Beat Eicke

Grundlagen der Datenanalyse anhand praktischer Beispiele

29. Oktober 2016

FHNW Windisch

Ziele

- Anhand von Beispielen die gemäss RLP zu behandelnden Themen vorstellen:
  - Datenerhebung
  - Grundlagen
  - Diagramme
  - Masszahlen
- Hinweise auf Einsatzmöglichkeiten von TI-Nspire CX CAS geben; Genaueres dazu folgt im Workshop.

Datenerhebung (Meinungsumfragen) I

- Was denken die in der Schweiz lebenden gut N=8’000’000 Menschen (Grundgesamtheit) zu einem bestimmten Thema?
- Befragt werden wegen des Aufwands tatsächlich nur n Leute, eine hoffentlich repräsentativen Auswahl (Stichprobe).

Wie viele Personen muss man befragen?

- Annahmen: Zufällige Auswahl der Befragten, und die Differenz zwischen den Meinungen der Grundgesamtheit und der Stichprobe soll mit 95% höchstens 3% betragen:

$$n \geq \frac{N \cdot 1067}{N + 1067} \approx \frac{N \cdot 1000}{N + 1000}$$

Datenerhebung (Meinungsumfragen) II

- Auswahl der befragten Personen durch Zufallsprinzip oder Quotenverfahren
- Wenige einfache, klare Fragen! Die Formulierung der Frage kann das Ergebnis beeinflussen.
- Strassenumfrage zu verschiedenen Zeiten an verschiedenen Orten, Postversand, Telefonumfrage, Internet…
- Was muss man über die Befragten wissen? (Alter, Wohnort, Religion, Partei, …)
- Die Befragten sind bei persönlichen Themen nicht immer ehrlich: Drogenkonsum, Wahlverhalten, sexuelle Präferenzen, kleinere oder grössere Vergehen usw.

Beispiel 1: Klausur in 3 Klassen

Urliste: Daten, gewonnen durch Messungen, Umfragen, …
hier: Noten der Studierenden in beliebiger Reihenfolge

Strichliste:

Note	Häufigkeit		
	Klasse A	Klasse B	Klasse C
6	II		II
5½	III		IIIII I
5	IIII	IIIII	III
4½	IIIII	III	III
4	III	IIII	II
3½	II	III	II
3	III	I	
2½	II		
2			I
1½			I
1	I		

Evtl. sortieren: Mit TI Nspire CX CAS: sorta, sortd

Beispiel 1: Auswertung I

Lagemasse sagen aus, in welcher Grössenordnung die Daten insgesamt liegen. Also: Welche Klasse ist insgesamt am besten?

- Mittelwert: berücksichtigt alle Werte; «Gesamtschau»
- Median: mittlerer Wert der sortierten Liste; «Mittelmass»
- Modus: häufigster Wert; «typischer Wert»
- 1. Quartil: wird von ¼ aller Werte unterschritten oder erreicht
- 3. Quartil: wird von ¾ aller Werte unterschritten oder erreicht

Streuungsmasse sagen aus, wie stark die Daten voneinander abweichen.

- Statistische Standardabweichung σ (Grundgesamtheit)
- Empirische Standardabweichung s (Stichprobe)
- Quartilsdifferenz: 3. Quartil – 1. Quartil

Beispiel 1: Auswertung II

Note	Häufigkeit		
	Klasse A	Klasse B	Klasse C
6	II		II
5½	III		IIIII I
5	IIII	IIIII	III
4½	IIIII	III	III
4	III	IIII	II
3½	II	III	II
3	III	I	
2½	II		
2			I
1½			I
1	I		

	Klasse A	Klasse B	Klasse C
Klassengrösse	25	16	20
Median	4.5	4.25	**5**
Modus	4.5	5	**5.5**
Mittelwert	4.2	4.25	**4.6**
1. Quartil	3.25	3.75	**4**
3. Quartil	5	5	**5.5**
Stat. Stand. σ	1.2	0.637	1.2
Emp. Stand. s	1.225	0.658	1.231
Quartilsdiff.	1.75	1.25	1.5

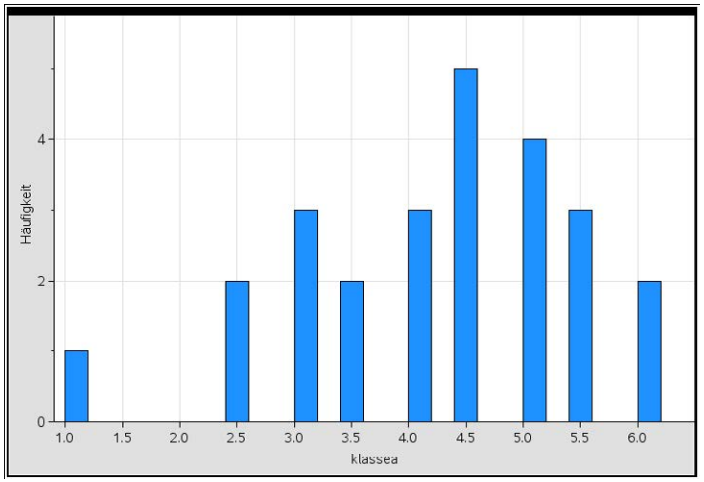
Beispiel 1: Auswertung III

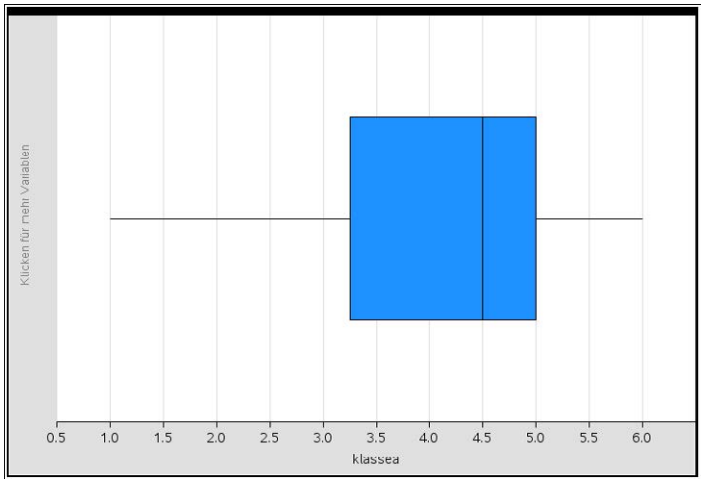
Gesamtauswertung (ohne den Modus) für Klasse A mit TI-Nspire CX CAS:
onevar und stat.results

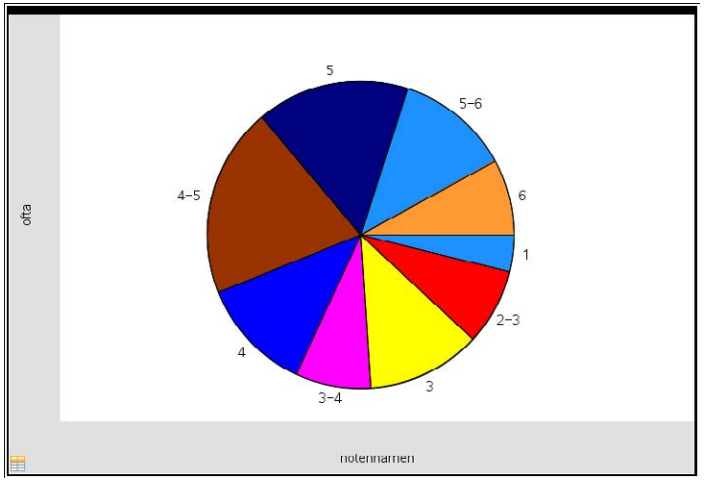
Spezielle Befehle:
mean, median, stdevsamp, stdevpop

"Titel"	"Statistik mit einer Variable"
"$\bar{x}$"	4.2
"Σx"	105.
"Σx^2"	477.
"$sx := sn\text{-}1x$"	1.22474
"$\sigma x := \sigma nx$"	1.2
"n"	25.
"MinX"	1.
"Q_1X"	3.25
"MedianX"	4.5
"Q_3X"	5.
"MaxX"	6.
"$SSX := \Sigma(x-\bar{x})^2$"	36.

Beispiel 1: Graphische Darstellungen







Beispiel 2: Steuerbares Einkommen

Die Spitzenreiter unter den Gemeinden (2014):

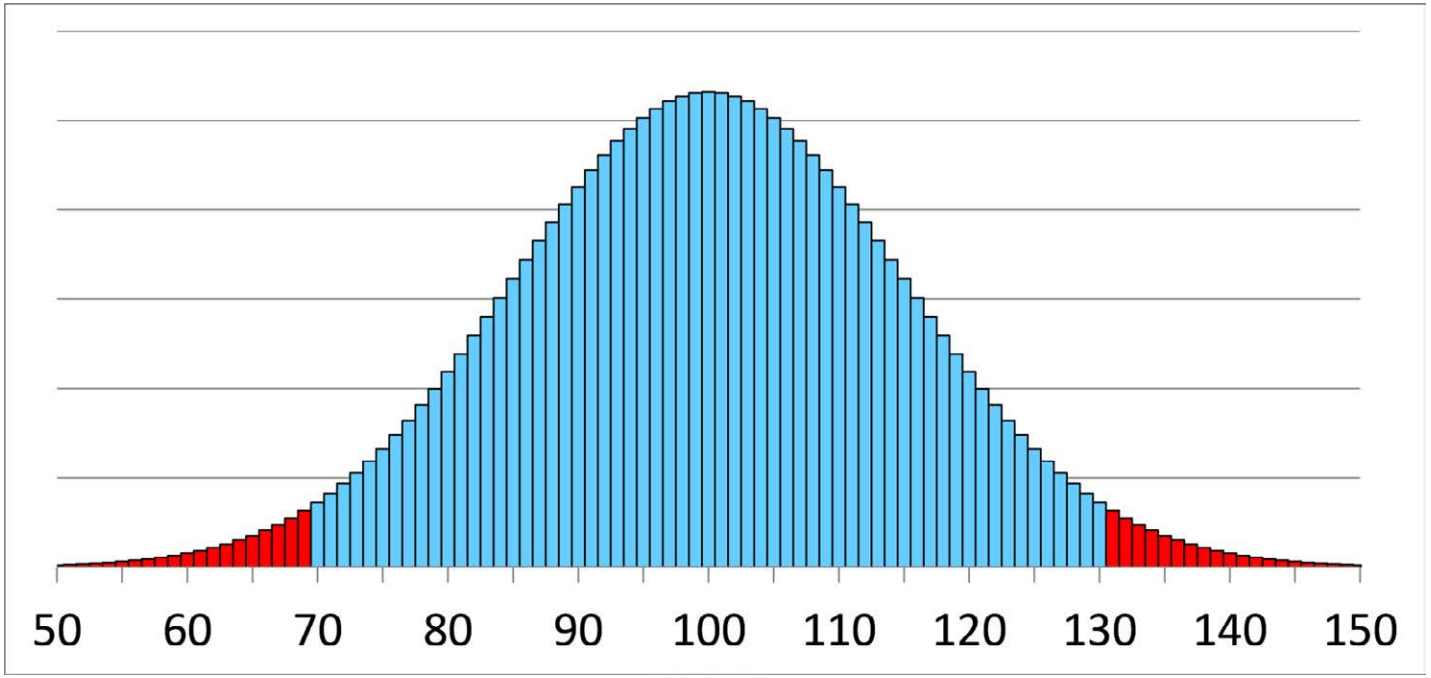
	Mittelwert [CHF]	Median [CHF]
Rüschlikon ZH	536’056	70’100
Vaux-sur-Morges VD	533’312	64’100
Walchwil ZG	494’455	75’250

Welches Lagemass sagt mehr über das Einkommen des «normalen» Einwohners dieser kleinen Gemeinden aus?

Spitzenreiter beim Median:
Chavannes-des-Bois (VD) 93’450 CHF

Beispiel 3: Intelligenztests

Intelligenztests sind geeicht: Mittelwert 100, Standardabw. 15.



Abweichungen von mehr als 2 Standardabweichungen vom Mittelwert kommen selten vor (IQ<70 oder IQ>130).

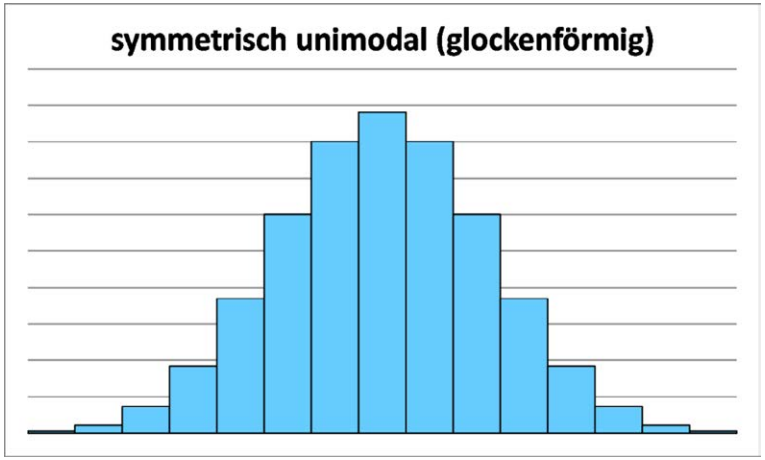
Beispiel 4: Plausibilitätskontrolle

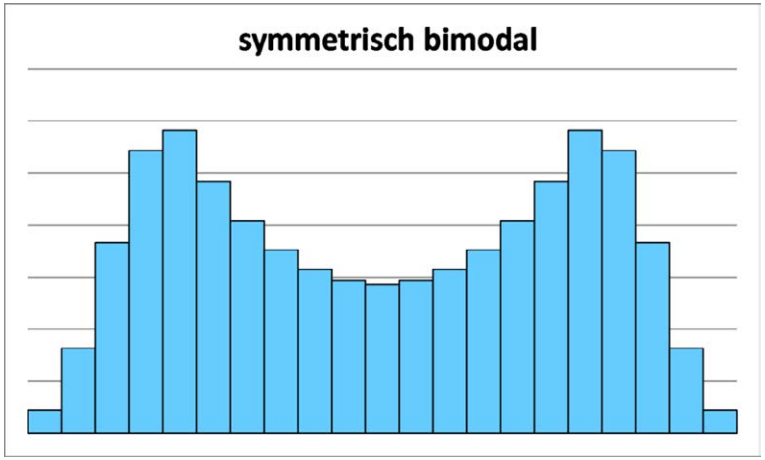
Abweichungen von mehr als 2 Standardabweichungen vom Mittelwert kommen selten vor. Dies ermöglicht eine Kontrolle:

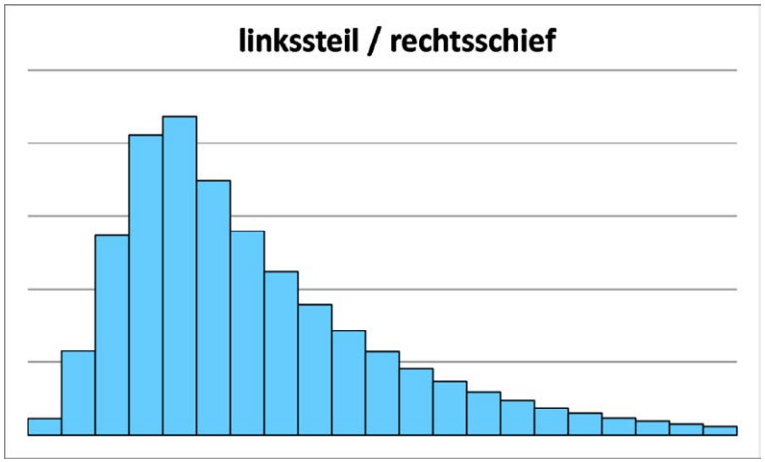
- Unterlief bei der Dateneingabe ein Tippfehler (IQ 39 statt 93)?
- Liegt ein Messfehler vor?
- Wurde das Experiment richtig durchgeführt?
- Liegt ein Ausreisser vor?

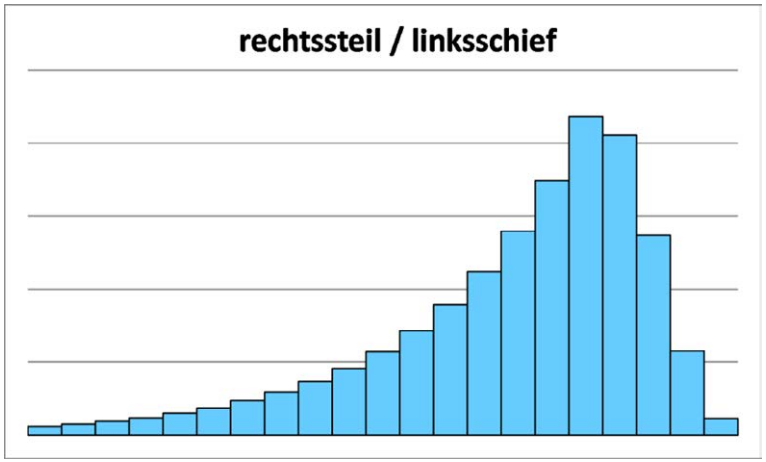
Führt man ein Experiment mehrmals unter gleichen Bedingungen durch, sollten die Messergebnisse nicht allzu verschieden sein. Eine grosse Standardabweichung kann auf Fehler bei der Durchführung hinweisen.

Diagrammtypen I



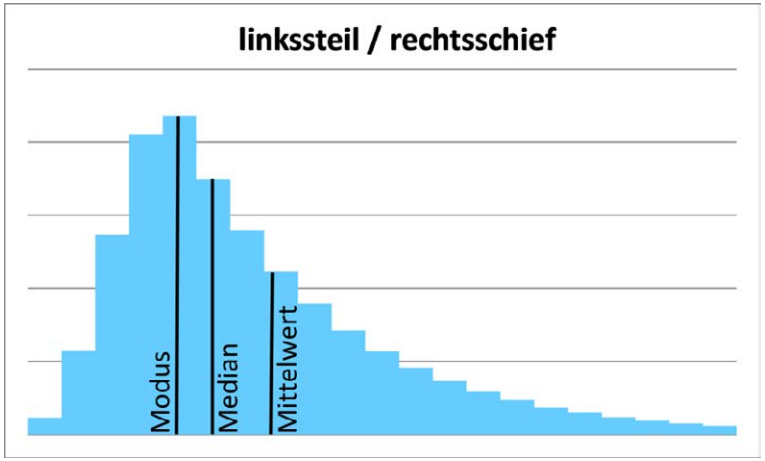






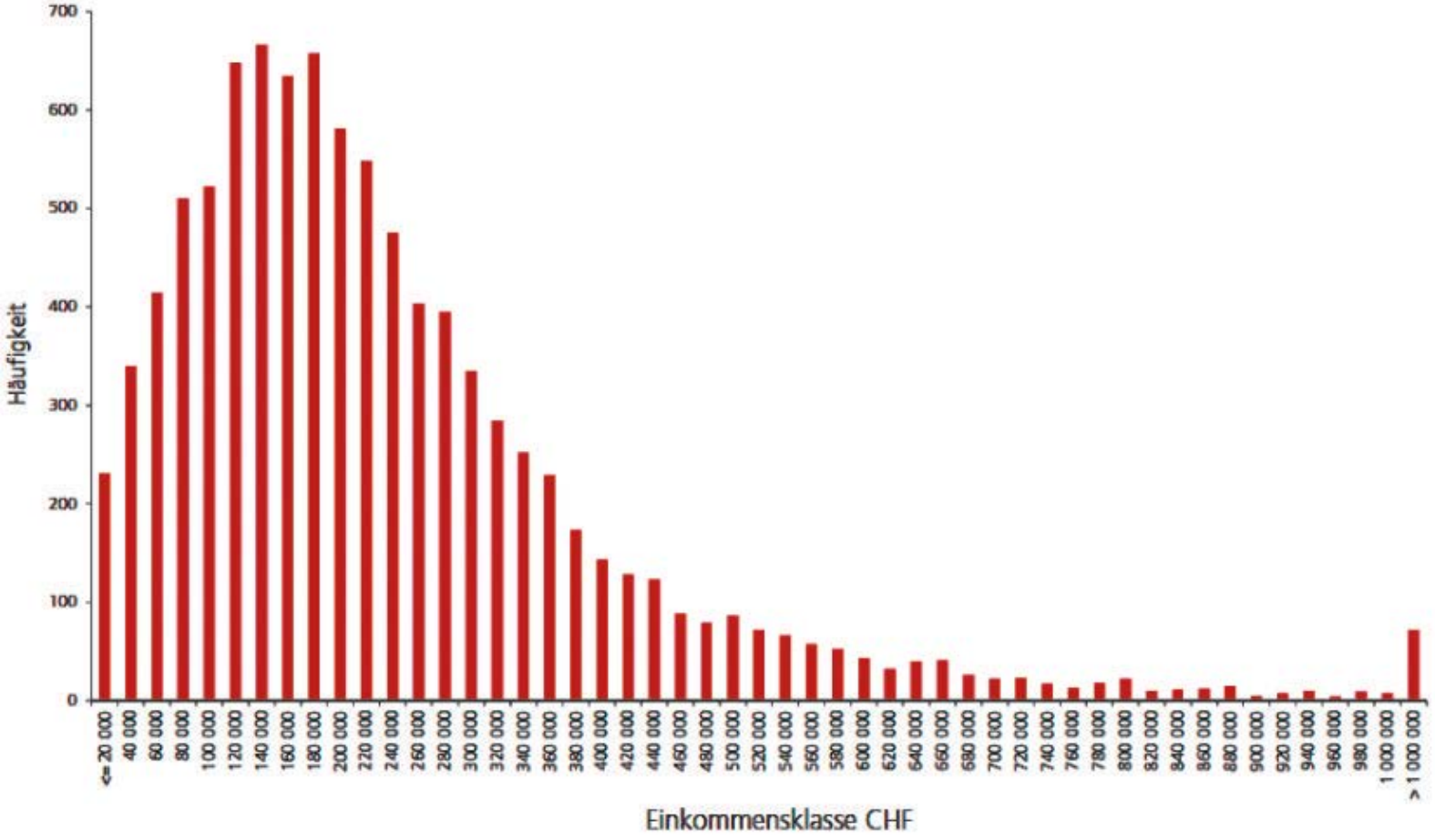
Diagrammtypen II

- symmetrisch unimodal, glockenförmig):
  Mittelwert = Median = Modus
- symmetrisch bimodal:
  Mittelwert = Median; die beiden Modi weichen davon ab
- rechtsschief:
  Modus < Median < Mittelwert
- linksschief:
  Mittelwert < Median < Modus

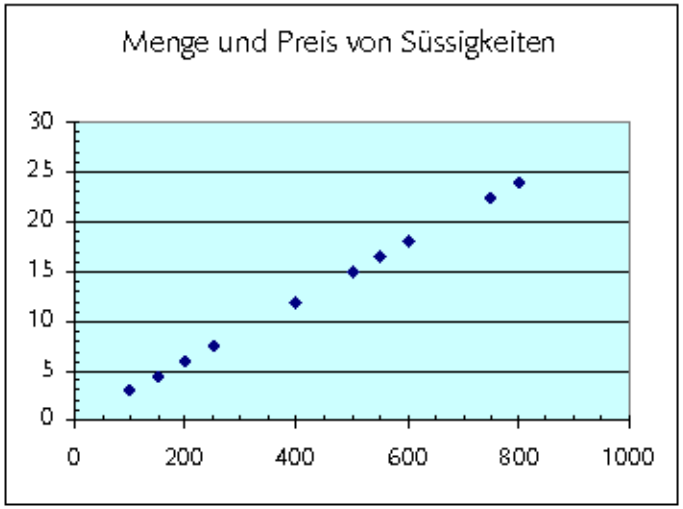


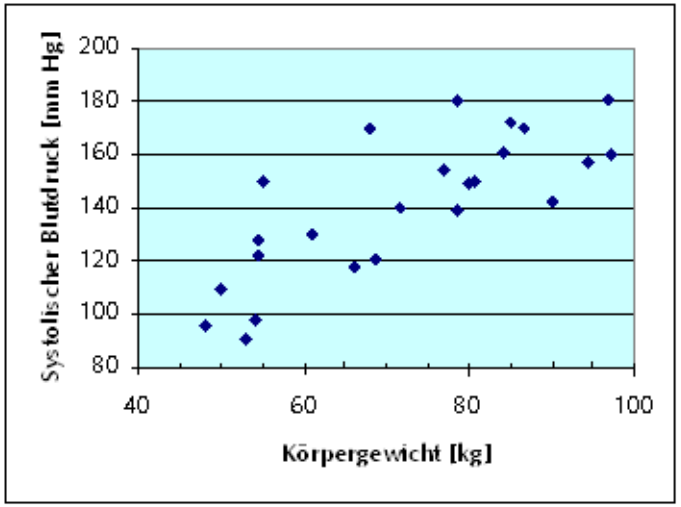
Beispiel 5: Ärzteeinkommen

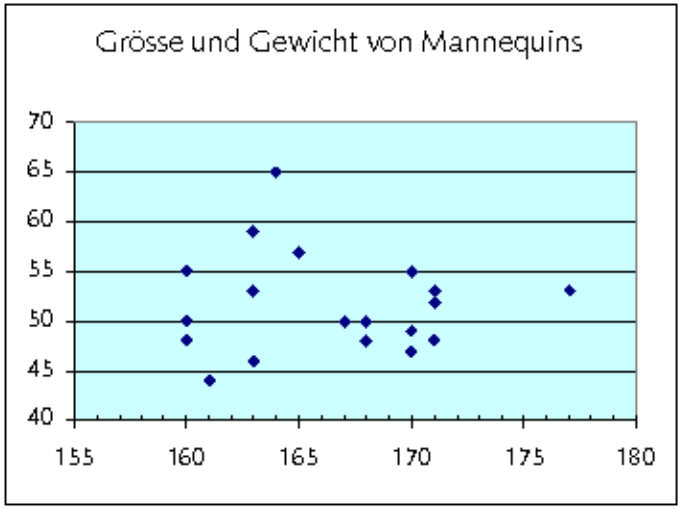
Einkommen aus freier Praxistätigkeit (2008)



Beziehungen zwischen 2 Merkmalen



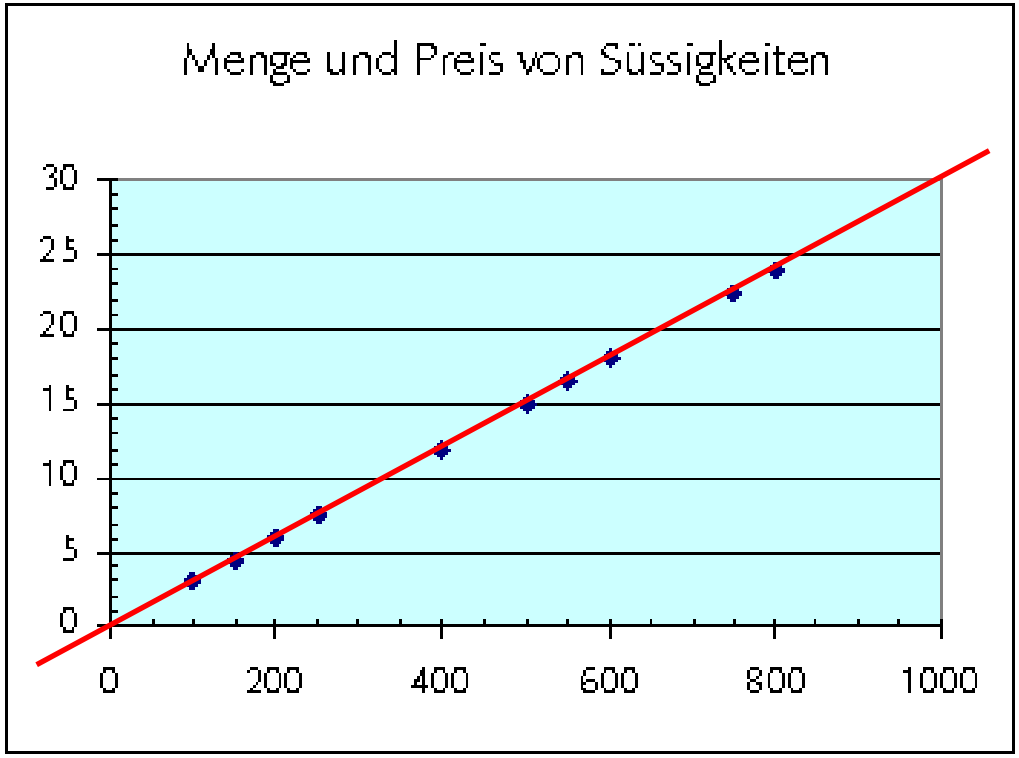




Beispiel 6: Süssigkeiten und ihr Preis

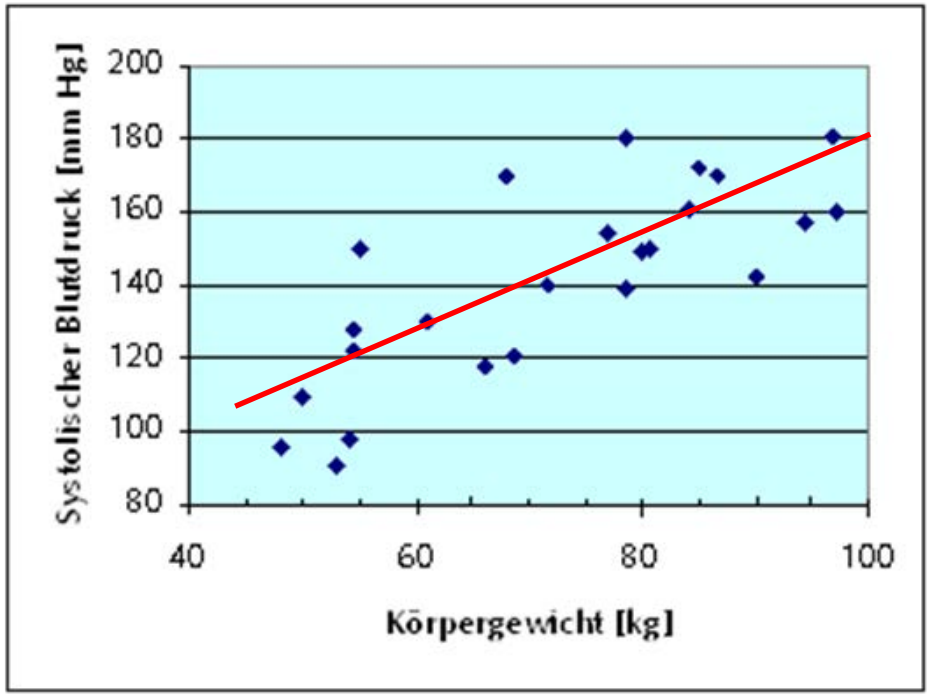
Alle Punkte liegen genau auf einer steigenden Geraden.

Je mehr Süssigkeiten, desto höher ihr Preis.



Beispiel 7: Gewicht und Blutdruck

Die Punkte liegen nicht exakt auf einer Geraden. Vielleicht passt eine gebogene Kurve besser?
Im Grundsatz gilt: Je schwerer, desto höher der Blutdruck.

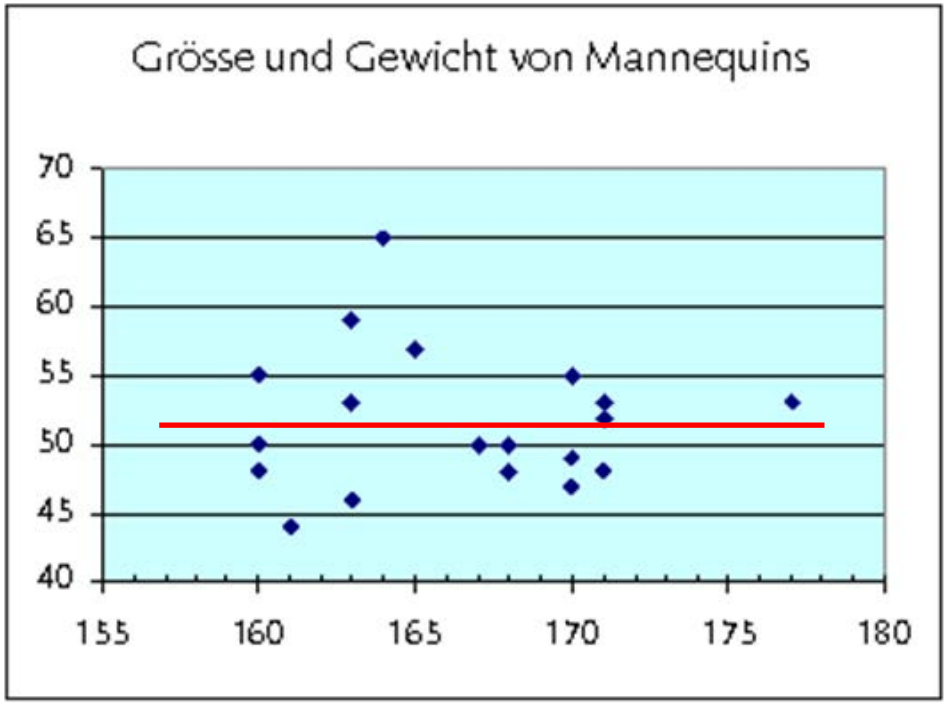


TI-Nspire CX CAS:
Bestimmen der «insgesamt möglichst gut passenden» Geraden oder Kurve durch die eingezeichneten Punkte.

Beispiel 8: Grösse und Gewicht

Die Punkte liegen nicht exakt auf einer Geraden.

Es ist nicht erkennbar, ob grössere Mannequins tendenziell schwerer sind oder nicht.

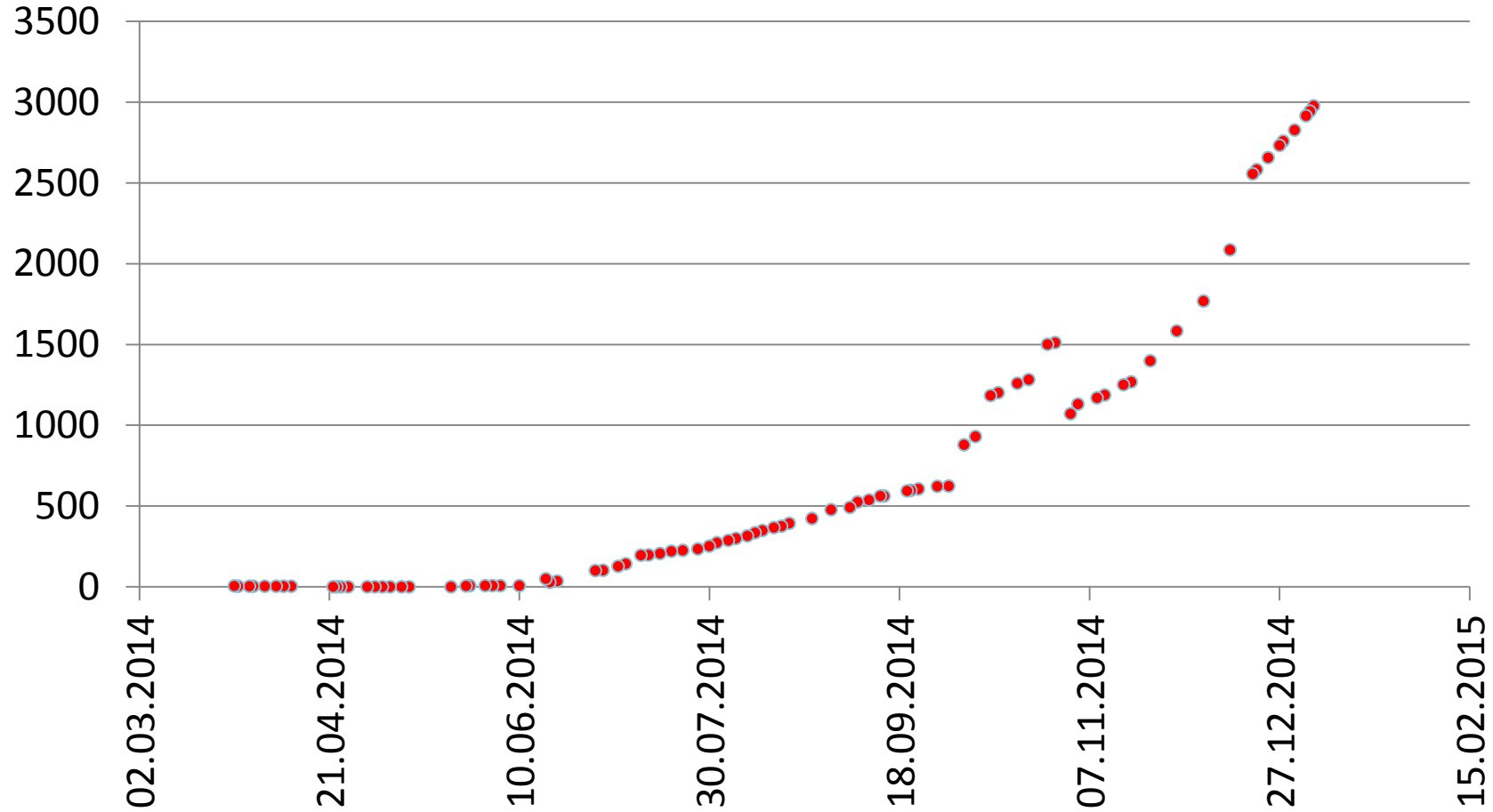


Vorsicht vor falschen Schlüssen!

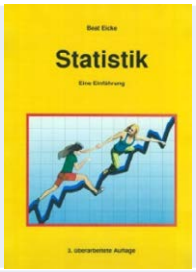
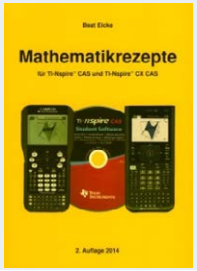
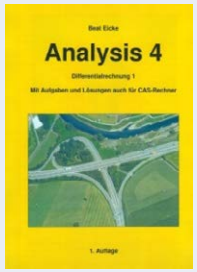
Eine hohe Korrelation zwischen 2 Merkmalen deutet auf einen hohen *statistischen* Zusammenhang hin – und oftmals, aber *nicht immer,* auf einen *sachlichen* Zusammenhang.

Wer das nicht beachtet, kann höchst absurde Vermutungen «beweisen».

Beispiel 9: Ebola-Tote in Sierra Leone



Literaturhinweise

	**Statistik – Eine Einführung**
	**Mathematikrezepte** **für TI-Nspire CAS und TI-Nspire CX CAS**
	**Analysis 4** **Differentialrechnung 1** (Teil einer auf 7 Bände angelegten Buchreihe)

Alle Bücher sind auf www.pythagoras.ch erhältlich.

Danke für Ihre Aufmerksamkeit!