

Einführung in freie statistische Software

Christian Reinboth, Dipl.-Wi.Inf.(FH)

Sommersemester 2017

Bachelorstudiengang Betriebswirtschaftslehre

Agenda

- Einführung
 - Motivation
 - Freie Software
 - PAST-Grundlagen
- Deskriptive Statistik
 - Lagemaße
 - Streuungsmaße
 - Ausreißeridentifikation
 - Korrelationskoeffizienten
- Explorative Statistik
 - Box-Plot
- Induktive Statistik
 - Lineare Regression
- Was kann andere (freie) Software?
- Kleine Vorschau auf Statistik II

Wichtiger Hinweis

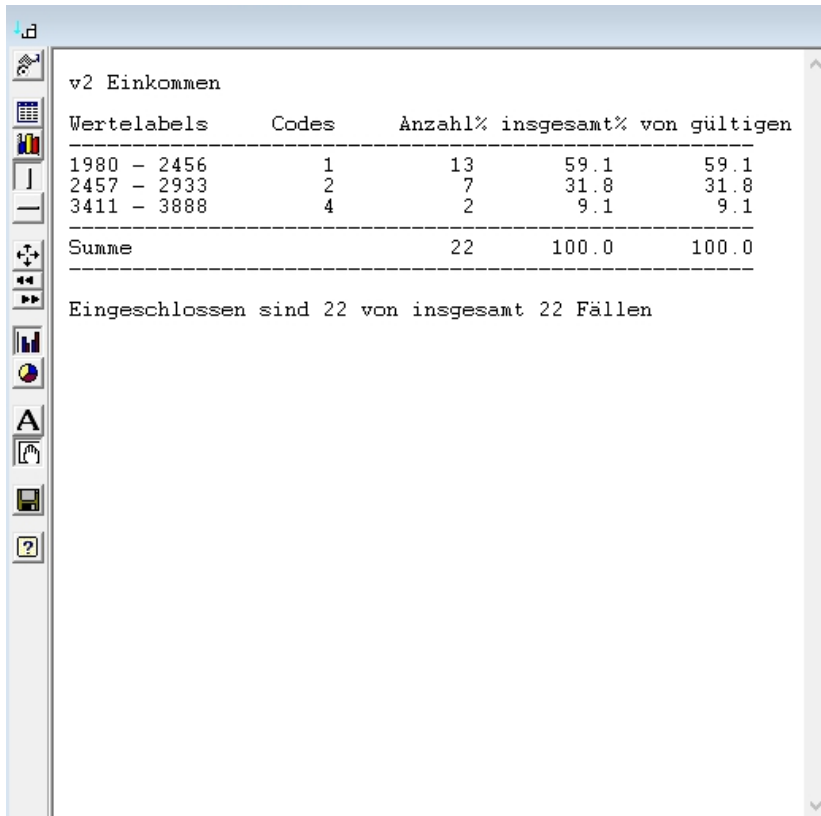
...für alle, die diese Folien „nur“ lesen

- Dieser Foliensatz ergänzt die Vorlesung „Statistik“ im berufsbegleitenden Bachelor-Studiengang Betriebswirtschaftslehre an der Hochschule Harz
- Dieser Foliensatz beinhaltet Übungen und Software-Tutorials für den Umgang mit der statistischen Analysesoftware PAST
 - aber keine Wiederholung theoretischer Grundlagen
- Wer nach einer Einführung in die Grundlagen der Statistik sucht, sei deshalb auf den Hauptfoliensatz zu dieser Vorlesung verwiesen

Download des Hauptfoliensatzes unter: <https://www.hs-harz.de/creinboth/lehre/>

Warum eine gesonderte Software-Einführung?

(Nur weil wir nicht per Hand rechnen wollen?)



Wertelabels	Codes	Anzahl	% insgesamt	% von gültigen
1980 - 2456	1	13	59.1	59.1
2457 - 2933	2	7	31.8	31.8
3411 - 3888	4	2	9.1	9.1
Summe		22	100.0	100.0

Eingeschlossen sind 22 von insgesamt 22 Fällen

- Praxisnah: In keinem Betrieb würde eine lineare Regressionsanalyse noch „per Hand“ durchgeführt
- Vorbereitung: Wer im Rahmen der BA empirisch arbeiten möchte, wird hierfür Software einsetzen müssen

Und warum freie Software?

- Eine einfache SPSS-Lizenz kostet 1.168,00 EUR pro User und Jahr
- Freie Software ist ohne Kosten in Studium und Beruf einsetzbar

Was ist SPSS?

Statistical Package for Social Sciences

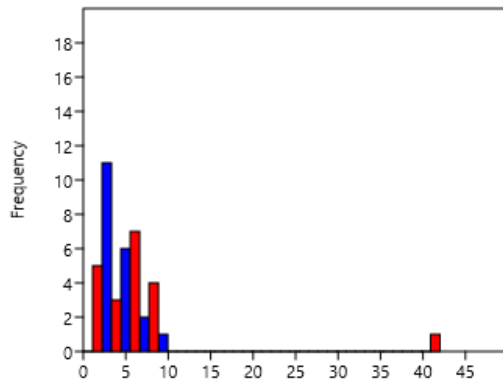
- **SPSS** ist eines der **marktführenden Softwareprodukte** für statistische Analysen in der Sozial- und Gesundheitswissenschaft sowie in der Markt- und Meinungsforschung
- Es wurde 1983 von SPSS Inc., einer Ausgründung der Stanford University, entwickelt
- Der Name wechselte mehrfach von „Statistical Package for Social Sciences“ über „Superior Performing Software System“ und „Predictive Analysis Software“ (PASW) bis zu IBM SPSS STATISTICS seit der Übernahme von SPSS Inc. durch IBM in 2009



www.ibm.com/software/de/analytics/spss/

Empfehlenswerte freie Statistik-Software

(Kategorie: Allgemeine Datenanalyse)



PAST (Windows, Mac)

- Paleontological Statistics Software Package for Education and Data Analysis (Universities of Copenhagen and Oslo)

<http://folk.uio.no/ohammer/past/>

9: Var0002

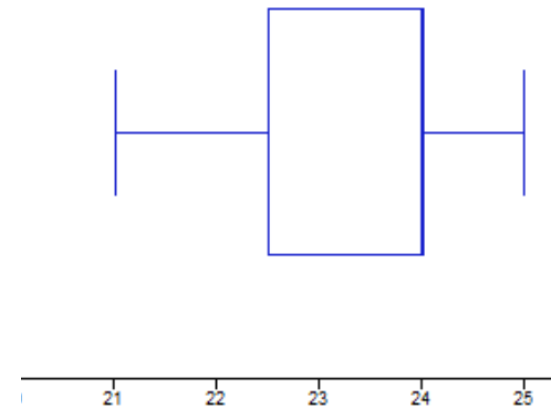
Fall	Var0001	Var0002
1	32,00	34,00
2	34,00	34,00
3	23,00	34,00
4	243,00	34,00
5	334,00	43,00
6	43,00	34,00
7	34,00	34,00
8	43,00	34,00
9	43,00	44,00
10		

Menu options: Fälle sortieren..., Transponieren..., Aggregieren..., Datei aufteilen..., Fälle auswählen..., Fälle gewichten...

PSPP (Windows, Mac, Linux)

- Open Source-„Nachbau“ von SPSS
- Identische Funktionen und Bedienung, „Look & Feel“ ist sehr gut vergleichbar

<https://www.gnu.org/software/pspp/>



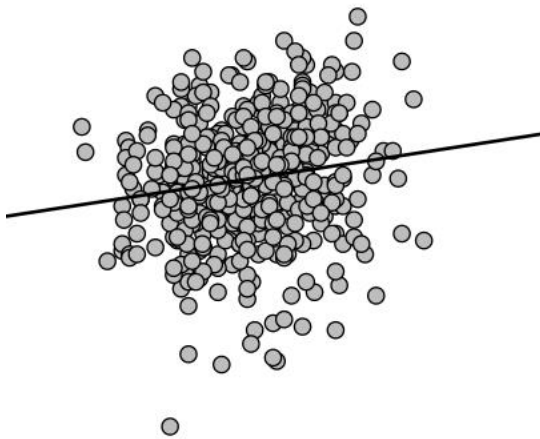
SSP (Windows, Mac)

- Smith's Statistical Package
- „Ein-Mann-Entwicklung“ von Prof. Gary Smith vom Pomona College

<http://economics-files.pomona.edu/GarySmith/StatSite/ssp.html>

Empfehlenswerte freie Statistik-Software

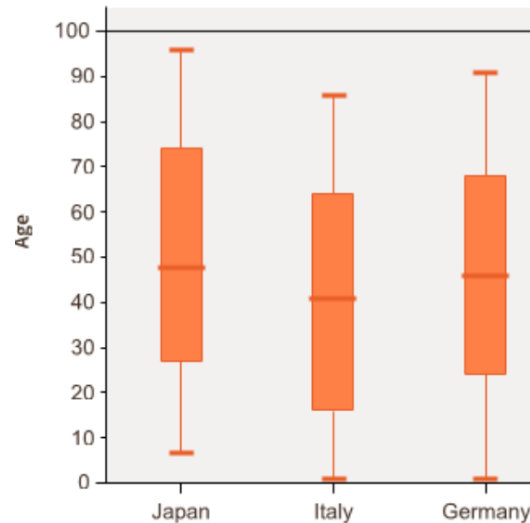
(Kategorie: Spezielle Anforderungen)



JASP (Windows, Mac, Linux)

- Just Another Stats Program
- Bietet liquiden Output, der sich mit jedem Klick ändert (ideal für Lerner)

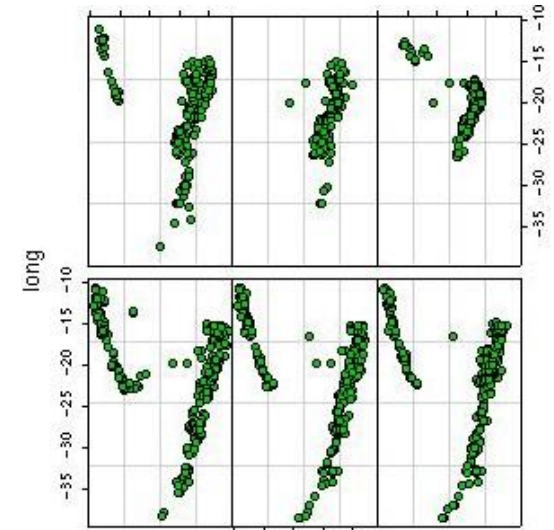
<https://jasp-stats.org>



SOFA (Windows, Mac, Linux)

- Statistics Open For All
- Bietet vielfältige Möglichkeiten der grafischen Aufbereitung von Daten

<http://www.sofastatistics.com>



MacANOVA (Windows, Mac, Linux)

- Entwickelt an der Uni Minnesota
- Der Schwerpunkt der Software liegt auf der Varianzanalyse (ANOVA)

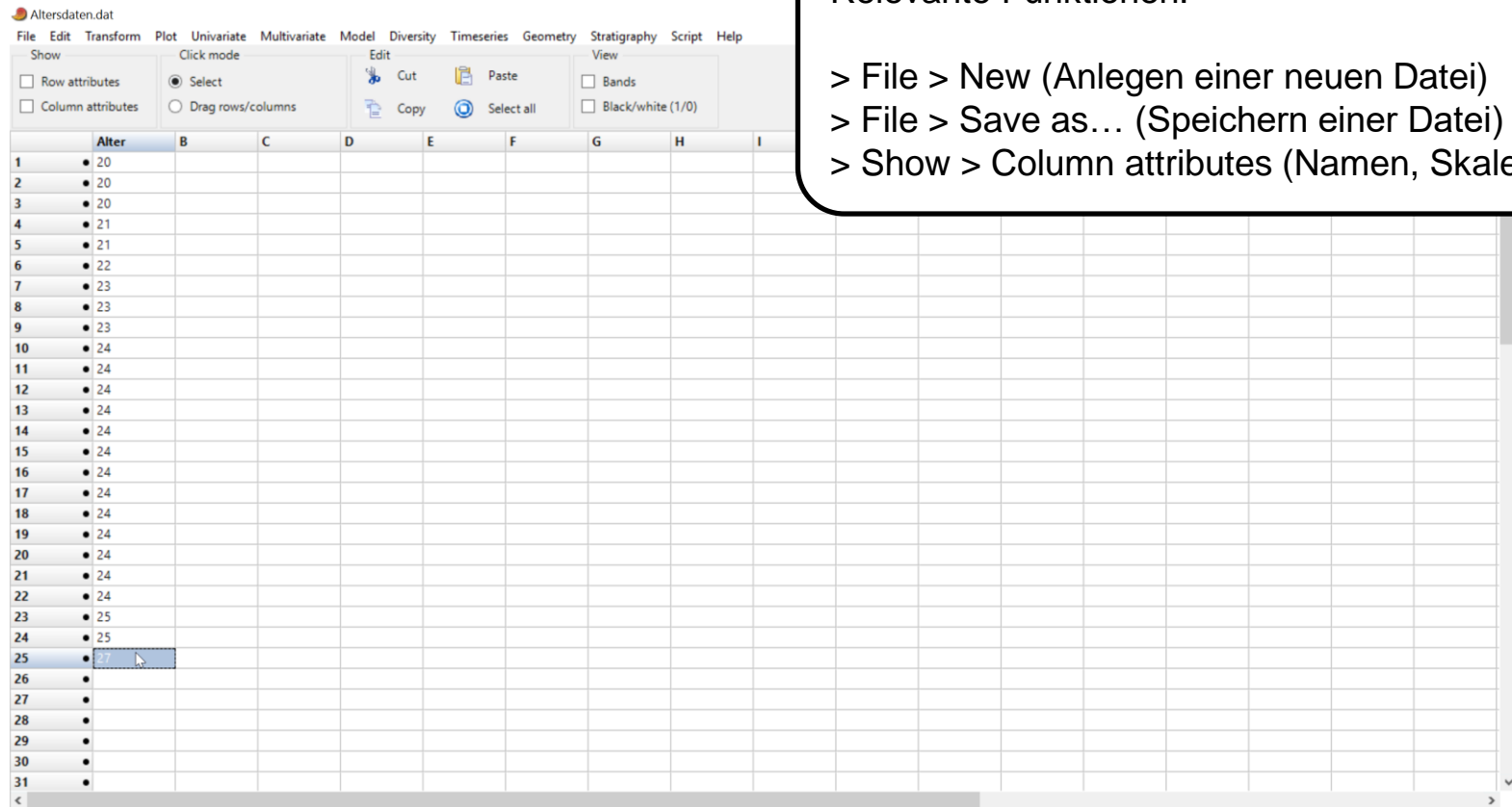
<http://www.stat.umn.edu/macanova/>

Unser zentraler Beispieldatensatz (bereits aus der Hauptvorlesung bekannt)

Ausprägung	abs. Häufigkeit	rel. Häufigkeit	in %
20 Jahre	3	0,12	12,00%
21 Jahre	2	0,08	8,00%
22 Jahre	1	0,04	4,00%
23 Jahre	3	0,12	12,00%
24 Jahre	13	0,52	52,00%
25 Jahre	2	0,08	8,00%
26 Jahre	0	0,00	0,00%
27 Jahre	1	0,04	4,00%
Σ	25	1,00	100,00%

Wie bekommen wir diese
Daten nun in PAST?

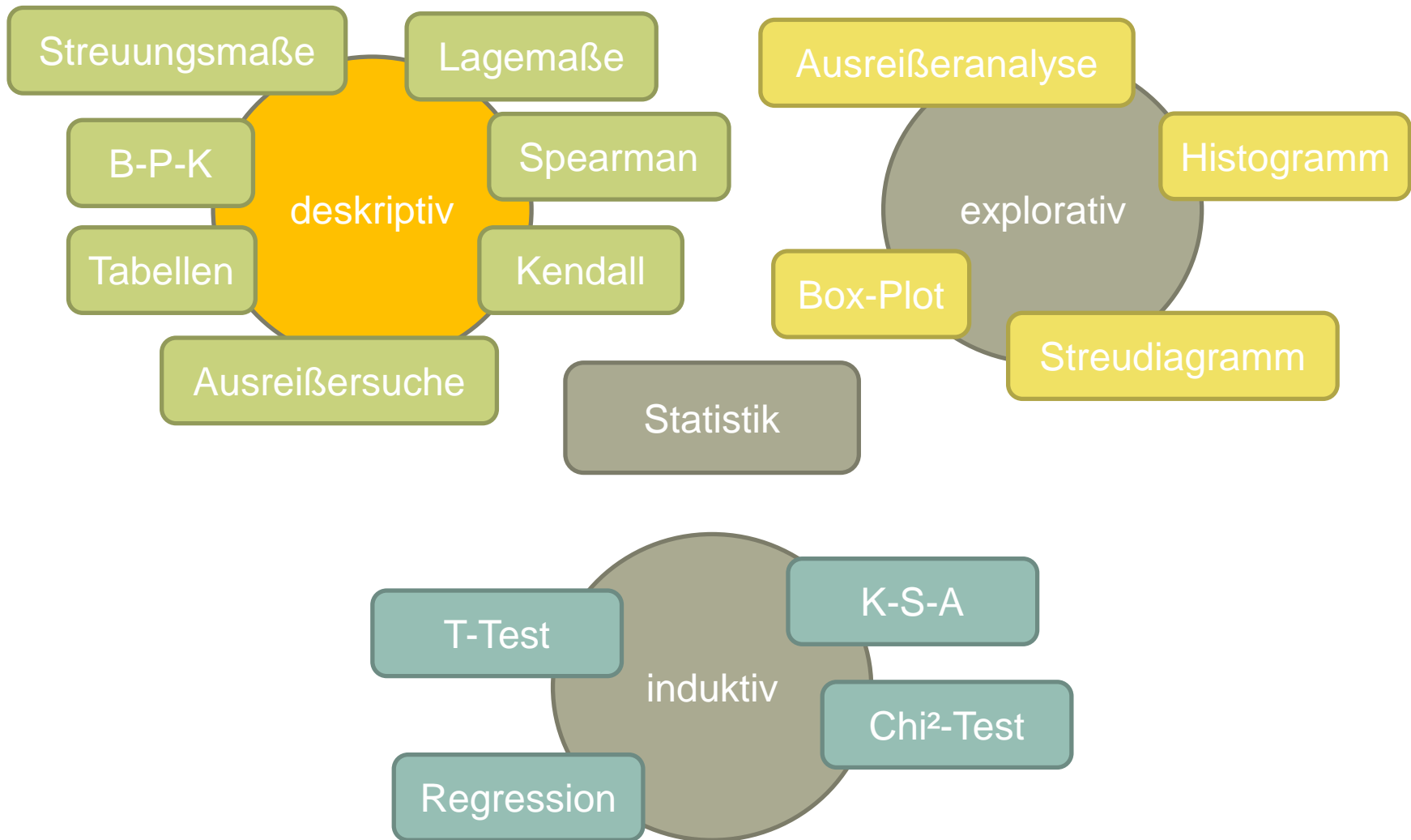
Eingabe von Daten in PAST



Relevante Funktionen:

- > File > New (Anlegen einer neuen Datei)
- > File > Save as... (Speichern einer Datei)
- > Show > Column attributes (Namen, Skalen)

Wo befinden wir uns?



Lagemaße und Streuungsmaße

> Univariate > Summary statistics

Was ist hier was?

N = Anzahl der Werte

Min = kleinster Wert

Max = größter Wert

Mean = arithmetisches Mittel

Geom. mean = Geometrisches Mittel

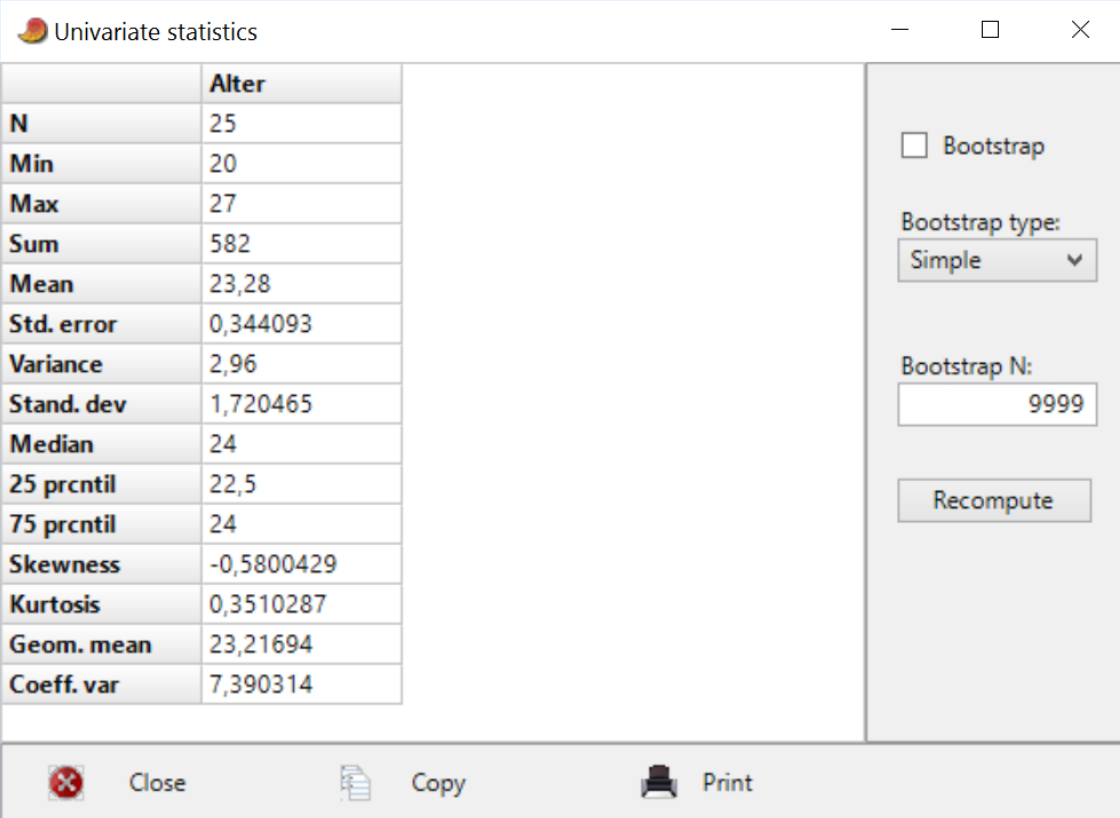
25 prcntil = Unteres Perzentil

Median = Mittleres Perzentil

75 prcntil = Oberes Perzentil

Variance = Varianz

Stand dev. = Standardabweichung



	Alter
N	25
Min	20
Max	27
Sum	582
Mean	23,28
Std. error	0,344093
Variance	2,96
Stand. dev	1,720465
Median	24
25 prcntil	22,5
75 prcntil	24
Skewness	-0,5800429
Kurtosis	0,3510287
Geom. mean	23,21694
Coeff. var	7,390314

Das „SPSS-Analyseproblem“

- Software führt JEDE Analyse unabhängig von den Voraussetzungen durch!
- ...also auch die Berechnung des arithmetischen Mittels
 - ... aus Schulnoten
 - ... aus Geschlechtern
 - ... aus Kontonummern
 - ... aus Telefonnummern
 - ... aus Präferenzrängen
- Bei komplexen Verfahren sind noch weit schlimmere „Vergehen“ denkbar
- Die fachlichen Kenntnisse der Anwender/innen sind daher entscheidend
- Darum: KEINE Analyse ohne vorherige Prüfung der Voraussetzungen!



Warum ergeben sich andere Streuungsmaße?

- In der Vorlesung haben wir die Standardvarianz als Durchschnitt der quadrierten Abweichungen berechnet:

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- $s^2 = 2,8416 \mid s = 1,6875$

- Mit Hilfe von PAST berechnen wir die sog. Stichprobenvarianz mit den Freiheitsgraden (n-1) im Vorfaktor:

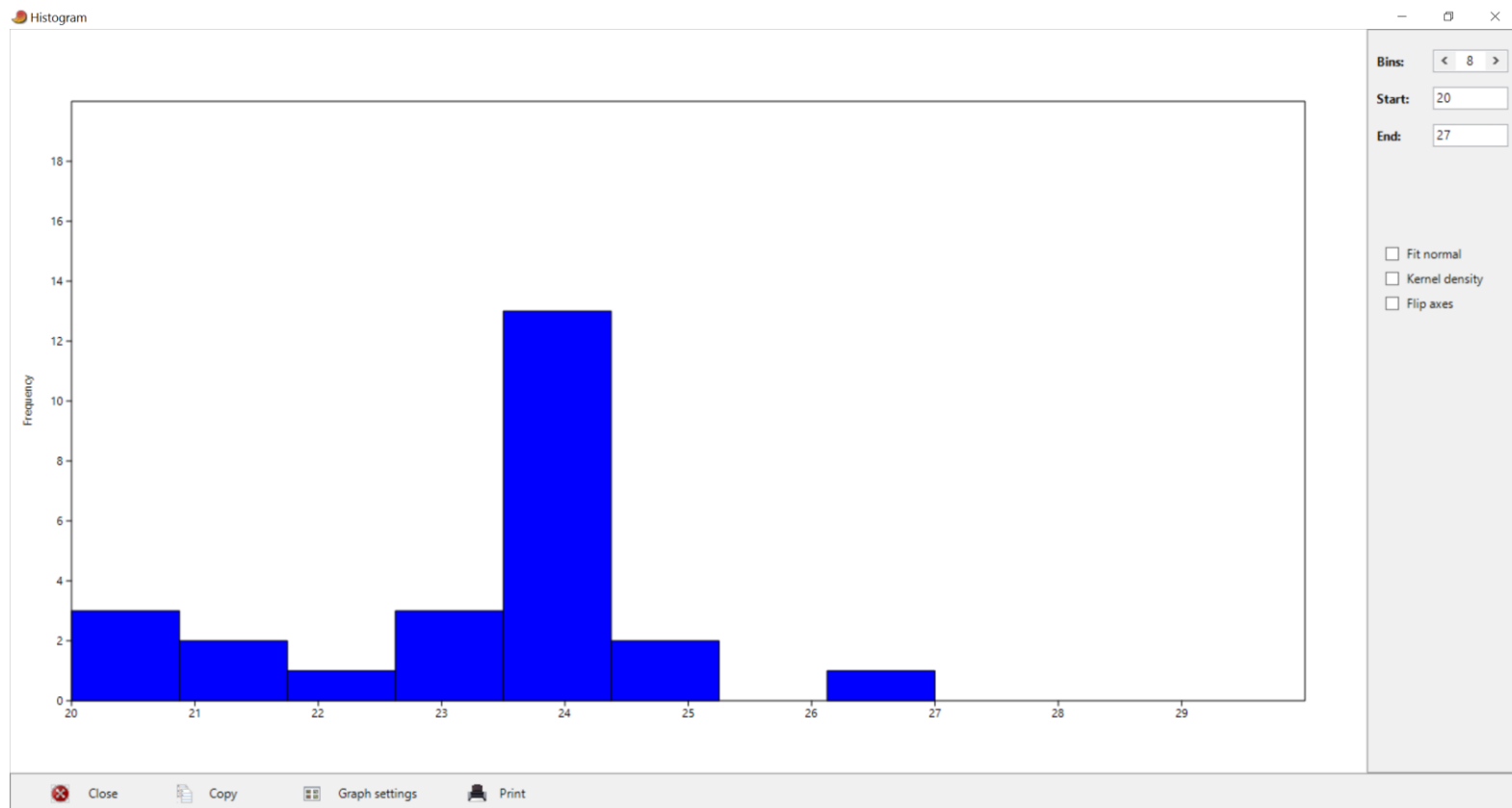
- $s^2 = 2,96 \mid s = 1,72$

Ist die Wahl der Formel eher für große oder eher für kleine Datensätze relevant?

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

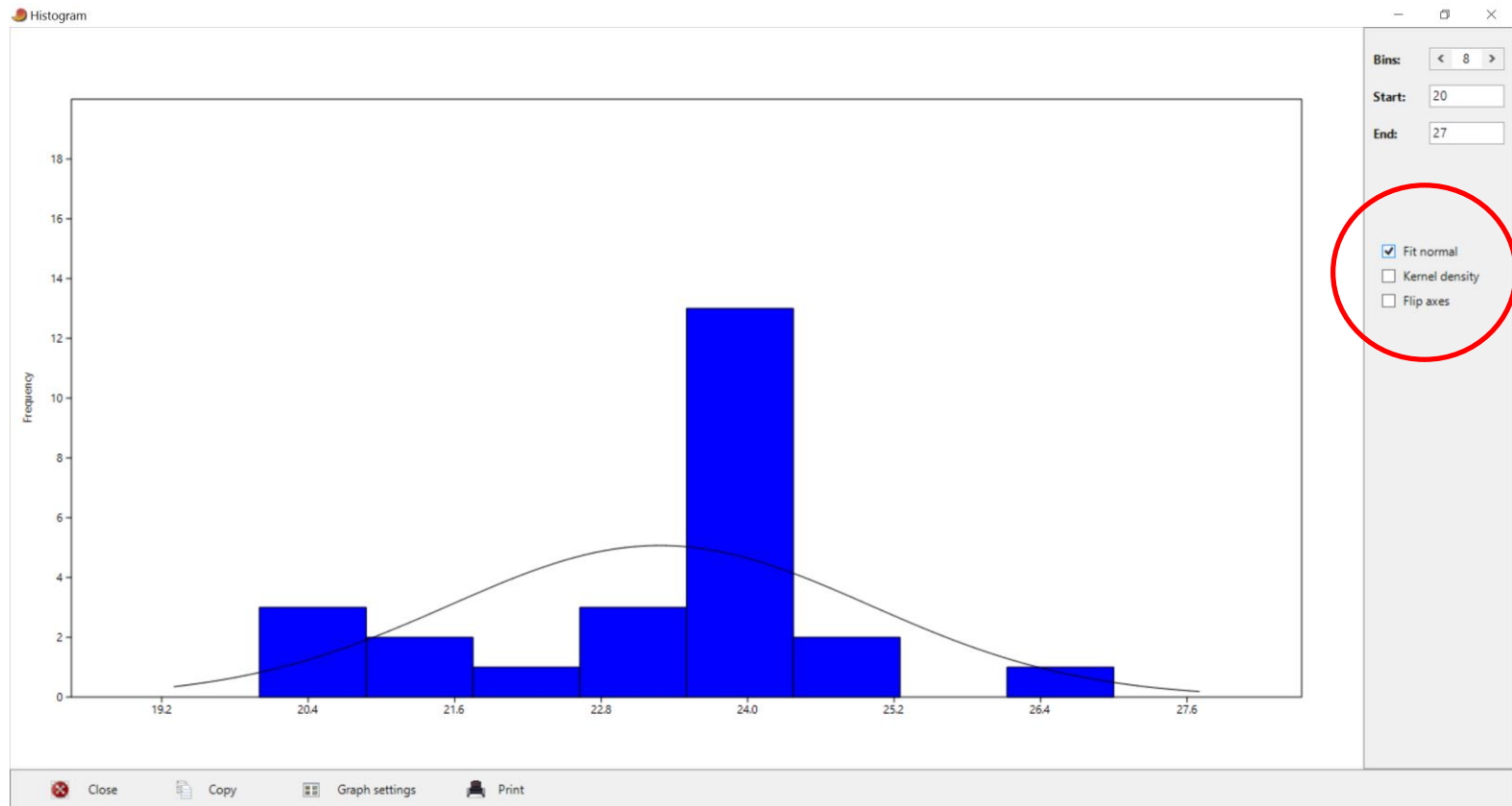
Gibt es einen Modus?

> Plot > Histogram



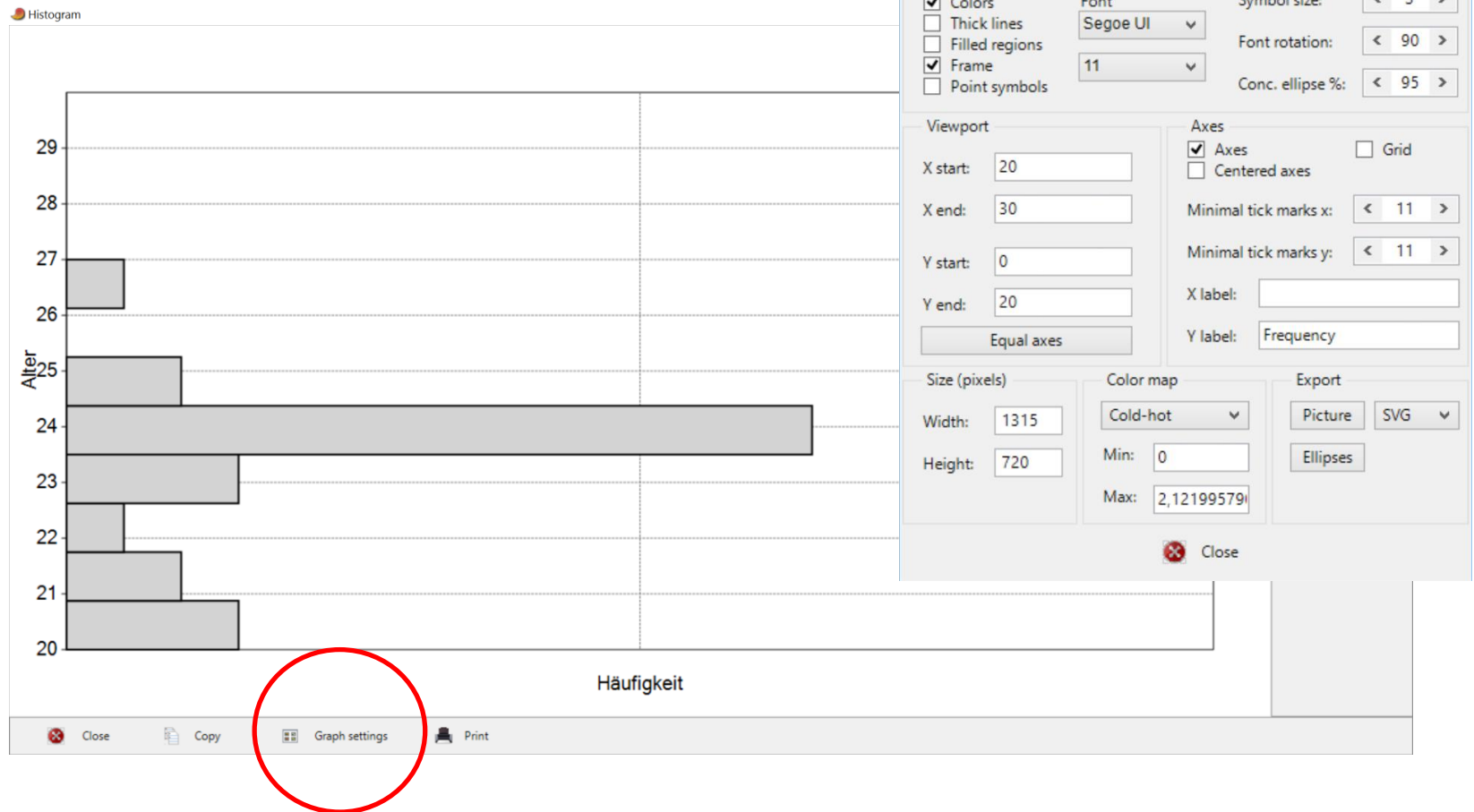
Folgt die Verteilung einer Normalverteilung?

> Plot > Histogram



Lässt sich die Grafik noch individualisieren?

> Plot > Histogram > Graph settings

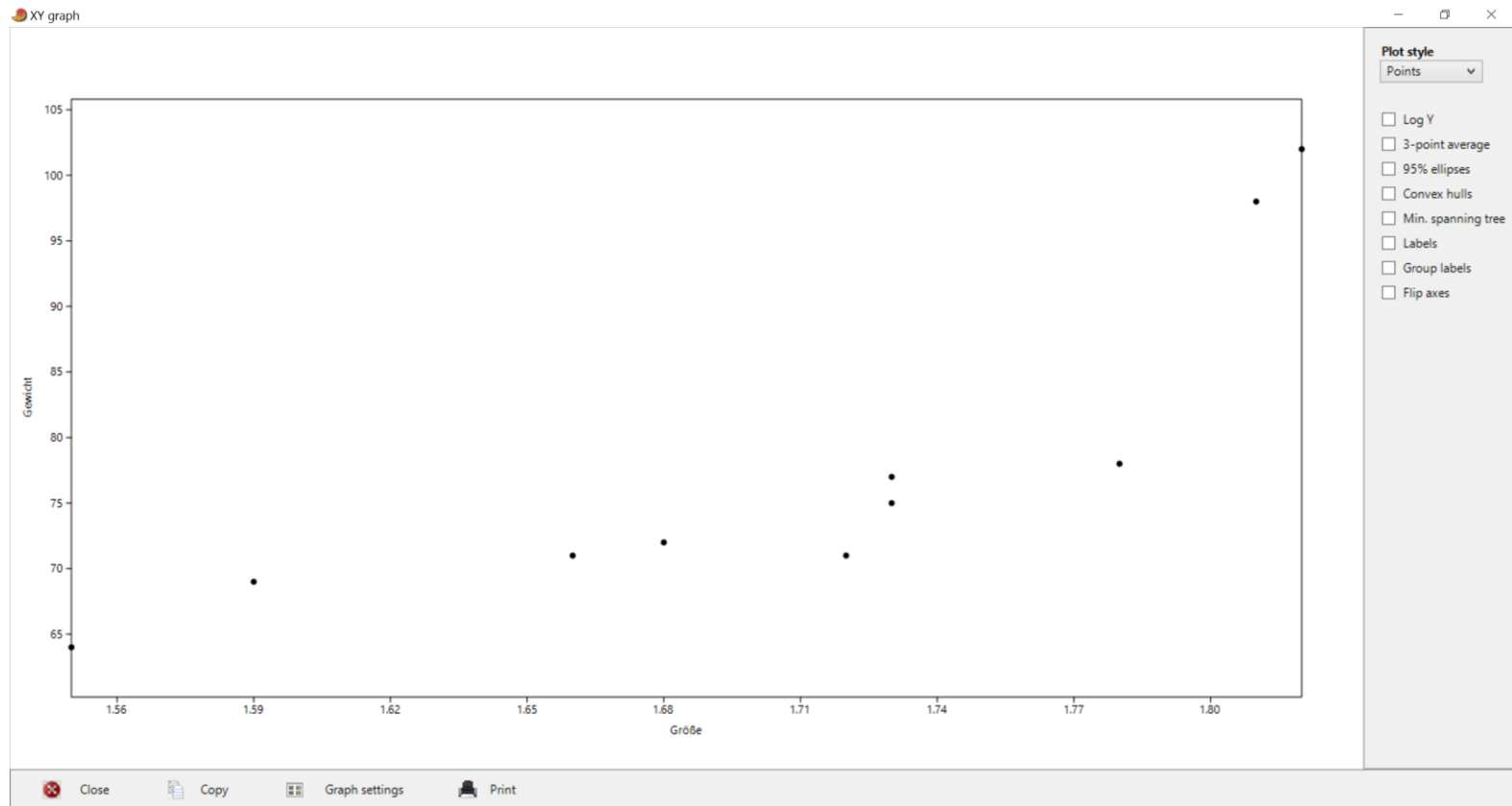


Bivariater Datensatz für Korrelationsanalysen (ebenfalls aus der Hauptvorlesung bekannt)

Befragte/r	Größe (m)	Gewicht (kg)
1	1,55	64
2	1,68	72
3	1,72	71
4	1,73	75
5	1,82	102
6	1,81	98
7	1,66	71
8	1,78	78
9	1,73	77
10	1,59	69

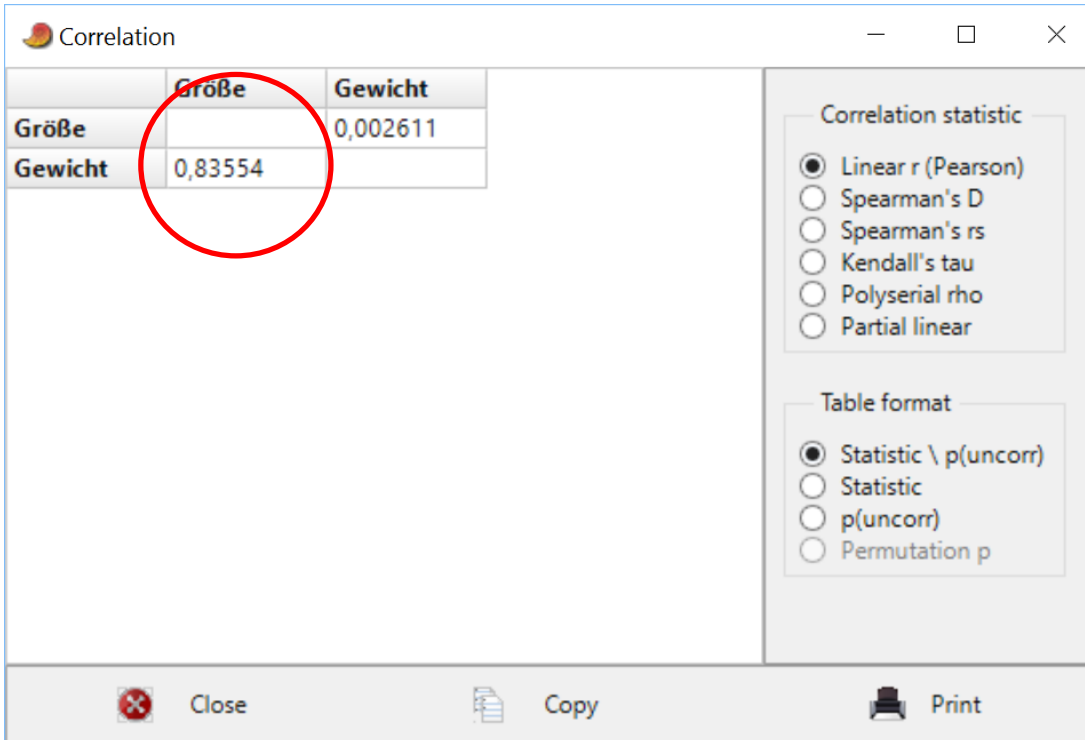
Ist ein Zusammenhang grafisch plausibel?

> Plot > XY graph



Berechnung von Korrelationskoeffizienten

> Univariate > Correlation



	Größe	Gewicht
Größe		0,002611
Gewicht	0,83554	

Correlation statistic

- ☒ Linear r (Pearson)
- ☐ Spearman's D
- ☐ Spearman's rs
- ☐ Kendall's tau
- ☐ Polyserial rho
- ☐ Partial linear

Table format

- ☒ Statistic \ p(uncorr)
- ☐ Statistic
- ☐ p(uncorr)
- ☐ Permutation p

Was ist hier was?

Kendall's tau =
Konkordanzkoeffizient nach Kendall

Linear r (Pearson) =
Bravais-Pearson-Korrelationskoeffizient

Spearman's rs =
Rangkorrelationskoeffizient nach Spearman

Interpretation des Betrags von x

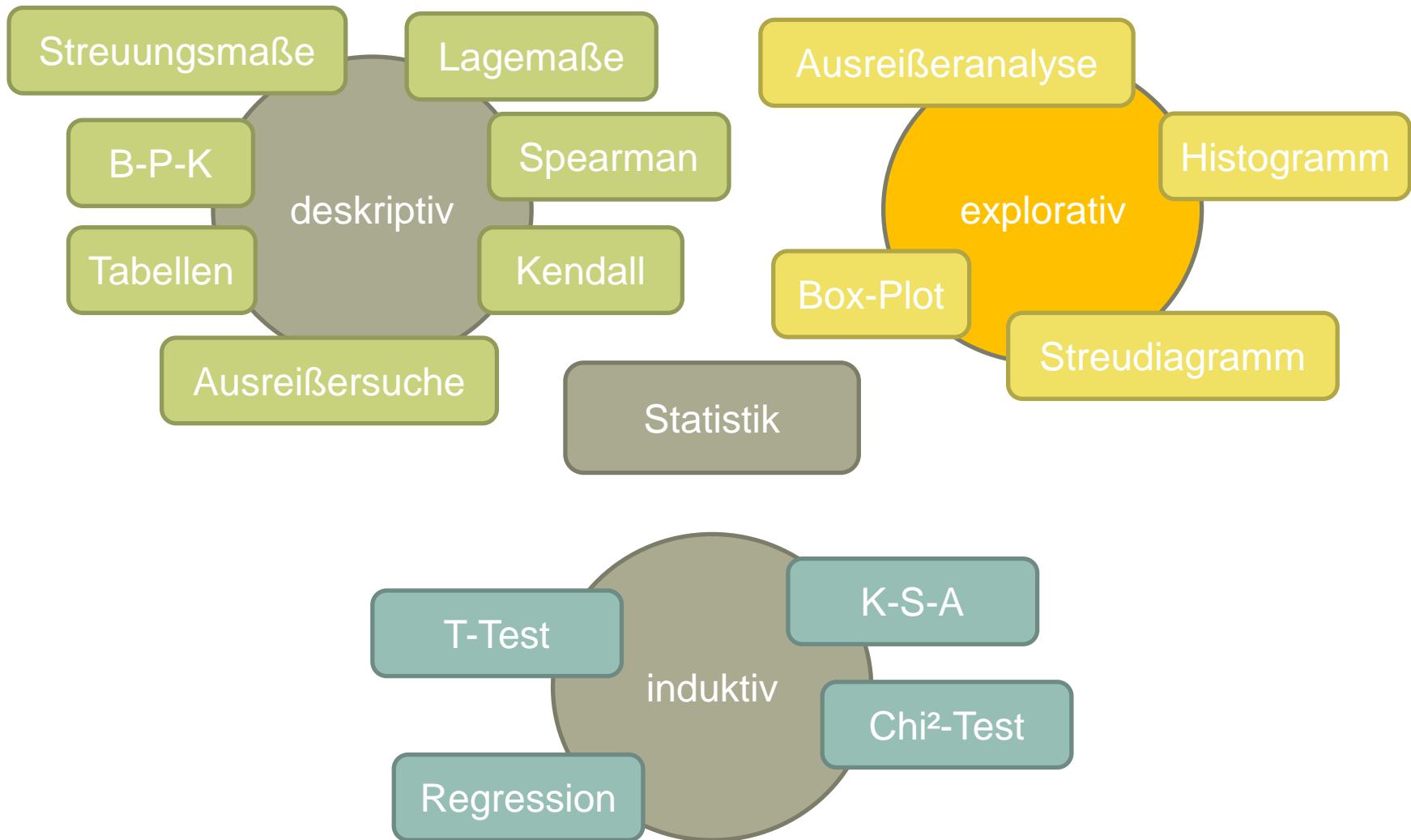
x = 0 = keine Korrelation
0 < x < 0,5 = schwache Korrelation
0,5 ≤ x < 0,8 = mittlere Korrelation
0,8 ≤ x < 1 = starke Korrelation
x = 1 = perfekte Korrelation

Korrelation ist nicht gleich Kausalität

- Eine über einen Korrelationskoeffizienten identifizierte Korrelation sollte näher untersucht, dabei jedoch **niemals inhaltlich interpretiert werden**
- Grund dafür ist, dass eine Korrelation nicht notwendigerweise auf einem Ursache-Wirkungs-Zusammenhang beruht – auch wenn es in vielen Fällen leider äußerst verführerisch ist, diese Annahme zu treffen
- Tatsächlich kann es verschiedene Erklärungen für Korrelationen geben
 - Einseitiger Zusammenhang: X beeinflusst Y bzw. Y beeinflusst X
 - Beidseitiger Zusammenhang: X und Y beeinflussen sich gegenseitig
 - Es handelt sich um einen reinen Zufallseffekt in den Daten (Scheinkorrelation)
 - Eine dritte Variable (Z) beeinflusst X und Y gleichermaßen (Scheinkorrelation)
- Ein klassisches Beispiel für eine Scheinkorrelation ist die Korrelation zwischen Storchenzahl und Geburtenquote (verbunden über die Variable „Urbanisierung“)

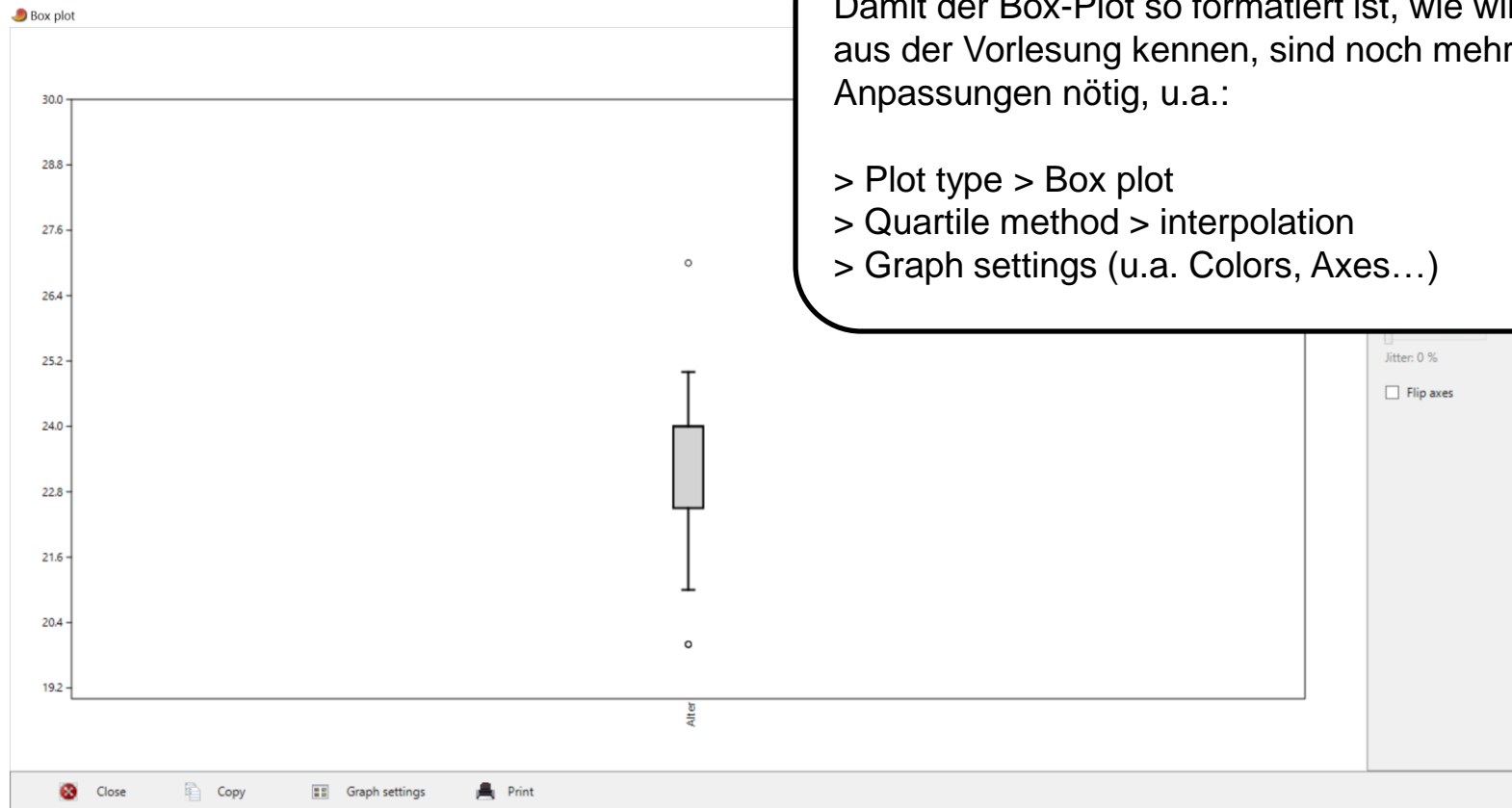


Wo befinden wir uns?



Erstellung eines Box-Plots

> Plot > Barchart/Boxplot



Erstellung vergleichender Box-Plots (nach Erweiterung des Datensatzes)

Altersgruppen.dat

File

Edit

Transform

Plot

Univariate

Multivariate

Model

Diversity

Timeseries

Geometry

Stratigraphy

Script

Help

Show

☐ Row attributes


☐ Column attributes

Click mode


☒ Select

☐ Drag rows/columns


Edit




Cut



Copy



Paste



Select all

View

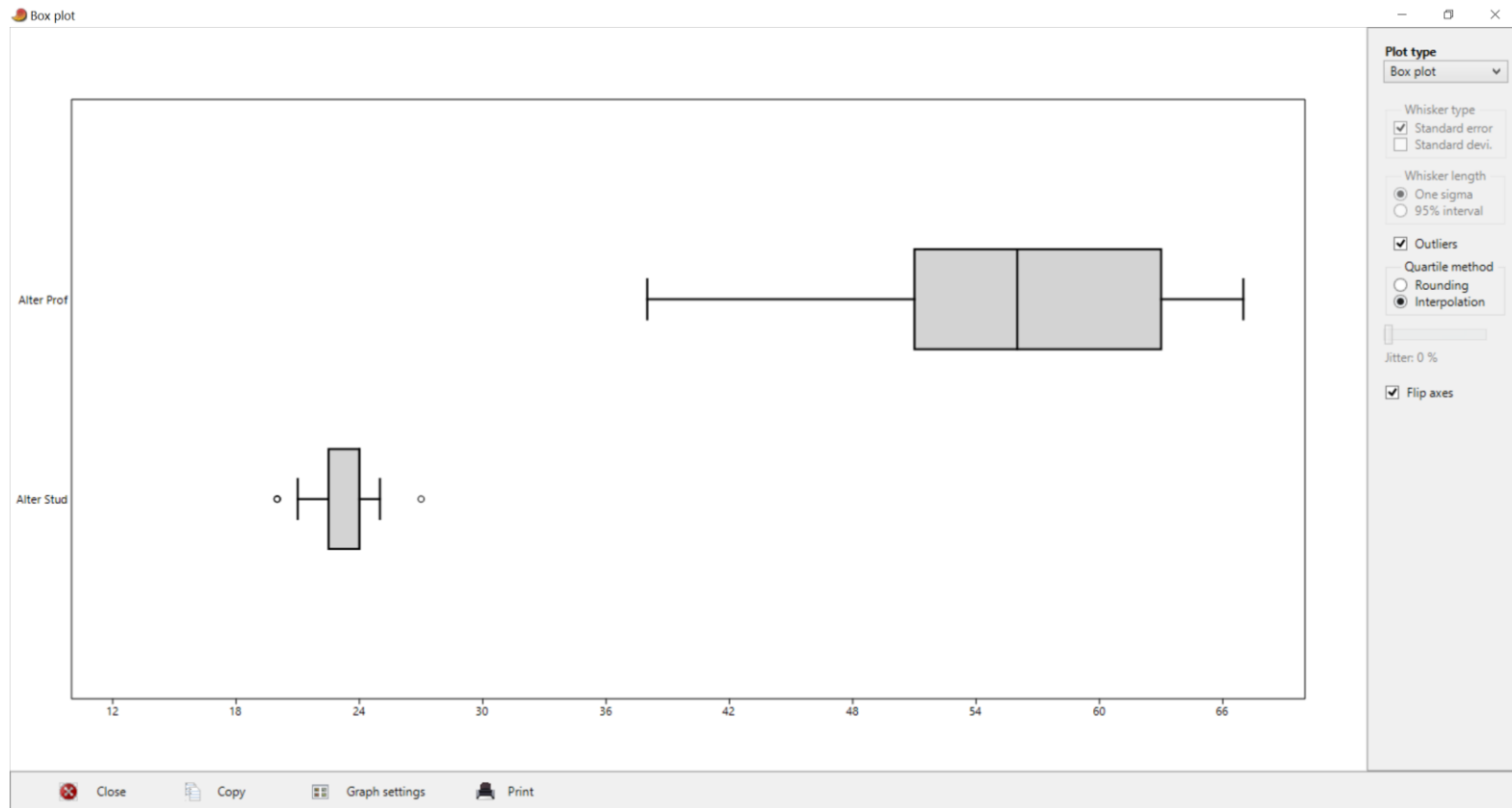
☐ Bands

☐ Black/white (1/0)

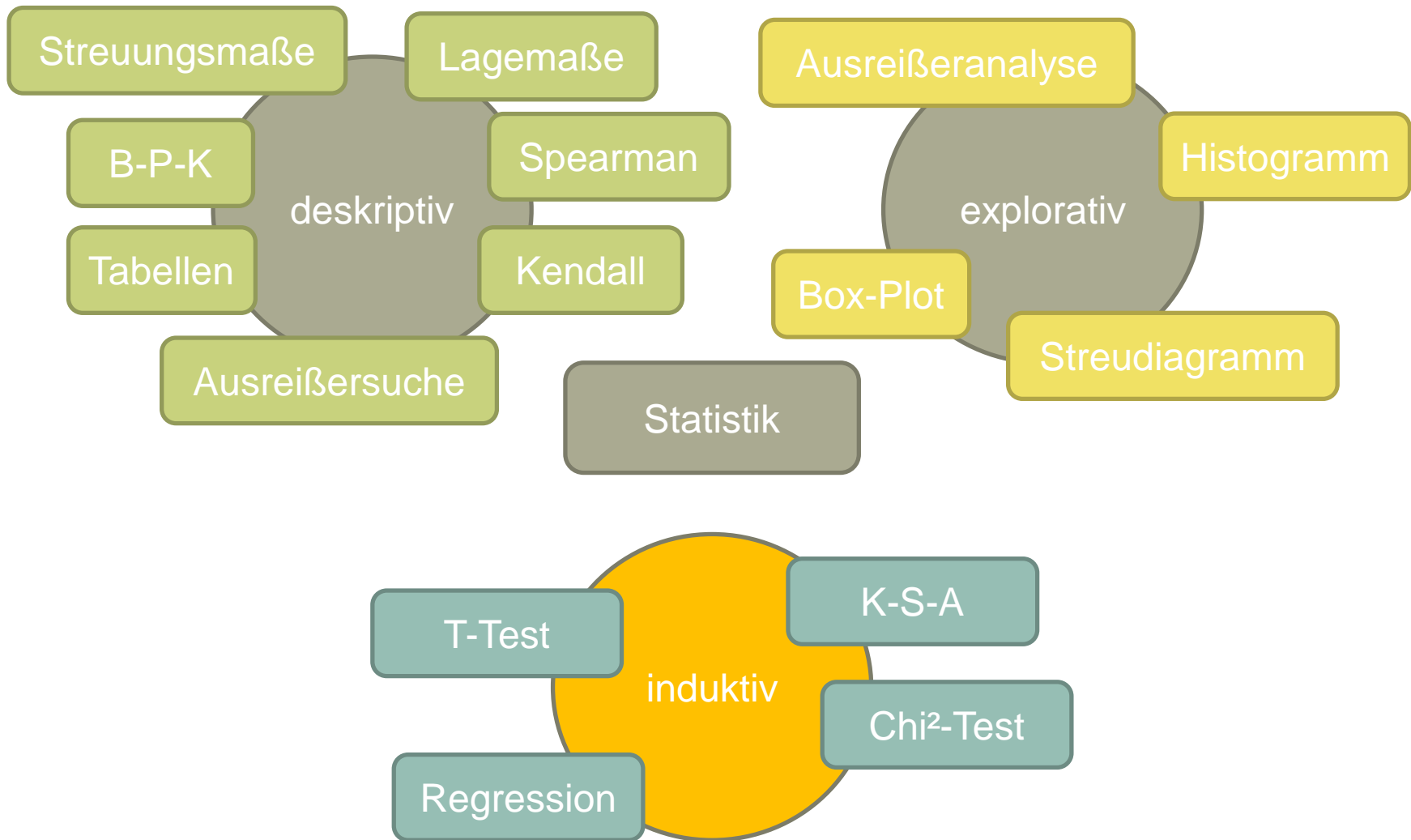
	Alter Studis	Alter Profs	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	• 20	38															
2	• 20	43															
3	• 20	44															
4	• 21	49															
5	• 21	49															
6	• 22	49															
7	• 23	53															
8	• 23	54															
9	• 23	54															
10	• 24	54															
11	• 24	56															
12	• 24	56															
13	• 24	56															
14	• 24	58															
15	• 24	58															
16	• 24	59															
17	• 24	59															
18	• 24	59															
19	• 24	62															
20	• 24	64															
21	• 24	64															
22	• 24	64															
23	• 25	66															
24	• 25	66															
25	• 27																
26	•																
27	•																
28	•																
29	•																
30	•																
31	•																

Erstellung vergleichender Box-Plots

> Plot > Barchart/Boxplot



Wo befinden wir uns?



Beispieldatensatz zur linearen Regression

Nr.	x	y
1	12	10000
2	15	15000
3	8	6000
4	11	11000
5	3	5000
6	17	23000
7	24	37000

Beispielfall mit bewusst gering gehaltener (Foliendarstellung...) Anzahl von Werten:

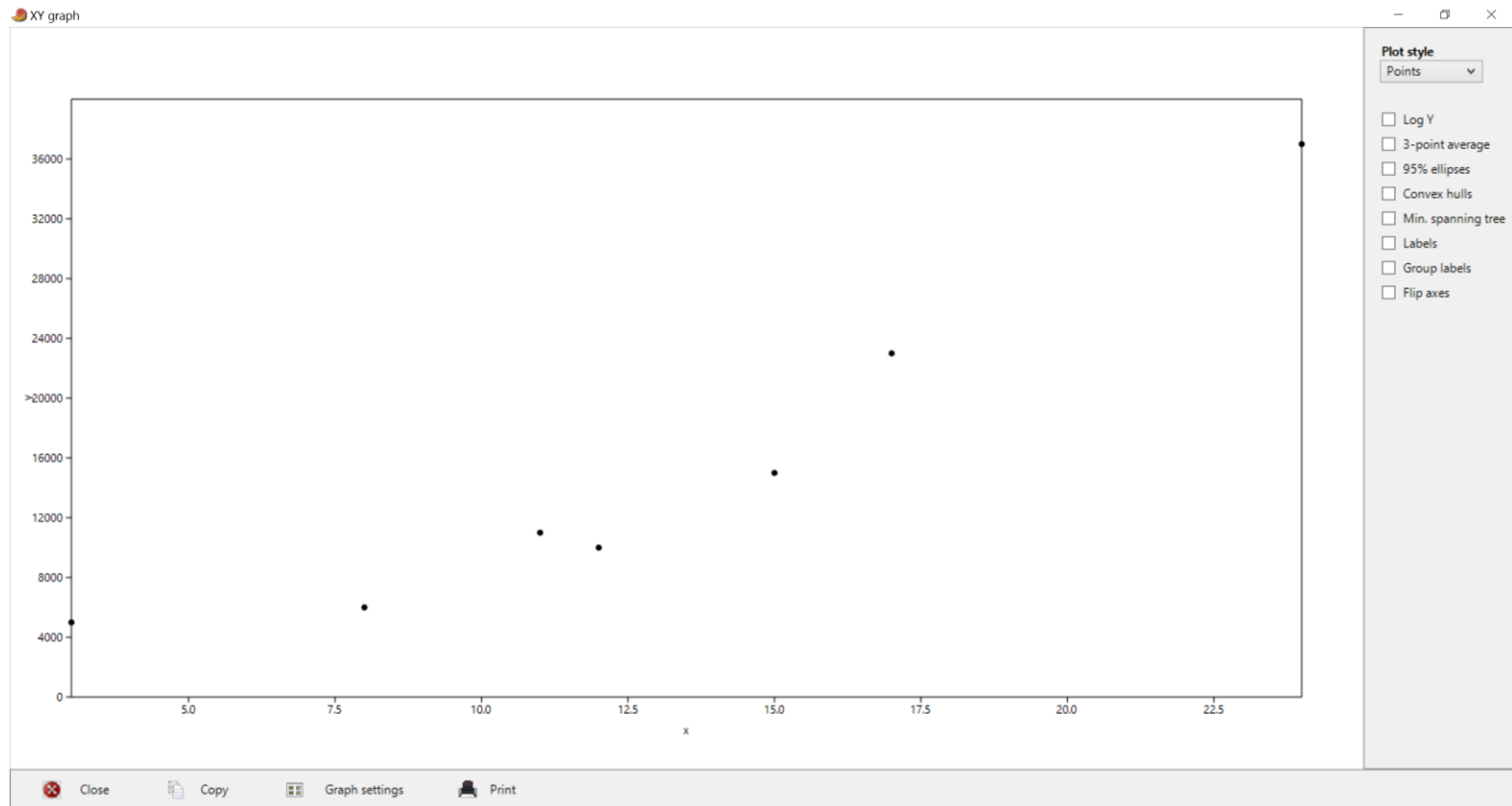
- x = Prozentualer Anteil des Werbebudgets eines Produkts am Gesamtbudget der Firma
- y = Verkaufte Einheiten des betrachteten Produkts in einem Untersuchungszeitraum
- Annahme: Das betrachtete Produkt, der Untersuchungszeitraum sowie das Gesamtbudget bleiben gleich

(ceteris paribus)

Wie lautet die Regressionsgleichung?

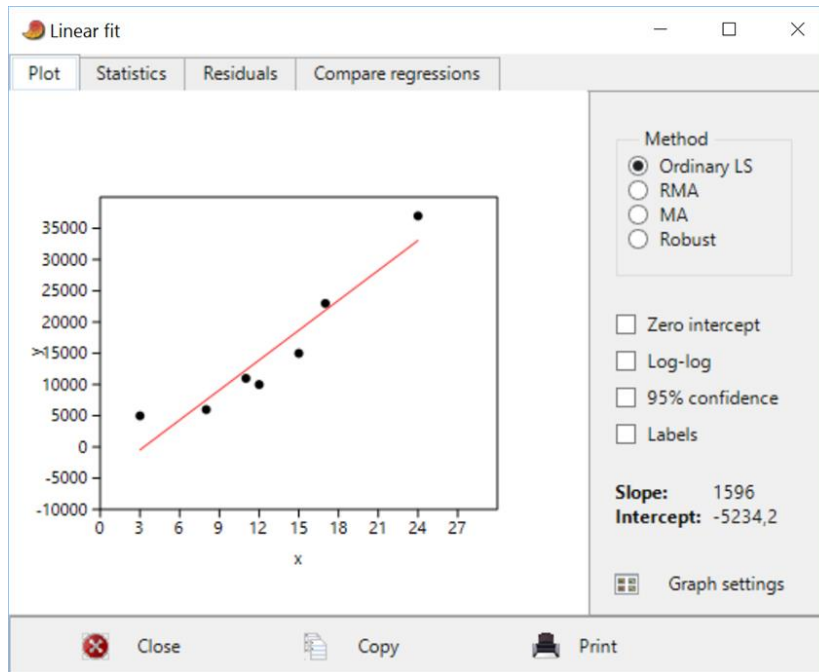
Ist ein Zusammenhang grafisch plausibel?

> Plot > XY graph

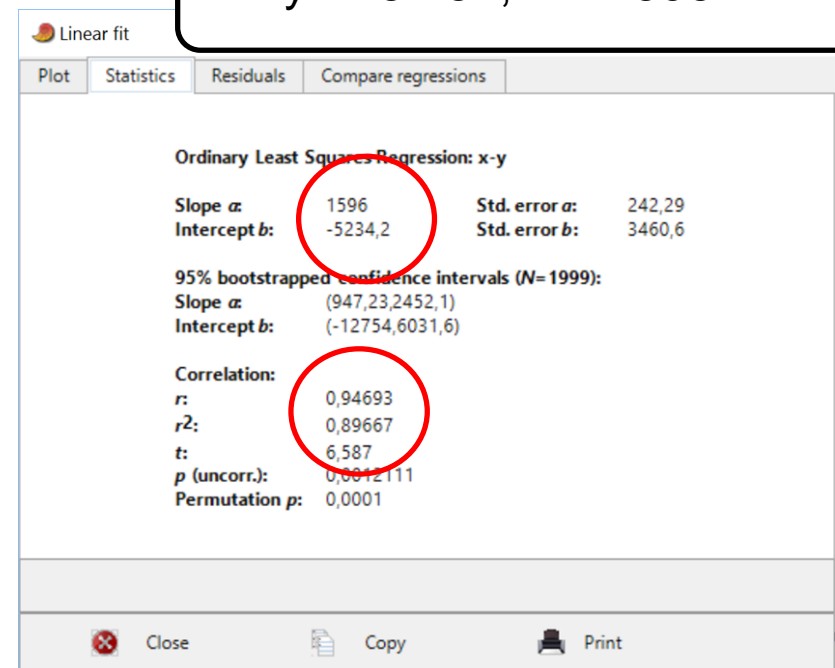


Erstellung und Bewertung des LR-Modells

> Model > Linear > Bivariate



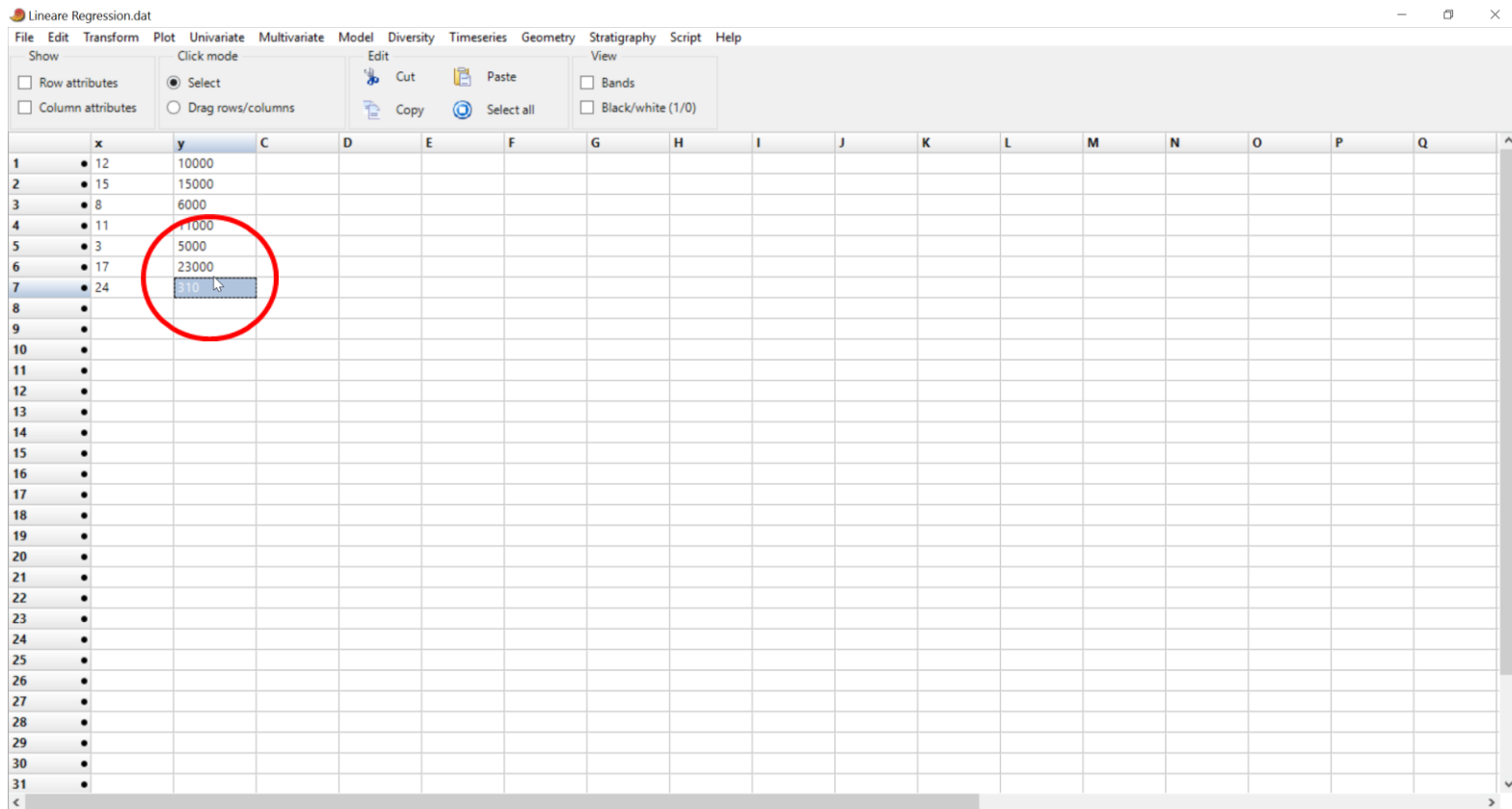
$$y = -5.234,2 + 1.596 * x$$



Was ist hier was?

Slope = Steigung (Regressionskoeffizient)
Intercept = Schnittpunkt mit der y-Achse
 r^2 = Bestimmtheitsmaß / Gütekriterium

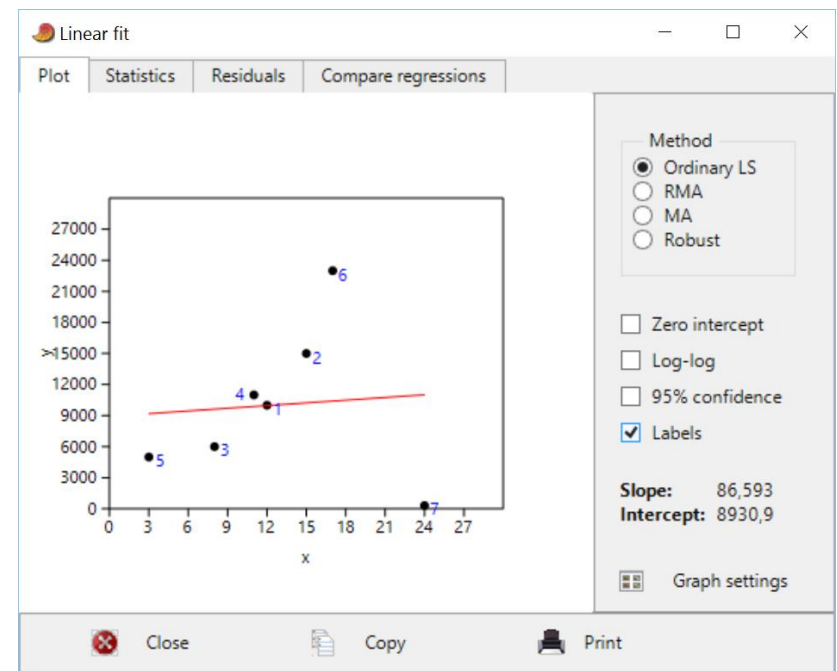
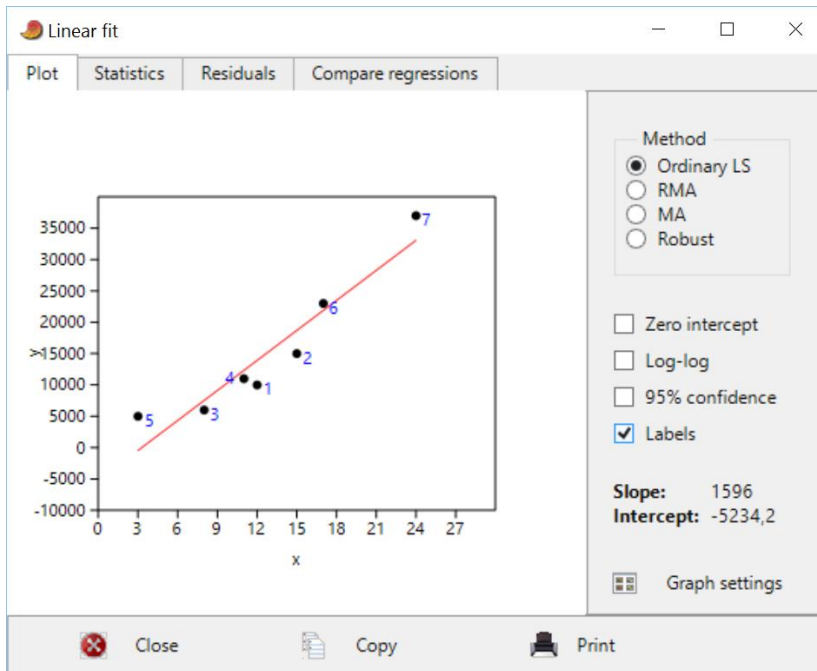
Sichtbarmachung des Leverage-Effekts (Was eine kleine Änderung bewirken kann...)



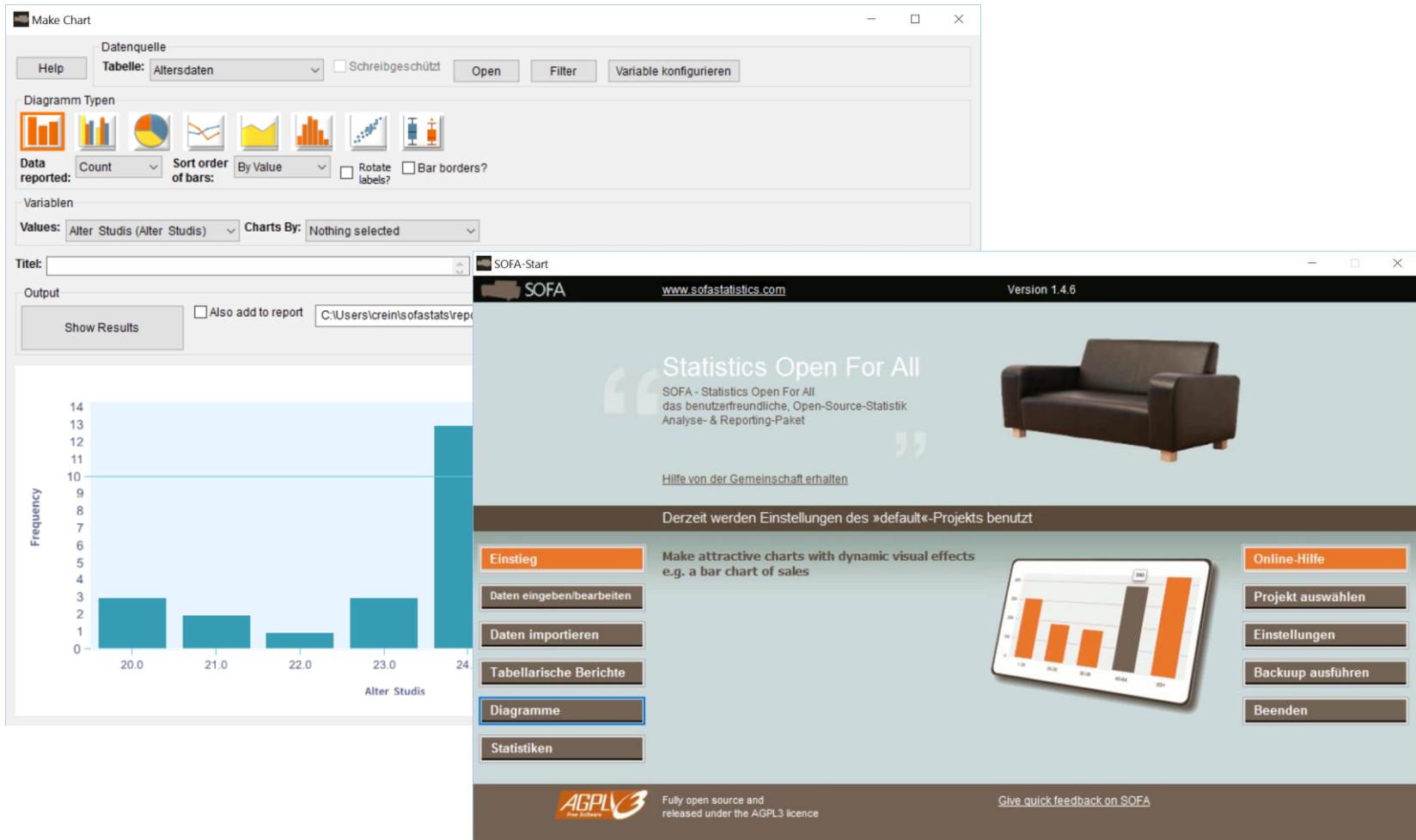
Sichtbarmachung des Leverage-Effekts

> Model > Linear > Bivariate

Wie deutlich verschlechtert sich hier r^2 ?

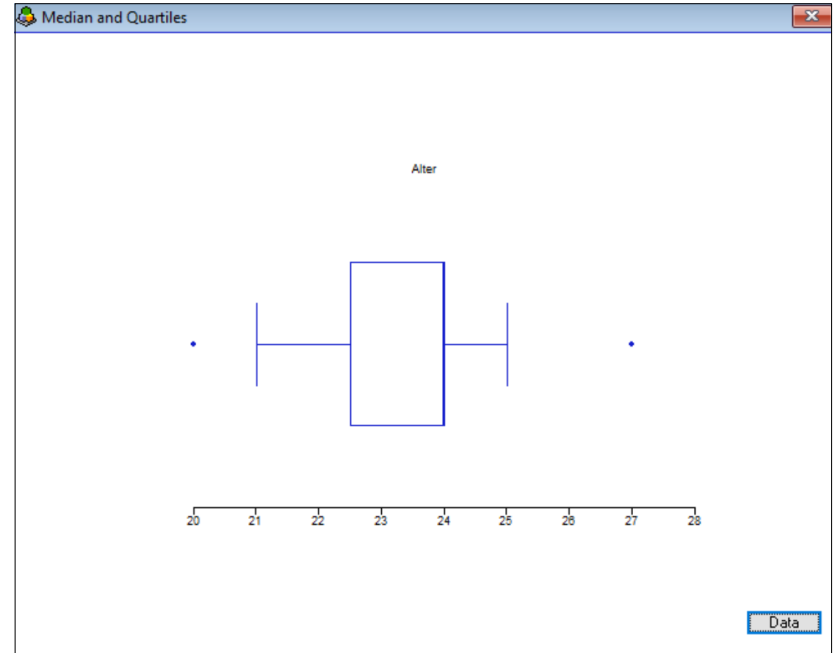
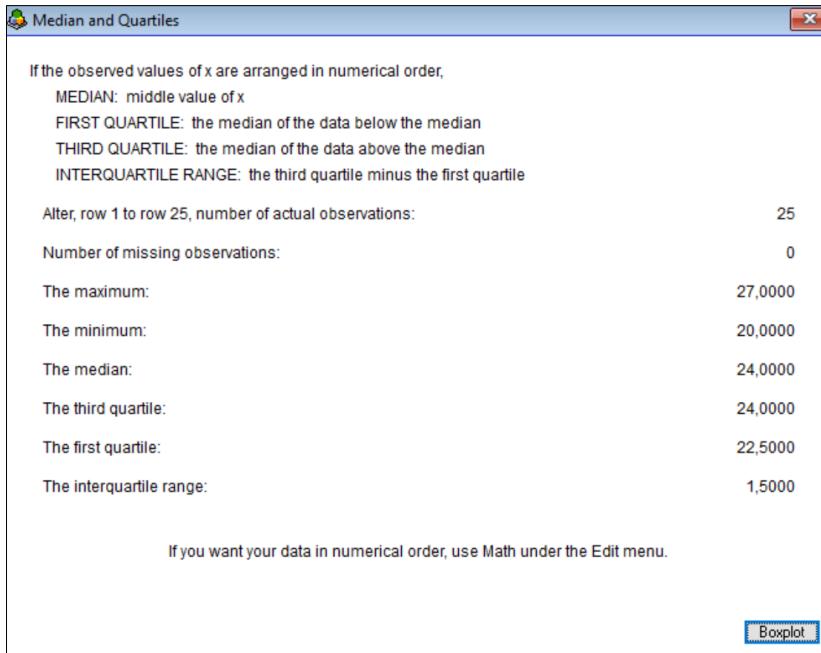


Was kann andere (freie) Software (besser)?



Erstellung von Box-Plots mit SSP

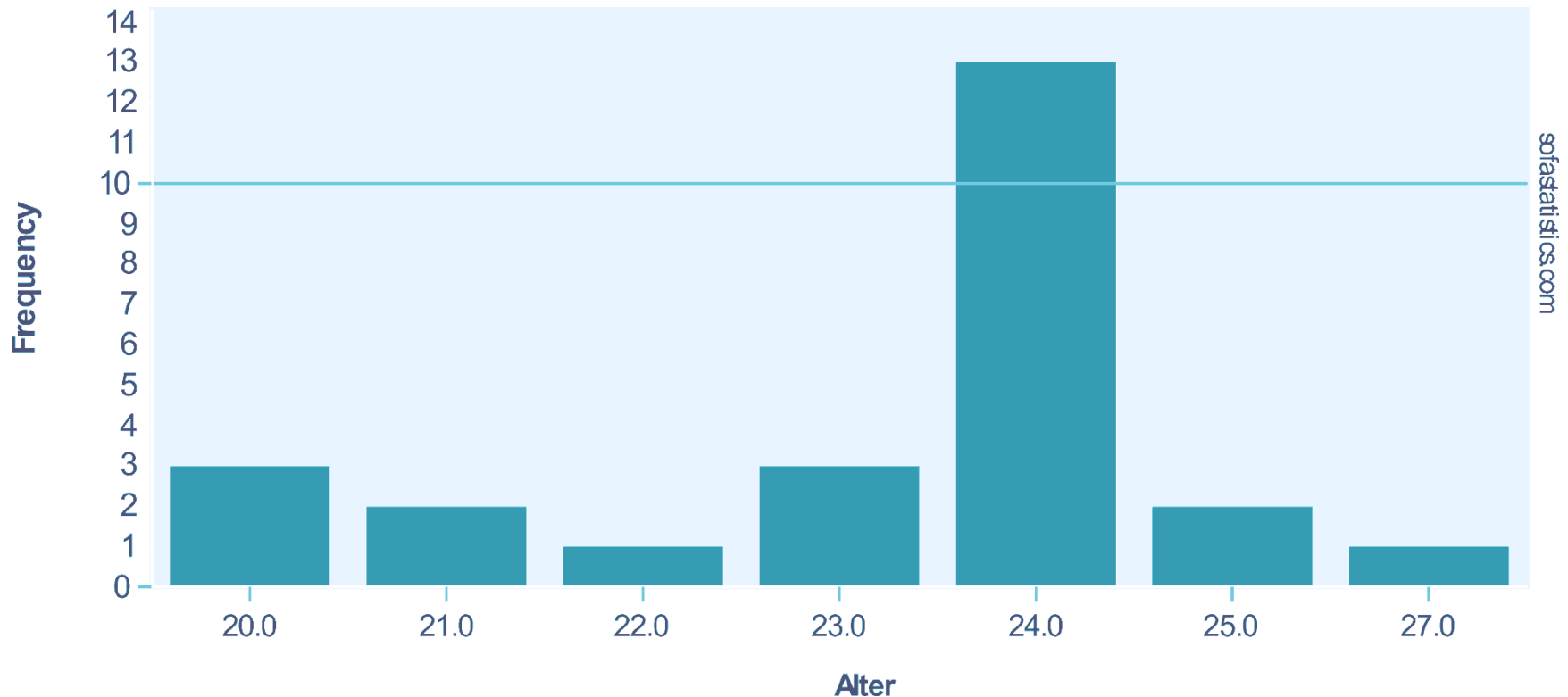
> Describing Data > Median, Quartiles > Box-Plot



Schöne Übersicht der Konstruktionsgrößen – weniger ansehnlicher Box-Plot

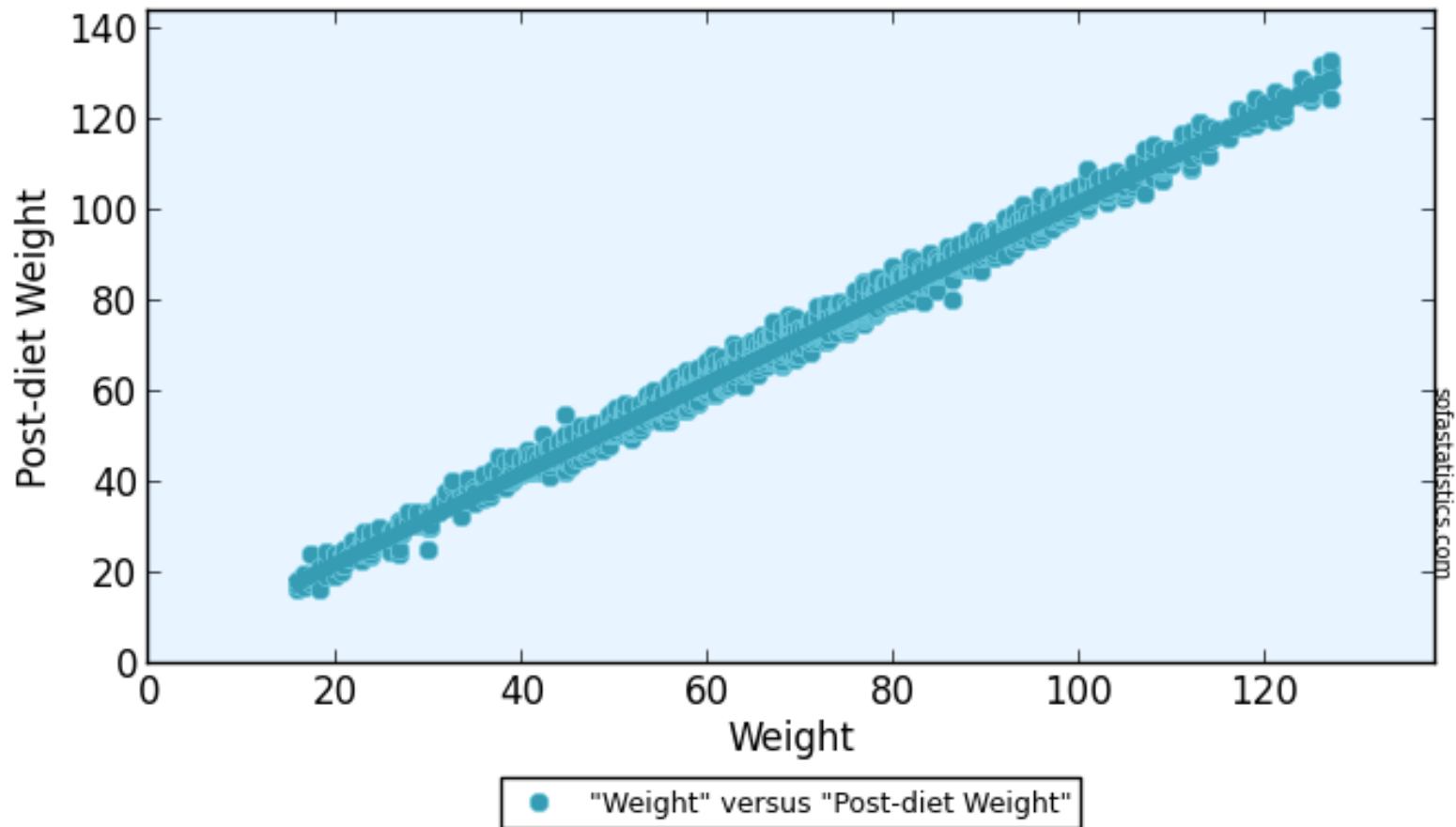
Erstellung „schöner“ Grafiken mit SOFA

> Diagramme > Balkendiagramm erstellen



Erstellung „schöner“ Grafiken mit SOFA

> Diagramme > Scatterplot erstellen



Detailergebnisse der Regression in PSPP

> Analysieren > Regression > Linear

Ziel — PSPP:RE Ausgabeanzeige

Datei Bearbeiten Fenster Hilfe

REGRESSION

REGRESSION
/VARIABLES= x
/DEPENDENT= y
/METHOD=ENTER
/STATISTICS=COEFF CI R ANOVA BCOV.

Modellzusammenfassung (y)

R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
.95	.90	.88	3987,59

ANOVA (y)


	Quadratsumme	df	Mittel der Quadrate	F	Sig.
Regression	689924352,02	1	689924352,02	43,39	,001
Residual	79504219,41	5	15900843,88		
Gesamt	769428571,43	6			

Koeffizienten (y)

	Unstandardisierte Koeffizienten		Standardisierte Koeffizienten	t	Sig.	95% Konfidenzintervall für B	
	B	Standardfehler	Beta			Untere Grenze	Obere Grenze
(Konstante)	-5234,18	3460,63	,00	-1,51	,181	-14130,01	3661,65
x	1595,99	242,29	,95	6,59	,001	973,16	2218,82

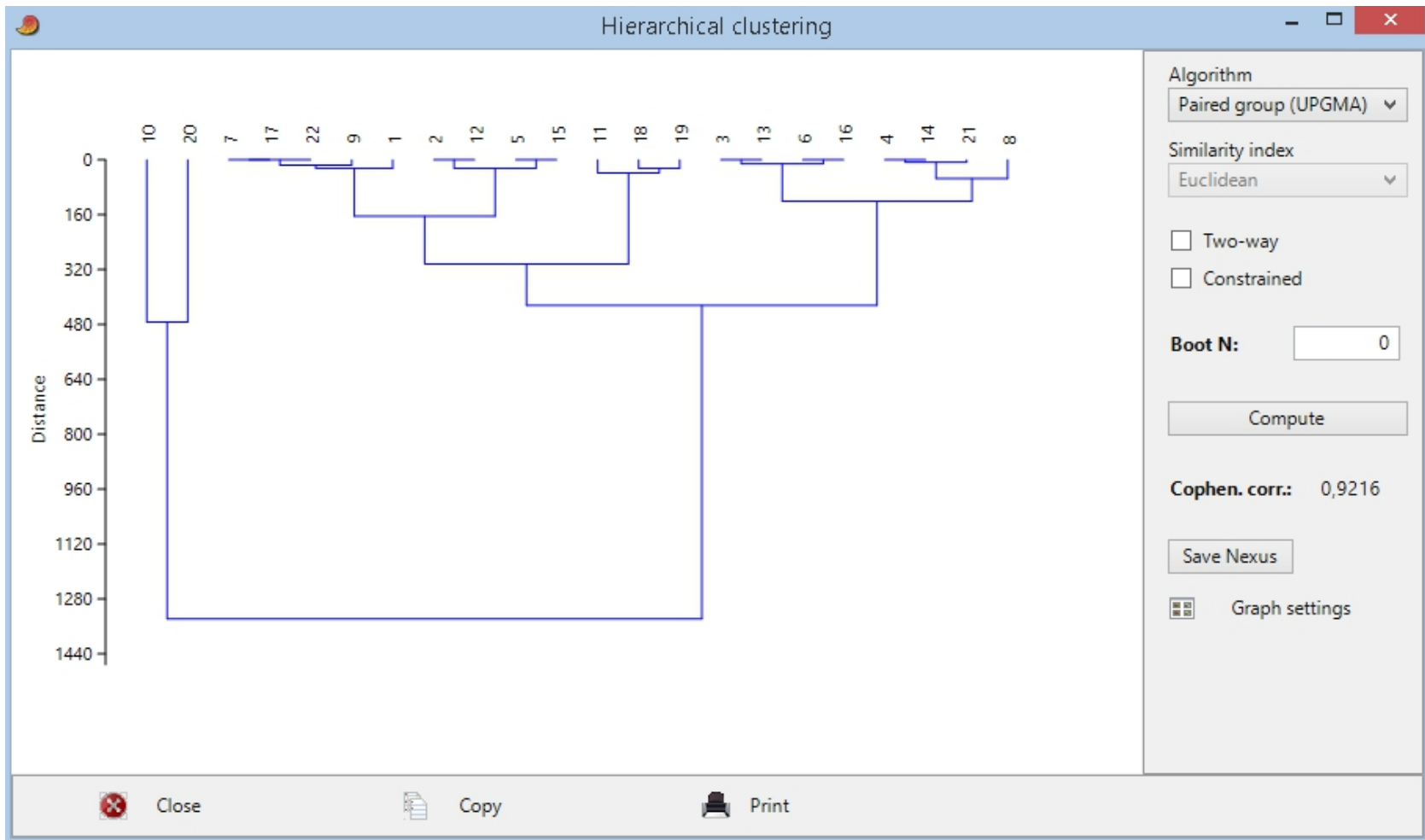
Korrelationen der Koeffizienten (y)

Modell	x
Kovarianzen	



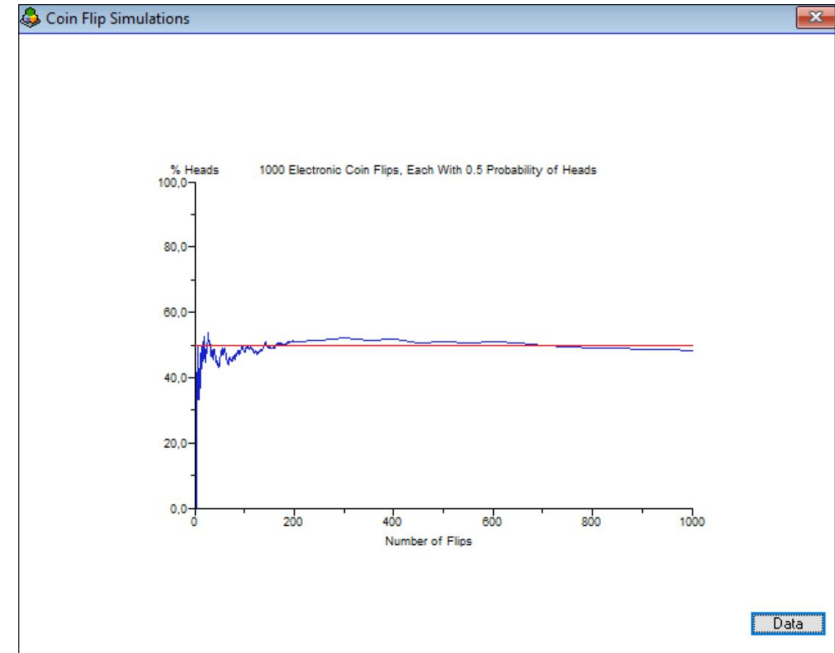
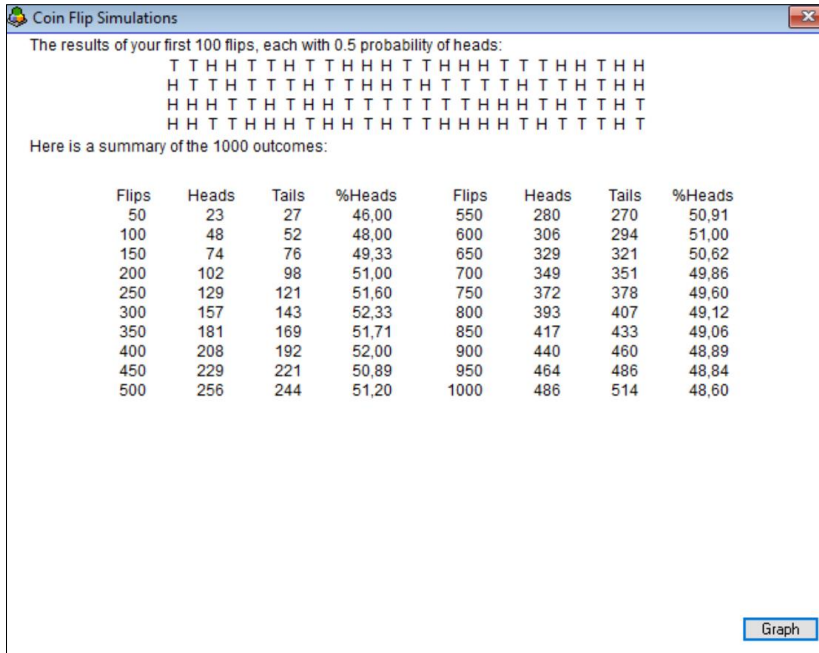
Tipp: PSPP eignet sich für die Vorbereitung der Statistik II-Klausur in besonderer Weise

Kleine Vorschau auf Statistik II



Simulation von Münzwürfen in SSP

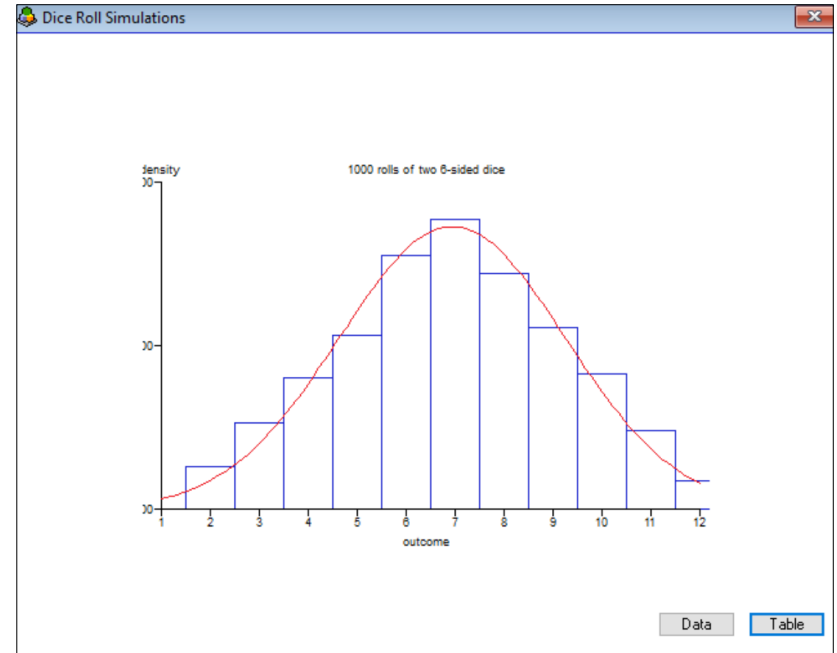
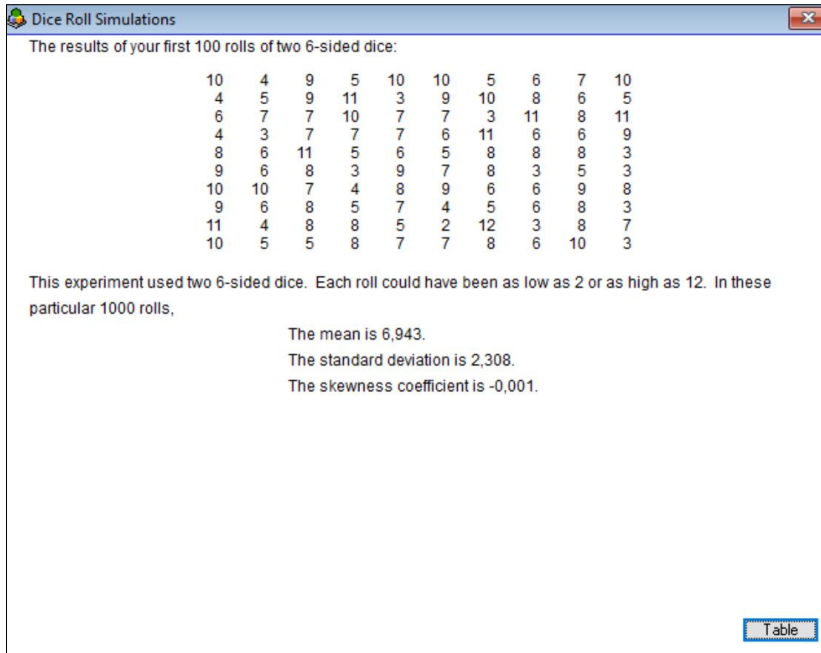
> Uncertainty > Coin Flip Simulation



Gesetz der Großen Zahlen: Die relative Häufigkeit eines Zufallsergebnisses stabilisiert sich um die theoretische Wahrscheinlichkeit eines Zufallsergebnisses, wenn das zu Grunde liegende Zufallsexperiment immer wieder unter denselben Voraussetzungen durchgeführt wird.

Simulation von Würfelwürfen in SSP

> Uncertainty > Dice Roll Simulation



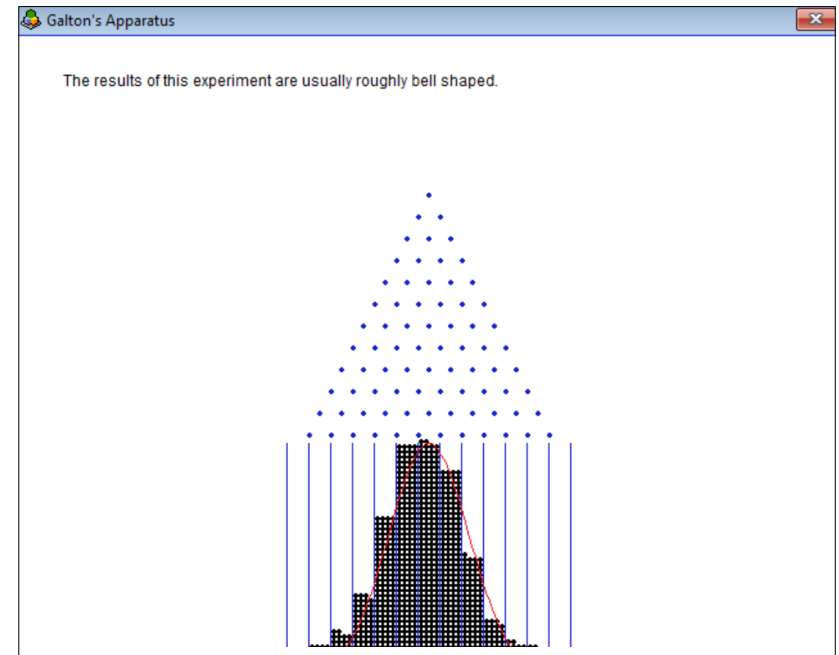
Gesetz der Großen Zahlen: Die relative Häufigkeit eines Zufallsergebnisses stabilisiert sich um die theoretische Wahrscheinlichkeit eines Zufallsergebnisses, wenn das zu Grunde liegende Zufallsexperiment immer wieder unter denselben Voraussetzungen durchgeführt wird.

Simulation eines Galtonbretts in SSP

> Uncertainty > Galton's Apparatus



Foto: Klaus-Dieter Keller; Lizenz: gemeinfrei; Quelle: Wikimedia



Mit Hilfe eines Galtonbretts lässt sich visuell demonstrieren, warum viele Zufallsvariablen der Binomialverteilung folgen.

Bestimmung der optimalen Stichprobengröße

$$n = \frac{\frac{Z^2 * p * q}{e^2}}{1 + \frac{\frac{Z^2 * p * q}{e^2} - 1}{N}}$$

- Was passiert bei....
 - größerer Grundgesamtheit?
 - kleinerer Grundgesamtheit?
 - bekannten Anteilswerten?
 - weniger Sicherheit?
 - mehr Sicherheit?

SampleSizer 1.2

Menü

Grundgesamtheit: 20000

Stichprobenanteil: 0,5

Wenn nicht bekannt p = 0,5 (50%-Schätzer)

Intervallbreite (+/-): 0,03

Die Breite muss im Format 0,0x angegeben werden

Bei einer Sicherheit des Konfidenzintervalls von 95%:

Stichprobengröße: 1015

<http://www.statistikberatung.eu>

Kostenloser Download unter:
[http://www.statistikberatung.eu/
SampleSizer.zip](http://www.statistikberatung.eu/SampleSizer.zip)

Einführung in freie statistische Software

**Vielen Dank für die
Aufmerksamkeit!**

Christian Reinboth

Telefon +49 3943 – 896

Telefax +49 3943 – 5896

E-Mail creinboth@hs-harz.de

Friedrichstraße 57 – 59

38855 Wernigerode