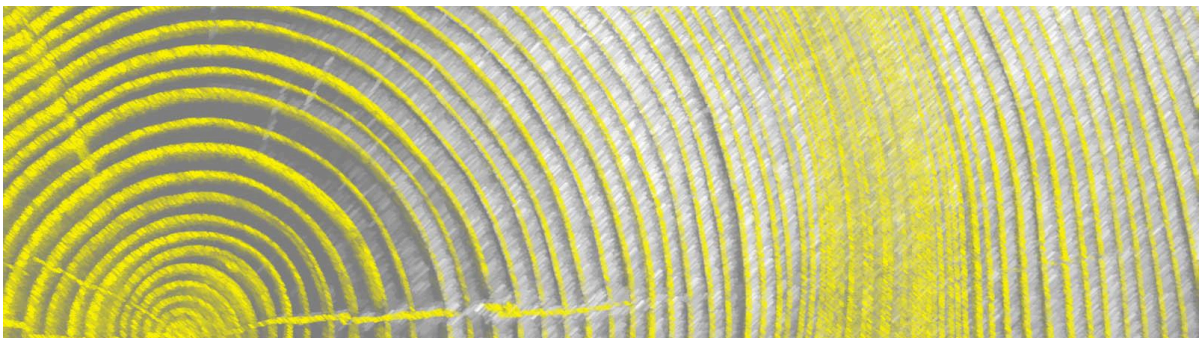
Beschreibende Statistik



Karl Entacher 

Demo Beispiele zur Beschreibenden Statistik

Bei der beschreibenden (deskriptiven) Statistik versucht man mit Hilfe unterschiedlicher Kennzahlen oder spezifischer grafischer Darstellung die experimentell oder empirisch gewonnenen Daten zu untersuchen bzw. deren Eigenschaften zu beschreiben.

Einfache Stichproben

Aufgabenstellung

Verwende eine Software deiner Wahl oder ein Blatt Papier, Stift und Taschenrechner und ermittle aus den gegebenen Beispieldaten die berechneten Lage- und Streuungsparameter sowie die nachfolgenden grafischen und tabellarischen Auswertungen.

Falls Nachhilfe benötigt wird, einfach die folgenden Demos ansehen: $Video_1$ (Mittelwert und Median), $Video_2$ (Quartile und Boxplot) und $Video_3$ (Standardabweichung) oder eine KI fragen ;-). Geogebra Files zur freien Verwendung: ggb_1, ggb_2, ggb_3.

Beispiel 1: Holzfeuchte

Im Rahmen einer Qualitätskontrolle wurde die Holzfeuchte von zwei Holzwerkstoffen unterschiedlicher Nutzungsklasse für jeweils $n = 25$ Proben getestet (ggb):

Produkt 1: Holzfeuchte in %				
10,1	10,5	10,3	9,2	8,4
12,0	11,3	14,9	12,4	9,4
11,7	11,6	11,3	11,5	10,6
9,3	11,7	10,0	12,7	10,3
10,5	8,4	11,4	8,3	12,2

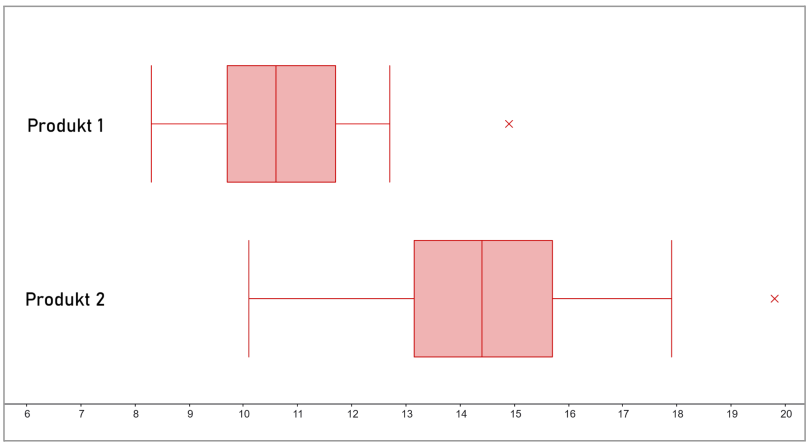
Produkt 2: Holzfeuchte in %				
15,5	10,1	14,4	14,5	13,6
13,4	11,5	17,9	11,9	15,7
12,8	17,4	14,2	14,1	15,7
15,7	16,6	14,1	13,5	15,1
12,9	19,8	15,6	12,8	16,1

Lage- und Streuungsparameter

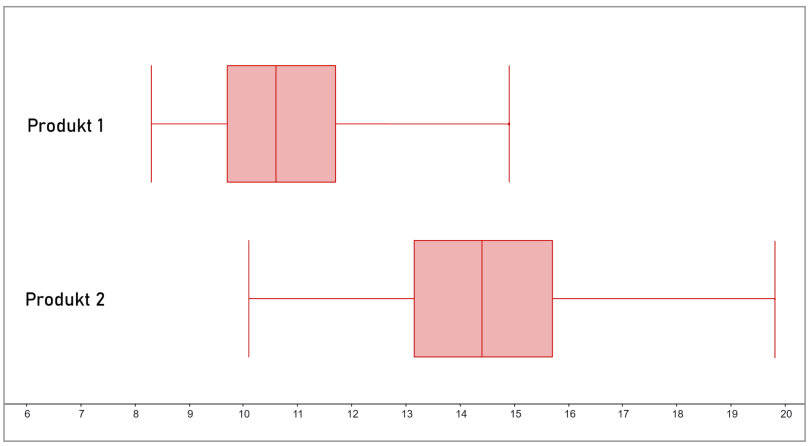
Diese Daten können durch unterschiedliche Parameter und verschiedene grafische Verfahren verglichen und untersucht werden. Hier eine Auswahl:

	Stichproben-größe *n*	Mittelwert	Median	Standard-abweichung
Produkt 1	25	10.8	10.6	1.53
Produkt 2	25	14.6	14.4	2.13
	Minimum	1. Quartile	3. Quartile	Maximum
Produkt 1	8.3	9.7	11.7	14.9
Produkt 2	10.1	13.15	15.7	19.8

Die beiden Datenreihen lassen sich mit Hilfe eines Boxplot vergleichen:



In dieser Version der Boxplots werden auch Ausreißer gekennzeichnet. Meistens wird eine einfachere Version verwendet, die keine Ausreißer markiert. Die Entfernung von Minimum zu Maximum nennt man Spannweite:



Quartile und Median "händisch"

Der Mittelwert (Durchschnitt, arithmetisches Mittel) lässt sich sehr einfach berechnen (Daten addieren und durch die Anzahl dividieren). Möchte man die Quartile oder den Median "händisch" ermitteln, muss man vorher die Datenreihe sortieren! Beispiel: $n = 25$ Holzfeuchten bei Produkt 1:

8,3	8,4	8,4	9,2	9,3	9,4	10,0	10,1	10,3	10,3	10,5	10,5	**10,6**	11,3	11,3	11,4	11,5	11,6	11,7	11,7	12,0	12,2	12,4	12,7	14,9

Der Median (= 2. Quartile) ist der Datenwert genau in der Mitte, also genau bei 10,5 (25 Daten). Will man die Datenreihe vierteln, dann geht sich das hier nicht exakt aus, da 25/4 =

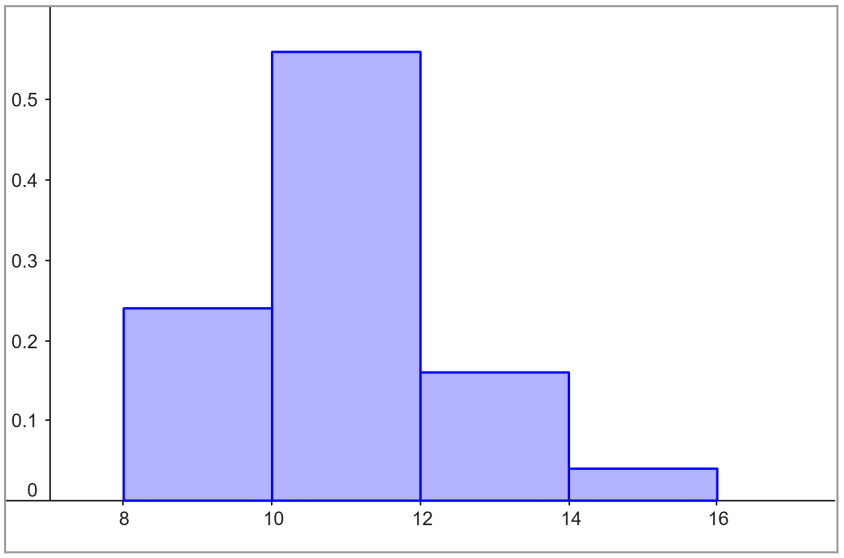
6,25. D.h. die 1. Quartile liegt zwischen 9,4 und 10,0. Man nimmt hier oftmals einfach den Mittelwert, d.h. $Q_1 = 9{,}7$. Bei der 3. Quartile ist das genau so, hier sind aber die beiden Zahlen genau gleich, somit ist $Q_3 = 11{,}7$. Bestimme auch die Quartile für die Holzfeuchte-Daten für Produkt 2.

Häufigkeiten und Histogramm

Die Verteilung der Daten kann man auch durch Gruppierung (Klassenbildung) und den entsprechenden Häufigkeiten zu diesen Klassen darstellen, hier am Beispiel Produkt 1:

Holzfeuchte Klassen (Bereiche)	absolute Häufigkeit	relative Häufigkeit
8 - 10 oder [8; 10[	6	0,24
10 - 12 oder [10; 12[	14	0,56
12 - 14 oder [12; 14[	4	0,16
14 - 16 oder [14; 16]	1	0,04

Ein Histogramm stellt diese Häufigkeiten grafisch dar, in folgender Darstellung die relativen:



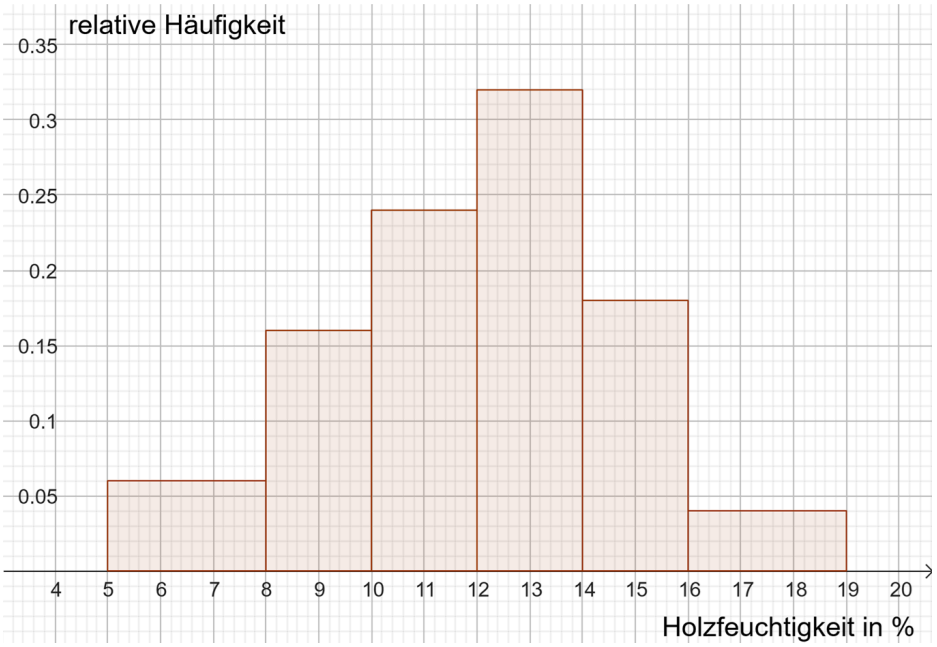
Beispiel 2: Holzfeuchte - Häufigkeitstabellen

Oftmals werden die Daten nicht direkt bereitgestellt, sondern durch Angabe einer Häufigkeitstabelle. Die folgende Tabelle enthält die relativen Häufigkeiten zu $n = 150$ gemessenen Holzproben. Angegeben werden nicht die tatsächlich gemessene Holzfeuchten, sondern nur die relativen Häufigkeiten zu den angegebenen Klassen:

Holzfeuchtigkeit in % Klassen	relative Häufigkeit	Klassenmitte

5 - 8	0,06	6,5
8 - 10	0,16	9
10 - 12	0,24	11
12 - 14	0,32	13
14 - 16	0,18	15
16 - 19	0,04	17,5

Diese Häufigkeiten kann man z.B. folgendermaßen grafisch darstellen:

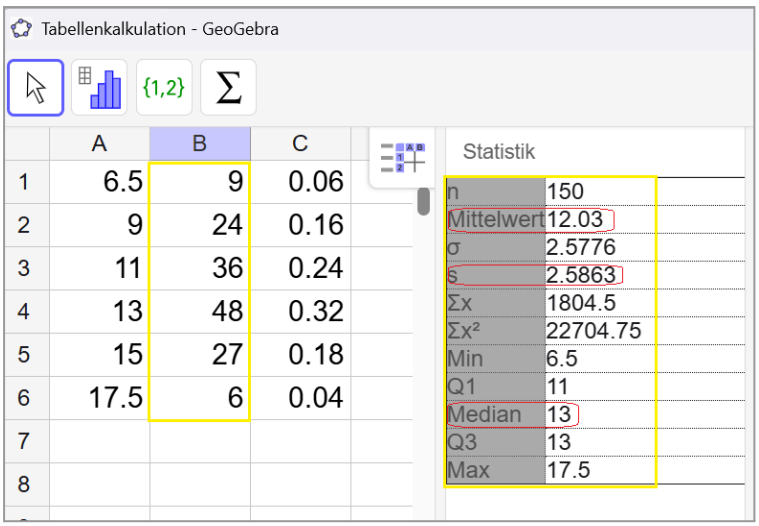


Mittelwert und Standardabweichung aus Häufigkeitstabellen

In diesem Fall kann man den Mittelwert und die Standardabweichung aus den Daten nicht direkt berechnen. Aber man kann einen Näherungswert ermitteln, wenn man davon ausgeht, dass die Daten in den einzelnen Klassen gleichmäßig verteilt sind. Dann entspricht der Mittelwert pro Klasse in etwa dem Mittelpunkt dieser Klasse, wie oben angegeben. Somit erhält man für diese Angaben folgende Parameter

- ➔ Mittelwert = 12.03 %
- ➔ Standardabweichung = 2.5863 %
- ➔ Median = 13 %

Anmerkung: mit Hilfe von Geogebra kann diese Aufgabe sehr einfach gelöst werden (Hinweis: wenn man die absoluten Häufigkeiten verwendet, dann können alle Parameter berechnet werden, bei den relativen Häufigkeiten können Quartile und Median nicht bestimmt werden).



Diese Werte können natürlich auch mit Hilfe der Algebra-Befehlszeile berechnet werden:

$m = \{6.5,\ 9,\ 11,\ 13,\ 15,\ 17.5\}$

$h = \{0.06,\ 0.16,\ 0.24,\ 0.32,\ 0.18,\ 0.04\}$

$H = h \cdot 150 = \{9,\ 24,\ 36,\ 48,\ 27,\ 6\}$

$a = \text{Mittel}(m, H) = 12.03$

$b = \text{stdev}(m, H) = 2.5863$

$c = \text{Median}(m, H) = 13$

Korrelation und Regression

Aufgabenstellung

Verwende eine Software deiner Wahl und ermittle aus den gegebenen Beispieldaten die entsprechenden Regressionsmodelle.

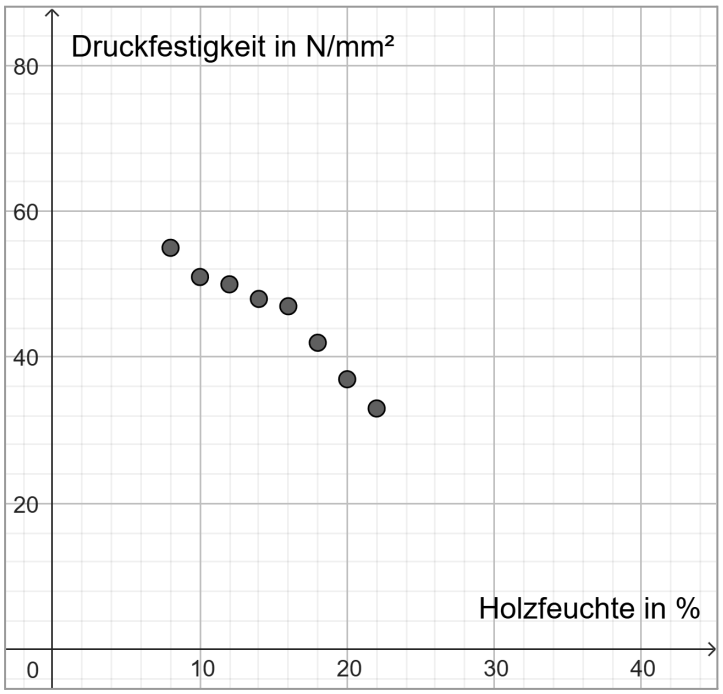
Falls Hilfe benötigt wird, einfach die folgenden Demos ansehen: Video$_4$ (Regression) oder eine KI fragen. Geogebra Files zur freien Verwendung: ggb$_4$

Beispiel 1: Lineare Regression - Holzfeuchte und Druckfestigkeit

Für jeweils $n = 8$ Holzproben wurde für jede der gegebenen Holzfeuchten die Druckfestigkeit in N/mm² gemessen und die mittlere Druckfestigkeit aus den 8 Daten bestimmt:

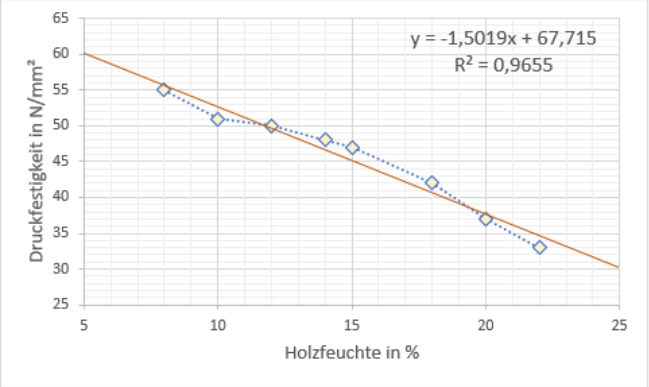
Holzfeuchte in %	8	10	12	14	15	18	20	22
Mittlere Druckfestigkeit in N/mm²	55	51	50	48	47	42	37	33

Den Zusammenhang zwischen Holzfeuchte und Druckfestigkeit kann man durch einen Punkt Plot (*xy*-Plot) darstellen:



Der **Korrelationskoeffizient** $r = -0.9729$ gibt die Güte des linearen Zusammenhangs zwischen Holzfeuchte und Druckfestigkeit an. Das negative Vorzeichen bedeutet eine negative Korrelation, d.h. mit steigender Holzfeuchte sinkt die Druckfestigkeit.

Bei der **Regressionsrechnung** wird ein mathematisches Modell, also eine Funktion angegeben, die den Zusammenhang beschreibt. In diesem Beispiel verwenden wir eine Lineare Regressionsfunktion, hier mit MS-Excel ermittelt:



Das **Bestimmtheitsmaß** R^2 beschreibt die Güte der Anpassung an die Daten → Video$_4$. Im linearen Fall ist das Bestimmtheitsmaß der Korrelationskoeffizient zum Quadrat. Bei der nichtlinearen Regression ist das nicht der Fall ;-)

Beispiel 2: Nichtlineare Regression - Bestandeshöhenkurve

Die folgende Tabelle enthält Daten für den Brusthöhendurchmesser (BHD) in cm und die Baumhöhe in m von $n = 25$ Fichten eines bestimmten Waldbestandes. Die geometrische Form, die diese Daten bilden, nennt man Bestandeshöhenkurve[1].

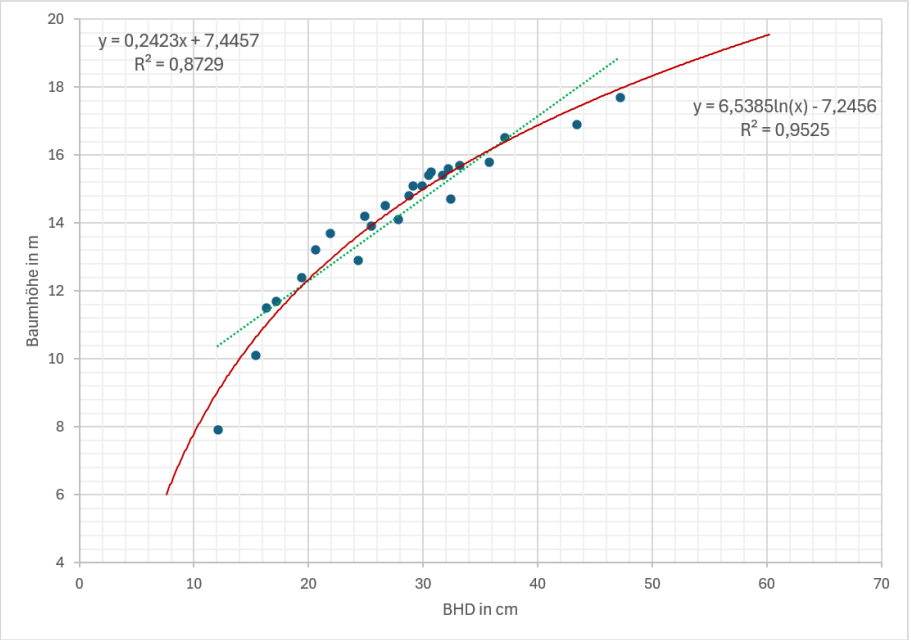
BHD	12,1	15,4	16,3	17,2	19,4	20,6	21,9	24,3	26,7	24,9	25,5	27,8	28,8	29,1	29,9	30,5	30,7	31,7	32,2	32,4	33,2	35,8	37,1	43,4	47,2
Höhe	7,9	10,1	11,5	11,7	12,4	13,2	13,7	12,9	14,5	14,2	13,9	14,1	14,8	15,1	15,1	15,4	15,5	15,4	15,6	14,7	15,7	15,8	16,5	16,9	17,7

Führt man hier eine Regressionsrechnung durch, dann sieht man sehr schön, dass der Korrelationskoeffizient mit $r = 0{,}9343$ zwar hoch ist (d.h. man könnte einen linearen Zusammenhang vermuten), aber ein Logarithmisches Modell viel besser zu den Daten passt als ein lineares (Berechnung mit MS-Excel).

---

[1] Schmidt, A.: (1968) Der rechnerische Ausgleich von Bestandeshöhenkurven. Forstwissenschaftliches Zentralblatt 86 (6), Verlag Paul Parey, Hamburg/Berlin, S. 370-382. Nagel, J.: (2001) Skript: Waldmesslehre, Niedersächsischen Forstlichen Versuchsanstalt, Abteilung Waldwachstum.

BHD in cm	Baumhöhe in m
12,1	7,9
15,4	10,1
16,3	11,5
17,2	11,7
19,4	12,4
20,6	13,2
21,9	13,7
24,3	12,9
26,7	14,5
24,9	14,2
25,5	13,9
27,8	14,1
28,8	14,8
29,1	15,1
29,9	15,1
30,5	15,4
30,7	15,5
31,7	15,4
32,2	15,6
32,4	14,7
33,2	15,7
35,8	15,8
37,1	16,5
43,4	16,9
47,2	17,7



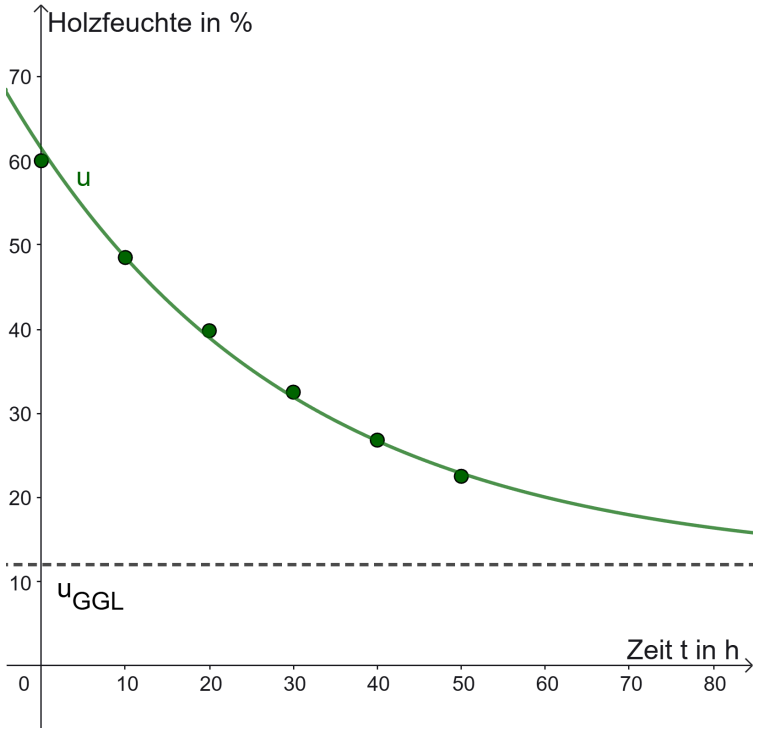
Beispiel 3: Nichtlineare Regression - Holztrocknung

In einer konventionellen Trockenkammer werden 50 mm Fichtenbohlen getrocknet. Um Energiekosten zu sparen und Risse zu vermeiden, sollte vorhergesagt werden, wann die gewünschte Endfeuchte in etwa erreicht ist, ohne den Prozess manuell zu unterbrechen. Die Holzfeuchte (u) wird zu verschiedenen Zeitpunkten (t) gemessen und dokumentiert:

Zeit t (Stunden)	Holzfeuchte u (%)
0	60.0
10	48.5
20	39.8
30	32.5
40	26.8
50	22.5

Bestimmen Sie eine exponentielle Regressionsfunktion f zur Beschreibung des Trocknungsverlaufes. Ziel ist es, dass die Regressionsfunktion gegen die Zielfeuchte der Trocknung $u_{GGL} = 12\%$ läuft.

Eine “normale” exponentielle Regressionsfunktion strebt aber gegen Null. Wie könnte man hier vorgehen, damit man die korrekte Regressionsfunktion aus den gegebenen Daten erhält?


Holzfeuchte in %
70
60
50
40
30
20
10
u
u_{GGL}
Zeit t in h
0
10
20
30
40
50
60
70
80

Begriffe

Statistische Begriffe

Beschreibende Zusatztexte sind teilweise mit Hilfe von KI (google gemini) generiert.

Statistische Software

Hier einige freie Softwareprodukte für Statistik:

➔ Tabellenkalkulation: Geogebra, MS-Excel, Google Sheets, Libreoffice, Gnumeric, …
➔ Komplexere Softwareprodukte: Jamovi, R, Jasp, …

Daten

Im Bereich der Statistik lassen sich Daten in verschiedene Formen einteilen, wobei die häufigste Unterscheidung zwischen **qualitativen** und **quantitativen** Daten getroffen wird.

Qualitative Daten

Qualitative Daten, auch kategoriale Daten genannt, beschreiben Eigenschaften oder Merkmale, die nicht numerisch sind. Sie werden oft in Kategorien oder Gruppen eingeteilt. Beispiele hierfür sind Geschlecht (männlich, weiblich), Nationalität oder Augenfarbe. Innerhalb der qualitativen Daten gibt es zwei Subtypen:

- **Nominaldaten**: Bei diesen Daten gibt es keine natürliche Rangfolge. Ein Beispiel ist die Haarfarbe (blond, braun, schwarz).
- **Ordinaldaten**: Hier gibt es eine sinnvolle Rangfolge. Ein Beispiel ist die Schulnotenvergabe (Sehr gut, Gut, Befriedigend).

Quantitative Daten

Quantitative Daten sind numerisch und können gemessen oder gezählt werden. Sie geben Auskunft über eine Menge oder Anzahl. Auch hier gibt es zwei Hauptkategorien:

- **Diskrete Daten**: Diese Daten können nur bestimmte, abzählbare Werte annehmen. Es sind oft ganze Zahlen, wie die Anzahl der Kinder in einer Familie oder die Anzahl der Autos in einem Parkhaus.
- **Stetige Daten**: Diese Daten können jeden Wert innerhalb eines bestimmten Bereichs annehmen, da sie durch Messungen entstehen. Beispiele sind Körpergröße, Gewicht oder Temperatur.

Hier im folgenden Beispiel “Demo-Beispiel Wirkung”, sieht man Variablen (Spalten) mit unterschiedlicher Ausprägung (Datentypen)

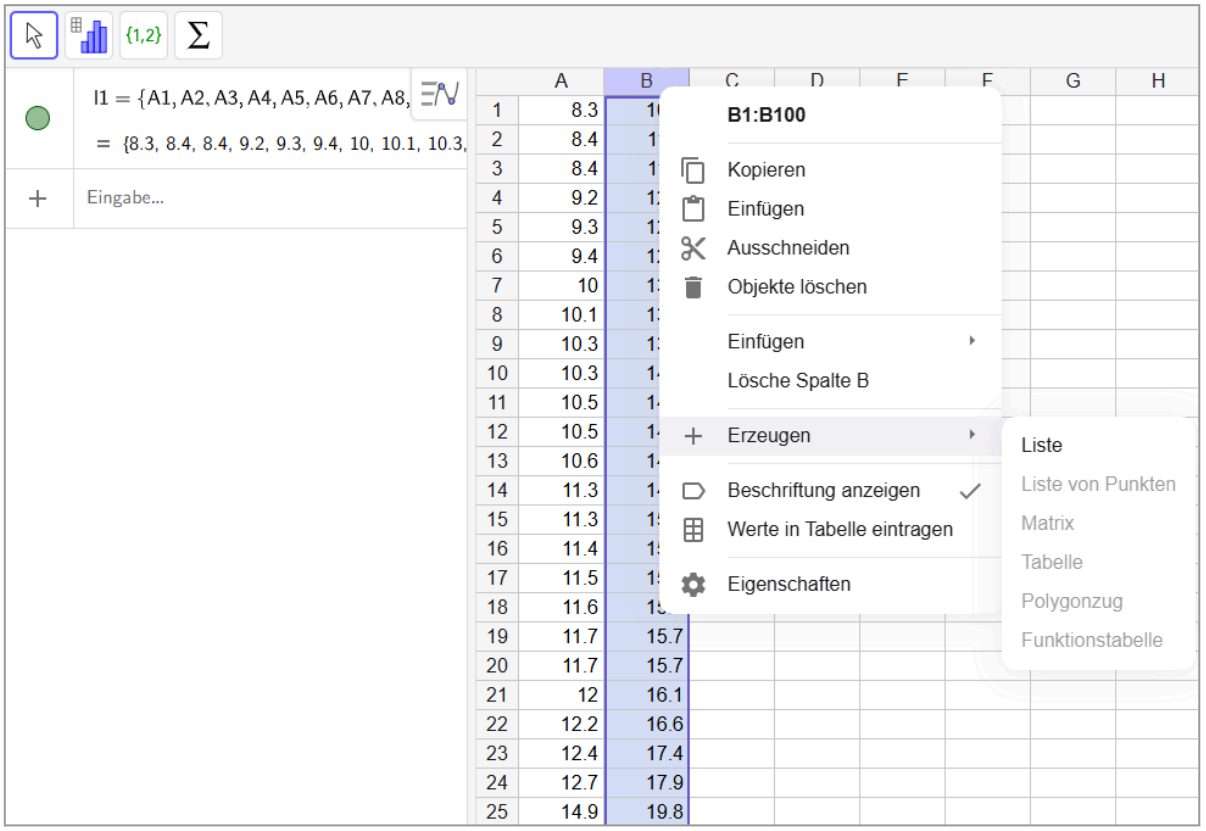
Nr	Behandlung	Wirkung	Nebenwirkung	Name	Bewertung	Datum
1	0	42,85	0	Berger	0	01.01.2019
2	3	45,20	1	Berger	5	17.03.2019
3	1	45,06	0	Berger	6	11.07.2019
4	3	42,20	1	Schreiber	4	02.08.2020
5	2	45,11	0	Berger	7	18.02.2020
6	3	50,04	0	Gratz	6	14.06.2019
	…	…	…	…	…	…
218	4	49,43	1	Ziegler	7	26.06.2020
219	3	48,26	0	Ziegler	6	24.11.2020
220	0	45,88	0	Ziegler	0	30.12.2020

Lösungen

Lösungen

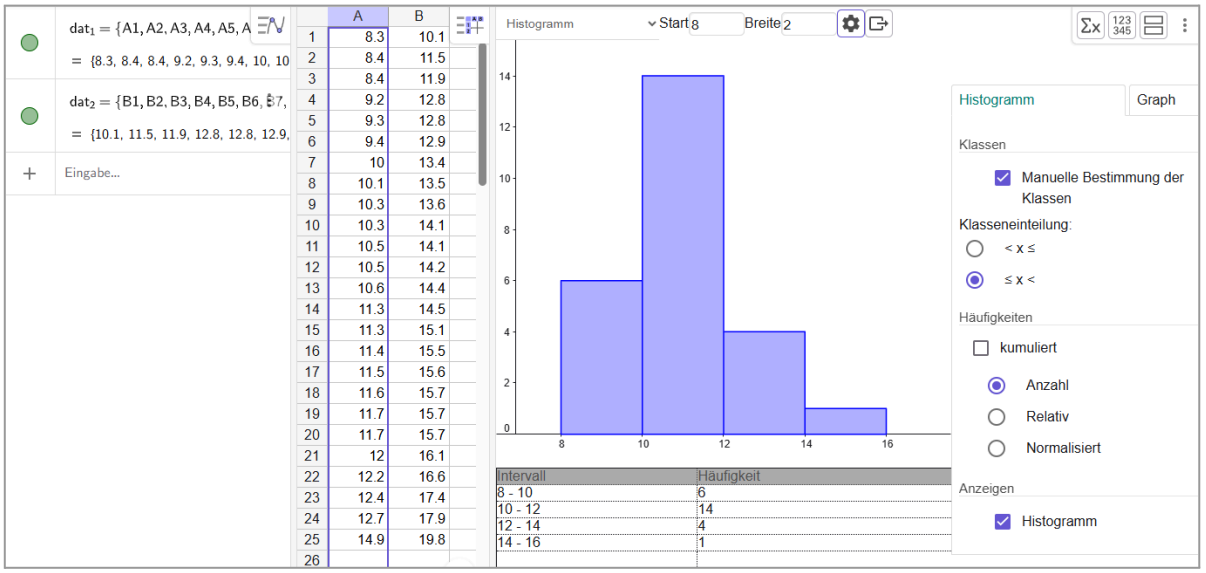
Beispiel 1 Geogebra:

Daten eingeben (kopieren). Man kann auch eine Liste der Daten erzeugen lassen

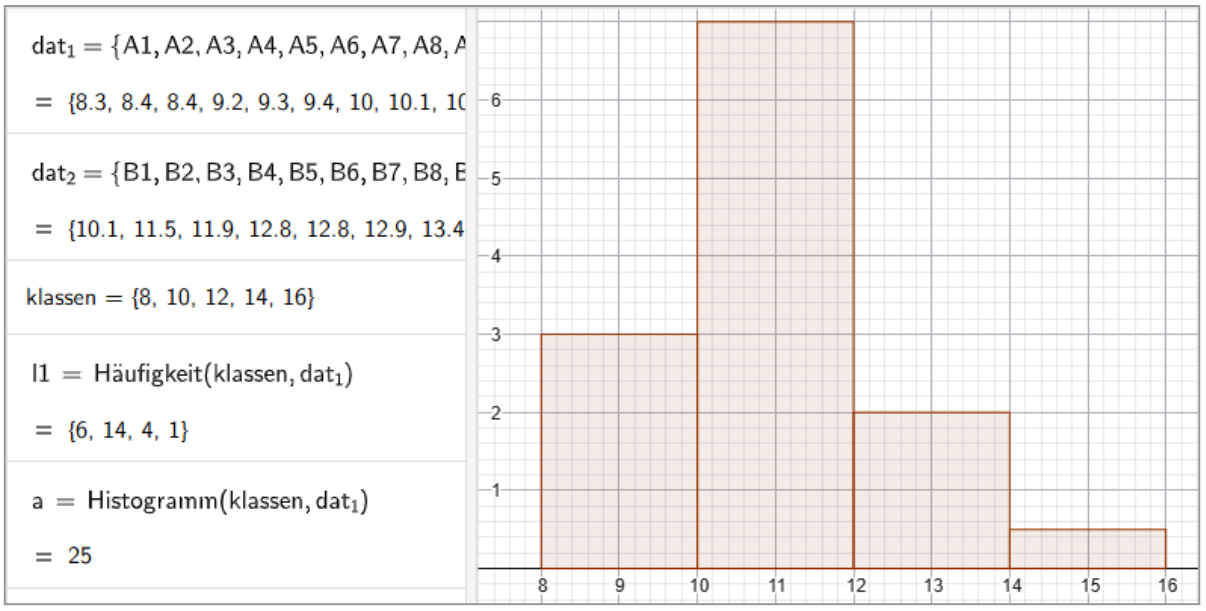


Analyse mehrerer Variablen:

Häufigkeiten → Analyse einer Variablen → Klassen manuell einstellen und Häufigkeitstabelle anzeigen



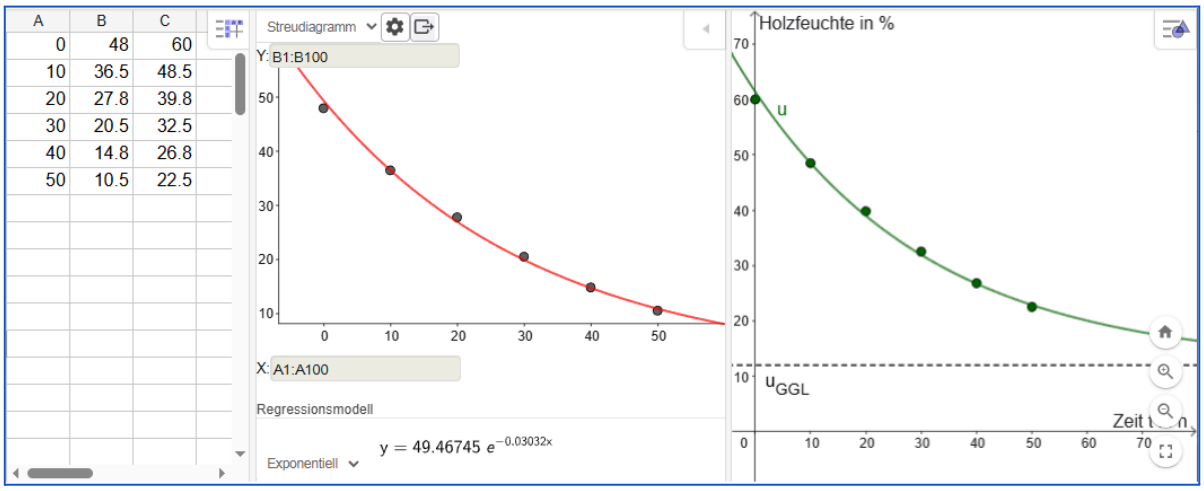
oder als Algebra Eingabe:



Beispiel 3

Man reduziert einfach die Daten um 12% und bestimmt die exponentielle Regressionsfunktion mit Hilfe von einer bekannten Software (MS-Excel, Google Sheets, Geogebra, …). Die Lösung ist dann einfach das errechnete Modell + 12:

$$u(t) = 49.46745 \cdot e^{-0.03032 \cdot t} + 12$$


A B C
0 48 60
10 36.5 48.5
20 27.8 39.8
30 20.5 32.5
40 14.8 26.8
50 10.5 22.5
Streudiagramm
Y: B1:B100
X: A1:A100
Regressionsmodell
Exponentiell
y = 49.46745 e^-0.03032x
Holzfeuchte in %
u
u_GGL
Zeit