

Applied Statistics for Data Science

Zusammenfassung

Stephan Stofer

2. November 2020

Inhaltsverzeichnis

1	Einleitung	5
1.1	Vierstufige Problemlösungsstrategie	5
2	Deskriptive Statistik - Eindimensionale Daten	6
2.1	Datensätze	6
2.1.1	Liste	6
2.1.2	Tabellen	6
2.2	Deskriptive Statistik	6
2.2.1	Bezeichnung von Daten	6
2.3	Kennzahlen	7
2.3.1	Arithmetisches Mittel	7
2.3.2	Empirische Varianz und Standardabweichung	7
2.3.3	Median	8
2.3.4	Quartile	9
2.3.5	Quartilsdifferenz	9
2.3.6	Quantile	10
2.4	Graphische Methoden	10
2.4.1	Histogramm	10
2.4.2	Boxplot	12
3	Deskriptive Statistik - Zweidimensionale Daten	14
3.1	Streudiagramme	14
3.1.1	Streudiagramm in R	14
3.2	Abhängigkeit und Kausalität	15
3.3	Einfache lineare Regression	15
3.3.1	Methode der kleinsten Quadrate	15
3.3.2	Empirische Korrelation	17
4	Wahrscheinlichkeit	18
4.1	Wahrscheinlichkeitsmodelle	18
4.1.1	Definition Wahrscheinlichkeitsmodelle	18
4.2	Disjunkte Ereignisse	19
4.3	Axiome und Rechenregeln der Wahrscheinlichkeitsrechnung	19
4.3.1	Rechenregeln	19
4.4	Diskrete Wahrscheinlichkeit	19
4.5	Laplace Wahrscheinlichkeit	20
4.6	Der Begriff der Unabhängigkeit	20
5	Zufallsvariable	21
5.1	Wahrscheinlichkeitsverteilung einer Zufallsvariable	21
5.2	Kennzahlen einer Verteilung	21
5.2.1	Standardabweichung mit R	22
5.3	Unterschied empirischer und theoretischer Kennzahlen	22
5.3.1	Unterschied Mittelwert und Erwartungswert	22

6	Bedingte Wahrscheinlichkeit	23
6.1	Rechenregeln der bedingten Wahrscheinlichkeit	23
6.2	Bayes Theorem und totale Wahrscheinlichkeit	23
6.2.1	Bayes' Theorem	23
6.2.2	Totale Wahrscheinlichkeit	23
7	Normalverteilung	24
7.1	Stetige Zufallsvariable und Wahrscheinlichkeitsverteilungen	24
7.1.1	Stetige Verteilungen	24
7.1.2	Wahrscheinlichkeitsdichte	24
7.1.3	Quantile	24
7.1.4	Kennzahlen von stetigen Verteilungen	25
7.2	Normalverteilung (Gaussverteilung)	25
7.2.1	Graphische Darstellung der Normalverteilung	25
7.2.2	Standardnormalverteilung	26

Abbildungsverzeichnis

2.1	Graphische Darstellung des arithmetischen Mittelwertes	7
2.2	“Grosse” und “kleine” Streuung von zwei Messreihen	7
2.3	Die Robustheit des Median	8
2.4	Vergleich der Histogramme mit verschiedener Klassenwahl	11
2.5	Bimodales Verhalten in zwei Histogrammen	11
2.6	Symmetrisches, rechts- und linksschiefes Histogramm	12
2.7	Schematischer Aufbau eines Boxplots	13
3.1	Streudiagramm für die Mortalität und Weinkonsum in 18 Länder	14
3.2	Residuen für Buchpreis in Abhängigkeit der Seitenanzahl	16
3.3	Streudiagramm mit Regressionsgeraden aus obigem R Code	16
4.1	Wahrscheinlichkeit für nicht disjunkte Ereignisse	19
7.1	Quantil q_α anhand der Dichte $f(x)$ für $\alpha = 0.75$	25

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 (z.B. 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 x_1, x_2, \dots, x_n 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 oder 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

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

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

```
1 # Vektor (Datensatz) bilden
2 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)
3 mean(waageA)
4 # [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 zweiten. Wir sprechen von unterschiedlicher *Streuung* der Daten um den Durchschnitt.

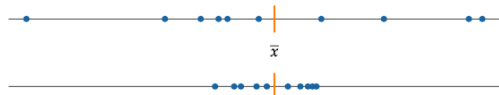


Abbildung 2.2: “Grosse” und “kleine” Streuung von zwei Messreihen

Die gebräuchlichsten Masse für die Streuung oder Variabilität 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 Büchern 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 $n - 1$.

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

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

```
1 var(waageA)
2 # [1] 0.000574359
3 sd(waageA) # sd = standard deviation
4 ## [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

```
1 median(waageA)
2 ## [1] 80.03
```

2.3.3.2 Bemerkungen 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, sondern 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

```
1 # Syntax fuer das untere Quartil: p = 0.25, type definiert den verwendeten Algorithmus
2 # https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile
3 quantile(waageA, p = 0.25, type = 2)
4 ## 25%
5 ## 80.02
6
7 # Syntax fuer das obere Quartil: p = 0.75
8 quantile(waageA, p = 0.75, type = 2)
9 ## 75%
10 ## 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

```
1 IQR(waageA, type = 2)
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 α -Quantile* ist derjenige Wert, bei dem $\alpha * 100$ der Datenpunkte kleiner oder gleich und $(1 - \alpha) * 100$ der Punkte grösser oder gleich diesem Wert sind.

2.3.6.1 Quantil in R

```
1 quantile(waageA, p = 0.1, type = 2)
2 ## 10%
3 ## 79.98
4
5 quantile(waageA, p = 0.7, type = 2)
6 ## 70%
7 ## 80.04
```

Weiteres Beispiel mit versch. Quantilen in einer Zeile

```
1 quantile(noten, p = seq(from = 0.2, to = 1, by = 0.2), type = 2)
2 ## 20% 40% 60% 80% 100%
3 ## 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 Histogramm in R

```
1 iq <- rnorm(n = 200, mean = 100, sd = 15)
2 hist(iq,
3     col = "darkseagreen3",
4     xlab = "Punkte im IQ-Test",
5     ylab = "Anzahl Personen",
6     main = "Verteilung der Punkte in einem IQ-Test",
7     breaks = "sturges" # default, sonst INT-Value
8 )
```

- `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.

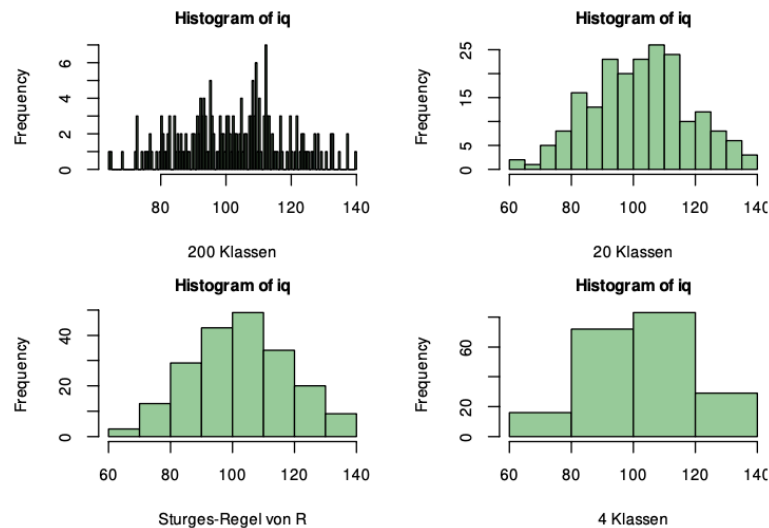


Abbildung 2.4: Vergleich der Histogramme mit verschiedener Klassenwahl

2.4.1.2 Bimodales Verhalten

Bimodales Verhalten ist sichtbar, wenn es zwei “Hügel” im Histogramm gibt

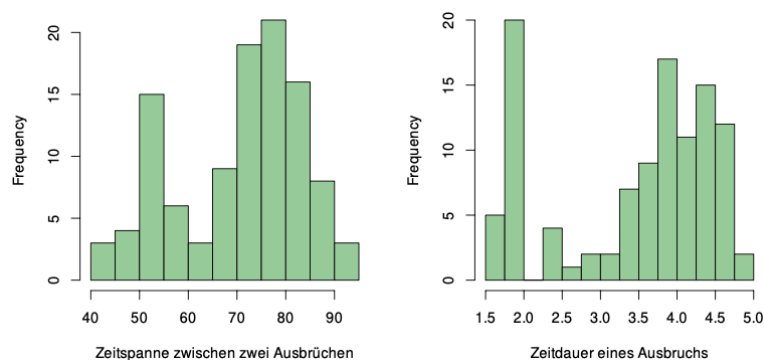


Abbildung 2.5: Bimodales Verhalten in zwei Histogrammen

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

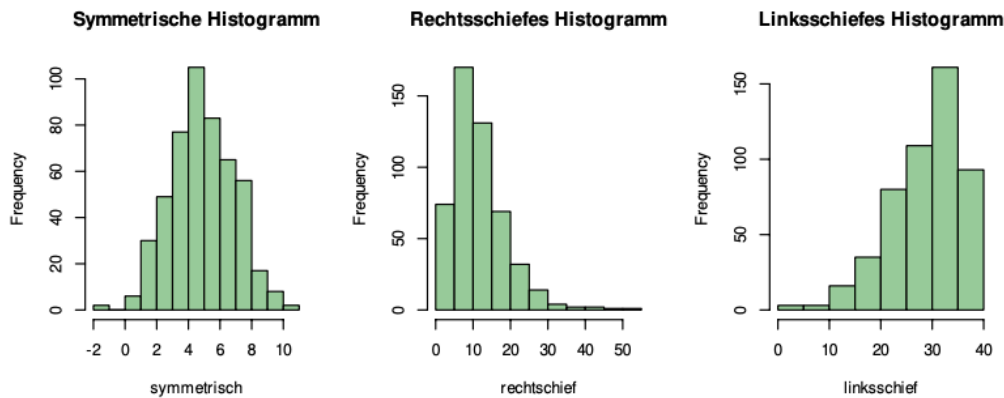


Abbildung 2.6: Symmetrisches, rechts- und linksschiefes Histogramm

- 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 jeweiligen 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).

```

1 hist(waageA,
2     freq = F,
3     main = "Histogramm von Waage A",
4     col = "darkseagreen3",
5     ylim = c(0, 25)
6 )
7 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 welchem Bereich die vertikale Achse gezeichnet werden soll
- `rect` zeichnet ein Rechteck in eine vorgegebene Grafik. Die ersten beiden 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

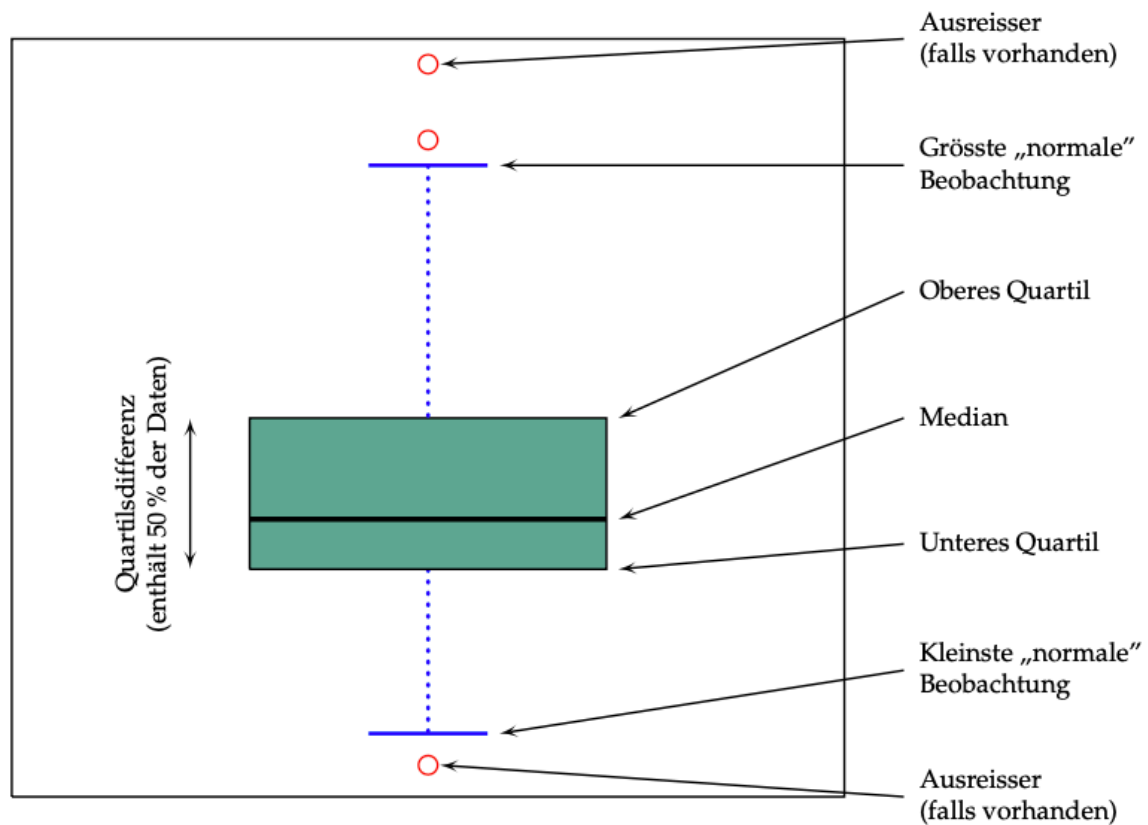


Abbildung 2.7: Schematischer Aufbau eines Boxplotes

2.4.2.1 Boxplot in R

```

1 boxplot(waageA,
2   col = "darkseagreen3"
3 )

```

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

3 Deskriptive Statistik - Zweidimensionale Daten

Bei zweidimensionalen Daten werden an *einem* Versuchsobjekt jeweils *zwei* verschiedene Grössen ermittelt. Als Beispiel dient uns das *Versuchsobjekt* Mensch mit den Messungen zu der *Körpergrösse* und *Körpergewicht*.

3.1 Streudiagramme

Zweidimensionale Daten werden häufig mit *Streudiagrammen* (Scatterplot) dargestellt. Dabei werden die beiden Messungen einer Versuchseinheit als *Koordinaten* in einem Korrdinatensystem interpretiert und dargestellt. Sind die Daten in dieser Form gegeben, interessieren wir uns in erster Linie für die *Zusammenhänge* und *Abhängigkeiten* zwischen den beiden Variablen.

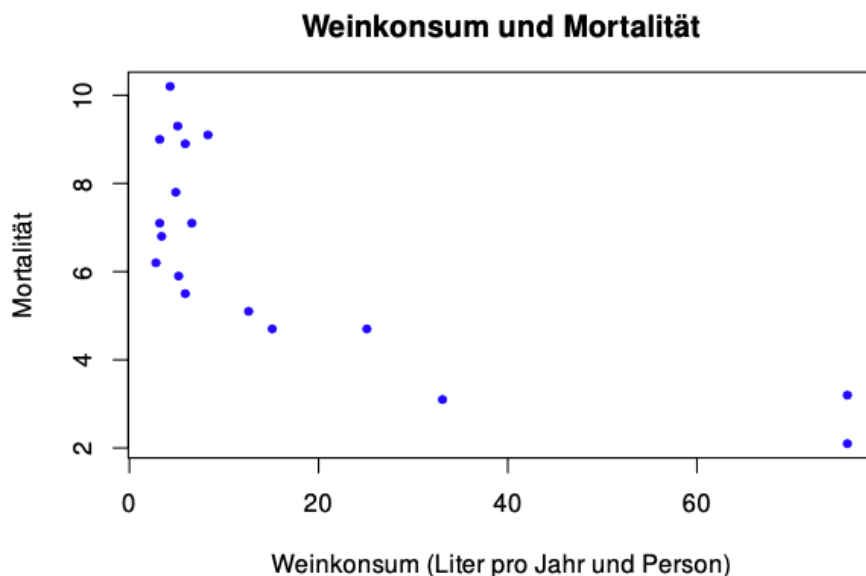


Abbildung 3.1: Streudiagramm für die Mortalität und Weinkonsum in 18 Länder

Dabei ist ein Punkt als (x_1, y_1) mit den Grössen $x = \text{Weinkonsum}$ und $y = \text{Mortalität}$ zu interpretieren.

- Einfluss auf andere Körperorgane wird hier nicht berücksichtigt
- *Kausaler* Zusammenhang muss nicht zwingend zwischen den beiden Grössen vorhanden sein
- Die Zuweisung Grösse und X/Y-Achse könnte auch vertauscht werden. Die Entscheidung hängt von der Problemstellung ab
- Die Punkte im Streudiagramm werden auch als *Punktwolke* bezeichnet

3.1.1 Streudiagramm in R

```
1  wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9, 6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9)
2  mort <- c(6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5, 7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1)
3  plot(wein,
```

```

4   mort,
5   xlab = "Weinkonsum (Liter pro Jahr und Person)",
6   ylab = "Mortalität",
7   main = "Zusammenhang zwischen Weinkonsum und Mortalität", pch = 20,
8   col = "blue"
9 )

```

3.2 Abhängigkeit und Kausalität

Bei Streudiagrammen müssen wir aufpassen, dass *Abhängigkeit* und *Kausalität* nicht miteinander verwechselt werden. Nur weil eine Gesetzmässigkeit vorhanden ist, heisst das nicht, dass diese Gesetzmässigkeit auch kausal erklärt werden kann. Man muss sich *bewusst* sein, auf *welchen Daten* das Streudiagramm basiert. Man soll sich *nie* blindlings auf Grafiken verlassen. Die Daten müssen auf anderen Weg auf einen kausalen Zusammenhang untersucht werden.

3.3 Einfache lineare Regression

Weil wir wissen möchten, *wie* sich Daten verhalten, versuchen wir einem Muster eine Form zu geben. Dies kann eine Gerade sein. Die Beschreibung dieser Form geschieht in der Sprache der Mathematik. Dabei wird auch von einer *Modellierung* der Daten gesprochen.

Liegt ein Modell vor, können wir *Vorhersagen* machen. Dieses Modell erlaubt uns, Daten die nicht exakt *gemessen* vorliegen abzuschätzen. Als Beispiel die Wettervorhersage oder den Preis einer Ware bei x-Stücken.

3.3.1 Methode der kleinsten Quadrate

Wie finden wir nun die Gerade die *möglichst gut* zu allen Punkten passt? Dazu verwenden wir den Begriff **Residuum**. Ein *Residuum* r_i ist die vertikale Differenz zwischen einem Datenpunkt (x_i, y_i) und dem Punkt $(x_i, a + bx_i)$ auf der gesuchten Geraden:

$$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

Bei der Methode der kleinsten Quadrate werden die Summen der *Quadrate der Abweichungen* aufsummiert $r_1^2 + r_2^2 + \dots + r_n^2 = \sum_i r_i^2$. Die Parameter a und b werden so gewählt, dass die Summe minimal wird. Eine Gerade passt also dann am besten zu den Punkten im Streudiagramm, wenn die Summe der Quadrate der vertikalen Abweichungen minimal ist (Optimierungsproblem).

3.3.1.1 Gerade mit R

```

1   seite <- c(seq(50, 500, 50))
2   preis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5, 26.1, 29.1)
3   plot(seite,
4        preis,
5        xlab = "Seitenzahl",
6        ylab = "Buchpreis",
7        pch = 16,
8        col = "blue"
9   )

```

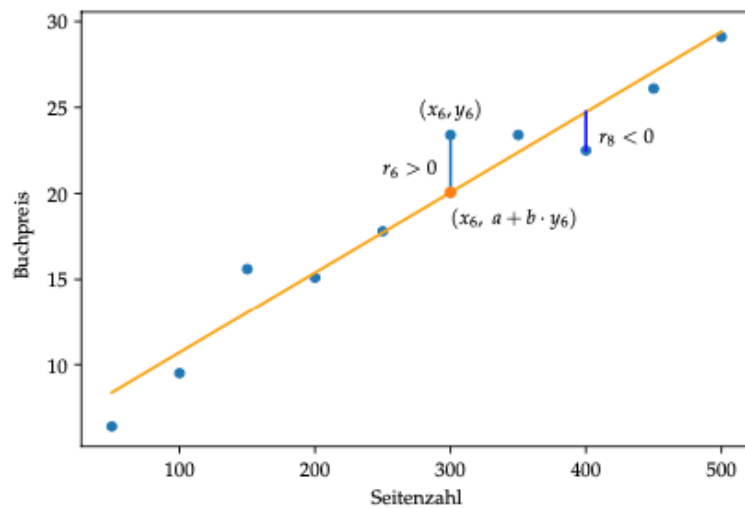


Abbildung 3.2: Residuen für Buchpreis in Abhängigkeit der Seitenanzahl

```

10 abline(lm(preis ~ seite), col = "orange")
11
12 lm(preis ~ seite)
13 # Call:
14 # lm(formula = preis ~ seite)
15 #
16 # Coefficients:
17 # (Intercept)    seite
18 #   6.04000    0.04673
19 # y = 6.04 + 0.047x

```

- `lm()` steht für *linear model*
- Mit `lm(y ~ x)` passt R ein Modell der Form $y = a + bx$ an die Daten an
- Die Variablen sind verglichen mit dem `plot(x,y)` vertauscht
- Die Gerade wird *Regressionsgerade* genannt

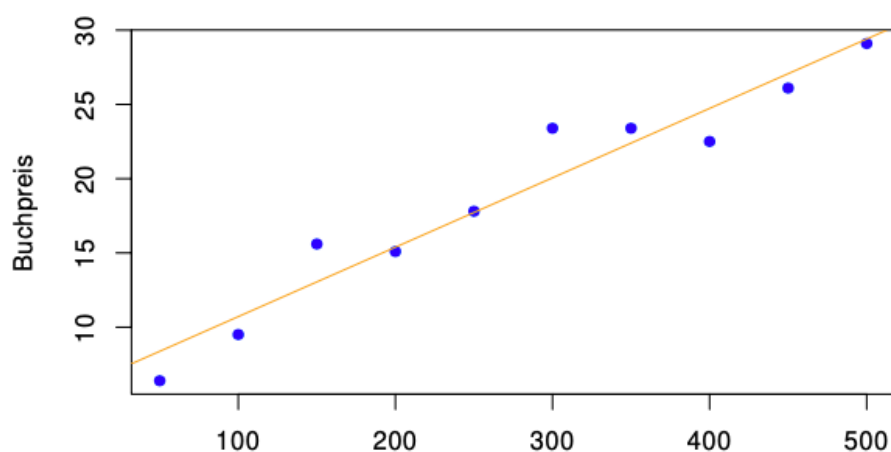


Abbildung 3.3: Streudiagramm mit Regressionsgeraden aus obigem R Code

3.3.1.2 Extrapolationen

Extrapolationen sind Vorhersagen der y -Werte des Modelles, die *ausserhalb* des Bereiches (x) liegen, womit das Modell erstellt wurde. Für Extrapolationen, die weit ausserhalb des gültigen Bereichs liegen, können die Vorhersagen problematisch wenn nicht sogar sinnlos werden.

3.3.1.3 Interpolationen

Interpolationen sind Vorhersagen der y -Werte des Modelles, die *innerhalb* des Bereiches liegt, womit das Modell erstellt wurde. Die Interpolationen sind unproblematischer, aber auch nur *begrenzt* gültig.

3.3.2 Empirische Korrelation

Die *empirische Korrelation* (r als Kennzahl oder auch \hat{p} ist die quantitative Zusammenfassung der *linearen* Abhängigkeit von zwei Grössen. Es ist eine dimensionslose Zahl zwischen -1 und 1 und misst die Stärke und die Richtung der *linearen Abhängigkeit* zwischen den Daten x und y . *Wichtig:* Der Korrelationskoeffizient misst (erkennt) nur den linearen Zusammenhang!

- Ist $r = +1$, dann liegen Punkte auf einer steigenden Geraden ($y = a + bx$ mit $b > 0$)
- Ist $r = -1$, dann liegen die Punkte auf einer fallenden Geraden ($y = a + bx$ mit $b < 0$)
- Sind x und y unabhängig (es besteht kein Zusammenhang), so ist $r = 0$. Die Umkehrung gilt Allgemein aber nicht!

3.3.2.1 Empirische Korrelation in R

```
1 cor(seite, preis)
2 # [1] 0.968112
```

Dieser Wert liegt sehr nahe bei 1 und somit besteht ein *enger* linearen Zusammenhang. Dazu ist der Wert positiv, was einem “je mehr, desto mehr” entspricht.

4 Wahrscheinlichkeit

4.1 Wahrscheinlichkeitsmodelle

Wir verwenden oft Modelle um die Wahrscheinlichkeit zu bestimmen. Dazu treffen wir Annahmen (z.B. dass ein Würfel fair ist) die dann im Modell umgesetzt/berechnet werden. Mit Hilfe des Modells können wir dann auch untersuchen, ob dieser Würfel fair ist. Wenn wir einen wiederholt werfen und oft die Zahl 2 vorkommt, können wir annehmen, dass der Würfel nicht fair ist.

4.1.1 Definition Wahrscheinlichkeitsmodelle

Wir betrachten *Zufallsexperimente*, bei denen der Ausgang *nicht exakt* vorhersagbar ist (z.B. # Anrufe in Callcenter). Das Wahrscheinlichkeitsmodell beschreibt grob welche Ergebnisse möglich sind und beinhaltet die Wahrscheinlichkeiten, wie die Ergebnisse eintreten können.

Das Wahrscheinlichkeitsmodell besteht aus folgenden Komponenten:

- *Grundraum* Ω , der aus der Menge der *Elementarereignissen* ω besteht
- *Elementarereignisse* sind mögliche Ergebnisse oder Ausgänge des Experiments, die alle zusammen den Grundraum bilden.
- *Ergebnisse* A, B, C, \dots als Teilmengen von Ω
- *Wahrscheinlichkeiten* P , die zu den Ereignissen A, B, C, \dots gehören

4.1.1.1 Grundraum, Elementarereignisse

Bei einem Experiment wird aus dem Grundraum *ein* Elementarereignis *zufällig* gewählt. Als Beispiel das Würfeln. Grundraum gegeben durch

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Element ω ist ein Elementarereignis und bedeutet, dass bei Würfeln die Zahl 2 geworfen wird.

4.1.1.2 Ereignis

Unter einem Ereignis versteht man eine Teilmenge von Ω . Das Ereignis A bedeutet, dass das Ergebnis ω des Experiments zu A gehört.

Beispiel für ein Ereignis A : “eine ungerade Augenzahl würfeln”, dann ist $A = \{1, 3, 5\}$ und tritt ein, wenn eine der drei Zahlen gewürfelt wird. Eine Leere Menge ist auch ein Ereignis $\{\} \subset \Omega$.

4.2 Disjunkte Ereignisse

Zwei Ereignisse A und B heissen *disjunkt*, wenn sich A und B gegenseitig ausschliessen und nicht gemeinsam eintreten können. Dann gilt

$$A \cap B = \{\}$$

und ist somit unmöglich.

4.3 Axiome und Rechenregeln der Wahrscheinlichkeitsrechnung

Die Wahrscheinlichkeitsrechnung baut auf die folgenden drei Grundregeln (Kolmogorov Axiome):

- A1: $P(A) \geq 0$
- A2: $P(\Omega) = 1$
- A3: $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \{\}$

4.3.1 Rechenregeln

Für dieses Modul relevante Rechenregeln:

1. $P(\bar{A}) = 1 - P(A)$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Die zweite Regel für disjunkte Ereignisse. Die Schnittmenge wird doppelt gezählt, weshalb wir diese einmal abziehen müssen.

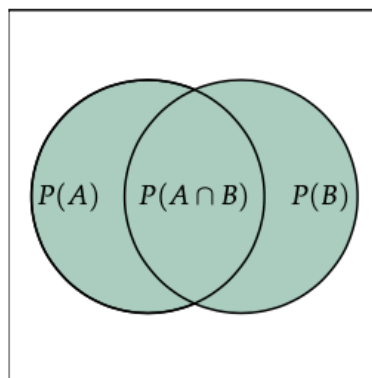


Abbildung 4.1: Wahrscheinlichkeit für nicht disjunkte Ereignisse

4.4 Diskrete Wahrscheinlichkeit

Mit diskret sind endliche und unendliche Mengen gemeint, welche ganzzahlige Elemente in Ω enthalten.

4.5 Laplace Wahrscheinlichkeit

Beim *Modell von Laplace* wird für jedes Elementarereignis die gleiche Wahrscheinlichkeit angenommen. Um diese Wahrscheinlichkeit zu berechnen zählen wir die Anzahl der *günstigen* Elementarereignisse durch die Anzahl der *möglichen* Elementarereignisse. Wenn alle Ereignisse E gleich wahrscheinlich sind, ist das Eintreten des Ereignisses E nach dem Laplace-Modell:

$$P(E) = \frac{|E|}{|\Omega|}$$

4.6 Der Begriff der Unabhängigkeit

Hat der Ausgang von Ereignis A keinen Einfluss auf den Ausgang des Ereignisses B , sind die Ereignisse Ereignisse A und B stochastisch unabhängig. Dann gilt

$$P(A \cap B) = P(A) * P(B)$$

Als Beispiel: A sei mit einem fairen Würfel eine eis oder zwei zu würfeln und Ereignis B sei Kopf beim Werfen einer fairen Münze. Weil die beiden Ereignisse keinen Einfluss aufeinander haben gilt obige Formel.

Sind Ereignisse *nicht* stochastisch unabhängig, gibt es keine allgemeine Formel für die Brechnung der Wahrscheinlichkeit von zwei Ereignissen.

5 Zufallsvariable

Bei einem Zufallsexperiment mit dem Grundraum Ω ordnen wir mit der *Funktion* X jedem Elementarereignis ω von Ω einen *Zahlwert* zu. Die Funktion X wird als *Zufallsvariable* bezeichnet und ordnet damit jedem Element eine *Zahl* zu.

Eine Wertemenge bezeichnet die Werte, welche die Zufallsvariable annehmen kann.

Bemerkungen

- die *Zufallsvariable wird mit einem Grossbuchstaben bezeichnet (X)
- der entsprechende *Kleinbuchstabe* x stellt einen konkreten Wert dar, den die Zufallsvariable annehmen kann (die Zahl)
- für das Ereignis, welches X annimmt, schreiben wir $X = x$
- bei der Zufallsvariable ist nicht die Funktion X zufällig, sondern das Argument ω . Je nach Ausgang erhalten wir einen anderen Wert $X(\omega) = x$
- x wird auch als *Realisierung* der Zufallsvariable X bezeichnet

5.1 Wahrscheinlichkeitsverteilung einer Zufallsvariable

Berechnen wir für *jede* Realisierung einer Zufallsvariable die zugehörige Eintretenswahrscheinlichkeit, so bilden alle diese Wahrscheinlichkeiten zusammen die *Wahrscheinlichkeitsverteilung* dieser Zufallsvariablen. Dabei gilt

$$P(X = x_1) + P(X = x_2) + \dots + P(X = x_n) = 1$$

oder

$$\sum_{i=1}^n P(X = x_n) = 1$$

5.2 Kennzahlen einer Verteilung

Eine beliebige *diskrete* Verteilung kann durch zwei Kennzahlen, den *Erwartungswert* $E(X)$ und die *Standardabweichung* $\sigma(X)$. Der Erwartungswert beschreibt die mittlere Lage der Verteilung:

$$E(X) = x_1 * P(X = x_1) + x_2 * P(X = x_2) + \dots + x_n * P(X = x_n) = \sum_{i=1}^n x_i * P(X = x_i)$$

Die Standardabweichung oder *Varianz* beschreibt die Streuung der Verteilung.

$$\begin{aligned} Var(X) &= \sum_{i=1}^n (x_n - (E(X))^2 * P(X = x_n)) \\ \sigma(X) &= \sqrt{Var(x)} \end{aligned}$$

5.2.1 Standardabweichung mit R

```
1 x <- 1 : 6
2 p <- 1 / 6
3 E_X <- sum(x * p)
4 var_X <- sum((x - E_X)^2 * p)
5 sd_X <- sqrt(var_X)
6 sd_X
7 ## [1] 1.707825
```

Beispiel eines nicht-fairen Würfels auch mit R berechnet:

```
1 x <- 1 : 6
2 p <- c(4, 2, 1, 3, 1, 1) / 12
3 E_X <- sum(x * p)
4 E_X
5 ## [1] 2.833333
6 var_X <- sum((x - E_X)^2 * p)
7 sd_X <- sqrt(var_X)
8 sd_X
9 ## [1] 1.674979
```

Der Erwartungswert ist 2.8333 und die Standardabweichung ist 1.675 (die Werte weichen im “Durchschnitt” so viel ab).

5.3 Unterschied empirischer und theoretischer Kennzahlen

5.3.1 Unterschied Mittelwert und Erwartungswert

Der arithmetische Mittelwert \bar{x} wird aus *konkreten* Daten gewonnen. Wir haben also Messwerte x_1, \dots, x_n und können \bar{x} berechnen. Der Erwartungswert σ_X ist ein *theoretischer* Wert, der sich aus dem Modell der Wahrscheinlichkeitsverteilung ergibt.

Die Hoffnung ist, dass sich das arithmetische Mittel für immer mehr Versuche an den theoretischen Wert annähert, sofern sich die konkreten Daten der Wahrscheinlichkeitsverteilung von X folgen.

Derselbe Unterschied wie für Mittelwert und Erwartungswert gilt auch für die *empirische Standardabweichung* s_X und die *Standardabweichung* σ_X .

6 Bedingte Wahrscheinlichkeit

Die *bedingte Wahrscheinlichkeit* betrachtet nicht die gesamte Grundmenge, sondern nur einen Teil davon. Die neue Grundmenge in der Formel ist die Variable nach dem Längsstrich. Die bedingte Wahrscheinlichkeit bedeutet die Wahrscheinlichkeit von A unter der Voraussetzung von B.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Da wir wissen, dass B schon eingetreten ist, haben wir einen neuen Grundraum $\Omega' = B$. Damit müssen wir nur noch den Teil anschauen, der auch in B auftritt (daher $A \cap B$). Dies muss dann noch in Relation zur Wahrscheinlichkeit von B (neue Grundmenge) gesetzt werden.

6.1 Rechenregeln der bedingten Wahrscheinlichkeit

! [Rechenregeln der bedingten Wahrscheinlichkeit] (rechenregeln_bedWahrscheinlichkeit.png) {width=60%}

6.2 Bayes Theorem und totale Wahrscheinlichkeit

6.2.1 Bayes' Theorem

Das Theorem ist oft sehr nützlich. Es erlaubt uns die Wahrscheinlichkeit $P(A|B)$ zu berechnen, falls $P(B|A)$.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

6.2.2 Totale Wahrscheinlichkeit

Bei der totalen Wahrscheinlichkeit wird erst eine Menge A in Mengen A_1, \dots, A_k unterteilt, die untereinander keine Schnittmenge haben und zusammen (Vereinigung) die ganze Menge A bilden. Dies wird *Partitionierung* genannt.

$$P(B) = P(B|A_1) * P(A_1) + \dots + P(B|A_k) * P(A_k) = \sum_{i=1}^k P(B|A_i) * P(A_i)$$

7 Normalverteilung

7.1 Stetige Zufallsvariable und Wahrscheinlichkeitsverteilungen

Der Unterschied einer *stetigen* und *diskreten* Zufallsvariable besteht darin, dass die stetige *jeden* Wert eines bestimmten Bereiches annehmen. Diskrete hingegen aus einer definierten und endlicher Menge (z.B. von 0 bis 100, jeweils ganze Zahlen). Wichtig auch, es kann kein Wert zwischen zwei Werten aus einer Wertemenge ausgewählt werden. Dabei gilt:

- die Variable X ist eine Funktion
- die Variable x ist ein konkreter Wert (*Realisierung*) von X

7.1.1 Stetige Verteilungen

Stetig sind die Verteilungen wenn keine Lücken in einem Bereich vorhanden sind. Die jeweilige Wahrscheinlichkeit ist $P(X = x) = 0$. Diese werden *Punktwahrscheinlichkeiten* genannt. Diese Wahrscheinlichkeit bringt uns aber nicht weiter. Weshalb wird die Wahrscheinlichkeit zwischen zwei Punkten berechnen. Dazu wird das Konzept der *Wahrscheinlichkeitsdichte* angewendet.

7.1.2 Wahrscheinlichkeitsdichte

Wahrscheinlichkeitsdichten können unter der Berücksichtigung folgender Einschränkungen fast beliebig aussehen. Relevant für uns sind aber lediglich die Normalverteilung und die dazu verwandte t -Verteilung.

7.1.2.1 Eigenschaften Wahrscheinlichkeitsdichte

Für eine Wahrscheinlichkeitsdichte $f(x)$ gelten folgende Eigenschaften:

- es gilt $f(x) \geq 0 \rightarrow$ Kurve auf oder oberhalb der x-Achse
- die Wahrscheinlichkeit $P(a < X \leq b)$ entspricht der Fläche zwischen a und b unter $f(x)$
- die gesamte Fläche unter der Kurve ist 1 \rightarrow W'keit dass *irgendein* Wert gemessen wird.

[Dichte einer Zufallsvariable und der Wahrscheinlichkeit in ein Intervall $[a,b]$ zu fallen (grüne Fläche).](dichte_zv.png){width=60%}

Bei stetigen Wahrscheinlichkeitsverteilungen entsprechen Wahrscheinlichkeiten den Flächen unter der Dichtefunktion.

7.1.3 Quantile

Bei stetigen Verteilungen ist das α -Quantil q_α derjenige Wert, wo die Fläche unter der Dichtefunktion von $-\infty$ bis q_α gerade α entspricht. Das 50%-Quantil heisst der *Median*.

Dies bedeutet, dass 75% einer Menge maximal diese Messgrösse erreichen.

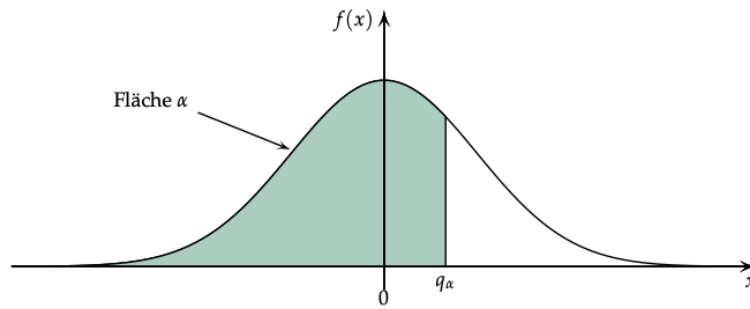


Abbildung 7.1: Quantil q_α anhand der Dichte $f(x)$ für $\alpha = 0.75$

7.1.4 Kennzahlen von stetigen Verteilungen

Der Erwartungswert $E(X)$ und die Standardabweichung σ_X werden gleich wie im diskreten Fall interpretiert:

- $E(X)$ beschreibt die mittlere Lage der Verteilung
- σ_X beschreibt die Streuung der Verteilung um den Erwartungswert

Im stetigen Fall, sind die beiden Funktionen aber mit Integralen statt Summen definiert.

7.2 Normalverteilung (Gaussverteilung)

Die *Normalverteilung* ist die häufigste Verteilung für Messwerte und spielt vor allem für Durchschnitte von Messwerten eine wichtige Rolle. Wichtige Wahrscheinlichkeitsverteilung in der Statistik.

Normalverteilung

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Es ist folgende Schreibweise für die Verteilung einer normalverteilten Zufallsvariable X mit Parameter μ und σ zu verwenden:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

- $E(X) = \mu$
- $Var(X) = \sigma^2$
- $\sigma_X = \sigma$

Die bedeutet, dass die Parameter μ und σ^2 eine natürliche Interpretation als Erwartungswert und Varianz einer Verteilung haben.

7.2.1 Graphische Darstellung der Normalverteilung

Alle Normalverteilungskurven haben folgende Eigenschaften:

- Die Kurven sehen aus wie “Glocken”, weswegen man auch von *Glockenkurven* spricht
- Die Wahrscheinlichkeitsdichtefunktion der Normalverteilung ist symmetrisch um den Erwartungswert μ
- Der Parameter μ verschiebt den Graphen um μ -Einheiten nach links ($\mu < 0$) oder nach rechts ($\mu > 0$)
- Je grösser σ , desto flacher bzw. breiter wird die Dichtekurve. Je näher bei 0, wird die Kurve umso spitzer.

Der Grund dafür, weil σ die Streuung um den Erwartungswert μ angibt. Je grösser σ um so mehr sind die Werte um den Erwartungswert μ verteilt, die Kurve wird also breiter. Ist σ nahe bei 0, so weichen die Werte wenig von μ ab, die Kurve wird schmaler und höher.

7.2.1.1 Wahrscheinlichkeiten mit R

```
1 pnorm(q = 130, mean = 100, sd = 15)
2 ## [1] 0.9772499
```

Dieser Befehl ermittelt die Fläche (Wahrscheinlichkeit von $-\infty$ bis $q = 130$ unter der Normalverteilungskurve mit $\mu = 100$ und $\sigma = 15$

```
1 qnorm(p = c(0.025, 0.975), mean = 100, sd = 15)
2 ## [1] 70.60054 129.39946
```

`qnorm()` bestimmt die Quantile für die Normalverteilung. Bei diesem Beispiel haben wir den Wertebereich von 95% aller Beobachtungen eruiert. Diese liegen zwischen 70 bis 130.

Ist eine Zufallsvariable normalverteilt, so liegen etwa zwei Drittel aller Messerte etwa eine Standardabweichung um den Erwartungswert.

Für *alle* Normalverteilungen $\mathcal{N}(\mu, \sigma^2)$ gilt, dass die Wahrscheinlichkeit, dass eine Beobachtung höchstens *eine* Standardabweichung vom Erwartungswert abweicht, ist etwa $\frac{2}{3}$:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx \frac{2}{3}$$

Die Wahrscheinlichkeit, dass eine Beobachtung höchstens *zwei* Standardabweichungen vom Erwartungswert abweicht ist:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

Aus diesen beiden Wahrscheinlichkeiten lassen sich auch als Flächen interpretieren. Die Fläche der Normalverteilung über dem Intervall $[\mu - \sigma, \mu + \sigma]$ ist ca. $\frac{2}{3}$. Die Fläche über dem Intervall $[\mu - 2\sigma, \mu + 2\sigma]$ ist ca. 0.95.

7.2.2 Standardnormalverteilung

Die Normalverteilung $\mathcal{N}(0, 1)$ mit Mittelwert 0 und Standardweichung 1 heisst *Standardnormalverteilung*. Falls $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist die standardisierte Zufallsvariable wieder normalverteilt, hat nun aber den Erwartungswert 0 und die Varianz 1. Man erhält also die Standardnormalverteilung:

$$Z = \frac{Z - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$