# Applied Statistics for Data Science

**Zusammenfassung**

Stephan Stofer

25. September 2020

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# 1 Einleitung

Applied statistics wenden wir Statistik auf konkrete Alltagsprobleme an. Dazu wenden wir die vierstufige Problemlösungsstrategie an.

## 1.1 Vierstufige Problemlösungsstrategie

1. Erste Schritte Es ist nicht immer klar was die effizienteste Antwort auf das Problem ist. Informationen organisieren/sammeln. Problem gegebenenfalls mit eigenen Worten formulieren. Sind alle Infos da, die wir für die Lösung des Problems brauchen?
2. Plan ausarbeiten Herausfinden, welche Schritte nötig sind, um das Problem zu lösen.
3. Plan ausführen Die Schritte in den im Punkt 2 definierten Abfolge ausführen.
4. Resultat interpretieren Überprüfung ob das Resultat möglich und sinnvoll ist. Interpretation des Resultats in den Worten der Problemstellung.

# 2 Deskriptive Statistik - Eindimensionale Daten

## 2.1 Datensätze

Datensätze sind Zusammenstellungen von Daten die in vielen Formen vorkommen können.

### 2.1.1 Liste

Eine Liste von Daten ist die einfachste Variante eines Datensatzes. Sie enthält zum Beispiel die Körpergrösse in Meter von fünf Personen.

$$1.75, 1.80, 1.72, 1.65, 1.54$$

Solche Listen heissen auch *eindimensionale Datensätze* oder *Messreihen*

### 2.1.2 Tabellen

Die häufigste Form von Datensätzen sind Tabellen oder *zweidimensionale Datensätze*. Bei Einträgen mit Zahlen spricht man von *quantitativen* Daten, sprich Messwerte und können theoretisch jede beliebigen Zahlenwert annehmen. Andere (zB. ein Spalte Geschlecht oder Nationalität) sind sogenannte *qualitative* Daten und können nur eine bestimmte Anzahl Werte annehmen. Sie können aber auch Zahlen sein.

## 2.2 Deskriptive Statistik

Die deskriptive Statistik befasst sich mit der *Darstellung* von Datensätzen (lat. describere, beschreiben). Dabei werden Datensätze durch gewisse Zahlen charakterisiert (z.B. Mittelwert) und/oder graphisch dargestellt. Quantitative Daten werden organisiert und zusammengefasst. Die Interpretation und darauffolgende statistische Analyse soll vereinfacht werden. Wir erledigen dies mit Hilfe von:

- graphischen Darstellungen wie Histogramme und Boxplots
- *Kennzahlen*, die Daten numerisch zusammenfassen, die Durchschnitt und Standardabweichung

  Daten sollten immer mit Hilfe von graphischen *und* Kennzahlen dargestellt werden. Nur auf diese Weise können (teils unerwartete) Strukturen und Besonderheiten entdeckt werden.

Man muss sich ausserdem bewusst sein: Werden *Daten zusammengefasst*, gehen *Informationen verloren*!

### 2.2.1 Bezeichnung von Daten

Im Folgenden werden Daten mit $x1, x2, ..., xn$ bezeichnet, wobei $n$ der *Umfang* der Messreihe genannt wird.

## 2.3 Kennzahlen

Meistens ist es sinnvoll, Datensätze durch eine Zahl, also numerisch zusammenzufassen und damit zu beschreiben. Sie werden dabei auf eine odere mehrere Zahlen reduziert. Die zwei wichtigsten sind:

- Lageparameter oder Lagemasse; beschreiben *wo* sich Daten befinden. Beschreibt als Bsp. die *mittlere Lage* der Messwerte (muss nicht Durchschnitt gemeint sein)
- Streuungsparameter oder Streuungsmasse; beschreiben *wie* sich die Daten um die mittlere Lage verteilen. Die *Variabilität* oder *Streuung* der Messwerte gibt die durchschnittliche Abweichung von der mittleren Lage an

### 2.3.1 Arithmetisches Mittel

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

Bekannteste Grösse für eine mittlere Lage ist der *Durchschnitt* oder das **Arithmetische Mittel** $\bar{x}$. In der Notation $\bar{x}_n$ beschreibt $n$ wieder den Umfang der Messreihe.


Abbildung 2.1: Graphische Darstellung des arithmetischen Mittelwertes

#### 2.3.1.1 Arithmetisches Mittel mit R

```r
# Vektor (Datensatz) bilden
waageA <- c(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)
mean(waageA)
# [1] 80.02077
```

### 2.3.2 Empirische Varianz und Standardabweichung

Das arithmetische Mittel beschreibt einen Datensatz nur unvollständig. In der Abbildung 2.2 ersichtlich, dass zwei Datensreihen dasselbe arithmetische Mittel haben. Allerdings liegen die Punkte der ersten Datenreihe weiter vom Mittelpunkt entfernt, als die Punkte der ersten. Wir sprechen von unterschiedicher *Streuung* der Daten um den Durchschnitt.

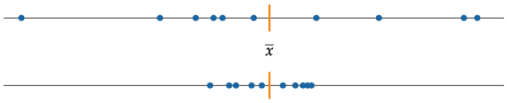


Abbildung 2.2: “Grosse” und “kleine” Streuung von zwei Messreihen

Die gebräuchlichsten Masse für die Streuung oder Variablität von Messwerten sind die *empirische Varianz* und *empirische Standardabweichung*.

#### 2.3.2.1 Mathematische Definition Empirische Varianz

Bei der Varianz werden die Abweichungen quadriert, dadurch können sich diese nicht gegenseitig aufheben. In einigen Bücher steht bei der Definition für die Varianz im Nenner $n$, anstatt $n-1$. Für kleine Datensätze spielt dies eine Rolle, bei grossen jedoch vernachlässigbar. r verwendet mit dem Befehl `r var(x)` auch $n-1$.

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

#### 2.3.2.2 Mathematische Definition Empirische Standardabweichung

Durch das Quadrieren erhalten die Werte eine neue Einheit (z.B. $cm^2$). Durch das Wurzelziehen führen wir diese wieder ihrer ursprünglichen Einheit zu und erhalten damit die Standardabweichung.

$$s_x = \sqrt{Var(x)}$$

> Nur die Standardabweichung $s_x$ lässt sich korrekt interpretieren. Der Wert der empirischen Varianz hat keine physikalische Bedeutung. Wir wissen nur, je grösser der Wert, umso grösser die Streuung.

#### 2.3.2.3 Empirische Varianz bzw. Standardabweichung mit R

```r
var(waageA)
# [1] 0.000574359
sd(waageA) # sd = standard deviation
## [1] 0.02396579
```

### 2.3.3 Median

Der Median ist ein Lagemass für denjenigen Wert, bei dem rund die Hälfte der Messwerte kleiner oder gleich und die andere Hälfte grösser oder gleich diesem Wert sind. Um den *Median* zu bestimmen, müssen alle Daten erst geordnet werden. Ist die Anzahl der Daten *ungerade*, gibt es eine *mittlere* Beobachtung. Bei einer *geraden* Anzahl gibt es zwei *gleichwertige mittlere* Beobachtungen. Als Median benützen wir in diesem Fall den Durchschnitt der beiden mittleren Beobachtungen $\frac{m_1+m_2}{2}$.

- Der Median muss *kein* Wert aus dem Datensatz sein
- Er wird auch *Zentralwert* oder *mittlerer Wert* genannt
- der Median ist sehr *robust*, dies bedeutet dass er weniger stark durch extreme Beobachtungen (Ausreisser) beeinflusst wird als das arithmetische Mittel.

In der Abbildung 2.3 erkennt man, wie durch den blauen Punkt (zweiter Zahlenstrahl ganz rechts) der Durchschnitt von $x_n$ zu $x_n^*$ verändert wird.


Abbildung 2.3: Die Robustheit des Median

#### 2.3.3.1 Median mit R

```r
median(waageA)
## [1] 80.03
```

#### 2.3.3.2 Bermerkungen zum Median

Der Median ist gerechter. Weshalb er auch für das berechnen des mittleren Einkommens verwendet wird. Die beiden Lagemasse für die mittlere Lage sollten immer gemeinsam betrachtet werden. Eine grosse Abweichung zwischen den Werten deutet auf besondere Verteilung der Daten hin.

### 2.3.4 Quartile

Die Quartile sind analog dem Median definiert, aber nicht für 50% der Daten die grösser oder kleiner sind, sonder für 25% bzw. 75% der Daten. Das *untere* Quartil ist derjenige Wert, bei welchem 25% aller Beobachtungen kleiner oder gleich und 75% grösser oder gleich diesem Wert sind. Entsprechend ist das *obere* Quartil derjenige Wert, bei dem 75% aller Beobachtungen kleiner oder gleich und 25% grösser oder gleich diesem Wert sind.

Hat eine Messreihe 13 Messpunkte sind 25% davon 3.25. Wir runden jeweils auf –> der vierte Wert wird dann zum unteren Quartil.

#### 2.3.4.1 Quartil in R

```r
# Syntax fuer das untere Quartil: p = 0.25, type definiert den verwendeten Algorithmus
# https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile
quantile(waageA, p = 0.25, type = 2)
## 25%
## 80.02

# Syntax fuer das obere Quartil: p = 0.75
quantile(waageA, p = 0.75, type = 2)
## 75%
## 80.04
```

### 2.3.5 Quartilsdifferenz

Die Quartilsdifferenz ist definiert als die Differenz der beiden Quartile: $oberesQuartil - unteresQuartil$ Sie ist ein Streuungsmass für die Daten. Es misst die Länge des Intervalls, das etwa die Hälfte der mittleren Beobachtungen enthält. Je kleiner dieses Mass, umso näher liegt die Hälfte aller Werte beim Median und umso kleiner ist die Streuung. Dieses Streuungsmass ist robust.

#### 2.3.5.1 Quartilsdifferenz in R

```r
IQR(waageA, type = 2)
## [1] 0.02
```

Dies bedeutet, dass die Hälfte aller Messwerte in einem Bereich der Länge 0.02 liegen.

### 2.3.6 Quantile

Mit den *Quantilen* kann das Konzept der Quartile auf jede beliebige Prozentzahl verallgemeinert werden. Das *empirische* $\alpha$*-Quantile* ist derjenige Wert, bei dem $\alpha * 100$ der Datenpunkte kleiner oder gleich und $(1 - \alpha) * 100$ der Punkte görsser oder gleich diesem Wert sind.

#### 2.3.6.1 Quantil in R

```r
quantile(waageA, p = 0.1, type = 2)
## 10%
## 79.98

quantile(waageA, p = 0.7, type = 2)
## 70%
## 80.04
```

Weiteres Beispiel mit versch. Quantilen in einer Zeile

```r
quantile(noten, p = seq(from = 0.2, to = 1, by = 0.2), type = 2)
## 20% 40% 60% 80% 100%
## 3.6 4.2 5.0 5.6 6.0
```

Rund 20% der Lernenden haben also eine 3.6 oder waren schlechter und rund 80 % der Lernenden waren gleich oder besser als dieser Wert. Genau 20% der Lernenden ist nicht möglich, da dies 4.8 Lernenden entsprechen würde. Das 60%-Quantil besagt, dass rund 60 Prozent der Lernenden Noten von 5 oder weniger haben. Folglich haben rund 40% eine 5 oder sind besser.

## 2.4 Graphische Methoden

Daten graphisch dazustellen ist ein sehr wichtiger Aspekt der Datenanalyse.

### 2.4.1 Histogramm

Histogramme helfen bei der Frage, in welchem *Wertebereich* besonders viele Datenpunkte liegen. Besonders dann, wenn die Datenmenge gross ist und es keinen Sinn macht, alle Werte einzeln zu betrachten.

#### 2.4.1.1 Histgramm in R

```r
iq <- rnorm(n = 200, mean = 100, sd = 15)
hist(iq,
   col = "darkseagreen3",
   xlab = "Punkte im IQ-Test",
   ylab = "Anzahl Personen",
   main = "Verteilung der Punkte in einem IQ-Test",
   breaks = "sturges" # default, sonst INT-Value
)
```

- `rnorm(n = 200, mean = 100, sd = 15)` wählt zufällig 200 normalverteilte Daten mit Mittelwert 100 und einer Standardabweichung von 15 aus
- `hist(iq, ...)` zeichnet das Histogramm für `iq`

- xlab ist das x-Label
- ylab ist das y-Label
- col definiert die Farbe
- main steht für Haupttitel

Beim Aufbau eines Histogramm werden die Daten in Klassen eingeteilt. Dabei wird die *Anzahl* der Klassen (Balken) anhand verschiedenen Faustregeln gebildet. Bei weniger als 50 Messungen sind es 5 bis 7, bei mehr als 250 wählt man 10 bis 20 Klassen. Die Wahl der Anzahl ist relevant für die Aussagekraft eines Histogrammes. Es gibt keine allgemeine Grundregel für die Wahl.

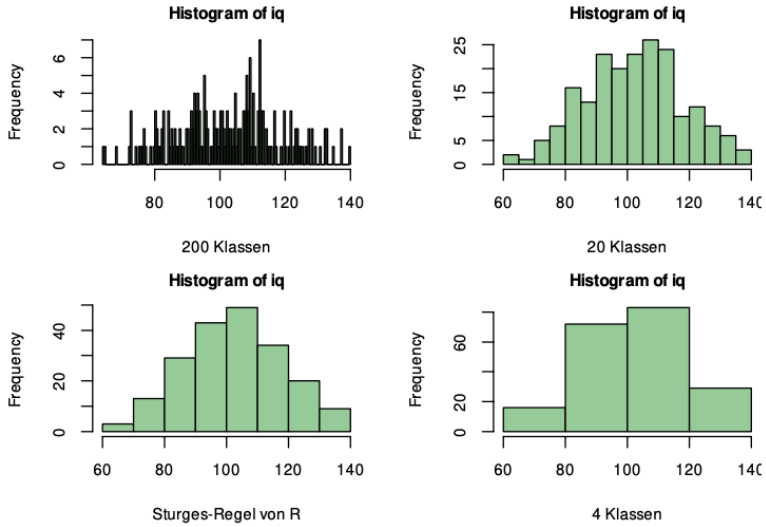


Abbildung 2.4: Vergleich der Histogramme mit verschiedener Klassenwahl

#### 2.4.1.2 Bimodales Verhalten

*Bimodales* Verhalten ist sichtbar, wenn es zwei "Hügel" im Histogramm gibt

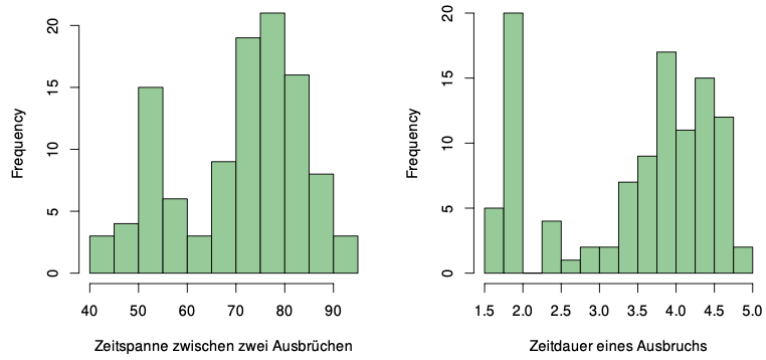


Abbildung 2.5: Bimodales Verhalten in zwei Histogrammen

#### 2.4.1.3 Schiefe von Histogrammen

Wir betrachten die Histogramme in Abbildung 2.6

- Das Histogramm links ist symmetrisch bezüglich 5. Die Daten sind um 5 auf beiden Seiten ähnlich verteilt.
- In einem *rechtsschiefen* Histogramm sind die meisten Daten links im Histogramm

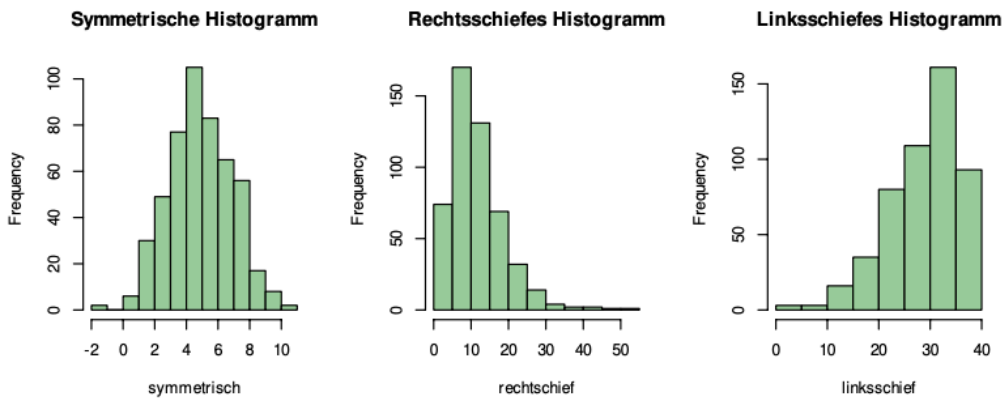


Abbildung 2.6: Symmetrisches, rechts- und linksschiefes Histogramm

- In einem *linksschiefen* Histogramm sind die meisten Daten rechts im Histogramm

Die Bezeichnung "rechts" und "links" bezieht sich immer auf die Richtung von *weniger* Daten sind.

#### 2.4.1.4 Normiertes Histogramm

In den vorherigen Histogrammen ist die Höhe der Balken gerade der Anzahl der Beobachtungen in einer Klasse. In einem normierten Histogramm wird die Balkenhöhe so gewählt, dass die *Balkenfläche* dem Anteil der jeweilgen Beobachtungen an der Gesamtanzahl entspricht. Die Gesamtfläche der Balken muss dann gleich eins sein. Auf der vertikalen Achse ist dann die *Dichte* aufgetragen (entspricht *nicht* Prozentwerten).

```
hist(waageA,
  freq = F,
  main = "Histogramm von Waage A",
  col = "darkseagreen3",
  ylim = c(0, 25)
)
rect(80.02, 0, 80.04, 23.1, col="darkseagreen4")
```

- mit `freq = F` (frequency false) wird das Histogramm normiert gezeichnet
- Die Option 'ylim = c(0, 25) gibt an, in welcem Bereich die vertikale Achse gezeichnet werden soll
- `rect` zeichnet ein Rechteck in eine vorgegebene Grafik. Die ersten beieden Zahlen sind die Koordinaten des Punktes links unten und die zweiten beiden Zahlen die Koordinaten des Punktes rechts oben.

Mit Hilfe der normierten Histogrammen lassen sich insbesondere solche Datenstämme vergleichen, die sehr unterschiedlich viele Messpunkte enthalten.

### 2.4.2 Boxplot

Ein Boxplot ist in Abbildung 2.7 schematisch dargestellt. Er besteht aus:

- einem Rechteck dessen Höhe vom empirischen 25%- und 75%-Quantil begrenzt wird (grüne Fläche)
- horizontalem Strich in der Box für den Median (schwarz)
- *whiskers*, blaue Linien, die zum kleinsten und grössten "normalen" Beobachtung führen (normal heisst *höchstens* 1.5 mal die Quartilsdifferenz von oberen und unteren Quartil)
- kleine roten Kreisen, die Ausreisser, welche ausserhalb der normalen Beobachtungen liegen

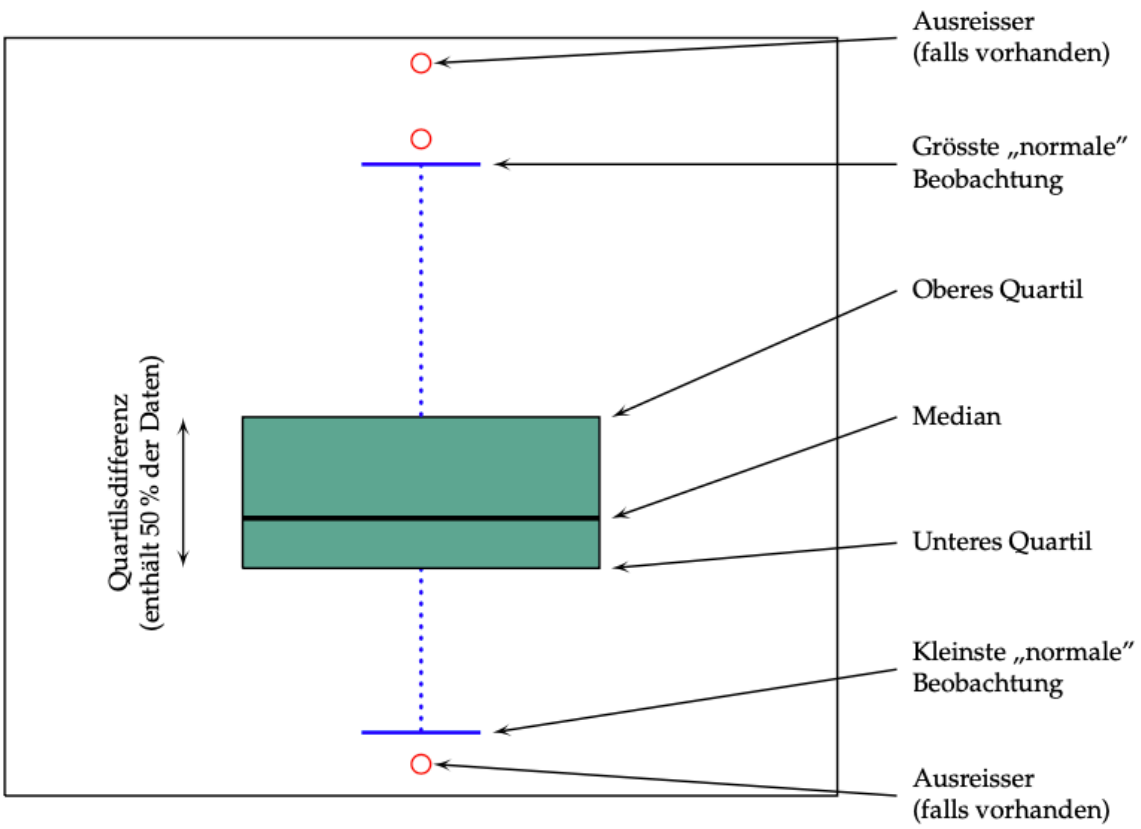


Abbildung 2.7: Schematischer Aufbau eines Boxplotes

#### 2.4.2.1 Boxplot in R

```
boxplot(waageA,
    col = "darkseagreen3"
)
```

Boxplotte sind vorallem dann geeignet, wenn die Verteilung der Daten in verschiedenen Gruppen (versch. Versuchsbedingungen) verglichen werden sollen.