Deskriptive Statistik Eindimensionale Daten

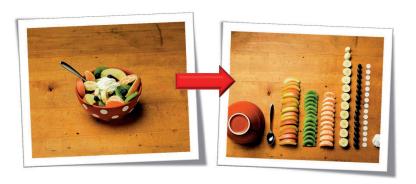
Peter Büchel

HSLU TA

Stat: SW02

Ziele der Deskriptiven Statistik

- Daten zusammenfassen durch numerische Kennwerte
- Graphische Darstellung der Daten



Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 1/56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 2/

Daten

- In diesem Modul \rightarrow meist reale Daten
- Datensatz Wiederholte Messungen: Freigesetzte Wärme beim Übergang von Eis bei $-0.7\,^{\circ}\mathrm{C}$ zu Wasser bei $0\,^{\circ}\mathrm{C}$
 - ightarrow 13 Werte (siehe Skript) (in cal/g) ightarrow Methode A

79.98; 80.04; 80.02; ... 80.02; 80.00; 80.02

- Basierend auf den Daten: Diverse *Kennwerte* berechnen bzw. Daten *graphisch darstellen*
- Warnung: Wann immer wir einen Datensatz "reduzieren" (durch Kennzahlen oder Graphiken), geht *Information verloren*!

25980996 0.37021603 0.07884733 0.71977404 0.07237495 0.68020504 0.48657579 0.53165132 0.59685485 0.78909487 0.93854889 0.95425422 0.5002 74579848 0.30692408 0.05351679 0.2853162 0.39888676 0.39349628 0.61886139 0.73188697 0.42457447 0.31000296 0.156226 0.50062453 0.4875 82994033 0.83220426 0.9372354 0.73133803 0.96199504 0.55862717 0.32692428 0.61868638 0.56245289 0.71896155 0.34543829 0.75111871 0.1583 92944405 0.64783158 0.60979875 0.52364734 0.26584028 0.40918689 0.16443477 0.25090652 0.04425809 0.06631721 0.45026614 0.96015307 0.5999 1.3322061 0.87182226 0.22334968 0.45692102 0.38131123 0.91921094 0.56080453 0.42412237 0.79812259 0.12081416 0.18896155 0.2448978 0.4241 97712468 0.50452793 0.57458309 0.02272522 0.12008212 0.68844427 0.93512611 0.35232595 0.54222107 0.74300188 0.1006917 0.22498337 0.6473 57467084 0.16038595 0.20683896 0.58934436 0.55401355 0.78000419 0.67956489 0.09056988 0.68952151 0.00707904 0.26790229 0.42494747 0.6355; 50684686 0.14058675 0.07426667 0.6377913 0.44437689 0.32789424 0.38075527 0.28287319 0.55515924 0.17444947 0.44069165 0.35637294 0.2464 72021194 0.52889677 0.51331006 0.20434876 0.5249763 0.71545814 0.61285279 0.87822767 0.53536095 0.28884442 0.69949788 0.84420515 0.7418 47268391 0.3610854 0.31014 04257944 0.09101231 0.10635 33114474 0.80847503 0.589571 17245673 0.67983345 0.231912 40573334 0.59170081 0.718914 15493105 0.51554705 0.81666845 0.734327 0.79912816 0.67877946 0.22687246 0.40043241 0.61701288 0.49018961 0.03681597 0.2230552 0.9720 38415242 0.04575544 0.18294704 0.07535783 0.49763891 0.15634616 0.47553336 0.39954434 0.49785766 0.19208229 0.03939701 0.50543817 0.1786 0.747484 0.7417904 0.48776921 0.34229175 0.65785054 0.77978943 0.20129577 0.62714576 0.46987345 0.69996167 0.48786104 0.99177657 0.6729 71427139 0.83346645 0.50236863 0.59062007 0.29268677 0.67964115 0.09614286 0.14222698 0.66263698 0.42537685 0.64928539 0.5648649 0.2613 96293853 0.6974188 0.85632265 0.45947964 0.00242453 0.68051404 0.20703925 0.87558209 0.679752 0.45999782 0.8722821 0.04547348 0.8243 0.5989028 0.87059205 0.12444579 0.26178908 0.8533065 0.20800837 0.90760418 0.06746495 0.61181415 0.37402957 0.36137753 0.83494 1.5616472 0.78210485 0.26718637 0.74856241 0.93690527 0.51338037 0.94582627 0.60380999 0.19747357 0.34424067 0.05237252 0.91349594 0.8796 71333452 0.28822987 0.65203382 0.49709346 0.70379359 0.27200958 0.85341908 0.15968767 0.34960955 0.6796046 0.34255204 0.62727145 0.9353 33192659 0.72932196 0.07036634 0.31364757 0.31615678 0.62072333 0.68964657 0.47503972 0.80823875 0.9708966 0.32082118 0.11199293 0.2306; 91696324 0.46608963 0.38554788 0.09440939 0.18995497 0.19254922 0.8299711 0.63238203 0.87524562 0.38170458 0.40120436 0.12882023 0.0850 0.8707509 0.06485663 0.22943682 0.41974316 0.9098332 0.86713599 0.88315761 0.31558244 0.63788522 0.48528904 0.17606219 0.17009773 0.4134 06291977 0.05277628 0.48101212 0.1043349 0.30497809 0.0559275 0.64358846 0.19723847 0.74347764 0.6704249 0.26325428 0.04458277 0.4040 22521559 0.30987268 0.99622375 0.94174692 0.28813039 0.20353298 0.84322955 0.54332297 0.34110065 0.68044315 0.87158643 0.41122531 0.8023

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 3/56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 4/56

Überblick

• Bekannt: *n* beobachtete Datenpunkte (Messungen)

$$X_1, X_2, \ldots, X_n$$

(z.B. Verkehrsaufkommen an *n* verschiedenen Tagen)

- Unterscheidung zwischen Lage- und Streuungsparametern
- Lageparameter ("Wo liegen die Beobachtungen auf der Mess-Skala?")
 - Arithmetisches Mittel ("Durchschnitt")
 - Median
 - Quantile
- Streuungsparameter ("Wie streuen die Daten um ihre mittlere Lage?")
 - ► Empirische Varianz / Standardabweichung
 - Quartilsdifferenz

Arithmetisches Mittel

Definition:

Arithmetisches Mittel

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Umgangsprachlich: Durchschnitt
- ullet Beispiel Schmelzwärme: Arithmetische Mittel der n=13 Messungen

$$\overline{x} = \frac{79.98 + 80.04 + \ldots + 80.03 + 80.02 + 80.00 + 80.02}{13} = 80.02077$$

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

5 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik I

Stat: SW02

6 / E

Arithmetisches Mittel

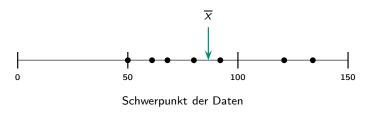
• Python-Befehl

```
from pandas import Series
import pandas as pd

methodeA = Series([79.98, 80.04, 80.02, 80.04, 80.03,
80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])

methodeA.mean()
## 80.02076923076923
```

Arithmetische Mittel: Anschaulich



Streuung

- Arithmetisches Mittel: "Wo ist die "Mitte" der Daten?"
- Aber: Beispiel von (fiktiven) Schulnoten:

2; 6; 3; 5 und 4; 4; 4; 4

- Beide Mittelwert 4, aber Verteilung der Daten um Mittelwert ziemlich unterschiedlich
 - ▶ 1. Fall: Zwei gute und zwei schlechte Schüler
 - ▶ 2. Fall: Alle Schüler gleich gut
- Datensätze haben eine verschiedene Streuung um den Mittelwert

Peter Büchel (HSLU TA)

Deskriptive Statistik I

Stat: SW02

7 / 56

Peter Büchel (HSLU TA)

eskrintive Statistik

Stat: SW02

/02 8 / 56

Streuung numerisch

- 1. Idee: Durchschnitt der Unterschiede zum Mittelwert
 - 1. Fall:

$$\frac{(2-4)+(6-4)+(3-4)+(5-4)}{4}=\frac{-2+2-1+1}{4}=0$$

Zweiter Fall auch 0 \rightarrow Keine Aussage

ullet Problem: Unterschiede können *negativ* werden ullet Können sich aufheben

Streuung

• Nächste Idee: Unterschiede durch die Absolutwerte ersetzen 1. Fall:

$$\frac{|(2-4)|+|(6-4)|+|(3-4)|+|(5-4)|}{4} = \frac{2+2+1+1}{4} = 1.5$$

- D.h.: Noten weichen im Schnitt 1.5 vom Mittelwert ab
- 2. Fall: Dieser Wert natürlich auch 0
- Je grösser dieser Wert (immer grösser gleich 0), desto mehr unterscheiden sich die Daten bei gleichem Mittelwert untereinander
- Dieser Wert für die Streuung: mittlere absolute Abweichung
- Aber: Theoretische Nachteile

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02 9 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

10 / 50

Empirische Varianz und Standardabweichung

- Besser: Empirische Varianz und empirische Standardabweichung für das Mass der Variabilität oder Streuung der Messwerte verwendet
- Definition:

Empirische Varianz Var(x) und Standardabweichung s_X

$$Var(x) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

und

$$s_x = \sqrt{\operatorname{Var}(x)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2}$$

Eigenschaften der Varianz

- Bei Varianz: Abweichungen $x_i \overline{x}$ quadrieren , damit sich Abweichungen nicht gegenseitig aufheben können
- Nenner n-1, anstelle von $n \rightarrow mathematisch begründet$
- Standardabweichung ist die Wurzel der Varianz
- Durch Wurzelziehen wieder dieselbe Einheit wie bei Daten selbst
- Ist empirische Varianz (und damit die Standardabweichung) gross, so ist die Streuung der Messwerte um das arithmetische Mittel gross
- Wert der empirische Varianz hat keine physikalische Bedeutung
 Man weiss nur, je grösser der Wert umso grösser die Streuung

Beispiele: Schmelzwärme

- Arith. Mittel der n = 13 Messungen ist $\overline{x} = 80.02$ (siehe oben)
- Empirische Varianz:

$$Var(x) = \frac{(79.98 - 80.02)^2 + (80.04 - 80.02)^2 + \dots + (80.00 - 80.02)^2 + (80.02 - 80.02)^2}{13 - 1}$$

$$= 0.000574$$

• Empirische Standardabweichung:

$$s_x = \sqrt{0.000574} = 0.024$$

 \bullet D.h.: "mittlere" Abweichung vom Mittelwert 80.02 cal/g ist 0.024 cal/g

Beispiele: Schmelzwärme

Von Hand sehr mühsam. Mit pandas-Methoden:
 Varianz:

```
methodeA.var()
## 0.0005743589743590099
```

Standardabweichung:

```
methodeA.std()
## 0.023965787580611863
```

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

13 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

-- /--

Median

- ullet Ein weiteres Lagemass für die "Mitte" ullet Median
- Sehr vereinfacht gesagt: Wert, bei dem die Hälfte der Messwerte unter diesem Wert liegen
- Beispiel: Prüfung in der Schule ist Median 4.6
- D.h.: Hälfte der Klasse liegt unter dieser Note
- Umgekehrt liegen die Noten der anderen Hälfte über dieser Note
- Obige Interpretation des Medians ist sehr vereinfacht dargestellt. Die exakte Definition folgt nun.

Geordnete Strichprobe

- Datensatz in aufsteigender Reihenfolge ordnen
- Bezeichnung der geordneten Daten mit $x_{(i)}$:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$$

• Beispiel: $x_1 = 3, x_2 = 7, x_3 = 2$:

$$x_{(1)} = x_3 = 2,$$
 $x_{(2)} = x_1 = 3,$ $x_{(3)} = x_2 = 7$

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

15 / 56

Peter Büchel (HSLU TA)

skrintive Statisti

Stat: SW02

Median

• Bestimmung *Median*: Daten zuerst der Grösse nach ordnen:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

• Für die Daten der Methode A heisst dies

 $79.97; \ 79.98; \ 80.00; \ 80.02; \ 80.02; \ 80.02; \ 80.03; \ 80.03; \ 80.03; \ 80.04; \ 80.04; \ 80.04; \ 80.05$

- Median ist nun sehr einfach zu bestimmen
- Unter diesen 13 Messungen: Wert der mittleren Beobachtung
- Dies ist in diesem Fall der Wert der 7. Beobachtung:

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

Median

- Median des Datensatzes der Methode A ist 80.03
- D.h.: Knapp die Hälfte der Messwerte, nämlich 6 Beobachtungen sind kleiner oder gleich 80.03
- Ebenso sind 6 Messwerte grösser oder gleich dem Median
- ullet Ungerade Anzahl Messungen o Genau eine mittlere Messung

Peter Büchel (HSLU TA)

Deskriptive Statistik I

Stat: SW02

17 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

10 / 56

Median

- Vorher: Anzahl der Daten ungerade und damit ist die mittlere Beobachtung eindeutig bestimmt
- ullet Anzahl Daten gerade ullet Keine mittlere Beobachtung
- Definition Median: Mittelwert der beiden mittleren Beobachtungen
- Beispiel: Datensatz der Methode B hat 8 Beobachtungen
- Ordnen den Datensatz: Median Durchschnitt von der 4. und 5. Beobachtung

79.94; 79.95; 79.97; 79.97; 79.97; 79.94; 80.02; 80.03 $\frac{79.97 + 79.97}{2} = 79.97$

Median

• Mit pandas erhalten wir für die Methode A

```
methodeA.median()
## 80.03
```

und für die Methode B

```
methodeB = Series([80.02, 79.94, 79.98, 79.97, 79.97,
80.03, 79.95, 79.97])
methodeB.median()
## 79.97
```

- Als Median kann Wert auftreten, der in Messreihe nicht vorkommt
- Annahme: Mittlere Beobachtungen der Methode *B* sind Werte 79.97 und 79.98:

 $\frac{79.97+79.98}{2}=79.975$

Peter Büchel (HSLU TA)

Deskriptive Statistik

19 / 56

Stat: SW02

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

Median vs. arithmetisches Mittel

- Zwei Lagemasse für die Mitte eines Datensatzes
- Welches ist nun "besser"?

Peter Büchel (HSLU TA)

- Dies kann man so nicht sagen, das kommt auf die jeweilige Problemstellung an. Am besten werden beide Masse gleichzeitig verwendet.
- Eigenschaft des Medians: Robustheit
- Das heisst: Wird viel weniger stark durch extreme Beobachtungen beeinflusst als das arithmetisches Mittel

Deskriptive Statistik

Median vs. arithmetisches Mittel

- Beispiel: Bei der grössten Beobachtung ($x_9 = 80.05$) ist ein Tippfehler passiert und $x_9 = 800.5$ eingegeben worden
- Das arithmetische Mittel ist dann

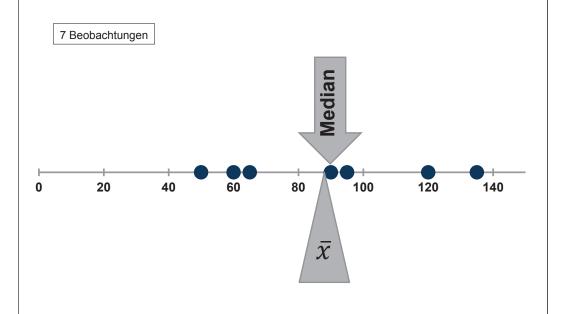
$$\bar{x} = 135.44$$

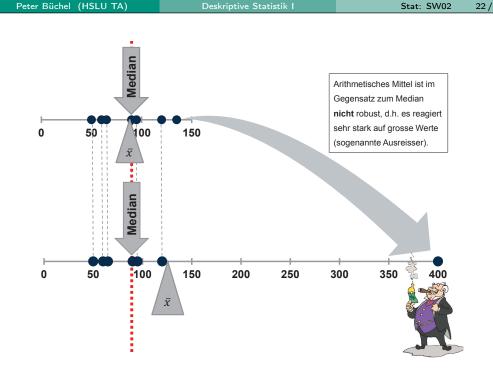
Der Median ist aber nach wie vor

$$x_{(7)} = 80.03$$

- Arithmetisches Mittel: Durch Veränderung einer Beobachtung sehr stark beeinflusst
- ullet Median hier gleich bleibt o robust

Arith. Mittel vs. Median: Einkommen [k CHF]





Stat: SW02

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 23/56 Peter Büchel (HSLU TA)

Stat: SW02



"Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?"

Quartile

- Median: Wert, wo die Hälfte der Beobachtungen kleiner (oder gleich) wie dieser Wert sind
- Analoge Überlegung: Unteres und oberes Quartil
- Unteres Quartil: Wert, wo 25 % aller Beobachtungen kleiner oder gleich und 75 % grösser oder gleich sind wie dieser Wert
- Oberes Quartil: Wert, wo 75 % aller Beobachtungen kleiner oder gleich und 25 % grösser oder gleich wie dieser Wert sind
- Achtung: Meist gibt es nicht exakt 25 % der Beobachtungen
- Man definiert Wert für das untere Quartil bzw. obere Quartil

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

25 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

26 / 5

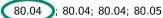
Beispiel: Schmelzwärme

- Methode A hat n = 13 Messpunkte \rightarrow 25 % davon ist 3.25
- Man wählt nächstgrösseren Wert $x_{(4)}$ als unteres Quartil:

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04; 80.04; 80.04; 80.05

- Unteres Quartil ist 80.02
- Knapp ein Viertel der Messwerte ist gleich oder kleiner 80.02
- Oberes Quartil: Wählen $x_{(10)}$, da für $0.75 \cdot 13 = 9.75$ die Zahl 10 der nächsthöhere Wert ist

79.97; 79.98; 80.00; 80.02; 80.02; 80.02; 80.03; 80.03; 80.03; 80.04);



Knapp drei Viertel der Messwerte sind kleiner oder gleich 80.04

- Methode B: 25% der Werte $2 \rightarrow \text{ganze Zahl}$
- Wählen dann nächste Beobachtung $x_{(3)}$ als unteres Quartil

79.94; 79.95; 79.97; 79.97; 79.97; 79.94; 80.02; 80.03

• Unteres Quartil der Methode B ist also 79.97

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 27/56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 28/5

Bemerkungen

- Hier jeweils aufgerundet, falls 25 % bzw. 75 % der Anzahl Beobachtungen nicht ganz ist
- ullet Hätten auch abrunden können ullet Andere Werte für die Quartile
- Für grosse Datensätze: Spielt praktisch keine Rolle, ob auf- oder abgerundet oder gerundet wird
- Aber: Es gibt keine einheitliche Definition für die Quartile

pandas

- pandas kennt keine eigenen Befehle für die Quartile
- Allgemeinerer Befehl quantile (Quantile kommen gleich)
- Python: Quartile nach unserer Definition Option interpolation="lower"
- Für das untere Quartil der Methode A lautet der Befehl

```
methodeA.quantile(q=.25, interpolation="lower")
## 80.02
und für das obere
```

```
methodeA.quantile(q=.75, interpolation="lower")
## 80.04
```

Peter Büchel (HSLU TA)

Deskriptive Statistik I

Stat: SW02

29 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

30 / 56

pandas

• Methode *B*:

```
methodeB.quantile(q=.25, interpolation="lower")
## 79.95
```

- Dies entspricht *nicht* unserem oben bestimmten Wert 79.97
- Python führt noch Korrekturfaktor zur Berechnung der Quantile ein
- Dieser ist aber für grosse Datensätze irrelevant

Quartilsdifferenz

- Quartilsdifferenz ist ein Streuungsmass für die Daten oberes Quartil – unteres Quartil
- Es misst die Länge des Intervalls, das etwa die Hälfte der "mittleren"
 Beobachtungen enthält
- Je kleiner dieses Mass, umso n\u00e4her liegt die H\u00e4lfte aller Werte um den Median und umso kleiner ist die Streuung
- Dieses Streuungsmass ist robust
- Quartilsdifferenz der Methode A

$$80.04 - 80.02 = 0.02$$

Peter Büchel (HSLU TA)

Deskriptive Statistik I

Stat: SW02

31 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik I

Stat: SW02

pandas

Mit pandas

```
q75, q25 = methodeA.quantile(q = [.75, .25],
interpolation="lower")

iqr = q75 - q25
iqr
## 0.020000000000010232
```

• Die (oder ungefähr die) Hälfte der Messwerte liegt also in einem Bereich der Länge 0.02

Quantile

- Quartile auf jede andere Prozentzahl verallgemeinern: Quantile
- 10 %-Quantil: Wert, wo 10 % der Werte kleiner oder gleich und 90 % der Werte grösser oder gleich diesem Wert sind
- Empirische α -Quantil: Wert, wo $\alpha \times 100 \%$ der Datenpunkte kleiner oder gleich und $(1-\alpha) \times 100 \%$ der Punkte grösser oder gleich sind
- Definition

```
Empirische \alpha-Quantile (0 < \alpha < 1) x_{(\alpha n+1),} \text{ falls } \alpha \cdot n \text{ eine natürliche Zahl ist} x_{(k)}, \qquad \text{wobei } k \text{ die Zahl } \alpha \cdot n \text{ aufgerundet ist, falls } \alpha \cdot n \notin \mathbb{N}
```

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 33/56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 34/56

- Wie bei den Quartilen:
 - ► Aufgerunden, falls die entsprechende Prozentzahl der Beobachtungen nicht ganz;
 - sonst nächsthöheren Wert
- Empirischer Median ist empirisches 50 %-Quantil
- Empirisches 25 %-Quantil ist unteres Quartil
- Empirisches 75 %-Quantil ist oberes Quartil

• pandas: 10 %— und 70 %—Quantil der Methode A:

```
methodeA.quantile(q=.1, interpolation="lower")
methodeA.quantile(q=.7, interpolation="lower")
## 79.98
## 80.03
```

- \bullet Knapp 10 % der Messwerte sind kleiner oder gleich 79.98
- Entsprechend: Knapp 70 % der Messwerte kleiner oder gleich 80.03

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 35 / 56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 36 / 56

Beispiel

- Noten an Prüfung in Schulklasse mit 24 SchülerInnen:
 - 4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1
- Verschiedene Quantile mit pandas:

- D.h.: Knapp 20 % der SuS sind schlechter als 3.6
- 20 % SchülerInnen nicht möglich, da dies 4.8 SuS wären
- 60 %-Quantil: (Knapp) diese Anzahl Prozent der SuS waren schlechter oder gleich einer 4.9

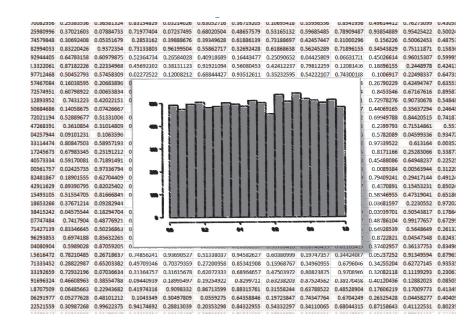
Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

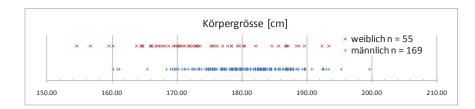
2 37 / 56

Graphische Darstellungen



Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 38 / 56

Eindimensionales Streudiagramm



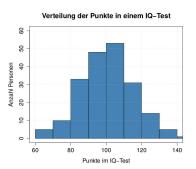
- Guter Überblick, falls nicht zu viele Daten vorhanden sind
- Achtung bei diskret verteilten Daten (Punkte liegen aufeinander!)

Histogramm

- Histogramm: Graphischer Überblick über die auftretenden Werte
- Aufteilung des Wertebereichs in k Klassen (Intervalle)
- Faustregel:
 - ▶ bei weniger als 50 Messungen ist die Klassenzahl 5 bis 7
 - ▶ bei mehr als 250 Messungen wählt man 10 bis 20 Klassen
- Zeichne für jede Klasse einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 39 / 56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 40 / 56

Beispiel: IQ-Test



- Histogramm von IQ-Test Ergebnis von 200 Personen
- Breite der Klassen: 10 IQ-Punkte; für jede Klasse gleich
- Höhe der Balken: Anzahl Personen, die in diese Klasse fallen
- Beispiel: ca. 14 Personen fallen in die Klasse zwischen 120 130 IQ-Punkten

Peter Büchel (HSLU TA) Deskriptive Statistik Stat: SW02

41 / 56

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

42 / 56

44 / 56

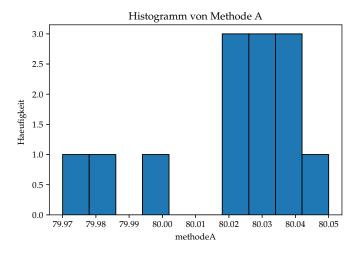
Python

Es wurde mit folgendem Code erzeugt:

```
import pandas as pd
from pandas import DataFrame, Series
import matplotlib.pyplot as plt
methodeA = Series([79.98, 80.04, 80.02, 80.04, 80.03, 80.03,
80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02])
methodeB = Series([80.02, 79.94, 79.98, 79.97, 79.97, 80.03,
79.95, 79.97])
methodeA.plot(kind="hist", edgecolor="black")
plt.title("Histogramm von Methode A")
plt.xlabel("methodeA")
plt.ylabel("Haeufigkeit")
plt.show()
```

Python

Für die Methode A sieht das Histogramm wie folgt aus:



Bemerkungen

- Methode A 13 Messungen → 10 Balken (pandas-default)
- Bedeutung der Anzahlen (Frequency):
 - ▶ 10 Klassen mit Werten im Bereich [79.97, 80.05] → Balkenbreite

$$\frac{80.05 - 79.97}{10} = 0.008$$

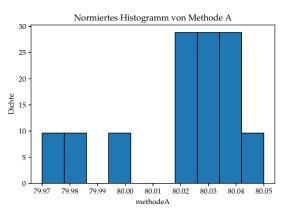
- ▶ 1. Klasse 79.97-79.978: Anzahl Beobachtungen 79.97 berücksichtigt
- 2. Klasse Werte 79.98: usw.
- pandas selbst keine Graphiken \rightarrow Bibliothek matplotlib
- Pandas-Attribut plot für Plots → Option kind="hist"
- Mit dem Python -Befehl lassen sich auch die Anzahl Klassen festlegen, Überschriften ändern, usw. (siehe Übungen)

Peter Büchel (HSLU TA) Deskriptive Statistik Stat: SW02 43 / 56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02

Histogramm: Dichte

- Histogramm oben: Höhe der Balken entspricht Anzahl Beobachtungen in Klasse
- Andere Form des Histogramms

```
methodeA.plot(kind="hist", normed=True,
edgecolor="black")
```



Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

45 / 56

Gesamtfläche der Balken ist 1

• Auf der vertikalen Achse sind nun die Dichten angegeben

• Herauslesen: über

$$(80.018 - 80.026) \cdot 28.846 = 0.23$$

also etwa 23 % der Daten zwischen 80.018 und 80.026 befinden

- Balkenhöhe: Anzahl Beobachtungen in einem Balken mit $\frac{1}{n}$ multiplizieren, diese Zahl durch die Balkenbreite dividiert
- Unser Beispiel: 3 Beobachtungen im Intervall [80.018, 80.026]
- Balkenhöhe:

$$\frac{\frac{1}{13} \cdot 3}{0.008} = 28.8462$$

 Vorteil: Messungen mit unterschiedlichen Umfängen besser miteinander vergleichbar

Peter Büchel (HSLU TA)

Deskriptive Statistik

Stat: SW02

46 / 56

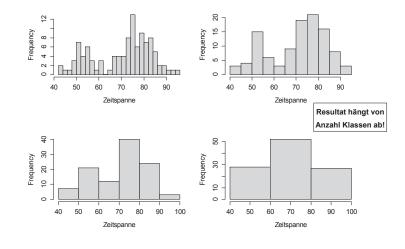
Old Faithful Geysir (Yellowstone NP): Daten

- Zeitspanne [min] zwischen Ausbrüchen
- Eruptionsdauer [min]
- Daten finden Sie auf ILIAS



A	Α	В	С	D
1	Tag	Zeitspanne	Eruptionsdauer	
2	1	78	4.4	
3	1	74	3.9	
4	1	68	4	
5	1	76	4	
6	1	80	3.5	
7	1	84	4.1	
8	1	50	2.3	
9	1	93	4.7	
10	1	55	1.7	
11	1	76	4.9	
12	1	58	1.7	
13	1	74	4.6	
14	1	75	3.4	
15	2	80	4.3	
16	2	56	1.7	
17	2	80	3.9	
18	2	69	3.7	
19	2	57	3.1	
20	2	90	4	
21	2	42	1.8	
22	2	91	4.1	
23	2	51	1.8	

Histogramme der Zeitspanne (verschiedene Anzahl Klassen)



Interaktiv Klassen verändern (bei anderen Daten):

http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html

Peter Büchel (HSLU TA) Deskriptive Statistik I

Stat: SW02

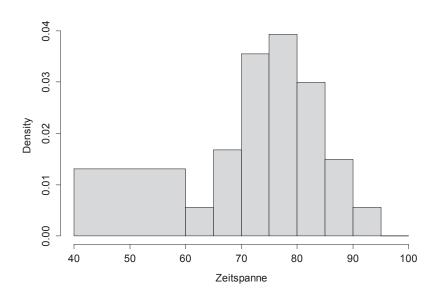
47 / 56

Peter Büchel (HSLU TA)

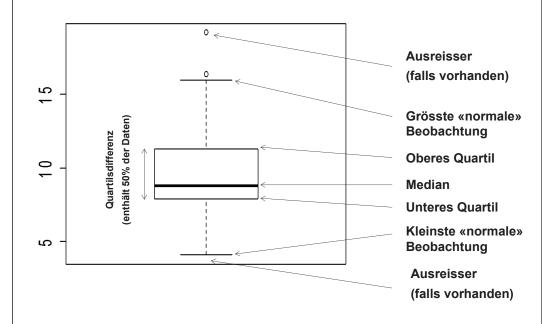
Deskriptive Statistik

Stat: SW02

Histogramm der Zeitspanne mit unterschiedlicher Intervallbreite



Boxplot: Schematischer Aufbau



Deskriptive Statistik

Boxplot: Schematischer Aufbau

Peter Büchel (HSLU TA)

• Grösste normale Beobachtung: Grösste Beobachtung, die höchstens $1.5 \cdot r$ vom oberen Quartil entfernt ist (r: Quartilsdifferenz)

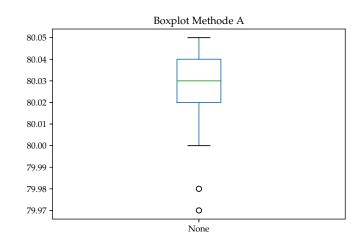
Deskriptive Statistik

- Kleinste normale Beobachtung: Analog definiert mit dem unteren Quartil
- Ausreisser sind Punkte, die ausserhalb dieser Bereiche liegen

Python

Peter Büchel (HSLU TA)

methodeA.plot(kind="box", title="Boxplot Methode A")



Peter Büchel (HSLU TA)

Stat: SW02

Stat: SW02

49 / 56

51/56

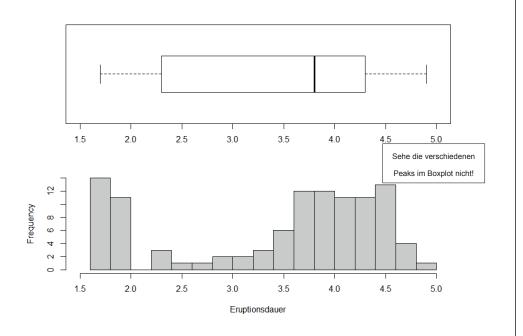
Peter Büchel (HSLU TA)

Stat: SW02

Stat: SW02

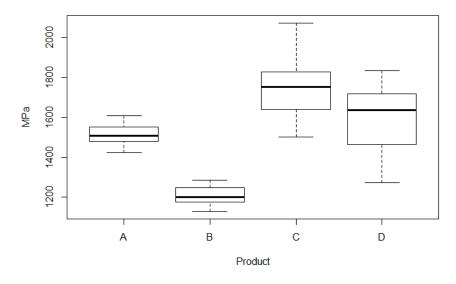
50 / 56

Boxplot und Histogramm der Eruptionsdauer



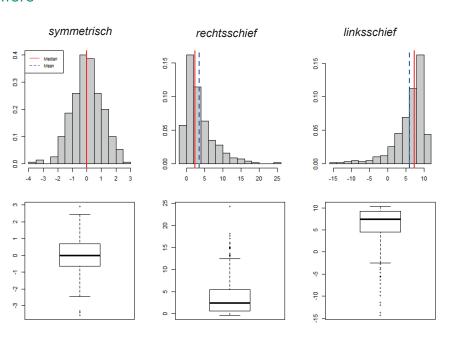
Mehrere Boxplots

Mit mehreren Boxplots kann man einfach und schnell die Verteilung von verschiedenen Gruppen (Methoden, Produkte, ...) vergleichen



Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 53/56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 54/56

Schiefe



Boxplot: Bemerkungen

- Im Boxplot sind ersichtlich:
 - ▶ Lage
 - Streuung
 - Schiefe
- Man sieht aber z.B. nicht, ob eine Verteilung mehrere "Peaks" hat

Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 55/56 Peter Büchel (HSLU TA) Deskriptive Statistik I Stat: SW02 56/56