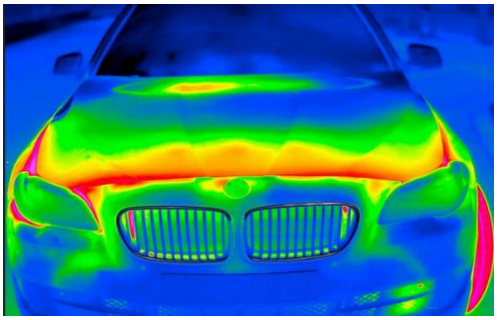




Nichtlineare Regression (Trendlinie) in Excel

Beispiel:

Das Kühlwasser der Motorkühlung hat nach der Fahrt eine Temperatur von 80°C, das sind 60°C über der Temperatur in der Garage. Die unten angegebenen Daten geben den Abkühlungsprozess wieder.

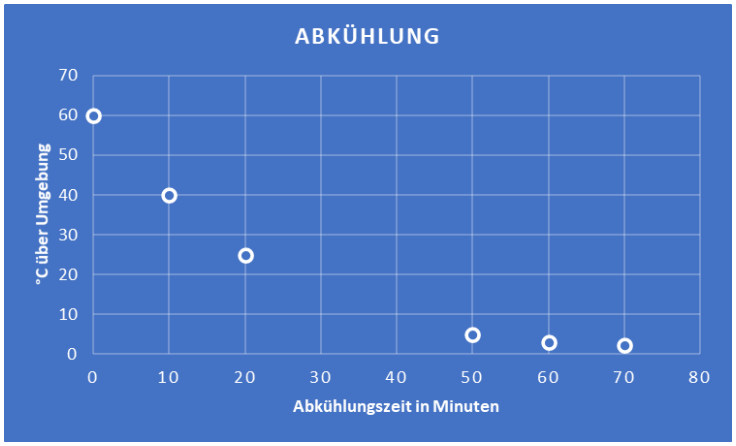


a) Stelle die Daten graphisch dar und bestimme eine geeignete Regressionsfunktion!
b) Interpoliere mit Hilfe der Regressionsfunktion Temperaturwerte für die Zeiten 30 min und 40 min!

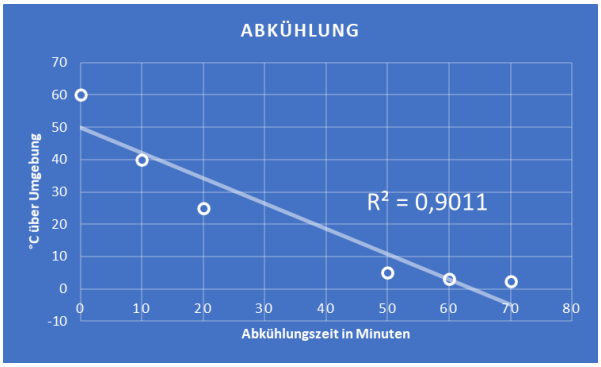
Lösung:

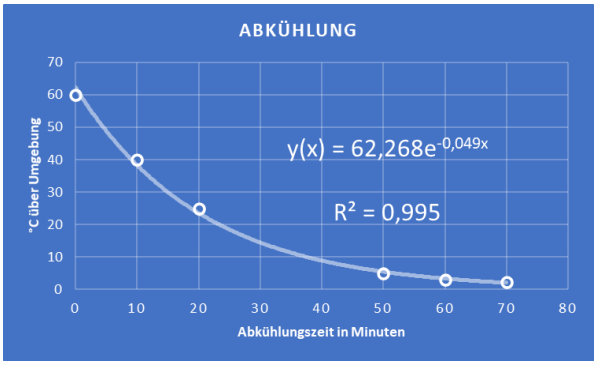
a) Die Tabelle in Excel übertragen und als Punktdiagramm darstellen. Mit einem Rechtsklick auf einen der Datenpunkte im Kontextmenü die „Trendlinie“ öffnen.

Zeit in Minuten	Temperatur-unterschied in °C
0	60
10	40
20	25
50	5
60	3
70	2



Die lineare Trendlinie ist wie immer vor ausgewählt. Wir sehen aber gleich, sie passt nicht gut. Das Bestimmtheitsmaß ist mit nur etwa 0,9011 auch nicht berauschend. Wir vermuten ja ohnehin eher einen exponentiellen Abfall. Wählen wir unter den Trendlinienoptionen „Exponentiell“, so sieht die Sache schon wesentlich besser aus.



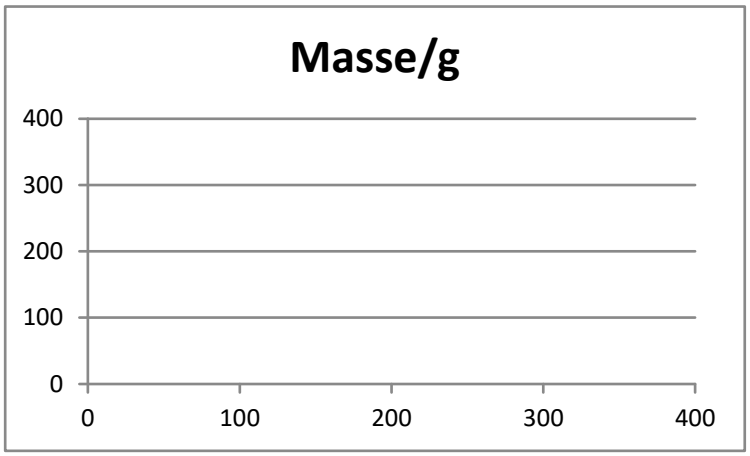


b) $y(30) = 62{,}268 \cdot e^{-0{,}049 \cdot 30} = 14{,}3$°C über Raumtemperatur, also 34,3°C Kühlwassertemperatur.
$y(40) = \cdots$

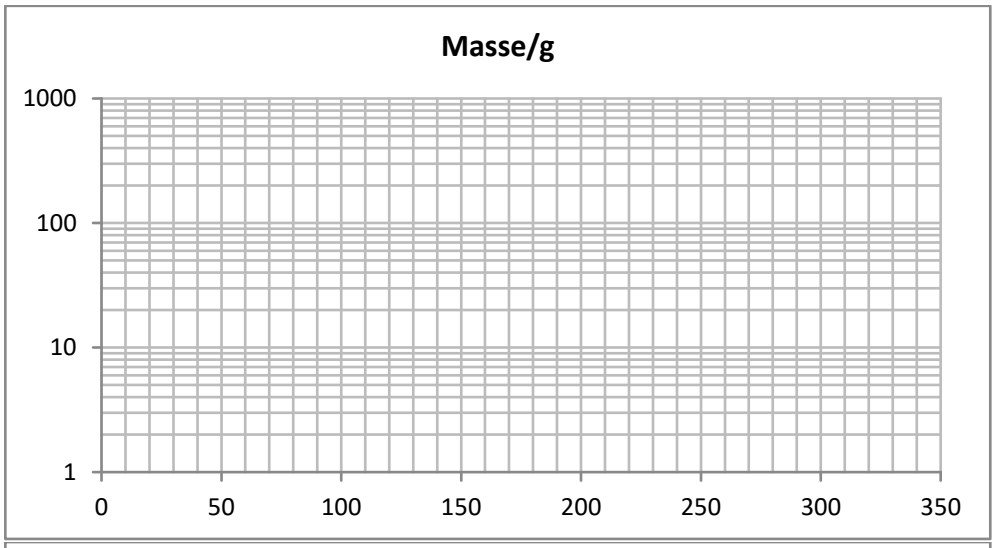
Weitere Übungen

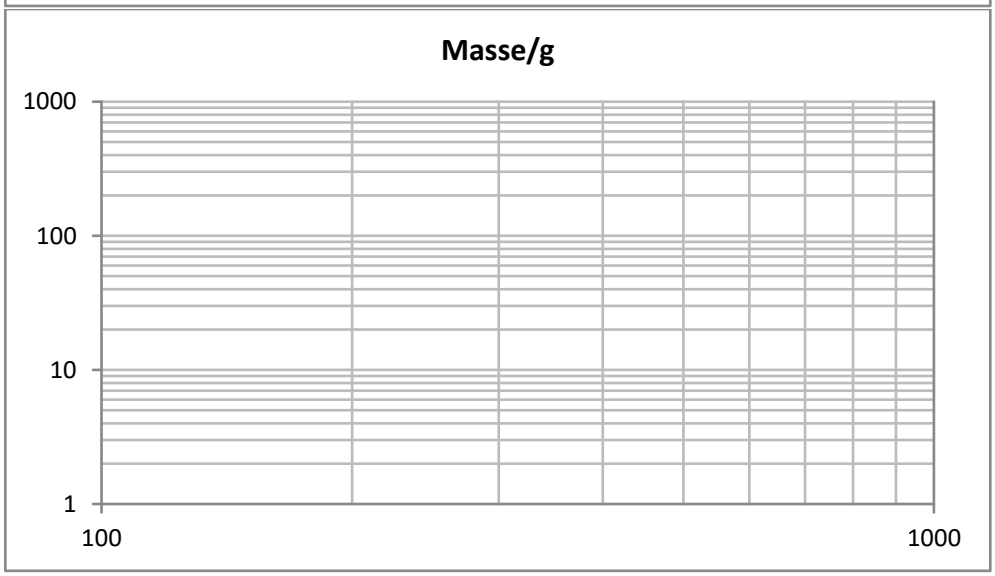
Die Bachforelle 

Länge/cm	Masse/g
140	31
160	45
180	52
200	79
220	122
240	154
260	184
280	210
300	263
320	360



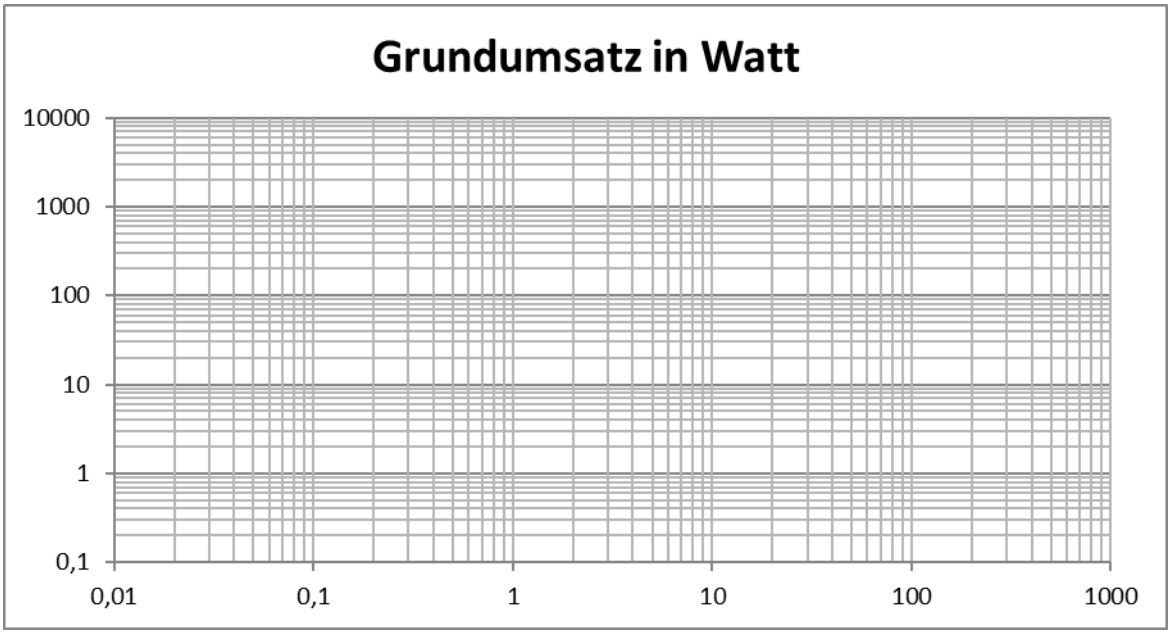
Aufgabe: Übertrage die Daten in obiges Diagramm ein. Der Graph legt einen Zusammenhang gemäß einer Potenzfunktion nahe. Verwende dafür das dazu passende logarithmische Diagramm aus und bestimme die Funktionsgleichung!





Der Grundumsatz von Tieren

	Masse in kg	Grundumsatz in Watt		Masse in kg	Grundumsatz in Watt
Mouse	0,02	0,16	**Sheep**	52	59
Rat	0,44	2,04	**Man**	70	73
Guinea Pig	0,45	1,54	**Pig**	144	114
Cat	3,4	6,5	**Boar**	265	258
Monkey	5	11,5	**Cow fem.**	525	275
Dog	16	24	**Horse**	525	400
Goat	38	52	**Bull**	676	548
Chimpanzee	44	39	**Elephant**	3806	2116

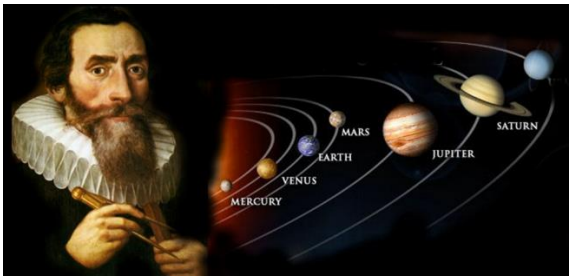


Der freie Fall

Ein senkrechter Sprung vom 10 m-Brett dauert etwa 1,4 Sekunden. Misst man alle 0,2 Sekunden die bereits durchfallenen Wegstrecke, so ergibt sich folgende Tabelle.

Zeit/s	Weg/m
0,00	0,00
0,20	0,21
0,40	0,80
0,60	1,75
0,80	3,08
1,00	4,78
1,20	6,86
1,40	9,29

Kepler III



1618 entdeckte Johannes Kepler, sein drittes Gesetz über die Planetenbahnen. Es stellt einen Zusammenhang zwischen der Umlaufzeit und der mittleren Entfernung zur Sonne her. Mit den heutigen Daten ergibt sich folgende Tabelle.

	Merkur	Venus	Erde	Mars	Jupiter	Saturn	Uranus	Neptun
Mittl. Abstand zur Sonne in Mio km	58	108	150	228		1427	2871	4498
Umlaufdauer in siderischen Jahren	0,24	0,62	1	1,88	11,86	29,45	84,02	164,79

Es ist wesentlich einfacher, die Umlaufzeit eines Planeten um die Sonne zu bestimmen, als seine Entfernung. Bei Jupiter fehlt der entsprechende Wert in der Tabelle. Ermittle ihn mit Hilfe einer geeigneten Trendlinie.

Lsg: 778