

Applied Statistics for Data Science

Zusammenfassung

Stephan Stofer

13. Januar 2021

Inhaltsverzeichnis

1	Einleitung	6
1.1	Vierstufige Problemlösungsstrategie	6
2	Deskriptive Statistik - Eindimensionale Daten	7
2.1	Datensätze	7
2.1.1	Liste	7
2.1.2	Tabellen	7
2.2	Deskriptive Statistik	7
2.2.1	Bezeichnung von Daten	7
2.3	Kennzahlen	8
2.3.1	Arithmetisches Mittel	8
2.3.2	Empirische Varianz und Standardabweichung	8
2.3.3	Median	9
2.3.4	Quartile	10
2.3.5	Quartilsdifferenz	10
2.3.6	Quantile	11
2.4	Graphische Methoden	11
2.4.1	Histogramm	11
2.4.2	Boxplot	13
3	Deskriptive Statistik - Zweidimensionale Daten	15
3.1	Streudiagramme	15
3.1.1	Streudiagramm in R	15
3.2	Abhängigkeit und Kausalität	16
3.3	Einfache lineare Regression	16
3.3.1	Methode der kleinsten Quadrate	16
3.3.2	Empirische Korrelation	18
4	Wahrscheinlichkeit	19
4.1	Wahrscheinlichkeitsmodelle	19
4.1.1	Definition Wahrscheinlichkeitsmodelle	19
4.2	Disjunkte Ereignisse	20
4.3	Axiome und Rechenregeln der Wahrscheinlichkeitsrechnung	20
4.3.1	Rechenregeln	20
4.4	Diskrete Wahrscheinlichkeit	20
4.5	Laplace Wahrscheinlichkeit	21
4.6	Der Begriff der Unabhängigkeit	21
5	Zufallsvariable	22
5.1	Wahrscheinlichkeitsverteilung einer Zufallsvariable	22
5.2	Kennzahlen einer Verteilung	22
5.2.1	Standardabweichung mit R	23
5.3	Unterschied empirischer und theoretischer Kennzahlen	23
5.3.1	Unterschied Mittelwert und Erwartungswert	23

6	Bedingte Wahrscheinlichkeit	24
6.1	Rechenregeln der bedingten Wahrscheinlichkeit	24
6.2	Bayes Theorem und totale Wahrscheinlichkeit	24
6.2.1	Bayes' Theorem	24
6.2.2	Totale Wahrscheinlichkeit	24
7	Normalverteilung	25
7.1	Stetige Zufallsvariable und Wahrscheinlichkeitsverteilungen	25
7.1.1	Stetige Verteilungen	25
7.1.2	Wahrscheinlichkeitsdichte	25
7.1.3	Quantile	26
7.1.4	Kennzahlen von stetigen Verteilungen	26
7.2	Normalverteilung (Gaussverteilung)	26
7.2.1	Graphische Darstellung der Normalverteilung	27
7.2.2	Standardnormalverteilung	27
8	Durchschnitte und Summen von Zufallsvariablen	28
8.1	Unabhängigkeit und i.i.d. Annahme	28
8.2	Kennzahlen von S_n und \bar{X}_n	28
8.2.1	Varianz und Standardabweichung der Summe	29
8.2.2	Erwartungswert des Durchschnittes	29
8.3	Verteilungen von S_n und \bar{X}_n	30
9	Hypothesentest für Messdaten	32
9.1	Statistische Tests und Vertrauensintervall für eine Stichprobe bei normalverteilten Daten	32
9.1.1	Ziel des Hypothesentests	32
9.2	Hypothesentest	32
9.2.1	Modell	33
9.2.2	Nullhypothese	33
9.2.3	Alternativhypothese	33
9.2.4	Teststatistik	33
9.2.5	Signifikanzniveau α	34
9.2.6	Verwerfungsbereich	34
9.2.7	p-Wert	34
9.2.8	p-Wert und Statistischer Test	34
9.3	t-Test	35
9.3.1	t-Verteilung	35
10	Vertrauensintervall, Zweistichprobentest und Wilcoxon-Test	36
10.1	Vertrauensintervall für μ	36
10.2	Der Wilcoxon-Test	36
10.3	Statistische Tests bei zwei Stichproben	36
10.3.1	Gepaarte Stichproben	36
10.3.2	Ungepaarte Stichproben	37
10.4	Tests mit R bei zwei Stichproben	37
11	Lineare Regression	38
11.1	Das einfache Regressionsmodell	38
11.2	Vertrauensintervall im linearen Modell	39
11.3	Hypothesentest im linearen Modell	39
11.4	Abschätzung der Genauigkeit des Modells	40
11.4.1	R^2 -Statistik	40

11.4.2 Alternative Darstellung von R^2	40
12 Multiple lineare Regression	42
12.1 Graphische Interpretation	42
12.2 Koeffizienten mit R	43
12.3 Wichtige Fragestellungen	43
12.3.1 Zusammenhang zwischen erklärenden Variablen und der Zielgrösse	43
12.3.2 Passung des Modells zu den Daten	44
12.3.3 Vorhersagen	44
12.4 Erweiterung des linearen Modells	44
12.5 Interpretation des Interaktionsterm	45
12.6 Qualitative erklärende Variablen	45
13 Variablenselektion	46
13.1 Schrittweise Vorwärtsselektion	46
13.2 Schrittweise Rückwärtsselektion	46
13.3 Anzahl Variablen / Abbruchbedingung	47

Abbildungsverzeichnis

2.1	Graphische Darstellung des arithmetischen Mittelwertes	8
2.2	«Grosse» und «kleine» Streuung von zwei Messreihen	8
2.3	Die Robustheit des Median	9
2.4	Vergleich der Histogramme mit verschiedener Klassenwahl	12
2.5	Bimodales Verhalten in zwei Histogrammen	12
2.6	Symmetrisches, rechts- und linksschiefes Histogramm	13
2.7	Schematischer Aufbau eines Boxplots	14
3.1	Streudiagramm für die Mortalität und Weinkonsum	15
3.2	Residuen für Buchpreis in Abhängigkeit der Seitenanzahl	17
3.3	Streudiagramm mit Regressionsgeraden aus obigem R Code	17
4.1	Wahrscheinlichkeit für nicht disjunkte Ereignisse	20
6.1	Rechenregeln der bedingten Wahrscheinlichkeit	24
7.1	Dichte einer Zufallsvariable und der Wahrscheinlichkeit in ein Intervall (a, b) zu fallen (grüne Fläche)	25
7.2	Quantil q_α anhand der Dichte $f(x)$ für $\alpha = 0.75$	26
8.1	Regeln für die Kennzahlen von S_n und \bar{X}_n	29
8.2	Gesetz der grossen Zahlen	30
8.3	Standardfehler des arithmetischen Mittels	30
8.4	Zentraler Grenzwertsatz	30
8.5	4 Histogramme vom Durchschnitt von 16, 64, 256 und 1024 Versuchen mit je 1000 Ziehungen, inkl. Dichtekurven	31
9.1	Normalverteilungskurve eines Hypothesentests	33
9.2	Wahrscheinlichkeit, unter Gültigkeit der Nullhypothese, das erhaltene Ergebnis oder ein extremeres zu erhalten	34
11.1	Null- und Alternativhypothese graphisch	39
11.2	Definition von R^2	40
12.1	Datenpunkte im Raum mit zwei erklärenden Variablen	42

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, wie 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 ist 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.

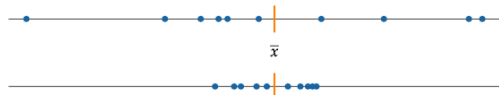


Abbildung 2.2: «Grosse» und «kleine» Streuung von zwei Messreihen

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, ist bei grossen jedoch vernachlässigbar. R verwendet mit dem Befehl `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 $\alpha * 100$ 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.

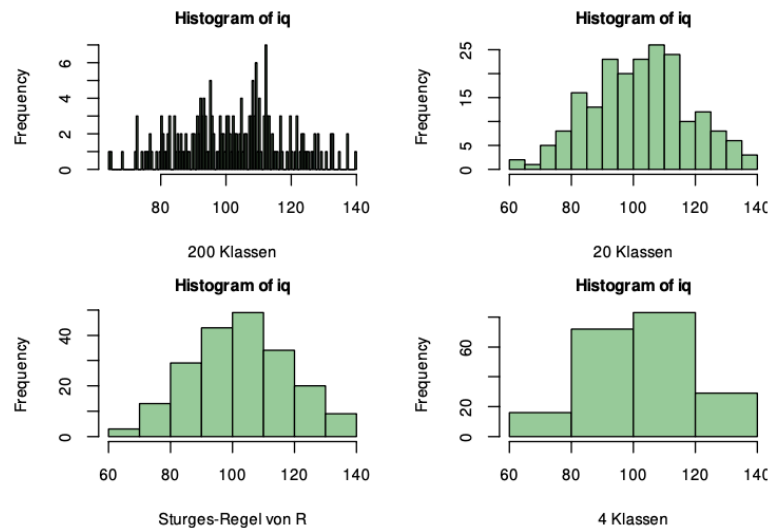


Abbildung 2.4: Vergleich der Histogramme mit verschiedener Klassenwahl

2.4.1.2 Bimodales Verhalten

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

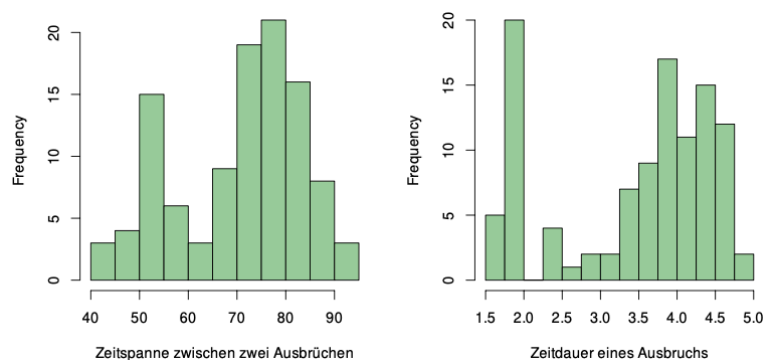


Abbildung 2.5: Bimodales Verhalten in zwei Histogrammen

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

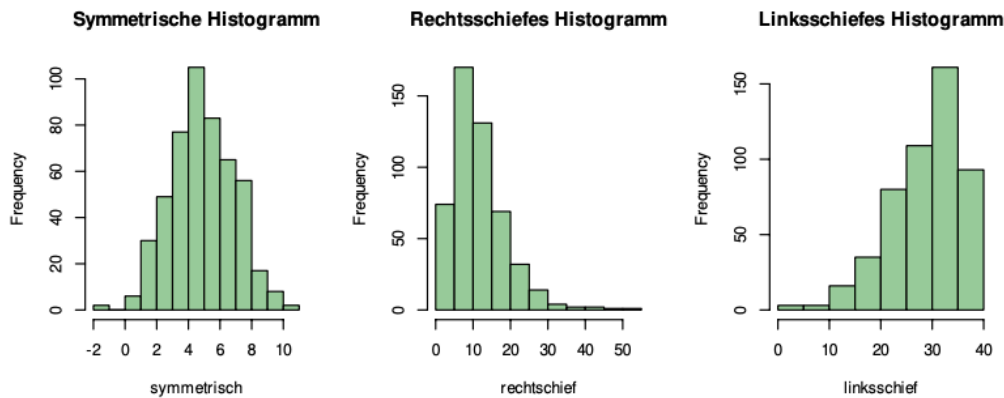


Abbildung 2.6: Symmetrisches, rechts- und linksschiefes Histogramm

- 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)
- oberhalb des Medians sind 25% der Daten bis zum oberen Quantil, 25% der Messungen von Median zum unteren Quantil
- *whiskers*, blaue Linien, die zur kleinsten und grössten «normalen» Beobachtung führen (normal heisst *höchstens* 1.5 mal die Quartilsdifferenz von oberen und unteren Quantil)

- kleine roten Kreise, für die Ausreisser, welche ausserhalb der normalen Beobachtungen liegen

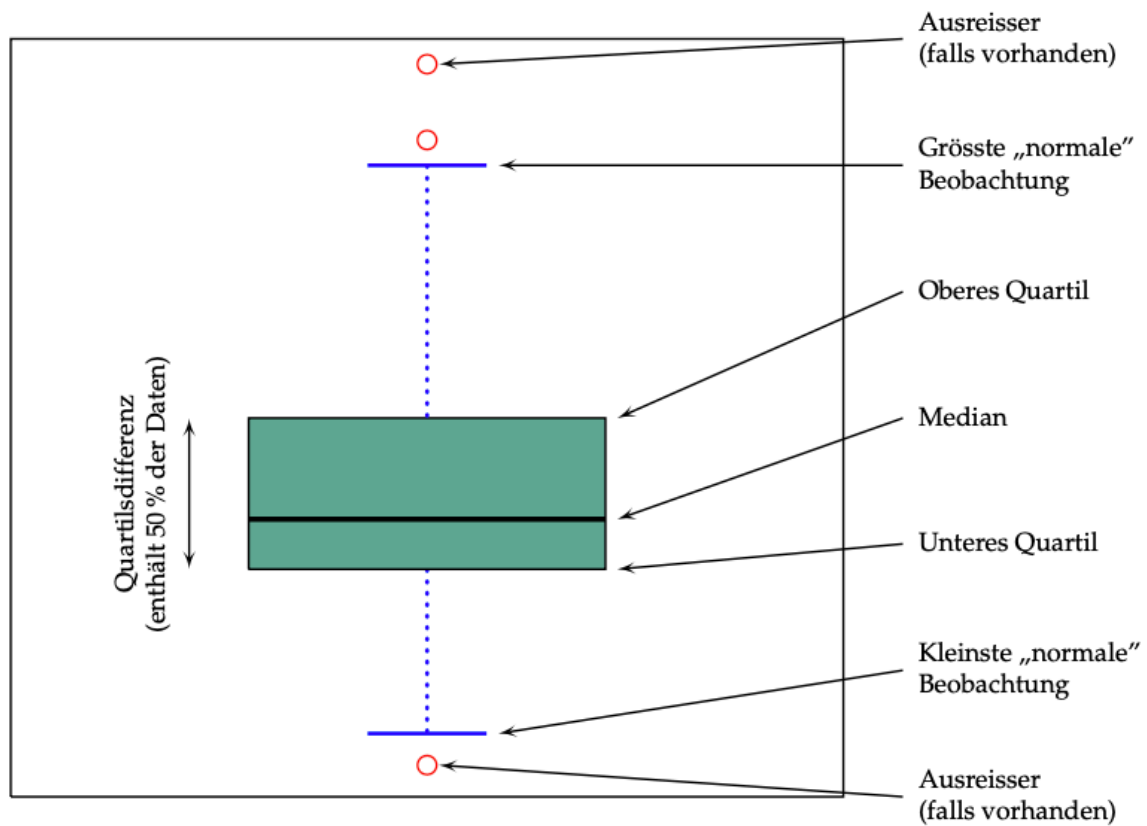


Abbildung 2.7: Schematischer Aufbau eines Boxplots

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.

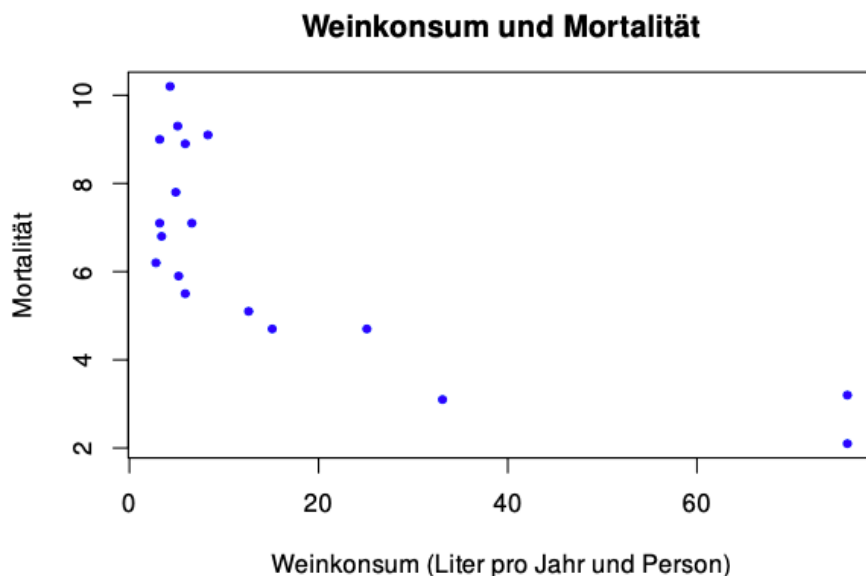


Abbildung 3.1: Streudiagramm für die Mortalität und Weinkonsum

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 der 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 )

```

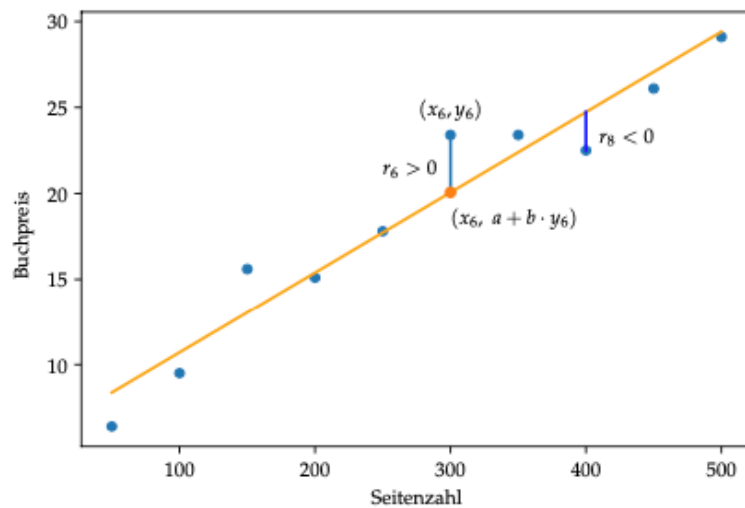



Abbildung 3.2: Residuen für Buchpreis in Abhängigkeit der Seitenanzahl

```

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 #    6.040000    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

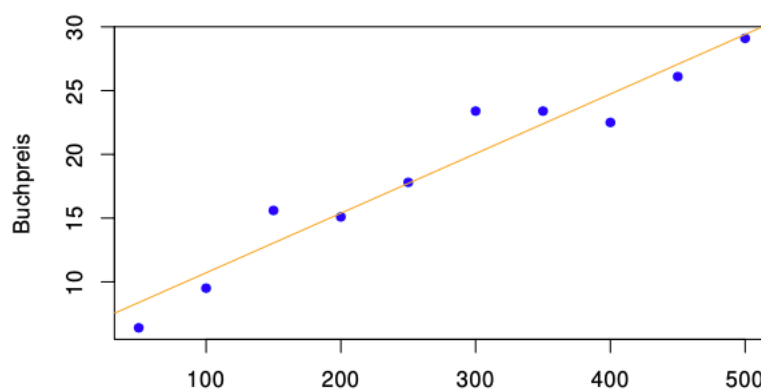


Abbildung 3.3: Streudiagramm mit Regressionsgeraden aus obigem R Code

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 Modells, die *innerhalb* des Bereichs 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 Würfel 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

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

Dieses Ereignis 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 nicht disjunkte Ereignisse. Die Schnittmenge wird doppelt gezählt, weshalb wir diese einmal abziehen müssen.

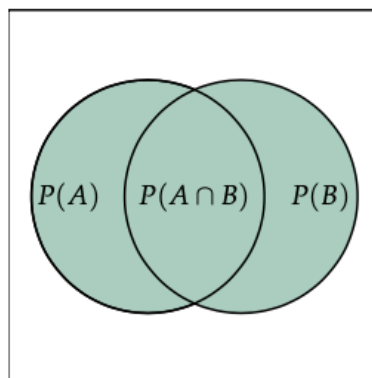


Abbildung 4.1: Wahrscheinlichkeit für nicht disjunkte Ereignisse

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 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 Berechnung 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* $\sigma(X)$ zusammengefasst werden. 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 das B eingetreten ist.

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

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

6.1 Rechenregeln der bedingten Wahrscheinlichkeit

$0 \leq P(A B) \leq 1$	für jedes Ereignis A
$P(B B) = 1$	für jedes Ereignis B
$P(A_1 \cup A_2 B) = P(A_1 B) + P(A_2 B)$	für A_1, A_2 disjunkt ($A_1 \cap A_2 = \{\}$)
$P(\bar{A} B) = 1 - P(A B)$	für jedes Ereignis A

Abbildung 6.1: Rechenregeln der bedingten Wahrscheinlichkeit

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 $P(B|A)$ bekannt ist.

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

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. Deshalb wird die Wahrscheinlichkeit zwischen zwei Punkten berechnet. 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.

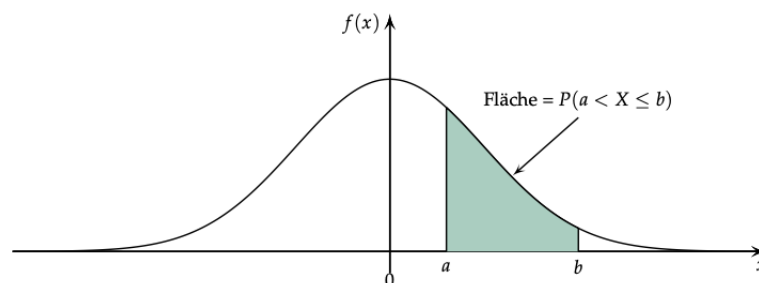


Abbildung 7.1: Dichte einer Zufallsvariable und der Wahrscheinlichkeit in ein Intervall (a, b) zu fallen (grüne Fläche)

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*.

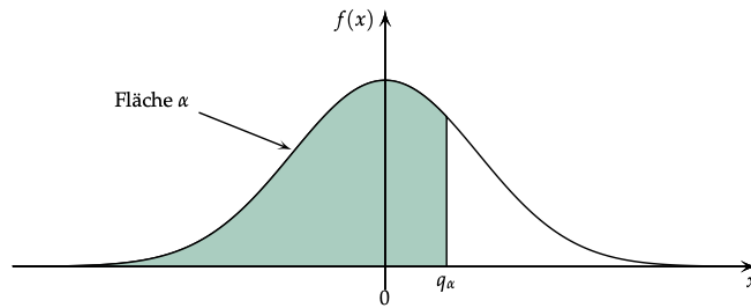


Abbildung 7.2: Quantil q_α anhand der Dichte $f(x)$ für $\alpha = 0.75$

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

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 mit 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. Ist die wichtigste 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:

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

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

Dies 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 im Abstand von etwa einer 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 Standardabweichung 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:

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

8 Durchschnitte und Summen von Zufallsvariablen

Bisher haben wir lediglich untersucht, wie *eine* Zufallsvariable verteilt ist. In Vielen Anwendungen haben wir es mit *mehreren* Zufallsvariablen zu tun. Überlicherweise messen wir die *gleiche* Grösse mehrmals und untersuchen, wie die Summe und der Durchschnitt von *mehreren* Zufallsvariablen verteilt sind.

Die Messungen bezeichnen wir mit x_1, x_2, \dots, x_n und fassen diese als Realisierungen der Zufallsvariablen X_1, X_2, \dots, X_n auf. Dies ist die Zufallsvariable für das *arithmetische Mittel*. Das arithmetische Mittel \bar{x}_n ist also eine Realisierung der Zufallsvariablen \bar{X}_n .

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

Analoges gilt für die *Summe* S_n :

$$S_n = X_1 + \dots + X_n = \sum_{i=1}^n X_i$$

8.1 Unabhängigkeit und i.i.d. Annahme

Wir treffen oft die Annahme, dass die Zufallsvariablen *unabhängig* voneinander sind. Es gibt keine gemeinsamen Faktoren, die den Ausgang der verschiedenen Messungen beeinflussen. Damit ist gemeint, dass eine Messung keinen Einfluss auf das Resultat der nachfolgenden Messung hat. Als Beispiel für *nicht* unabhängige Zufallsvariablen die Temperaturmessung an benachbarten Sommertage. Die Tage werden ähnliche Temperaturen haben nicht 28°C und am nächsten -5°C.

Wenn die Zufallsvariablen X_1, \dots, X_n unabhängig sind und alle *dieselbe* Verteilung haben, schreiben wir das kurz als

$$X_1, \dots, X_n \quad \text{i.i.d.}$$

Die Abkürzung i.i.d. steht für:

independent, **i**dentically, **d**istributed

Sind Zufallsvariablen i.i.d., so wird dasselbe unter den gleichen Bedingungen gemessen.

Die Unabhängigkeit spielt insofern eine Rolle bei den *Regeln für Erwartungswerte und Varianzen* von Summen. Die Beziehung

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

gilt immer,

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$$

jedoch nur wenn X_1 und X_2 unabhängig sind.

8.2 Kennzahlen von S_n und \bar{X}_n

Wir nehmen an, dass X_1, \dots, X_n i.i.d. wegen dem zweiten “i” in i.i.d., *jedes* X_i dieselbe Verteilung und dieselben Kennzahlen hat.

8.2.1 Varianz und Standardabweichung der Summe

Die Varianz und Standardabweichung nimmt mit zunehmender Anzahl Würfeln zu. Je mehr Ereignisse, umso grösser wird der Wertebereich. Die Summen verteilen sich auf mehr Zahlen und damit nimmt auch die Streuung zu. Das Gesetz für die Varianz und Standardabweichung der Summe lautet:

$$\text{Var}(S_n) = n\text{Var}(S_1)$$

und für die Standardabweichung gilt

$$\sigma_{S_n} = \sqrt{n}\sigma_X$$

8.2.2 Erwartungswert des Durchschnittes

Da wir gleichbleibende Ergebnisse erwarten, ob nun 2 oder 50 Experimente gemacht werden, entspricht der Durchschnitt in etwa dem Erwartungswert.

$$E(\bar{X}_n) = \mu$$

Für die Varianz und die Standardabweichung des Durchschnitts gilt folgendes Gesetz:

$$\text{Var}(\bar{X}_n) = \frac{\text{Var}(X)}{n}$$

$$\sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$

Allgemein gilt:

Kennzahlen von S_n
$E(S_n) = n\mu$
$\text{Var}(S_n) = n\text{Var}(X_i)$
$\sigma(S_n) = \sqrt{n}\sigma_X$
Kennzahlen von \bar{X}_n
$E(\bar{X}_n) = \mu$
$\text{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}$
$\sigma(\bar{X}_n) = \frac{\sigma_X}{\sqrt{n}}$
Die Standardabweichung von \bar{X}_n heisst auch der <i>Standardfehler</i> des arithmetischen Mittels.

Abbildung 8.1: Regeln für die Kennzahlen von S_n und \bar{X}_n

Die Standardabweichung der Summe wächst mit wachsendem n , aber langsamer, als die Anzahl Beobachtungen n . Der Erwartungswert von \bar{X}_n ist gleich demjenigen einer einzelnen Zufallsvariable X_i , die *Streuung nimmt jedoch mit wachsendem n ab*.

Gesetz der grossen Zahlen

Für $n \rightarrow \infty$ geht die Streuung gegen null. Es gilt das **Gesetz der grossen Zahlen**: Falls X_1, \dots, X_n i.i.d., dann

$$\bar{X}_n \rightarrow \mu \quad \text{für } n \rightarrow \infty$$

Abbildung 8.2: Gesetz der grossen Zahlen

Standardfehler

Die Standardabweichung des arithmetischen Mittels (*Standardfehler*) ist jedoch *nicht* proportional zu $1/n$, sondern nimmt nur ab mit dem Faktor $1/\sqrt{n}$

$$\sigma_{\bar{X}_n} = \frac{1}{\sqrt{n}} \sigma_X$$

Um den *Standardfehler* zu halbieren, braucht man also *viermal* so viele Beobachtungen. Dies nennt man auch das \sqrt{n} -Gesetz.

Abbildung 8.3: Standardfehler des arithmetischen Mittels

8.3 Verteilungen von S_n und \bar{X}_n

Die Verteilung der *Mittelwerte* \bar{X}_n (oder auch der Summen) nähert sich mit wachsendem n einer Normalverteilung an. Sind die X_i 's i.i.d., dann gilt der *Zentrale Grenzwertsatz*.

Zentraler Grenzwertsatz

Falls X_1, \dots, X_n i.i.d. mit irgendeiner Verteilung mit Erwartungswert μ und Varianz σ^2 , dann gilt (ohne Beweis)

$$S_n \approx \mathcal{N}(n\mu, n\sigma_X^2)$$

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

Abbildung 8.4: Zentraler Grenzwertsatz

Wobei die Approximation im allgemeinen besser wird mit grösserem n . Überdies ist die Approximation besser, je näher die Verteilung von X_i an der Normalverteilung $\mathcal{N}(\mu, \sigma_x^2)$ ist.

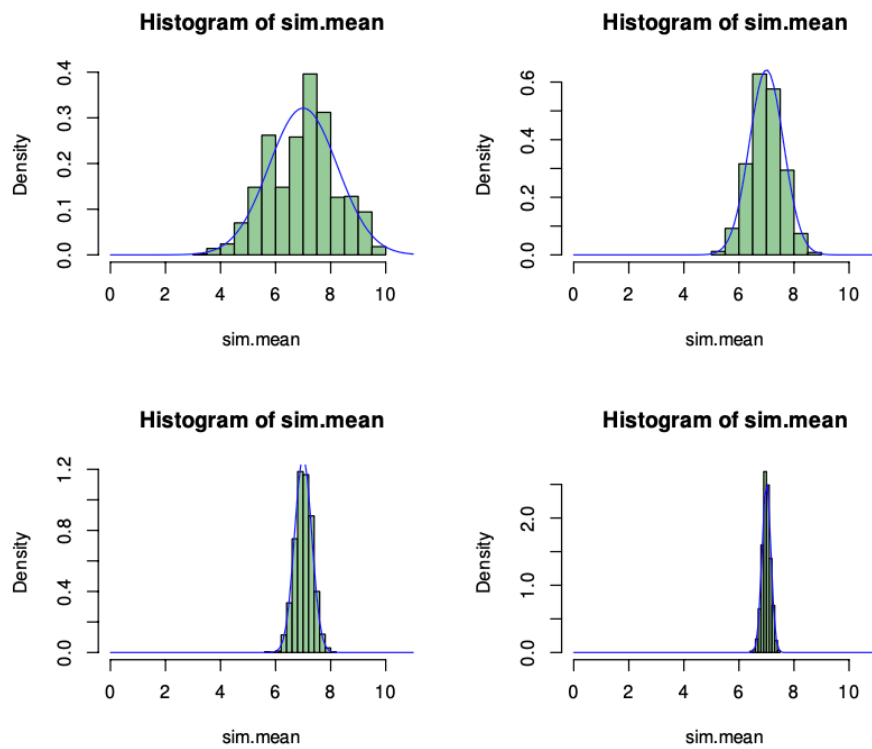


Abbildung 8.5: 4 Histogramme vom Durchschnitt von 16, 64, 256 und 1024 Versuchen mit je 1000 Ziehungen, inkl. Dichtekurven

9 Hypothesentest für Messdaten

Mittels eines Hypothesentests wird eine Angabe (z. B. Inhaltsmenge einer Pet-Flasche) überprüft, ob diese wahr ist. Solche Tests sind standardisiert und ist ein reproduzierbares Verfahren. Diese geben auch ein klares Kriterium an, wann ein Durchschnitt zu weit von einer Angabe entfernt ist. Der Hypothesentest ist *nie* ein Beweis, dass eine Angabe wahr ist oder falsch, sondern lediglich ob die Angabe mit einer gewissen Wahrscheinlichkeit stimmt oder nicht.

9.1 Statistische Tests und Vertrauensintervall für eine Stichprobe bei normalverteilten Daten

Messungen können wir als Realisierungen von unabhängigen, identisch verteilten Zufallsvariablen X_i betrachten und die Kennzahlen $E(X_i) = \mu$ und $Var(X_i) = \sigma_X^2$ erfassen. Typischerweise sind diese (und andere) Kennzahlen *unbekannt*. Trotzdem möchten wir eine Aussage über das wahre, aber eben unbekannte μ und σ^2 machen. Das Ziel ist es die Daten μ und σ^2 anzunähern und sprechen dabei von einer *Schätzung* der Parameter μ und σ^2 . Geschätzte Werte werden mit einem $\hat{\cdot}$ bezeichnet, zum Beispiel $\hat{\mu}$. Für die (Punkt-)Schätzungen für den Erwartungswert und Varianz gilt:

$$\hat{\mu} = \bar{X}_n = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$
$$\hat{\sigma}_X^2 = \frac{(X_1 - \bar{X}_n)^2 + \dots + (X_n - \bar{X}_n)^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Beachte, dass $\hat{\mu}$ und $\hat{\sigma}_X^2$ selbst Zufallsvariablen sind und für jede neue Messung sich neue $\hat{\mu}$ und $\hat{\sigma}_X^2$ ergeben. Obwohl im Allgemeinen $\hat{\mu} \neq \mu$ und $\hat{\sigma}_X^2 \approx \sigma_X^2$ ist die Hoffnung, dass $\hat{\mu} \rightarrow \mu$ und $\hat{\sigma}_X^2 \rightarrow \sigma_X^2$ und damit annähert.

9.1.1 Ziel des Hypothesentests

Das Ziel des Hypothesentests ist das feststellen, ob der wahre Mittelwert wahr ist. Entstehen bei den Schätzungen grössere Standardabweichungen, muss die Angabe hinterfragt werden.

9.2 Hypothesentest

Hypothesentests sind ein wichtiges statistisches Mittel um zu entscheiden, ob eine Messreihe zu einer gewissen Grösse passt. Wir gehen davon aus, dass wir den wahren Mittelwert *nicht* kennen, gehen aber von einem *Idealwert* oder einem vermuteten Wert aus.

Unter der Annahme, dass die Daten normalverteilt sind, wird überprüft, ob eine Messreihe, unter der Annahme von $\mu = a$ (der Mittelwert), wahrscheinlich ist oder nicht.

9.2.1 Modell

Eine Anzahl Messwerte sind die Realisierung der Zufallsvariablen X_1, X_2, \dots, X_n , wobei X_i eine kontinuierliche Messgrösse ist. Dabei soll gelten:

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_x^2)$$

9.2.2 Nullhypothese

$$H_0 : \mu = \mu_0 = a$$

Die Nullhypothese ist eine Annahme über den wahren Mittelwert. Diese Annahme wird mit dem Mittelwert $\hat{\mu}$ überprüft, ob er sich dem wahren Mittelwert annähert.

9.2.3 Alternativhypothese

$$H_A : \mu \neq \mu_0 = a$$

oder

$$H_A < \text{oder } >$$

Wenn die Annahme nicht gleich μ ist

9.2.4 Teststatistik

Es wird getestet, ob die Verteilung mit der Annahme $\mu = a$ gerechtfertigt ist. Oder Mathematisch: Die Verteilung der Teststatistik T unter Nullhypothese H_0

$$T = \bar{H}_n \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

Ist die Wahrscheinlichkeit kleiner als 2.5% ist sie zu klein und der Mittelwert zu unwahrscheinlich, als dieser zur Ausgangsgrösse a passen könnte.

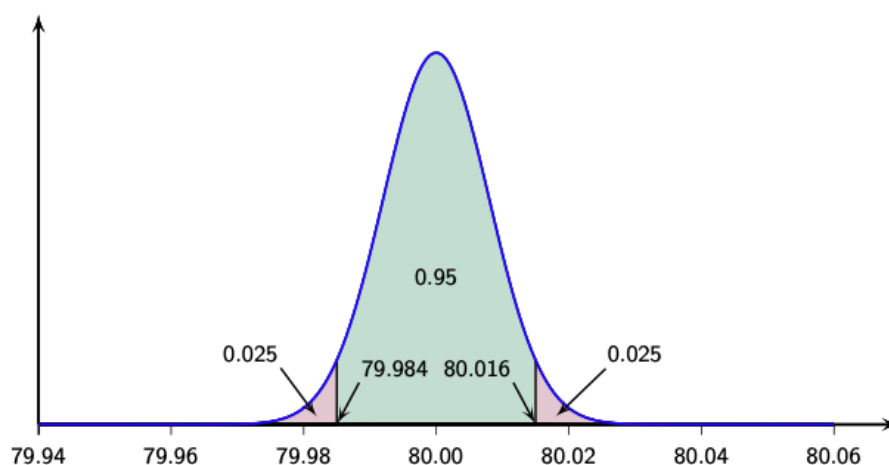


Abbildung 9.1: Normalverteilungskurve eines Hypothesentests

Die *Abmachung*, dass die Grenze 2.5% gilt, kommt daher, dass der symmetrische Teil um den Mittelwert 95% betragen soll.

9.2.5 Signifikanzniveau α

Das Signifikanzniveau α gibt an, wie hoch das Risiko ist, eine falsche Entscheidung zu treffen. Normalerweise 0.05 bzw. 0.01.

9.2.6 Verwerfungsbereich

Liegt der gemessene Mittelwert im roten Bereich (der Abbildung), so zweifelt man an der Nullhypothese und *verwerfen* diese. Diese werden in einem Intervall angegeben:

$$K = (-\infty, a - \alpha] \cup [a + \alpha, \infty)$$

9.2.7 p-Wert

Der P-Wert ist die Wahrscheinlichkeit, unter der Nullhypothese ein mindestens so extremes Ereignis (in Richtung der Alternative) zu beobachten, wie das aktuell beobachtete. Damit wird angedeutet, wie extrem das Ergebnis ist. Je kleiner der p-Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese.

- 0: passt gar nicht
- 1: passt sehr gut

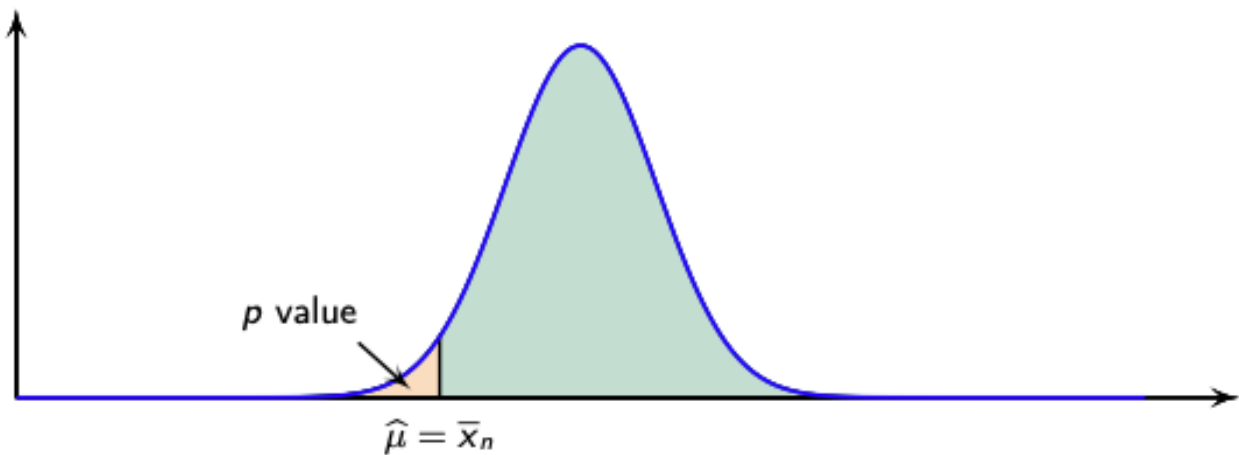


Abbildung 9.2: Wahrscheinlichkeit, unter Gültigkeit der Nullhypothese, das erhaltene Ergebnis oder ein extremeres zu erhalten

9.2.8 p-Wert und Statistischer Test

Bei einem vorgegebenen Signifikanzniveau α gilt aufgrund der Definition des p -Werts für einen einseitigen Test:

- Verwerfe H_0 falls $p\text{-Wert} \leq \alpha$
- Belasse H_0 falls $p\text{-Wert} > \alpha$

Computerprogramme liefern den Testentscheid nur mit p -Wert und *immer* auf Signifikanzniveau.

9.2.8.1 Signifikanz

- p-Wert $\approx 0.05 \Rightarrow$ schwach signifikant, “.”
- p-Wert $\approx 0.01 \Rightarrow$ signifikant, “*”
- p-Wert $\approx 0.001 \Rightarrow$ stark signifikant, “**”
- p-Wert $\leq 10^{-4} \Rightarrow$ äusserst signifikant, “***”

9.3 t -Test

Entgegen dem bisherigen Verfahren (z -Test) wo die Standardabweichung bekannt ist, setzt der t -Test keine Standardabweichung voraus. Dies ist auch praktisch kaum der Falls, dass die SD vorliegt.

9.3.1 t -Verteilung

Die Verteilung der Teststatistik beim t -Test unter der Nullhypothese

$$H_0 : \mu = \mu_0$$

ist gegeben durch

$$T = \bar{X}_n \sim T_{n-1}(\mu, \frac{\hat{\sigma}_X}{\sqrt{n}})$$

wobei t_{n-1} eine Verteilung mit $n - 1$ Freiheitsgraden ist.

Die Normalverteilung wird also durch eine t -Verteilung ersetzt. Gleicht aber der Normalverteilung. Ist aber flacher, wegen der grösseren Unsicherheit. Dies hängt von der Anzahl Beobachtungen ab.

10 Vertrauensintervall, Zweistichprobentest und Wilcoxon-Test

10.1 Vertrauensintervall für μ

Das Intervall gibt an, wo der wahre Mittelwert μ mit einer bestimmten 95%-Wahrscheinlichkeit liegt. Bei der Bestimmung des Verwerfungsbereiches beim z -Test gehen wir von einem wahren (aber unbekannten) Wert μ aus mit einer bekannten Standardabweichung.

```
1 qnorm(p=c(0.025, 0.975), mean= 5, sd =2)
2 ## 1.080072 8.919928
```

In obiger Ausgabe sehen wir das **Vertrauensintervall**. In diesem Intervall liegt mit bestimmter Wahrscheinlichkeit der **wahre** Mittelwert. Liegt nun \bar{x}_n im Vertrauenintervall, wird H_0 nicht verworfen. Liegt der Mittelwert ausserhalb, wird H_0 verworfen.

Dies ist eine weitere Möglichkeit für einen Testentscheid. R gibt das Vertrauenintervall auch im t -Test als **confidence interval** aus. Dieses besagt, dass bei einem Signifikanzniveau von 5% das wahre μ zu 95% in diesem Intervall liegt. Je schmalere das Vertrauensintervall ist, umso sicherer sind wir, wo sich der wahre Mittelwert befindet. Bei einem grossen Intervall besteht eine grosse Unsicherheit wo das wahre μ liegt.

10.2 Der Wilcoxon-Test

Der Wilcoxon-Test ist eine Alternative zum t -Test, setzt dabei aber weniger voraus. Er wird vorallem bei **nicht-normalverteilten** Daten eingesetzt. Grundsätzlich hat er die grössere *Macht*. Macht ist die Wahrscheinlichkeit, dass die Nullhypothese richtigerweise verworfen wird. Einzige Voraussetzung bei einem Wilcoxon-Test ist, dass die Verteilung unter der Nullhypothese *symmetrisch* bezüglich μ_0 ist.

Der Test berechnet den V -Wert, der die sogenannte *Rangsumme* repräsentiert. Ist der V -Wert zu weit weg vom Median, wird die Nullhypothese verworfen, ansonsten beibehalten.

10.3 Statistische Tests bei zwei Stichproben

Wird ein Vergleich zwischen zwei Proben gemacht, spricht man von gepaarten und ungepaarten Stichproben.

10.3.1 Gepaarte Stichproben

Um eine gepaarte Stichprobe zu sein müssen folgende Voraussetzungen gelten:

- beide Versuchsbedingungen müssen an derselben Versuchseinheit eingesetzt werden
- jeder Versuchseinheit aus der einen Gruppe kann genau eine Versuchseinheit aus der anderen Gruppe zugeordnet werden
- die Stichprobengrösse n ist für beide Versuchsbedingungen dieselbe
- x_i und y_i sind abhängig, weil die Werte von der gleichen Versuchseinheit kommen

10.3.1.1 Beispiele gepaarte Stichproben

- Messung vor und nach einem Ereignis
- Zwei Messungen vom selben Punkt/Objekt (Zuordnung ist eindeutig)

10.3.1.2 Statistischer Test für gepaarte Stichproben

Bei der Analyse arbeitet man mit den Differenzen innerhalb der Paare $d_i = x_i - y_i (i = 1, \dots, n)$ welche wir als Zufallsvariablen D_1, \dots, D_n auffassen. **Kein** Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[D_i] = 0$. Falls die Daten normalverteilt sind eignet sich ein t -Test, sonst Wilcoxon-Test. Bei beiden Tests kann für eine gepaarte Stichprobe die Option `paired = TRUE` mitgegeben werden.

Hinweise

- man kann auch die Differenzen d_i berechnen und dann den Test durchführen
- das erste Argument im Funktionsaufruf bezieht sich auf das **nachher**, das zweite auf die Messung **vorher**

10.3.2 Ungepaarte Stichproben

Oft kann nicht jede Messung einer Messung aus der zweiten Gruppe eindeutig zuordnen. Dann spricht man von einer **ungepaarten** Stichprobe. Im Allgemeinen ist die Anzahl der Messungen beider Gruppen unterschiedlich (muss aber nicht). Entscheidend ist dass x_i und y_i zu verschiedenen Versuchseinheiten gehören und unabhängig angenommen werden.

10.3.2.1 Beispiele für ungepaarte Stichproben

- Messung wird nacheinander von zwei versch. Geräten gemacht (nicht gleichzeitig)
- verschieden lange Messreihen
- Zufällige Zuordnung in zwei Gruppen

10.4 Tests mit R bei zwei Stichproben

Folgende Parameter können bei t -Test oder Wilcoxon-Test mitgegeben werden:

- `x`: erste Messreihe (nachher)
- `y`: zweite Messreihe (vorher)
- `alternative=`: "less", "greater" oder "two.sided" als Varianten, je nach Test
- `mu=`: gibt an welcher Unterschied in den Mittelwerten der beiden Gruppen getestet werden soll. Wenn Test prüfen soll ob Gruppenmittelwerte gleich, dann `mu=0`
- `paired=`: TRUE oder FALSE für gepaarte/ungepaarte Messreihen
- `conf.level=`: definiert Vertrauensintervall - default 0.05

11 Lineare Regression

Mit Hilfe der lineare Regression soll ein *Modell* entwickelt werden, womit eine *Vorhersage* über den Verlauf gemacht werden kann. Dabei suchen wir nach der *Zielgrösse* oder *Outputvariable* Y , welche wir nur approximativ darstellen können, weil die Datenpunkte aus mathematischer Sichtweise keinen Graph einer Funktion darstellt, sondern Streudiagramme. Aus diesem Grund gilt

$$Y \approx f(X_1, X_2, X_n) + \varepsilon$$

wobei X_n als *Prädiktoren*, *Inputvariablen* oder *erklärende Variablen* bezeichnet werden. ε ist ein *zufälliger Fehlerterm* der von X_n unabhängig ist und den Mittelwert 0 hat.

Die Funktion f ist im Allgemeinen unbekannt und wir müssen das optimale Modell wählen. Polynome n -ten Grades passen meistens zwar besser, sind aber deutlich komplizierter als lineare Abbildungen. Es kann auch zu *Overfitting* führen, einem Phänomen welches Fehler oder Ausreisser zu stark berücksichtigt. Der grosse Vorteil von linearen Modellen ist das einfache geometrische interpretieren. β_0 entspricht dem y -Achsenabschnitt und β_1 der Steigung der Geraden.

11.1 Das einfache Regressionsmodell

Das *einfache lineare Regressionsmodell* ist ein einfaches Verfahren, um einen quantitativen Output Y auf der Basis einer einzigen Inputvariable X vorherzusagen. Es wird eine annähernd lineare Beziehung zwischen X und Y angenommen.

$$Y \approx \beta_0 + \beta_1 X$$

Es wird eine Gerade gesucht, welche möglichst nahe an der wahren Gerade anliegt. Um die *Nähe* jedes Datenpunktes zu messen, wird meistens die *Methode der kleinsten Quadrate* verwendet. Der erhaltene Wert ist das i -te *Residuum*, welches der Differenz zwischen dem i -ten *beobachteten* Wert der Zielgrösse und dem i -ten vom linearen Modell *vorhergesagten* Wert der Zielgrösse entspricht. Die Residuen oberhalb der Geraden sind positiv, diejenigen unterhalb der Geraden negativ. Die *Summe* der Residuen ist kein guter Wert für die Nähe der Punkte zur Gerade, weil sich diese aufheben. Dazu eignet sich die Summe der *Quadrate* der Residuen (Residual Sum of Squares - RSS).

$$RSS = r_1^2 + r_2^2 + \dots + r_n^2$$

oder äquivalent

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Mit der Methode der kleinsten Quadrate wird nun $\hat{\beta}_0$ und $\hat{\beta}_1$ so gewählt, dass RSS *minimal* wird.

In R wird die Regressionsgerade mit dem Befehl `lm(y~x)` berechnet. Wobei der *intercept* den y -Achsenabschnitt bestimmt. Die Inputvariable bestimmt die Steigung der Gerade.

11.2 Vertrauensintervall im linearen Modell

Mit dem Vertrauensintervall bestimmen wir wie *genau* unsere Schätzungen für die Koeffizienten sind. Meistens hat man in der Statistik nicht alle nötigen Daten zur Verfügung um eine Vorhersage zu machen. Um einen ungefähren Wert von μ zu erhalten, wählen wir aus einer Messreihe z. B. 1000 Werte und verwenden den Durchschnitt \bar{y} und damit $\mu \approx \hat{\mu} = \bar{y}$.

Mit R erhalten wir nun für β_0 und β_1 das 95%-Vertrauensintervall:

```
confint(lm(y~x), level=0.95)
```

11.3 Hypothesentest im linearen Modell

Der häufigste Hypothesentest besteht aus dem Testen der *Nullhypothese* von

$$H_0 = \text{Es gibt keinen Zusammenhang zwischen X und Y}$$

gegen die *Alternativhypothese*

$$H_A = \text{Es gibt einen Zusammenhang zwischen X und Y}$$

Mathematisch entspricht dies $H_0 : \beta_1 = 0$ gegen $H_A : \beta_1 \neq 0$. Ist $\beta_1 = 0$, dann gilt $Y = \beta_0 + \varepsilon$ und damit hängt Y *nicht* von X ab.

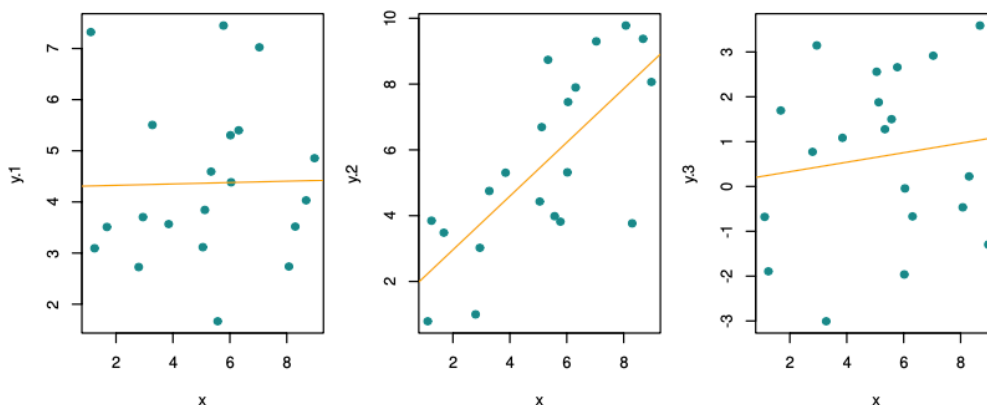


Abbildung 11.1: Null- und Alternativhypothese graphisch

Bei der Abbildung links ist $\hat{\beta}_1$ praktisch 0. Es gibt *keinen* Zusammenhang zwischen den Variablen x und y_1 . Das bedeutet, egal welches x wir wählen, der y -Wert bleibt immer gleich.

Im Graph in der Mitte ist $\hat{\beta}_1$ ungleich 0. Es *gibt* einen Zusammenhang zwischen den Variablen x und y_2 . Das heisst je grösser wir x wählen, umso grösser wird der y -Wert.

Im dritten Graph ist die Gerade zwar leicht steigend, aber es ist schwer einen Zusammenhang im Streudiagramm zu erkennen. Die Abweichung kann zufällig sein.

Die Frage ist nun, wie stark steigend oder fallend muss die Regressionsgerade sein, damit die Steigung *statistisch signifikant* ungleich 0 ist? Um die Nullhypothese zu testen wenden wir einen Hypothesentest (t -Test) an.

```
summary(lm(y~x))
```

Ist der Wert von $Pr(>|t|)$ unter dem Eintrag *Coefficients* kleiner als 0.05, wird die Nullhypothese verworfen. Dann gibt es einen Zusammenhang zwischen den beiden Variablen.

11.4 Abschätzung der Genauigkeit des Modells

Nachdem die Nullhypothese verworfen wurde, stellt sich die Frage wie genau passt das Modell zu den wahren Daten. Die Qualität einer linearen Regression wird typischerweise durch den *residual standard error* (RSE) und die R^2 -Statistik abgeschätzt.

11.4.1 R^2 -Statistik

Die R^2 -Statistik ist ein Wert zwischen 0 und 1. Sie gibt an, welcher Anteil der Variabilität in Y mit Hilfe des Modells durch X erklärt werden. Ein Wert nahe bei 1 bedeutet, dass ein grosser Anteil der Variabilität durch die Regression erklärt wird. Das Modell beschreibt die Daten sehr gut. Ein Wert nahe bei 0 bedeutet, dass die Regression die Variabilität der erklärenden Variablen nicht erklärt. Wurde nun die Nullhypothese verworfen, stellt sich die Frage in welchem Ausmass das Modell zu den Daten passt.

Der R^2 -Wert wird durch zwei Varianzen berechnet. Er ergibt sich aus dem Durchschnitt der quadrierten Länge vom Abstand des Mittelwertes von y zum *Modell*, geteilt durch den den Durchschnitt der quadrierten Länge vom Abstand des Mittelwertes von y zum *Sample* (Messreihe).

Siehe in der folgenden Abbildung. Die roten Linien sind die Abstände zum Modell, die pinken die Abstände zu den Datenpunkten.

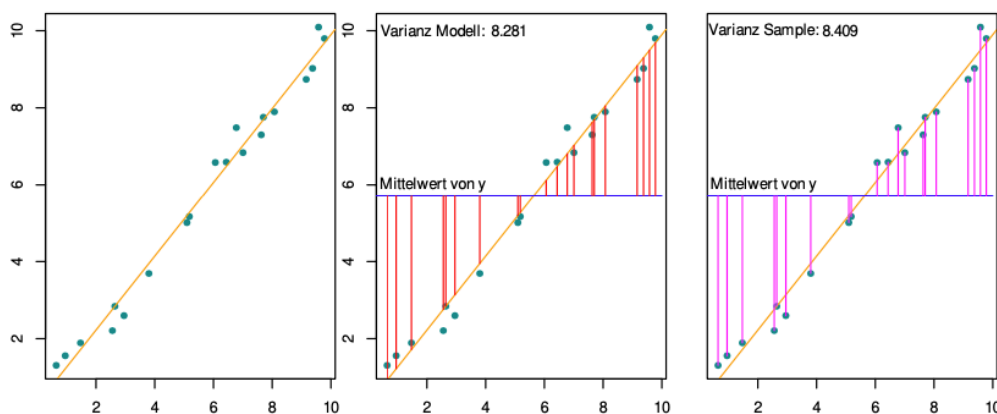


Abbildung 8.9. : Definition von R^2

Abbildung 11.2: Definition von R^2

Der R^2 -Wert wird definiert durch:

$$R^2 = \frac{\text{Varianz Modell}}{\text{Varianz Sample}}$$

```
summary(lm(y~x))$r.squared
```

11.4.2 Alternative Darstellung von R^2

R^2 lässt sich auch durch die Differenzen der Daten zum Modell berechnen. Dazu wird der Durchschnitt der Quadrate der Differenzen von Modell zu den Daten berechnet. Dadurch ergibt sich folgende Definition:

$$R^2 = 1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$$

Die Berechnung mit R ist identisch wie vorhin.

Der Vorteil dieser Variante ist, dass sich der R^2 -Wert einfacher interpretieren lässt

- Varianz Differenz: Ist die Varianz des Samples, die *nicht* durch das Modell erklärt wird
- $\frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: Ist der *Anteil* der Varianz vom Sample, der *nicht* vom Modell erklärt wird
- $1 - \frac{\text{Varianz Differenz}}{\text{Varianz Sample}}$: *Anteil* der Varianz vom Sample, der vom Modell erklärt wird
- R^2 : *Anteil* der Varianz vom Sample, der vom Modell erklärt wird.

Die Varianz lässt sich in R einfach berechnen

`var(y)`

Liegt die Varianz der Differenz nahe bei 0, passt das Modell gut zu den Daten. Die jeweiligen Abweichungen sind sehr klein. Der R^2 -Wert liegt dann nahe bei 1.

12 Multiple lineare Regression

Die einfache lineare Regression ist ein nützliches Vorgehen, um eine Vorhersage einer einzelnen erklärenden Variable zu treffen. Allerdings hängt die Vorhersage in der Praxis meist von mehr als einer erklärenden Variable ab. Man könnte die Vorhersage mit drei separate linearen Regressionen treffen. Allerdings ist dann nicht klar, wie eine geeignete Vorhersage für die drei erklärenden Variablen auszusehen hat, weil jeweils der Input durch eine andere Regressionsgleichung mit der Zielvariable verknüpft ist. Zweitens ignoriert jede der Regressionsgleichungen die anderen erklärenden Variablen zur Bestimmung der Regressionskoeffizienten. Die führt zu irreführenden Schätzungen der Wirkung der erklärenden Variablen.

Um dies zu umgehen empfiehlt es sich die einfache lineare Regression mit allen erklärenden Variablen zu erweitern. Allgemein gehen wir davon aus, dass wir p verschiedene erklärenden Variablen haben. Daraus ergibt sich die Form für das *multiple lineare Regressionsmodell*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

12.1 Graphische Interpretation

Die graphische Interpretation entfällt bei multipler linearen Regression vollends, da wir mehr als zwei Variablen haben und so nicht in einem Koordinatennetz dargestellt werden können. Pro Variable wäre eine Achse nötig.

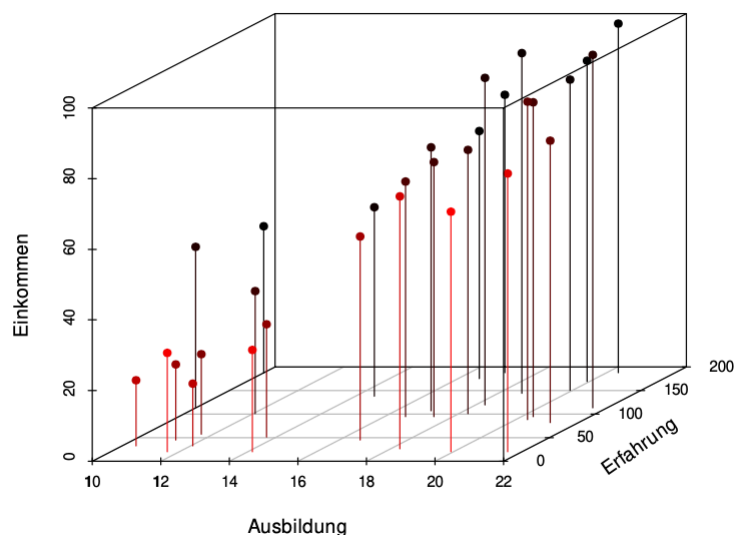


Abbildung 12.1: Datenpunkte im Raum mit zwei erklärenden Variablen

Analog zum einfachen linearen Regressionsmodell wird hier einfach eine *Ebene* gesucht, die am «besten» zu den Datenpunkten passt.

12.2 Koeffizienten mit R

Um die Koeffizienten $\hat{\beta}_0$, $\hat{\beta}_1$ und $\hat{\beta}_p$ zu schätzen, kann folgender R-Befehl verwendet werden. Beachte das hinzufügen der weiteren erklärenden Variable mit +.

```
coef(lm(y~x+z))
```

Um die multiple lineare Regression auszuführen verwenden wir den Befehl

```
summary(lm(y~x+z))
```

Es ist möglich, dass sich eine Grösse mit fremden Lorbeeren schmückt. Wobei wir entsprechend den p -Wert genau prüfen, ob die Grösse signifikanten Einfluss auf das Ergebnis hat. Es kann nämlich sein, dass bei der Evaluation einer einzelnen Variable eine Steigung der Gerade resultiert, diese aber eigentlich nicht von dieser Variable, sondern einer anderen die nicht im Modell betrachtet wird, ausgeht. Dies heisst, man muss bei der Interpretation genau aufpassen und die Aussage der Daten überdenken (Beispiel mit TV, Radio, Zeitung).

12.3 Wichtige Fragestellungen

Bei der Durchführung einer multiple linearen Regression, möchten wir folgende Fragen beantworten:

1. Ist mindestens eine der erklärenden Variablen nützlich, um die Zielgrösse vorherzusagen?
2. Spielen alle erklärenden Variablen für die Vorhersage von Y eine Rolle, oder ist es nur eine Teilmenge?
3. Wie gut passt das Modell zu den Daten?
4. Welche Zielgrösse können wir aufgrund konkreter Werte der erklärenden Variablen vorhersagen und wie genau ist diese Vorhersage?

12.3.1 Zusammenhang zwischen erklärenden Variablen und der Zielgrösse

Bei der multiplen linearen Regression mit p erklärenden Variablen müssen wir uns fragen ob *alle* Regressionskoeffizienten ausser β_0 Null sind. Wie bei der einfachen linearen Regression verwenden wir dazu einen Hypothesentest und testen die Nullhypothese

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

gegen die Alternativhypothese

$$H_A : \text{min. ein } \beta_i \neq 0$$

Bei dem Hypothesentest wird die F -Statistik und ein F -Wert berechnet. Ist dieser «weit» weg von 0, wird die Nullhypothese verworfen. Anhand des F -Wertes wird auch ein p -Wert berechnet. Damit wird der Testentscheid getroffen.

```
summary(lm(y~x+z))
```

Obiger Befehl liefert die F -Statistik unter F -statistic. Ist er grösser als 1 liefert er ein überzeugender Hinweis gegen die Nullhypothese. Die bedeutet, dass mindestens eine erklärende Variable Einfluss auf die Zielgrösse hat. Wie nahe der F -Wert sein muss, ist von n und p abhängig. Ist n gross, kann F -Wert wenig grösser als 1 genügend starken Hinweis gegen H_0 sein. Wenn n klein, muss F -Wert gross sein.

Auf Basis des p -Wertes können wir entscheiden, ob wir die Nullhypothese verwerfen oder nicht (muss unter Signifikanzniveau liegen). Damit ist mindestens eine erklärende Variable für die Zunahme der Zielgrösse verantwortlich.

Der F -Wert ist aber nur aussagekräftig bei wenigen p 's, so müssen zusätzlich die t -Werte der einzelnen Variablen betrachtet werden. Sind diese signifikant, haben auch nur diese Variablen Einfluss auf die Zielgrösse. Die F - und p -Wert ergeben dabei kein signifikantes Resultat.

12.3.2 Passung des Modells zu den Daten

Wie beim einfachen linearen Modell deutet ein R^2 -Wert nahe bei 1 an, dass ein grosser Anteil der Varianz in der Zielvariable durch das Modell gut erklärt werden kann. Der Wert von R^2 nimmt mit der Anzahl Variablen zu, sogar dann, wenn diese keinen oder kaum Einfluss auf die Zielgrösse haben. Der Grund ist, weil die Anpassung an die Daten genauer wird. Bei unrelevanten Variablen ist die Zunahme aber *sehr klein*.

12.3.3 Vorhersagen

Auf Basis des multiplen linearen Regressionsmodells können wir die Zielgrösse vorhersagen. Allerdings gibt es dabei *Ungewissheiten*, wie stark Y von \hat{Y} abweichen wird. Als Antwort darauf kann das *Prognoseintervall* berechnet werden. Dieses ist breiter als das Vertrauensintervall, weil es den Fehler in der Schätzung von $f(X_1, \dots, X_p)$ (reduzierbarer Fehler) und die Unsicherheit, wie weit ein einzelner Punkt in der wahren Regressionsebene abweicht (irreduzierbarer Fehler), beinhaltet.

Vertrauensintervall (95%) für eine Zielgrösse mit R mit dem Einsatz von z. B. CHF 100000 für x und CHF 20000 für z .

```
predict(lm(y~x+z), interval = "confidence", data.frame(x = 100, z = 20))
```

$$[lwr, upr]$$

Die Interpretation lautet: 95% aller Intervalle dieser Form enthalten den wahren Wert in diesem Intervall. Wenn wir die Prognose in einem spezifischen Gebiet machen möchten, ersetzen wir im R-Befehl den Parameter `interval` mit *prediction*.

```
predict(lm(y~x+z), interval = "prediction", data.frame(x = 100, z = 20))
```

Normalerweise ist dann der Prognosebereich wesentlich breiter. Dies widerspiegelt die grössere Unsicherheit in Bezug auf eine breite Anwendung gegenüber einer Anwendung in einem spezifischen Gebiet.

12.4 Erweiterung des linearen Modells

Die Beziehung zwischen den erklärenden Variablen und der Zielgrösse beinhaltet im Standardregressionsmodell einschneidende Einschränkungen:

1. *Additivität*: Die Annahme hat den Effekt, dass eine Änderung einer erklärenden Variable unabhängig von den Werten der anderen erklärenden Variablen ist. Dies heisst, dass kein Synergieeffekt auf die andere Variable entsteht.
2. *Linearität*: Die Annahme hat den Effekt, dass die Veränderung der Zielvariable konstant ist, unabhängig der Änderung von X_j

In der Statistik spricht man dabei von einem *Interaktionseffekt*. Um das Modell zu erweitern, fügt man einen *Interaktionsterm* hinzu, welcher sich aus dem Produkt von X_1 und X_2 zusammensetzt. Dies resultiert im folgendem Modell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Das Modell bleibt aber bestimmten Einschränkungen unterworfen. Sofern ein X_j auf Null gesetzt wird, können Vorhersagen in einem Bereich gemacht werden, für welche das Modell nicht geeignet ist (unrealistisch). Dieses Problem wird als *Extrapolation* bezeichnet.

12.5 Interpretation des Interaktionsterm

Die Koeffizienten für ein solches Modell können mit folgenden R-Befehl ausgegeben werden:

```
summary(lm(y~x+z+x*z))
```

Sofern der p -Wert von β_3 unter dem Signifikanzniveau liegt, deutet dies darauf hin, dass der wahre Zusammenhang nicht additiv ist. Das Modell, welches nur die *Haupteffekte* beinhaltet, ist dem erweiterten Modell unterlegen. Sofern alle p -Werte signifikant sind, können sie im Modell enthalten bleiben. Es ist allerdings auch möglich, dass die p -Werte der Haupteffekte nicht klein sind. Das *hierarchische Prinzip* besagt aber, dass falls ein Interaktionsterm im Modell vorhanden ist, dass dann auch deren Haupteffekte vorhanden sein sollten, auch wenn ihre p -Werte nicht klein sind.

12.6 Qualitative erklärende Variablen

Erklärende Variablen sind nicht nur *quantitativ* sondern auch *qualitativ*. Zum Beispiel kann eine Variable *gender* nur spezifische Werte annehmen. Eine entsprechende Variable wird *Indikatorvariable* genannt. Nachfolgend die Geschlechtervariable mit zwei *Levels*:

$$x_i = \begin{cases} 0, & \text{falls } i\text{-te Person weiblich} \\ 1, & \text{falls } i\text{-te Person männlich} \end{cases}$$

Daraus resultiert folgendes Modell

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{falls } i\text{-te Person weiblich} \\ \beta_0 + \varepsilon_i, & \text{falls } i\text{-te Person männlich} \end{cases}$$

Die Kodierung (1,0/w,m) kann beliebig gewählt werden und hat keinen Einfluss auf das Regressionsmodell. Lediglich die Interpretation der Koeffizienten ist unterschiedlich. Die Anzahl Levels kann beliebig gewählt werden, es gibt aber folgendes zu beachten:

- Es gibt immer eine Indikatorvariable weniger, als es Levels hat
- Der Level ohne Indikatorvariable heisst auch *Baseline*

13 Variablenselektion

Mit Hilfe der Variablenselektion haben wir eine Methode, wie wir eine bessere *Vorhersagegenauigkeit* und *Modellinterpretierbarkeit* erreichen.

- *Vorhersagegenauigkeit*: Der wahre Zusammenhang zwischen den X_i und Y ist annähernd linear, wenn n sehr viel grösser als p ist. Die Varianz ist klein und Vorhersagen entsprechend gut. Durch *Schrumpfen* oder auch *Shrinking* kann die Varianz oft verkleinert und die Vorhersagekraft gesteigert werden.
- *Modellinterpretierbarkeit*: Oft haben einige Variablen keinen Zusammenhang mit der Zielvariable und können weggelassen werden. Das Modell kann einfacher interpretiert werden.

Die Methode der kleinsten Quadrate liefert keine Koeffizienten die exakt 0 sind. Folgende Verfahren entfernen automatisch *irrelevante* Variablen aus dem multiplen Regressionsmodell.

13.1 Schrittweise Vorwärtsselektion

Ist eine rechnerisch effiziente Methode, um Variablen zu eliminieren. Sie beginnt mit einem Modell ohne erklärenden Variablen und fügt die Variablen hinzu, welche die *grösste zusätzliche* Verbesserung mit sich bringen. Die Methode startet mit dem *Nullmodell* M_0 .

$$M_0 = \beta_0 + \varepsilon$$

Danach wird die *beste* Variable ausgewählt und hinzugefügt. Das Auswahlkriterium ist der kleinste RSS-Wert (Summe der Quadrate der Residuen). Je kleiner dieser Wert, umso besser passen die Daten zum System. Die Variable wird mit dem R-Befehl `add1` hinzugefügt. Danach ist ein `update` notwendig. Jede weitere hinzugefügte Variable ergibt ein neues Modell M_i . Mit Hilfe des AIC-Wertes wird entschieden, welches das beste Modell unter den M_i 's ist.

```
1 reg <- regsubsets(y~., data = list, method = "forward")
2 reg.sum <- summary(reg)
3 reg.sum$which
```

Obiger Befehl führt gesamtes Verfahren automatisch ohne `add1` und `update` durch. Sobald ein *TRUE* erscheint, wird die Variable verwendet.

13.2 Schrittweise Rückwärtsselektion

Funktioniert ähnlich wie die Vorwärtsselektion, allerdings wird mit einem vollen Modell begonnen, welches alle erklärenden Variablen enthält. Schrittweise wird die Variable vom Modell entfernt, welche am wenigsten nützlich ist. Es wird wiederum der RSS-Wert (kleinste) verwendet, um zu entscheiden welche Variable entfernt wird.

13.3 Anzahl Variablen / Abbruchbedingung

Die Vorwärts- und Rückwärtsselektion beschreibt nur in welcher Reihenfolge Variablen ausgewählt werden, jedoch nicht wieviele. Das Problem bei der Vorwärtsselektion ist, dass der RSS-Wert mit jeder weiteren Variable abnimmt. Aber auch der R^2 ist kein Abbruchkriterium, weil dieser mit jeder Variable zunimmt. Der *adjusted* R^2 ist abhängig von der Anzahl Variablen und nimmt zu Beginn zu. Er erreicht irgendwann ein Maximum und würde dann wieder abnehmen. Weiter könnte auch der AIC (Akaike information criterion) verwendet werden. Ein kleiner AIC ist besser und die Variablen werden addiert, solange der AIC-Wert abnimmt. Sobald er zunimmt, werden Die restlichen Variablen weggelassen.