Scriptum - Fortgeschrittene Statistik

Robert Rein

Invalid Date

Table of contents

Vorwort

Dies ist das Skriptum für den Master-Statistikkurse Fortgeschrittene Statistik und ist die Vorlage für die Kurse LTC4 und SBG4. Es werden in den Kursen nicht alle Themen des Skriptums behandelt. Das Skriptum befindet sich derzeit noch in einem frühen Stadium, so dass die Inhalte noch nicht vollständig ausgearbeitet sind.

Part I

Statistik

Die erste Frage die sich im Umgang mit der Anwendung von Verfahren der Statistik stellt ist: Wofür benötigen wir Statistik überhaupt?

Beispielsweise wurden ein Datensatz gesammelt, bei dem zwei Gruppen miteinander verglichen werden, eine Treatmentgruppe (TRT) und eine Kontrollgruppe (CON). In beiden Gruppen wurden jeweils $N_i = 20$ Personen untersucht. Es wurde das folgende Ergebnis erhalten (siehe Figure 1).

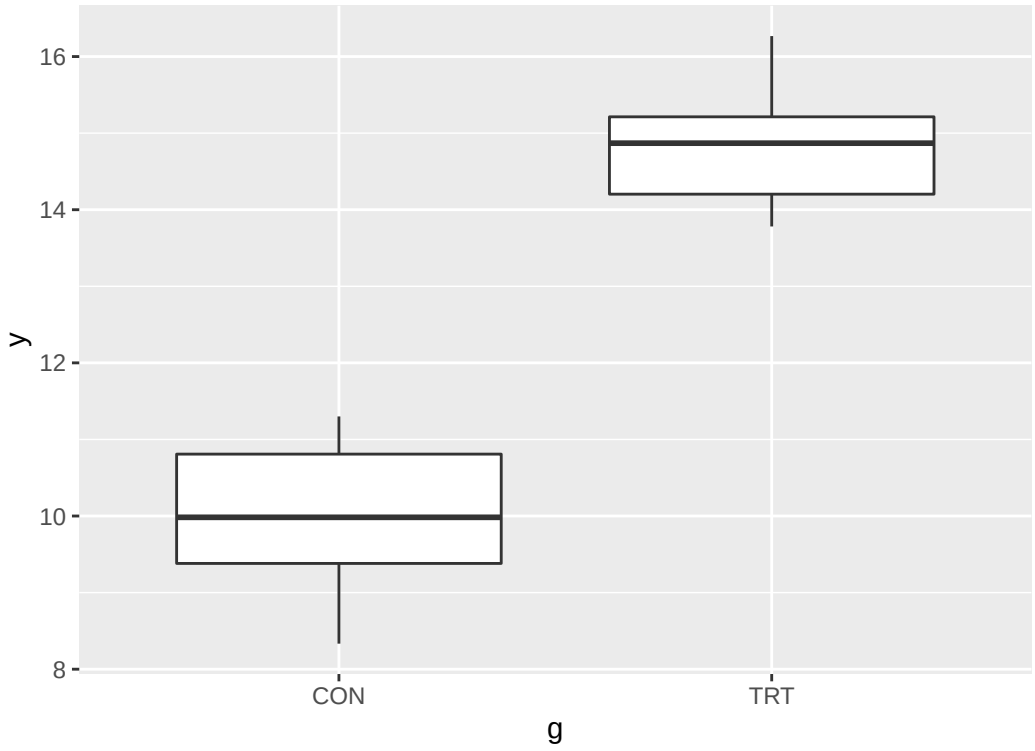


Figure 1: Boxplot der Kontroll- und der Treatmentgruppe bezüglich einer abhängigen Variable

Offensichtlich sind die Werte in der Treatmentgruppe deutlich höher als diejenigen in der Kontrollgruppe. Warum ist es nicht ausreichend das offensichtliche zu dokumentieren? Warum ist eine statistische Analyse der Daten notwendig?

Diese Fragestellung wird in dem folgenden Abschnitt untersucht. Gleichzeitig werden die notwendigen Werkzeuge entwickelt um die verschiendenen Schritte die einer statistische Analyse von Daten zugrundeliegenen zu verstehen und anwenden zu können.

1 Eine kleine Welt der Unsicherheit

Beginnen wir mit eine einfachen Modell. Dazu nehmen wir eine kleine Welt die nur aus 20 Personen besteht. In Figure 1.1 können wir alle Personen einzeln sehen. Die Gesamtheit aller Personen (allgmeine Objekte) über die wir eine Aussage treffen woll bezeichnen wir als eine Population.

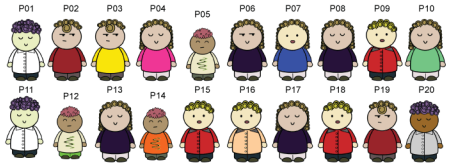


Figure 1.1: Eine kleine Welt

Definition 1.1 (Population)**.** Eine Population oder auch die Grundgesamtheit ist Gesamtheit aller Objekte/Dinge/Personen über die eine Aussage getroffen werden soll.

1.1 Ein Experiment

Wir wollen nun eine Krafttrainingsstudie durchführen um die Beinkraft zu erhöhen. Wir haben allerdings nur sehr wenige Ressourcen (bzw. wir sind faul) und können insgesamt nur sechs Messungen durchführen. Aus einem kürzlich durchgeführten Census haben wir aber die Kraftwerte der ganzen Population. Wir stellen die Kraftwerte zunächst mittels einer Tabelle dar (siehe Table 1.1)

Table 1.1: Kraftwerte (in Newton) der Lummerländer an der einbeinigen Beinpresse

ID	Kraft[N]	ID	Kraft[N]
P01	2414	P11	2243
P02	2462	P12	2497
P03	2178	P13	1800
P04	2013	P14	2152
P05	2194	P15	2089
P06	2425	P16	2090

ID	Kraft[N]	ID	Kraft[N]
P07	2305	P17	3200
P08	2117	P18	2196
P09	2298	P19	2485
P10	2228	P20	2440

Selbst bei 20 Werten ist diese Darstellung wenig übersichtlich. Wir könnten zwar Zeile für Zeile durchgehen und nach etwas notieren und suchen würden wir sehen das der Maximalwert bei 3200N für P17 und der Minimalwert von Person P13 bei 1800N liegt. Aber wirklich einfach ist diese Darstellung nicht. Für solche univariaten Daten (uni = eins) kann eine übersichtlichere Darstellung mittels eines sogenannten Dotplots erreicht werden (siehe Figure 1.2).

Figure 1.2: Dotplot der Lummerlandkraftdaten

Hier kann deutlich schneller abgelesen werden was das Minimum und das Maximum der Daten ist, sowie es kann auch direkt abgeschätzt werden in welchem Bereich sich der Großteil der Daten befindet. Allerdings wird durch diese Art der Darstellung die Information über welche Person die jeweiligen Werte besitzt nicht mehr dargestellt. Dies stellt in den meisten Fällen allerdings kein Problem dar, da wir in den meisten Fällen aussagen über die Gruppe und weniger über einzelne Personen machen wollen.

Gehen wir jetzt von der folgenden Fragestellung aus. Wir wollen den Gesundheitsstatus unserer Lummerländer verbessern und wollen dazu ein Krafttraining durchführen. Da evidenzbasiert arbeiten wollen, möchten wir überprüfen ob wirklich ein Verbesserung der Kraft durch das Training stattgefunden hat. Da es sich aber gleichzeitig um unsere selbst geschaffene Welt handelt führen wir natürlich ein perfektes Krafttraining, eine perfekte Intervention, durch. D.h wir stellen uns immer wieder als unwissend da und geben vor das wir gar nicht wissen, das das Training perfekt effektiv ist.

D.h. wir führen gleichzeitig ein Gedankenexperiment durch. Wir führen ein Krafttraining für die Beine durch. Das Training ist perfekt und verbessert die Kraftleistung um genau +100N. Dieser Kraftzuwachs unabhängig davon welche Person aus unserer Population das Training durchführt (Warum ist das keine realistische Annahme?). Wir wollen zwei Gruppen miteinander vergleichen eine Interventionsgruppe und eine Kontrollgruppe. In beiden Gruppen

sollen jeweils $n_{\text{TRT}} = n_{\text{CON}} = 3$ TeilnehmerInnen bzw. Teilnehmer einbezogen werden da wir nicht mehr Ressourcen für mehr ProbandInnen haben.

Die erste Frage die sich nun stellt ist wie wählen wir die sechs Personen aus unserer Population aus und wie teilen wir die sechs Personen in die beiden Gruppen? Nach etwas überlegen kommen wir darauf, dass wir am besten eine zufällige Stichprobe ziehen sollten (Warum?).

Definition 1.2 (Stichprobe)**.** Eine Stichprobe ist eine Teilmenge der Objekte aus der Population.

Definition 1.3 (Zufallsstichprobe)**.** Eine Zufallsstichprobe ist eine Teilmenge der Objekte aus der Population die *zufällig* ausgewählt wurde.

Diese sechs Personen, unsere Stichprobe, wird dann wiederum zufällig auf die beiden Gruppen aufgeteilt.

Ein Zufallszahlengenerator hat die Zahlen $i = \{3, 7, 8, 9, 10, 20\}$ gezogen. Die entsprechenden Personen werden aus der Population ausgewählt und wiederum zufällig in die beiden Gruppen aufgeteilt (siehe Table 1.2).

Table 1.2: Zufällig ausgewählte Stichprobe der Kontrollgruppe (CON) und der Interventionsgruppe (TRT).

ID	Kraft[N]	Gruppe
P08	2117	CON
P09	2298	CON
P03	2178	CON
P07	2305	TRT
P10	2228	TRT
P20	2440	TRT

Mit diesen sechs Personen führen wir jetzt unser Experiment durch. Die drei Personen aus der Kontrollgruppe, unterlaufen im Interventionszeitraum nur ein Stretchtraining während die Interventionsgruppe zweimal die Woche für 12 Wochen unser perfektes Krafttraining durchführt. Nach diesem Zeitraum messen wir alle Personen aus beiden Gruppen und erhalten das folgende Ergebnis (siehe Table 1.3).

Für beide Gruppen ist jeweils der Mittelwert berechnet worden, um die Wert miteinander vergleichen zu können. Später werden wir noch weitere Maße kennenlernen die es ermöglichen zwei Mengen von Werten miteinander zu vergleichen.

Table 1.3: Ergebnis der Intervention in Experiment 1 für die Kontroll- und die Interventionsgruppe.

(a) Kontrollgruppe

ID	Kraft[N]
P08	2117
P09	2298
P03	2178
$\bar{K}$	2198

(b) Interventionsgruppe

ID	Kraft[N]
P07	2405
P10	2328
P20	2540
$\bar{K}$	2424

Definition 1.4 (Mittelwert)**.** Der Mittelwert über n Werte berechnet sich nach der Formel:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

Der Mittelwert wird mit einem Strich über der Variable dargestellt.

Damit lernen wir direkt auch ein neues Konzept kennen. Nämlich das der Statistik. Ein Wert der auf der erhobenen Stichprobe berechnet wird, wird als Statistik bezeichnet.

Definition 1.5 (Statistik)**.** Ein auf einer Stichprobe berechnet Wert, wird als Statistik bezeichnet.

Um jetzt Unterschied zwischen den beiden Gruppen zu untersuchen berechnen wir die Differenz D zwischen den beiden Mittelwerten $D = \bar{K}_{\text{TRT}} - \bar{K}_{\text{CON}}$. Die Differenz kann natürlich auch in die andere Richtung berechnet werden und es würde sich das Vorzeichen ändern. Hier gibt es keine Vorgaben, sondern die Richtung kann frei bestimmt werden. Wenn bekannt ist in welcher Richtung der Unterschied berechnet wird, dann stellt dies keine Problem dar. Im vorliegenden Fall ziehen wir die Interventionsgruppe von der Kontrollgruppe ab, da wir davon ausgehen, dass die Intervention zu einer Krafterhöhung führt und wir dadurch einen positiven Unterschied erhalten (vgl. Equation 1.2)

$$D = 2424N - 2198N = 226N \tag{1.2}$$

Da der Wert D, wiederum auf den Daten der Stichprobe berechnet wird, handelt es sich ebenfalls um eine Statistik.

In Figure 1.3 sind die Werte der beiden Gruppen, deren Mittelwerte $\bar{K}_{\text{CON}}$ und $\bar{K}_{\text{TRT}}$ und der Unterschied D zwischen diesen abgebildet. Wie erwartet zeigt die Interventionsgruppen den höheren Kraftwert im Vergleich zu der Kontrollgruppe. Allerdings ist der Wert mit $D = 226$ größer als der tatsächliche Zuwachs von $\Delta_{\text{Training}} = 100$ (Warum ist das so?).

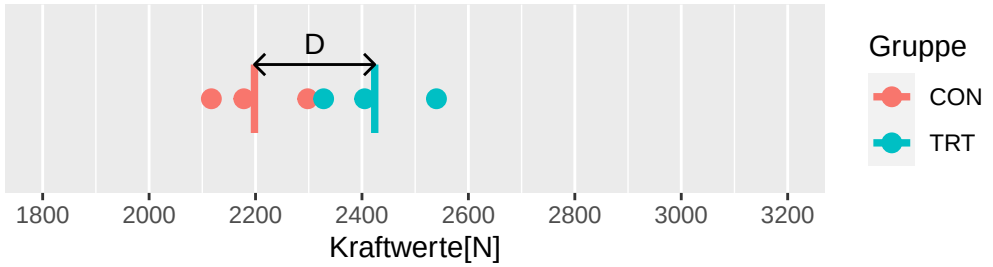


Figure 1.3: Dotplot der beiden Stichproben. Senkrechte Striche zeigen die jeweiligen Mittelwerte an.

Der Unterschied zwischen den beiden Gruppen ist natürlich auch zum Teil auf die Unterschiede die zwischen den beiden Gruppen vor der Intervention bestanden haben zurück zu führen. Was wäre denn passiert, wenn wir eine andere Stichprobe gezogen hätten?

Sei $i = \{12, 2, 19, 4, 8, 16\}$ eine zweite Stichprobe. Dies würde zu den folgenden Werten führen nach der Intervention führen.

Table 1.4: Ergebnis der Intervention in Experiment 2 für die Kontroll- und die Interventionsgruppe.

ID	Kraft[N]	Gruppe
P08	2117	CON
P09	2298	CON
P03	2178	CON
P07	2405	TRT
P10	2328	TRT
P20	2540	TRT

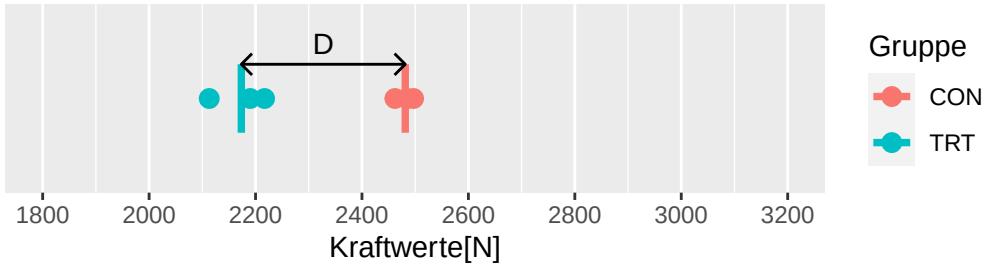


Figure 1.4: Dotplot der beiden Stichproben in Experiment 2. Senkrechte Striche zeigen die jeweiligen Mittelwerte an.

In Figure 1.4 sind wiederum die Datenpunkte, Mittelwerte und der Unterschied abgetragen. In diesem Fall ist allerdings die Differenz zwischen den beiden Gruppen genau in der anderen Richtung $D = -308$, so dass die Interpretation des Ergebnisses genau in der anderen Richtung wäre. Nämlich, nicht nur hat das Krafttraining zu keiner Verbesserung in der Kraftfähigkeit geführt, sondern zu einer Verschlechterung!

Es hätte aber auch sein können, das wir noch eine andere Stichprobe gezogen hätten, z.B. $i = \{6, 5, 7, 20, 14, 16\}$. Dies würde zu dem folgenden Ergebnis führen (siehe Table 1.5).

Table 1.5: Mittelwertsdaten aus Experiment 3 und der Unterschied D zwischen den beiden Gruppenmittelwerten

Gruppe	Kraft[N]
CON	2308
TRT	2327
D	19

In diesem Fall haben wird zwar wieder einen positiven Unterschied zwischen den beiden Gruppen in der zu erwartenden Richtung gefunden. Der Unterschied von $D = 19$ ist allerdings deutlich kleiner als das tatsächlichen $\Delta = 100$. Daher würden wir möglicherweise das Ergebnis so interpretieren, führen, dass wir das Krafttraining als ineffektiv bewerten würden und keine Empfehlung ausprechen.

Zusammengenommen, ist keines der Ergebnisse 100% korrekt. Entweder der Unterschied zwischen den beiden Gruppen ist deutlich zu groß, oder in der anderen Richtung oder deutlich zu klein. Das Ergebnis des Experiments hängt ursächlich damit zusammen, welche Stichprobe gezogen wird. Diese Einsicht gilt in jedem Fall generell für jedes Ergebnis eines Experiments.

Das Phänomen, das der Wert der berechneten Statistik zwischen Wiederholungen des Experiments schwankt wird als Stichprobenvariabilität bezeichnet.

Definition 1.6 (Stichprobenvariabilität)**.** Durch die Anwendung von Zufallsstichproben, variiert eine auf den Daten berechnete Statistik. Die Variabilität wird als Stichprobenvariabilität bezeichnet.

Streng genommen, führt die Stichprobenvariabilität für sich genommen noch nicht dazu, das sich die Statistik zwischen Wiederholungen des Experiments verändert, sondern die zu untersuchenden Werte in der Population müssen selbst auch noch eine Streuung aufweisen. Wenn wir eine Population untersuchen würden, bei der alle Personen die gleiche Beinkraft hätten, würden unterschiedliche Stichproben immer den gleichen Mittelwert haben und wiederholte Durchführung des Experiment würden immer wieder zu dem selben Ergebnis führen. Dieser Fall ist in der Realität aber praktisch nie gegeben und sämtlich Parameter für die wir uns hier interessieren zeigen immer eine natürlich Streuung in der Population. Diese Streuung in der Population führt daher zu dem besagten Ergebnis, das das gleiche Experiment mehrmals wiederholt zu unterschiedlichen Zufallsstichproben führt und dementsprechend immer zu unterschiedlichen Ergebnissen führt.

Daher ist eine der zentrale Aufgabe der Statistik mit dieser Variabilität umzugehen und die Forscherin trotzdem in die Lage zu versetzen rationale Entscheidungen zu treffen. Eine implizite Kernannahme dabei ist, das wir mit Hilfe von Daten überhaupt etwas über die Welt lernen können. D.h. das uns die Erhebung von Daten überhaupt auch in die Lage versetzt rationale Entscheidungen zu treffen. Entscheidungen wie ein spezialisiertes Krafttraining mit

einer klinischen Population durchzführen oder eine bestimmte taktische Variante mit meiner Mannschaft zu trainieren um die Gegner besser auszuspielen. Alle diese Entscheidungen sollten rational vor dem Hintergrund von Variabilität getroffen werden und auch möglichst oft korrekte Entscheidungen zu treffen. Wie wir sehen werden, kann uns die Statistik leider nicht garantieren immer die korrekte Entscheidungen zu treffen. Nochmal auf den Punkt gebracht nach Wild and Seber (2000, 28)

> The subject matter of statistics is the process of finding out more about the real world by collecting and then making sense of data.

Untersuchen wir jedoch zunächst unsere Einsicht, das Wiederholungen des gleichen Experiments zu unterschiedlichen Ergebnissen führt, weiter. In unserem Beispiel aus Lummerland haben wir nämlich den Vorteil, das uns die Wahrheit bekannt ist. In Figure 1.5 ist die Verteilung unsere bisheringen drei Ds abgetragen.

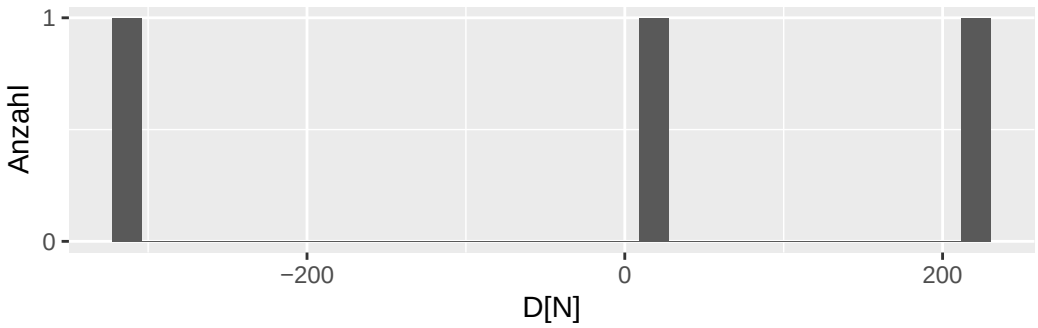


Figure 1.5: Bisherige Verteilung der Unterschiede D

Die drei Werte liegen ja relativ weiter auseiander. Eien Anschlussfrage könnte jetzt sein: "*Welche weiteren Werte sind denn überhaupt möglich mit der vorliegenden Population?*".

1.2 Die Stichprobenverteilung

Wir können jetzt ja einfach mal das Experiment anfangen zu wiederholen. In Figure 1.6 sind mal 15 verschiedene Stichproben abgetragen. Wir haben in jeder Zeile jeweils sechs TeilnehmerInnen gezogen. Drei für die Kontrollgruppe und drei für die Inervationsgruppe. Für jede dieser Zeilen können wir jeweils den Gruppenmittelwert berechnen und den Unterschied D bestimmen.

Warum eigentlich bei 15 aufhören. Wir haben ja den Vorteil, das unsere Population relativ übersichtlich ist. Vielleicht können wir uns ja noch aus unserer Schulezeit an Kombinatorik erinnern. Da haben wir den Binomialkoeffizienten kennengelernt. Die Anzahl der möglichken Kombination von k Elementen aus einer Menge von n Elementen berechnet sich nach:

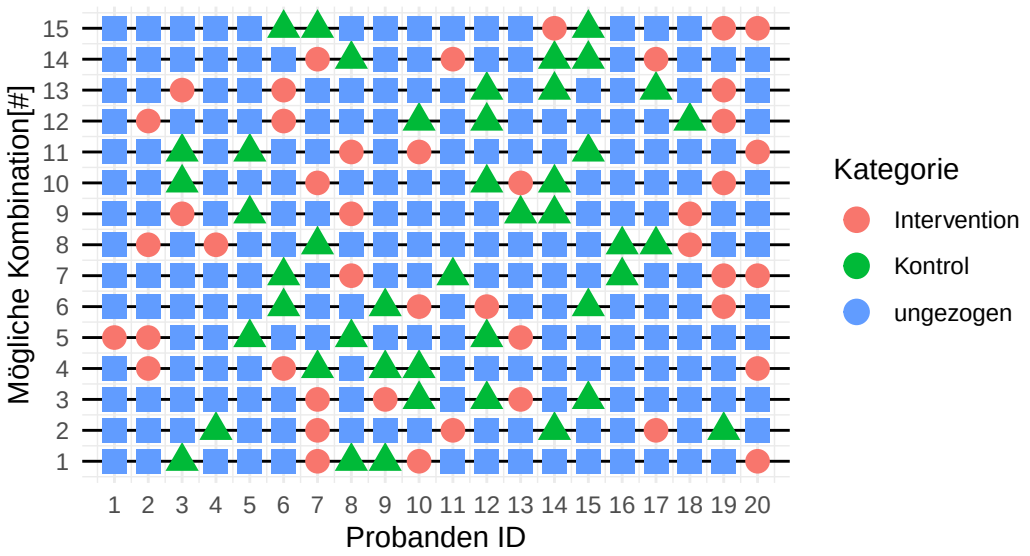


Figure 1.6: Beispiele für verschiedene Möglichkeiten zwei Stichproben mit jeweils $n_i = 3$ aus der Population zu ziehen

$$\text{Anzahl} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.3}$$

In unserem Fall wollen wir zunächst sechs Elemente aus $N = 20$ auswählen und dann drei Elemente aus den sechs gezogenen Elementen auswählen um diese entweder der Interventionsgruppe oder der Kontrollgruppe zu zuweisen (Warum brauchen wir uns nur eine Gruppe anzuschauen?). Die Anzahl der möglichen Stichprobenkombinationen ist folglich:

$$\text{Anzahl} = \binom{20}{6}\binom{6}{3} = 7.752 \times 10^5 \tag{1.4}$$

Das sind jetzt natürlich selbst bei dieser kleinen Population ein große Menge von einzelnen Experimenten, aber dafür sind Computer da, die können alle diese Experiment in kurzer Zeit durchführen. In Figure 1.7 ist die Verteilung aller möglichen Experimentausgänge, d.h. alle Differenzen D zwischen der Interventions- und der Kontrollgruppe, abgebildet.

Auf der x-Achse sind die möglichen Differenzen D abgetragen, während auf der y-Achse die relative Häufigkeit, d.h. die Häufigkeit für einen bestimmten D-Wert geteilt durch die Anzahl 7.752×10^5 aller möglichen Werte. Die Verteilung der D's wird als Stichprobenverteilung bezeichnet.

Definition 1.7. Die Stichprobenverteilung kennzeichnet die Verteilung der beobachteten Statistik.

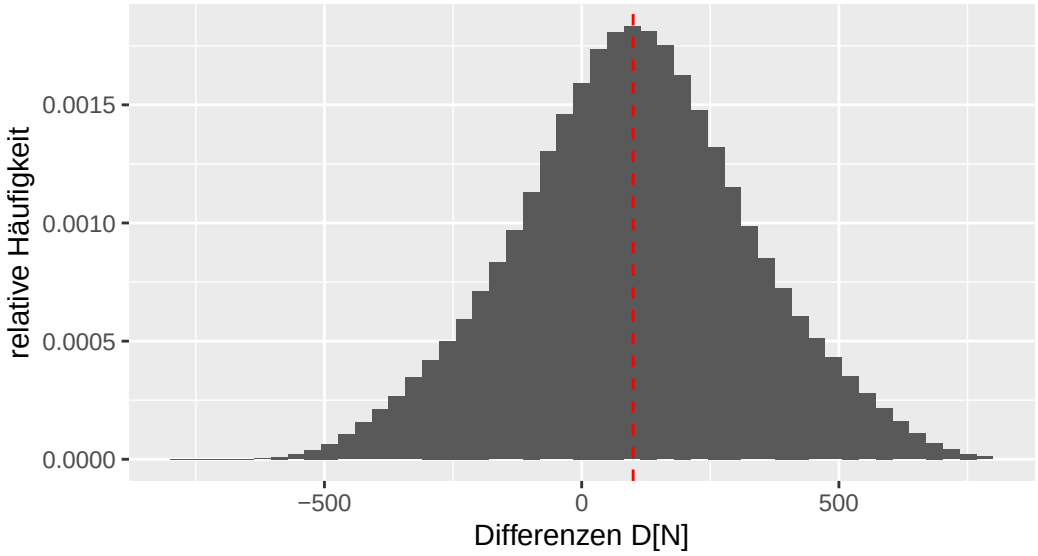


Figure 1.7: Verteilung aller möglichen Differenzen zwischen Kontroll- und Interventionsgruppe bei einer Intervention mit $\Delta = 100$ (im Graphen mittels der roten Linie angezeigt).

Die Figure 1.7 zeigt, dass die überwiegende Anzahl der Ausgänge tatsächlich auch im Bereich von $\Delta = 100$ liegen. Noch präziser das Maximum der Verteilung, also die höchste relative Häufigkeit liegt genau auf der roten Linie. Dies sollte uns etwas beruhigen, denn es zeigt, das unsere Art der Herangehensweise mittels zweier Stichproben auch tatsächlich in den meisten Fällen einen nahezu korrekten Wert ermittelt. Allerdings zeigt die Stichprobenverteilung auch das Werte am rechten Ende die deutlich zu hoch sind wie auch Werte am linken Ende der Verteilung die deutlich in der falschen Richtung möglich sind. Das bedeutet, wenn wir das Experiment nur einmal durchführen wir uns eigentlich nie sich sein können, welches dieser vielen Experimente wir durchgeführt haben. Es ist zwar warscheinlicher, dass wir eins aus der Mitte der Verteilung durchgeführt haben, einfach da die Anzahl größer ist, aber wir haben keine 100% Versicherung, das wir nicht *Pech* gehabt haben und das Experiment ganz links mit $D = -500$ oder aber das Experiment ganz rechts mit $D = 700$ durchgeführt haben. Diese Unsicherheit wird leider keine Art von Experiment vollständig auflösen können. Eine weitere Eigenschaft der Verteilung ist ihre Symmetrie bezüglich des Maximums mit abnehmenden relativen Häufigkeiten umso weiter von Maximum D entfernt ist (Warum macht das heuristisch Sinn?).

Die Darstellungsform von Figure 1.7 wird als Histogramm bezeichnet und eignet sich vor allem dazu die Verteilung einer Variablen z.B. x darzustellen. Dazu wird der Wertebereich von x zwischen dem Minimalwert $x_{\min}$ und dem Maximalwert $x_{\max}$ in k gleich große Intervalle unterteilt und die Anzahl der Werte innerhalb jedes Intervalls wird abgezählt und durch die Anzahl der Gesamtwerte geteilt um die relative Häufigkeit zu erhalten.

Zum Beispiel für die Werte:

$$x_i \in \{1, 1.5, 1.8, 2.1, 2.2, 2.7, 2.8, 3.5, 4\}$$

könnte das Histogram ermittelt werden, indem der Bereich von $x_{\text{min}} = 1$ bis $x_{\text{max}} = 4$ in vier Intervalle unterteilt wird und dann die Anzahl der Werte in den jeweiligen Intervallen ermittelt wird (siehe Figure 1.8). Die ermittelte Anzahl würde dann noch durch die Gesamtanzahl 9 der Elemente geteilt um die relative Häufigkeit zu berechnen.

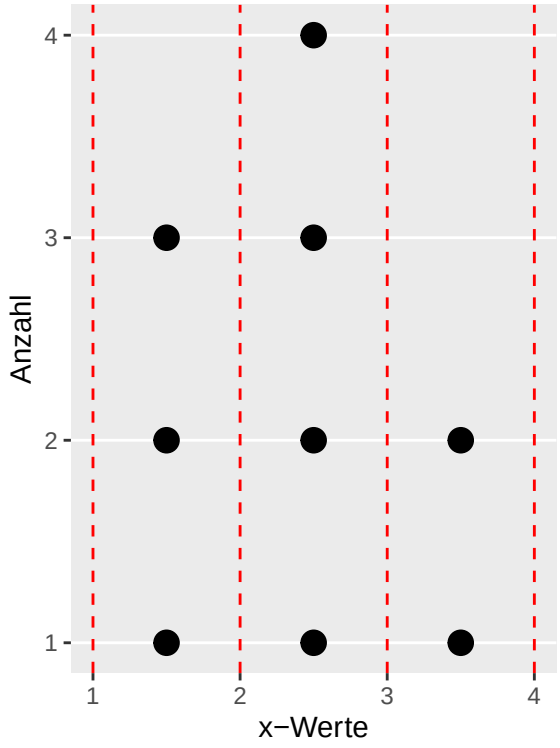


Figure 1.8: Beispiel für die Darstellung eines Histogramms für die Daten x_i.

Die Form des Histogramms hängt davon ab wie viele Intervalle verwendet werden, so wird die Auflösung mit mehr Intervallen besser, aber es die Anzahl wird geringer und andersherum wird die Auflösung mit weniger Intervallen geringer aber die Anzahl der Elemente pro Intervall wird größer und somit stabiler. Daher sollte in den meisten praktischen Fällen die Anzahl variiert werden um sicher zu gehen, das nicht nur zufällig eine spezielle Darstellung verwendet wird.

Zurück zu unserer Verteilung von D unter $\Delta = 100\text{N}$ in Figure 1.7. Wie schon besprochen sind alle Werte zwischen etwa $D = -500N$ und $D = 700\text{N}$ plausibel bzw. möglich. Schauen wir uns doch einmal an, was passiert wenn das Training überhaupt nichts bringen würde und es keine Verbesserung gibt, also $\Delta = 0$.

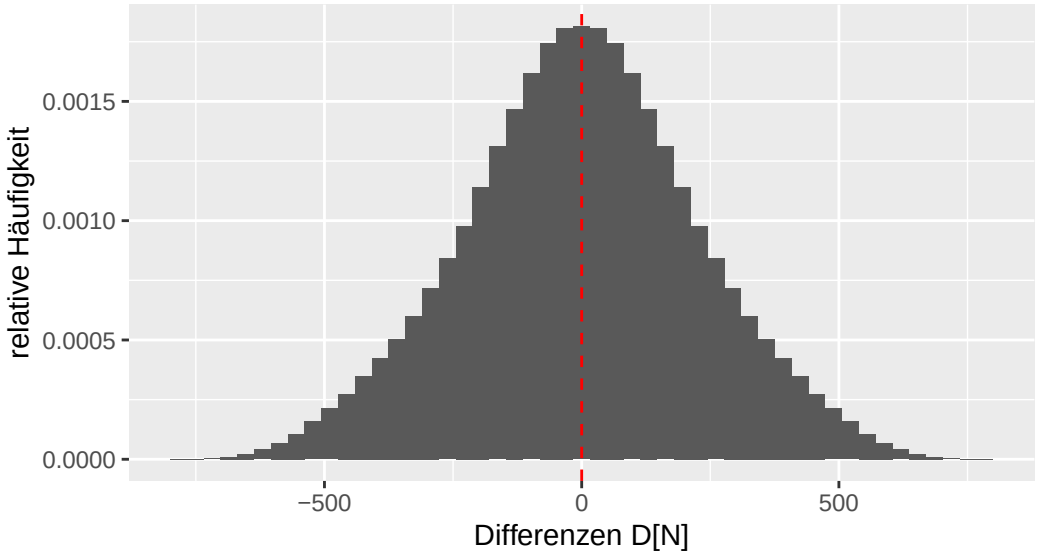


Figure 1.9: Verteilung aller möglichen Differenzen zwischen Kontroll- und Interventionsgruppe wenn $\Delta = 0$ (rote Linie).

Die Verteilung in Figure 1.9 sieht praktisch genau gleich aus, wie diejenige für $\Delta = 100$. Der einzige Unterschied ist lediglich das sie nach links verschoben ist und zwar scheinbar genau um die 100N Unterschied zwischen den beiden Δs. Dies ist letztendlich auch nicht weiter verwunderlich, bei der Berechnung des Unterschied D zwischen den beiden Gruppen kommen in beiden Fällen genau die gleichen Kombination vor. Bei $\Delta = 100$ wird aber zu der Interventionsgruppe das Δ dazuaddiert bevor die Differenz der Mittelwerte berechnet wird. Da aber gilt:

$$D = \frac{1}{3}\sum_{i=1}^{3} x_{\text{KON}i} - \frac{1}{3}\sum_{j=1}^{3}(x_{\text{TRT}j} + \Delta) = \bar{x}_{\text{KON}} - \bar{x}_{\text{TRT}} + \Delta$$

Daher bleibt die Form der Verteilung immer genau gleich und wird lediglich um den Wert Δ im Vergleich zur Nullintervention verschoben. Wobei mit Nullintervention Umgangssprachlich die Intervention bezeichnet, bei der nichts passiert also $\Delta = 0$ gilt.

1.3 Unsicherheit in Lummerland

Das führt jetzt aber zu einem Problem für uns. Gehen wir jetzt nämlich von diesen beiden Annahmen aus, das entweder die Intervention effektiv ist $\Delta = 100$ gilt oder das die Intervention nichts bringt also $\Delta = 0$ gilt. Wenn wir diese beiden Verteilungen übereinander legen erhalten

wir Figure 1.10. Wir haben die Darstellung jetzt etwas verändert und eine Kurve durch die relativen Häufigkeiten gelegt. Dieser Graphen wird jetzt nicht mehr als Histogramm sondern als Dichtegraph bezeichnet.

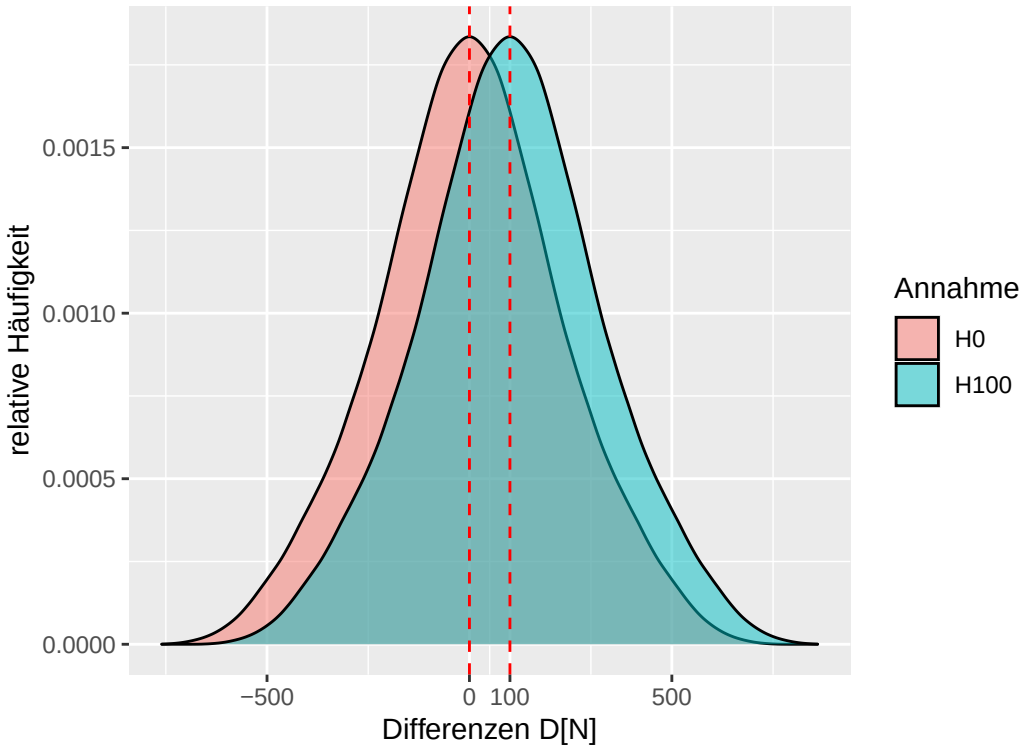


Figure 1.10: Verteilung aller möglichen Differenzen zwischen Kontroll- und Interventionsgruppe wenn $\Delta = 0$ und $\Delta = 100$.

In Figure 1.10 ist klar zu sehen, dass die beiden Graphen zu großen Teilen überlappen und dazu noch in einem Bereich wo beide Ergebnisse ihrer höchsten relativen Häufigkeiten, also auch die größte Wahrscheinlichkeit haben unter den jeweiligen Annahmen aufzutreten. Unser Problem besteht jetzt darin, dass wir in der Realität gar nicht diese Information haben welchen Effekt unser Training auf die Stichprobe ausführt. Wenn wir dies wüssten, dann müssten wir das Experiment ja gar nicht durchführen. Wir haben im Normalfall nur ein einziges Ergebnis, nämlich den Ausgang unseres einen Experiments.

Wenn wir jetzt unser Experiment einmal durchgeführt haben und ein einziges Ergebnis für D erhalten haben, sei zum Beispiel $D = 50$ dann haben wir ein Zuweisungsproblem (siehe Figure 1.11). Wie weisen wir unser Ergebnis jetzt den beiden möglichen Realität zu? Einmal kann es sein, das das Krafttraining aber auch gar nichts gebracht hat und wir haben lediglich eine der vielen möglichen Stichprobenkombination beobachtet haben die zu einem positiven Wert für D führt. Oder aber das Krafttraining ist effektiv gewesen und hat zu einer Verbesserung von $\Delta = 100$N geführt und wir haben lediglich ein Stichprobenkombination aus den vielen

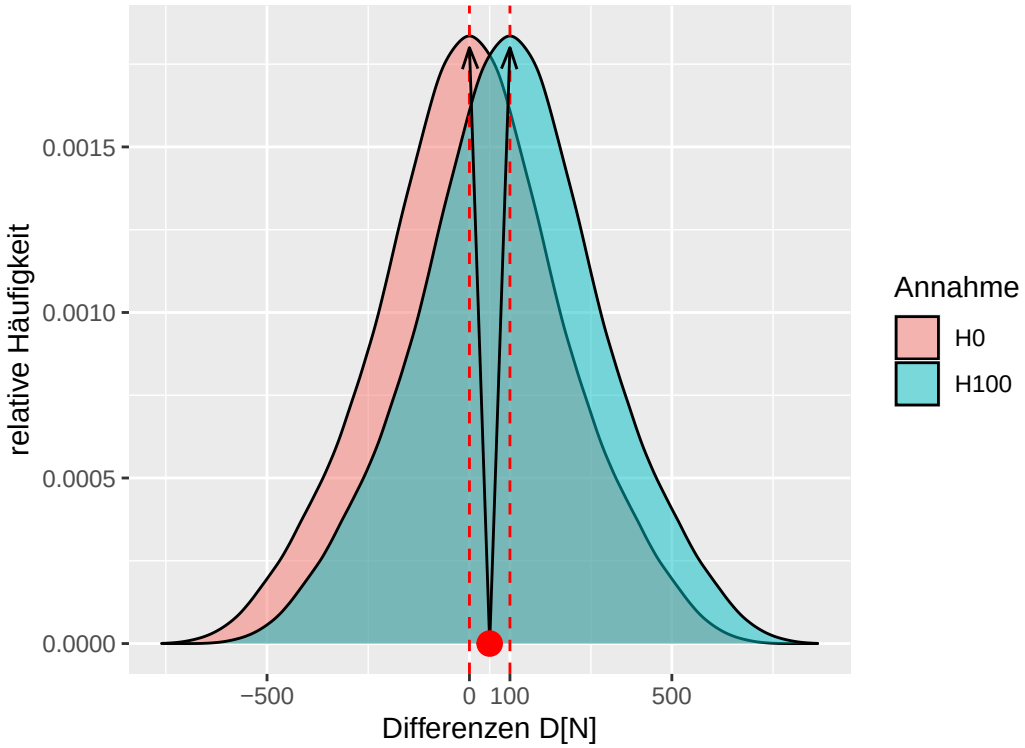


Figure 1.11: Zuweisung eines beobachteten Unterschieds D nach einem Experiment

Table 1.6: Entscheidungsmöglichkeiten wenn entweder H_0 oder H_1 zutrifft.

	Realität	
	H_0	H_1
H_0	korrekt	β
H_1	α	korrekt

möglichen Stichprobenkombination gezogen die zu einem Ergebnis von $D = 50$ führt. Noch mal, in der Realität wissen wir nicht welche der beiden Annahmen korrekt ist und können es auch nie vollständig wissen. Denn egal wie viele Experimente wir machen, wir können immer den zwar unwahrscheinlichen aber nicht unmöglichen Fall haben, das wir nur Werte beispielsweise aus dem linken Teil der Verteilung beobachten. Das heißt wir haben immer mit einer Ungewissheit zu kämpfen. Wir können nicht im Sinne eines Beweises zeigen, das das Training effektiv ist.

Die Methoden der Statistik liefern uns nun Werkzeuge an die Hand um trotzdem rational zu Entscheiden welche der beiden Annahmen möglicherweise wahrscheinlicher ist. Gleichzeitig ermöglicht uns die Statistik abzuschätzen respektive zu berechnen wie groß die Unsicherheit in dieser Entscheidung ist. Die Statistik sagt dabei immer nur etwas über die beobachteten Daten aus. Die Statistik sagt jedoch nichts über die zugrundeliegenden wissenschaftlichen Theorien aus.

Schauen wir uns jetzt als vorläufig letzten Punkt an welche Entscheidungsmöglichkeiten wir haben.

1.4 Eine Entscheidung treffen

Wir hatten im Beispiel zwei verschiedene Annahmen, einmal das das Training nichts bringt und keine Verbesserung der Kraftfähigkeit folgt $\Delta = 0N$. Andererseits hatten wir das Beispiel gestartet damit, dass die Kraftfähigkeit um $100N$ zunimmt, also $\Delta = 100N$. Wie bezeichnen jetzt diese beiden Annahmen als Hypothesen und bezeichnen $\Delta = 0N$ als die Nullhypothese H_0 und $\Delta = 100N$ als die Alternativhypothese H_1.

Wenn wir jetzt das Experiment durchgeführt haben, können wir uns also entweder für die H_0 oder die H_1 entscheiden. Aus Gründen der Symmetrie ist dies gleichbedeutend wenn wir uns nur auf die H_0 fokussieren und entweder die H_0 annehmen bzw. beibehalten oder verwerfen also uns gegen H_0 entscheiden.

In Table 1.6 sind die verschiedenen Entscheidungsmöglichkeiten abgetragen. In der Realität gehen wir, wie gesagt, von zwei Fällen aus. Entweder trifft die H_0 oder die H_1 zu. Wenn die $H_=$ zutrifft und wir uns für die H_0 entscheiden, dann haben wir eine korrekte Entscheidung getroffen. Wenn H_0 zutrifft und wir allerdings die H_0 ablehnen, also uns für die H_1 entscheiden

ist unsere Entscheidung falsch und wir begehen einen Fehler. Dieser Fehler wird als Fehler 1. Art bzw. α-Fehler bezeichnet. Trifft in der Realität dagegen die H_1 zu und wir entscheiden uns gegen die H_0 und für die H_1, dann haben wir wiederum eine korrekte Entscheidung getroffen. Zuletzt, wenn die H_1 zutrifft und wir uns aber für die H_0 entscheiden, also die H_0 beibehalten bzw. uns gegen die H_1 entscheiden, treffen wir wieder eine falsche Entscheidung. Dieser Fehler wird als Fehler 2. Art, bzw. β-Fehler bezeichnet.

Definition 1.8. Wenn eine Entscheidung gegen die H_0 getroffen wird, obwohl die H_0 korrekt ist, wird dies als α-Fehler bezeichnet.

Definition 1.9. Wenn eine Entscheidung gegen die H_1 getroffen wird, obwohl die H_1 korrekt ist, wird dies als β-Fehler bezeichnet.

2 Statistische Signifikanz, p-Wert und Power

Im vorherigen Kapitel haben wir gesehen, wie Unsicherheit ein zentrales Problem bei der Interpretation von Ergebnissen von Experimenten oder Daten allgemein ist. Im nun folgenden Abschnitt wollen wir eine Prozess aufbauen, der es uns vor dem Hintergrund dieser Unsicherheit eine Entscheidung zu treffen.

2.1 Wie treffe ich eine Entscheidung?

In unserem kleine Welt Bespiel waren wir in der komfortablen Position, das wir genau wussten was passiert bzw. welcher Prozess unseren beobachteten Datenpunkt erzeugt hat. D.h wir kannten den datengenerieren Prozesses.

Definition 2.1 (Datengenerierender Prozess (DGP))**.** Der Prozess in der realen Welt der die beobachteten Daten und damit die daraus folgende Statistik erzeugt wird als datengenerierender Prozess bezeichnet.

Letztendlich zielt unsere Untersuchung, unser Experiment, darauf ab, Informationen über den DGP zu erhalten, weil diese Information uns erlaubt Aussagen über die reale Welt zu treffen. Dabei muss allerdings beachtet werden, dass dieser Prozess in den allermeisten Fällen ein starke Vereinfachung des tatsächlichen Prozesses in der Realität darstellt. Meistens sind die Abläufe in der Realität zu komplex um sie ins Gänze abzubilden. Somit wird fast immer nur ein Modell verwendet.

Zurück zu unseren Problem, wenn wir ein Experiment durchführen, dann haben wir normalerweise nur eine einzige beobachtete Statistik. In unseren bisherigen Beispiel also den berechneten Unterschied D in der Kraftfähigkeit nach der Intervention zwischen der Kontroll- und der Interventionsgruppe.

In Figure 2.1 ist der beobachtete Wert, $D = 50$ abgetragen. Wir wissen von vorne herein, dass dieser Wert beeinflusst ist durch die zufällige Wahl der Stichprobe und die daran geknüpfte Streuung der Werte in der Population. Wie können wir den nun überhaupt eine Aussage treffen darüber, ob das Krafttraining was bringt oder vielleicht nur einen sehr kleinen Effekt zeigt oder möglicherweise sogar schädlich ist also zu einer Abnahme der Kraft führt?

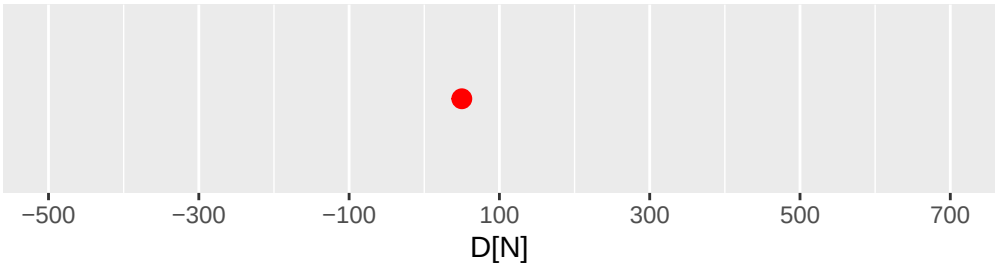


Figure 2.1: Beobachteter Unterschied nach der Durchführung unseres Experiments

Überlegen wir uns zunächst, welche Prozesse unseren beobachteten Wert zustande gebracht haben könnten. Wir haben schon zwei Prozesse kennengelernt, einmal den Prozess mit $\Delta = 100$ wie auch den Prozess mit $\Delta = 0$

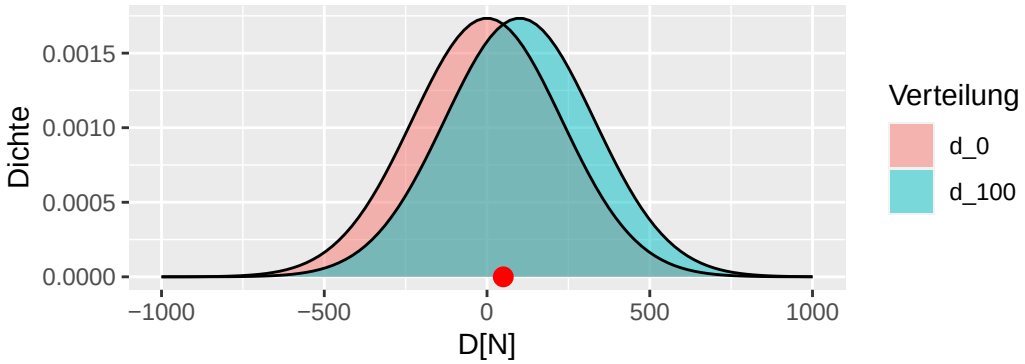


Figure 2.2: Mögliche datengenerierende Prozesse für den beobachteten Unterschied D (rot)

In Figure 2.2 ist wieder unser beobachteter Wert $D = 50$ und die beiden Verteilungen abgetragen. Leider können wir nicht eineindeutig sagen, welche der beiden Verteilungen, bzw. deren zugrundeliegende Prozesse, unseren beobachteten Wert erzeugt haben könnte. Da unser beobachteter Wert D genau zwischen den beiden Maxima der Verteilungen liegt. Etwas motiviertes Starren auf die Abbildung wird uns allerdings auf die Idee bringen, dass der beobachtete Wert nicht nur von diesen beiden Verteilungen erzeugt worden sein muss, sondern durchaus noch mehr Verteilungen in Frage kommen.

Figure 2.3 zeigt, dass selbst die Verteilung mit $\Delta = -250N$ und $\Delta = 350N$ nicht unplausibel sind den beobachteten Wert erzeugt zu haben. Warum aber bei diesen fünf Verteilungen aufhören, warum sollte Δ nicht -50 oder 127 sein. Und überhaupt, keiner kann behaupten die Natur kennt nur ganzzahlige Werte (siehe π). Warum sollte D also nicht auch $123.4567N$ sein?

Wenn diese Überlegung weitergeführt wird, dann wird schnell klar, dass letztendlich eine un-

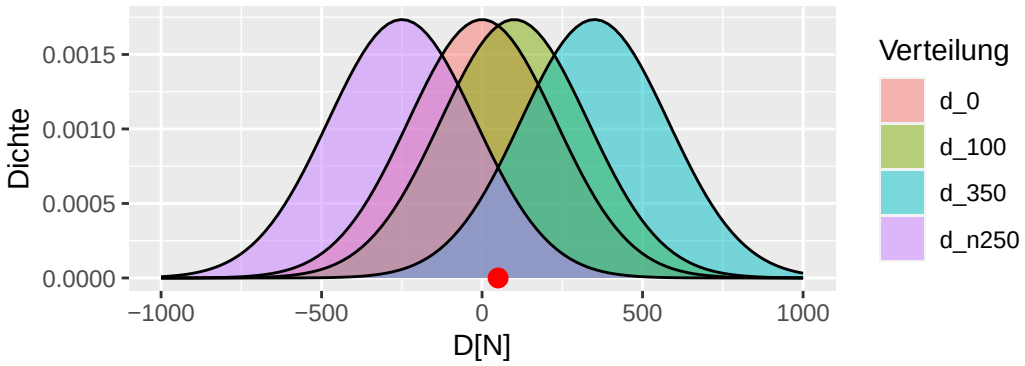


Figure 2.3: Beispiele für weitere mögliche Verteilungen als DGP.

endliche Anzahl von Verteilung in der Lage ist unseren beobachteten Wert plausibel zu generieren. D.h. wir haben ein Experiment durchgeführt und den ganzen Aufwand betrieben und haben wochenlang mit unseren ProbandInnen Krafttraining durchgeführt und sind hinterher eigentlich keinen Schritt weiter da wir immer noch nicht wissen was der datengenerierende Prozess ist. Also können wir selbst nach dem Experiment nicht sagen ob unser Krafttraining tatsächlich wirksam ist.

Zum Glück werden wir später sehen, das unser Unterfangen nicht ganz so aussichtslos ist. Schauen wir uns zum Beispiel die Verteilung für $\Delta = -350N$ an (Figure 2.4).

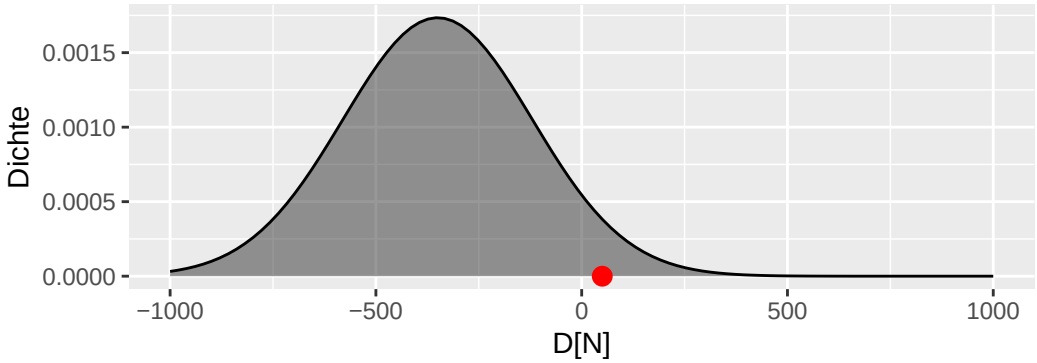


Figure 2.4: Verteilung für $\Delta = -350N$ und der beobachtete Wert D

Unser beobachteter Wert unter der Annahme das $\Delta = -350N$ ist nicht vollkommen unmöglich, aber so richtig *wahrscheinlich* erscheint er auch nicht. Der Wert liegt relativ weit am Rand der Verteilung. Die Kurve ist dort schon ziemlich nahe bei Null. D.h. der beobachtete Wert ist zwar durchaus möglich, aber es wäre schon überraschend wenn wir bei einer Durchführung des Experiments ausgerechnet so einen Wert beobachten würden wenn unsere angenommenes Δ korrekt ist.

Wenn wir jetzt dagegen von der Annahme ausgehen, dass dem DGP der Wert $\Delta = 50N$ zugrundeliegen würde, hätten wir die Verteilung in Figure 2.5. Zunächst ist dieser Wert möglich unter der Annahme. Zusätzlich liegt der beobachtete Wert mitten drin in dem Teil der Verteilung der auch zu erwarten wäre. D.h. der beobachtete Wert ist durchaus plausibel unter der Annahme und bei der einmaligen Durchführung des Experiments würde uns der beobachtete Wert nicht unbedingt überraschen.

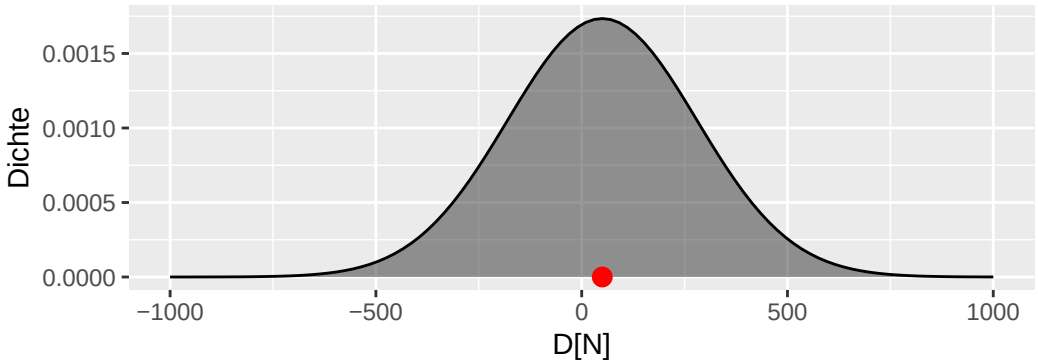


Figure 2.5: Verteilung für $\Delta = 50N$ und der beobachtete Wert D

Diesen Ansatz können wir verwenden um mit Hilfe unseres Experiments doch etwas über den DGP auszusagen. Allerdings müssen wir uns noch einmal etwas eingehender mit Verteilungen auseinandersetzen um z.B. genauer zu bestimmen welche Ergebnisse uns überraschen würden. D.h. wir müssen uns erst ein mal ein paar neue Konzepte erarbeiten.

2.2 Lage- und Skalenparameter

In Figure 2.3 hatten wir mehrere Verteilungen abgebildet. Die Verteilung haben die gleiche Form sind aber gegeneinander verschoben. D.h. sie unterscheiden sich bezüglich ihrer Position bzw. Lage. Der Parameter der bei einer Verteilungen die Lage steuert ist der sogenannte Erwartungwerts μ der auch als Mittelwert bezeichnet wird. Dieser Mittelwert μ unterscheidet sich allerdings von dem uns bereits bekannten Mittelwert $\bar{x}$ in der Stichprobe. In einem späteren Abschnitt werden wir uns genauer anschauen wie der Mittelwert μ berechnet wird.

2.2.1 Mittelwert μ der Population

Da der Mittelwert μ die Position der Verteilung bestimmt, ist μ ein Parameter der Verteilung. Die Beschreibung als Parameter der Verteilung bedeutet somit, dass die Verteilung von μ abhängt, oder formaler das die Verteilung eine Funktion von μ ist. Wenn wir uns an Funktionen aus der Schule zurück erinnen wo wir Funktionen f von x kennengelernt haben und als $f(x)$ dargestellt haben. Übertragen auf die Verteilung könnte dies mittels $f(\mu)$ dargestellt werden.

Betrachten wir zwei Verteilungen die sich bezüglich ihrer Mittelwerte μ unterscheiden. Zum Beispiel sei $\mu_1 = 0$ und $\mu_2 = 3$. Wie in Figure 2.6 zu sehen ist, führt dies dazu, das die beiden Verteilungen gegeneinander verschoben sind.

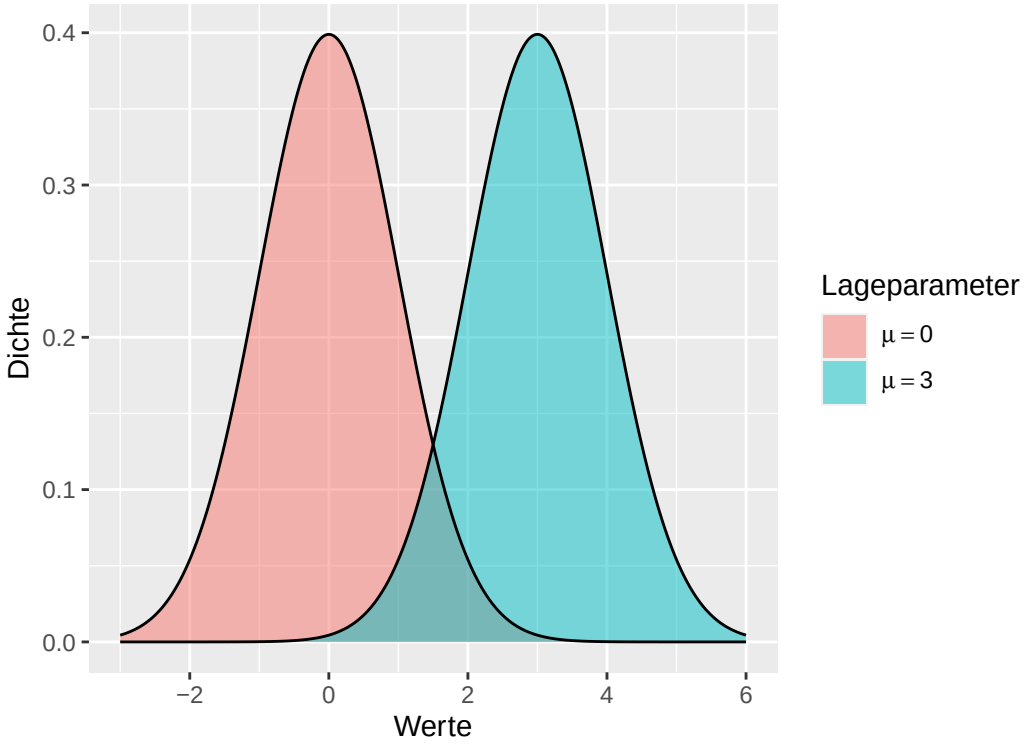


Figure 2.6: Verteilungen mit zwei unterschiedlichen Mittelwerten

Wie bereits erwähnt, wird der Mittelwert μ der Verteilung auch als Erwartungswert bezeichnet. Dies kann dahingehend interpretiert werden, das wenn Stichproben aus dieser Verteilungen gezogen werden, im Mittel der Wert μ erwartet werden kann. Soweit ist dies eigentlich noch nichts wirklich Neues, sondern hatten dies schon vorher gesehen, als wir alle möglichen Unterschiede zwischen der Kontrollgruppe und der Interventionsgruppe ermittelt haben. Hier war der Mittelwert der Verteilung genau derjenige Wert von Δ.

An dieser Stelle nochmal der Unterschied zwischen μ und $\bar{x}$. Der Mittelwert μ ist eine Eigenschaft der Population, also letztendlich ein Wert den wir niemals kennen werden ohne die gesamte Population zu untersuchen. Der Mittelwert $\bar{x}$ ist eine Eigenschaft der Stichprobe aus der Population. Also der konkrete Wert den wir anhand der Stichprobe berechnen. In vielen Fällen versuchen wir über $\bar{x}$ einen Rückschluss auf μ zu ziehen.

2.2.2 Standardabweichung σ der Population

Als zweite Eigenschaft von Verteilungen schauen wir uns jetzt die Streuung in der Population an. Die Streuung in der Population wird als Varianz bezeichnet und wird mit dem Symbol σ^2 bezeichnet. Schauen wir uns zunächst an, welchen Einfluss σ^2 auf die Form der Verteilung hat. In Figure 2.7 sind wieder zwei Verteilungen abgetragen. Dieses Mal ist μ in beiden Fällen gleich, aber die Varianzen σ^2 sind mit $\sigma_1^2 = 2$ und $\sigma_2^2 = 1$ unterschiedlich.

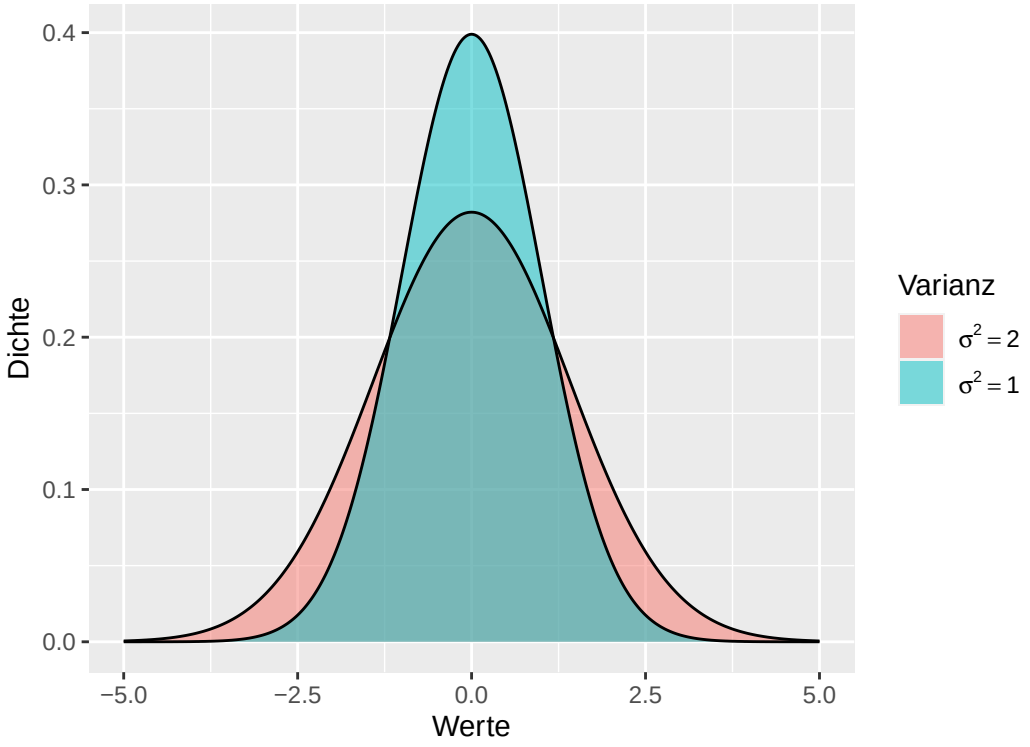


Figure 2.7: Verteilungen mit unterschiedlichen Varianzen

In Figure 2.7 ist zu sehen, dass beide Verteilungen ihren Mittelpunkt an der gleichen Stelle haben, aber die rote Verteilung mit $\sigma_1^2 = 2$ breiter ist als die andere Verteilung. Dies bedeutet das die Werte in der Verteilung stärker um den Mittelwert herum streuen. Wenn wir Werte aus der türkisen Verteilung ziehen, dann sollten diese näher um den Mittelwert $\mu = 0$ liegen, als dies bei der roten Verteilung der Fall ist.

Die Varianz σ^2 ist ebenfalls wie der Mittelwert ein Parameter der Verteilung. Sie bestimmt die die Form der Verteilung. D.h. wenn wir wieder unsere Schreibweise von eben verwenden und die Funktion f die Verteilung beschreibt, dann gilt $f(\sigma^2)$ oder eben zusammen mit dem Mittelwert μ, $f(\mu, \sigma^2)$.

Wenn aus der Varianz σ^2 die Wurzel gezogen wird, dann wird der resultierende Wert σ als

Standardabweichung bezeichnet. Da die Varianz σ^2 nur positive Werte annehmen kann, ist die Wurzelfunktion bzw. deren Umkehrung die Quadierung eineindeutig. Wenn wir die Standardabweichung kennen, dann kennen wir auch die Varianz und umgekehrt.

In der Stichprobe wird die Standardabweichung meistens mit dem Zeichen s bezeichnet und mittels der folgenden Formel berechnet:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{2.1}$$

D.h. die Standardabweichung ist die mittlere quadrierte Abweichung vom Mittelwert (siehe Formel (2.1)). Die Standardabweichung wird verwendet um die Streuung der Daten zu beschreiben. Die Standardabweichung hat den Vorteil, dass sie die gleiche Einheit hat wie der Mittelwert. Da die Abweichungen quadriert werden, also die quadrierten Einheiten haben, hat die Standardabweichung s die gleiche Einheit wie der Mittelwert $\bar{x}$. Da die Varianz die quadrierte Standardabweichung ist, hat die Varianz der Stichprobe s^2 daher die quadrierten Einheiten.

Wenn wir uns an unsere erstes Beispiel aus der kleinen Welt erinnern, dort hatten wir in der Kontrollgruppe die Personen $i = \{3, 8, 9\}$ gezogen, berechnen wir für diese Stichprobe die Standardabweichung erhalten mit dem Mittelwert $\bar{x} = 2198$:

$$s = \sqrt{\frac{(2178 - 2198)^2 + (2117 - 2198)^2 + (2298 - 2198)^2}{2}} = 92$$

Wir erhalten einen Wert von $s = 92N$. Wenn dieser Wert größer wird, dann streuen die Wert entsprechend weiter um den Mittelwert herum und entsprechend verringert sich die Streuung wenn die Standardabweichung s abnimmt.

2.3 Entscheidungen und μ und σ

Zeichnen wir in eine Verteilung die Standardabweichung ein, ergibt sich vollgendes Bild (siehe Figure 2.8).

Ein großteil der Werte liegt in dem Bereich $\mu \pm 1 \times \sigma$. Der Bereich $\mu \pm 2 \times \sigma$ beinhaltet schon fast alle Werte, während der Bereich $\mu \pm 3 \times \sigma$ fast alle Werte. Wenn wir die Verteilung noch etwas weiter nach links und rechts abtragen würden, würden wir sehen, dass auch noch Werte jenseits von $\mu \pm 3 \times \sigma$ liegen, aber nur noch sehr wenige. Diese Einsicht können wir dazu benutzen umgekehrt zu denken, wenn wir annehmen, das unsere Statistik dieser Verteilung folgt, welche Werte würde uns den *überraschen*. Welche Werte würden wir als Evidenz sehen um zu folgern: *Ich glaube nicht, dass die beobachtete Statistik aus der angenommen Verteilung stammt*!?

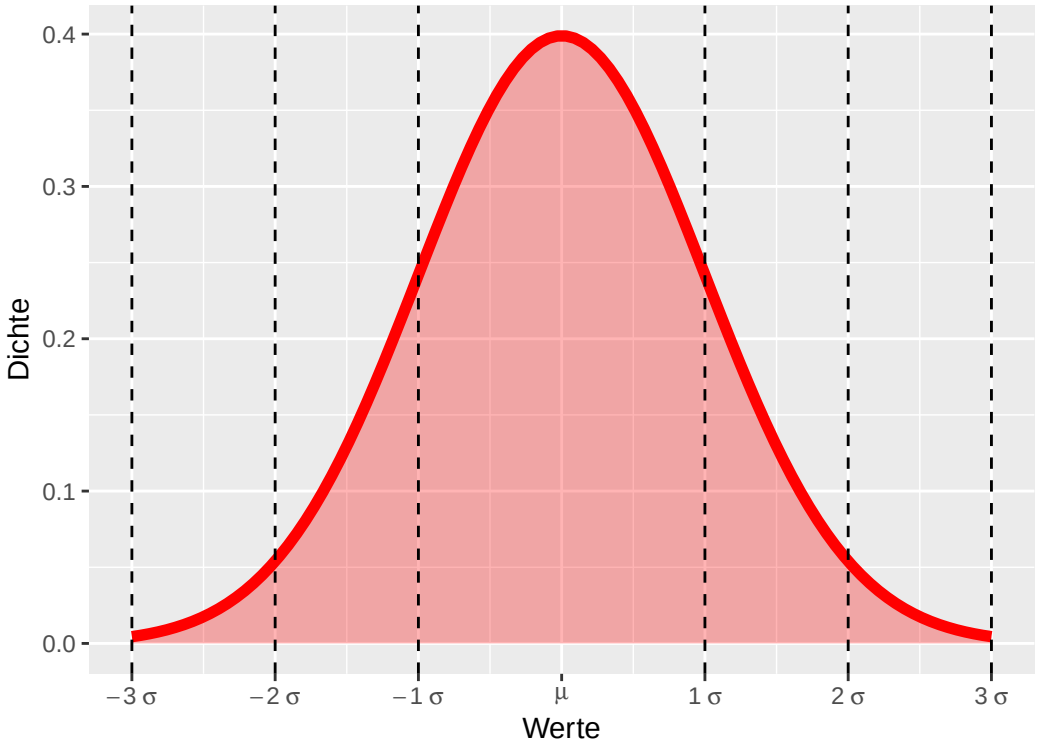


Figure 2.8: Verteilung mit verschiedenen mehrfachen der Standardabweichung σ\$

Table 2.1: Parameter einer Verteilung und deren Schätzer

Population	Stichprobe
Mittelwert μ	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Varianz σ^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standardabweichung σ	$s = \sqrt{s^2}$

Nun, zum Beispiel wenn der Wert mehr als $3 \times \sigma$ vom Mittelwert μ entfernt ist, dann wäre das zwar nicht unmöglich, aber es wäre schon ziemlich unwahrscheinlich so einen Wert zu beobachten. Vielleicht ist uns das aber ein zu schwer zu erreichender Wert, ein Kompromiss könnte ein Wert jenseits von $2 \times \sigma$ von μ entfernt, könnte auch schon als überraschen bezeichnet werden. Tatsächlich ist, die Wahrscheinlichkeit einen Wert jenseits von $2 \times \sigma$ zu beobachten etwa 5%. D.h. wir könnten einen Entscheidungsprozess erstellen bei dem wir sagen, wenn wir eine bestimmte Stichprobenverteilung für unsere Statistik annehmen. Wenn wir bei unserer Ausführung einen Wert beobachten der weiter als $2 \times \sigma$ von μ entfernt sind. Dann sind wir überrascht und sehen das als Evidenz gegen die Verteilungsannahme an.

Oder als Liste:

1) Setze eine Verteilung der Statistik mit definierten μ und σ als Annahme an.
2) Ziehe eine Zufallsstichprobe.
3) Berechne die Statistik auf der Stichprobe.
4) Überprüfe wie viele Standardabweichungen σ die Statistik von μ entfernt liegt.

2.3.1 Detour - Schätzer

Schauen wir uns noch einmal den Mittelwert μ der Population und den Mittelwert $\bar{x}$ der Stichprobe und deren Zusammenhang an. Der Mittelwert $\bar{x}$ der Stichprobe wird als sogenannter Schätzer verwendet. Diesen Begriff werden wir später noch genauer untersuchen. Im Moment reicht es sich zu merken, dass ein Schätzer eine Statistik ist, mit der wir einen Parameter der Population, z.B. μ, abschätzen wollen. Wie schon mehrmals erwähnt, den wahren Wert μ aus der Population werden wir mittels unserer Stichprobe niemals 100% korrekt bestimmen wir können aber mittels geschickt gewählter Statistiken Schätzer konstruieren die bestimmte Eigenschaften haben.

Nehmen wir zum Beispiel den Mittelwert $\bar{x}$. In unserer kleinen Welt kennen wir den Mittelwert μ unserer Population. Der Wert beträgt $\mu = 2291.3$. Schauen wir uns einmal an, was passiert, wenn wir alle möglichen Stichproben der Größe $N = 10$ unserer kleinen Welt bestimmen und die Verteilung der Mittelwert abtragen (siehe Figure 2.9).

In Figure 2.9 sehen wir, dass im Mittel der Stichprobenmittelwert $\bar{x}$ tatsächlich um den wahren Populationsmittelwert μ herum zentriert ist. Einzelne Ausgänge des *Experiments* können zwar

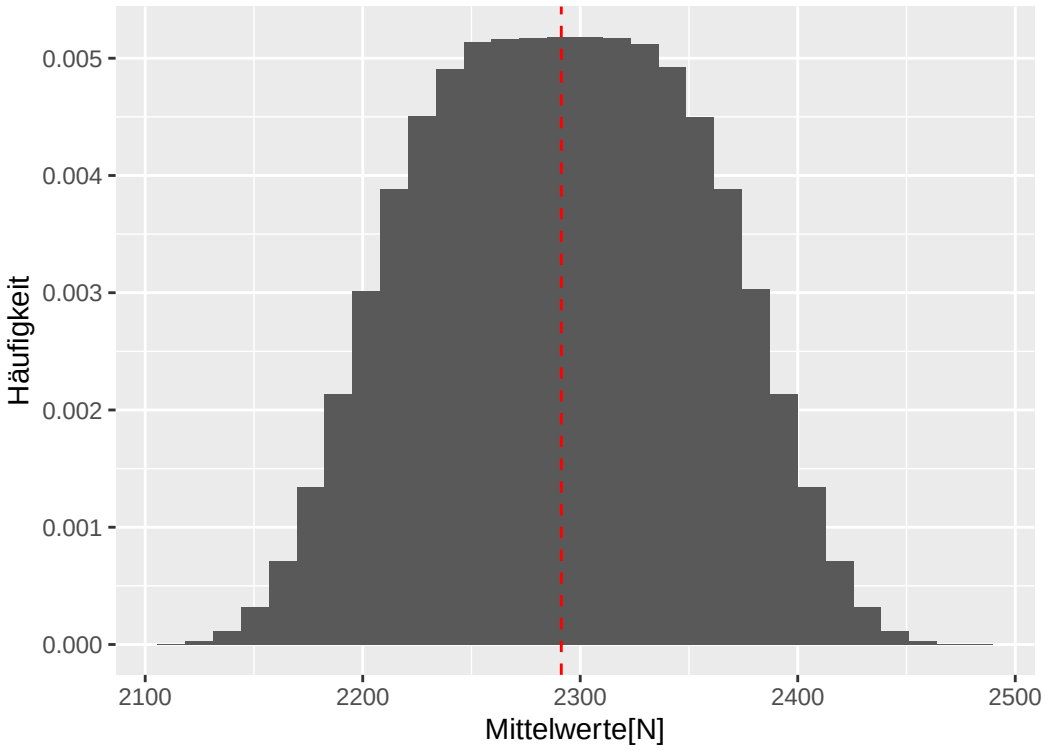


Figure 2.9: Verteilung der Mittelwerte von Stichproben der Größe $n = 10$, Kleine Welt Population μ (rot)

daneben liegen, der Großteil der Experiment gruppiert sich jedoch um μ herum. Der Stichprobenmittelwert $\bar{x}$ ist daher eine *gute* Statistik um den tatsächlichen Populationsmittelwert μ abzuschätzen.

2.4 Welche Verteilung setzen wir an?

Kommen wir aber wieder zurück zu unserem Ausgangsproblem, dass wir anhand unserer beobachteten Stichprobe etwas über die Effektivität der Kraftintervention aussagen wollen. Wie hilft uns jetzt die Kenntnis von Mittelwert μ oder $\bar{x}$ und der Standardabweichung σ bzw. s weiter? Wenn die Verteilung unserer Statistik der Form folgt wie sie bisher jetzt mehrmals beobachtet haben, dann können wir davon ausgehen, dass wenn wir eher Wert in der Nähe des Mittelpunkts erwarten würden. Wie werden selten genau den Mittelpunkt beobachten aber wir würde schon sehr überrascht sein, wenn wir Werte weit ab des Mittelwerts beobachten würden. Ab welcher Weite diese Werte als überraschen eingestuft werden hängt dabei von der Streuung der Verteilung an. Wenn σ groß ist, überraschen uns weit entfernte Werte weniger als wenn σ klein ist.

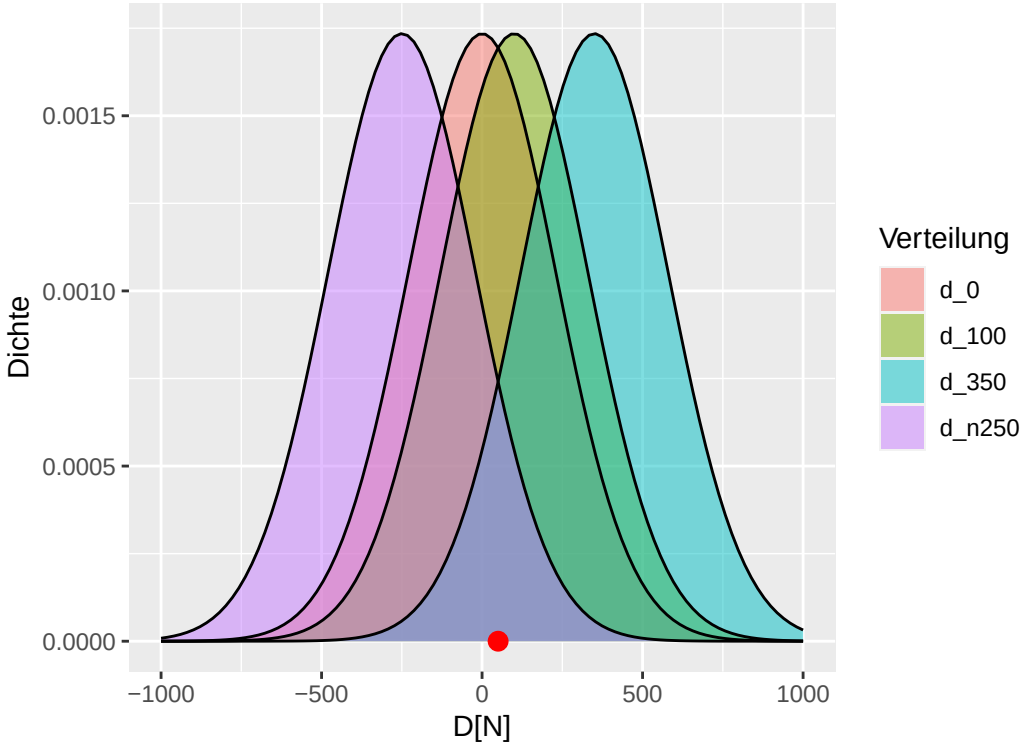


Figure 2.10: Welche Verteilung nehmen wir?

Spielen wir verschiedene Möglichkeiten einmal durch. Wir vernachlässigen zunächst einmal σ und konzentrieren uns auf μ. Wir benötigen eine einzelne Referenzverteilung um unseren beobachteten Wert Δ, den Unterschied zwischen den beiden Gruppen, mit der Verteilung in Beziehung zu setzen. Wir könnten zum Beispiel sagen, dass wir davon ausgehen, dass der Unterschied zwischen den beiden Gruppen $\Delta_{\text{wahr}} = 75N$ ist. D.h. dies wäre der *wahre* Unterschied zwischen den beiden Gruppen. Wir treffen ihn nicht genau, da wir eine Zufallsstichprobe gezogen haben und die Stichprobenvariabilität dazu führt, dass wir nicht genau den Unterschied treffen. Allerdings, wird wieder einmal etwas starren auf den Wert $75N$ zu der Einsicht führen, dass 75 vollkommen willkürlich ist. Warum nicht $85N$ oder 25 oder warum überhaupt ganzzahlig, π ist schließlich auch keine ganzzahlige Zahl, also könnten wir genauso gut $74.1234N$ nehmen. Schnell wird daher klar, dass keine Zahl so richtig gut begründet werden kann. Wir brauchen aber eine Zahl um unseren Apparatus mit Verteilungen ansetzen zu können. Tatsächlich gibt es eine Zahl die zwar auch willkürlich ist, aber doch etwas besser begründet werden kann, nämlich die Zahl $\Delta_{\text{wahr}} = 0$. Warum ist der Wert 0 in diesem Fall speziell. Nun, er bedeutet, dass wir davon ausgehen, dass zwischen den beiden Gruppen kein Unterschied besteht, also die Intervention überhaupt nichts gebracht hat. Dies ist zwar keine wirklich interessante Annahme, aber sie hat trotz ihr Willkürlichkeit doch etwas mehr Gewicht als eine beliebige andere Zahl. Wir bezeichnen diese Annahme jetzt auch noch als die H_0-Hypothese. Die 0 bei H bedeutet dabei nicht unbedingt, dass die H_0 davon ausgeht, dass nicht passiert, sondern nur, das das unsere Ausgangsannahme ist. In vielen Fällen hat die H_0 tatsächlich auch die Annahem das nichts passiert, dies muss aber nicht immer der Fall sein. Daher ist unsere Referenzverteilung für die Stichproben in unseren Fall die Hypothese (siehe Formel (2.2)):

$$H_0 : \Delta = 0 \tag{2.2}$$

oder graphisch (siehe Figure 2.11)

Diese Referenzverteilung können wir nun verwenden um eine Entscheidung bezüglich unseres beobachteten Werts zu treffen. Die Streuung in der Referenz- bzw. Stichprobenverteilung wird als Standardfehler bezeichnet im Gegensatz zur Streuung in der Population σ und in der Stichprobe s.

Definition 2.2.

Standardfehler

{Standardfehler}

Die Streuung der Stichprobenverteilung wird als Standardfehler σ_e bezeichnet. Wir dieser Wert anhand der Stichprobe abgeschätzt hat der Standardfehler das Symbol s_e.

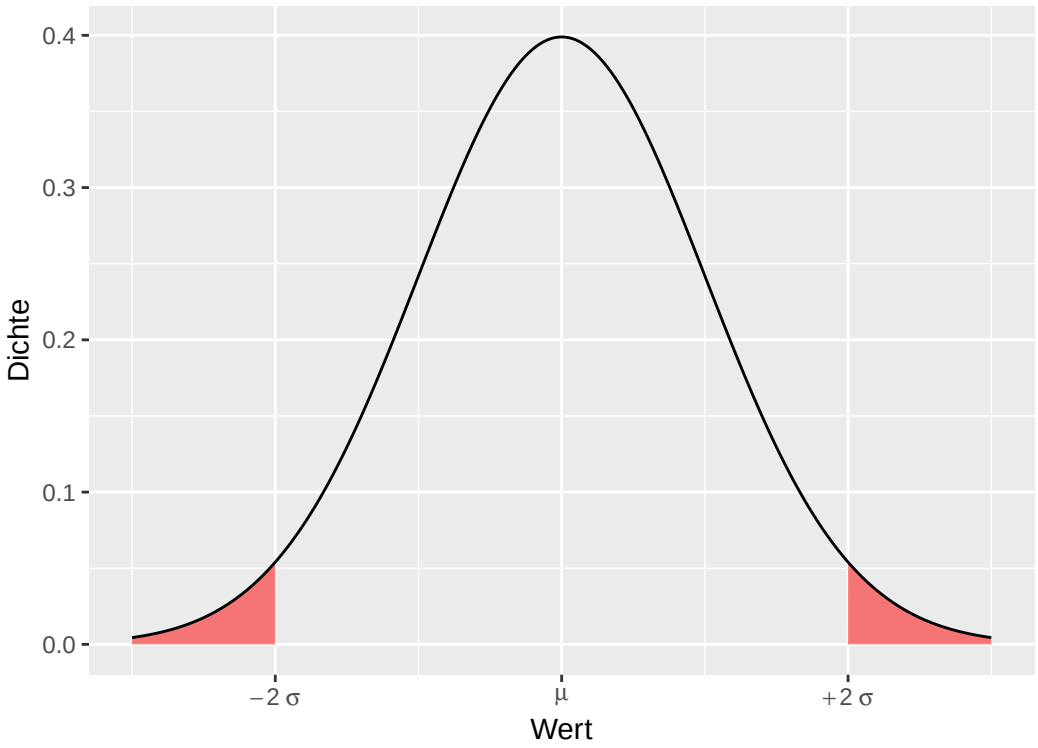


Figure 2.11: Verteilung wenn nichts passiert mit den beiden Bereichen jenseits von zwei Standardfehlern ausgezeichnet.

2.5 Statistisch *signifikanter* Wert

Kommen wir nun zu dem wichtigen Konzept des statistisch signifikanten Werts. Im vorhergehend Abschnitt haben wir eine Stichprobenverteilung für unsere Statistik, den Unterschied zwischen den Mittelwerten der beiden Gruppen, hergeleitet. Wir gehen von der Verteilung aus, bei der es keinen Unterschied $H_0 : \Delta = 0$ zwischen den beiden Gruppen gibt. $\Delta = 0$ hat somit die Bedeutung, das das Krafttraining nicht effektiv war. Dazu haben wir als Kriterium hergeleitet, dass wir Werte die mehr als 2 Standardabweichungen von Mittelwert entfernt sind, als unwahrscheinlich ansehen, da diese Werte etwa eine Wahrscheinlichkeit von 5% haben. Präziser, Werte die mehr als zwei Standardfehler vom Mittelwert entfernt sind. Da, unserer angenommenere Mittelwert mit $\Delta = 0$ zu $\mu = 0$ wird, bedeutet dies, das wir Werte die entweder kleiner als $-2\times$ Standardfehler oder größer als $2\times$ Standardfehler sind, als unwahrscheinlich unter der Annahme von $H_0 : \mu = 0$ betrachten. Als Entscheidungsregel:

$$|\text{beobachteter Wert }| > 2 \times \sigma_e \Rightarrow \text{ Evidenz gegen } H_0$$

In Figure 2.13 ist die Entscheidungsregel noch einmal graphisch anhand der Verteilung dargestellt. Wir haben unsere Stichprobenverteilung unter der $H_0 : \mu = \Delta = 0$ und schneiden rechts und links jeweils einen Bereich der Verteilung ab. Diesen Bereich bezeichnen wir als kritischen Bereich. Wenn unser beobachteter Wert im kritischen Bereich liegt, sehen wir dies als Evidenz gegen die Korrektheit der Annahme H_0 an.

Wenn der Stichprobenwert der Statistik in der *kritischen* Region auftritt, dann wird von einem **statistisch** signifikanten Effekt gesprochen. *Unter der H_0 bin ich überrascht diesen Wert zu sehen!* Allerdings, dieser Wert ist **nicht unmöglich**, sondern lediglich unwahrscheinlich unter der Annahme H_0. Unwahrscheinlich ist dabei kein absolutes Maß, sondern nur eine **willkürliche** Festsetzung die wir getroffen haben.

Wir hatten vorhin vorhin gesagt, dass Werte jenseits von $2 \times \sigma_e$ etwa eine Wahrscheinlichkeit von 5% haben. Dies Bedeutet nun, dass die Wahrscheinlichkeit Werte aus dem kritischen Bereich zu beobachten bei etwas 5 liegt, wenn die H_0 zutrifft. Oder anders, wenn die H_0 in der Realität zutrifft, also den DGP korrekt beschreibt, und ich das Experiment $100\times$ wiederhole, dann würde ich etwa 5 Experimente erwarten bei denen der beobachtete Wert im kritischen Bereich liegt. werde ich mich in 5 Fällen, irrtümlich gegen die H_0 entscheiden, obwohl sie korrekt ist. D.h. in 5 Fällen würde mich irren, da ich einen Wert im kritischen Bereich beobachtet habe, trotzdem die H_0 zutrifft. Daher wird die Wahrscheinlichkeit die ich benutze um einen kritschen Bereich ausweisen als Irrtumswahrscheinlichkeit bezeichnet. Da die Irrtumswahrscheinlichkeit ein zentrales Konzept in der Statistik ist, erhält sie auch ein eigenes Symbol α.

Definition 2.3 (Irrtumswahrscheinlichkeit α)**.** Die Wahrscheinlichkeit mit der fälschlicherweise eine korrekte H_0-Hypothese abgelehnt wird, wird als Irrtumswahrscheinlichkeit bezeich-

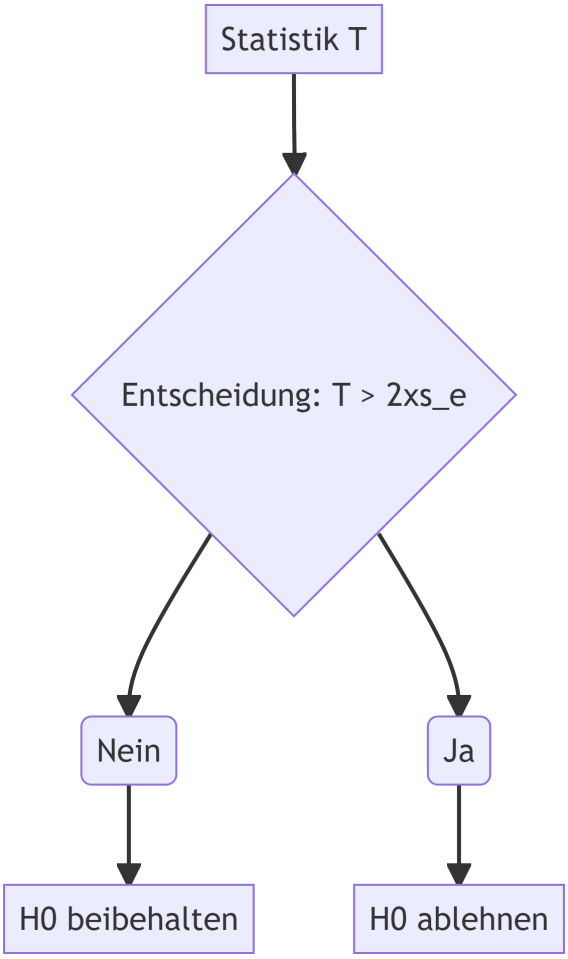


Figure 2.12: Entscheidungsregel zur H_0

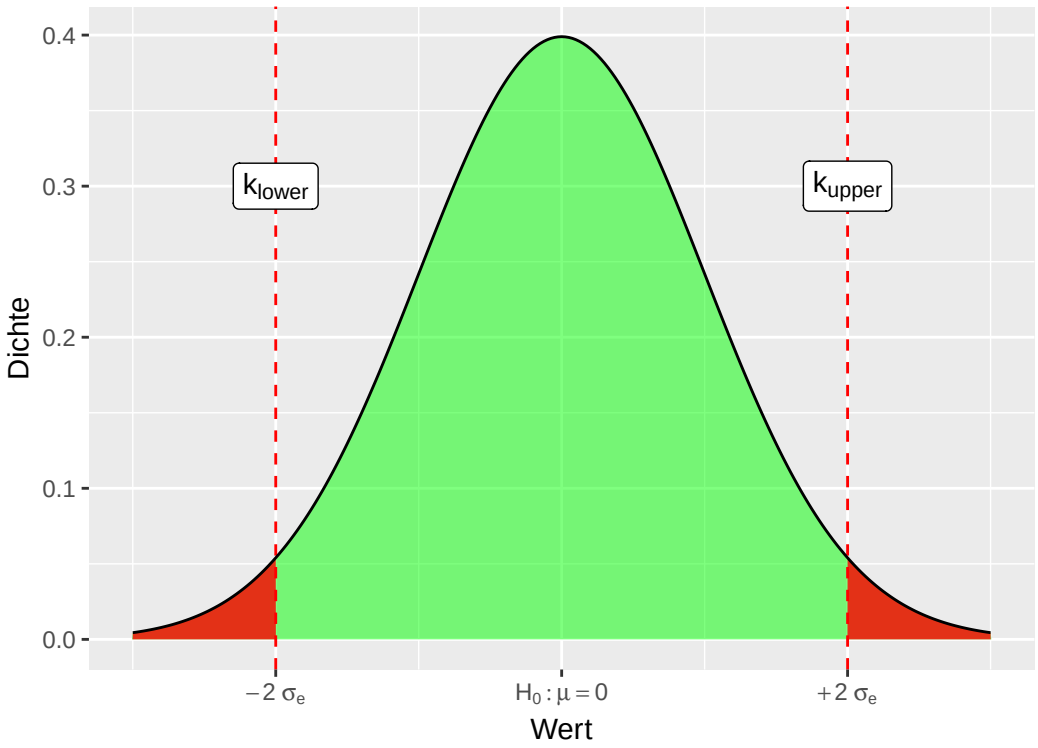


Figure 2.13: Die H_0 Verteilung wenn nichts passiert unterteilt in Regionen die zur Entscheidung für die H_0 (grün) und gegen die H_0 (rot, kritische Regionen) führen.

net. Die Irrtumswahrscheinlichkeit wird mit Symbol α bezeichnet und auch als Fehler I. Art bezeichnet.

Eines der grundlegenden Probleme, das oftmals nicht beachtet wird bei der Interpretation von **statistisch** signifikanten Ergebnis bezieht sich darauf, dass ich nicht weiß, welches der 100 Experimente ich durchgeführt habe. Zusätzlich, dies ist keine Aussage über die Wahrscheinlichkeit mit der die H_0 in der Realität zutrifft. Ob die H_0 zutrifft hat die Wahrscheinlichkeit entweder $P(H_0) = 1$ oder $P(H_0) = 0$. Entweder sie trifft zu oder eben nicht. Darüber wird hier keine Aussage gemacht, sondern nur ob unter der Annahme das H_0 zutrifft, der beobachtete Wert in einem *wahrscheinlichen* oder einem *unwahrscheinlichen* Bereich liegt. Und nochmal, *wahrscheinlich* war eine willkürliche Festlegung unsererseits.

2.6 Der p-Wert

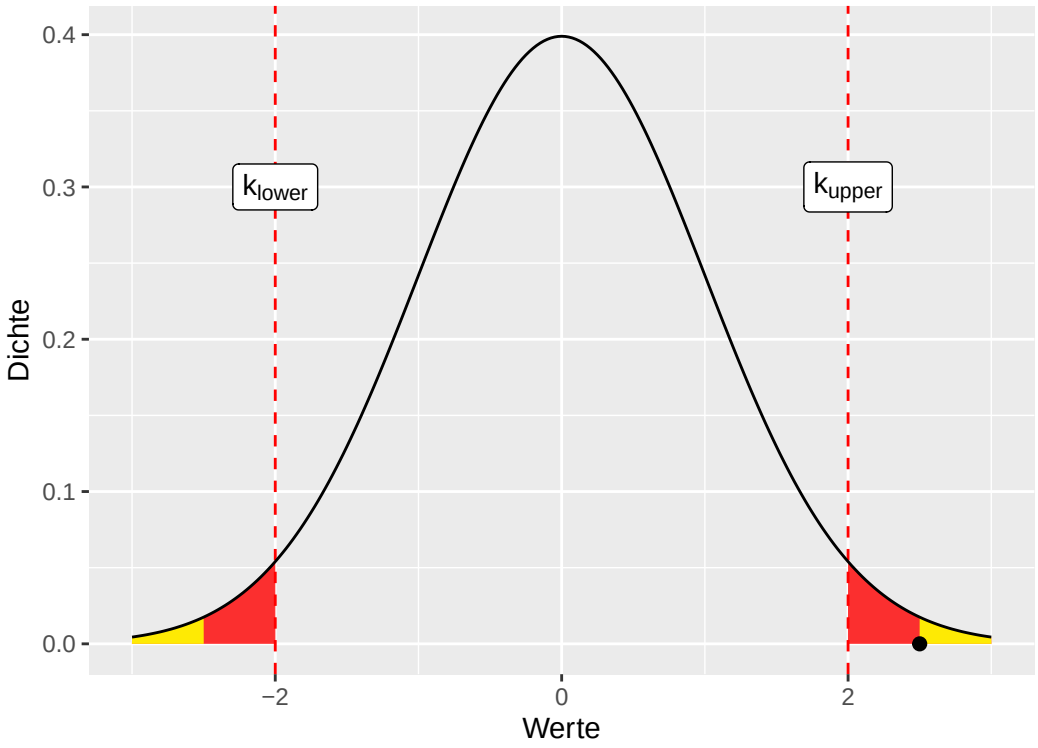


Figure 2.14: Der gelben Flächen zeigen den p-Wert für den Wert der Statistik von d = 2,5 an.

Der p-Wert gibt die Wahrscheinlichkeit für den gefundenen oder einen noch extremeren Wert unter der H_0 an.

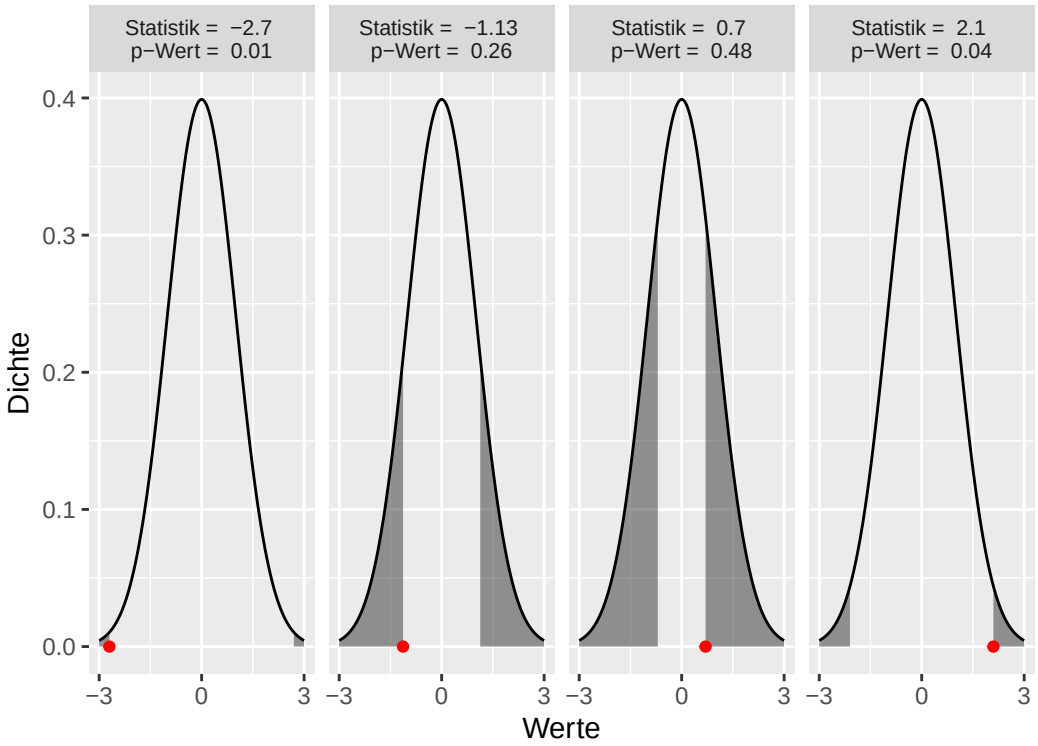


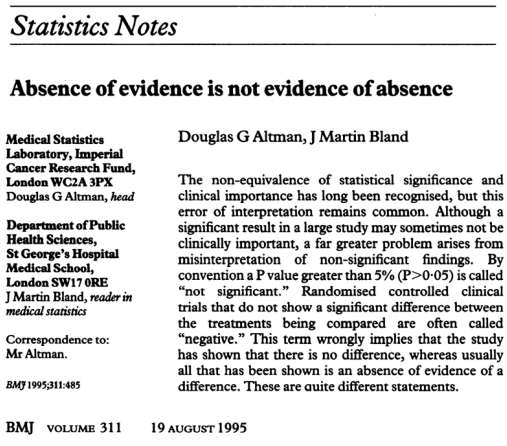
Figure 2.15: Verschiedene P-Werte

"[A] p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." (Wasserstein and Lazar 2016, 131)

"[T]he P value is the probability of seeing data that are as weird or more weird than those that were actually observed." (Christensen 2018, 38)

2.6.1 Signifikanter Wert - Das Kleingedruckte

- **Vor** dem Experiment wird für ein H_0 ein α-Level angesetzt (per Konvention $\alpha = 0,05 = 5\%$)
- Anhand des α-Levels können **kritische Werte** (k_{lower}, k_{upper}) bestimmt werden. Diese bestimmen die Grenzen der **kritischen Regionen**.
- Wenn der gemessene Wert w der Statistik in die kritische Region fällt, also $w \leq k_{lower}$ oder $w \geq k_{upper}$ gilt, dann wird von einem **statistisch** signifikanten Wert gesprochen und die dazugehörige Hypothese wird **abgelehnt**. Äquivalent: Der p-Wert ist kleiner als α.
- Da in α-Fällen ein Wert in der kritischen Region auftritt, auch wenn die H_0 zutrifft, wird in α-Fällen ein α-Fehler gemacht.
- Wenn der Wert w der Statistik nicht in den kritischen Regionen liegt, oder gleichwertig der p-Wert größer als α ist, wird die H_0 **beibehalten**. D.h. nicht, dass **kein Effekt** vorliegt, sondern lediglich, dass anhand der Daten keine Evidenz diesbezüglich gefunden werden konnte!
- Die **statistische** Signifikanz sagt nichts über die Wahrscheinlichkeit der Theorie aus!
- Ein p-Wert von $p = 0.0001$ heißt nicht, dass mit 99,99% Wahrscheinlichkeit ein Effekt vorliegt!
- *Statistisch* signifikant heißt nicht automatisch *praktisch* relevant!



Statistics Notes

Absence of evidence is not evidence of absence

Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX
Douglas G Altman, *head*

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE
J Martin Bland, *reader in medical statistics*

Correspondence to:
Mr Altman.

BMJ 1995;311:485

Douglas G Altman, J Martin Bland

The non-equivalence of statistical significance and clinical importance has long been recognised, but this error of interpretation remains common. Although a significant result in a large study may sometimes not be clinically important, a far greater problem arises from misinterpretation of non-significant findings. By convention a P value greater than 5% (P>0·05) is called "not significant." Randomised controlled clinical trials that do not show a significant difference between the treatments being compared are often called "negative." This term wrongly implies that the study has shown that there is no difference, whereas usually all that has been shown is an absence of evidence of a difference. These are quite different statements.

BMJ VOLUME 311 19 AUGUST 1995

Figure 2.16: Ausschnitt aus D. G. Altman and Bland (1995)

Eine weitere Erklärung für den p-Wert nach Wasserstein and Lazar (2016)

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

2.7 Was passiert nun aber wenn die "andere" Hypothese zutrifft?

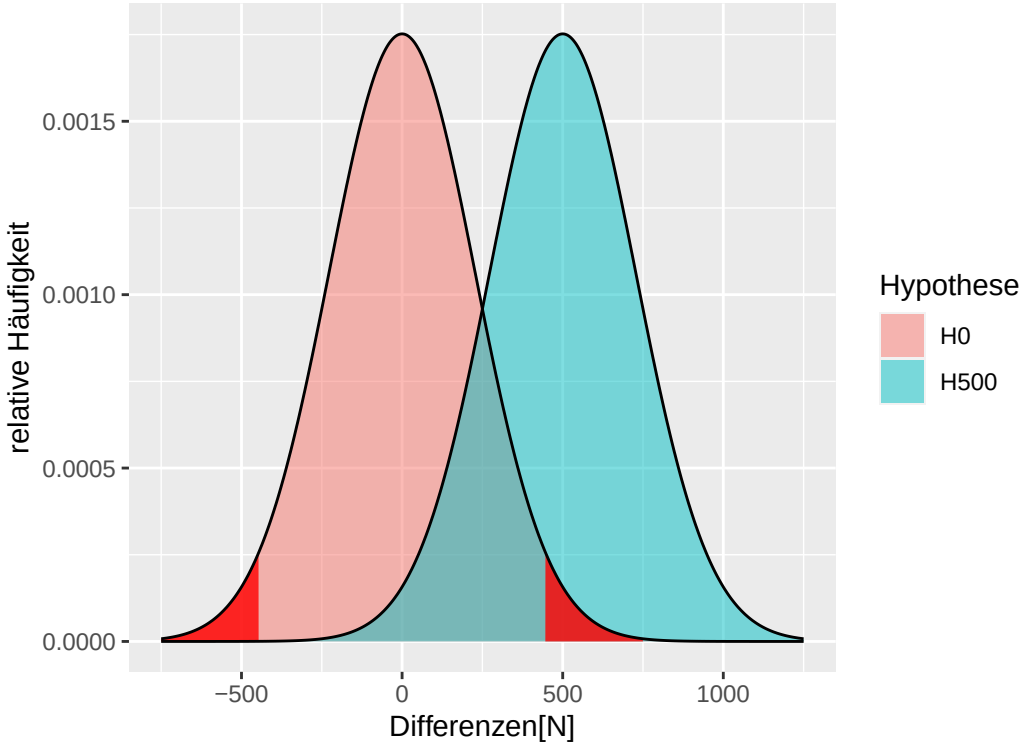


Figure 2.17: Differenzen mit kritischen Regionen (rot) mit einer Wahrscheinlichkeit von α wenn H_0 zutrifft.

2.8 Wir machen einen β-Fehler!

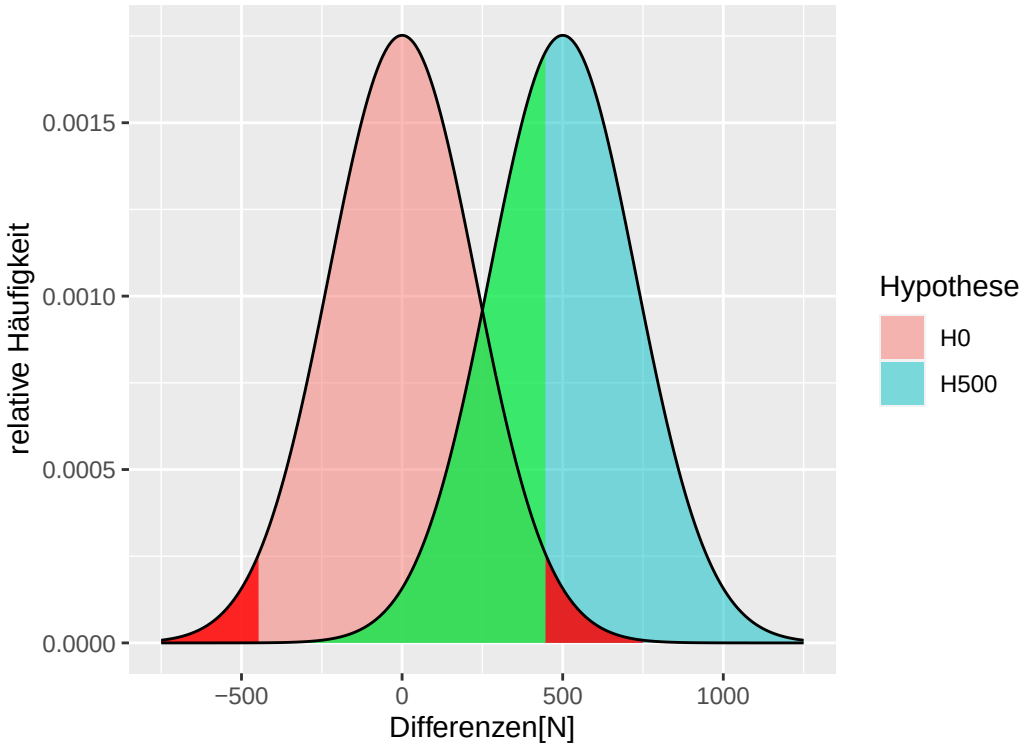


Figure 2.18: Differenzen mit kritischen Regionen (rot) mit einer Wahrscheinlichkeit von α wenn H_0 zutrifft und β (grün) wenn H_1 zutrifft.

2.9 Snap!(1989) - The Power

2.10 Terminologie noch mal

- α: Die Wahrscheinlichkeit sich gegen die H_0 zu entscheiden, wenn die H_0 zutrifft. α-Level wird vor dem Experiment festgelegt um zu kontrollieren welche Fehlerrate toleriert wird.
- β: Die Wahrscheinlichkeit sich gegen die H_1 zu entscheiden, wenn die H_1 zutrifft.
- Power $:= 1 - \beta$: Die Wahrscheinlichkeit sich für die H_1 zu entscheiden, wenn die H_1 zutrifft. Sollte ebenfalls **vor** dem Experiment festgelegt werden.

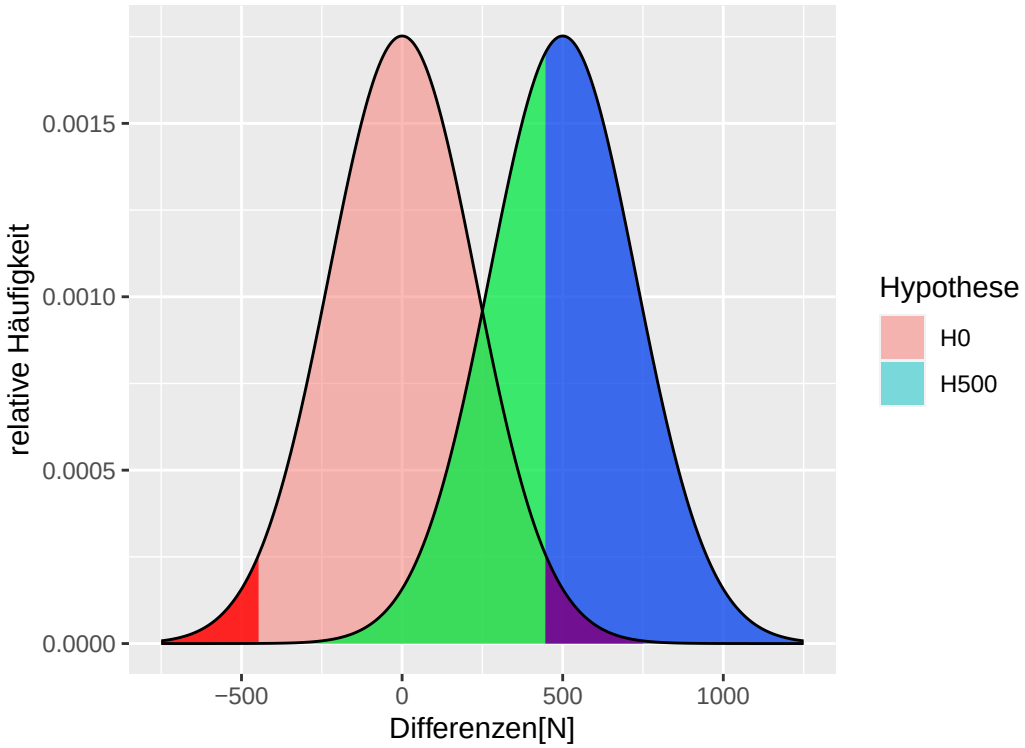


Figure 2.19: $1 - \beta$ = Power des Tests (blaue Fläche).

2.11 Wie können wir die Power erhöhen?

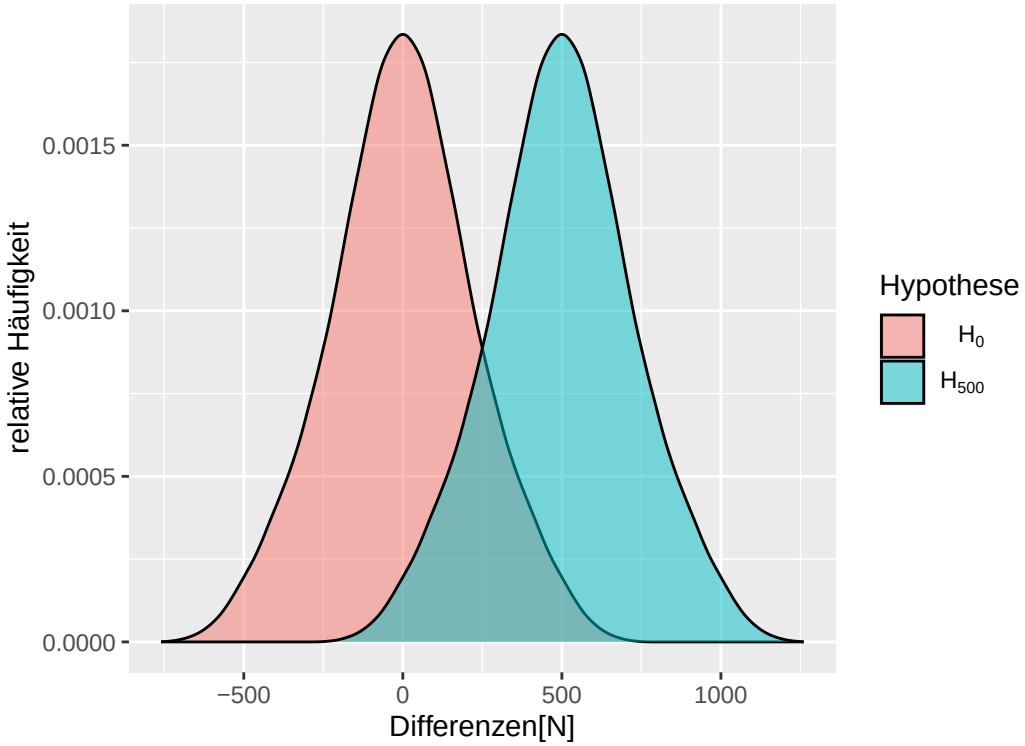


Figure 2.20: Verteilungen wenn δ=500 und δ=0 in unserem kleine Welt Beispiel mit n = 3.

2.12 Stichprobengröße von n = 3 auf n = 9 erhöhen?

2.13 Standardfehler

Die Standardabweichung der Stichprobenverteilung wird als **Standardfehler** s_e bezeichnet[1]. Der Standardfehler ist nicht gleich der Standardabweichung in der Population bzw. der Stichprobe. Es gilt für den Mittelwert:

[1]Der Standardfehler schätzt die Reliabilität der Statistik ab (Cohen (1988))

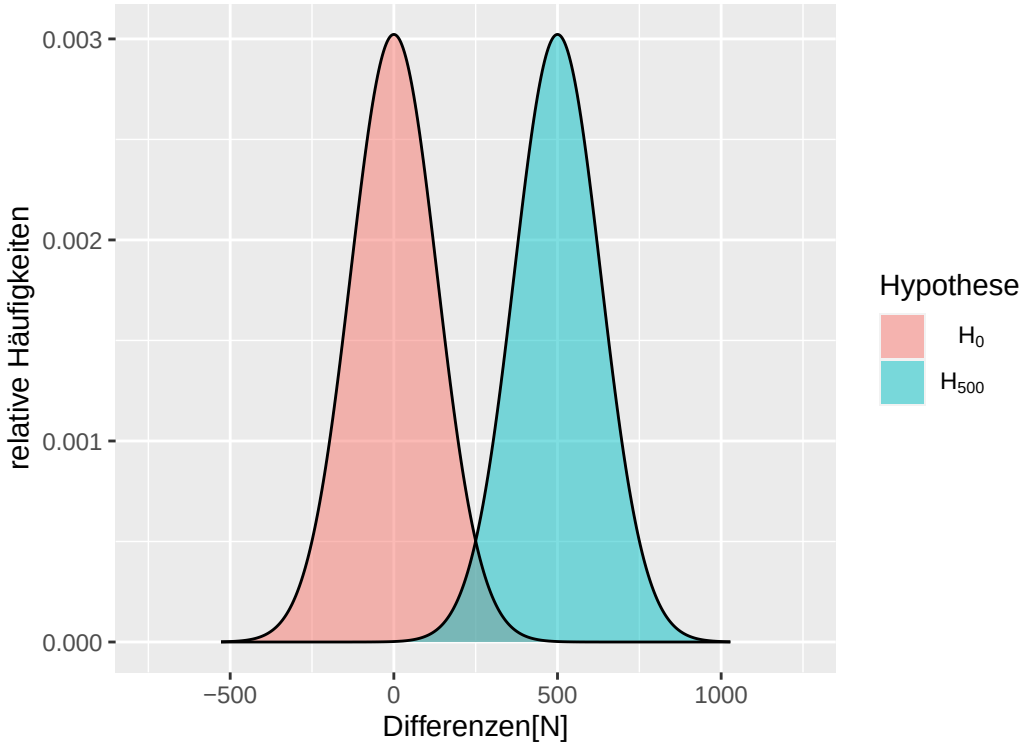


Figure 2.21: Stichprobenverteilungen der Differenz unter H_0 und $H_1 : \delta = 500$N bei einer Stichprobengröße von n = 9

Table 2.2: Standardfehler des Mittelwerts, n = Stichprobengröße

Population	Stichprobe
$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$	$s_e = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$

3 Parameterschätzung

3.1 Problem bei einer dichotomen Betrachtung der Daten

Only two studies have evaluated the therapeutic effectiveness of a new treatment for insomnia. Lucky (2008) used two independent groups each of size $N = 22$, and Noluck (2008) used two groups each with $N = 18$. Each study reported the difference between the means for the new treatment and the current treatment.

Lucky (2008) found that the new treatment showed a statistically significant advantage over the current treatment: *M(difference)* = 3.61, *SD(difference)* = 6.97, $t(42) = 2.43$, $p = .02$. The study by Noluck (2008) found no statistically significant difference between the two treatment means: *M(difference)* = 2.23, *SD(difference)* = 7.59, $t(34) = 1.25$, $p = .22$.

Figure 3.1: Auszug aus Cumming (2013, 1)

3.2 Wie groß ist der Effekt?

3.3 Schätzung der Populationsparameter

Kleine Welt: Experiment wird einmal mit n = 9 durchgeführt

3.3.1 Beobachtete Stichprobenkennwerte

$$\begin{aligned} d = \bar{x}_{treat} - \bar{x}_{con} &= 350 \\ s &= 132 \\ s_e &= 44 \end{aligned}$$

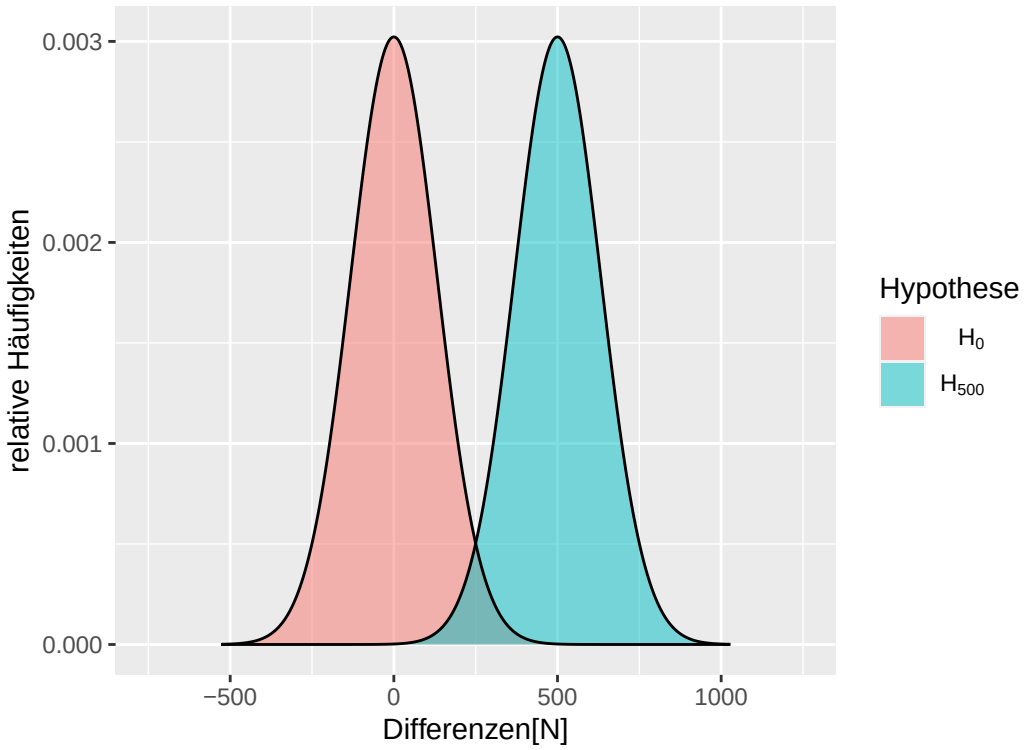


Figure 3.2: Stichprobenverteilungen der Differenz unter H_0 und $H_1 : \delta = 500$N bei einer Stichprobengröße von n = 9

Wie präzise ist meine Schätzung und welche anderen Unterschiedswerte sind anhand der beobachteten Daten noch plausibel?

3.4 Welche δs sind plausibel für $d = 350$?

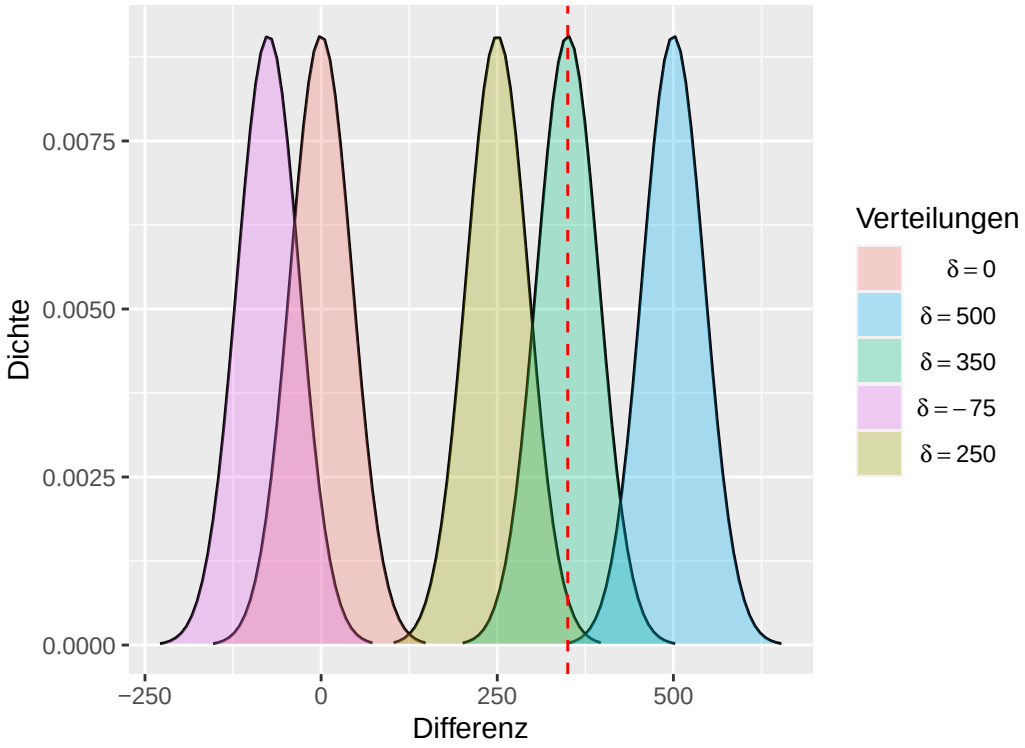


Figure 3.3: Verschiedene Verteilungen von Gruppendifferenzen, beobachteter Unterschied (rot)

Plausibel unter einem gegebenem α**-Level**!

3.5 Alle möglichen δs die plausibel sind

3.6 Was passiert wenn ich das Experiment ganz oft wiederhole?

3.7 Konfidenzintervall - Das Kleingedruckte

- Das Konfidenzintervall für ein gegebenes α-Niveau gibt nicht die Wahrscheinlichkeit an mit der der *wahre* Parameter in dem Intervall liegt.

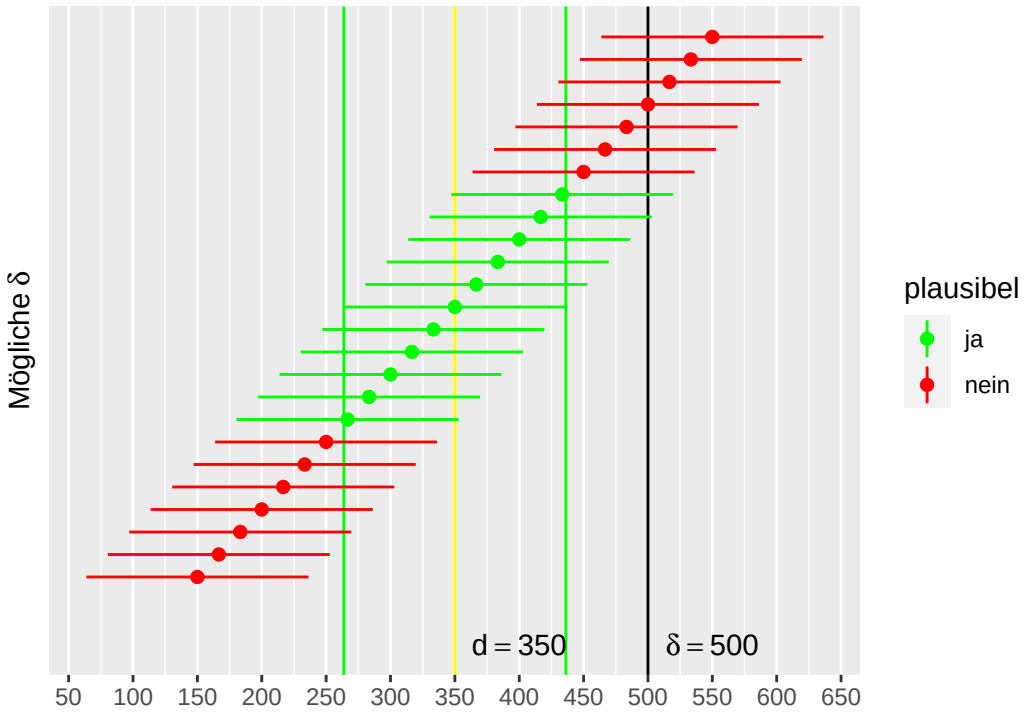


Figure 3.4: Konfidenzintervall (grün), Populationsparameter δ und α-Level für die beobachtete Differenz (gelb).

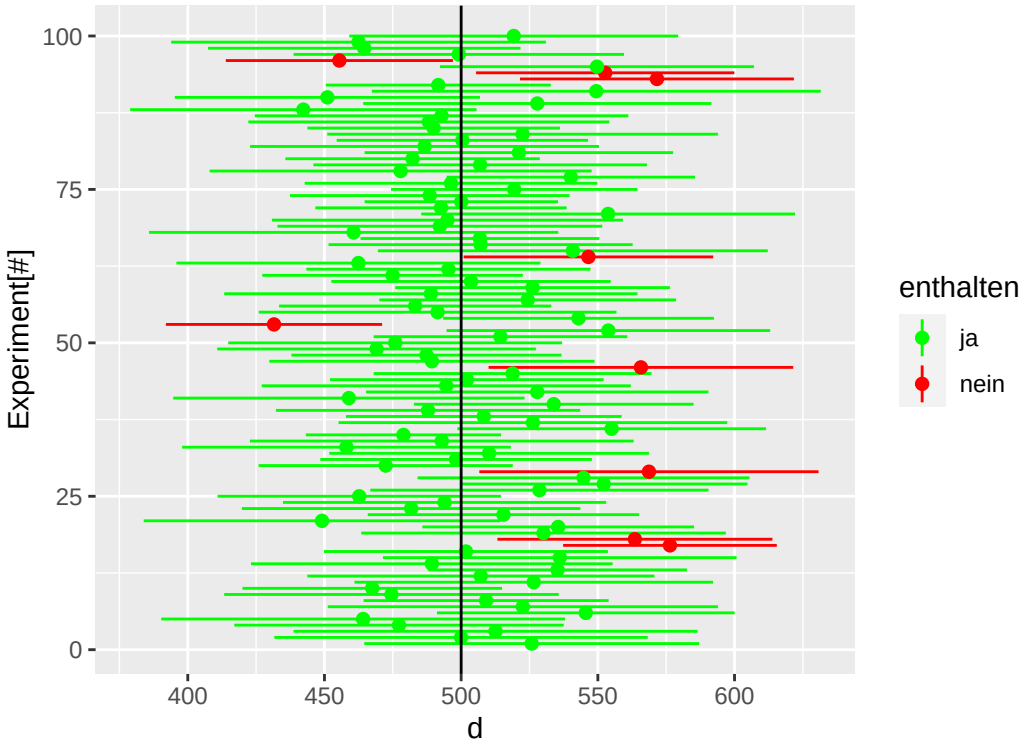


Figure 3.5: Simulation von $n = 100$ Konfidenzintervallen.

- Das Konfidenzintervall gibt alle mit den Daten kompatiblen Populationsparameter an.
- Das α-Niveau des Konfidenzintervalls gibt an bei welchem Anteil von Wiederholungen davon auszugehen ist, das das Konfidenzintervall den wahren Populationsparameter enthält.

3.8 Konfidenzintervall herleiten nach Spiegelhalter (2019, 241)

1. We use probability theory to tell us, for any particular population parameter, an interval in which we expect the observed statistic to lie with 95% probability.
2. Then we observe a particular statistic.
3. Finally (and this is the difficult bit) we work out the range of possible population parameters for which our statistic lies in their 95% intervals. This we call a "95% confidence interval".
4. This resulting confidence interval is given the label "95%" since, with repeated application, 95% of such intervals should contain the true value.[1]

All clear? If it isn't, then please be reassured that you have joined generations of baffled students.

3.9 Konfidenzintervall berechnen (Vorschau)

$$\text{CI}_{1-\alpha} = \bar{x} \pm z_{\alpha/2} \times s_e$$

3.10 Dualität von Signifikanztests und Konfidenzintervall

Wenn das Konfidenzintervall mit Niveau $1 - \alpha\%$ die H_0 nicht beinhaltet, dann wird auch bei einem Signifikanztest die H_0 bei einer Irrtumswahrscheinlichkeit von α abgelehnt.

[1] Strictly speaking, a 95% confidence interval does ***not*** mean there is a 95% probability that this particular interval contains the true value [...]

4 Verteilungen

4.1 Die Verteilung - 1. deep dive

Wir versuchen jetzt als erstes zu Verstehen was nochmal genau der Graph der Verteilung bedeutet. Auf der x-Achse werden die verschiedenen möglichen Werte der jeweiligen Statistik abgebildet. In unserem bisherigen Beispiel was das die Unterschiede D zwischen der Kontroll- und der Treatmentgruppe. Der Wert auf der y-Achse was zunächst die relative Häufigkeit was auch Sinn gemacht hatte, da wir nur eine bestimmte endliche Anzahl von möglichen Unterschieden D (ihr erinnert auch an die Zahl) vorliegen hatten. Was passiert aber wenn wir tatsächlich eine kontiuierliche Statistik haben, also eine Statistik die alle Werte innerhalb eines Intervalls einnehmen kann. Um den Fall zu verstehen fangen wir aber erst mal wieder mit einem einfachen Modell an.

4.1.1 Der Münzwurf

Wir fangen mit dem einfachsten Experiment an: dem Münzwurf. Beim Münzwurf haben wir zwei mögliche Ausgänge unseres Experiments, entweder Kopf oder Zahl. Wir gehen von einer perfekten Münze aus, d.h. die Münze ist vollkommen symmetrisch auf beiden System und keine der Seiten ist in irgendeiner Form schwere oder beeinflusst in einer Art den Ausgang.

Wenn wir uns an die Schule zurück erinnern, dann haben wir in Wahrscheinlichkeitstheorie schon mal was gehört, das im Fall gleichwahrscheinlicher Ereignisse die Wahrscheinlichkeit für ein bestimmtes Ereignis, mittels der Anzahl der vorteilhaften Ausgänge geteilt durch die Anzahl der möglichen Ausgänge berechnet wird. Also beim einmaligen Münzwurf haben wir zwei Ausgänge $\{\text{Kopf}, \text{Zahl}\}$ und jeweils nur vorteilhaften Ausang als entweder Kopf oder Zahl, daher folgt daraus.

$$P(\text{Kopf}) = \frac{1}{2} \tag{4.1}$$

$$P(\text{Zahl}) = \frac{1}{2} \tag{4.2}$$

Wenn wir das jetzt als Graphen in Form einer Wahrscheinlichkeitsverteilung abtragen, dann sieht das noch wenig interessant aus (siehe Figure 4.1). Das Muster ist aber trotzdem wichtig,

damit wir später wissen worauf wir hier eigentlich schauen. Auf der x-Achse haben wir die möglichen Ausgänge, Kopf oder Zahl, und auf der y-Achse haben wir die Wahrscheinlichkeit abgetragen.

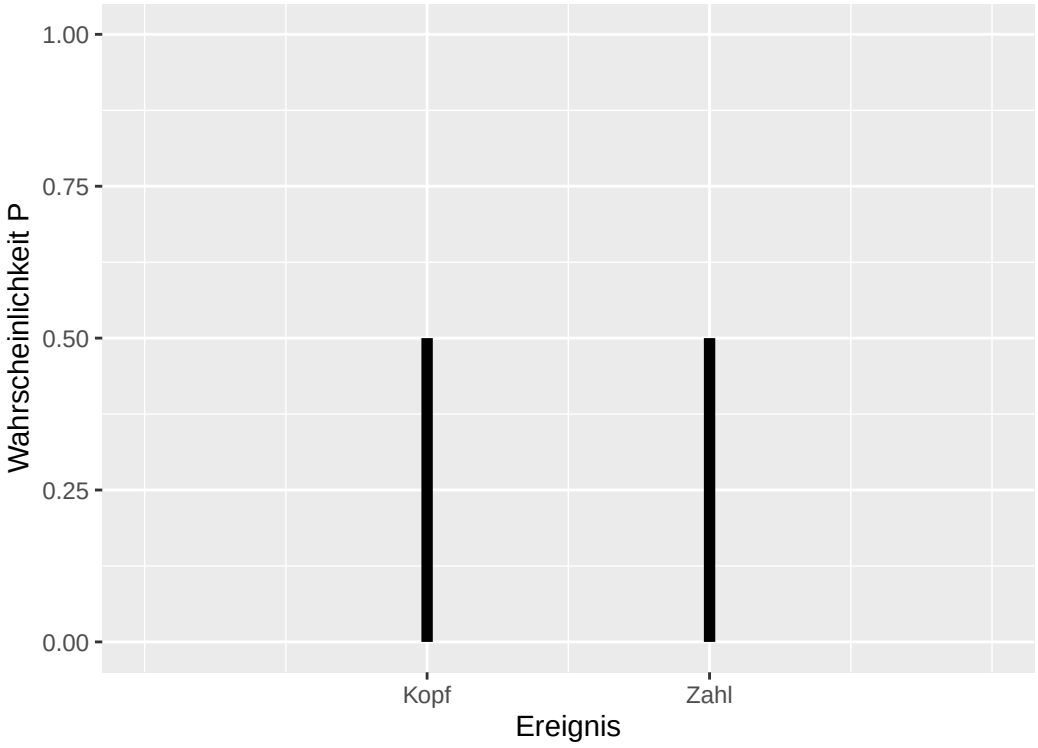


Figure 4.1: Wahrscheinlichkeitsverteilung des einmaligen Münzwurfes

Da sich mit einem Münzwurf aber so wenig anfangen lässt, machen wir das Ganze jetzt etwas komplizierter und schauen uns an, wie unser Experiment aussieht wenn wir zwei Münzwwürfe uns anschauen. Rein operational, wir schmeißen unsere Münze in die Luft, schreiben uns das Ergebnis auf, und machen das Ganze noch ein zweites Mal und schreiben uns das Ergebnis auf. D.h. was auch immer im ersten Durchgang passiert, hat keine Auswirkungen auf das Ergebnis des zweiten Wurfs. Wir könnten auch zwei Münzen nehmen und beide gleichzeitig in die Luft werfen. Das wäre das gleiche Experiment. Welche Ausgänge haben wir jetzt beim zweimaligen Münzwurf? Zunächst einmal haben wir jetzt nicht mehr nur einen einzelnen Ausgang sondern wir haben ein Ausgangstupel, eine Liste mit zwei Elementen. Etwas motiviertes krizteln auf einem Schmierblatt wird wahrscheinlich relativ schnell zu folgender Tabelle führen (siehe Table 4.1)

Table 4.1: Mögliche Ausgänge bei einem zweimaligen Münzwurf

Ausgang 1. Wurf	Ausgang 2. Wurf	Tupel
Kopf	Kopf	(Kopf, Kopf)
Kopf	Zahl	(Kopf, Zahl)
Zahl	Kopf	(Zahl, Kopf)
Zahl	Zahl	(Zahl, Zahl)

Jetzt können wir uns wieder fragen, was die Wahrscheinlichkeit für die jeweiligen Ereignistupel ist. Eine direkte Methode wäre, wieder mittels der Symmetrie zu argumentieren. Es gibt vier verschiedene Ausgänge von denen jetzt keiner in irgendeiner Weise bevorzugt ist, daraus würde folgen das alle vier Ausgänge eine Wahrscheinlichkeit von $P = \frac{1}{4}$ haben.

Eine weitere Möglichkeit wäre mit den Wahrscheinlichkeiten aus dem einfachen Wurf an das Problem heran zu gehen. Wir betrachten die beiden Münzwürfe jetzt wieder sequentiell (siehe **?@fig-sts-sig-coin-toss-tree**). Im ersten Schritt können wir entweder Kopf oder Zahl beobachten. Beide Wahrscheinlichkeiten sind $P = \frac{1}{2}$. Darauf folgend können wir wieder zwei verschiedene Ausgänge beobachten, eben Kopf oder Zahl, wieder mit der Wahrscheinlichkeit $P = \frac{1}{2}$.

Da die Münzwürfe voneinander unabhängig sind und keinen Einfluss aufeinander ausüben, folgt daraus, dass die Wahrscheinlichkeiten für jede spezielle Folge von Kopf oder Zahl sich berechnet nach:

$$P(\text{Ausgang}) = P(1.\ \text{Wurf}) \times P(2.\ \text{Wurf}) \tag{4.3}$$

Also in unseren Fall:

$$P(\text{Ausgang}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \tag{4.4}$$

Womit wir wieder beim gleichen Ergebnis wie vorher angekommen sind. Der Vorteil dieser Herangehensweise ist jedoch, dass wir damit eine einfache Möglichkeit gefunden haben das Ergebnis auf mehr als nur zwei Würfe zu verallgemeinern. Nehmen wir zum Beispiel den dreifachen Münzwurf, dann können wir die Wahrscheinlichkeit für die Folge $P(\text{KKZ}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$ direkt angeben.

Bleiben wir aber erst noch mal kurz beim zweimaligen Münzwurf und schauen uns die Wahrscheinlichkeitsverteilung an. Hier stoßen wir nämlich auf ein Problem in der Darstellung. Wenn wir bei dem Muster aus Figure 4.1 bleiben wollen und auf der x-Achse die möglichen Ergnisse und auf der y-Achse die dazugehörende Wahrscheinlichkeit abtragen wollen, dann ist nicht ganz klar wie wir die Ergebnisse ordnen sollen. Eine mögliche Lösung ist in Figure 4.2 zu sehen.

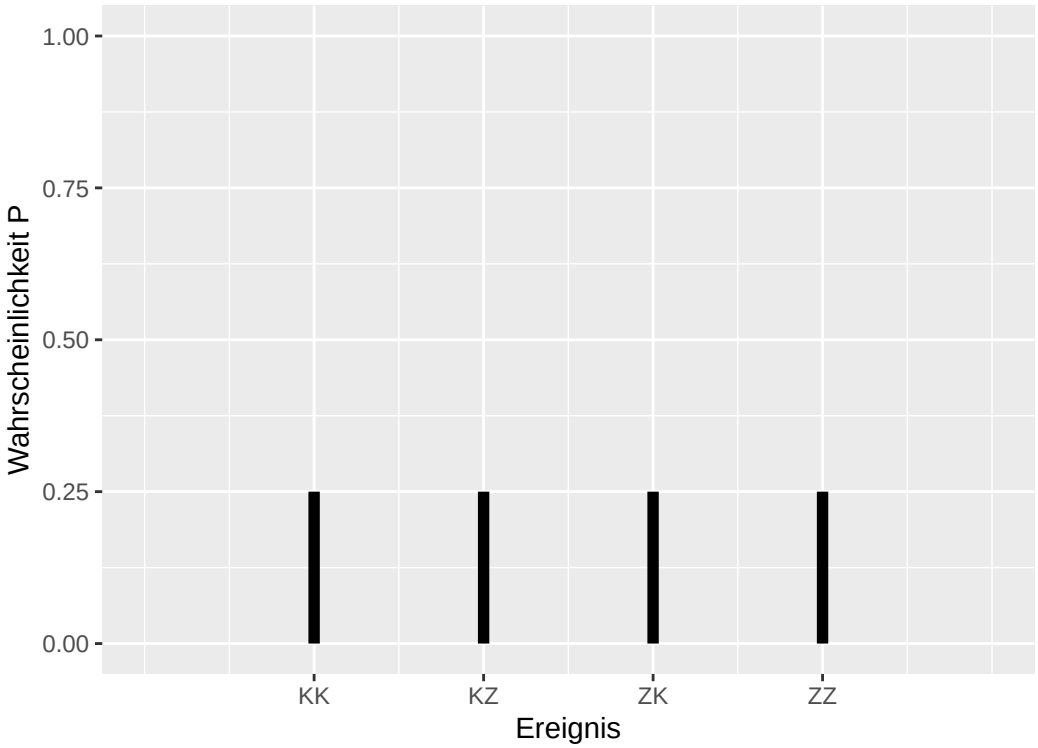


Figure 4.2: Wahrscheinlichkeitsverteilung des zweimaligen Münzwurfes (K: Kopf, Z: Zahl)

Dies ist natürlich nicht die einzige Möglichkeit wie wir die Ereignisse ordenen können sondern wahrscheinlich ist jede der 24 möglichen Anordnungen gleich sinnig. Wir könnten auch beispielsweise nicht mehr die beiden einzelnen Ausgänge als Ereignisse wählen, sondern könnten zum Beispiel nur noch die Anzahl der Köpfe in unseren zwei Würfen zählen. Dies würde zu der folgenden Zuordnung führen (siehe Table 4.2).

Table 4.2: Zuordnung der Anzahl der Köpfe zu den Ereignissen beim zweimaligen Münzwurf

Ereignisse	Anzahl der Köpfe
(Kopf, Kopf)	2
(Kopf, Zahl)	1
(Zahl, Kopf)	1
(Zahl, Zahl)	0

Wir verliegen bei dieser Zuordnung nachtürlich die Information bei welchem Wurf die Zahl beobachtet wurde, aber eigentlich interessiert uns das sowieso nicht so brennend. In der Terminologie der Wahrscheinlichkeitstheorie wird die Anzahl der Köpfe als Zufallsvariable bezeichnet.

Definition 4.1 (Zufallsvariable)**.** Eine Zufallsvariable ist die Abbildung eines Zufallsereignisses auf eine Zahl.

Anders dargestellt, ist eine Zufallsvariable eine Funktion, die einem Ereignis eine Zahl zuordnet (siehe **?@fig-sts-sig-random-variable**.

Wenn wir uns jetzt die Wahrscheinlichkeiten für unsere Zufallsvariable anschauen, dann sehen wir aber, dass wir nicht mehr vier verschiedne Ausgänge haben, sondern nur noch drei und das die gleiche Wahrscheinlichkeit für nicht gleich sind.

Table 4.3: Wahrscheinlichkeitstabelle für Zufallsvariable “Anzahl der Köpfe beim zweimaligen Münzwurf”.

Ereignisse	Zufallsvariale	Wahrscheinlichkeit
(Zahl, Zahl)	Keine Köpfe	$\frac{1}{4}$
(Kopf, Zahl)(Zahl,Kopf)	1 Kopf	$\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
(Kopf,Kopf)	2 Köpfe	$\frac{1}{4}$

Jetzt können wir wieder eine Wahrscheinlichkeitsverteilung für unsere Zufallsvariable abtragen (siehe Figure 4.3).

Nur um nebenbei noch einmal das offensichtliche Anzusprechen. Die Summe aller Wahrscheinlichkeiten aller Ereignisse muss 1 sein. Das sollte auch direkt einsichtig sein. Wenn ich alle

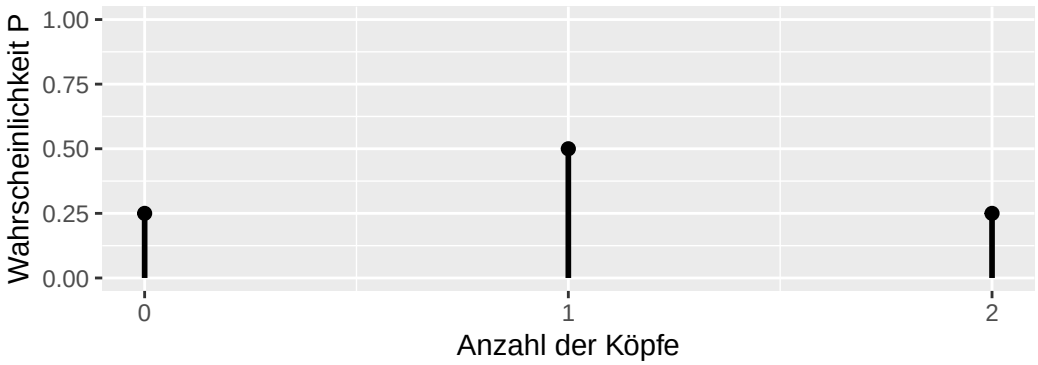


Figure 4.3: Wahrscheinlichkeitsverteilung für die Anzahl der Köpfe beim zweimaligen Münzwurf

möglichen Ereignisse abfrage also: "*Was ist die Wahrscheinlichkeit das ich keine Köpfe, 1 Kopf oder 2 Köpfe beim zweimaligen Münzwurf erhalte*", dann sind das alle möglichen Ausgänge und dementsprechend sollte die Wahrscheinlichkeit dafür "1" sein oder mathematisch ausgedrückt:

$$P(0 \text{ Köpfe} \cup 1 \text{ Kopf} \cup 2 \text{ Köpfe}) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

Jetzt gehen wir zum nächst komplizierteren Fall. Die Anzahl der Köpfe bei drei Münzwürfen. Welche Möglichkeiten gibt es hier? Nun bei drei Würfen kann entweder $0, 1, 2$ oder 3 Kopf auftreten. Wenn wir die Wahrscheinlichkeiten für diese vier Ereignisse berechnen wollen, können wir aber nicht einfache $\frac{1}{4}$ für jedes Ereignis als Wahrscheinlichkeit ansetzen (Warum?). Schauen wir uns erst einmal wieder die möglichen Tupel, oder auch die Elemenarereignisse, den wir erinnern uns, dass die Anzahl der Köpfe eine Zufallsvariable ist. Also eine Abbildung der 3-fach Tupel auf eine der Zahlen $\{0, 1, 2, 3\}$.

Table 4.4: Abbildung der 3-fach Tupel auf die Anzahl Kopf beim dreifachen Münzwurf

Elementarereignis	Anzahl Kopf
(Z,Z,Z)	0
(K,Z,Z)	1
(Z,K,Z)	1
(Z,Z,K)	1
(K,K,Z)	2
(Z,K,K)	2
(K,Z,K)	2
(K,K,K)	3

Die Elementarereignisse in Table 4.4 sind wieder alle gleichwahrscheinlich, daher können wir jetzt wieder einfache abzählen. Es gibt insgesamt 8 mögliche Ausgänge, davon haben jeiweils einer 0-mal oder 3-mal Kopf und jeweils 3 Ausgänge haben 1-mal oder 2-mal Kopf. Daraus folgt für die Wahrscheinlichkeitsfunktion (siehe Table 4.5).

Table 4.5: Wahrscheinlichkeitsfuntion für den dreifachen Münzwurf

Anzahl Kopf	P
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

Das Ganze auch wieder als Graph (siehe **?@fig-sts-coin-toss-3**)

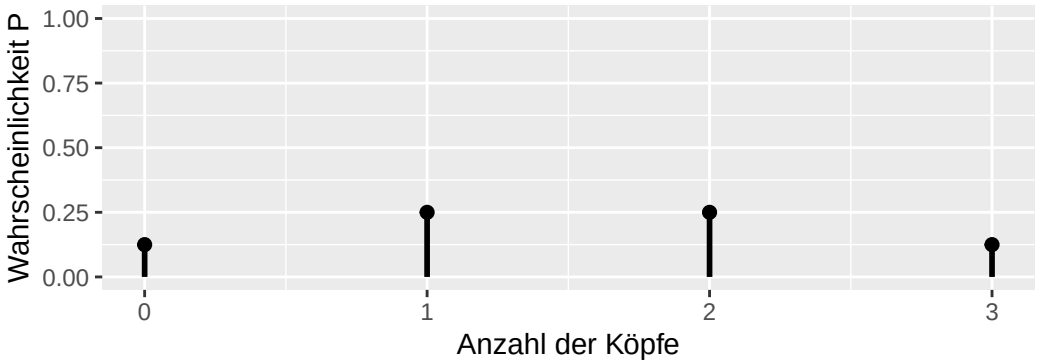


Figure 4.4: Wahrscheinlichkeitsverteilung für die Anzahl der Köpfe beim dreimaligen Münzwurf

Bleiben wir noch einmal kurz bei dem Beispiel und versuchen uns die Wahrscheinlichkeiten anders herzuleiten. Sollten wir zum Beispiel einmal in die Verlegenheit kommen und 20 Münzwürfe untersuchen wollen, dann wir die Tabelle relative schnell relativ unhandlich.

Sei N die Anzahl der Würfe die wir durchführen. Wenn wir N kennen, wissen wir auch direkt welche möglichen Ausgänge bei dem Experiment möglich sind, nämlich alle Zahlen zwischen 0 und N. 0 wenn wir kein Kopf geworfen haben, und N wenn wir nur Kopf geworfen haben. Dementsprechend sind alle Zahlen dazwischen auch noch möglich.

Schauen wir uns jetzt noch mal den dreimaligen Münzwurf an. Wenn wir kein Kopf werfen in 3 Würfen und betrachten die Würfe wieder sequentiell, dann haben wir $\frac{1}{2}$ für die erste Zahl, $\frac{1}{2}$ für die zweite Zahl und $\frac{1}{2}$ für die dritte Zahl. Also insgesamt $P(1 \text{ Kopf}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. Aber diese Wahrscheinlichkeit hat ja jedes Elementarereignis egal ob es (K,K,K) oder (K,Z,K) oder (Z,Z,K) usw. ist. Jetzt haben wir aber das Problem, das wir für 1× oder 2× Kopf nicht

nur eine Möglichkeit vorhanden diese Anzahl an Kopf zu beobachten. In Table 4.4 haben wir bereits gezeigt, dass jeweils drei verschiedene Möglichkeiten, Kombination von Kopf und Zahl, möglich sind. D.h. wir haben jetzt ein Abzählproblem. Können wir irgendwie direkt bestimmen wie viele unterschiedliche Möglichkeiten es gibt?

Schauen wir uns den Fall $1\times$ Kopf im 3-fach Tupel an. Auf wie viele Arten können wir 3-fach Tupel erzeugen mit nur einem Kopf. Nun, der Kopf ist entweder an der ersten, der zweiten oder der dritten Stelle und die jeweils anderen Position im Tupel sind mit Zahl besetzt. Das hört sich aber ähnlich wie ein Problem an wie wie etwas was wir schon vorher einmal gehört haben. Als wir uns die Anzahl der möglichen Stichproben aus unserer kleinen Welt angeschaut haben. Dort hatten wir das Problem, das wir bestimmen wollten auf wie viele Möglichkeiten wir zwei Stichproben mit jeweils drei Personen aus 20 Personen ziehen können. Dabei sind wir auf den Binomialkoeffizienten gestoßen Equation 1.3.

$$\text{Anzahl} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Formal berechnet der Binomialkoeffizient die Möglichkeiten k Objekte aus n Objekten zu ziehen. Wenden wir das mal auf unseren Dreifachwurf an mit $n = N = 3$ und $k = 1$. Ausgeschrieben, auf wie viele Arten können wir $1\times$ Kopf aus drei Positionen auswählen.

$$\text{Kombinationen mit } 1\times \text{ Kopf} = \binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3\times 2\times 1}{1\times 2\times 1} = 3$$

Passt. Probieren wir das auch direkt mit dem Ereignis $2\times$ Kopf, also mit $N = 3$ und $k = 2$, aus.

$$\text{Kombinationen mit } 2\times \text{ Kopf} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3\times 2\times 1}{2\times 1\times 1} = 3$$

Passt auch. Jetzt müssen wir noch nur die beiden Fälle $0\times$ und $3\times$ Kopf behandeln. Wenn wir in einem Mathebuch den Binomialkoeffizienten nachschlagen, dann sind dort die beiden folgenden Definition zu finden für die Fälle $k = 0$ und $k = n$.

$$\binom{N}{N} = 1$$
$$\binom{N}{0} = 1$$

Wenn wir diese Definition für die anderen beiden verbleibenden Fälle anwenden, erhalten wir:

$$\text{Kombinationen mit } 0\times \text{ Kopf} = \binom{3}{0} = 1$$
$$\text{Kombinationen mit } 3\times \text{ Kopf} = \binom{3}{3} = 1$$

Damit können wir nun für alle möglichen Ausgängen die Anzahl der möglichen Elementarereignisse mittels bestimmen. Allgemein erhalten wir dadurch eine Formel für die Wahrscheinlichkeiten der Ereignisse für den dreifachen Münzwurf.

$$P(k \times \text{Kopf}) = \binom{3}{k}\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \binom{3}{k}\left(\frac{1}{2}\right)^3 \tag{4.5}$$

Weil wir natürlich sofort nach einer allgemeinen Lösung streben führen wir jetzt noch ein paar Symbole ein. Die Zufallsvariable, also die Anzahl von Kopf, bezeichnen wir mit dem Großbuchstaben Y. Einen speziellen Ausgang bezeichnen wir mit dem Kleinbuchstaben y. Damit würden allgemein die Wahrscheinlichkeit für irgend eines der Ereignisse mit $Y = y$ bezeichnen. Und wenn wir sagen wir das Ereignis 2× Kopf bezeichnen, mit $y = 2$. Also, die Wahrscheinlichkeit für 3× Kopf mit:

$$P(Y = 3) = \binom{3}{3}\left(\frac{1}{2}\right)^3$$

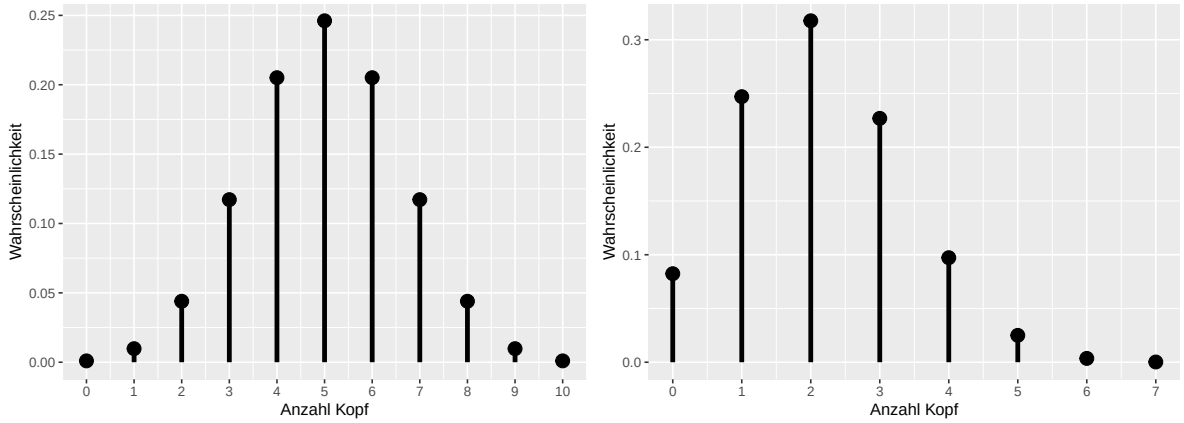
Die nächste Verallgemeinerung die wir Vornehmen ist dass wir für die Wahrscheinlichkeit das Kopf auftritt das Symbol p benutzen. So könnten wir auch modellieren, wenn wir eine unfaire Münze haben. Wenn jetzt aber $p \neq \frac{1}{2}$ gilt, also zum Beispiel die Wahrscheinlichkeit für Kopf $p = \frac{2}{3}$ wäre, dann ist die Wahrscheinlichkeit für Zahl nicht mehr die Gleiche wie für Kopf. Die Wahrscheinlichkeit für Zahl wäre dann $1 - p$. Wenn wir für die Wahrscheinlichkeit für das Auftreten von Zahl das Symbol q einführen, muss die Wahrscheinlichkeit für Kopf oder Zahl gleich 1 sein, formal:

$$p + q = 1$$

Daraus folgt, dass $q = p - 1$. Wenn wir das auf unseren Münzwurf übertragen, müssen wir das dementsprechend berücksichtigen. Wir können uns aber zunutze machen, dass wir wissen wie viele Würfe durchgeführt wurden, nämlich N, und wie viele davon Kopf waren, nämlich y. Damit wissen wir automatisch auch die Anzahl von Zahl, $N - y$. Jedes Kopf, hat die Wahrscheinlichkeit p und jede Zahl hat die Wahrscheinlichkeit $q = 1 - p$. Das gilt unabhängig von der Reihenfolge, da z.B. die Wahrscheinlichkeiten $KKZK$ und $ZKKK$ gleich $ppqp = qppp$ sind. Insgesamt haben wir $y \times K$ und $(n - y) \times Z$ also p^y und q^{n-y}. Diesen Zusammenhang können wir in eine Formel stecken.

$$P(Y = y) = \binom{N}{y} p^y (1-p)^{N-y} = \binom{N}{y} p^y q^{N-y} \tag{4.6}$$

Damit haben wir jetzt auch direkt unsere erste theoretische Verteilung kennengelernt, die in der Statistik eine zentrale Rolle spielt. Die Verteilung in Formel (4.6) wird als die Binomialverteilung bezeichnet. Da die Formel (4.6) von den Parametern p und n abhängt, wird die Binomialverteilung als eine Familie von Verteilungen bezeichnet.



(a) $p = 0.5, n = 10$ (b) $p = 0.3, n = 7$

Figure 4.5: Beispiel für verschiedene Binomialverteilungen

Schauen wir uns aber noch mal ob wir mit den ganzen Symbolen wirklich unseren dreifachen Münzwurf zurückbekommen. Es gilt $N = 3, p = \frac{1}{2}$. Daraus folgt das $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$. Wenn wir uns noch an $x^a x^b = x^{a+b}$ aus der Schule erinnern folgt:

$$\begin{aligned}
P(Y = 0) &= \binom{3}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^3 = \binom{3}{0} \left(\frac{1}{2}\right)^3 = 1 \left(\frac{1}{2}\right)^3 \\
P(Y = 1) &= \binom{3}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^2 = \binom{3}{1} \left(\frac{1}{2}\right)^3 = 3 \left(\frac{1}{2}\right)^3 \\
P(Y = 2) &= \binom{3}{0} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 = \binom{3}{2} \left(\frac{1}{2}\right)^3 = 3 \left(\frac{1}{2}\right)^3 \\
P(Y = 3) &= \binom{3}{0} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^0 = \binom{3}{3} \left(\frac{1}{2}\right)^3 = 1 \left(\frac{1}{2}\right)^3
\end{aligned}$$

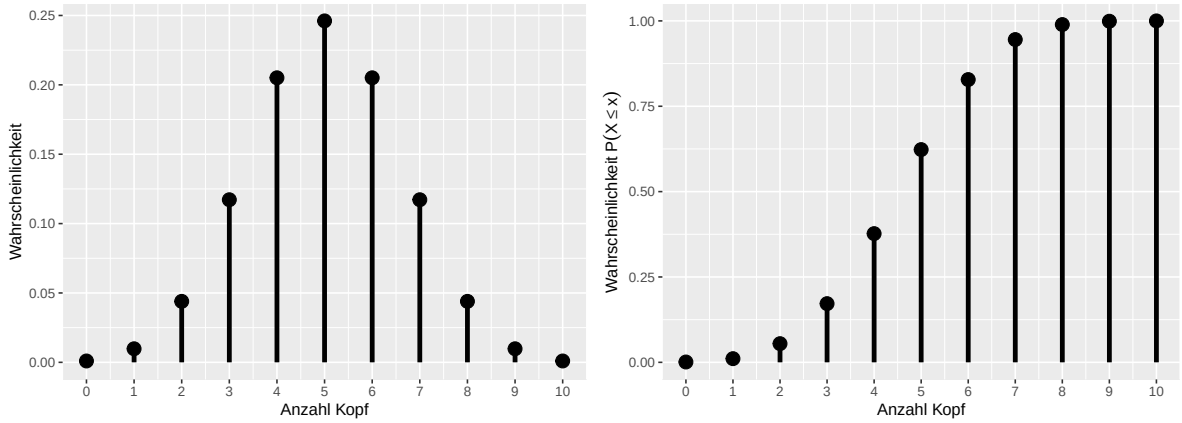
Tatsächlich können wir unser Ergebnis von oben wiedergewinnen. Die Funktion der Binomialverteilung (Formel (4.6)) wird als Wahrscheinlichkeitsfuntion bezeichnet.

Definition 4.2 (Wahrscheinlichkeitsfunktion)**.** Eine Wahrscheinlichkeitsfunktion ist eine mathematische Funktion, die die Wahrscheinlichkeiten für alle möglichen Ausgänge eines diskreten Zufallsexperiments angibt. Sie wird auch als diskrete Wahrscheinlichkeitsverteilung bezeichnet. Eine Wahrscheinlichkeitsfunktion ordnet jedem möglichen Ausgang x eines Experiments eine Wahrscheinlichkeit $P(X = x)$ zu. Die Wahrscheinlichkeit liegt zwischen 0 und 1. Die Summe aller Wahrscheinlichkeiten für alle möglichen Ergebnisse muss gleich 1 sein. Eine Wahrscheinlichkeitsfunktion kann als Tabelle oder als Formel dargestellt werden

Für die Eigenschaften einer Verteilung gibt es einer weitere Darstellungsform, die Verteilungsfunktion.

Definition 4.3 (Verteilungsfunktion)**.** Die Verteilungsfunktion gibt die Wahrscheinlichkeit P an, dass eine Zufallsvariable X einen Wert kleiner oder gleich einem bestimmten Wert x annimmt, formal $P(X \leq x)$. Sie wird daher auch als kumulative Verteilungsfunktion bezeichnet.

Um die Definition der Verteilungsfunktion leichter nachzuvollziehen schauen wir uns das Ganze graphisch an (siehe Figure 4.6).



(a) Wahrscheinlichkeitsfunktion $P(X = x)$ (b) Verteilungsfunktion $P(X \leq x)$

Figure 4.6: Zusammenhang zwischen der Wahrscheinlichkeits- und der Verteilungsfunktion bei $p = 0.5, n = 10$

Die Wahrscheinlichkeitsfunktion gibt, wie schon bekannt, die Wahrscheinlichkeit für eine bestimmtes Ereignis an. Zum Beispiel, die Wahrscheinlichkeit bei $p = 0.5, n = 10, 5\times$ Kopf zu sehen ist etwas unter 0.25. Wir könnten uns aber auch fragen, was die Wahrscheinlichkeit ist 5 oder weniger Köpfe zu beobachten. Diese Wahrscheinlichkeit setzt sic zusammen aus $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$. Genau diesen Wert gibt die Verteilungsfunktion (siehe Figure 4.6b) an.

Die beiden Funktionen sind dabei eineindeutig aufeinander abbildbar. Wenn die Verteilungsfunktion bekannt ist, dann kann daraus die Wahrscheinlichkeitsfunktion berechnet werden

und anders herum wenn die Wahrscheinlichkeitsfunktion bekannt ist, dann kann, wie wir eben gesehen haben, die Verteilungsfunktion berechnet werden. Später bei den kontinuierlichen Verteilungen lernen wir noch die Dichtefunktion kennen, welche die Funktion der Wahrscheinlichkeitsfunktion einnimmt.

Für unser Ausgangsproblem ist jetzt aber mit der Verteilungsfunktion die Möglichkeit gegeben, das wir bestimmte Wahrscheinlichkeitsbereiche unserer Verteilung auszeichnen können. Denn die Wahrscheinlichkeitsfunktion liefert uns die Antwort auf die Frage, welchen Wertebereich wir für eine gegebene Verteilung eher nicht erwarten würden. Schauen wir uns zum Beispiel die Verteilung bei $p = 0.5$ und $n = 30$.

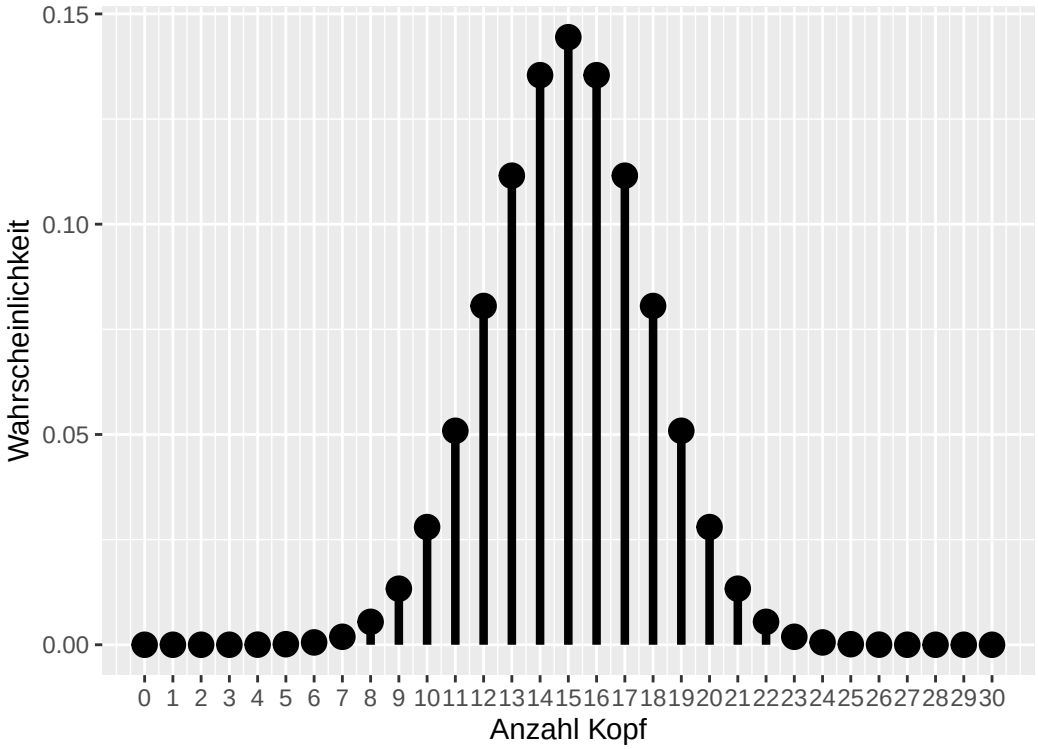


Figure 4.7: Wahrscheinlichkeitsfunktion bei $p = 0.5$ und $n = 30$

In Figure 4.7 sehen wir, dass wir zum Beispiel recht überrascht wären, wenn wir bei einem Durchgang von 30 Münzwürfen einen Wert von z.B. $x = 29\times$ Kopf beobachten würden. Es ist nicht unmöglich, aber es wäre schon überraschend. Diesen Grad der Überraschung können wir als Kriterium nehmen, um zu entscheiden ob wir eine bestimmt Beobachtung dazu verwenden würden diese als Evidenz für oder gegen eine bestimmte Verteilungsannahme zu sehen.

Setzen wir unser Kriterium z.B. bei 2% an. Die Entscheidung wird jetzt folgendermaßen getroffen. Wenn wir einen Wert beobachten der unter der Annahme einer fairen Münze die

wir 30× aus dem Bereich der Werte von $\leq 2\%$ kommt. Dann sehen wir dies als gegen die Annahme an.

Im Folgenden werden vier verschiedene Verteilungen noch einmal etwas genauer vorgestellt, da diese Verteilung immer wieder im weiteren Verlauf auftauchen werden. Dies sind die Normalverteilung, die t-Verteilung, die χ^2-Verteilung und die F-Verteilung. Dabei ist es, außer bei der Normalverteilung, weniger wichtig sich die Formeln einzuprägen sondern es soll eher darum gehen die Form der Verteilung, den Wertebereich und die Parameter der Verteilung zu kennen. Also zum Beispiel wird die Normalverteilung durch zwei Parameter μ und σ^2 spezifiziert während die χ^2-Verteilung nur über einen einzelnen Parameter den Freiheitsgrad df bestimmt wird. Streng genommen wird auch nicht über vier Verteilungen gesprochen, sondern es handelt sich um jeweils Verteilungsfamilien, da es beispielsweise nicht die eine Normalverteilung gibt, sondern die Form wie eben beschrieben von den beiden Parametern abhängt. Dies gilt in gleich3em Maßen ebenfalls für die anderen behandelten Verteilungen.

4.2 Normalverteilung

Beginnen wir mit der Normalverteilung.

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

Die Normalverteilung ist eine symmetrische Verteilung und hat die uns schon oft begegnete Glockenform (siehe Figure 4.8).

Der Wertebereich der Normalverteilung ist $X \in [-\infty, \infty]$. Das Maximum liegt genau beim Erwartungswert μ der dementsprechend die Verteilung in die linken 50% und die rechten 50% unterteilt. Das Abfallen der Flanken wird über die Varianz σ^2 geregelt. Wird σ^2 größer, fallen die Flanken flacher ab, wird σ^2 kleiner, fallen die Flanken schneller ab (siehe Figure 4.9).

Die Standardabweichung kann dazu verwendet werden, die Dichtfunktion in verschiedene Abschnitte zu unterteilen. Es gelten die folgenden Zusammenhänge (siehe Table 4.6):

Table 4.6: Wahrscheinlichkeiten für verschiedene Intervalle um μ in Abhängigkeit von σ

$x \in$	P
$[-\sigma, \sigma]$	0.682
$[-2\sigma, 2\sigma]$	0.955
$[-3\sigma, 3\sigma]$	0.997

Übertragen auf den Dichtegraphen folgt (siehe Figure 4.10):

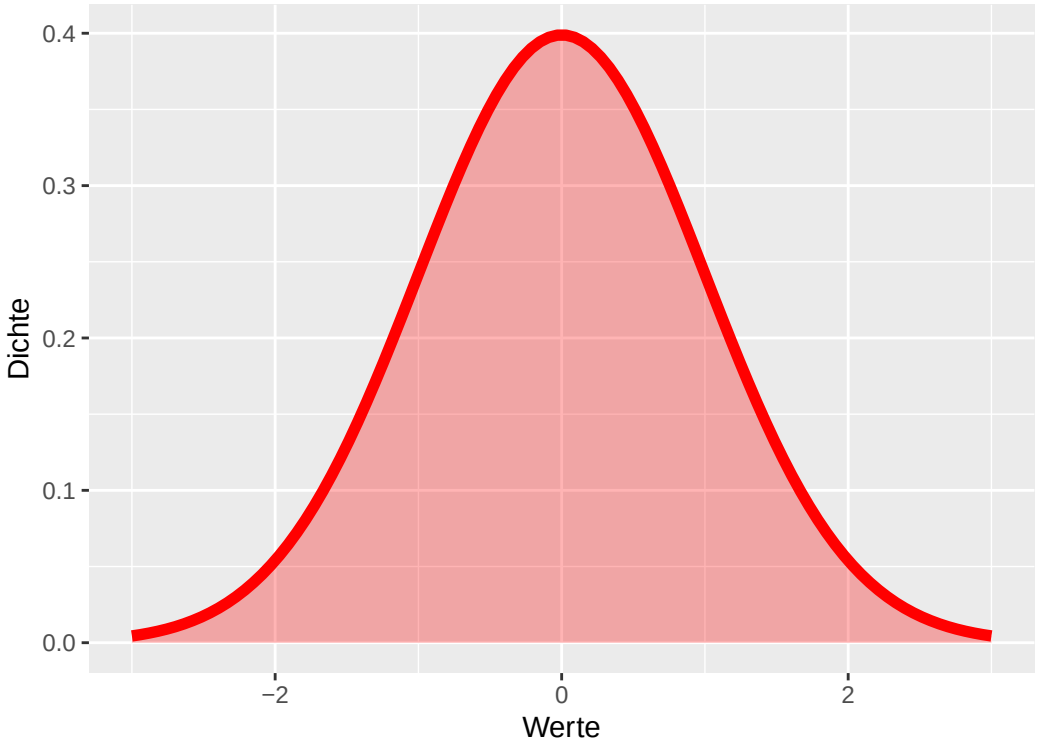
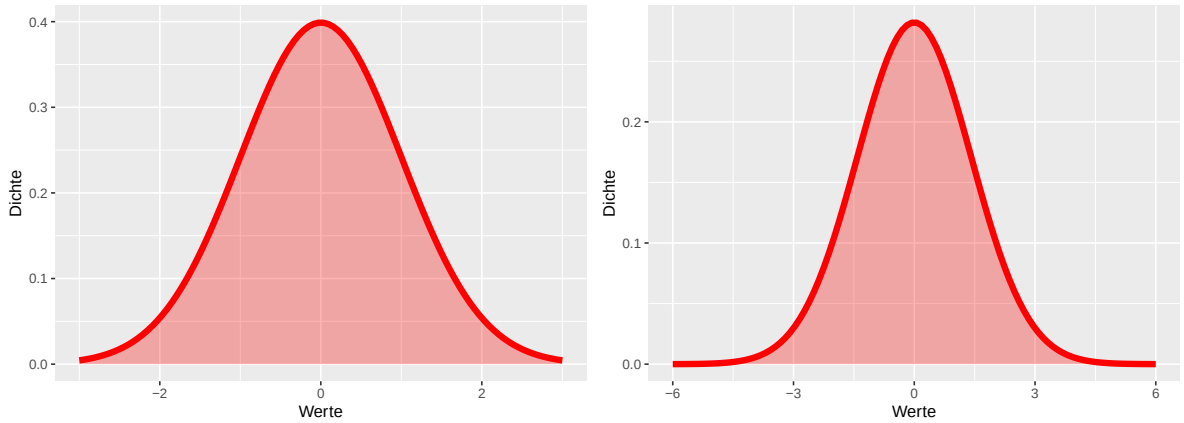


Figure 4.8: Dichtefunktion der Normalverteilung mit den Parametern $\mu = 0$ und $\sigma = 1$.



(a) $\sigma^2 = 1$ (b) $\sigma^2 = 2$

Figure 4.9: Veränderung der Dichtefunktion bei unterschiedlichen Varianzen σ^2

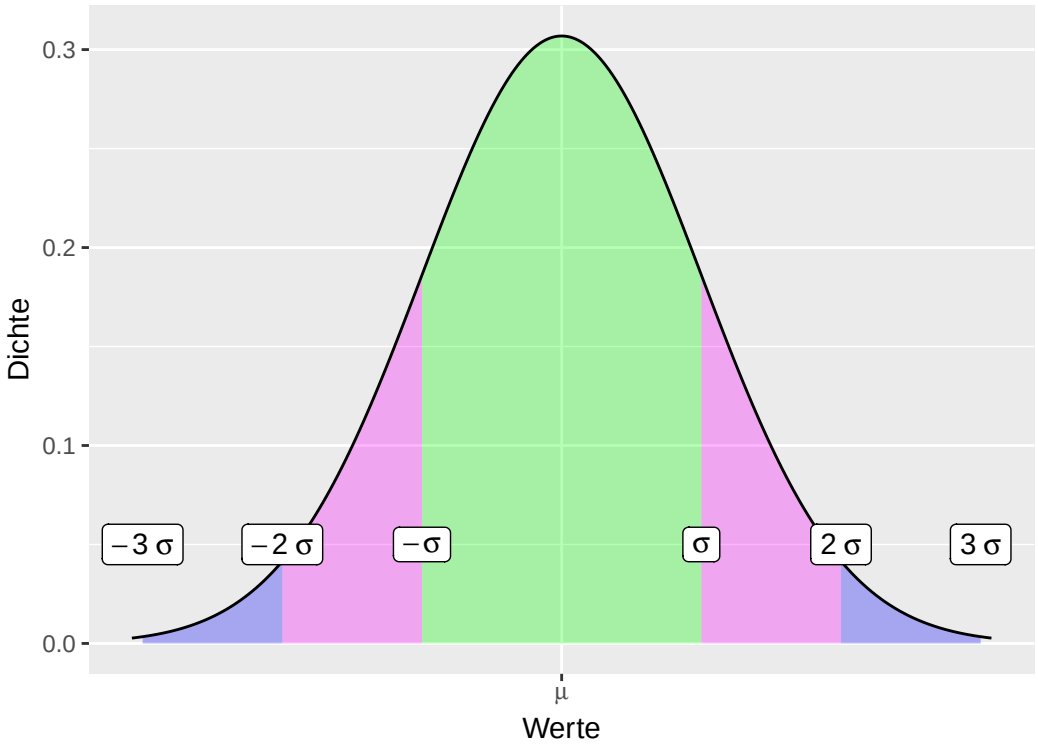


Figure 4.10: Dichtefunktion von $\mathcal{N}(\mu, \sigma^2)$

Wie in Table 4.6 zu sehen ist, hat der Bereich $[-2\sigma, 2\sigma]$ eine Wahrscheinlichkeit von etwas über 0.95. Daher, wenn ich einen Bereich um den Erwartungswert μ auszeichnen möchte, der genau eine Wahrscheinlichkeit von 0.95 hat, dann muss σ mit einem kleineren Wert als 2 multipliziert werden, nämlich 1.96. Das wird hier noch mal speziell erwähnt, da die Zahl 1.96 später immer wieder auftaucht. Formal:

$$P(x \in [\mu - 1.96\sigma, \mu + 1.96\sigma]) = 0.95$$

Anders herum, wenn es darum geht in Konfidenzintervall abzuschätzen, dann funktioniert auch die Faustregel, Teststatistik $\pm 2\times$ Standardfehler.

4.2.1 Die Standardnormalverteilung

Eine Sonderrolle in der Familie der Normalverteilungen spielt die Standardnormalverteilung mit $\mu = 0$ und $\sigma^2 = 1$. Tatsächlich taucht diese so oft aus, dass die Mathematiker ihr ein eigenes Symbol spendiert haben $\phi(x)$

$$\phi(x) = \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

Im Fall der Standardnormalverteilung nehmen Table 4.6 und Figure 4.10 besonders einfache Formen an da die Intervalle jeweils $[-1, 1]$, $[-2, 2]$ und $[-3, 3]$ sind (siehe Figure 4.11).

4.2.2 *z*-Transformation

Es besteht mittels einer einfachen Möglichkeit jede beliebiege Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ auf die Standardnormalverteilung $\mathcal{N}(0, 1)$ abzubilden. Die Transformation wird als z-Transformation bezeichnet und hat die folgende Form:

$$z = \frac{X - \mu_X}{\sigma_X} \tag{4.7}$$

D.h. der Mittelwert der Verteilung von X wird von X abgezogen und die Differenz wird durch die Standardabweichung der Population σ_X geteilt. Die Umkehrfunktion ist dementsprechend:

$$X = \mu_X + z\sigma_X \tag{4.8}$$

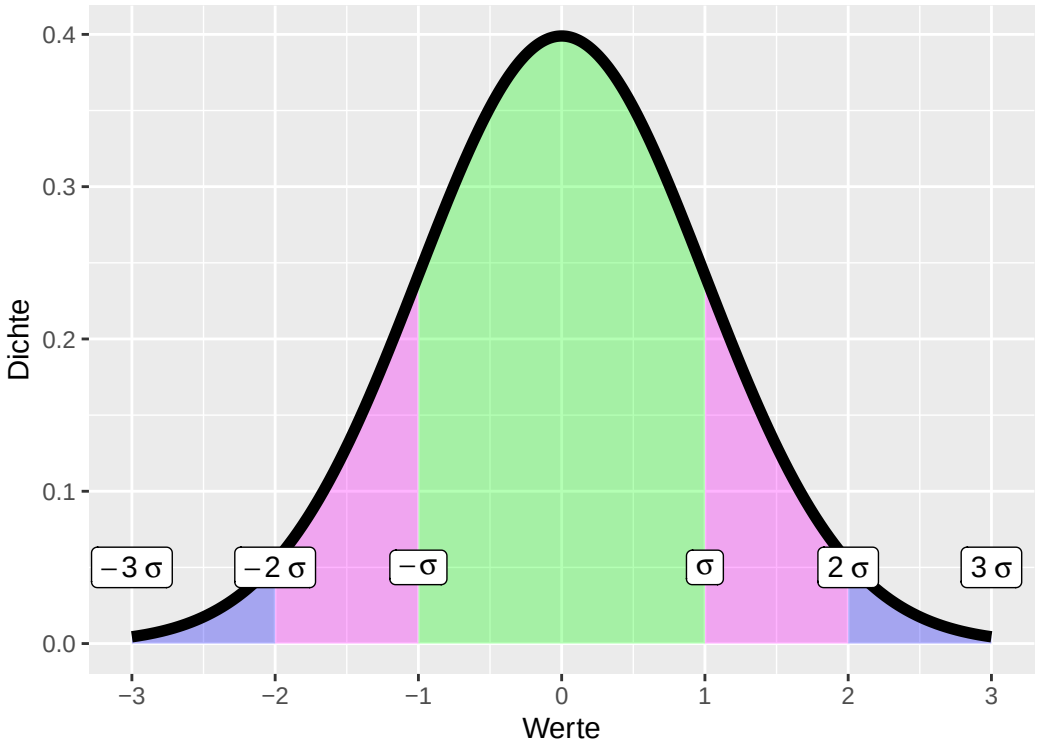


Figure 4.11: Dichtefunktion der Standardnormalverteilung $\phi(x)$ mit $\mu = 0$ und $\sigma^2 = 1$

4.2.3 Zentraler Grenzwertsatz

Die Normalverteilung spielt in der Wahrscheinlichkeitstheorie und der Statistik aus verschiedenen Gründen eine Spezialrolle. Ein Grund dafür ist der sogenannte Zentrale Grenzwertsatz, den wir hier nicht beweisen sondern nur kurz diskutieren.

Proposition 4.1 (Zentraler Grenzwertsatz)**.** *Seien $X_1, X_2, \ldots, X_n$ n unabhängige, gleichverteilte Zufallsvariablen mit $E[X_i] = \mu$ und $Var[X_i] = \sigma^2$ endlich.*

$$\lim_{n\to\infty} \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \;\rightarrow\; \mathcal{N}(\mu = 0, \sigma^2 = 1)$$

In Worten besagt der Zentrale Grenzwertsatz, dass egal welche Ursprungsform die Verteilung einer Zufallsvariablen X hat, wenn die Stichprobengröße gegen unendlich geht, die konvergiert die Differenz des Stichprobenmittelwerts und des Mittelwert der Verteilung geteilt durch den Stichprobenstandardfehler gegen die Standardnormalverteilung. Grenzwertsätz sind manchmal etwas schwierig zu interpretieren, da hier noch keine Aussage gemacht wird, wie groß die Stichprobe sein muss, damit diese Abschätzung valide ist. In der Praxis wird oft ab einer *gefühlt* großen Stichproben diese Abschätzung als zulässig angesehen.

5 Verteilungszoo

5.1 t-Verteilung

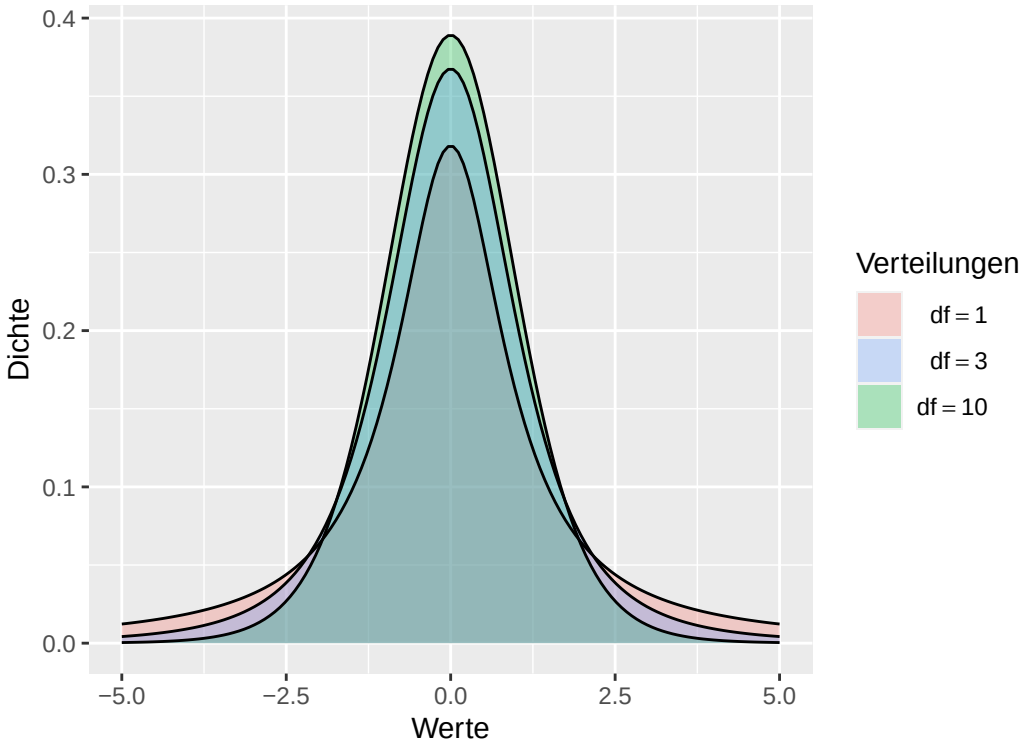


Figure 5.1: Beispiel für verschiedene Dichtefunktionen der t-Verteilung

5.2 χ^2-Verteilung

5.3 F-Verteilung

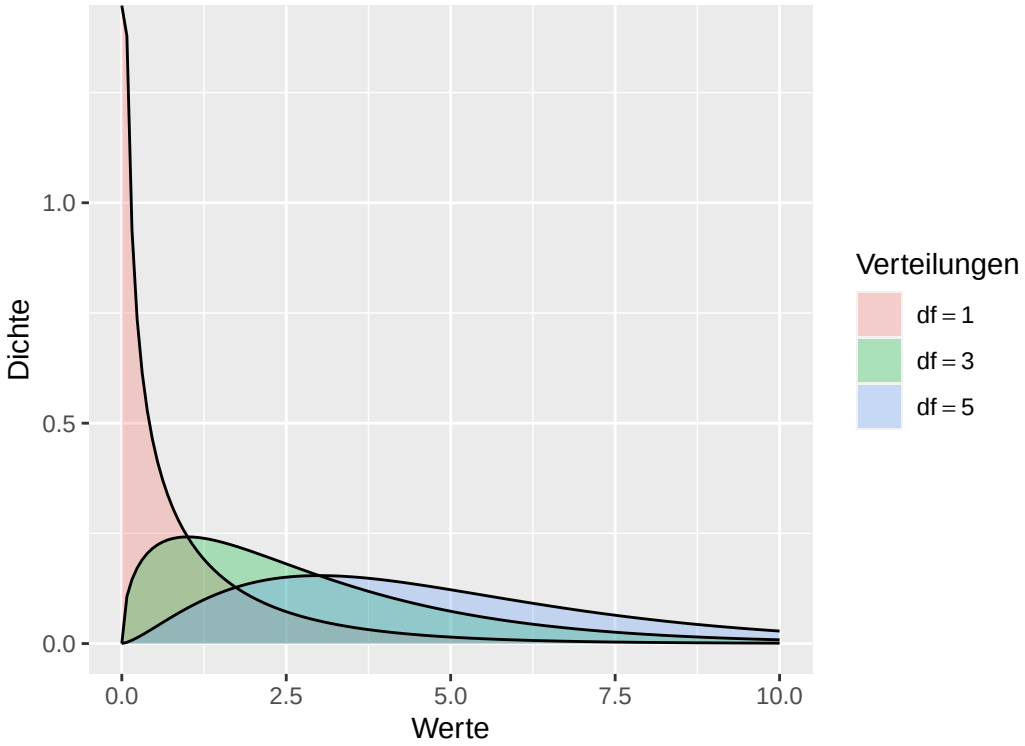


Figure 5.2: Beispiele für verschiedene Dichtefunktion der χ^2-Verteilung.

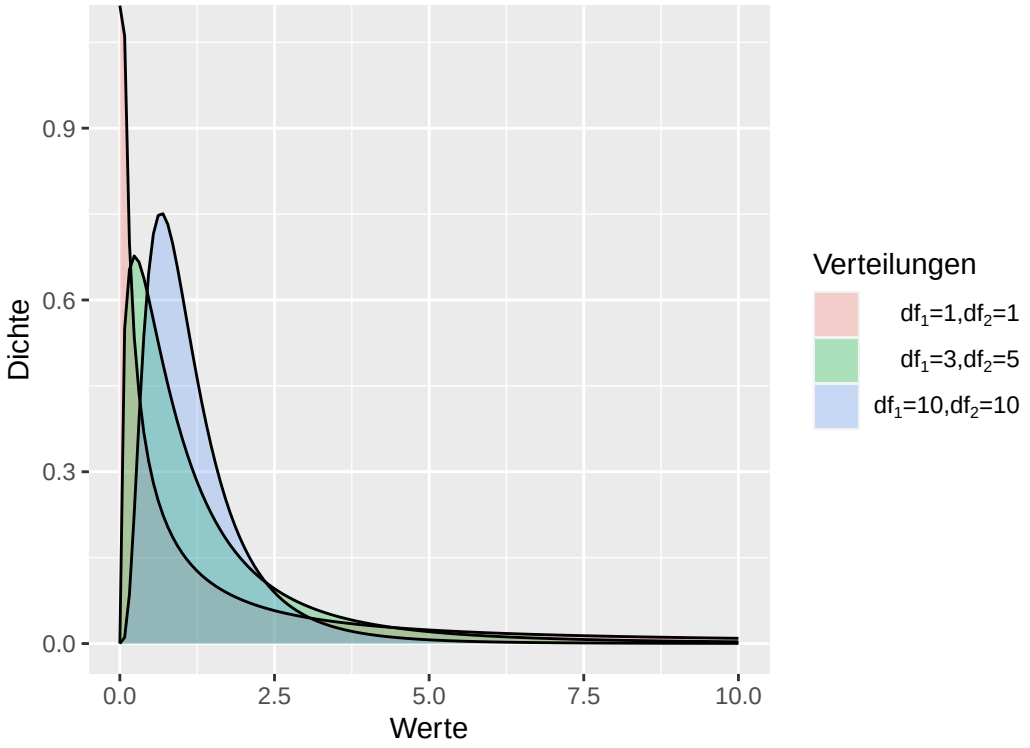


Figure 5.3: Beispiele für verschiedene Dichtefunktion der F-Verteilung.

6 Hypothesen testen

6.1 Wahrscheinlichkeitstheorie

6.2 Rechenregeln zum Erwartungswert und der Varianz

6.2.1 Erwartungwert

Für eine diskrete Zufallsvariable X auf einer endlichen Menge $\{x_i, i = 1, \dots, n\}$ mit n Elementen ist der Erwartungswert definiert mit:

$$E[X] = \sum_{i=1}^{n} x_i P(x_i)$$

D.h. jedes mögliche Ereignis wird mit seiner Wahrscheinlichkeit multipliziert und die Summe über alle diese Möglichkeiten wird gebildet. Da der eine zentrale Rolle in der Wahrscheinlichkeitstheorie und der Statistik spielt, hat er ein eigenes Symbol bekommen μ. Daher wird uns immer wieder die Schreibweise:

$$E[X] = \mu_X$$

Oder wenn der Zusammenhang klar ist und nur von einer bestimmten Zufallsvariablen gesprochen wird, dann auch nur μ. Es hat sich eingebürgert, die Größe μ als den Mittelwert der Population zu bezeichnen auch wenn es sich dabei nicht unbedingt um den Mittelwert handelt wie er üblicherweise verstanden wird und z.B. bei der Stichprobe berechnet wird ($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$). Bei dem Erwartungswert handelt es sich um den gewichteten Mittelwert und wird daher manchnal die Unterscheidung vorgenommen wenn von dem Mittelwert der Population μ und dem Mittelwert der Stichprobe $\bar{x}$ gesprochen wird.

Im folgenden werden verschiedene Rechenregeln mit dem Erwartungswert aufgelistet. Diesen Regeln werden wir immer wieder begegnen wenn wir später Erwarungswerte für Statistiken berechnen. Die erste Regel bezieht sich darauf, wenn eine Zufallsvariable mit einer Konstanten a multipliziert wird. Konstant heißt, bei a handelt es sich nicht um eine Zufallsvariable und a hat immer den gleichen Wert. Der Erwartungswert berechnet sich dann mittels:

$$E[aX] = \sum_{i=1}^{n} ax_i P(x_i) = a \sum_{i=1}^{n} x_i P(x_i) = aE[X]$$

In den meisten Fällen sind wir nicht an einer einzelnen Zufallsvariablen interessiert, sondern, beispielsweise wenn wir eine Stichprobe untersuchen, es liegen mehrere Zufallsvariablen vor. Im einfachsten Fall starten wir mit zwei unabhängigen Zufallsvariablen X und Y. Die beiden Variablen können auf der gleichen Ereignismenge definiert sein, können aber auch auf unterschiedlichen Ereignismengen, z.B. $\{x_i, i = 1, \dots, n\}$ und $\{y_j, j = 1, \dots, m\}$ definiert sein. Wollen wir den Mittelwert von X und Y berechnen und davon den Erwartungswert berechnen, müssen wir verstehen wie sich die Addition unabhängiger Zufallsvariblen auf den Erwartungswert auswirkt. Tatsächlich ist diese Operation relativ einfach zu verstehen, der Erwartungswert von $E[X + Y]$ berechnet sich mittels:

$$E[X + Y] = \sum_{i=1}^{n} x_i P(x_i) + \sum_{j=1}^{m} y_j P(x_j) = E[X] + E[Y]$$

Diese Formel generalisiert für unabhängige $X_i, i = 1, \dots, n$ zu:

$$E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n]$$

In Kombination mit der Regel für konstante Terme mit den Konstanten $a_1, a_2, \dots, a_n$ folgt:

$$E[a_1 X_1 + a_2 X_2 + \dots + a_n X_n] = a_1 E[X_1] + a_2 E[X_2] + \dots + a_n E[X_n]$$

Beispiele

Nehmen wir zur Veranschaulichung ein einfaches Beispiel mit einer Zufallsvariable X welche die folgende Verteilung hat (siehe Table 6.1):

Table 6.1: Verteilung der Zufallsvariablen X

x	0	1	2	3
$P(x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Dann berechnet sich der Erwartungswert $E[X]$ mittels:

$$E[X] = \sum_{i=1}^{4} x_i P(x_i) = \frac{1}{8} \cdot 0 + \frac{1}{8} \cdot 1 + \frac{5}{8} \cdot 2 + \frac{1}{8} \cdot 3 = 1.25$$

Hier kann auch eine interessante Eigenschaft des Erwartungswerts beobachtet werden, nämlich das der berechnete Wert gar nicht in der Menge der möglichen Werte der Zufallsvariablen vorkommen muss. In der Ereignismenge von X sind nur ganzzahlige Werte.

Haben wir eine zweite Zufallsvariable Y mit der Verteilung (siehe Table 6.2)

Table 6.2: Verteilung der Zufallsvariablen X

y	0	1	2	3
$P(y)$	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{3}{8}$

Mit $E[Y]$:

$$E[Y] = \sum_{i=1}^{4} y_i P(y_i) = \frac{2}{8} \cdot 0 + \frac{2}{8} \cdot 1 + \frac{1}{8} \cdot 2 + \frac{3}{8} \cdot 3 = 1.625$$

Dann folgt für den Erwarungswert von $E[X + Y]$:

$$E[X + Y] = E[X] + E[Y] = 1.25 + 1.625 = 2.875$$

Definieren wir eine neue Zufallsvariable Z mit $Z := a \cdot X$ mit der Konstanten $a := 2$. Dann folgt für den Erwartungswert von $E[Z]$:

$$E[Z] = E[aX] = aE[X] = 2 \cdot 1.25 = 2.5$$

Ein ganz anderes Beispiel, welches noch mal den Begriff Erwartungswert veranschaulicht, bezieht sich auf ein Glückspiel mit dem Namen Chuck-a-Lcuk. Das Beispiel ist Gross, Harris, and Riehl (2019) entnommen. Das Spiel wird mit einem 1 € Einsatz gespielt. Es werden drei Würfel geworfen und die folgende Regeln bestimmen den Gewinn (siehe Table 6.3).

Table 6.3: Gewinnauschüttung bei Chuck-a-Luck

Ausgang	Gewinn
keine 6	0 EU
min. eine 6	2 EU
3 x 6	27 EU

Die Frage die sich nun stellt, ist ob dieses Spiel fair ist bzw. lohnt es sich einen 1 € Einsatz zu setzen? Diese Frage kann mit dem Erwartungswert beantwortet werden. Um den

Erwartungswert zu berechnen benötigen wir allerdings zunächst die Wahrscheinlichkeiten für die verschiedenen Ausgänge.

Die Wahrscheinlichkeit keine 6 zu werfen ist für jeden Würfel einzeln $\frac{5}{6}$, dementsprechend, da die Würfel unabhängig voneinander sind, kann diese Wahrscheinlichkeit dreimal miteinander multipliziert werden.

$$P(0 \times 6) = \left(\frac{5}{6}\right)^3 = \frac{125}{216} \approx 0.579$$

D.h. in knapp 60% der Fälle wird beim dem Spiel kein Gewinn ausgeschüttet. Berechnen wir zunächst den Fall, dass drei Sechsen geworfen werden. Der ist Parallel zu keiner Sechs, nur das jetzt für einzelnen Würfel die Wahrscheinlichkeit $\frac{1}{6}$ ist. Es folgt.

$$P(3 \times 6) = \left(\frac{1}{6}\right)^3 = \frac{1}{216} \approx 0.005$$

D.h. die Wahrscheinlichkeit für 3×6 ist gerade einmal ein halbes Prozent. D.h. in 500 Spielen, würde wir dieses Ereignis nur ein einziges Mal erwarten.

Letzlich bleibt noch das Ereignis mindestens eine 6. Hier nehmen wir das Komplementärereignis zu mindestens eine Sechs heißt, nämlich keine Sechs und ziehen dessen Wahrscheinlichkeit von 1, dem sicheren Ereignis, ab. Da diese Menge auch die drei Sechsen beinhaltet, müssen wir dessen Wahrscheinlichkeit auch noch abziehen.

$$P(\text{min. eine 6}) = 1 - P(0 \times 6) - P(3 \times 6) = \frac{216}{216} - \frac{125}{216} - \frac{1}{216} = \frac{90}{216} = 0.41\bar{6}$$

Die Wahrscheinlichkeit für mindestens eine Sechs ist dementsprechend etwas über 40%. Jetzt wenden wir wieder die Formel für den Erwartungswert an um die zu erwartende Gewinnsumme zu bestimmen. Die Gewinnsumme nimmt jetzt den Wert der Zufallsvariablen ein.

$$E[X] = \frac{125}{216} \times 0 + \frac{90}{216} \times 2 + \frac{1}{216} \times 27 = \frac{207}{216} \approx 0.958$$

Im Mittel erwarten wir bei dem Spiel einen Gewinn von 0.958€ bei einem Einsatz von 1 €. Daher wird im Mittel ein Verlust bei dem Spiel gemacht.

Als letztes Beispiel schauen wir uns den Erwartungswert des Mittelwerts $\bar{x}$ an.

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

6.2.2 Varianz

6.3 Schätzer

Erwartungstreue

6.4 Hypothesentestung

6.4.1 Der t-Test

Das Verhältnis einer standardnormalverteilten Variable z und eine χ^2-verteilten Variable s folgt einer t-Verteilung.

$$T = \frac{\hat{\Delta}}{\hat{s}_e(\hat{\delta})} \sim t\text{-Verteilung}$$

6.4.2 χ^2-Test der Varianz

Sei $\hat{\sigma}^2$ ein Schätzer für eine Varianz und $H_0 : \sigma^2 = \sigma_0^2$ die Nullhypothese, dann lässt sich eine Teststatistik über die folgende Formel konstruieren:

$$T = d\frac{\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2(d \text{ Freiheitsgrade})$$

6.4.3 F-Test von Varianzverhältnissen

Seien zwei normalverteilte Stichproben gegeben und deren Varianzen über $\hat{\sigma}_A^2$ und $\hat{\sigma}_B^2$ abgeschätzt werden dann kann eine Teststatisk über die Gleichheit der beiden Varianzen $\sigma_A^2 = \sigma_B^2$ über die folgende Formel konstruiert werden.

$$T = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} \sim F(df_A, df_B)$$

Die beiden Varianzen folgen dabei jeweils einer χ^2 Verteilung mit Freiheistgraden df_A und df_B, so dass die Statistik T einer F-Verteilung mit (df_A, df_B) Freiheitsgeraden folgt und die H_0 lautet $H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1$

Part II

Das einfache Regressionmodell

Wir beginnen nun mit dem einfachen Regressionsmodell. Das Modell knüpft an unsere Vorkenntnisse aus der Schule mit linearen Gleichungen an. Ausgehend von diesem Modell werden schrittweise neue Konzept eingeführt. Diese Herangehensweise hat den Vorteil, dass eine einfaches mentales Template immer wieder auf die neuen Konzepte abgebildet werden kann. Diese stetige Aufbau vollzieht sich über den ersten Teil des einfachen Regressionsmodells und wird dann im folgenden Teil, welcher die multiple Regression behandelt, fortgeführt. Dabei wird auch gezeigt, wie vorher voneinander unabhängig gelernte Methoden, wie die Regression und die ANOVA letztendlich aus dem gleichen Ansatz entstehen und es eigentlich keinen Unterschied zwischen den beiden Ansätzen gibt.

7 Einführung

7.1 Back to school

Wir beginnen mit ein Konzept mit dem wir sehr gut umgehen können. Nämlich der Punkt-Steigungsform aus der Schule (siehe Equation 7.1).

$$y = mx + b \tag{7.1}$$

Wir haben eine abhängige Variable y und eine lineare Formel $mx + b$ die den funktionalen Zusammenhang zwischen den Variablen y und x beschreibt. Um das Ganze einmal konkret zu machen setzen wir $m = 2$ und $b = 3$ fest. Die Formel Equation 7.1 wird dann zu:

$$y = 2x + 3 \tag{7.2}$$

Um ein paar Werte für y zu erhalten setzen wir jetzt verschiedene Wert für x ein indem wir x in Einserschritten zwischen $[0, \ldots, 5]$ erhöhen. Um die Werte darzustellen verwenden wir zunächst eine Tabelle (vlg. Table 7.1)

Table 7.1: Tabelle der Daten

x	y
0	3
1	5
2	7
3	9
4	11
5	13

Wenig überraschend nimmt y für den Wert $x = 0$ den Wert 3 an und z.B. für den Wert $x = 3$ nimmt y den Wert $2 \cdot 3 + 3 = 9$ an.

Eine andere Darstellungsform ist naturlich eine graphische Darstellung in dem wir die Werte von y gegen x auf einem Graphen abtragen (siehe Figure 7.1).

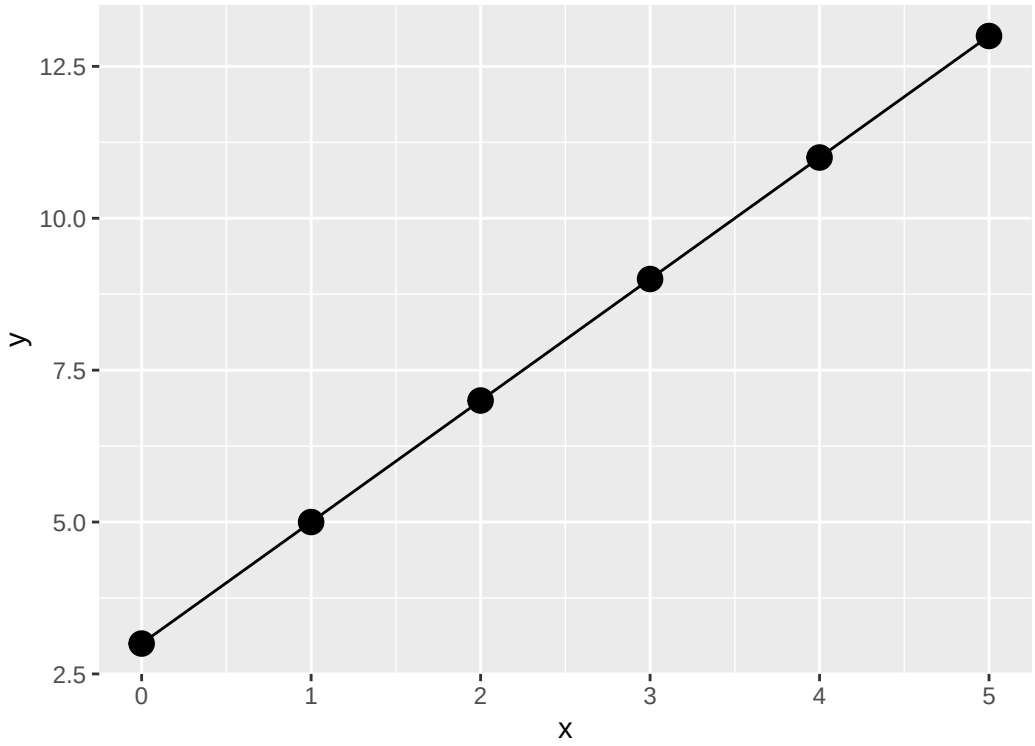


Figure 7.1: Graphische Darstellung der Daten aus Table 7.1

Wiederum wenig überraschen sehen wir einen linearen Zuwachs der y-Wert mit den größerwerdenden x-Werte. Da in der Definition der Formel Equation 7.2 nirgends festgelegt wurde, dass diese nur für ganzzahlige x-Werte gilt, haben wir direkt eine Gerade durch die Punkte gelegt. Hier wird auch die Bedeutung von m und b direkt klar. Die Variable m bestimmt die Steigung der Gleichung während b den y-Achsenabschnitt beschreibt.

Definition 7.1 (y-Achsenabschnitt)**.** Der y-Achsenabschnitt ist der Wert den y einnimmt wenn x den Wert 0 annimmt. Sei y durch eine lineare Gleichung $y = mx + b$ definiert, dann wird der y-Achsenabschnitt durch den Wert b bestimmt.

Die Variable m dahingehend bestimmt die Steigung der Gerade.

Definition 7.2 (Steigungskoeffizient)**.** Wenn y durch eine lineare Gleichung $y = mx + b$ definiert ist, dann bestimmt die Variable m die Steiung der dazugehörenden Gerade. D.h. wenn sich die Variable x um einen Einheit vergrößert (verkleinert) wird der Wert von y um m Einheiten größer (kleiner). Gilt $m < 0$ dann umgekehrt.

Diese beiden trivialen Konzepte mit eigenen Definitionen zu versehen erscheint im ersten Moment vielleicht etwas übertrieben. Wie sich allerdings später zeigen wird, sind diese beiden Einsichten immer wieder zentral wenn es um die Interpretation von linearen statistischen Modellen geht.

Soweit so gut. Führen wir direkt ein paar Symbole ein, die uns später noch behilflich sein werden. Sei jetzt die Menge der x-Werte geben $x = [0, 1, 2, 3, 4, 5]$. Strenggenommen handelt es sich wieder um ein Tupel, da wir jetzt die Reihenfolge nicht mehr ändern. Wir führen nun einen Index i ein, um einzelne Werte in dem Tupel über ihre Position zu bestimmen und wir hängen diesen Index i an x an. Dann wird aus x, x_i.

Table 7.2: x-Werte und ihr Index i

Index i	x-Wert
1	0
2	1
3	2
4	3
5	4
6	5

Damit können wir jetzt einen speziellen Wert zum Beispiel den dritten Wert mit $x_3 = 2$ bestimmen. Wenden wir unseren Index auf unsere Equation 7.1 an, folgt daraus, dass y jetzt auch einen Index i erhält.

$$y_i = mx_i + b \qquad i \text{ in } [1, 2, 3, 4, 5, 6]$$

Wir bezeichnen die beiden Variablen m, die Steigung, und b, den y-Achsenabschnitt, jetzt auch mit neuen Variablen die auch noch einen Index erhalten. Aus m wird β_1 und aus b wird β_0. Damit wird der y-Achsenabschnitt mit β_0 bezeichnet und die Steigung wird mit β_1 bezeichnet. Dann wir aus unserer Gleichung:

$$y_i = \beta_0 + \beta_1 x_i \tag{7.3}$$

Das ist immer noch unsere einfache Punkt-Steigungsform, wir haben lediglich den Index i eingeführt um unterschiedliche $y - x$-Wertepaare zu bezeichnen und wir haben den y-Achsenabschnitt und die Steigung mit neuen Symbolen versehen.

Bei dem bisherigen Zusammenhang handelt es sich um einen funktionalen Zusammenhang zwischen den beiden Variablen x und y. Funktional deswegen, weil wir eine definiertes mathematisches Modell angeben können, d.h. wir haben eine mathematische Funktion welche die Beziehung zwischen den beiden Variablen beschreibt. Wenn wir den Wert für x kenne, dann können wir den präzisen Wert für y ausreichen, indem wir ihn in Equation 7.1 einsetzen. Aus der Schule kennen wir auch noch die Darstellung $y = f(x)$. Streng genommen ist diese Darstellung für Equation 7.1 nicht ausreichend, denn um den Wert für y auszurechnen benötigen wir auch noch Kenntnis über die Werte m und b, bzw. in unsere weiteren Darstellung β_0 und β_1. Daher sollte der Zusammenhang eigentlich mit $y = f(x, \beta_0, \beta_1)$ bezeichnet werden. Es gilt aber immernoch, für gegebene x, β_0 und β_1 ist der Wert für y fest determiniert.

Wenn wir mit realen Daten arbeiten, dann funktioniert dieser Ansatz leider nicht ganz. Selbst wenn wir ein Experiment gleich durchführen werden wir immer etwas unterschiedliche Werte im Sinne der Messungenauigkeit messen. Wenn wir biologische Systeme messen, kommt dazu das diese in den seltensten Fällen zeitstabil sind sondern immer bestimmte Veränderungen von einem Zeitpunkt zum nächsten auftauchen. In Figure 7.2 sind Sprungweiten von mehreren Weitspringerinnen gegen die Anlaufgeschwindigkeit abgetragen. Bei der Betrachtung der Daten erscheint ein linearer Zusammenhang zwischen diesen beiden Variablen durchaus als plausibel.

In Figure 7.2 sind zwei Punkte rot markiert. Die beiden Werte haben praktisch die gleichen x-Werte allerdings unterscheiden sich die y-Werte deutlich von einander. Und dies sind nicht die einzigen Beispielpaare bei denen die x-Werte nahe beiandern liegen, während die y-Werte deutlich weiter voneiander entfernt liegen als bei einen funktionalen Zusammenhang nach Equation 7.1 zu erwarten wäre. Diese Abweichungen kommen durch zufällige Einflussfaktoren wie eben zum Beispiel die Veränderungen angesprochener biologischer Faktoren, Messunsicherheiten, beim Weitsprung draußen sind auch immer externe Einflüsse mögliche, vielleicht wenn es sich um den gleichen Springer handelt, hat er auch beim zweiten Mal keine Lust mehr gehabt. Wenn die Punkte zwei unterschiedliche Springer sind, dann kommt auch dazu, dass zwei Weitspringer bei identischer Anlaufgeschwindigkeit unterschiedliche Sprungfähigkeiten haben oder

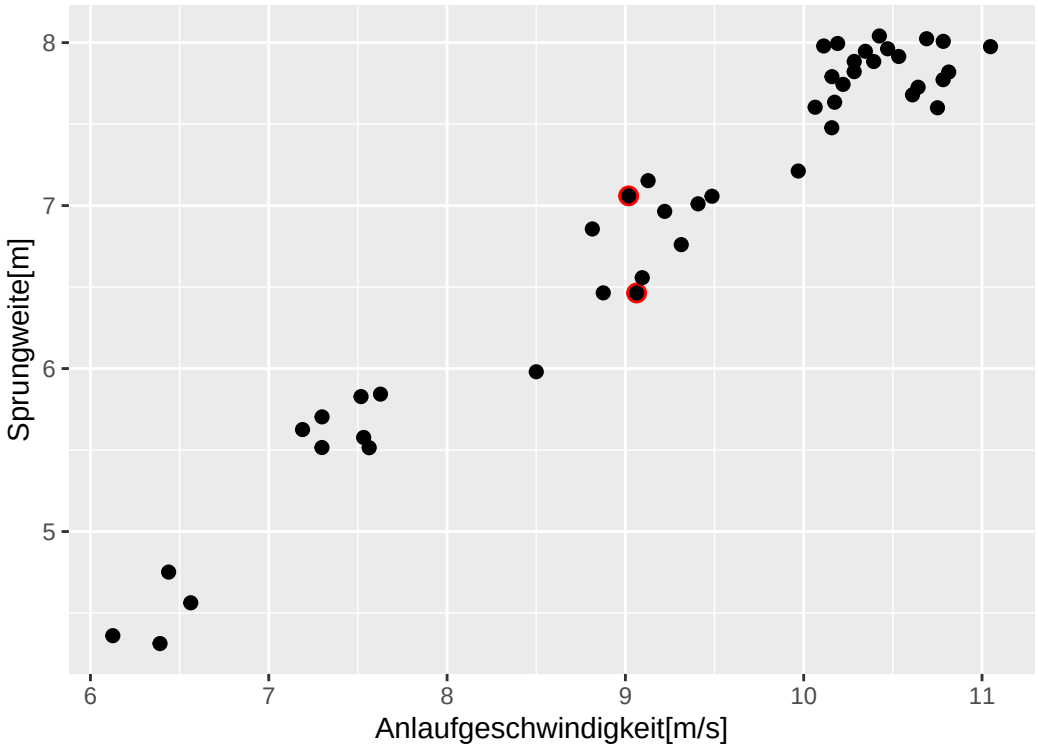


Figure 7.2: Zusammenhang der Anlaufgeschwindigkeit und der Sprungweite beim Weitsprung

auch technisch nicht gleich gesprungen sind und so weiter und so fort. Insgesamt führen alle diese Einflüsse dazu, dass wir nicht mehr einen streng funktionalen Zusammenhang zwischen unseren beiden Variablen x der Anlaufgeschwindigkeit und y der Sprungweite vorfinden. Wie wir mit diesen Einflüssen umgehen ist das zentrale Thema des nächsten Abschnitts und markiert auch unseren Eingang zur einfachen linearen Regression.

7.2 Die einfache lineare Regression

Bleiben wir bei unserem Beispiel aus Figure 7.2 und interpretieren das als praktisches Problem. Wir sind eine Weitsprungtrainerin und stehen jetzt vor der Aufgabe in unserem Training etwas zu verändern um die Weitsprungleistung zu verbessern. Wir haben wir haben uns dazu entschlossen am Anlauf etwas zu verbessern wissen jetzt aber nicht ob, das wirklich lohnenswert ist. Von einer befreundeten Trainerin haben wir einen Datensatz bekommen von Anlaufgeschwindigkeiten und den dazugehörigen Sprungweiten. Schauen wir uns zunächst die einmal die Struktur der Daten an.

Table 7.3: Ausschnitt der Sprungdaten

jump_m	v_ms
4.36	6.13
4.31	6.39
4.56	6.56
4.75	6.44
5.52	7.30
5.63	7.19
5.70	7.30

In Table 7.3 ist ein Ausschnitt Sprungdaten abgebildet. Wir haben eine einfache Struktur der Daten. Wir haben eine Tabelle mit zwei Spalten. `jump_m` bezeichnet die Sprungweiten und `v_ms` die Anlaufgeschwindigkeiten. Damit wir die Datenpaare voneinander unterscheiden bzw. identifzieren können führen wir unseren bereits besprochenen Index i und können so einzelne Paare ansprechen.

Table 7.4: Ausschnitt der Sprungdaten

i	jump_m	v_ms
1	4.36	6.13
2	4.31	6.39
3	4.56	6.56
4	4.75	6.44

i	jump_m	v_ms
5	5.52	7.30
6	5.63	7.19
7	5.70	7.30

Das waren bisher aber nur Formalitäten. Wir wollen jetzt denn Zusammenhang zwischen den beiden Variablen modellieren. Wir könnten wahrscheinlich auch einfach Pi-mal-Daumen abschätzen wie groß der Zusammenhang ist. Wenn wir jetzt aber einen unserer Läufer haben, der z.B. etwa $9m/s$ anläuft, welchen Vergleichswerte nehmen wir dann aus Figure 7.2. Den unteren oder den oberen der beiden roten Werte? Oder vielleicht den Mittelwert? Welchen Wert nehmen wir wenn unserer Athlete $9.7m/s$ anläuft. Da haben wir leider keinen Vergleichswert in unserer Tabelle. Daher wäre es schon ganz praktisch eine Formel nach dem Muster von Equation 7.3 zu haben. Wie wir allerdings schon festgestellt haben, geht dies nicht so einfach da wir eben das Problem mit den Einflussfaktoren haben, die dazu führen, dass die Werte eben nicht streng auf eine Gerade liegen. Somit liegt die Herausforderung nun eine Gerade zu finden die möglichst *genau* die Daten wiederspiegelt.

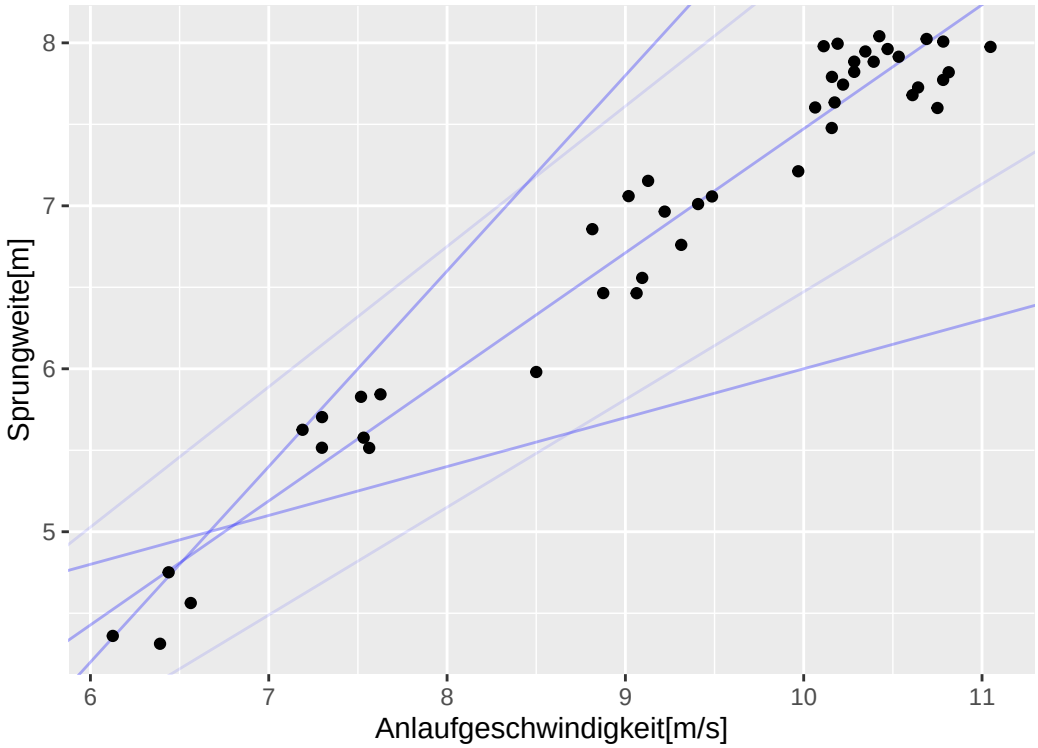


Figure 7.3: Mögliche Geraden um den Zusammenhang der Anlaufgeschwindigkeit und der Sprungweite zu modellieren

In Figure 7.3 sind die Daten zusammen mit verschiedenen möglichen Geraden abgebildet. Eine kurze Überlegung macht schnell klar, dass es im Prinzip unendlich viele unterschiedliche Geraden gibt die durch die Datenpunkte gelegt werden können. D.h. es gibt unendlich viele Kombinationen von β_0 und β_1, die die jeweiligen Geraden bezeichnen. Daher muss jetzt eine Kriterium gefunden werden, welches ermöglicht aus diesen unendlich vielen Geraden eine auszuwählen die im Sinne des Kriterium optimal ist.

Tatsächlich gibt es dort auch verschiedene Möglichkeiten Kriterien anzuwenden, dasjenige dass jedoch am weitesten verbreitet ist aus verschiedenen Gründen sind die quadratierten Abweichungen von der Gerade. Schauen wir uns die Herleitung dazu schrittweise an. In Figure 7.4 ist zur Übersicht nur ein Ausschnitt der Daten zusammen mit einer möglichen Gerade eingezeichnet. Die senkrechten Abweichungen der Geraden zu den jeweiligen Datenpunkten sind rot eingezeichnet. Es ist ersichtlich, dass für diese Wahl der Geraden es zwei Punkte gibt die tatsächlich auch ziemlich genau auf der Geraden liegen während die anderen Punkte zum Teil oberhalb bzw. unterhalb der Geraden liegen. Das Kriterium wäre jetzt dementsprechen die jenige Geraden aus den unendlich vielen zu finden, bei der diese Abweichung ein Minimum annehmen.

$$\min \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i$$

Unglücklicherweise haben die einfachen Abweichungen die unhandliche Eigenschaft, dass dann die *Gerade* $y_i = \hat{y}$ optimal ist.

$$\sum_{i=1}^{n} y_i - \hat{y} = \sum_{i}^{n} y_i - \sum_{i=1}^{n} \hat{y} = \sum_{i=1}^{n} y_i - n\hat{y} = \sum_{i=1}^{n} y_i - n\frac{1}{n}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i = 0$$

Wir können das Kriterium aber auch noch etwas schärfer machen. Wenn wir sagen, dass wir größere Abweichungen stärker gewichten wollen als kleinere Abweichungen. D.h. große Abweichungen zwischen der Gerade und den Datenpunkten sollten stärker berücksichtigt werden, als kleine Abweichungen. Dies können wir erreichen indem wir die Abweichungen noch zusätzlich quadrieren. Dies hat auch noch den Vorteil noch verschiedene andere mathematische Vorteile, unter anderem führt dies dazu, dass wir eine Gerade erhalten, die auch tatsächlich die Steigung der Punkte berücksichtigt und nicht einfache nur eine horizontale Gerade durch die Punkte zeichnet. Dementsprechend erhalten wir die folgende Funktion, die es zu minimieren gilt:

$$\min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{7.4}$$

Die Abweichungen zwischen der zu findenden Gerade und den Datenpunkten werden als Residuen e_i bezeichnet. Dementsprechend ist die Minimierungsgleichung auch als:

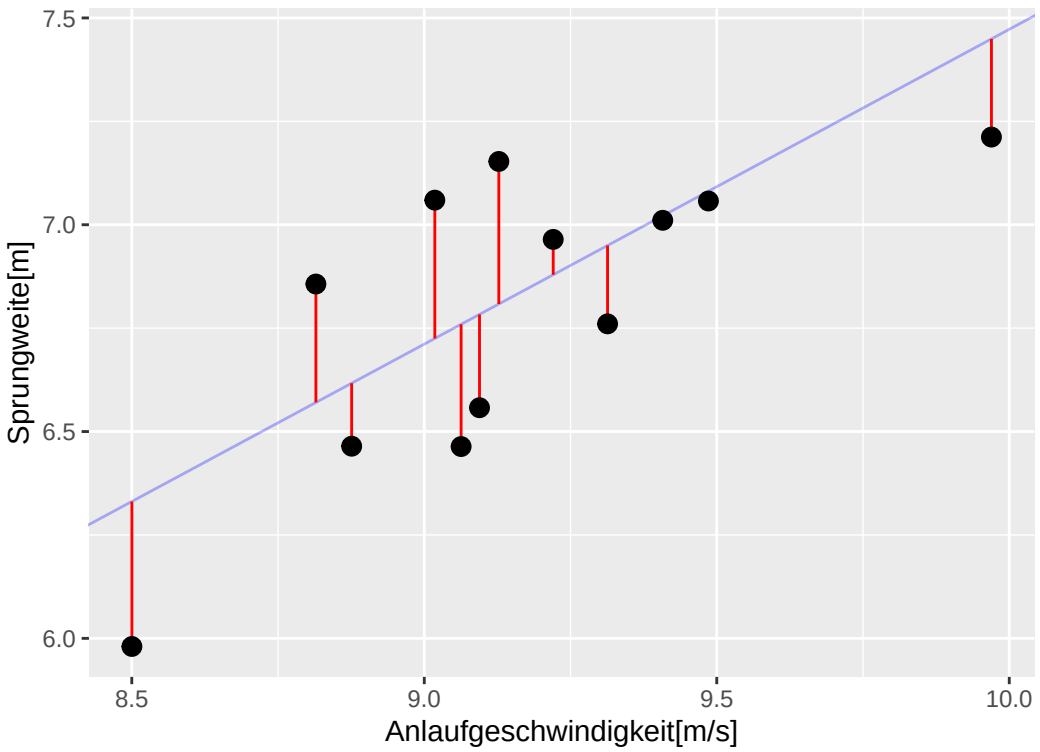


Figure 7.4: Abweichungen der Gerade von der Datenpunkten für die Daten mit eine Anlaufgeschwindigkeit zwischen $8m/s$ und $10m/s$.

$$\min \sum_{i=1}^{n} e_i^2$$

darzustellen, mit $e_i := y_i - (\beta_0 + \beta_1 x_i)$. Führen wir noch eine weitere Bezeichnung E ein, mit der wir die Minimierungsfunktion bezeichnen (E nach englisch error).

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Das Minimum läßt sich finden, indem die partiellen Ableitungen von E nach β_0 und β_1 berechnet werden und, wie wir es aus der Schule kennen, die Ableitungen gleich Null gesetzt werden.

$$\frac{\partial E}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{\partial E}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Diese Gleichungen lassen sich umstellen und nach β_0 und β_1 auflösen:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{7.5}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{7.6}$$

$\bar{x}$ und $\bar{y}$ sind wieder die Mittelwerte von x_i und y_i. Diese beiden Gleichungen werden als die Normalengleichungen bezeichnet.

Wir führen noch einen weiteren Term ein, den vorhergesagten Wert $\hat{y}_i$ von y_i anhand der Geradengleichung. Das Hütchen über y_i ist dabei immer das Signal dafür, das es sich um einen abgeschätzten Wert handelt. Wenn wir β_0 und β_1 anhand der Normalengleichung bestimmen, dann sind das mit großer Wahrscheinlichkeit nicht die *wahren* Werte aus der Population, sondern wir haben sie nur anhand der Daten abgeschätzt. Daher bekommen die berechneten Werte ebenfalls ein Hütchen $\hat{\beta}_0$ und $\hat{\beta}_1$. Insgesamt nimmt die lineare Geradengleichung dann die folgende Form an:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

Graphisch sind die $\hat{y}_i$s die Werte auf der Geraden für die gegebenen x_i-Werte.

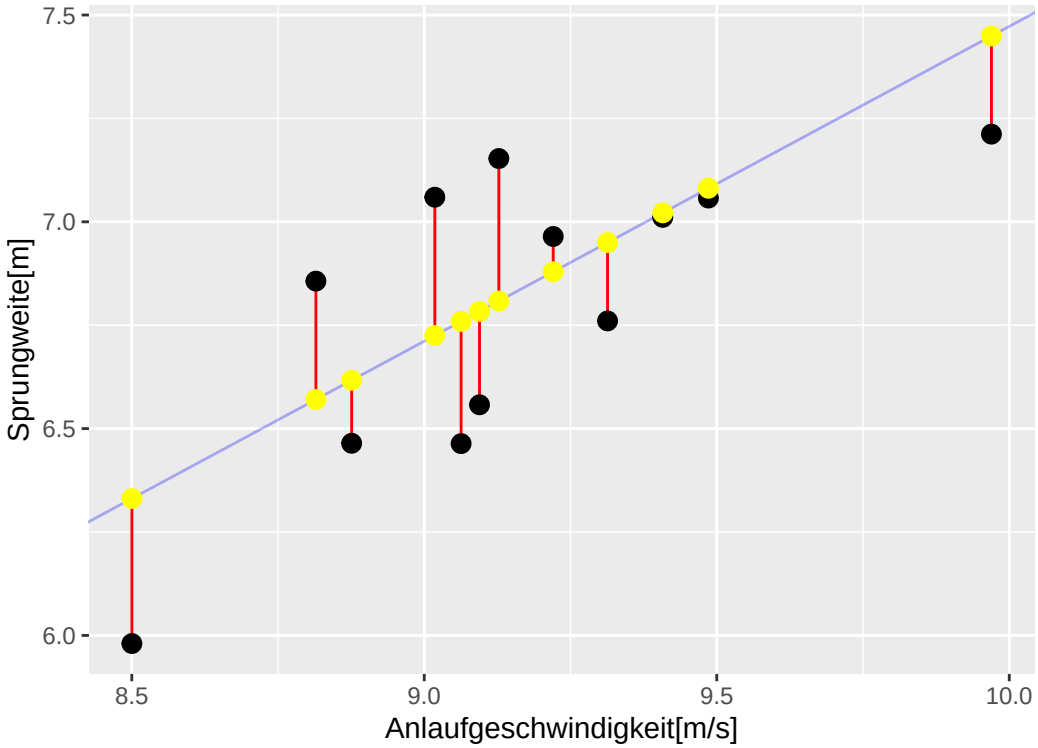


Figure 7.5: Die vorhergesaten Werte $\hat{y}_i$ auf der Gerade.

Für den vorliegenden Fall der Weitsprungdaten erhalten wir die Werte für die Koeffizienten nach Einsetzen der beobachteten Werte in Formel (7.6) mit $\hat{\beta}_0 = -0.14$ und $\hat{\beta}_1 = 0.76$. Somit folgt für die Geradengleichung:

$$\hat{y}_i = -0.14 + 0.76 \cdot x_i$$

Wir erhalten die graphische Darstellung der Geradengleichung indem die x_i-Werte eingesetzt werden und eine Gerade durch die Punkte gezogen wird. Oder auch einfacher für den größten und den kleinsten x_i-Wert.

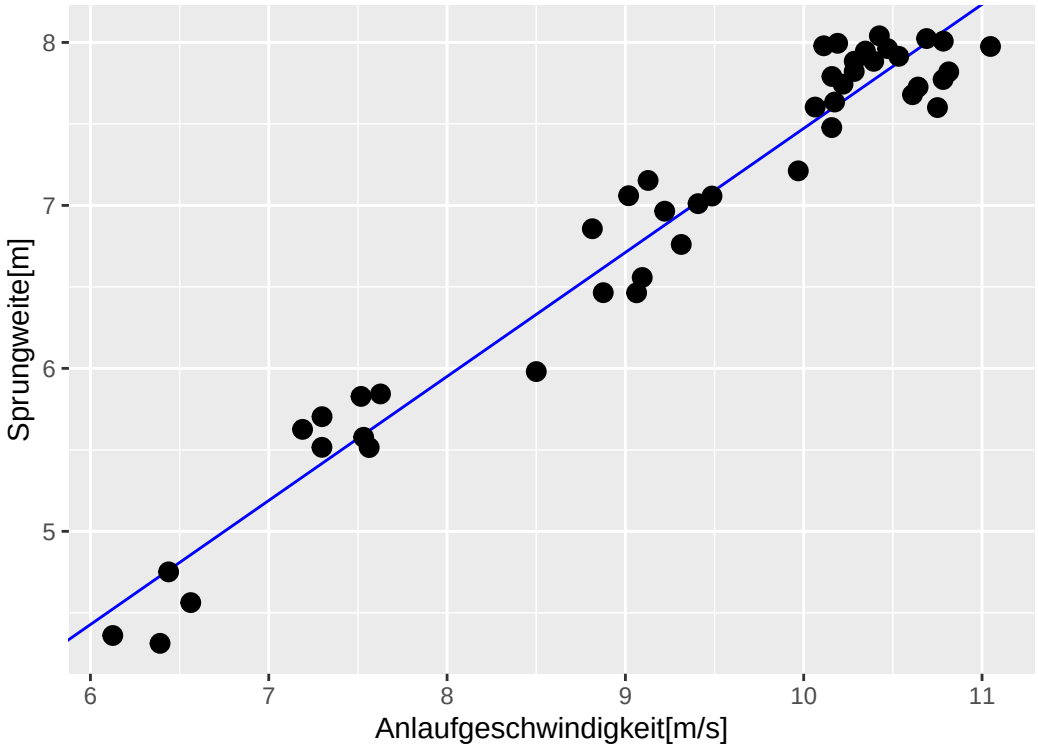


Figure 7.6: Die Regressionsgerade der Sprungdaten.

Um uns auch zu vergewissern, dass unsere Berechnungen korrekt sind, schauen wir uns noch einmal an, wie sich E verhält, wenn wir unterschiedliche Kombinationen von Werten für β_0 und β_1 in die lineare Gleichung einsetzen.

In Figure 7.7 sind verschiedene Werte für E in Form einer heatmap dargestellt. Die Abweichungen wurden *log*-transformiert (d.h. der Logarithmus der E-Werte wurde berechnet), da sonst die Unterschiede in der diagnaolen Bildrichtung zu schnell wachsen und die Unterschiede nicht mehr so einfach zu erkennen sind. Werte näher an Weiß bedeuten kleine Werte und Werte näher an Rot bedeuten größere Werte von E. Das berechnete Paar für $(\hat{\beta}_0, \hat{\beta}_1)$ mit $\hat{\beta}_0 = -0.14$

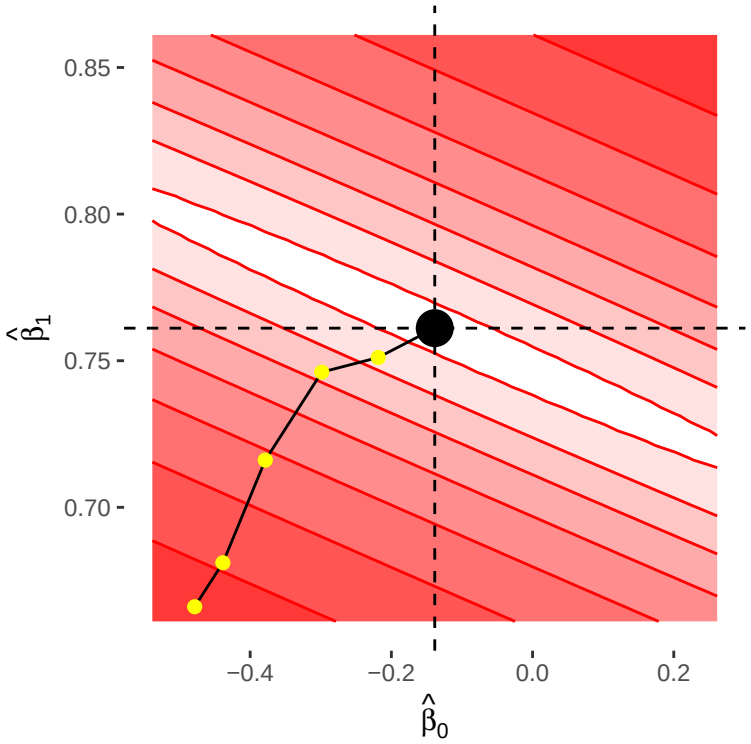


Figure 7.7: Heatmap von $log(E)$ für verschiedene Werte von β_0 und β_1

und $\hat{\beta}_1 = 0.76$ ist schwarz eingezeichnet. Die Abbildung zeigt, dass dieses Wertepaar tatsächlich ein Minimum bezüglich der Funktion E ist, da in alle Richtung weg von dem schwarzen Punkt die Werte für E zunehmen. Da wir nur einen Ausschnitt der möglichen Werte sehen, handelt es sich zunächst um eine lokales Minimum aber es lässt sich zeigen, dass es sich dabei auch um ein globales Minimum handelt. Diese Eigenschaft hängt mit der Form der Funktion E zusammen. In Table 7.5 sind beispielhaft ein paar Werte für $log(E)$ für Paare von β_0 und β_1 angezeigt, die in Figure 7.7 gelb eingezeichnet sind.

Table 7.5: Werte von $log(E)$ für verschiedenen Kombinationen von β_0 und β_1.

β_0	β_1	$log(E)$
-0.48	0.67	70.34
-0.44	0.68	51.85
-0.38	0.72	22.04
-0.30	0.75	6.46
-0.22	0.75	3.77
-0.14	0.76	2.41

7.2.1 Schritt-für-Schritt Herleitung der Normalengleichungen

Um die Herleitung der Normalengleichungen Schritt-für-Schritt nachvollziehen zu können benötigen wir zunächst einmal ein paar algebraische Tricks.

Für den Mittelwert gilt:

$$\bar{x} = \frac{1}{n} \sum x_i \Leftrightarrow \sum x_i = n\bar{x}$$

Bei Summen und konstanten a konstant gilt:

$$\sum a = na \tag{7.7}$$

$$\sum ax_i = a \sum x_i \tag{7.8}$$

$$\sum (x_i + y_i) = \sum x_i + \sum y_i \tag{7.9}$$

Wenn eine Summe abgeleitet wird, kann in die Ableitung in die Summe reingezogen werden.

$$\frac{d}{dx} \sum f(x) = \sum \frac{d}{dx} f(x)$$

Hier ein zwei Umformungen bei Summen und dem Kreuzprodukt bzw. dem Quadrat.

$$\begin{aligned} & \sum (x_i - \bar{x})(y_i - \bar{y}) \\ \Leftrightarrow \quad & \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}) \\ \Leftrightarrow \quad & \sum x_i y_i - \sum \bar{x} y_i - \sum x_i \bar{y} + \sum \bar{x}\bar{y} \\ \Leftrightarrow \quad & \sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\ \Leftrightarrow \quad & \sum x_i y_i - n\bar{x}\bar{y} \end{aligned} \tag{7.10}$$

$$\begin{aligned} & \sum (x_i - \bar{x})^2 \\ \Leftrightarrow \quad & \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ \Leftrightarrow \quad & \sum x_i^2 - 2\bar{x} \sum x_i + \sum \bar{x}^2 \\ \Leftrightarrow \quad & \sum x_i^2 - 2\bar{x} n \bar{x} + n\bar{x}^2 \\ \Leftrightarrow \quad & \sum x_i^2 - n\bar{x}^2 \end{aligned} \tag{7.11}$$

7.2.2 Herleitung

Zurück zu unserem Problem. Es gilt E zu minimieren:

$$\begin{aligned} E = \sum e_i^2 &= \sum (y_i - \hat{y}_i)^2 \\ \Leftrightarrow \quad & \sum (y_i - (\beta_0 + \beta_1 x_i))^2 \\ \Leftrightarrow \quad & \sum (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned} \tag{7.12}$$

Die Gleichung hängt von zwei Variablen β_0 und β_1. Um das Minimum der Gleichung zu erhalten, verfährt man wie in der Schule, indem man die Ableitung gleich Null setzt. Der vorliegenden Fall ist jedoch etwas komplizierter, da die Gleichung von zwei Variablen abhängt. Daher müssen wir die partiellen Ableitungen $\frac{\partial}{\partial \beta_0}$ und $\frac{\partial}{\partial \beta_1}$ verwendet. Wir erhalten dadurch ein Gleichungssystem mit zwei Gleichungen (die jeweiligen Ableitungen) in zwei Unbekannten (β_0 und β_1). Die Lösung erfolgt, indem zuerst eine Gleichung nach der einen Unbekannten umgestellt wird und das Ergebnis dann in die andere Gleichung eingesetzt wird.

Wir beginnen mit der partiellen Ableitung nach β_0 für den y-Achsenabschnitt. (Zurück an die Schule erinnern: Äußere Ableitung mal innere Ableitung)

$$\frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0} \tag{7.13}$$

$$\Leftrightarrow \quad \sum \frac{\partial}{\partial \beta_0}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\Leftrightarrow \quad \sum 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$\Leftrightarrow \quad -2 \sum (y_i - \beta_0 - \beta_1 x_i)$$

Zum minimieren gleich Null setzen.

$$-2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \quad \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \quad \sum y_i - \sum \beta_0 - \sum \beta_1 x_i = 0$$

$$\Leftrightarrow \quad n\bar{y} - n\beta_0 - \beta_1 n\bar{x} = 0$$

$$\Leftrightarrow \quad \bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

$$\Leftrightarrow \quad \bar{y} - \beta_1 \bar{x} = \beta_0$$

$$\Leftrightarrow \quad \beta_0 = \bar{y} - \beta_1 \bar{x} \tag{7.14}$$

Es folgt nach dem gleichen Prinzip die Herleitung für die Steigung β_1 und indem die Lösung für β_0 eingesetzt wird.

$$\frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1} \tag{7.15}$$

$$\Leftrightarrow \quad \sum \frac{\partial}{\partial b}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\Leftrightarrow \quad \sum 2(y_i - \beta_0 - \beta_1 x_i) - x_i$$

$$\Leftrightarrow \quad -2 \sum (y_i - \beta_0 - \beta_1 x_i) x_i \tag{7.16}$$

Wiederum gleich Null setzen.

$$-2 \sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\Leftrightarrow \quad \sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\Leftrightarrow \quad \sum (y_i x_i - \beta_0 x_i - \beta_1 x_i x_i) = 0$$

$$\Leftrightarrow \quad \sum y_i x_i - \beta_0 \sum x_i - b \sum x_i^2 = 0$$

$$\Leftrightarrow \quad \sum y_i x_i - n\beta_0 \bar{x} - \beta_1 \sum x_i^2 = 0$$

Einsetzen der Lösung für β_0 führt zu:

$$
\begin{aligned}
\Leftrightarrow \quad & \sum y_i x_i - n(\bar{y} - \beta_1 \bar{x})\bar{x} - \beta_1 \sum x_i^2 = 0 \\
\Leftrightarrow \quad & \sum y_i x_i - n\bar{y}\bar{x} + n\beta_1 \bar{x}^2 - \beta_1 \sum x_i^2 = 0 \\
\Leftrightarrow \quad & \sum y_i x_i - n\bar{y}\bar{x} = \beta_1 \sum x_i^2 - \beta_1 n\bar{x}^2 \\
\Leftrightarrow \quad & \sum (x_i - \bar{x})(y_i - \bar{y}) = \beta_1 (\sum x_i^2 - n\bar{x}^2) \\
\Leftrightarrow \quad & \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum x_i^2 - n\bar{x}^2} = \beta_1 \\
\Leftrightarrow \quad & \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\end{aligned}
$$

Somit erhält man die beiden Normalengleichungen der Regression.

Über diese beiden Gleichungen erhalten wir die gewünschten Koeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$. Die Methode wird als die als die Methode der kleinsten Quadrate bezeichnet oder im Englischen Root-Mean-Square (RMS).

7.3 Was bedeuten die Koeffizienten?

Gehen wir zurück nun zu unseren Ausgangsproblem der Weitspringer, was haben wir jetzt durch die Berechnung der Gerade eigentlich gewonnen? Dazu müssen wir erst einmal verstehen was die beiden Koeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ bedeuten. Wenn wir zurück zu Equation 7.1 gehen, haben die beiden Koeffzienten den y-Achsenabschnitt und die Steigung der Geraden beschrieben. In unserem Beispiel haben wir anhand der Daten einen y-Achsenabschnitt $\hat{\beta}_0$ von -0.14 berechnet. D.h ein Weitspringer der mit einer Anlaufgeschwindigkeit von $x = 0$ anläuft, landet 14cm hinter der Sprunglinie. Dies macht offensichtlich nicht viel Sinn (warum?). Der Grund warum hier ein offensichtlich unrealistischere Wert berechnet wurde, werden wir später noch genauer betrachten. Wir können trotzdem zwei Eigenschaften von $\hat{\beta}_0$ beobachten. 1) der Koeffizient hat eine Einheit, nämlich die gleiche Einheit wie die Variable y. 2) Ob der Wert zu interpretieren ist, hängt von der Verteilung der Daten ab. Schauen wir uns nun den Steigungskoeffizienten $\hat{\beta}_0$ an. Der Steigungskoeffizient in Equation 7.1 zeigt an, wie sich der y-Wert verändert, wenn sich der x-Wert um einen Einheit verändert. In unserem Fall welcher Unterschied zu erwarten ist zwischen zwei Weitspringern die sich in der Anlaufgeschwindigkeit um eine m/s unterscheiden. D.h. der Steigungskoeffizient ist ebenfalls in der Einheit der y-Variable zu interpretieren.

Unsere Trainerin kann jetzt die berechnete Gerade dazu nehmen um zu überprüfen ob es sich lohnen würde Trainingszeit in den Anlauf zu stecken und welche Verbesserung dort zu erwarten sind. Allerdings fehlt dazu noch etwas, wir wissen nämlich noch nicht ob die berechnete Gerade auch wirklich die Daten gut wiederspiegelt. Im Beispiel erscheint dies anhand der Grafik als relativ plausibel. Das muss aber nicht immer so sein. Wir können nämlich für alle möglichen

Daten eine Gerade berechnen ohne das diese Gerade die Daten wirklich auch nur annährend korrekt wiedergibt. In Formel (7.6) steht nirgends für welche Daten die Berechnung nur erlaubt ist.

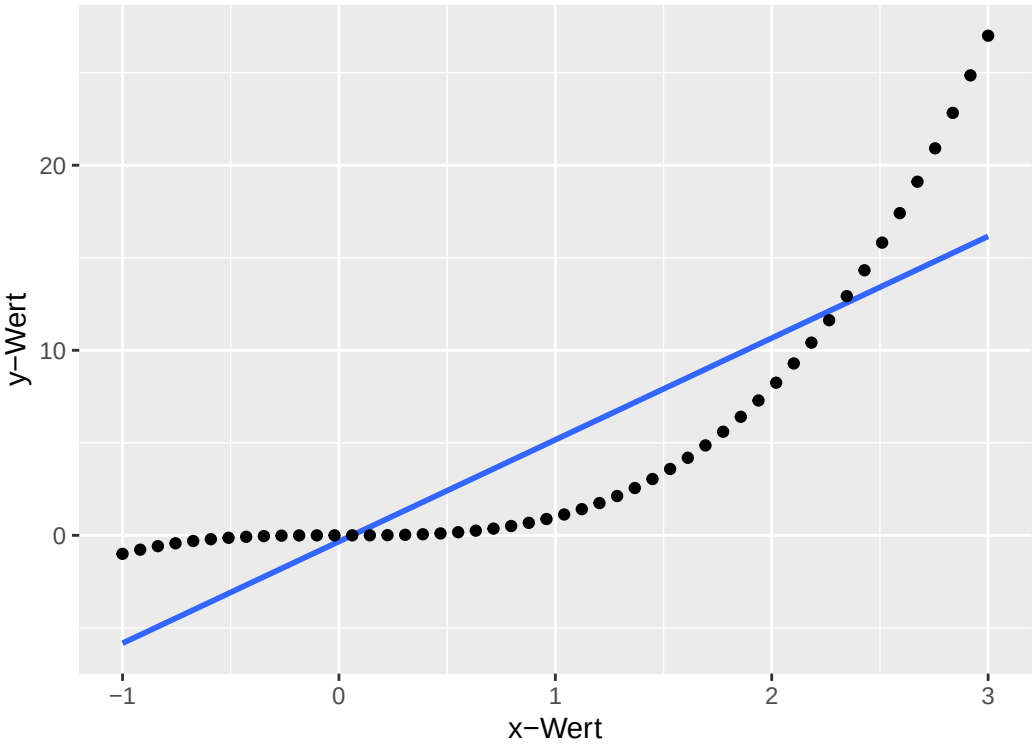


Figure 7.8: Gefittete Gerade durch die Daten einer Funktion $f(x) = x^3$.

In Figure 7.8 sind synthetische Daten der Funktion $f(x) = x^3$ abgebildet und die mittels **?@eq-slm-basics-norm1** berechneten Gerade eingezeichnet. Die Gerade ist zwar in der Lage die ansteigenden Werte zu modellieren aber eben nicht Schwingungen die durch die kubische Abhängigkeit zustande kommen. Aber, nichts verhindert die Anwendung der Formel auf die Daten.

Der gleiche Effekt ist auch in Figure 7.9 wieder zu beobachten. Hier besteht eine sinusförmige Abhängigkeit zwischen y und x. Wir können wieder (7.6) anwenden und erhalten auch ein Ergebnis für $\hat{\beta}_0$ und $\hat{\beta}_1$. Allerdings repräsentiert die Gerade in keinster Weise den tatsächlichen Zusammenhang zwischen den Daten.

Im nächsten Kapitel werden wir uns daher damit beschäftigen die Repräsentation der Daten näher zu betrachten und zu präzisieren.

Wir nehmen noch eine weitere Eigenschaft der Gerade mit, die zunächst nichts mit der Interpretation der Koeffizienten zu tun hat, aber später noch mal von Interesse sein wird. Die

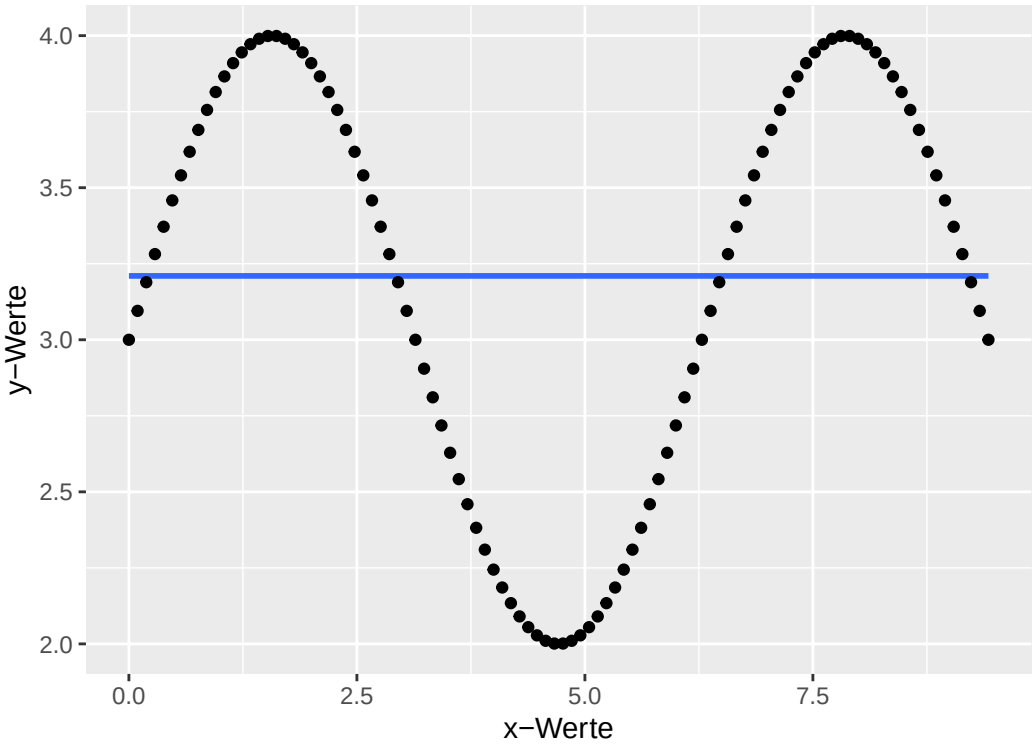


Figure 7.9: Gefittete Gerade durch die Daten einer Funktion $f(x) = sin(x) + 3$.

Gerade hat nämlich die Eigenschaft durch den Punkt $(\bar{x}, \bar{y})$ zu gehen. Dies kann daran gesehen werden wenn in die Gleichung $\bar{x}$ für x_i eingesetzt wird. Anhand der Normalgleichungen kann die Geradengleichung in der Form.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i = \underbrace{\bar{y} - \hat{\beta}_1 \bar{x}}_{\text{Def. } \hat{\beta}_0} + \hat{\beta}_1 \cdot x_i$$

Wird jetzt für x_i der Wert $\bar{x}$ eingesetzt folgt daher.

$$y_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

D.h. für den Wert $\bar{x}$ nimmt die Geradengleichung der Wert $\bar{y}$ an. Für die Sprungdaten ist die auch noch mal in Figure 7.10 graphisch dargestellt.

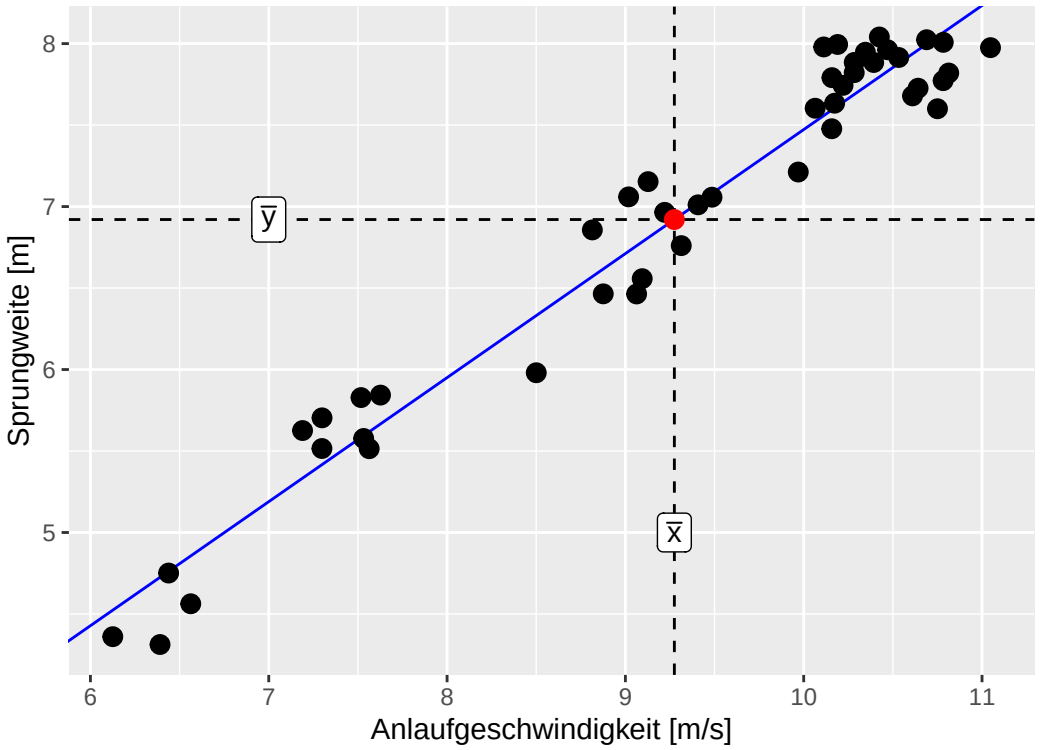


Figure 7.10: Regressionsgerade der Sprungdaten und der Punkt $(\bar{x}, \bar{y})$

Eine Eigenschaft die im weiteren Verständnis immer wieder auftaucht bezieht sich auf die x-Werte. Bei der Regression wird im Allgemeinen davon ausgegangen, dass die beobachteten x-Werte fixiert sind. D.h. trotzdem die x-Werte bei einem Experiment zufällig sein können,

werden diese in den nachfolgenden Schritten als fixiert angesehen. Daher ist in der Formel $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ auch nur ϵ_i die einzige zufällige Variable.

7.4 Die einfache lineare Regression in `R`

In `R` wird eine Regression mit der Funktion `lm()` berechnet. Die für uns zunächst wichtigsten Parameter von `lm()` sind der erste Parameter `formula` und der zweite Parameter `data`. Mit der Formel wird der Zusammenhang zwischen den Variablen beschrieben, dabei können die Namen bzw. Bezeichner aus dem `tibble()` benutzt werden, die an den zweiten Parameter `data` übergeben werden. D.h. die Spaltennamen aus dem `tibble()` werden in `formula` verwendet.

In unserem Weitsprungbeispiel konnten wir in Table 7.3 sehen, das das `tibble()` zwei Spalten mit den Namen `v_ms`, den Anlaufgeschwindigkeiten, und `jump_m`, den Weitsprungweiten enthielt. Dementsprechend, müssen wir diese beiden Bezeichner in `formula` verwenden, um unser Regressionsmodell zu beschreiben. Die Form der Modellbeschreibung folgt, dabei einer bestimmten Syntax die wir uns zunächst anschauen müssen. Zentrales Element der Syntax ist das Tilde Zeichen `~` (Win: ALTGR++, MacOS: _____), welches interpretiert wird als *modelliert mit*. Der Term der auf der linken Seite steht bezeichnet die abhängige Variable während die Terme auf der rechten Seite der Tilde stehen die unabhängige Variablen spezifizieren. Dementsprechend kann der Satz *"Y wird mittels X modelliert"* in die Formelsyntax mit `Y ~ X` übersetzt. Die komplette Syntax orientiert sich an eine Arbeit von Wilkinson and Rogers (1973).

Wenn ein konstanter in der Syntax benötigt wird, dann wird dieser mit einer 1 bezeichnet. Also zum Beispiel wenn wir Equation 7.3 modellieren wollen benutzen wir die Syntax `y ~ 1 + x`. Die beiden Koeffizienten β_0 und β_1 brauchen wir nicht explizit anzugeben, sondern `R` generiert uns automatisch anhand der Bezeichner Koeffizienten, die allerdings die Namen der Bezeichner bekommen. Dazu kommt noch eine Besonderheit, dass `R` bei einer Regressionsgleichung automatisch davon ausgeht, dass ein konstanter Term verwendet werden soll, d.h. der Term `+1` wird automatisch dazugeführt. Wenn wir ein Modell ohne einen y-Achsenabschnitt fitten wollen, dann müssen wir dies `R` explizit mitteilen, indem wir `-1` der linken Seite hinzufügen, also z.B. `y ~ x - 1`. Die Syntax generalisiert dann später einfach, wenn zusätzliche Terme in der multiplen Regression benötigt werden, in dem weitere unabhängige Variablen durch `+` dazugefügt werden. Dementsprechend würde sich die Formel `y ~ x_1 + x_1` übersetzen in *die abhängige Variable y wird mittels der unabhängigen Variablen `x_1` und `x_2` und einem konstaten Term modelliert.* In Table 7.6 sind weitere Beispiele für die Struktur der Formelsyntax für `lm()` gezeigt.

Table 7.6: Formelsyntaxbeispiele für `lm()` (y-Ab = y-Achsenabshnitt, StKoef = Steigungskoeffizient)

Modell	Formel	Erklärung
$y = \beta_0$	`y ~ 1`	y-Ab
$y = \beta_0 + \beta x$	`y ~ x`	y-Ab und StKoef
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	`y ~ x1 + x2`	y-Ab und 2 StKoe

Wenn wir jetzt also unsere Weitsprungdaten modellieren wollen, verwenden wir die folgenden Befehle.

```
lm(jump_m ~ v_ms, data = jump)
```

```
Call:
lm(formula = jump_m ~ v_ms, data = jump)

Coefficients:
(Intercept)         v_ms
    -0.1385       0.7611
```

Per default ist das Ergebnis von `lm()` nicht wirklich besonders hilfreich und es werden nur die beiden berechneten Koeffizienten ausgegeben. Dabei bezeichnet der Term `(Intercept)` den automatisch dazugefügten konstanten Term in der Formel, sprich den y-Achsenabschnitt $\hat{\beta}_0$ und mit `v_ms` den Steigungskoeffizienten $\hat{\beta}_1$. Um aus `lm()` mehr Informationen heraus zu bekommen, ist es sinnvoll das Ergebnis einen Variable zuzuweisen. In dem vorliegenden Arbeit wird dazu in den meisten Fällen eine Variante des Bezeichners `mod` benutzt, als Kurzform vom *model*. Diese Bezeichnung ist aber wie alle Bezeichner in `R` vollkommen willkürlich und entspringt nur der Tippfaulheit des Autors.

```
mod <- lm(jump_m ~ v_ms, data = jump)
```

Um jetzt mehr Informationen aus dem gefitteten `lm()`-Objekt zu bekommen werden Helferfunktion verwendet. Die wichtigste Funktion ist die `summary()`-Funktion (`?summary.lm`).

```
summary(mod)
```

```
Call:
lm(formula = jump_m ~ v_ms, data = jump)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.44314 -0.22564  0.02678  0.19638  0.42148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13854    0.23261  -0.596    0.555
v_ms         0.76110    0.02479  30.702   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2369 on 43 degrees of freedom
Multiple R-squared:  0.9564,    Adjusted R-squared:  0.9554
F-statistic: 942.6 on 1 and 43 DF,  p-value: < 2.2e-16
```

Hier bekommen wir schon deutlich mehr Informationen mitgeteilt. Als erstes die Formell die wir `lm()` übergeben haben. Dann folgt ein Abschnitt über die Residuen, gefolgt von den Koeffzienten und im unteren Abschnitt noch weitere Statistiken. Wir konzentrieren uns zunächst einmal nur auf die Tabelle im Abschnitt `Coefficients`. Hier begegnen uns wieder in der ersten Spalte die Bezeichner für die beiden βs in Form von β_0 `(Intercept)` und β_1 `v_ms`. In der zweiten Spalte daneben stehen die berechneten Koeffizienten die wir jetzt schon mehrmals gesehen haben. Die weiteren Spalten ignorieren wir hier zunächst. Im Laufe der folgenden Kapitel werden wir uns die weiteren Statistiken anschauen und deren Bedeutung verstehen.

Bei der Benutzung von `lm()` werden uns noch weitere Helferfunktionen begegnen, die den Umgang mit dem gefitteten Modell vereinfachen. Wollen wir zum Beispiel die beiden Koeffiziente aus dem Modell extrahieren können wir dazu die Funktion `coefficients()` oder auch nur kurz `coef()` verwenden. Koeffizienten und Standardschätzfehler

```
coef(mod)
```

```
(Intercept)        v_ms
 -0.1385361   0.7611019
```

Die Funktion `coef()` gibt einen Vektor benannten Vektor zurück der entweder über die Bezeichner oder einfach über die Position der Koeffizienten angesprochen werden kann. Möchte ich zum Beispiel den Steigungskoeffizienten verwendent werwende ich:

```
coef(mod)[1]
```

```
(Intercept)
 -0.1385361
```

oder

```
coef(mod)['v_ms']
```

```
     v_ms
0.7611019
```

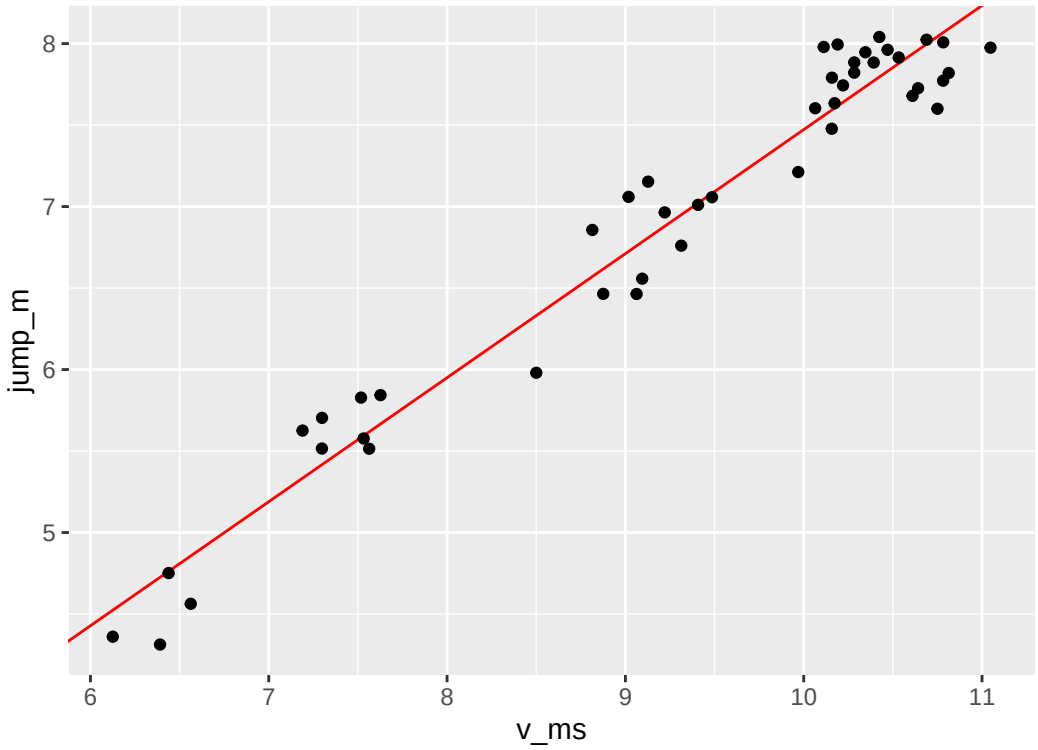
Ein etwas übersichtlicher Zugang ist wieder zunächst einmal das Ergebnis von `coef()` einer Variablen zuweisen und diese dann weiter benutzen.

```
jump_betas <- coef(mod)
jump_betas[1]
```

```
(Intercept)
 -0.1385361
```

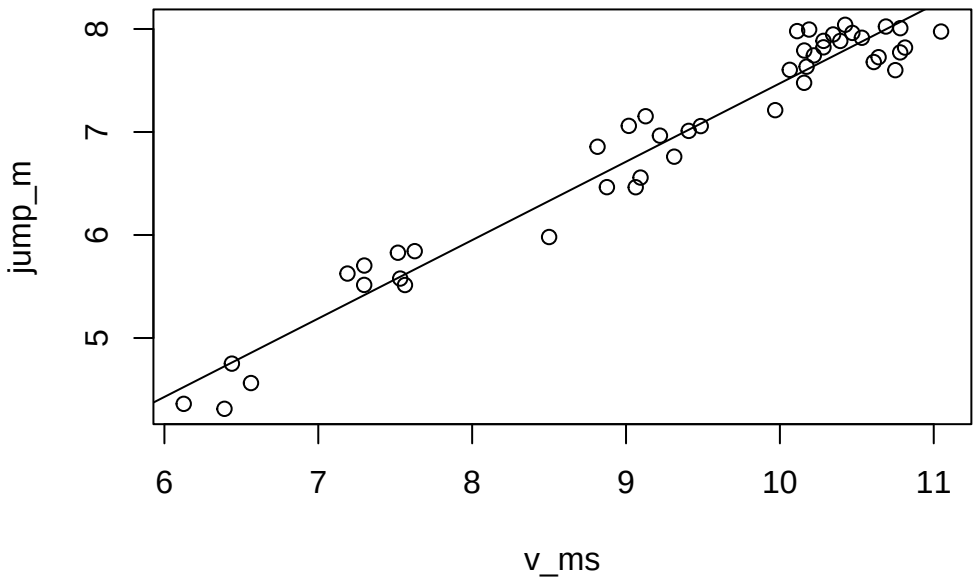
Die Koeffizienten kann ich zum Beispiel benutzen um die Regressionsgerade in ein Streudiagramm hinzuzufügen (Das `tibble()` mit den Sprungdaten hat den Bezeichner `jump`). Entweder mit dem `ggplot2()` Grafiksystem.

```
ggplot(jump,
       aes(x = v_ms, y = jump_m)) +
  geom_abline(intercept = jump_betas[1],
              slope = jump_betas[2],
              color = 'red') +
  geom_point()
```



Oder mit den Standard R-Grafiksystem. Hier kann der Funktion `abline()` das gefittete `lm()`-Objekt direkt übergeben werden und die Koeffizienten werden automatisch extrahiert.

```
plot(jump_m ~ v_ms, data = jump)
abline(mod, color = 'red')
```



Schauen wir uns noch mal ein ganz einfaches Beispiel, bei dem wir tatsächlich wissen welcher Zusammenhang zwischen den beiden Variablen. Wir halten das Beispiel ganz einfache und nehmen vier verschiedene x-Werte mit $x_i = i$. Wir setzen $\beta_0 = 1$ und $\beta_1 = 0.5$. Wir generieren die vier Werte mit `R`, speichern diese in einem `tibble()` mit dem Bezeichner `data` und berechnen die resultierenden Koeffizienten mittels `lm()`.

```r
data <- tibble(
  x = 1:4,
  y = 1 + 0.5 * x
)
mod <- lm(y ~ x, data)
coef(mod)
```

```
(Intercept)           x
        1.0         0.5
```

Und tatsächlich können wir die korrekten Koeffizienten mittels der einfachen linearen Regression wiedergewinnen. Diesen Ansatz mittels synthetisch generierten Daten die eingeführten Konzepte und Ansätze zu überprüfen werden wir im weiteren Verlauf des Skripts immer wieder

anwenden, da er die Möglichkeit bietet relativ einfach und nachvollziehbar das Verhalten verschiedener Ansätze auszutesten.

Zusammenfassend lässt sich sagen, das wir jetzt gelernt haben wie wir ein einfaches Regressionmodell der Form Equation 7.3 an einen beliebigen Datensatz fitten können. Die Berechnung der beiden Koeffizienten β_0 und β_1 erfolgt mittels **?@eq-slm-basics-norm1**. Dabei berechnen wir die Koeffizienten nicht von Hand sondern lassen die von `R` mittels der `lm()` durchführen. Die Berechnung ist dabei vollkommen mechanisch und die Koeffizienten per-se sagen nichts darüber aus, ob das lineare Modell die Daten tatsächlich auch widerspiegelt. Dazu müssen wir noch etwas mehr Theorie aufbauen um Aussagen darüber zu treffen ob das Modell adäquat ist. Dies gehen wir in den folgenden Abschnitten und Kapiteln an.

8 Inferenz

Nachdem wir im vorhergehenden Kapitel gelehrnt haben, wie wir ein Regressionsgerade an einen Datensatz fitten. Stellt sich nun die Frage ob die Regressionsgerade tatsächlich einen relevanten Zusammenhang zwischen den beiden Variablen beschreibt. Da das einfache lineare Modelle zwei Parameter β_0 und β_1 beinhaltet kann diese Fragestellung auf beide Koeffizienten angewendet werden. D.h. wir fragen uns ob das Modell einen statistisch signifikanten Zusammenhang zwischen den beiden Variablen beschreibt. Bezogen auf die beiden Parameter, ist der Parameter $\hat{\beta}_0$ statistisch signifikant und ist der Parameter $\hat{\beta}_1$ statistisch signifikant? Um unseren Werkzeugsatz zu statistischer Signifikanz anwenden zu können brauchen wir aber erst einmal wieder eine Verteilung bei der wir kritische Bereiche markieren können um zu entscheiden ob eine beobachtete Statistik statistisch signifikant ist. Wie behalten dabei im Hinterkopf das statistische Signifikanz nicht das Gleiche ist wie praktische Relevanz.

8.1 Statistische Überprüfung von β_1 und β_0

Der erste Schritt um eine Verteilung zu bekommen ist allerdings, dass wir zunächst einmal eine Zufallsvariable benötigen. Bisher haben wir den Zusammenhang zwischen Variablen über die Formel

$$y_i = \beta_0 + \beta_1 \cdot x_i$$

beschrieben. In dieser Form ist allerdings noch kein zufälliges Element vorhanden. Für ein gegebenes x_i bekommen wir ein genau spezifiziertes y_i. Allerdings haben wir bei der Herleitung gesehen, dass die Daten in den seltensten Fällen genau auf der Gerade liegen, sondern wir die Parameter $\hat{\beta}_0$ und $\hat{beta}_1$ so gewählt haben, dass die quadrierten Abweichungen, die Residuen ϵ_i minimal werden. Dies Residuen verwenden wir jetzt um eine zufälliges Element in unsere Regression rein zu bekommen. Ein mögliche Annahme ist, das die Residuen beispielsweise Normalverteilt sind.

Warum könnte dies Sinn machen. In dem vorhergehenden Weitsprungbeispiel haben wir informell hergeleitet, dass die Weitsprungleistung von unzähligen weiteren Faktoren beeinflusst werden kann, welche dazu führen, dass für eine gegebene Anlaufgeschwindigkeit nicht immer die gleiche Weitsprungweite erzielt wird. Generell, ist diese Art der Begründung bei biologischen System meistens plausibel. In vorhergehenden Abschnitt haben wir dazu aber auch

noch gesehen, dass die Normalverteilung eben gut geeignet ist, um solche Prozesse, bei denen viele kleine additive Effekt auftreten. Dieser Argumentation folgend ist es plausibel diese Einflüsse auch beim Regressionsfall mittels einer Normalverteilung zu modellieren. Dazu führen wir noch eine weitere Annahme an, nämlich dass diese Einflüsse im Mittel in gleichen Maßen die Werte nach nach oben wie auch nach unten ablenken. D.h. die Werte nach oben und unten von der Regressionsgerade abweichen. Dies erlaubt uns jetzt die Annahme genau zu spezifizieren.

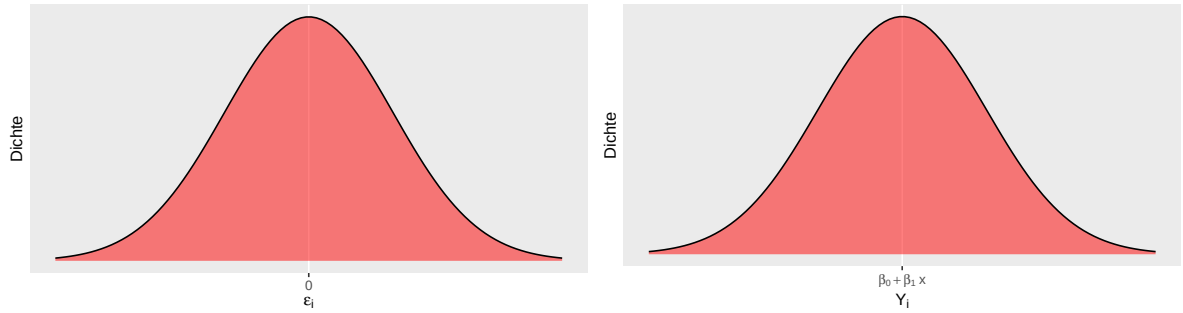
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

D.h also, wir gehen davon aus, dass die Residuen normalverteilt sind, mit einem Mittelwert von $\mu = 0$ und einer noch näher zu spezifizierenden Varianz σ^2. Das führt dann zu der folgenden Formulierung des Regressionsmodells.

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{8.1}$$

Y wird jetzt groß geschrieben, da es sich um eine Zufallsvariable handelt. Dies führt jetzt dazu, das das Regressionsmodell in zwei Teile unterschieden werden kann. Einmal eine deterministischen Teil $\beta_0 + \beta_1 \cdot x$ und einen stochastischen Teil ϵ_i. Dies führt dazu, dass Y_i ebenfalls stochastisch ist und zu einer Zufallsvariable wird.

Schauen wir uns weiter an, wie sich Y_i verhält, wenn wir x_i als Konstante x mit ein bestimmten Wert annehmen. Dann wird aus Equation 8.1 $Y_i = \beta_0 + \beta_i \cdot x + \epsilon_i$. Folglich bleibt der deterministische Teil immer gleich, wird zu einer Konstante. Da ϵ_i normalverteilt ist ist Y_i ebenfalls normalverteilt. Der Mittelwert der Normalverteilung von Y_i μ_{Y_i} ist allerdings nicht gleich Null, sondern die Normalverteilung von ϵ_i wird um die Konstante $\beta_0 + \beta_1 \cdot x$ verschoben (siehe Figure 8.1). Das führt dazu, dass Y_i der Verteilung $\mathcal{N}(\beta_0 + \beta_1 x)$ folgt.



(a) Verteilung von ϵ_i

(b) Verteilung von Y_i

Figure 8.1: Relation der Lageparameter von e_i und Y_i

Daraus folgt jetzt aber zusätzlich, dass für jedes gegebenes X die Y-Werte einer Normalverteilung folgen. Lediglich die Verschiebung des Mittelwert der jeweiligen Y-Normalverteilung hängt von X über die Formel $\beta_0 + \beta_1 \cdot X$ zusammen. Formal:

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Die Schreibweise $|X$ wird übersetzt für gegenbenes X und sagt aus, dass die Verteilung von Y von X abhängt. Es handelt sich dabei um eine bedingte Wahrscheinlichkeit. Die Varianz der jeweiligen Y-Werte ist dabei die zuvor angenommen Varianz der ϵ_i also σ^2. Eine wichtige Annahme die noch mal betont werden sollte, wir gehen davon aus, dass die einzelnen Punkte unabhängig voneinander sind. Im Weitsprungbeispiel würde dies bedeuten, dass jeder Sprung von einem anderen Athleten kommen muss.

Wenn wir die Verteilungen von Y graphisch führ beispielweise drei verschiedene X-Wert darstellen, dann folgt daraus die folgende Abbildung (siehe Figure 8.2). D.h. für jeden X-Wert werden mehrere Y-Werte beobachtet, die jeweils einer Normalverteilung folgen.

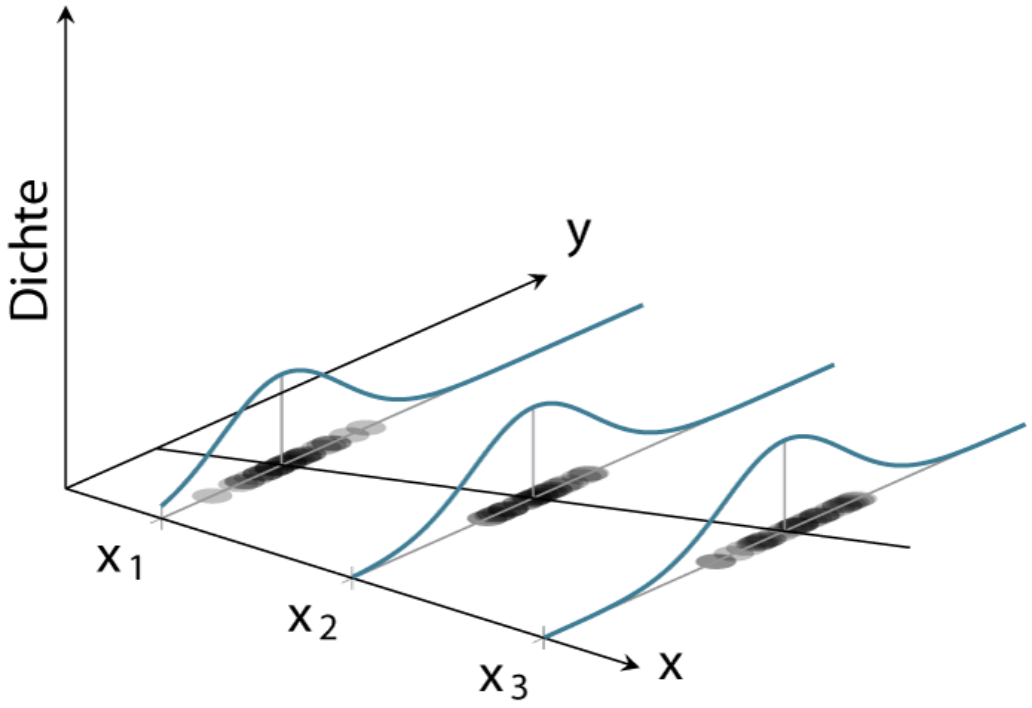


Figure 8.2: Verteilung der Daten für verschiedene x-Werte

In Figure 8.2 ist klar zu sehen, wie für jeden der drei Punkte von X die beobachteten Y-Werte einer Normalverteilung. Die Breite der Verteilung ist an jedem Punkte gleich, nämlich $= \sigma^2$ während der Mittelwert der Gleichung $\beta_0 + \beta_1 X$ folgend entlang der Regressionsgerade verschoben ist.

Wenn wir uns zurück an die Ausführungen zur statistischen Signifikanz erinnern, dann haben wir in dem Zusammenhang vom einem datengenerierenden Prozess gesprochen (Definition 2.1) (DGP). In unserem jetzigen Modell können wir dementsprechend zwei Komponenten als Teile des DGP identizifieren. Entsprechend Equation 8.1 besteht der DGP aus dem deterministischen Teil $\beta_0 + \beta_1 X$ und dem stochastischen Teil $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Diese Einsicht können wir verwenden um die Eigenschaften dieses Modells bezüglich Aussagen über statistische Signifikanz zu untersuchen.

Wir fokussieren uns jetzt auf ein vereinfachtes Modell bei dem wir zusätzlich noch $\beta_0 = 0$ setzen, und wir uns erst mal nur für die Eigenschaften von β_1 interessieren. Gehen wir nun davon aus, dass zwischen X und Y der Zusammenhang $\beta_1 = 1$ besteht. D.h. wenn X um eine Einheit vergrößert wird, dann wird Y ebenfalls um eine Einheit größer.

$$Y = 0 + 1 \cdot X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{8.2}$$

Jetzt müssen wir noch einen Wert für σ^2 festlegen. Sei dieser einfach einmal $\sigma = \frac{1}{2}$. Jetzt können wir `R` benutzen um *Experimente*, also Beobachtungen, anhand dieses DGP zu simulieren. Der Einfachheit halber legen wir ein übersichtliches $N = 12$ fest und nehmen uns jeweils drei X-Werte z.B. mit $X \in \{-1, 0, 1\}$, d.h. wir ziehen für jeden X-Wert vier Y-Werte.

```
N <- 12
beta_0 <- 0
beta_1 <- 1
sigma <- 1/2
dat_sim_1 <- tibble(
  x_i = rep(-1:1, each=4),
  y_i = beta_0 + beta_1 * x_i + rnorm(N, mean = 0, sd = sigma)
)
```

Wenn wir uns die generierten Daten anschauen, dann sehen wir wenig überraschend 12 verschiedene Werte für y_i und jeweils 3×4 verschiedene Werte für x_i (siehe Table 8.1).

Table 8.1: Eine Simulation des Modells Equation 8.2

x_i	y_i
-1	-1.01
-1	-0.92
-1	-0.96

x_i	y_i
-1	-1.75
0	-0.11
0	0.76
0	1.21
0	0.03
1	-0.24
1	2.04
1	1.16
1	1.45

Wenn wir die Daten graphisch darstellen erhalten wir (Figure 8.3):

```
ggplot(dat_sim_1, aes(x_i, y_i)) +
  geom_point()
```

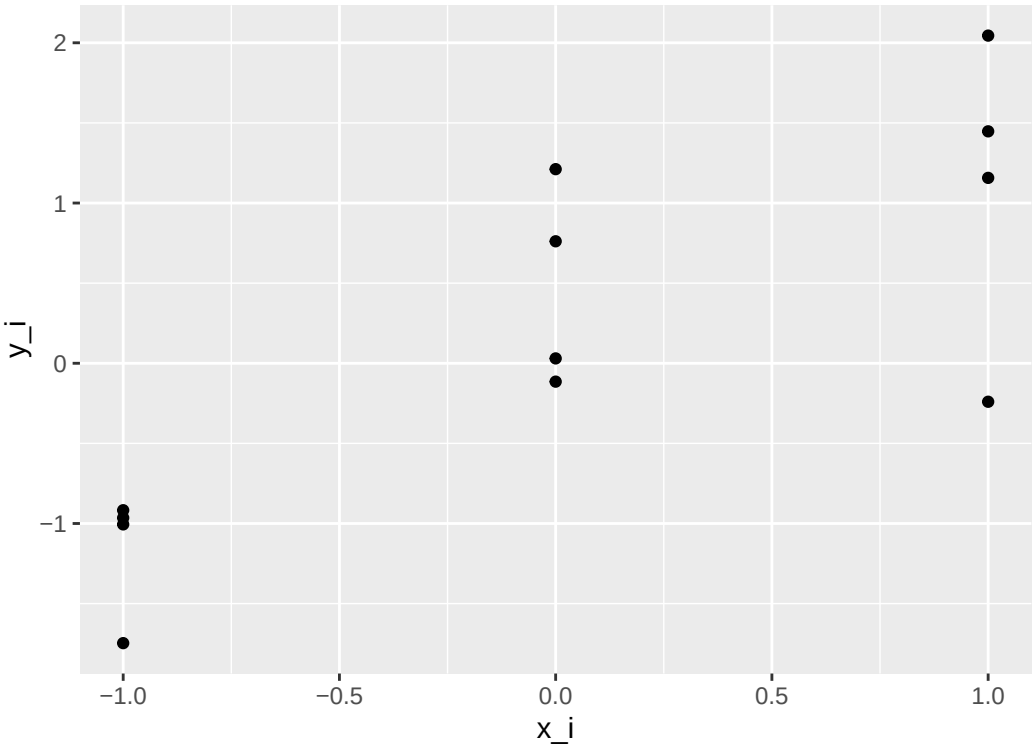


Figure 8.3: Streudiagramm der Daten aus Table 8.1

Ebenfalls wenig überraschend, die Punkte sind auf den x-Werten $-1, 0$ und 1 zentriert und liegen nicht alle aufeinander, da sie einer Zufallsstichprobe aus $\mathcal{N}(0, \frac{1}{4})$ entspringen.

Jetzt kann ich natürlich für diese Daten unsere Normalengleichungen anwenden und Werte für $\widehat{\beta}_0$ und $\widehat{\beta}_1$ berechnen. Oder eben direkt in `R`.

```
mod_sim_1 <- lm(y_i ~ x_i, dat_sim_1)
coef(mod_sim_1)
```

```
(Intercept)         x_i
  0.1386259   1.1304322
```

Wir sehen, dass die berechneten Werte für β_0 und β_1 schon in der Nähe der tatsächlichen Werte liegen (siehe **?@eq-slm-inf-mod-1**), aber auf Grund der Stichprobenvariabilität eben nicht genau auf diesen Werten. Was passiert denn jetzt, wenn ich das Ganze noch einmal durchlaufen lassen?

```
dat_sim_2 <- tibble(
  x_i = rep(-1:1, each=4),
  y_i = beta_0 + beta_1 * x_i + rnorm(N, mean = 0, sd = sigma)
)
mod_sim_2 <- lm(y_i ~ x_i, dat_sim_2)
coef(mod_sim_2)
```

```
(Intercept)         x_i
 0.09666798  0.73083795
```

Wieder wenig überraschend, da jedes Mal wenn ich `rnom()` eine neue Ziehung aus der Normalverteilung generiert wird, erhalte ich neue Werte für y_i und dementsprechend andere Werte für $\widehat{\beta}_0$ und $\widehat{\beta}_1$. Nochmal, warum? **Stichprobenvariabilität**! Jetzt sind wir wieder bei dem gleichen Prinzip, das wir im Rahmen der kleinen Welt ausgiebig behandelt haben. Schauen wir uns jetzt doch einfach mal was passiert wenn wir die Simulation nicht $2\times$ sondern z.B. $1000\times$ durchführen.

```
N_sim <- 1000
beta_1_s <- numeric(N_sim)
x_i <- rep(-1:1, each=4)
for (i in 1:N_sim) {
  daten_temporaer <- tibble(x_i,
                            y_i = beta_0 + beta_1 * x_i + rnorm(N, mean = 0, sd = sigma))
  model_temporaer <- lm(y_i ~ x_i, daten_temporaer)
```

```
  beta_1_s[i] <- coef(model_temporaer)[2]
}
```

Wir erhalten jetzt einen Vektor `beta_1_s` mit 1000 beobachteten $\hat{\beta}_1$. Da das etwas viele Werte sind um die uns einzeln anzuschauen, erstellen ein Histogramm der $\hat{\beta}_1$s. (Figure 8.4).

```
hist(beta_1_s, xlab = expression(hat(beta)[1]), main='')
abline(v = beta_1, col='red', lty=2)
```

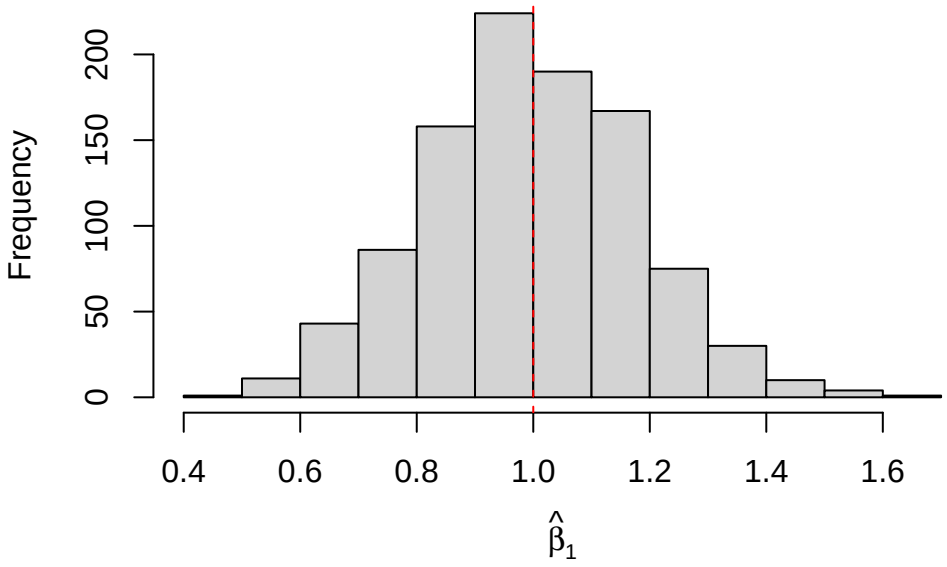


Figure 8.4: Histogram der auf den simulierten Daten berechneten $\hat{\beta}_1$. Wahrer Wert von β_1 rot eingezeichnet.

In Figure 8.4 begegnet uns zunächst einmal wieder unsere altbekannte Glockenkurve. Schön ist, dass deren Mittelwert im Bereich des wahren Werts von β_1 liegt und Werte mit größer werdender Abweichung vom wahren Wert in ihrer Häufigkeit abnehmen. Aber die Häufigkeit ist nicht Null, sondern eben nur geringer. Werte in der Nähe von β_1 weisen dagegen eine größere Häufigkeit aufweisen. Das sollte uns jetzt auch irgendwie zufrieden stimmen, denn dies bedeutet, dass wir in der Lage sind mit unserem Regressionsmodell im Mittel tatsächlich den korrekten Wert abzuschätzen. Allerdings, wie immer, bei einer einzelnen Durchführung

des Experiments können wir alles von perfekt spot-on bis komplett danebenliegen und würden es nicht wissen.

Wir können jetzt aber auch wieder ganz parallel zu unseren Herleitungen in der kleinen Welt einen Entscheidungsprozess spezifizieren. Wenn Figure 8.4 den DGP beschreibt und das die Verteilung der zu erwartenden $\hat{\beta}_1$ unter dem Modell sind. Bei der Dürchführung eines neuen Experiments, dann würden wir sagen, dass wenn unserer beobachteter Wert in den Rändern der Verteilung von Figure 8.4 liegt, das wir eher nicht davon ausgehen, dass unserer neues Experiment den gleichen DGP zugrundeliegen hat. D.h wir definieren uns jetzt Grenzen am oberen und am unteren Rand der Verteilung. Wenn jetzt ein neuer beobachteter Wert entweder unterhalb der unteren Grenze oder oberhalb der oberen Grenze liegt, dann sagen wir: *Wir sind jetzt aber sehr überrascht diesen Wert zu sehen, wenn der dem gleichen datengenerierenden Prozess entstammen soll. Daher glauben wir nicht, dass dieses Experiment den gleichen DGP besitzt.*

Um diese Entscheidung treffen zu können, müssen wir also Grenzen definieren. Dazu können wir zunächst einmal einfach die Quantilen der Verteilung nehmen und schneiden z.B. unten 2.5% und oben 2.5% ab. So kommen wir dann insgesamt auf 5%, um auf die übliche Irrtumswahrscheinlichkeit von $\alpha = 0.05$ zu kommen. Dazu benutzen wir `R` und zwar `quantile()`-Funktion^[Im folgenden Snippet werden die Werte auf zwei Kommastellen mit `round()` der besseren Darstellung wegen gerundet).

```
2.5% 97.5%
0.65  1.35
```

Mittels dieser Werte können wir zwei disjunkte Wertmenge definieren, einmal die Werte innerhalb von $\hat{\beta}_1 \in [0.65, 1.35]$ bei denen wir nicht überrascht sind, und die unter der Annahme $\beta_1 = 1$ erwartbar sind und die Werte $\hat{\beta}_1 \notin [0.65, 1.35]$ diejenigen Werte die uns überraschen würden unter der Annahme. Ins Histogramm übertragen (siehe Figure 8.5).

Führen wir nun ein Experiment noch einmal durch. Wir beobachten einen Wert für $\hat{\beta}_1$ von 1.46. Dieser Wert liegt außerhalb unseres definierten Intervalls $[0.65, 1.35]$, daher sehen wir diesen Wert als derart unwahrscheinlich unter dem angenommenen DGP, das wir sagen: *Wir glauben nicht, dass diesem Experiment nicht der angenommene DGP zugrunde liegt.* Graphisch wieder dargestellt (siehe Figure 8.6).

Daher würden wir diesen Wert als statisisch signifikant bezeichnen und würden unsere Annahme ablehnen.

Jetzt sind wir aber etwas hin und her zwischen Experiment, Annahmen und Schlussfolgerungen gesprungen. Normalerweise kennen wir die Stichprobenverteilung nicht vor dem Experiment, sondern, wir sind am dem Wert β_1 interessiert. Wenn wir den Wert schon wissen würden, dann müssten wir ja gar kein Experiment mehr durchführen. D.h. wir haben eigentlich noch keinen klaren Vorkenntnisse. Mit welcher Annahme gehen wir dann in das Experiment rein? Nun,

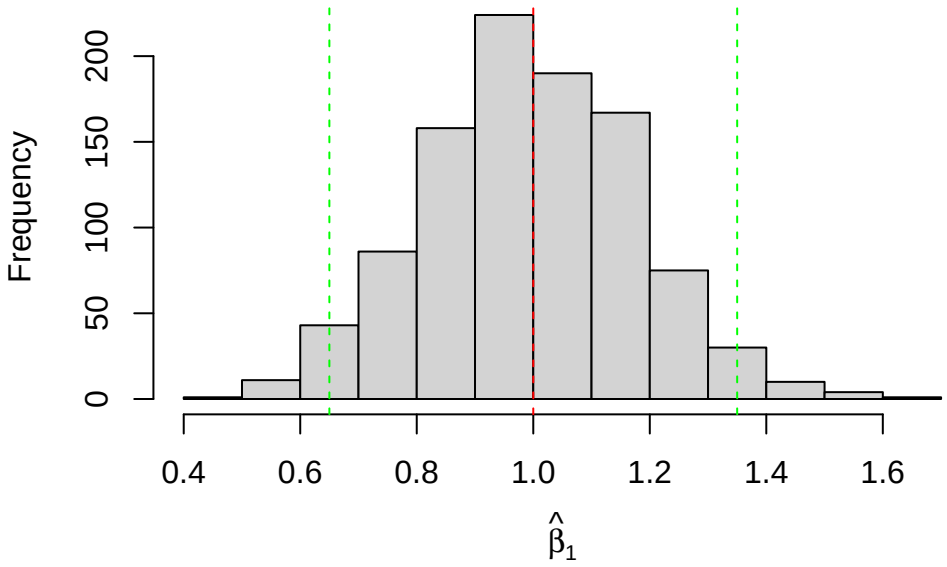


Figure 8.5: Histogram der auf den simulierten Daten berechneten $\hat{\beta}_1$. Wahrer Wert von β_1 rot eingezeichnet und kritische Werte grün.

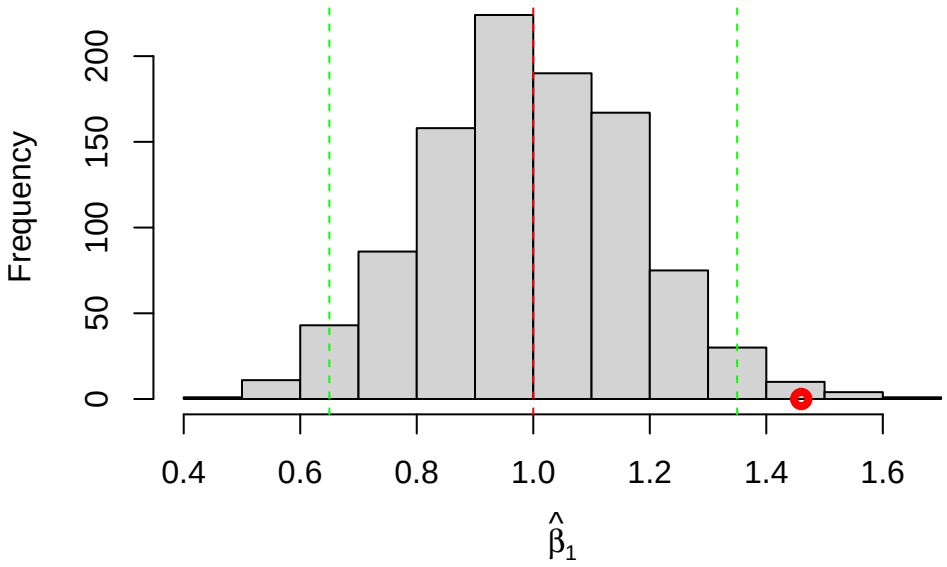


Figure 8.6: Histogram der auf den simulierten Daten berechneten $\hat{\beta}_1$. Wahrer Wert von β_1 rot eingezeichnet und kritische Werte grün und der beobachtete Wert als roter Punkt.

wir schon bei kleinen Welt Beispiel, starten wir mit der Annahme das zwischen den beiden Variablen kein Zusammenhang besteht. Übertragen auf die Modellparameter also, dass kein linearer Zusammenhang zwischen den beiden Variablen besteht.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Um die Stichprobenverteilung unter der H_0 formal Herleitung zu können, ist der Erwartungswert von $\hat{\beta}_1$ und dessen Standardfehler notwendig. Es lässt sich zeigen, dass die folgenden Zusammenhänge unter den gesetzten Annahmen bestehen:

$$E[\hat{\beta}_0] = \beta_0$$

Also der Schätzer von β_1 ist erwartungstreu (biased) und der Standardfehler des Schätzer lässt sich wie folgt bestimmen.

$$\sigma_{\beta_1} = \sqrt{\frac{\sigma^2}{\sum (X_i - \bar{X})^2}} \tag{8.3}$$

Hier taucht jetzt zum ersten Mal der Parameter σ^2 formal auf. Wo kommt diese Variance her? Sie gehört zu unserer Annahme der Verteilung der $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Bisher haben wir aber noch gar keine Möglichkeit kennen gelerntm, diese abzuschätzen. Wieder nach etwas motivierten Starren auf die verschiedenen Formeln, könnte heuristisch plausibel sein, dass die Varianz, also die Streuung der ϵ_i mit der Streuung unserer Werte um die Regressionsgerade zusammenhängen könnten. Formal hatten wir diese als Residuen bezeichnet und mit $e_i = \hat{y}_i - y_i$ bezeichnet. Vormals hatten wir diese Abweichungen als Fehler bezeichnet, aber unter den jetzt eingeführten Annahmen, handelt es sich nicht wirklich um Fehler, sondern die Abweichungen sind eine Folge davon, dass Y_i für jeden Wert von X_i nicht nur einen einzigen Wert hat, sondern eben einer Verteilung folgt $Y_i|X_i \sim \mathcal{N}(\beta_0 + beta_1, \sigma^2)$ deren Form über die ϵ_i bestimmt wird.

Die e_i sind tatsächlich die Schätzer für die *wahren* ϵ_i also $e_i = \hat{\epsilon_i} = \hat{y}_i - y_i$. Es lässt sich nun wieder zeigen, dass mittels dieser e_i ein erwartungstreuer Schätzer für σ^2 erzeugen lässt. Nämlich die mittleren quadrierten Abweichungen (MSE).

$$\hat{\sigma} = \frac{\sqrt{\sum_{i=1}^{N} e_i^2}}{N-2} = \frac{\text{SSE}}{N-2} = \text{MSE} \tag{8.4}$$

Da das später immer wieder auftauchen wird, hier auch noch mal in die zwei Komponenten zerlegt. Der Zähler wird als Summe der quadrierten Abweichungen (SSE) bezeichnet und

durch den Term $N-2$, der als Freiheitsgerade bezeichnet wird, geteilt. Dann mit die Formel und deren Bezeichnung *mittlere* Abweichung zusammenpasst, wäre es schöner wenn die Summe durch die Anzahl N der Terme geteilt wird, allerdings verhält sich das in diesem Fall ähnlich wie bei der Varianz einer Stichprobe wo die Summe auch durch $N-1$ geteilt wird (zur Erinnerung $s = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}$). Jetzt wird dementsprechend nicht durch $N-1$ sondern durch $N-2$ geteilt.

Für unser Problem der Stichprobenverteilung ist jetzt aber wichtiger, dass wir mittels Equation 8.4 den Standardfehler von $\hat{\beta}_1$ bestimmen können, indem wir für σ^2 das mittels der Daten ermittelte $\hat{\sigma}^2$ einsetzen.

$$\hat{\sigma}_{\beta_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}} \tag{8.5}$$

Dies erlaubt uns jetzt nach unserem bereits bekannten Muster eine Teststatistik für die H_0 herzuleiten:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}}$$

Unter der H_0 mit $\beta_1 = 0$ wird daraus

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}} \tag{8.6}$$

Diese Teststatistik folgt einer t-Verteilung mit $N-2$ Freiheitsgeraden. Da diese Formel wieder etwas aus der Luft gegriffen erscheint, hier noch mal eine Simulation zusammen mit der theoretischen Testverteilung.

```
N <- 45
n_sim <- 1000
x <- runif(N, -1, 1)
sigma <- 1
experiment <- function() {
  y <- rnorm(N, mean = 0, sd = sigma)
  mod <- lm(y~x)
  b <- coef(mod)[2]
  c(beta_0 = coef(mod)[1],
    beta_1 = coef(mod)[2],
    sigma = sigma(mod))
}
betas <- t(replicate(n_sim, experiment()))
```

```r
betas <- tibble(beta_0 = betas[,1],
                beta_1 = betas[,2],
                sigma = betas[,3]) |>
  mutate(
   s_e_beta_1 = sqrt(sigma**2/sum( (x - mean(x))**2)),
   t = beta_1 / s_e_beta_1)
t_theoretical <- tibble(
  t = seq(-3, 3, length.out = 150),
  p = dt(t, N - 2)
)

ggplot(betas, aes(t)) +
  geom_histogram(aes(y = ..density..), bins = 20) +
  geom_line(data = t_theoretical, aes(t, p), color = 'red') +
  labs(x = "t", y = 'Relative Häufigkeit')
```

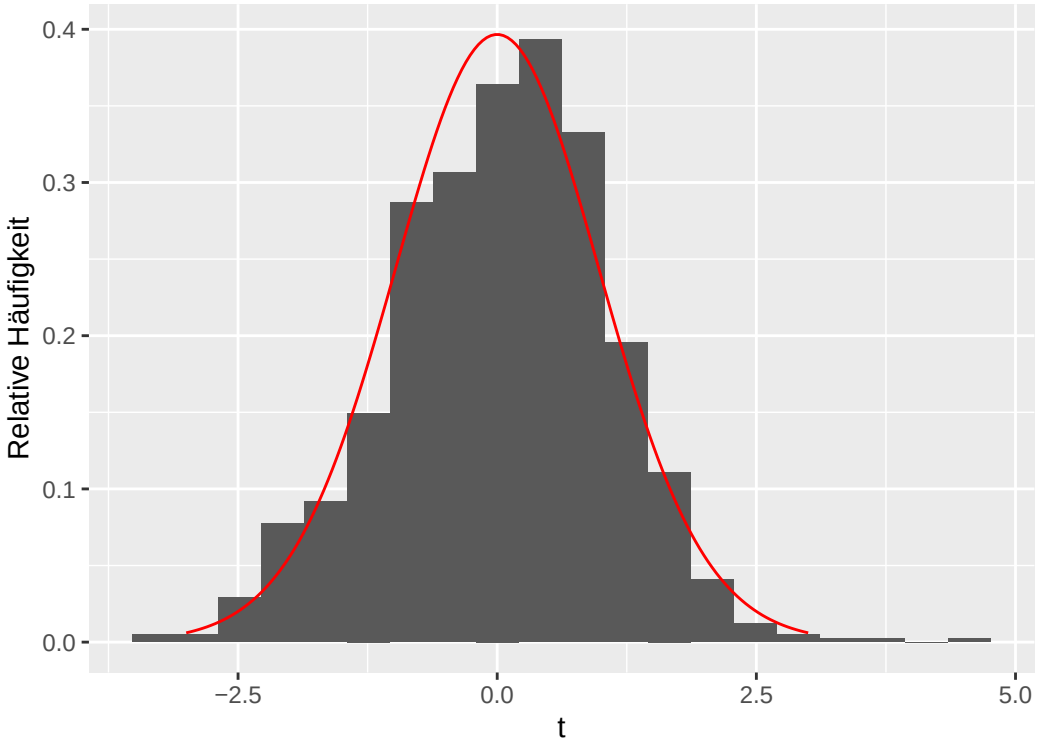


Figure 8.7: Verteilung von t bei 1000 Simulationen unter der Annahme der H_0 und die theoretische Verteilung von t (rot).

In Figure 8.7 können wir sehen, dass die theoretische Verteilung in rot die beobachtete Verteilung sehr gut abschätzt.

In `R` kann der Wert $\hat{\sigma}^2$ über die Funktion `sigma()` aus dem gefitteten `lm()`-Modell extrahiert werden.

```
sigma(mod)
```

```
[1] 0.2369055
```

Schauen wir uns die Stichprobenverteilung von $\hat{\sigma}^2$ anhand unserer Simulation an. Es ist wieder zu beobachten, das im Mittel der korrekte, im Modell definierte, Wert von $\sigma = 1$ beobachtet wird (siehe Figure 8.8).

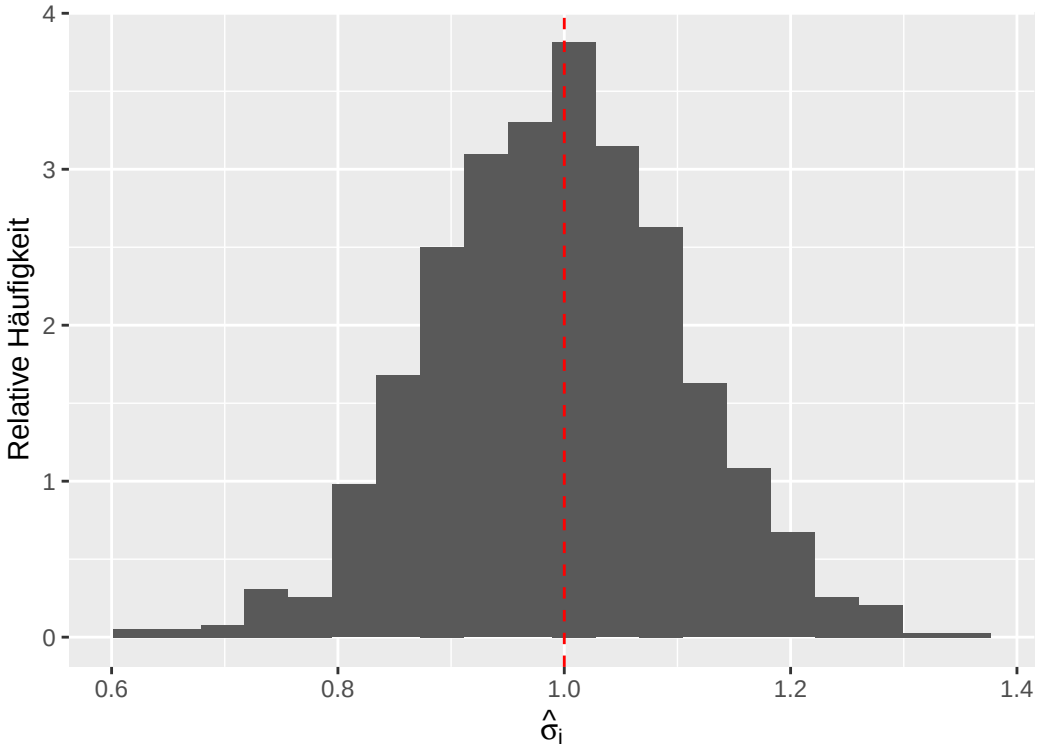


Figure 8.8: Verteilung von $\hat{\sigma}$ in der Simulation und der wahre Wert in rot eingezeichnet

Aber wie immer, leider steht uns bei einem realen Experiment diese Information nicht zur Verfügung und wir haben nur einen einzelnen Wert, der alles von komplett daneben bis ziemlich perfekt sein kann.

Schauen wir uns noch einmal die Ausgabe zu unserem Weitsprungmodell mittels `summary()` an. Unter Residual Standard Error sehen wir, dass hier $\hat{\sigma}$ angegeben wird. Dieser Wert wird auch als mittlerer Schätzfehler bezeichnet und kann als Maß verwendet werden, welche Abweichung das Modell im Mittel hat. Die Einheit sind wieder in den Einheiten der abhängigen Variable, so kann auch schon abgeschätzt werden mit welcher Präzision das Modell die Daten fittet.

```
summary(mod)
```

```
Call:
lm(formula = jump_m ~ v_ms, data = jump)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44314 -0.22564  0.02678  0.19638  0.42148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13854    0.23261  -0.596    0.555
v_ms         0.76110    0.02479  30.702   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2369 on 43 degrees of freedom
Multiple R-squared:  0.9564,    Adjusted R-squared:  0.9554
F-statistic: 942.6 on 1 and 43 DF,  p-value: < 2.2e-16
```

In unserem Fall beobachten wir $0.24m$. Diesen Wert muss jetzt unsere Trainerin im Sinne der Weitsprungleistung der deren Varianz interpretieren und ein Abschätzung treffen zu können.

Nach der Herleitung der Teststatistik für β_1, können wir jetzt auch weitere Teil der Ausgabe von `summary()` interpretieren. In der Tabelle stehen entsprechend die Standardfehler für $\hat{\beta}_1$ und $\hat{\beta}_0$. Für β_0 wird genau die gleiche Vorgehensweise wie auch bei β_1 angewendet. Die Nullhypothese H_0 ist hier ebenfalls das der Parameter standardmäßig als Null angesetzt wird. Der Standardfehler von β_0 errechnet sich nach:

$$\sigma^2[\beta_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \tag{8.7}$$

An Formel (**??**) ist zu erkennen, dass wenn die X-Werte den Mittelwert 0 haben, dass $\sigma^2[\beta_0]$ gleich dem Standardfehler für den Mittelwert SEM wird. Was auch wiederum Sinn macht, da in diesem Fall $\beta_0 = \bar{y}$ gilt.

Dies führt dies zu den beiden zu überprüfenden Hypothesen für β_0:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

Dementsprechend überprüft die Hypothesentestung ob der y-Achsenabschnitt gleich Null ist. Hier sollte berücksichtigt werden, dass diese Hypothese in den seltensten Fällen tatsächlich auch von Interesse ist und lediglich besagt, dass entweder der y-Achsenabschnitt durch den Nullpunkt geht, oder dass wenn tatsächlich $\beta_1 = 0$ gilt, der Mittelwert von y gleich Null ist, was ebenfalls in den seltensten Fällen von Interesse ist.

Die Spalten 3 und 4 in `summary()` unter `Coefficients:` können jetzt interpretiert werden, da es sich hierbei um die t-Teststatistik handelt und den entsprechenden p-Wert unter der jeweiligen H_0. Die Hypothesen sind ungerichtet.

8.2 Herleitung der Eigenschaften von $\hat{\beta}_1$

Um den Schätzer $\hat{\beta}_1$ für β_1 formal herzuleiten. Beginnen wir zunächst mit der folgenden Formel, wobei wir im folgenden den Schätzer mit b_1 bezeichnen.

$$b_1 = \sum k_i Y_i \tag{8.8}$$

D.h. wir zeigen zunächst, dass b_1 durch eine lineare Kombination der Y_i-Werte berechnet werden kann. Die Koeffizienten k_i der Summe sind dabei wie folgt definiert:

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \tag{8.9}$$

Der Grund für diese zunächst etwas uneinsichtige Definition wird im weiteren klarer werden. Zunächst haben die k_i verschieldene Eigenschaften die wir uns im Späteren zunutze machen wollen. Zunächst erst einmal noch ein paar Identitäten die wir später auch noch verwenden.

Die erste Identität bezieht sich auf das Kreuzprodukt der Abweichungen von X_i und Y_i von ihren jeweiligen Mittelwerten.

$$\begin{aligned}\sum (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum (X_i - \bar{X}) Y_i - \underbrace{\sum (X_i - \bar{X})}_{=0} \bar{Y} \\ &= \sum (X_i - \bar{X}) Y_i\end{aligned}$$

Wenn wir in der Formel $(Y_i - \bar{Y})$ durch $(X_i - \bar{X})$ austauschen, folgt noch eine weitere nützliche Identität:

$$\sum (X_i - \bar{X})^2 = \sum (X_i - \bar{X}) X_i$$

Werden die jeweiligen k_i mit den dazugehörigen X_i multipliziert und die Definition der k_i (siehe Equation 8.9) beachten, erhalten wir:

$$\sum k_i X_i = \frac{\sum (X_i - \bar{X}) X_i}{\sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} = 1$$

D.h. Die Summe der $k_i X_i$ ist gleich 1. Aus der Definition Equation 8.9 folgt weiterhin.

$$\sum k_i = \sum \left(\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right) = \frac{\sum (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \frac{0}{\sum (X_i - \bar{X})^2} = 0$$

D.h. die Summe der k_i ist gleich Null.

Wenn wir jetzt wieder die Definition unseres Schätzer für β_1 verwenden (siehe **?@eq-slm-basics-norm1**). Dann erhalten unter der Verwendung der Identität der Kreuzprodukte den gewünschten Zusammenhang zwischen b_1 und Y_i.

$$\begin{aligned} b_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \sum k_i Y_i \end{aligned}$$

Wenden wir jetzt den Erwartungswert auf Y_i an, dann werden die k_i als konstant angesehen und nur die Y_i sind Zufallsvariablen. Da aber $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ gilt und in dieser Formel wiederum nur ϵ_i eine Zufallsvariable mit β_0 und $\beta_1 X_i$ konstant ist und zudem die $\epsilon \sim \mathcal{N}(0, \sigma^2)$ also $E[\epsilon_i] = 0$ laut der Annahme gilt, folgt:

$$\begin{aligned} E[b_1] &= E\left[\sum k_i Y_i\right] = \sum k_i E[Y_i] = \sum k_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum k_i + \beta_1 \sum k_i X_i = \beta_1 \end{aligned}$$

D.h. **?@eq-slm-basics-norm1** ist ein erwartungstreuer Schätzer für β_1. Das gleiche gilt auch für den Schätzer b_0 für β_0.

Leiten wir noch eine weitere Identität über die Summe der k_i^2 her:

$$\sum k_i^2 = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 = \frac{\sum (X_i - \bar{X})^2}{\left[\sum (X_i - \bar{X})^2 \right]^2} = \frac{1}{\sum (X_i - \bar{X})^2}$$

Können wir auch noch die Varianz bzw. den Standardfehler unseres Schätzers für β_1 herleiten. Es gilt nämlich:

$$\begin{aligned} \sigma^2[b_1] &= \sigma^2 \left[\sum k_i Y_i \right] = \sum k_i^2 \sigma^2[Y_i] \\ &= \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2 \\ &= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2} \end{aligned}$$

Wir erhalten die bereits eingeführte Formel. Wiederum eine Einsicht aus der Herleitung der Formel folgt, dass die Varianz σ^2 als konstant angesehen wird, d.h. $\sigma_i^2 = \sigma^2$. Dies hat uns erlaubt im zweiten Schritt σ^2 aus der Summe heraus zu ziehen. Wenn die Varianz σ^2 nicht konstant ist, dann ist der berechnete Standardfehler für $\hat{\beta}_1 = b_1$ nicht korrekt.

Eine interessante Eigenschaft des Standardfehler von $\hat{\beta}_1$ ist in Formel (**??**) zu sehen. Im Nenner stehen die Abweichungen der X-Werte vom Mittelwert $\hat{X}$. D.h. wenn die X-Werte weiter auseinander sind, dann führt dies dazu, dass der Standardfehler $\sigma^2[b_1]$ kleiner wird. Intuitive macht dies auch Sinn, wenn ich eine Gerade bestimmen will, dann ist es einfacher die Gerade anhand weit auseinander liegenden Stütztwerten zu bestimmen im Vergleich zu wenn ich eng beinander liegende X-Werte verwende.

8.3 Maximum-likelihood Methode bei der einfachen linearen Regression

Ein anderer Herleitung für β_0 und β_1 kann über die sogenannten Maximum Likelihood durchgeführt werden. Dabei gehen direkt die Verteilungsannahmen direkt ein.

Für eine gegebene Zufallsvariable die jeweilige Dichte eines gegebenen Wertes über die Dichtefunktion berechnet werden. Wenn ein Zufallsvariable X einer Normalverteilung folgt, dann wird die Verteilung von X nach der bereits kennengelernte Dichtefunktion der Normalverteilung beschrieben.

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

Hier wird die Dichte von X als eine Funktion von μ und σ^2 aufgefasst. Es ist aber auch möglich, die Zufallsvariable X als gegeben anzusehen und die Dichte für verschiedene Werte von μ und σ^2 abzutragen. Der Einfachheit halber gehen wir davon aus, dass σ^2 gegeben sei und wir μ nicht kennen. Eine mögliche Fragestellung ist jetzt, für einen beobachteten Wert x, welcher Wert von μ ist am plausibelsten?

Tragen wir dazu verschiedene Dichtewerte für ein gegebenes x in Abhängigkeit von verschiedenen μ ab.

D.h. wir interpretieren die Funktion als:

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

Diese Funktion wird als die likelihood-Funktion bezeichnet. Das Maximum dieser Funktion kann als derjenige Wert interpretiert werden bei dem derjenige Wert von μ die maximal mögliche Dichte einnimmt.

Die Likelihood-Funktion ist eine Funktion, die die Wahrscheinlichkeit beschreibt, mit der eine gegebene Stichprobe, in Abhängikeit von den Parametern aus einer bestimmten Verteilung stammt. Die Likelihood-Funktion gibt also an, wie gut die beobachteten Daten zu einem bestimmten Satz von Parametern passen.

Formal wird die Likelihood-Funktion als die gemeinsame Wahrscheinlichkeitsdichte der Stichprobe beschrieben, betrachtet als Funktion der Parameter. Dabei werden die beobachteten Werte als festgelegt und die Parameter als Variablen betrachtet. Die Likelihood-Funktion ist also eine Funktion der Parameter, die die Wahrscheinlichkeit der beobachteten Daten als Funktion dieser Parameter beschreibt. Die Likelihood-Funktion ist dabei keine Dichtefunktion und beschreibt somit keine Wahrscheinlichkeiten. Dementsprechend ist gilt für das Integral der Likelihood-Funktion $\int L(\mu|X, \sigma^2)d\mu \neq 1$ bzw. ist $= 1$ per *Zufall*.

In unserem Regressionsfall nimmt die Likelihood-Funktion für einen einzelnen Wert die folgende Form an:

$$L(\beta_0, \beta_1, \sigma^2|y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

Bei unserer Regressionsanalyse haben wir jedoch nicht nur einen einzigen beobachteten Wert (y_i, x_i) sondern N beobachtete Werte. Da die Werte unabhängig voneinander sind (laut der Annahmen), werden die jeweiligen likelihoods miteinander multipliziert. Die resultierende Likelihood-Funktion nimmt dann die folgenden Form an:

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{N} f(y_i|x_i; \beta_0, \beta_1, \sigma^2) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(\sum_{i=1}^{N} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(\sum_{i=1}^{N} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)
\end{aligned}
$$

Die Idee ist jetzt wieder die Gleiche. Wir versuchen das Maximum dieser Funktion zu finden, da die Werte β_0, β_1 und σ^2 dann so gewählt sind, dass sie die höchste likelihood haben. Der Ansatz erfolgt wieder mechanisch ,indem wie bei der Herleitung der Normalengleichungen, die partiellen Ableitungen berechnet werden, diese gleich Null gesetzt werden und das resultierende Gleichungssystem gelöst wird. Zu beachten hierbei, wir haben in jedem Produktterm die gleichen Parameter $\beta_0, \beta - 1$ und σ^2 und die jeweiligen beobachteten (y_i, x_i) Tuple werden als gegeben angesehen.

Um die Berechnungn zu vereinfachen, bietet sich bei der Likelihoo-Funktion ein Trick an. Es wird nicht Likelihood-Funktion abgeleitet, sondern der Logarithmus der Likelihood-Funktion. D.h. die Funktion wird transformiert. Bei der Logarithmus-Funktion handelt es sich um eine sogenannte bijektive Funktion. Eine bijektive Funktion ist eine Funktion die jedem Element in der Ursprungsmenge genau ein Element in der Zielmenge zuordnet und ebenfalls umgekehrt. Dadurch kommt es zu keinen Kollisionen oder Auslassungen. Einfach gesagt, wenn die Funktion $y = f(x) = log(x)$ ist, dann wird jedem x genau ein y zugeordnet. Bzw. anders herum, wenn ich y kenne, dann kenne ich auch den Wert von x mit $f(x) = y$ bzw. $x = f^{-1}(y) = \exp(y)$. Dadurch, das die Logarithmus-Funktion bijektiv ist, führt dies dazu, dass das Maximum der ursprünglichen Funktion $L(\beta_0, \beta_1, \sigma^2)$ an der gleichen Stelle auftritt wie bei der transformierten Funktion $\ln L(\beta_0, \beta_1, \sigma^2)$.

Wenn jetzt die Eigenschaften der Logarithmusfunktion, speziell des natürlichen Logarithmus, beachtet werden, dann wird auch klar, warum es Sinn machen könnte die Likelihood-Funktion mit dem Logarithmus zu transformieren, da aus den Produkten Summen werden mit denen einfacher umgegangen werden kann:

$$
\begin{aligned}
\log(xy) &= \log(x) + \log(y) \\
\log\left(\frac{x}{y}\right) &= \log(x) - \log(y) \\
\log(x^n) &= n\log(x) \\
\log(\exp(x)) &= x \\
\log(1) &= 0
\end{aligned}
$$

Der Logarithmus angewendet auf $L(\beta_0, \beta_1, \sigma^2)$ resultiert dann in der folgenden Funktion:

$$
\begin{aligned}
\ell(\beta_0, \beta_1, \sigma^2) &= \ln L(\beta_0, \beta_1, \sigma^2) \\
&= \ln\left[\left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\sum_{i=1}^{N} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)\right] \\
&= \ln\left[\left(\frac{1}{2\pi\sigma^2}\right)^{N/2}\right] + \ln\left[\exp\left(-\sum_{i=1}^{N} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)\right] \\
&= \frac{N}{2}\ln\left[\left(\frac{1}{2\pi\sigma^2}\right)\right] - \sum_{i=1}^{N} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2
\end{aligned}
$$

Wir die Funktion $\ell(\beta_0, \beta_1, \sigma^2)$ wieder partiell nach β_0 und β_1 abgeleitet und gleich Null gesetzt erhalten wir das gleiche Gleichungssystem wie bei den vorhergehenden Herleitungen über die Abweichungen von der Regressionsgeraden. z.B.

$$
\begin{aligned}
\frac{\partial \ell(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \frac{2}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)
\end{aligned}
$$

Wenn dieser Ausdruck gleich Null gesetzt erhalten wir den gleichen Ausdruck wie unter

8.4 Konfidenzintervalle für die Koeffizienten

Wie wir im oberen Abschnitt gesehen haben, sind unsere Schätzer für die Koeffizienten β_0 und β_1 mit Unsicherheiten behaftet die sich in Form der Standardfehler ausdrücken. Wir können nun, diese standardfehler wiederum verwenden um Konfidenzintervalle für die Koeffizienten zu bestimmen.

$$\hat{\beta}_j \pm q_{t_{\alpha/2, df=N-2}} \times \hat{\sigma}_{\beta_j} \tag{8.10}$$

Wie in Formel(8.10) zu sehen berechnet sich das Konfidenzintervall nach dem üblichen Muster: Schätzer $\pm$ Quantile $\times$ Standardfehler. Im vorliegenden Falle wird die Quantile aus der t-Verteilung mit $N-2$ Freiheitsgarden bestimmt. Wie vorher bereits betont, das Konfidenzintervall erlaubt keine Aussage über die Wahrscheinlichkeit mit der der wahre Koeffizient in dem Intervall liegt, sondern gibt an welche H_0-Hypothesen mit den Daten kompatibel sind. Daher soll in der Ergebnisdokumentation das Konfidenzintervall angegeben und spätenstens in der Diskussion die obere und die untere Schranke diskutiert werden.

In `R` kann das Konfidenzintervall mit der Funktion `confint()` berechnet und ausgegeben werden.

```
confint(mod)
```

```
                2.5 %    97.5 %
(Intercept) -0.6076488 0.3305767
v_ms         0.7111082 0.8110957
```

Wie die Koeffizienten haben die Konfidenzintervall die gleiche Einheit wie die abhängige Variable und können daher direkt interpretiert werden. Im vorliegenden Fall sollte daher besprochen werden welche Bedeutung ein Koeffizient von $\beta_1 = 0.7$ bzw. von $\beta_1 = 0.8$ für die Interpretation des Modell hat.

Noch einmal zu erwähnen ist, dass die beiden Parameter $\hat{\beta}_0$ und $\hat{\beta}_1$ welche die Regressionsgerade beschreiben, Schätzer für die Parameter aus einer Population sind der die beiden Parameter β_0 und β_1 den zugrundeliegenden Zusammenhang zwischen den beiden Variablen beschreiben. Diese betrachtung ist parallel zu derjenigen, wenn wir z.B. anhand des Mittelwerts $\bar{x}$ den währenen Populationsmittelwert μ versuchen zu schätzen. D.h. wir haben eine Populationsregressionsgerade, die wir mit Hilfe der Daten versuchen zu schätzen. Die wahre Regressionsgerade werden wird aber niemals mit 100%-iger Sicherheit bestimmen, eben genausowenig wie wir den Populationsmittelwert μ nicht mittels $\bar{x}$ bestimmen können.

8.5 Weiteres Material

N. Altman and Krzywinski (2015b) und Kutner et al. (2005, 40–48)

9 Modellfit

Nachdem ein Modell mittels einer einfachen linearen Regression an die Daten gefittet wurde, sollte immer überprüft werden ob das Modell tatsächlich auch den Annahmen entspricht. Ein zentrales Mittel dazu ist eine Analyse der Residuen.

9.1 Residuen

Dazu schauen wir uns zunächst noch einmal an, was überhaupt Residuen e_i sind und gehen noch mal von den grundlegenden Modellannahmen aus (siehe Formel (9.1)).

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{9.1}$$

Das lineare Regressionsmodell geht von einem linearen Zusammenhang in den Koeffizienten zwischen der Variablen x_i und den Variablen y_i aus. Additiv kommt daz ein normalverteilter Fehler ϵ_i. Die Normalverteilung der ϵ_i habem einen Erwartungswert von $\mu = 0$ und eine Standardabweichung von σ. Die Standardabweichung σ ist zunächst unbekannt und muss über die Daten abgeschätzt werden. Dies führt dazu, dass y_i für jeden gegebenen Wert von x_i einer Normalverteilung mit $\mathcal{N}(\beta_0 + beta_1 x_i, \sigma^2)$ folgen und der bereits bekannten graphischen Darstellung (siehe Figure 9.1).

Für jeden gegebenen Wert von X sind die Y-Werte Normalverteilt. Die Varianz dieser Normalverteilungen ist gleich σ während der Mittelwert μ immer um den Wert der Regressionsgeraden verschoben ist. D.h. die Streuung von ϵ_i überträgt sich auf die Streuung von y_i für jeden gegebenen X-Wert. Ohne den zufälligen Einfluss der Fehlerwerte würden wir alle y_i-Werte perfekt auf der Regressionsgeraden erwarten. Dies deutet daher auch schon eine Möglichkeit an die Residuen ϵ_i mittels der Daten abzuschätzen. Man verwendet die Abweichungen der beobachteten Werten y_i von den vorhergesagten Werten $\hat{y}_i$ auf der Regressionsgeraden (siehe Formel (9.2)).

$$\hat{\epsilon}_i = e_i = y_i - \hat{y}_i \tag{9.2}$$

Diese Abweichungen e_i können als Schätzer $\hat{\epsilon}_i$ für die wahren Residuen ϵ_i verwendet werden (siehe Figure 9.2).

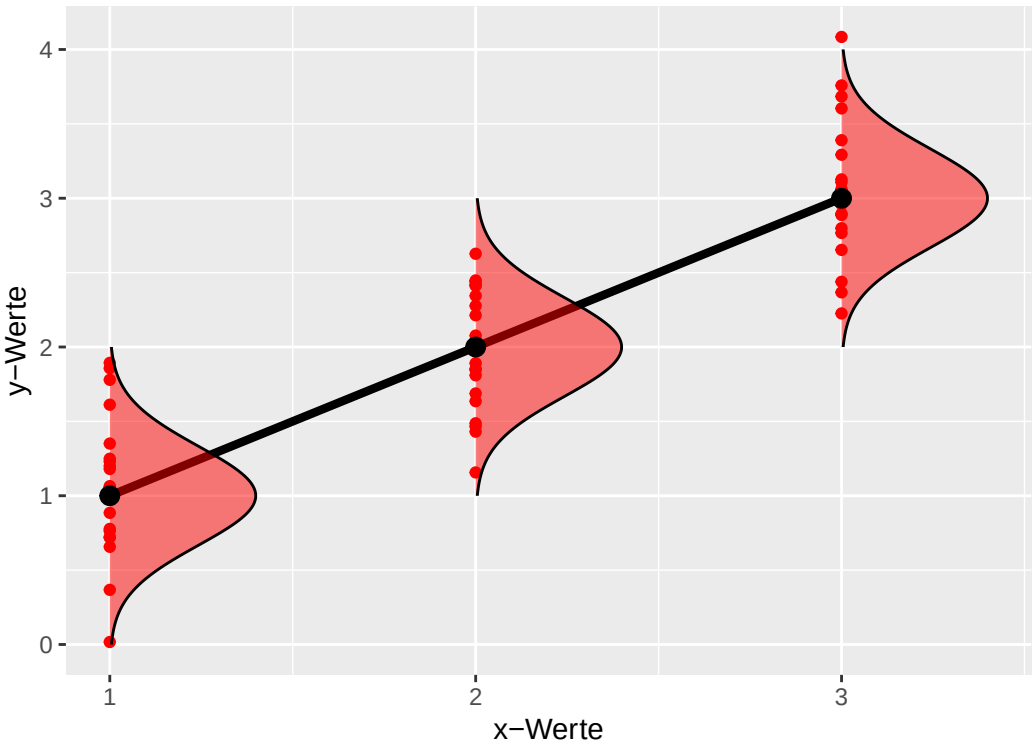


Figure 9.1: Beispiel einer Regressionsgeraden und der Verteilung der Residuen um den Vorhersagewert $\widehat{y}_i$

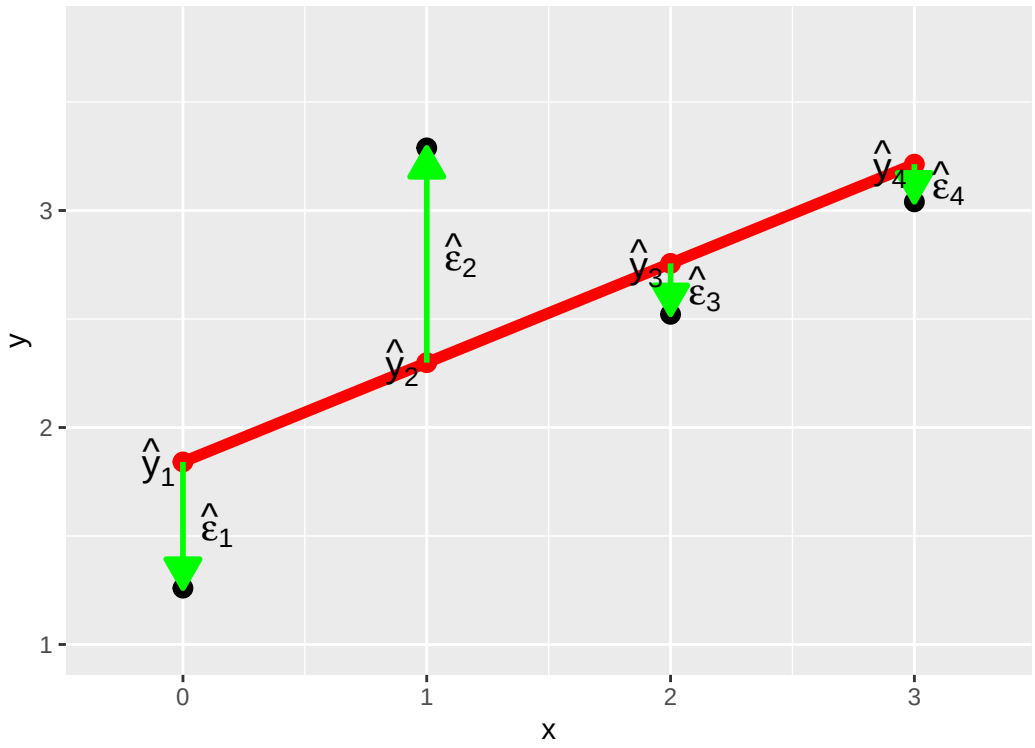


Figure 9.2: Examplarische Darstellung der Berechnung der Residuen e_i als Abweichung der beobachteten Werte y_i von den vorhergesagten Werten $\hat{y}_i$

Da die Normalverteilungen der ϵ_i für jeden X-Wert immer gleich sein sollten bis auf die Verschiebung von $\mu_{Y|X}$, deutet dies ebenfalls eine erste Möglichkeit an, die Modellannahmen graphisch zu überprüfen. Wenn die Residuen e_i geben die vorhergesagten Werte $\hat{y}_i$ abgetragen werden, dann sollte die Verteilung der Residuen e_i überall nahezu gleich sein, da die Streuung σ unabhängig von der *Position* auf der Regressionsgerade ist. In `R` können die Residuen mittels der Funktion `residuals()` bzw. der Kurzform `resid()` ermittelt werden. `residuals()` erwartet als Parameter das gefittete `lm()`-Objekt.

```
residuals(mod)
```

```
         1          2          3          4          5          6 
-9.3009275 -9.3682884 -11.2176585 -5.5721082 -6.3635647 -7.4162019 
         7          8          9         10         11         12 
-3.9665569 -8.7152962 -3.8032898 -0.4662810 -2.0491941 -2.1323841 
        13         14         15         16         17         18 
 0.1867102 -0.3382894 -2.7300208 -4.0317532 -6.1475804 -0.3782884 
        19         20         21         22         23         24 
 1.1267111 -0.4588401 -2.0417532 -2.5546673 -0.2276585  2.3352546 
        25         26         27         28         29         30 
-3.2075794  2.7949787  2.9982458  2.4379709  1.1162385  3.1894284 
        31         32         33         34         35         36 
 7.8049787 -0.9063196  4.0336013 11.3778918  7.6817897  8.6991516 
        37         38         39         40         41 
12.3052546  7.2595065 20.9634431 -1.3171838 -1.5994700 
```

Die anhand des Modells vorhergesagten Werte $\hat{y}_i$ werden der Funktion `predict()` berechnet. Diese Funktion werden wir uns im nächsten Kapitel noch ausführlich betrachten. Als Parameter wird wiederum das gefittete `lm()`-Modell übergeben.

```
predict(mod)
```

```
       1        2        3        4        5        6        7        8 
13.35093 14.39829 16.24766 13.61211 14.40356 15.45620 12.02656 16.77530 
       9       10       11       12       13       14       15       16 
12.82329 11.49628 13.07919 14.14238 12.82329 13.34829 15.72002 17.04175 
      17       18       19       20       21       22       23       24 
19.15758 14.39829 13.87329 15.45884 17.04175 17.57467 16.24766 14.66475 
      25       26       27       28       29       30       31       32 
20.20758 15.19502 15.99175 17.57203 18.89376 17.83057 15.19502 23.90632 
      33       34       35       36       37       38       39       40 
19.94640 13.61211 17.30821 17.31085 14.66475 20.74049 12.02656 13.32718 
```

```
      41
13.60947
```

Beide Funktionen, `resid()` und `predict()` geben die berechneten Werte in der Reihenfolge aus, in der die Originaldaten an `lm()` übergeben wurde. D.h. e_1 und $\hat{y}_1$ gehören zum ersten X-Wert x_1 aus den Originaldaten. Mit Hilfe dieser beiden Variablen kann nun ein Residuenplot erstellt werden (siehe Figure 9.3).

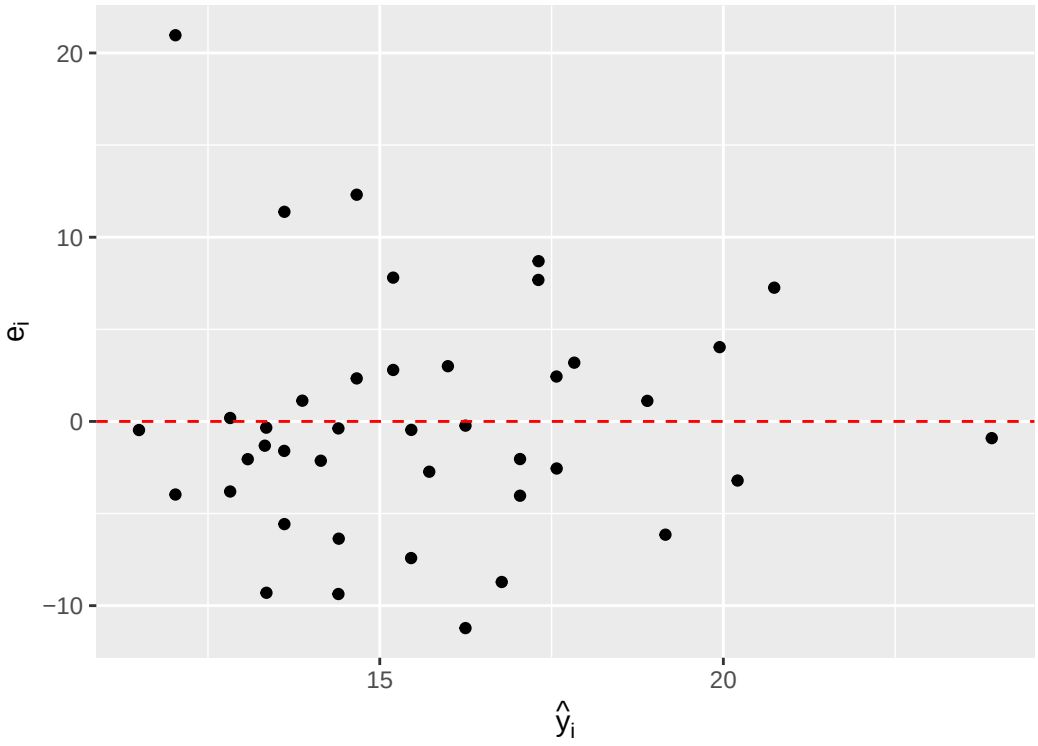
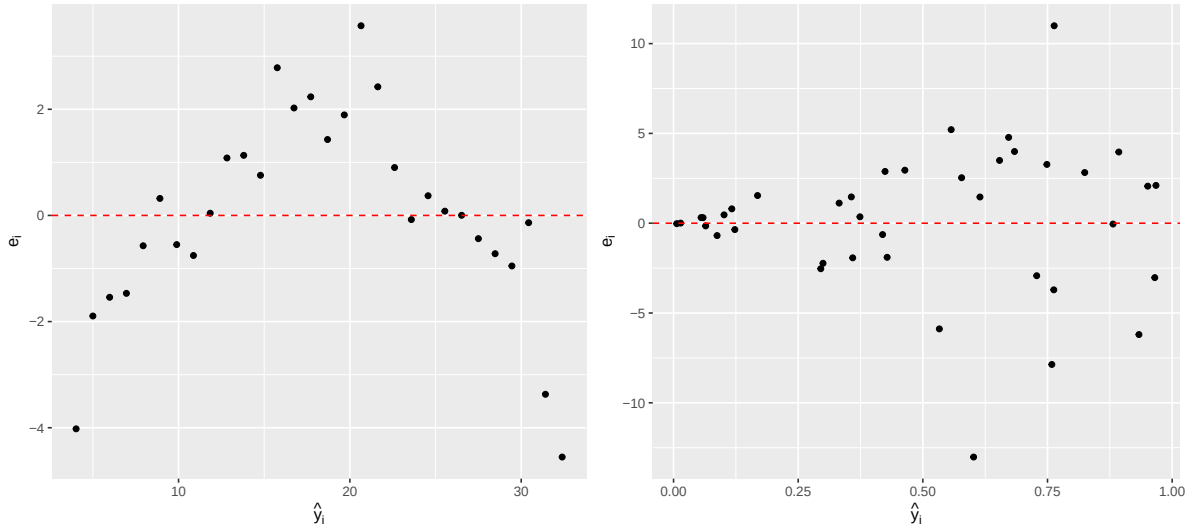


Figure 9.3: Residuenplot der Residuen e_i gegen die vorhergesagten Werte $\hat{y}_i$

Der Plot sollte im Optimalfall so aussehen, dass die Residuen e_i gleichmäßig oberhalb und unterhalb um die Nulllinie verteilt sind und keine weiteren Strukturen oder Muster im Zusammenhang mit $\hat{y}_i$ zu erkennen sind. In dem Residuenplot in Figure 9.3 ist zunächst einmal kein größeres Problem zu erkennen, bis auf den einen Wert links oben.

Um besser zu Verstehen wir Problem aussehen könnten, schauen wir uns zwei Residuenplots an, bei denen eine Struktur zu erkennen ist (siehe Figure 9.4)

In Figure 9.4a ist ein parabelförmiger Zusammenhang zwischen e_i und $\hat{y}_i$ zu erkennen. Für kleine und große $\hat{y}_i$ Werte sind die Residuen e_i negativ während für mittlere Werte von $\hat{y}_i$ die Residuen e_i positiv sind. Diese deutet darauf hin, das zusäztliche Struktur in den Daten nicht



(a) Parabelförmiger Zusammenhang zwischen e_i und $\hat{y}_i$

(b) Ansteigende Streuung mit größer werdendem $\hat{y}_i$

Figure 9.4: Residuenplots die Probleme anzeigen.

im Modell erfasst wird und führt dazu dass die Modellannahmen der Normalverteilung der ϵ_i nicht erfüllt sind.

In Figure 9.4b ist dagegen eine anderes Problem zu beobachten, die Residuen e_i zeigen zwar keine Struktur bezüglich positiv zu negativen Werten, allerdings werden die Abweichung von 0 mit größer werdenen $\hat{y}_i$ immer stärker. Dies deutet darauf hin, das die Streuung der Daten nicht gleich ist. Dies wird als Heteroskedastizität bezeichnet und deutet wiederum auf eine Verletzung der Annahmen bei der Homoskedastitzität ausgegangen wird. D.h. die Streuung soll über den gesamten Bereich von $\hat{y}_i$ gleich bleiben.

Definition 9.1 (Homoskedastizität)**.** Wenn die Größe der Varianz der Residuen ϵ_i in einem Regressionsmodell unabhängig von der Größe der Vorhersagevariable X_i ist, wird dies als Homoskedastizität bezeichnet. Die Streuung der Residuen ist dann für alle Werte X_i gleich. Wenn dies nicht der Fall ist, wird von Heteroskedastizität gesprochen.

Da die Varianz also konstant für alle Werte von X_i ist, trifft dies ebenfalls für die vorhergesagten Werte $\hat{Y}_i$ zu. Daher werden bei vielen der Plots die Residuen $\hat{\epsilon}_i$ gegen die $\hat{Y}_i$ abgetragen. Dies hat den Vorteil, dass später bei der multiplen Regression die gleiche Art von Graphen benutzt werden kann, um Homoskeastizität zu analysieren.

Eine weitere Möglichkeit die Verteilung der Residuen zu überprüfen ist die Anfertigung von sogenannten qq-Plots. Dies ermöglichen etwas strukturierter die Verteilung der Residuen zu überprüfen.

9.1.1 Quantile-Quantile-Plots

qq-Plot ist die Kurzform von Quantile-Quantile-Plot. D.h. es werden die Quantilen von zwei Variablen gegeneinander abgetragen. Um die Funktionsweise besser zu verstehen schauen wir uns erst einmal ein Spielzeugbeispiel an. In Table 9.1 ist eine kleiner Datensatz mit $n = 5$ Datenpunkten angezeigt.

Table 9.1: Spielzeugbeispieldaten mit $n = 5$

y
-2.0
5.0
-1.2
0.1
7.0

Wir wollen jetzt überprüfen ob dieser Datensatz einer Normalverteilung folgt (Wohlwissend das mit fünf Datenpunkten keine Verteilungsannahme überprüft werden kann). Dazu schauen wir uns zunächst noch einmal die bekannte Standardnormalverteilung $\Phi(z) = \mathcal{N}(\mu = 0, \sigma^2 = 1)$ an (siehe Figure 9.5).

Im ersten Schritt unterteilen wir die Standardnormalverteilung $\Phi(z)$ in $n + 1 = 6$ gleich große Flächen. D.h. die durch die Flächen bestimmten Abschnitte haben alle die gleiche Wahrscheinlichkeit (=Fläche unter der Dichtefunktion). Die Flächen werden durch jeweiligen Trennpunkte unterteilt die gleichzeitig die Quantilen sind.

In unserem Fall haben wir $n = 5$ Datenpunkte, unterteilen also unsere Verteilung in 6 Abschnitte die jeweils eine Fläche von $p = \frac{1}{6} = 0.17$ haben. D.h. $\frac{1}{6}$ der Werte von $\Phi(x)$ liegen links des ersten Trennpunktes, $\frac{2}{6}$ der Werte von $\Phi(x)$ liegen links des zweiten Trennpunktes, usw. D.h. die Trennpunkte bestimmen die jeweiligen Quantilen, oder genauer die theoretischen Quantilen die unter der Verteilungsannahme erwartet werden.

Die Idee hinter dem qq-Plot besteht nun darin, die empirischen Quantilen gegen die theoretischen Quantilen abzutragen (siehe Figure 9.7). Wenn die beobachteten Daten aus der gleichen Verteilung wie die theoretische Verteilung stammen, dann sollten die Punkte einer Geraden folgen. Die Steigung der Geraden ist 1, wenn es sich um die identischen Verteilungen handelt. Wenn die Steigung $\neq 1$ ist, dann kommen die Datenpunkte aus der gleichen Familie sind aber um einen Skalierungsfaktor unterschiedlich bzw. um den Mittelwert verschoben. Die Punkte sollten aber trotzdem auf einer Geraden liegen.

Um die empirischen Quartilen zu bestimmen, werden dazu zunächst die beobachteten Datenpunkte aus Table 9.1 aufsteigend nach der Größe sortiert (siehe Table 9.2). Diese Werte können als *empirische* Quantilen bezeichnet werden. Unter der Annahme, dass die Werte eine repräsentative Stichprobe aus der Verteilung darstellen, erwarten wir, dass wenn wir weitere

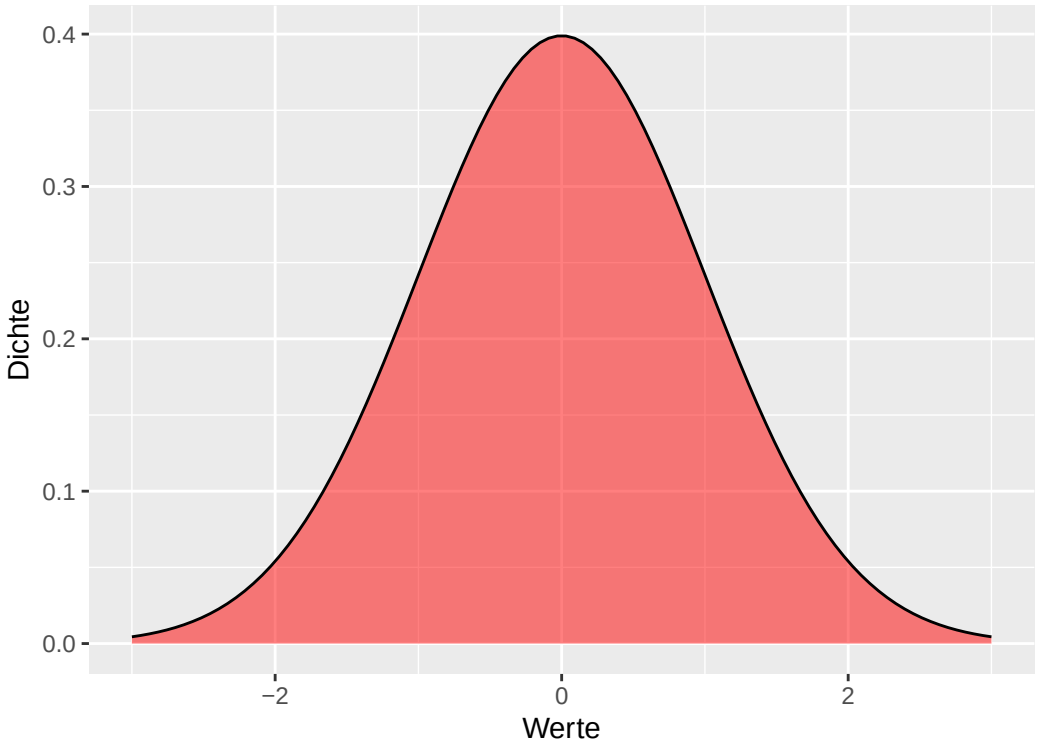


Figure 9.5: Dichtefunktion der Standardnormalverteilung

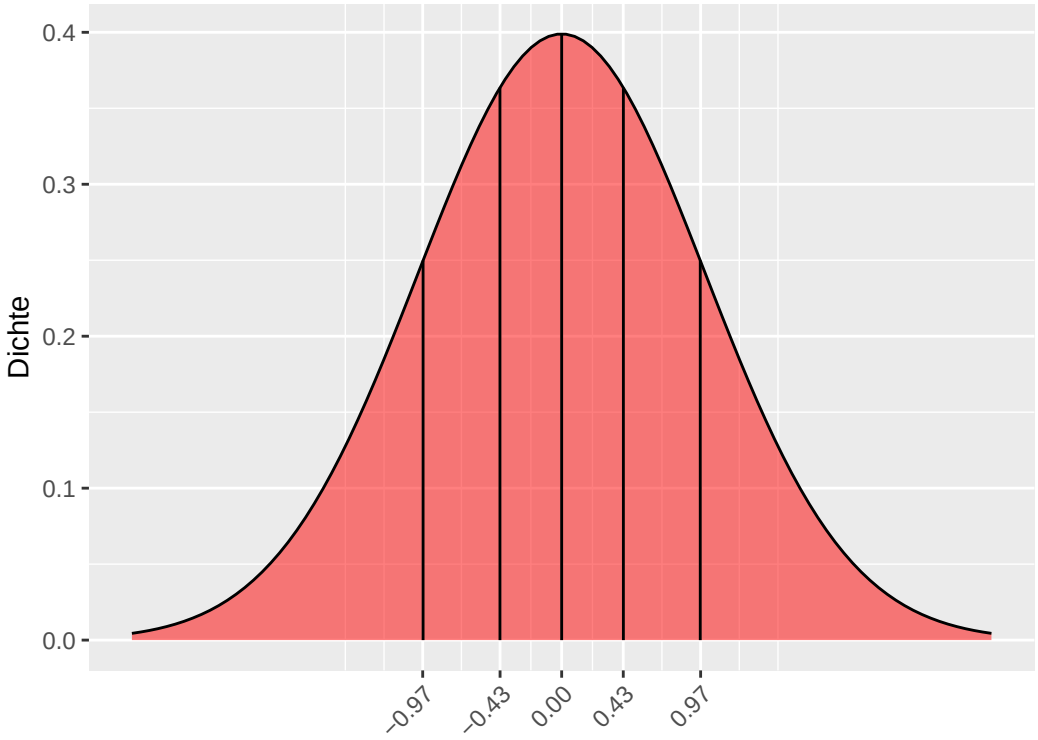


Figure 9.6: Unterteilung der Standardnormalverteilung in sechs gleich große Flächen

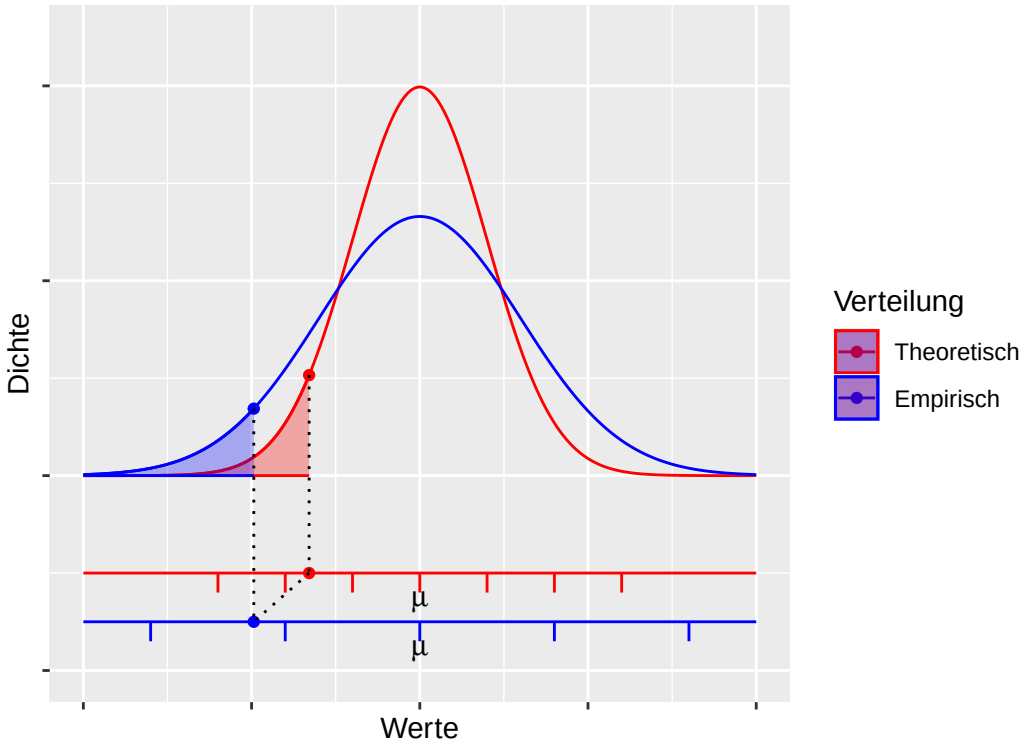


Figure 9.7: Skizze der theoretischen und der empirischen Verteilung mit unterschiedlicher Skalierung (Faktor 2×) aber aus der gleichen Verteilungsfamilie. In beiden Graphen ist die gleiche Quartile markiert

Werte beobachten würden, etwa $\frac{1}{6}$ der Werte kleiner als der kleinste Wert wären, $\frac{2}{6}$ der weiteren Werte kleiner als der 2. kleinste Wert wären und so weiter und so fort.

Table 9.2: Sortierte Datenwerte des Spielzeugbeispiels

kleinster	2.kleinster	mittlerer	2.größter	größter
-2	-1.2	0.1	5	7

Daher, wenn die beobachteten Werte der angenommenen theoretischen Verteilung folgen, dann sollte ein Graph der empirischen Quartilen gegen die theoretischen Quartilen nahezu (Stichprobenvariabilität) einer Geraden folgen.

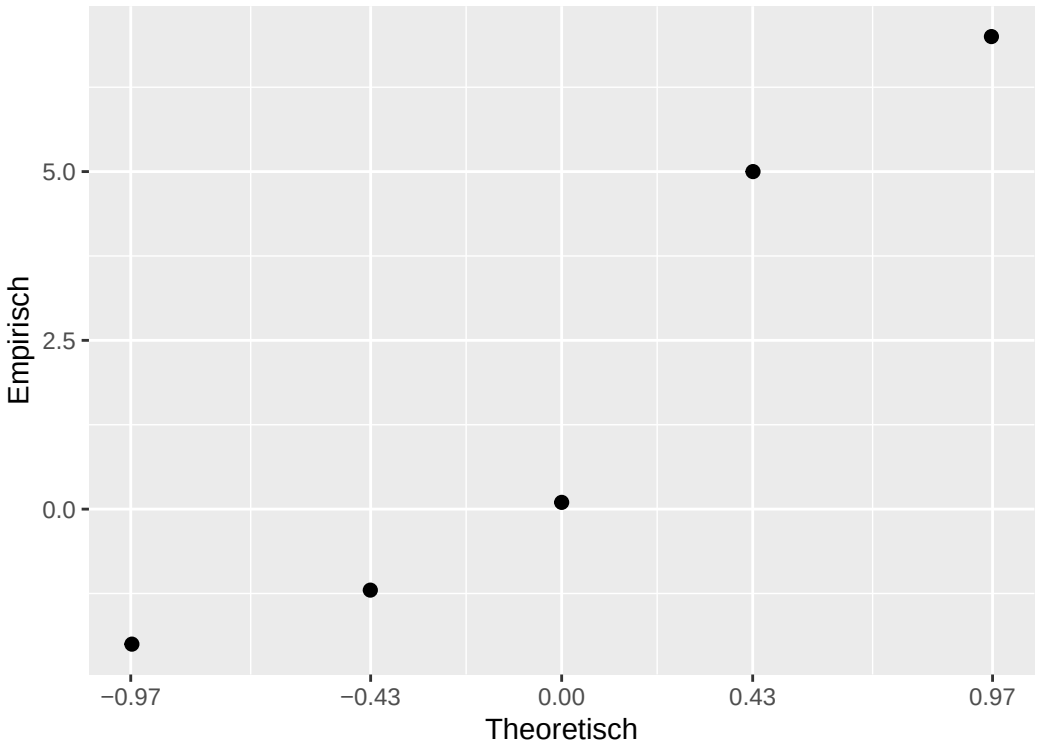
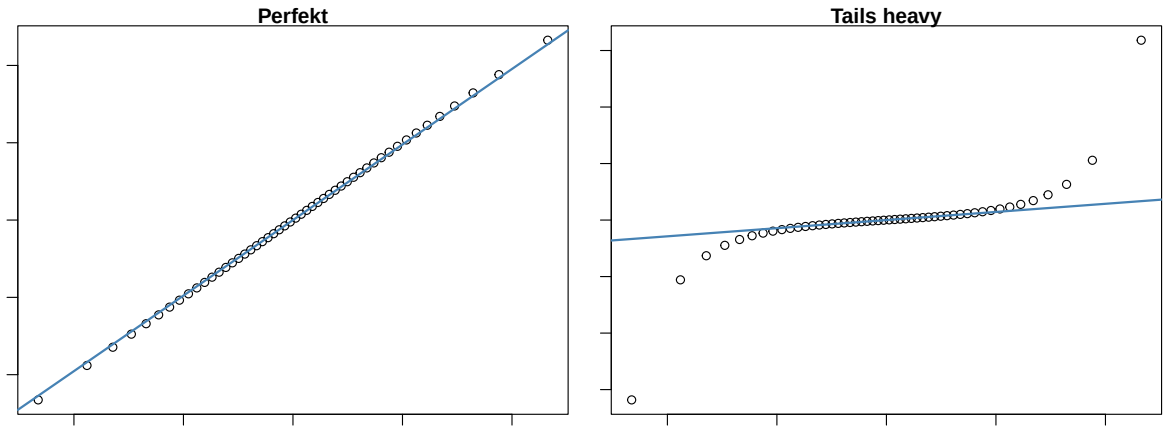


Figure 9.8: Streudiagramm der empirischen Werte gegen die theoretischen Quantilen

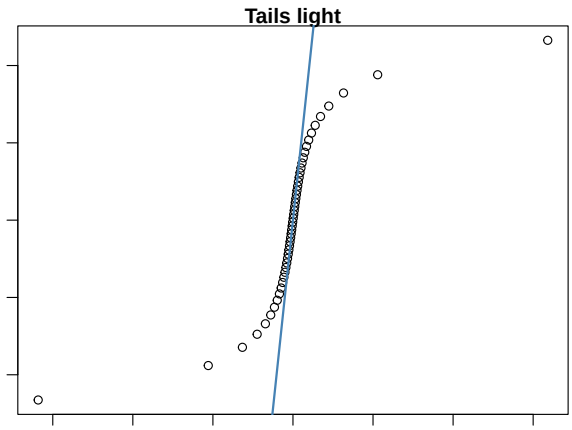
In Figure 9.8 sind die empirischen Quartilen gegen die theoretischen Quantilen für unser kleines Beispiel abgetragen. Tatsächlich ist es in diesem Fall schwierig eine Gerade zu erkennen bzw. von einer zu sprechen, da es sich nur um besagte fünf Wert handelt. Nochmals, mit $n = 5$ kann eine realistische Verteilungsannahme nicht überprüft werden.

Wenn der Datensatz größer ist, dann eignet sich ein qq-Plot allerdings sehr gut Abweichungen zu erkennen. In Figure 9.9 sind verschiedene Beispiele abgetragen.

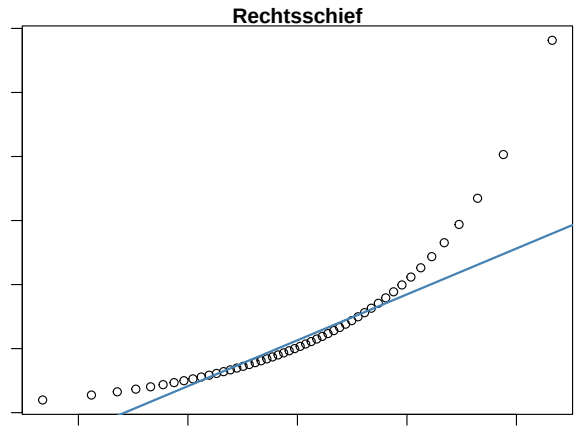


(a) Perfekt

(b) Enden schwer



(c) Enden leicht



(d) Rechtsschief

Figure 9.9: Beispiele für verschiedenen qq-Plots

In Figure 9.9a ist ein perfekter Zusammenhang zwischen den empirischen und den theoretischen Quantilen abgebildet. In diesem Falle wurden synthetisch für 50 normalverteiltet Zufallsdaten ein qq-Plot erstellt. Es ist zu sehen, das tatsächlich eine Gerade den Zusammenhang beschreibt. In Figure 9.9b ist dagegen ein Zusammenhang abgetragen, bei dem die empirischen und die theoretische Verteilung nicht zusammenpassen. In diesem Fall sind die haben die Randwerte der empirischen Vereteilung eine höhere Wahrscheinlichkeit als die unter der theoretischen Verteilung zu erwarten ist. D.h. extreme Werte kommen in der beobachteten Verteilung öfter in der theoretischen Verteilung vor. Dies deutet darauf hin, dass die Streuung der Daten möglicherweise nicht korrekt modelliert wurde. In diesem Fall, wird von einer *tail heavy* Verteilung gesprochen.

In Figure 9.9c ist der gegenteilige Effekt abgetragen. Hier hat die theoretische Verteilung mehr Wahrscheinlichkeitsmasse in den Randzonen als die empirische Verteilung. Die beobachtete Verteilung ist *tail light*. Entsprechend ist in Figure 9.9d ein Beispiel abgebildet, bei dem nur eine der Randzonen zu viel Wahrscheinlichkeitsmasse besitzt. Da die theoretische Verteilung wiederum die Normalverteilung ist und diese Symmetrisch ist, deutet diese darauf hin, das die empirische Verteilung ähnlich wie in Figure 9.9b in der rechten Randzone zu *viele* Werte hat und daher Rechtsschief ist.

Für unsere Daten ergibt sich das folgende qq-Diagramm (siehe Figure 9.10)

Der Graph sieht zunächst einmal gar nicht so schlecht aus. Allerdings deutet die Abweichung rechts oben darauf hin, das möglicherweise die Streuung nicht korrekt abgeschätzt wurde. Insbesondere ist ein Wert zu sehen, der im Verhältnis zu den anderen Werten schon relativ weit von der Gerade weg ist. Daher ist es hier angezeigt, diesen Wert noch einmal genauer zu untersuchen.

9.1.2 qq-Plot in `R`

In `R` gibt es zwei direkte Methoden einen qq-Plot zu erstellen. Mittels des Standardgrafiksystem können mit den Funktionen `qqnorm()` und `qqline()` qq-Plots mit der dazugehörigen Gerade erstellt werden. Für das `ggplot()`-System stehen die `geoms` `geom_qq()` und `geom_qq_line()` zur Verfügung. Wichtig ist hierbei, das in `aes()` der Parameter `sample` definiert werden muss. Für unser Spielzeugbeispiel sieht dies folgendermaßen aus:

```
df_toy <- tibble::tibble(y = c(-2, 5, -1.2, 0.1, 7))
ggplot(df_toy, aes(sample=y)) +
  geom_qq() +
  geom_qq_line()
```

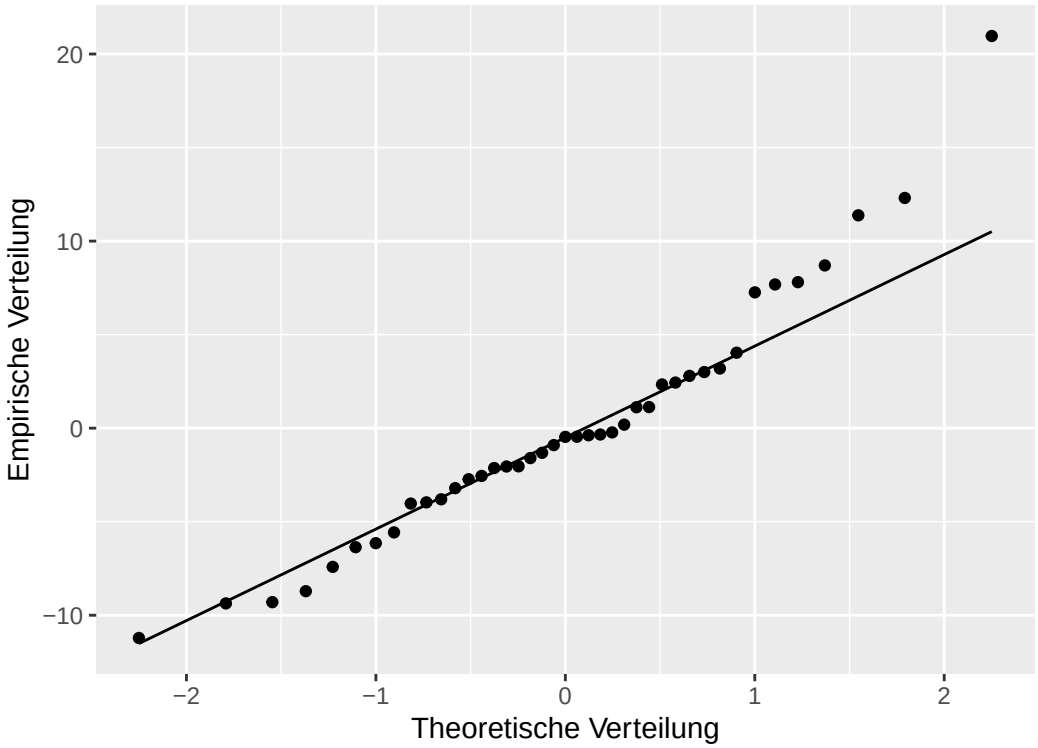


Figure 9.10: QQ-Diagramm der Residuen des ADAS-ADCS-Modells

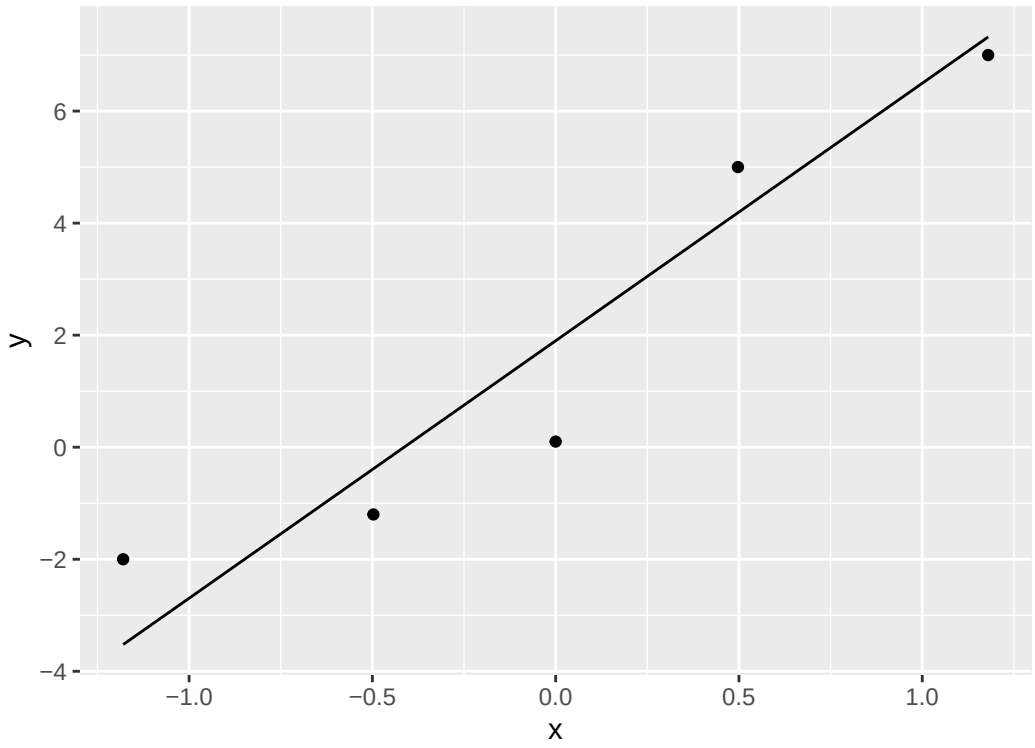


Figure 9.11: qq-Plot der Spielzeugdaten mittels `ggplot()`

9.1.3 Standardisierte Residuen

Eine Möglichkeit so einen Wert zu untersuchen, ist abzuschätzen wie *ungewöhnlich* der zu dem Residuen e_i gehörende y_i-Wert ist. Ein Problem der einfachen Residuen e_i ist, dass diese laut der Modellannahmen die gleiche Varianz σ^2 haben sollten. Allerdings, auf Grund der Art, wie die e_i berechnet werden, folgt die Randbedingung, dass die Summe der e_i gleich Null ist, $\sum_{i=1}^{n} e_i = 0$. Dies führt dazu, dass die einfachen Residuen nicht unanbhängig voneinander sind und nicht immer Homoskedastizität besitzen. Daher gibt es eine weitere Art Residuen anhand des Modell zu berechnen, die nicht unter diesen Beschränkungen leiden. Dies sind die standardisierten Residuen e_{Si}. Dazu müssen wir uns zunächst mit Hebelwerte h_i beschäftigen.

9.1.3.1 Hebelwerte

Wenn ein Modell an die Daten gefittet wird, dann haben nicht alle Werte den gleichen Einfluss auf die Modellparameter. Manche Werte üben einen stärkeren Einfluss auf das Modell aus

als andere Werte. In Figure 9.12 ist ein Beispiel abgebildet für einen Datensatz bei dem ein einzelner Punkt einen übermäßig großen Einfluss auf das Modell ausübt.

Figure 9.12: Beispiel für einen Datenpunkt mit einem großen Einfluss auf das Modell. Die resultierenden Regressionsgeraden sind mit dem Punkt (rot) und ohne den Punkt (grün) abgetragen.

Der einzelen Punkt rechts oben in Figure 9.12 hat einen großen Einfluss auf die resultierende Regressionsgerade wie in der Abbildung zu sehen ist. Der Einfluss ist zum Teil durch den großen Abstand des x_i-Wertes vom Mittelwert der x_i-Werte $\bar{x}$ bestimmt. Der Einfluss jedes einzelnen x-Wertes wird mittels der sogenannten Hebelwerte h_i bestimmt. Die genaue Berechnung der Hebelwerte h_i ist für das weitere Verständnis allerdings nicht wichtig, sondern mehr das Verständnis des Konzepts. Die Hebelwerte h_i können Werte in $h_i \in [1/n, 1]$ annehmen. In `R` können die Hebelwerte mit der Funktion `hatvalues()` berechnet werden.

Tragen wir in die Grafik die Hebelwerte in die Grafik Figure 9.12 ein (siehe Figure 9.13), dann ist zu sehen, dass der abgesetzte Wert auch den größten Hebelwert hat.

Eine Daumenregel für die Hebelwerte ist der Schwellenwert von $(2k+2)/n$, wobei k die Anzahl der unabhängigen Variablen ist. Für den Beispieldatensatz in Figure 9.13 würde sich daher ein Wert von $(2 \cdot 1 + 2)/30 = 0.13$ ergeben. Entsprechend wäre der abgesetzte Wert mit einem Hebelwert von $h_i = 0.54$ als problematisch einzustufen.

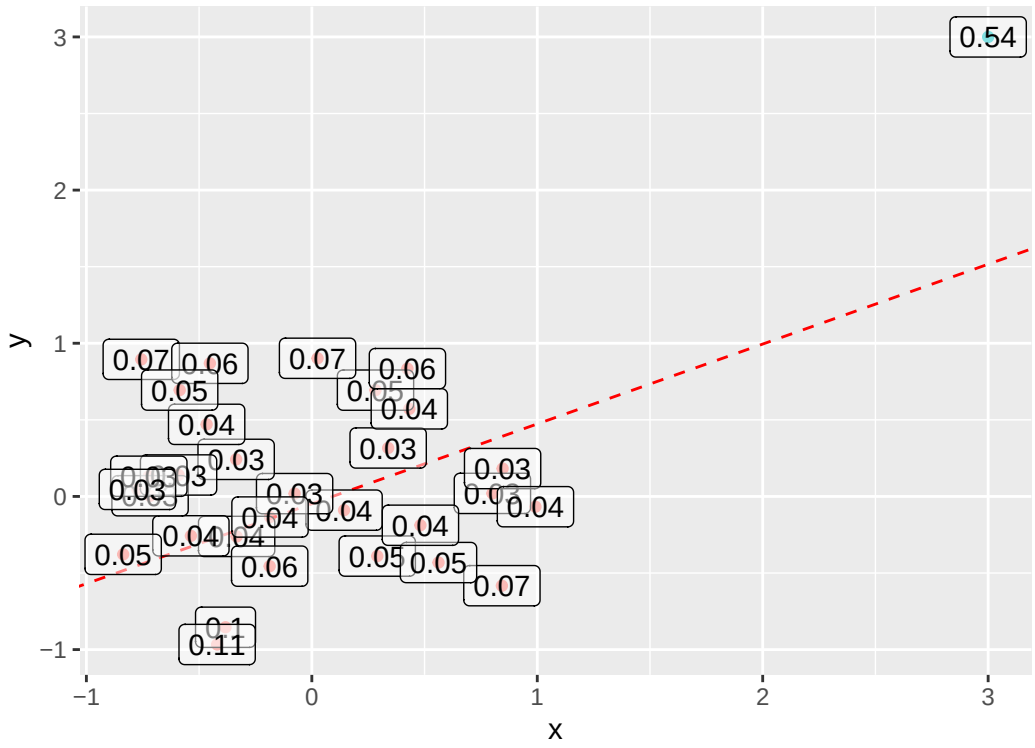


Figure 9.13: Beispiel für einen Datenpunkt mit einem großen Einfluss auf das Modell. Die Werte geben die jeweiligen Hebelwerte h_i der Datenpunkte wieder.

Nach diesem kurzen Exkurs zu den Hebelwerten h_i, schauen wir uns für unsere weitere Betrachtung der Residuen zunächst den Zusammenhang zwischen der Varianz der Residuen in der Population σ^2 und der Varianz der geschätzten Residuen $\sigma^2(\hat{\epsilon}_i) = \sigma^2(e_i)$ an. Es gilt:

$$\sigma^2(e_i) = \sigma^2(1 - h_i) \tag{9.3}$$

D.h. wenn ein Datenpunkt x_i einen kleineren Einfluss auf das Modell ausübt und dementsprechend einen kleinen Hebelwert h_i, dann wird die Varianz für diesen Wert nahezu korrekt eingeschätzt. Hat der Wert x_i allerdings, einen großen Hebelwert h_i, führt die dazu, dass die Varianz für diesen Wert stärker unterschützt wird. Dieser Zusammenhang kann dazu benutzt werden standardisierte Residuen zu erstellen.

$$e_{Si} = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}} \tag{9.4}$$

Die standardisierten Residuen e_{Si} haben dazu die Eigenschaft, dass sie eine Varianz und damit Standardabweichung von $\sigma^2(e_{Si}) = 1$ haben, also Standardnormalverteilt $\Phi(z)$ sein sollten. Dadurch können Abweichungen von den Modellannahmen leichter Identifiziert werden, da die Skala normiert ist. In `R` kann die standardiserten Residuen e_{Si} mittels der Funktion `rstandard()` berechnet werden. Eine Standardgrafik zum inspizieren der standardisierten Residuen ist wiederum eine Abbildung der e_{Si} gegen die $\hat{y}_i$.

Die Figure 9.14 sieht relativ ähnlich zu Figure 9.3 aus. Durch die Änderung der Skala ist jetzt aber leichter abschätzbar ob die Verteilung der erwarteten Normalverteilung folgt. D.h. etwa $\frac{2}{3}$ der Werte sollten zwischen -1 und 1 liegen und etwa 95% zwischen -2 und 2. Bis auf den einen Punkt oben rechts, sieht alles soweit unauffällig aus.

9.1.4 Studentized Residuals

Die letzte Art von Residuen sind die sogenannten Studentized Residuals e_{Ti}, die mittels der folgenden Formel berechnet werden.

$$e_{Ti} = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1 - h_i}} \tag{9.5}$$

Die Formel (9.5) ist sehr ähnlich zu derer für die standardisierten Residuen, der einzige Unterschied ist der Term $\hat{\sigma}_{(-i)}$. Dieser bezeichnet die Residualvarianz wenn dass Modell ohne den Datenpunkt i gefittet wird. D.h. wie stark verändert sich die Schätzung der Varianz wenn ein Datenpunkt weggelassen wird. Normalerweise sollte eine einzelner Punkt keinen übermäßigen Einfluss auf die geschätzte Varianz haben, daher können die Studentized Residuals dazu verwendet werden problematische Datenpunkte zu identifizieren. Wenn die tatsächlichen

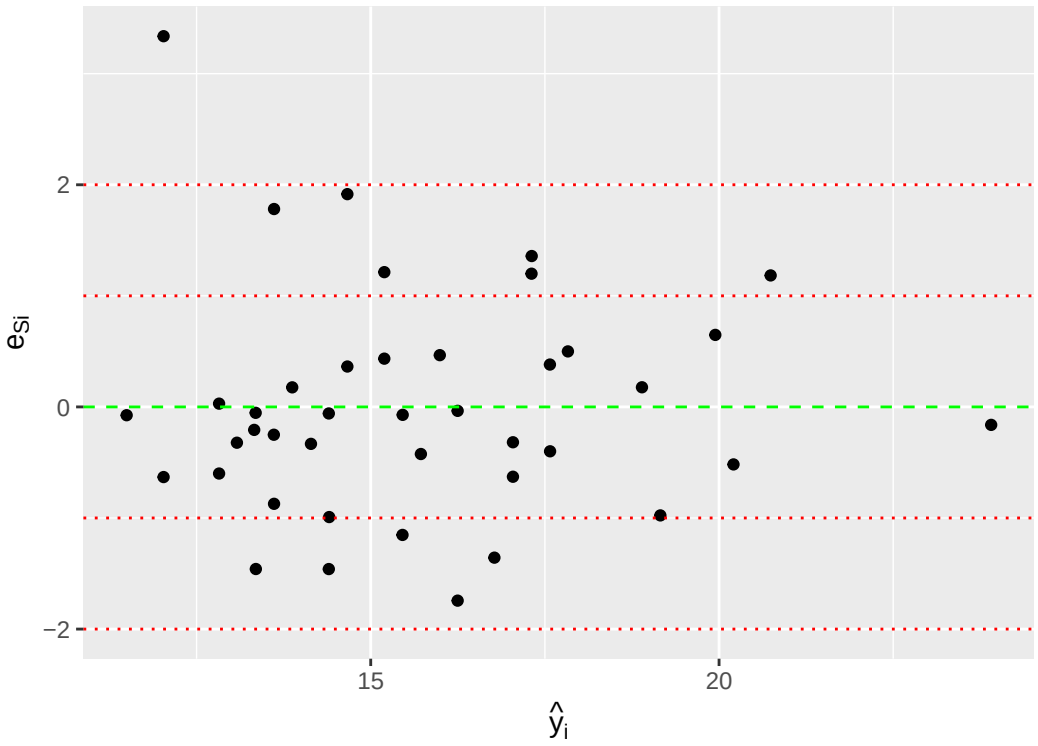


Figure 9.14: Grafik der standardisierten Residuen e_{Si} gegen die Vorhersagewerte $\hat{y}_i$ für das ADL-Modell.

Residuen einer Normalverteilung folgen, dann kann gezeigt werden, dass die Studentized Residuals einer t-Verteilung mit $N-k-2$ Freiheitsgeraden folgen. Daher könnte sogar ein formaler statistischer Test durchgeführt werden. In `R` können die Studentized Residuals e_{Ti} mittels der Funktion `rstudent()` berechnet werden und werden entsprechend den anderen Residuen in dem üblichen Graphen gegen die vorhergesagten Werte $\hat{y}_i$ abgetragen.

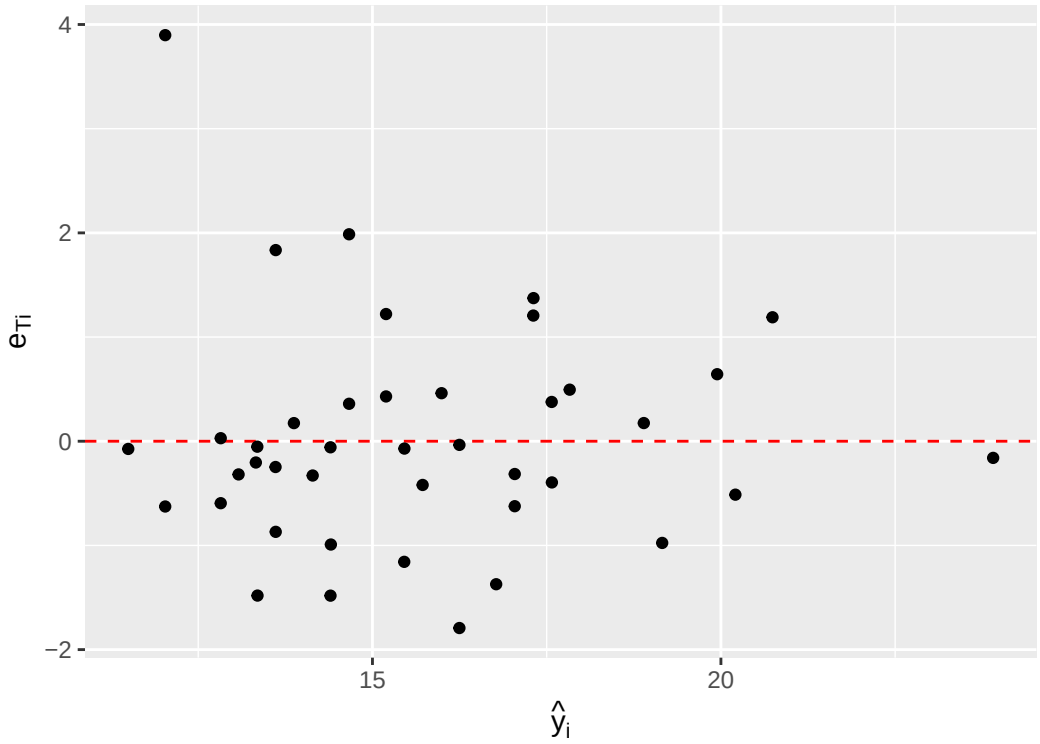


Figure 9.15: Graph der Studentized Residuals S_{Ti} gegen die vorhergesagten Werte $\hat{y}_i$ vor das `adl`-Modell

9.1.5 Übersicht über die Residuenarten

In Table 9.3 sind noch einmal die drei Arten von Residuen aufgelistet.

Table 9.3: Übersicht über verschiedene Arten von Residuen

Typ	Berechnung	Ziel
Einfache Residuen	$e_i = y_i - \hat{y}_i$	Verteilungsannahme
Standardisierte Residuen	$e_{Si} = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$	Verteilungsannahme
Studentized Residuen	$e_{Ti} = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$	Einfluss auf Modell

9.1.6 Ausgabe von `summary()` (continued)

Nach dieser Betrachtung der Residuen, die nach jedem Modellfit inspiziert werden sollten um zu überprüfen ob die Modellannahmen angemessen sind schauen wir uns noch einmal kurz die Ausgabe von `summary()` an.

```
Call:
lm(formula = adcs ~ adas, data = adl)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2177  -3.8033  -0.4663   2.7950  20.9634

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.5445     4.3052   6.166 3.05e-07 ***
adas         -0.2638     0.1015  -2.599   0.0131 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.516 on 39 degrees of freedom
Multiple R-squared:  0.1477,    Adjusted R-squared:  0.1258
F-statistic: 6.757 on 1 and 39 DF,  p-value: 0.01312
```

Nach der Wiedergabe des gefitten Modells erfolgt direkt eine Zusammenfassung der Residuen über Minimum und Maximum, Q1 und Q3 und den Median. Jetzt sollte daher auch besser nachvollziehbar sein, warum es sinnvoll ist diese Statistiken über die Residuen direkt anzugeben. Die beiden Extremwerte geben einen ersten Überblick auf mögliche Ausreißer, während die erste Quartile Q1 und die dritte Quartile Q3 möglich Asymmetrien in der Verteilung der Residuen anzeigen. Laut der Annahem der Residuen als Normalverteilt mit $\mu = 0$, sollten diese beiden Werte etwa gleich weit von Null entfernt sein. Dementsprechend sollte der Median nahe an Null dran sein. Was nah ist, kommt dabei immer auf die Einheit der abhängigen Variablen an, wenn der Abstand in Kilometern ist kann ein kleiner Wert schon problematisch sein, während wenn eine Sprungweite in Mikrometern angeben wird eine großer Wert unbedenklich sein kann. Der Schätzerwert für σ selbst, wir unten mit `Residual standard error` angegeben.

Im vorliegenden Fall des Modells für die `adl`-Daten ist der Median dementsprechend doch etwas weit von Null entfernt und der geschätzte Residualfehler $\hat{\sigma} = 6.52$ ebenfalls relativ groß. $\hat{\sigma}$ kann mittels der Funktion `sigma()` erhalten werden.

9.1.7 Zum Nachlesen

Zum weiteren Vertiefen der Inhalte findet ihr in Kutner et al. (2005, 100–114), N. Altman and Krzywinski (2016b) und Fox (2011, 285–96) noch einmal gute Zusammenfassungen.

9.2 Einflussmetriken

Um das gefittet Modell zu diagnostizieren reicht es allerdings nicht aus, sich nur die Residuen anzuschauen. Ein weiterer wichtiger Punkt ist die Analyse des Einflusses der einzelnen Datenpunkte auf das Modell. Wenn alles gut läuft sollte es keine einzelnen Datenpunkte geben, die einen übermäßig großen Einfluss auf das Modell ausüben. Anders ausgedrückt, die Anwesenheit bzw. Abwesenheit von einzelnen Datenpunkte sollte nicht dazu führen, dass die Aussage des Modells sich stark verändert. Im folgenden schauen wir uns dazu verschiedene Einflussmetriken die den Einfluss der Datenpunkte auf das Modell abschätzen. Die Idee der Einflussmetriken ist dabei die Gleiche wie schon bei den Studentized Residuals. Der Einfluss der Datenpunkte auf den Modellfit wird interpretiert indem ein Modell mit und ein Modell ohne den jeweiligen Datenpunkt gefittet wird. Der Einfluss auf verschiedene Modellparameter wird dann bestimmt und dementsprechend als möglicherweise bedenklich eingestuft. Die Meisten der im folgenden vorgestellten Ansätze verwenden in der einen oder anderen Form die Hebelwerte h_i die wir bereits kennengelernt haben.

9.2.1 DFFITS (difference in fits)

Das erst Maß, daß wir uns anschauen ist DFFITS (kurz für difference in fits). Das DFFITS-Maß wird getreent für jeden einzelnen Datenpunkt berechnet und der Einfluss des Datenpunkts auf den gefitteten Werte $\hat{y}_i$ für den jeweiligen Datenpunkt berechnet. Formal:

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}\sqrt{h_i}} \tag{9.6}$$

Im Zähler kommen von Formel(9.6) kommen zweimal die vorhergesagte y-Werte vor. $\hat{y}_i$ ist dabei der ganz normale Vorhersagewert der uns mittlerweile schon mehrfach begegnet ist. Der zweite Wert $\hat{y}_{i(i)}$ bezeichnet den vorhergesagten Wert aus dem Modell aus dem der Wert y_i weggelassen wurde. D.h, dass Modell ist mit einem Wert weniger gefittet worden. Daher misst die Differenz $\hat{y}_i - \hat{y}_{i(i)}$ den Unterschied in den Vorhersagewerte zwischen den zwei Modellen bei denen einmal der Wert y_i zum fitten verwendet wurde und einmal wenn y_i weggelassen wurde. Umso größer der Unterschied zwischen diesen beiden Werte umso größer ist der Einfluss des Wertes y_i auf den Modellfit. Im Nenner von Formel(9.6) wird wieder ein ähnlicher

Normierungswert wie bei den Studentizied Residuals angewendet. Insgesamt, wird mittels DFFITS daher für jeden Datenpunkt ein Wert ermittelt und umso größer dieser Wert ist umso größer ist der Einfluss des jeweiligen Datenpunktes auf den Modellfit.

Im idealen Fall sollte alle Datenpunkt ungefähr den gleichen Einfluss haben und einzelne Datenpunkte die einen übermäßig großen Einfluss auf das Modell haben sollten noch einmal genauer inspiziert werden.

> **Tip**
>
> Als Daumenregel, kann für kleine bis mittlere Datensätze ein DFFITS von ≈ 1 auf Probleme hindeuten, während bei großen Datensätzen $\approx 2\sqrt{k/N}$ als Orientierungshilfe verwendet werden kann (k := Anzahl der Prediktoren, N := Stichprobengröße).

> **Warning**
>
> Wenn ein Wert außerhalb der Daumenregel liegt, heißt das nicht, dass er automatisch ausgeschlossen werden muss/soll, sondern lediglich inspiziert werden sollte und das Modell mit und ohne diesen Wert interpretiert werden sollte.

In `R` können die DFFITS werden mittels der `dffits()`-Funktion berechnet werden. Als Parameter erwartet `dffits()` das gefittete `lm()`-Objekt. Ähnlich wie bei den Residuen, werden die DFFITS-Werte gegen die vorhergesagten y_i-Werte graphisch abgetragen um die Wert zu inspizieren und Probleme in der Modellspezifikation zu identifizieren.

In Figure 9.16 sind die DFFITS-Werte gegen die vorhergesagten Werte $\hat{y}_i$ abgetragen und zusätzlich die Daumenregel ± 1 eingezeichnet. Hier ist ein Wert nur gerade so außerhalb des vorgeschlagenen Bereichs. Hier könnte daher sich dieser Datenpunkt noch einmal genauer angeschaut werden, ob bei Ausschluß des Wertes es zu einer qualitativ anderen Interpretation der Daten kommt oder ob bespielsweise Übertragungsfehler für diesen Wert vorliegen oder sonstige Gründe.

9.2.2 Cook-Abstand

Während DFFITS den Einfluss des Datenpunktes i auf den jeweiligen Datenpunkt abschätzt, wird bei dem sogenanten Cook-Abstand der Einfluss des i-ten Datenpunktes auf alle n vorhergesagten Werte $\hat{y}_i$. Formal:

$$D_i = \frac{\sum_{j=1}^{N} (\hat{y}_j - \hat{y}_{j(i)})}{k\hat{\sigma}^2} \tag{9.7}$$

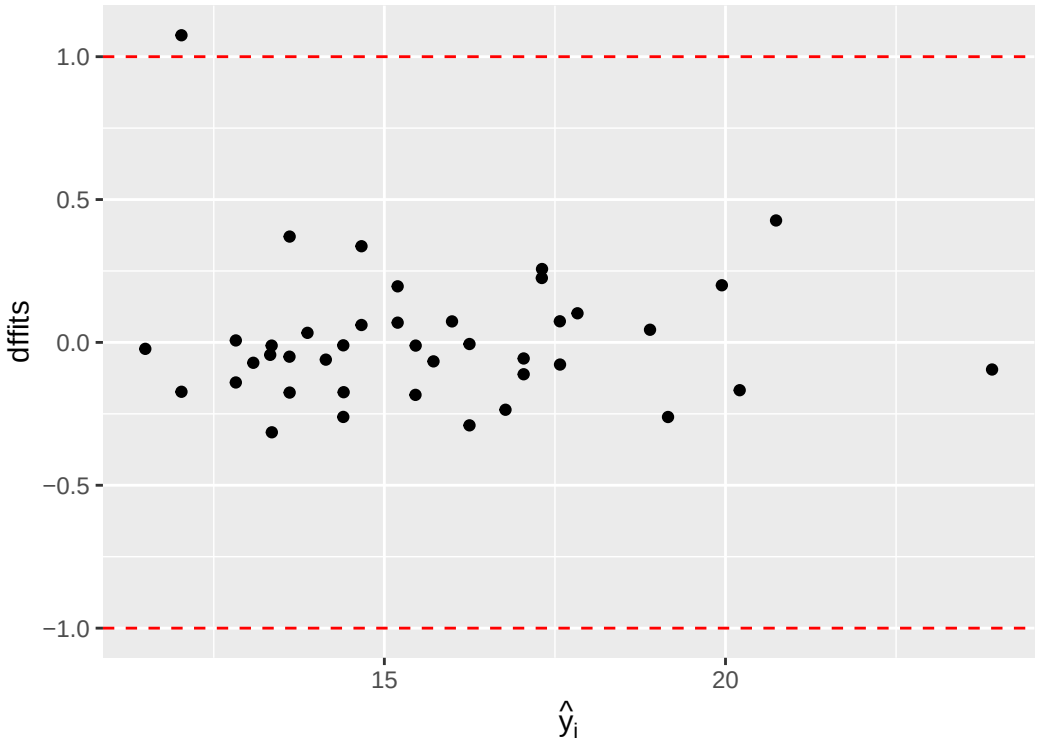


Figure 9.16: Graph der DFFITS-Werte gegen $\hat{y}_i$ für das `adl`-Modell.

Hier bedeutet die Syntax $\hat{y}_{j(i)}$ der vorhergesagte Wert für den Datenpunkt j wenn der i-te Datenpunkt ausgelassen wird. In `R` können die Cook-Abstände mit Hilfe der Funktion `cooks.distance()` berechnet werden.

> **Tip**
>
> Eine Daumenregel um einen *möglichen* Ausreißer zu identifzieren kann über $D_i > 1$ abgeschätzt werden.

In Figure 9.17 ist wiederum der übliche Graph gegen die vorhergesagten Werte $\hat{y}_i$ zu sehen. Anhand der abgebildeten Wert ist keiner der Datenpunkte als problematisch zu identifizieren.

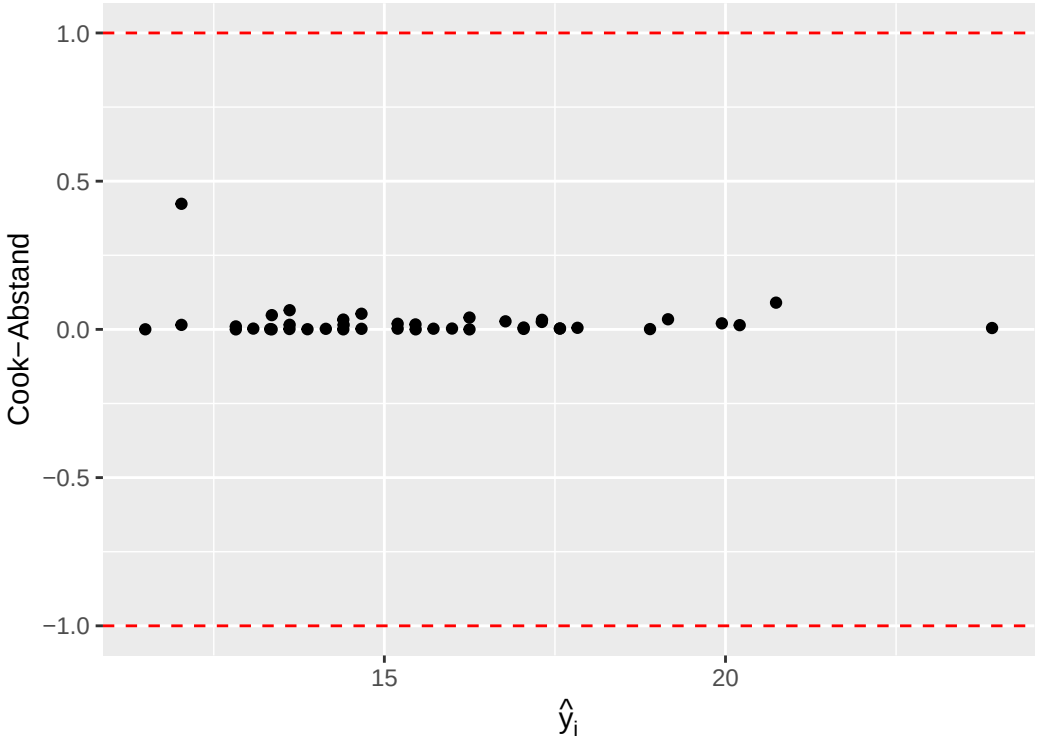


Figure 9.17: Cook's D_i gegen $\hat{y}_i$ für das `adl`-Modell.

9.2.3 DFBETAS

Als letztes Maß schauen wir uns noch DFBETAS an. DFBETAS berechnet ein Maß für die Veränderung der β-Koeffizienten auf Grund der einzelnen Datenpunkte i. D.h. es wird jetzt nicht nur ein Wert für jeden Wert berechnet, sondern ein Wert für den jeden Datenpunkt und

jeden β-Koeffizienten. In unseren Fall mit einem y-Achsenabschnitt β_0 und einem Steigungskoeffizienten β_1 werden entsprechend $2 \times x$ Werte berechnet. Formal:

$$(DFBETAS)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\hat{\sigma}^2 c_{kk}}} \tag{9.8}$$

Wie aus Formel (9.8) ersichtlich wird, wird die Veränderungen der Koeffizienten β_i bei weglassen des i-ten Datenpunktes abgeschätzt. Den Wert c_{kk} lassen wir unberücksichtigt, da er wiederum nur einen Normierungsfaktor darstellt.

> **Tip**
>
> Als Daumenregel gilt für kleine bis mittlere Datensätze ≈ 1, bzw. für große Datensätze $\approx 2/\sqrt{N}$

Wiederum gibt es eine spezielle Funktion in `R` um die DFBETAS zu berechnen `dfbeta()`. Dabei ist jedoch zu beachten das eine Matrize mit k-Spalten von `dfbeta()` zurück gegeben wird. Jede Spalte gibt den Wert für den jeweiligen β-Koeffizienten an.

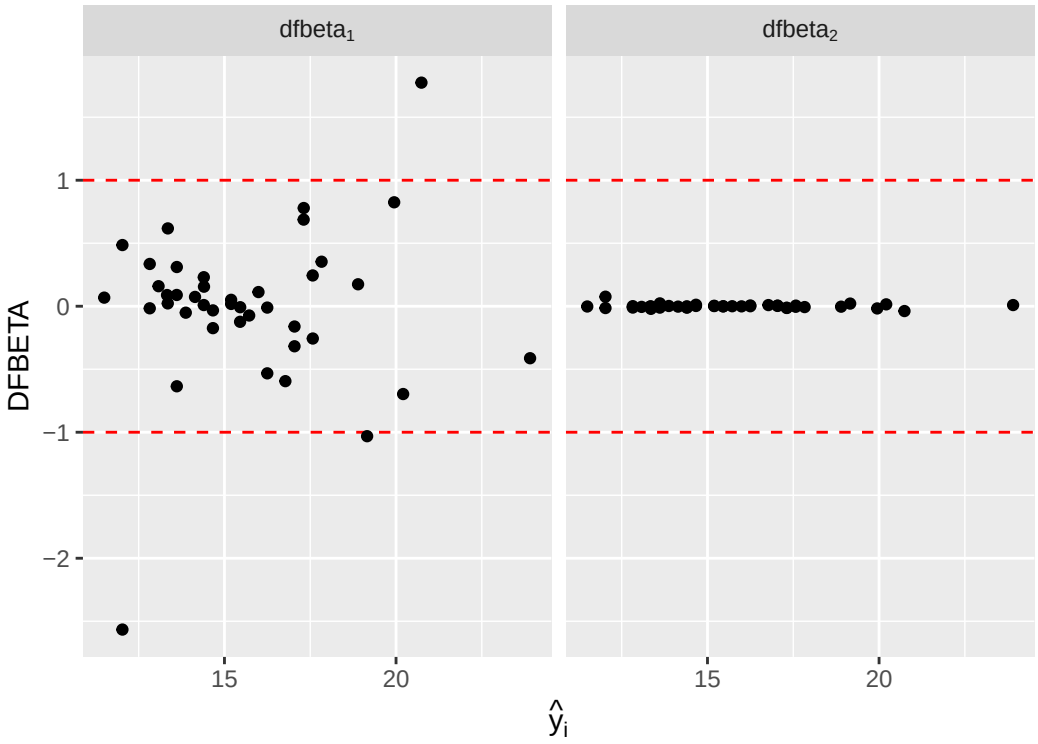


Figure 9.18: DFBETA-Werte für β_0 und β_1 gegen $\hat{y}_i$

In Figure 9.18 sind die DFBETAS für die beiden Koeffizienten $\hat{beta}_0$ und $\hat{beta}_1$ abgetragen. Hier ist zu sehen, dass die Wert für den Steigunsgkoeffizienten β_1 alle als unproblematisch anzusehen sind, während in Bezug auf β_0 ein paar wenige Fälle eine weiter Inspektion nach sich ziehen könnten. Allerdings sollte berücksichtigt werden, dass der y-Achsenabschnitt sehr stark durch die Verteilung der Datenpunkte in Bezug auf die x-Werte beeinflusst ist, da der Mittelwert der x-Werte bei $\bar{x} = 41.2$ liegt.

9.2.4 Übersicht über die Einflussmetriken

In Table 9.4 sind noch einmal die verschiedenen Methoden tabellarisch dargestellt.

Table 9.4: Übersicht über die verschiedene Einflussmaße zur Bewertung der Modellgüte

Typ	Veränderung	Daumenregel
$(DFFITS)_i$	Vorhersagewert i	$2\sqrt{k/N}$
Cook	Durchschnittliche Vorhersagewerte	> 1
$(DFBETAS)_{k(i)}$	Koeffizient i	$2\sqrt{N}$
e_{Ti}	Residuum i	t-Verteilung(n-k-2)

Nochmal, die Daumenregeln sind wirklich auch nur Daumenregeln und identifzieren nicht automatisch ein Problem im Datensatz.

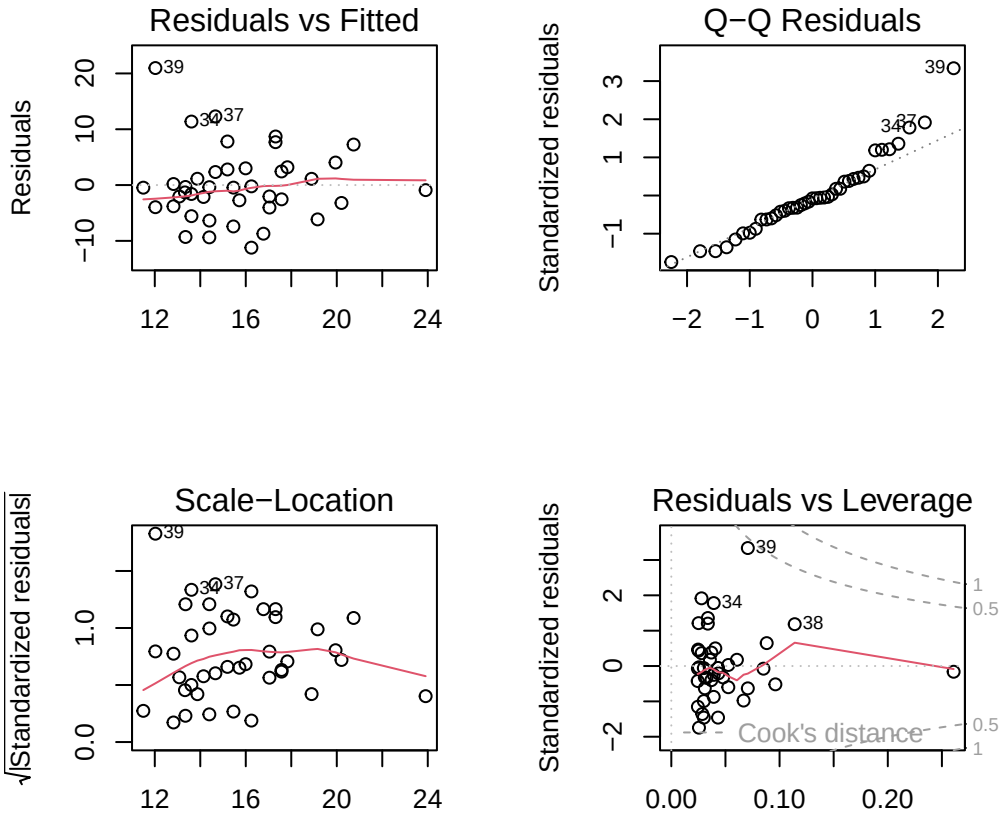
9.2.5 Zum Nacharbeiten

Noch mal weitere Informationen findet Ihr in N. Altman and Krzywinski (2016a), Fox (2011, 294–302) und Young (2019).

9.3 Diagnoseplots in R

Da die Diagnose eines gefitten Modell in jedem Fall durchgeführt werden soll und es sich dabei also um eine alltägliche Aufgabe handelt, gibt es mit `plot(mod)` einen short-cut um eine Reihe von Diagnoseplots direkt erstellen zu können.

```
plot(mod)
```



Eine weitere Möglichkeit ist das package `performance` das zahlreiche Funktion enthält rund um die Analyse von Modellfits (siehe beispielweise `performance::check_model()`).

10 Vorhersage

10.1 Vorhergesagte Werte $\hat{y}_i$

Wenn ein einfaches lineares Modell gefittet wurde ist eine zentrale Frage welche Vorhersagen anhand des Modell getroffen werden können. Die Vorhersagen $\hat{y}_i$ liegen auf der vorhergesagten Regressionsgerade und berechnen sich nach dem Modell für einen gegeben x-Wert.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_0 x$$

Wie schon mehrfach besprochen unterliegt die Regressionsgerade inherent der Unsicherheit bezüglich der geschätzen Modellkoeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$. Diese Unsicherheit überträgt sich auf die geschätzen Werte $\hat{y}_i$ und muss daher bei deren Interpretation berücksichtigt werden.

In Figure 10.1 sind die bereits behandelten Sprungdaten gegen die Anlaufgeschwindigkeiten zusammen mit der Regressionsgeraden und vorhergesagten Werten (rot) abgetragen.

In `R` können die vorhergesagten Werte des mittels `lm()` gefitteten Modells mit der Hilfsfunktion `predict()` bestimmt werden. Wenn der Funktion `predict()` keine weiteren Parameter außer dem `lm`-Objekt übergeben werden, berechnet `predict()` die vorhergesagten Werte $\hat{y}_i$ für alle die x-Werte die auch zum fitten des Modells benutzt wurden. Die Reihenfolge der Werte $\hat{y}_i$ enspricht dabei den Werten im Original-`data.frame()`.

```
predict(mod)[1:5]
```

```
       1        2        3        4        5
4.523537 4.725140 4.856256 4.761778 5.416207
```

Wir haben uns hier nur die ersten fünf Werte ausgeben lassen, da nur demonstriert werden soll wie die `predict()`-Funktion angewendet werden kann. Um eine Anwendung zu geben, so können mittels `predict()` die Residuen auch von Hand ohne die `resid()`-Funktion erhalten werden.

```
(jump$jump_m - predict(mod))[1:5]
```

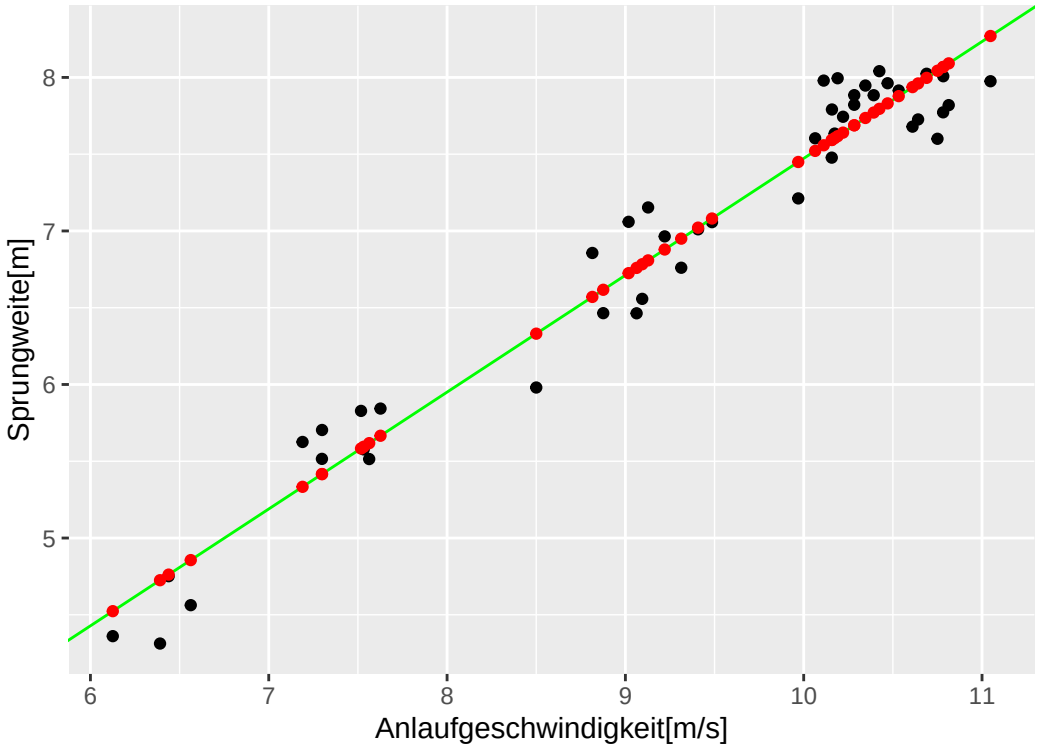


Figure 10.1: Vorhersagewerte $\hat{y}_i$ (rote Punkte) für die Sprungdaten.

```
          1           2           3           4           5
-0.16267721 -0.41248842 -0.29359256 -0.01047071  0.09927500
```

```
resid(mod)[1:5]
```

```
          1           2           3           4           5
-0.16267721 -0.41248842 -0.29359256 -0.01047071  0.09927500
```

Wiederum nur zur Demonstration die ersten fünf Wert um die Äquivalenz der beiden Methoden zu demonstrieren.

Meistens liegt das Interesse jedoch weniger auf den vorhergesagten Werten $\hat{y}_i$ für die gemessenen Werte, sondern es sollen Werte vorhergesagt werden für x-Werte die nicht im Datensatz enthalten sind. Operational ändert sich nichts, es wird immer noch das gefittete Modell verwendetet und es müssen lediglich neue x-Werte übergeben werden.

In `R` kann dies mittels des zweite Parameter in `predict()` erreicht werden. Soll zum Beispiel die Sprungweite für eine Anlaufgeschwindigkeit von $v = 11.5[m/s]$ berechnen werden, muss zunächst ein neues `tibble()` erstellt werden, welches den gewünschten x-Wert enthält. Dabei muss der Spaltenname in dem neuen `tibble()` demjenigen im Original-`tibble()` entsprechen. Ansonsten funktioniert die Anwendung von `predict()` nicht.

```
df <- tibble(v_ms = 11.5)
df
```

```
# A tibble: 1 x 1
   v_ms
  <dbl>
1  11.5
```

Dieses `tibble()` kann nun zusammen mit dem `lm()`-Objekt an `predict()` übergeben werden.

```
predict(mod, newdata = df)
```

```
       1
8.614136
```

D.h., bei einer Anlaufgeschwindigkeit von $v = 11.5[m/s]$ ist anhand des Modells eine Sprungweite von $8.6m$ zu erwarten.

10.2 Unsicherheit in der Vorhersage

Wie schon angesprochen ist unser Modell natürlich mit Unsicherheiten behaftet. Diese drücken sich in den Standardfehler für die beiden Koeffizienten $\hat{\beta}_0$ und $\hat{\beta}_1$ (siehe Table 10.1).

Table 10.1: Modellparameter und Standardfehler

	Schätzer	s_e
(Intercept)	-0.14	0.23
v_ms	0.76	0.02

Der vorhergesagte Wert $\hat{y}$ ist daher für sich alleine ist noch nicht brauchbar, da auch Informationen über dessen Unsicherheit notwendig sind um die Ergebnisse korrekt zu interpretieren.

Es können zwei unterschiedliche Anwendungsfälle voneinander unterschieden werden.

1. Der mittlere, erwartete Wert $\hat{\bar{y}}_{neu}$
2. Die Vorhersage eines einzelnen Wertes $\bar{y}_{neu}$

Im konkreten Fall werden damit zwei unterschiedliche Fragestellungen beantwortet. Im 1. Fall lautet die Frage, ich habe eine Trainingsgruppe und möchte wissen was der mittlere Wert der Gruppe anhand des Modells ist, wenn alle eine bestimmte Anlaufgeschwindigkeit v_{neu} haben. Im 2. Fall lautet die Frage welche Weite eine einzelne Athletin für die Anlaufgeschwindigkeit v_{neu} springen sollte. In beiden Fällen werden keiner genau den Wert des Regressionsmodells treffen, aber im 1. Fall der Gruppe werden sich Streuungen nach oben bzw. nach unten gegenseitig im Schnitt ausbalancieren während im 2. Fall der einzelnen Athletin dies nicht der Fall ist. Daher hat die Vorhersage im 2. Fall eine höhere Unsicherheit. Diese Unterschied sollte sich dementsprechend in den Varianzen der beiden Vorhersagen wiederspiegeln.

Wie bereits erwähnt, der vorhergesagte Wert $\hat{y}_{neu}$ ist in beiden Fällen gleich und entsprecht der oben beschriebenen Methode anhand des Modell $y_{neu} = \hat{\beta}_0 + \hat{\beta}_1 \times x_{\text{neu}}$.

Für den erwarteten Mittelwert errechnet sich die Varianz nach:

$$Var(\hat{\bar{y}}_{neu}) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_{neu} - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] = \hat{\sigma}^2_{\hat{\bar{y}}_{neu}} \tag{10.1}$$

Das dazugehörige Konfidenzintervall errechnet sich danach mittels:

$$\hat{\bar{y}}_{neu} \pm q_{t(1-\alpha/2;n-2)} \times \hat{\sigma}_{\hat{\bar{y}}_{neu}} \tag{10.2}$$

Die Varianz für die Vorhersage eines einzelnen Wertes errechnet sich:

$$Var(\hat{y}_{neu}) = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_{neu} - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] = \hat{\sigma}^2 + \hat{\sigma}^2_{\hat{\bar{y}}_{neu}} = \hat{\sigma}^2_{\hat{y}_{neu}} \tag{10.3}$$

Was wiederum zu dem folgenden Konfidenzintervall führt:

$$\hat{y}_{neu} \pm q_{t(1-\alpha/2;n-2)} \times \hat{\sigma}_{\hat{y}_{neu}} \tag{10.4}$$

In beiden Fällen ist der Term

$$\frac{(x_{neu} - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

enthalten. Anhand des Zählers kann abgeleitet werden, dass die Unsicherheit der Vorhersage mit dem Abstand vom Mittelwert der x-Werte zunimmt. Rein heuristisch macht dies Sinn, da davon ausgegangen werden kann, dass um den Mittelwert der x-Werte auch die meiste Information über y vorhanden ist und dementsprechend umso weiter die Werte sich vom $\bar{x}$ entfernen die Information abnimmt. Im Nenner ist wiederum wie auch beim Standardfehler σ_{β_1} des Steigungskoeffizienten β_1 zu sehen, dass die Varianz abnimmt mit der Streuung der x-Werte. Daher, wenn eine Vorhersage in einem bestimmten Bereich von x-Werten durchgeführt werden soll, dann sollte darauf geachtet werden möglichst diesen Bereich auch zu samplen um die Unsicherheit so klein wie möglich zu halten.

10.3 Vorhersagen in R mit `predict()`

10.3.1 Erwarteter Mittelwert

```
df <- data.frame(v_ms = 11.5) # oder tibble(v_ms = 11.5)
predict(mod, newdata = df, interval = 'confidence')
```

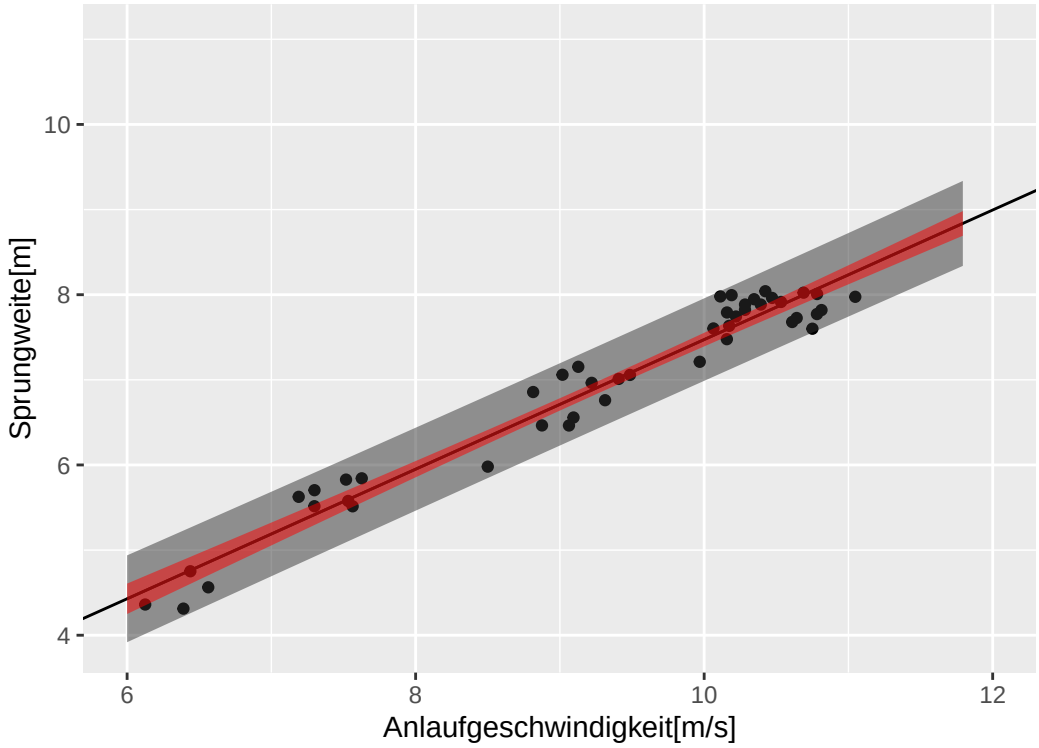
```
       fit      lwr      upr
1 8.614136 8.482039 8.746234
```

10.3.2 Individuelle Werte

```
predict(mod, newdata = df, interval = 'prediction')
```

```
       fit      lwr      upr
1 8.614136 8.118445 9.109827
```

10.4 Konfidenzintervalle graphisch



Weiterführende Literatur sind Kutner et al. (2005)

10.5 R^2 und Root-mean-square

10.6 Einfaches Modell

```
mod0 <- lm(y ~ x, simple)
summary(mod0)
```

```
Call:
lm(formula = y ~ x, data = simple)

Residuals:
      1       2       3       4
-0.5817  0.9898 -0.2345 -0.1736

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8414     0.7008   2.628    0.119
x             0.4574     0.3746   1.221    0.346

Residual standard error: 0.8376 on 2 degrees of freedom
Multiple R-squared:  0.4271,    Adjusted R-squared:  0.1406
F-statistic: 1.491 on 1 and 2 DF,  p-value: 0.3465
```

10.7 Nochmal Abweichungen

1. **Gesamtvarianz**:
$$SSTO := \sum_{i=1}^{N} (y_i - \bar{y})^2$$
2. **Regressionsvarianz**:
$$SSR := \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$
3. **Residualvarianz**:
$$SSE := \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

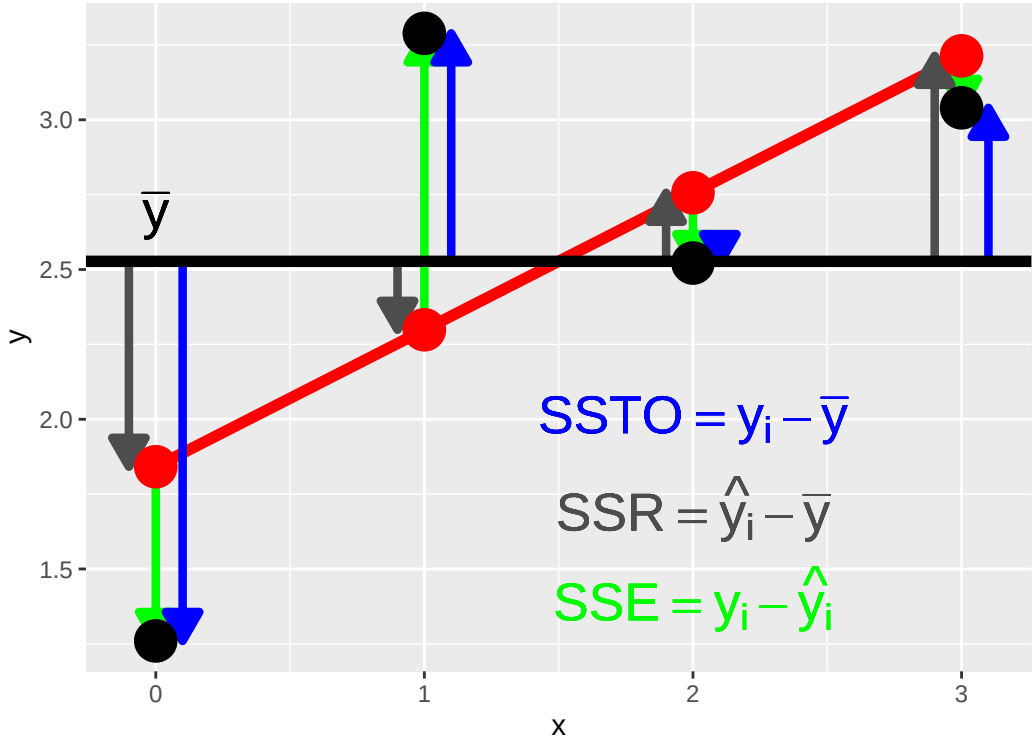


Figure 10.2: Minimalmodell der Abweichungen

10.8 Verhältnis von SSR zu $SSTO$

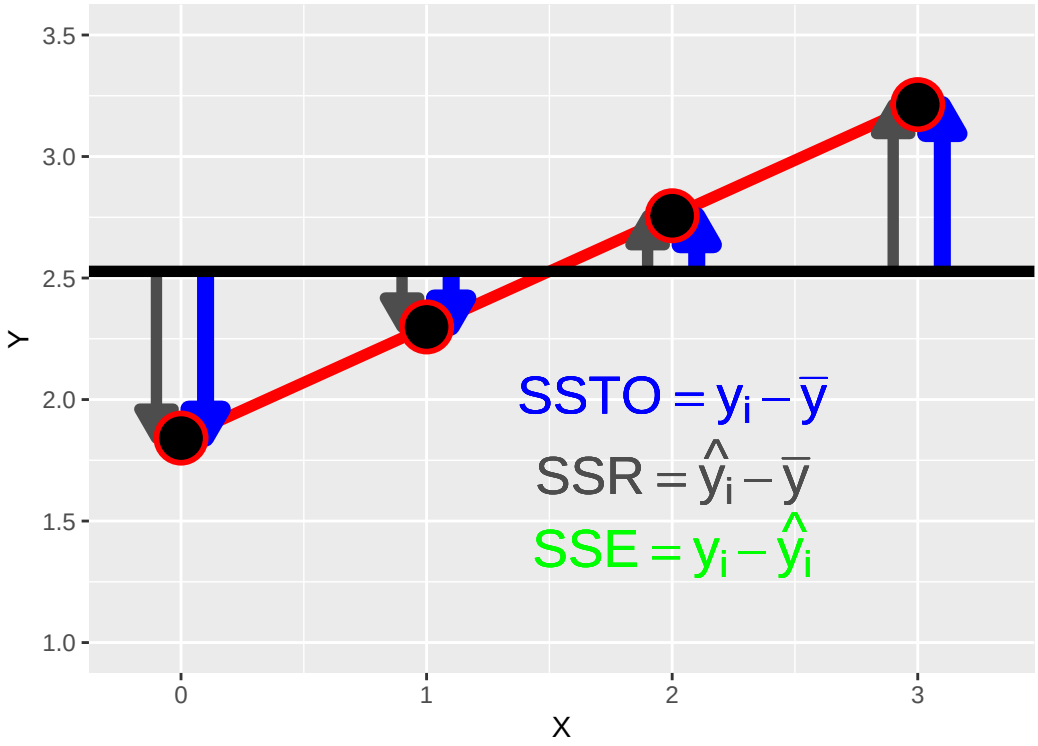


Figure 10.3: Perfekter Zusammenhang

$$\frac{SSR}{SSTO} = 1$$

$$\frac{SSR}{SSTO} = 0$$

10.9 Determinationskoeffizient R^2

Es gilt: $SSTO = SSR + SSE$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \in [0, 1]$$

[1]

[1] Bei der einfachen Regression gilt: $r_{xy} = \pm\sqrt{R^2}$

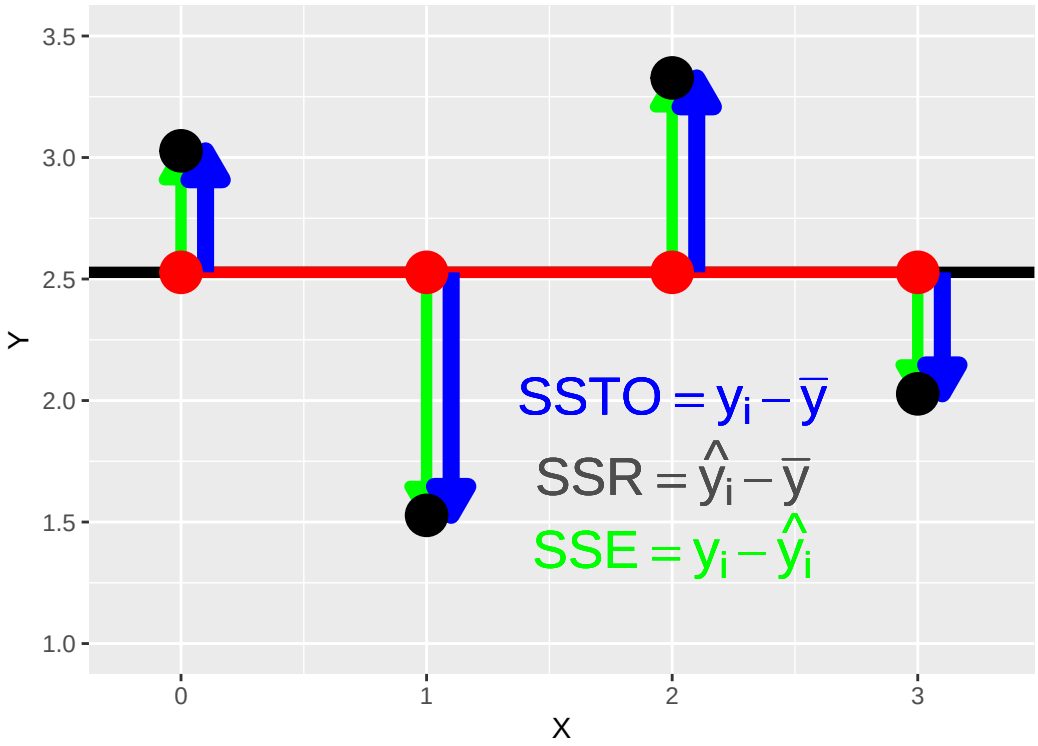


Figure 10.4: Kein Zusammenhang

10.9.1 Korrigierter Determinationskoeffizient R_a^2

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}$$

Part III

Multiple Regression

Im folgenden wird das Modell der einfachen linearen Regression erweitert indem zusätzliche Terme in das Modell aufgenommen werden. Die Prinzipien bleiben dabei jedoch weitestgehendst gleich und können direkt auf den komplizierteren Fall der multiplen Regression übertragen werden. Im Laufe der Erweiterung des Modells wird sich dabei wird herausstellen, dass neben mehreren kontinuierlichen Variablen auch nominale Faktoren in das Modell intergriert werden können. Daraus entsteht ein sehr flexibler Modellapparat, der in den verschiedensten Zusammenhängen angewendet werden kann.

11 Einführung

In vielen Fällen in der Praxis liegt selten der einfache Fall vor, dass eine abhängige Variable mitels nur einer einzigen Variable erklärt bzw. vorhergesagt werden soll. Sondern meisten sind mehrere Variablen an dem Prozess der modelliert werden soll beteiligt. Ein einfaches Beispiel aus der Literatur ist der Zusammenhang zwischen der Wurfgeschwindigkeit beim Handball in Abhängigkeit vom Körpergewicht und der Armspannweite. In Table 11.1 ist ein Ausschnitt aus einem möglichen Datensatz abgebildet.

Table 11.1: Datenausschnitt: Wurfgeschwindigkeit, Körpermasse und Armspannweite bei professionellen Handballern (angelehnt an Debanne & Laffaye, 2011).

Velocity[m/s]	body mass[kg]	arm span[cm]
15.8	70.7	189.2
17.2	63.7	182.0
18.3	76.2	192.1
18.4	64.9	171.1
18.4	63.0	181.1

Im Prinzip könnte der isolierte Einfluss der beiden Prädiktorvariablen Körpermasse und Armspannweite auf die Wurfgeschwindigkeit untersucht werden. Allerdings ist den meisten Fällen von größerem Interesse wie sich die beiden Variablen zusammen verhalten und ob durch die Kombination der beiden Variablen ein besseres Modell der Daten erstellt werden kann.

Aus dieser Problemstellung heraus ergibt sich die Notwendigkeit von der einfachen linearen Regression auf eine multiple multiple lineare Regression überzugehen. Formal, geschieht dies einfach dadurch, dass die Formel der einfachen Regression mit dem Prädiktor x um eine zweite Variable erweitert wird.

Dementsprechend wird aus:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{11.1}$$

die Formel für die multiple Regression mit:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i \tag{11.2}$$

Da bei der einfachen Regression nur eine einzige x-Variable in der Formel vorhanden war, ist kein zusätzlicher Index notwendig gewesen, bei der mutliplen Regression mit mehreren Prädiktorvariablen x wird jeder x Variabler ein zusätzlicher Index j angehängt um die Variablen eindeutig zu identifizieren. Per Konvention, wobei diese leider nicht global eingehalten wird, wird die Anzahl der Prädiktorvaiablen mit K bezeichnet. Der y-Achsenabschnitt erhält den Index $j = 0$ und die weiteren Steigungskoeffzienten β_1 bis β_K erhalten den Prädiktorvariablen x_j entsprechden Index.

In welcher Reihenfolge die Prädiktorvariablen mit $j = 1, j = 2, \ldots, j = K$ verteilt werden hat zunächst keine Auswirkung auf das Modell und regelt lediglich die Bezeichnung. In unserem konkreten Fall der Handballwurfdaten wäre zum Beispiel eine mögliche Zuordnung, das x_1 die Körpermasse und x_2 die Armspannweite kodiert.

i	Velocity[m/s]	body mass[kg] $j = 1$	arm span[cm] $j = 2$
1	15.8	70.7	189.2
2	17.2	63.7	182.0

i	Velocity[m/s]	body mass[kg] $j = 1$	arm span[cm] $j = 2$
3	18.3	76.2	192.1
4	18.4	64.9	171.1
5	18.4	63.0	181.1

Rein formal haben wir jetzt schon den Übergang zur multiple Regression vollzogen. Die Frage die sich natürlich direkt anschließt bezieht sich nun auf die Bedeutung der Koeffizienten $\beta_1, \dots, \beta_k$.

11.1 Bedeutung der Koeffizienten bei der multiplen Regression

Um die Bedeutung der Regressionskoeffzienten bei der multiple Regression besser zu verstehen ist es von Vorteil sich noch einmal die Bedeutung der Koeffizienten im einfachen Regressionsmodell zu vergegenwärtigen (siehe Figure 11.1).

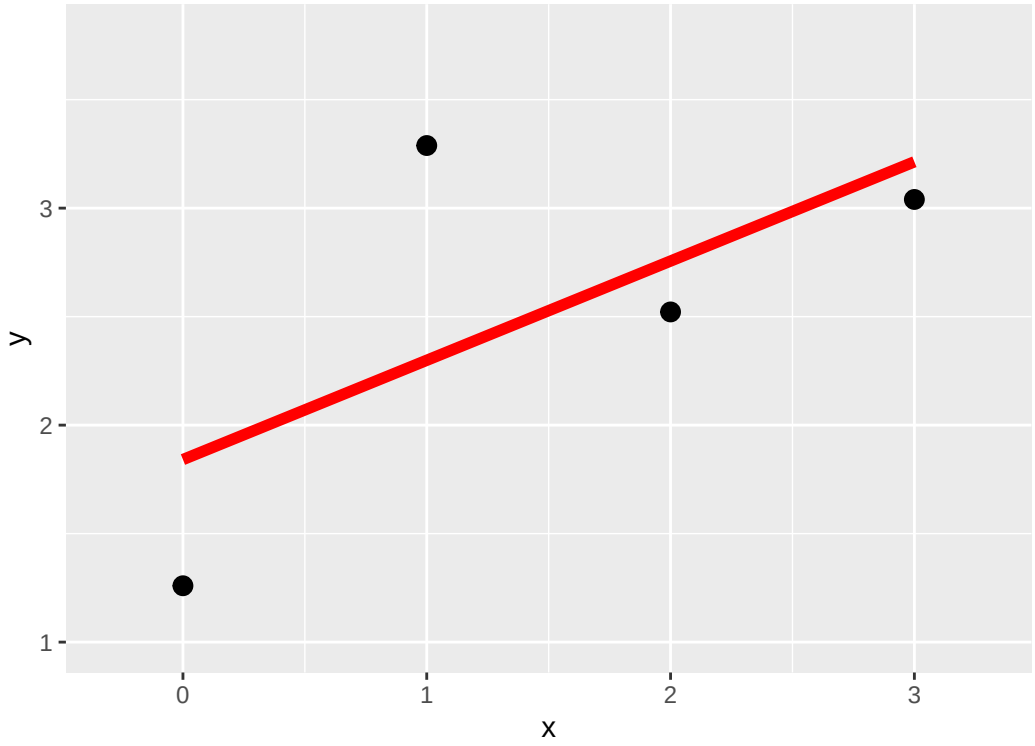


Figure 11.1: Beispiel für eine einfache Regression und der resultierenden Regressiongeraden

Bei der einfachen Regression haben mittels der Methode der kleinsten Quadrate eine Regressiongerade durch unsere Punktwolke gelegt. Dabei haben wir die Regressionsgerade so gewählt, dass die senkrechten Abstände der beobachteten Punkte von der Regressionsgerade minimiert werden bzw. die Abstände zwischen denen auf der Gerade liegenden, vorhergesagten Werte $\hat{y}_i$ und den beobachteten Wert y_i.

Wenn wir nun den Übergang von einer Prädiktorvariablenzum nächstkomplizierteren Fall nehmen mit zwei Prädiktorvariablen x_1 und x_2, dann wäre eine mögliche Darstellungsform der Daten eine Punktwolke im dreidimensionalen Raum (siehe Figure 11.2a).

(a) 3D Punktwolke

(b) 3D Punktwolke mit gefitteter Ebene

Figure 11.2: Punktwolken bei der multiple Regression

Da jetzt eine einzelne Gerade nicht mehr in der Lage ist die Daten zu fitten, ist die nächst Möglichkeit eine Ebene die in die Punktwolke gelegt wird (siehe Figure 11.2b). Dies ermöglicht dann genau die gleiche Herangehensweise wie bei der einfachen linearen Regression anzuwenden. Als Zielgröße wird aus den möglichen Ebenen diejenigen gesucht deren vorhergesagten, auf der Ebene liegenden Punkte $\hat{y}_i$ die geringsten senkrechten Abstand zu den beobachteten Punkten y_i haben. Anders, wir suchen diejenigen Ebene durch die Punktwolke deren Summe der quadrierten Residuen $e_i = y_i - \hat{y}_i$ minimal ist.

Diese Herangehensweise hat den Vorteil, dass sie zum einem die einfache lineare Regression als Spezialfall mit $K = 1$ beinhaltet und sich beliebig erweitern lässt mit der Einschränkung, dass bei $K > 2$ die dreidimenionale Darstellung mittels einer Grafik nicht mehr möglich ist. Das Prinzip der Minimierung der Abweichungen von $\hat{y}_i$ zu y bleibt aber immer erhalten. Zusammenfassend hat dieser Ansatz somit die folgenden Vorteile:

- Die Berechnungen bleiben alle gleich
- Abweichungen $\hat{\epsilon_i}$ sind jetzt nicht mehr Abweichungen von einer Gerade sondern von einer K-dimensionalen Hyperebene. Die Eigenschaften der Residuen bleiben aber alle erhalten.
- Die Modellannahmen bleiben gleich: Unabhängige y_i und $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ iid
- Inferenz für die Koeffizienten mittels $t_k = \frac{\hat{\beta}_k}{s_k} \sim t(N - K - 1)$ (Konfidenzintervall dito)
- Konzepte für die Vorhersage bleiben erhalten
- Modelldiagnosetools bleiben alle erhalten

Als nächster Schritt versuchen wir nun die Interpretation der Koeffizienten im multiplen Regressionsmodell besser zu verstehen.

11.2 Einfaches Beispiel

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \epsilon_i \\
\beta_0 &= 1, \beta_1 = 3, \beta_2 = 0.7 \\
\epsilon_i &\sim N(0, \sigma = 0.5)
\end{aligned}
$$

```
N <- 50 # Anzahl Datenpunkte
beta_0 <- 1
```

```r
beta_1 <- 3
beta_2 <- 0.7
sigma <- 0.5
set.seed(123)
df <- tibble(
  x1 = runif(N, -2, 2),
  x2 = runif(N, -2, 2),
  y = beta_0 + beta_1*x1 + beta_2*x2 +
    rnorm(N, 0, sigma))
```

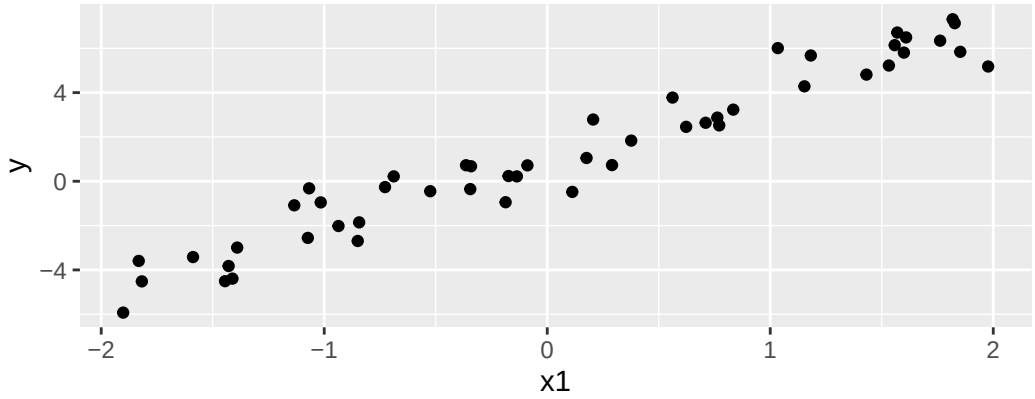


Figure 11.3: Einfacher Zusammenhang y~x1

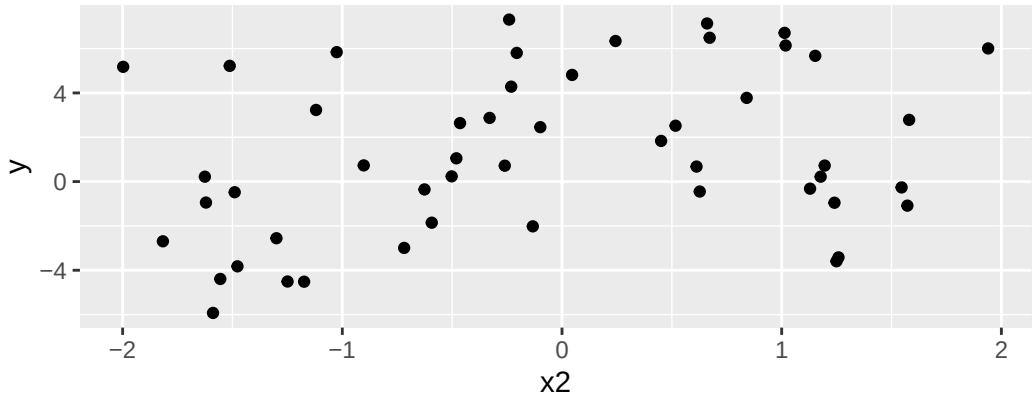


Figure 11.4: Einfacher Zusammenhang y~x2

11.3 Wie sieht der Fit aus?

```
Call:
lm(formula = y ~ x1 + x2, data = df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.20883 -0.26741 -0.00591  0.27315  1.01322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.07674    0.06552   16.43  < 2e-16 ***
x1           2.96537    0.05604   52.91  < 2e-16 ***
x2           0.70815    0.05961   11.88 9.27e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4604 on 47 degrees of freedom
Multiple R-squared:  0.9849,    Adjusted R-squared:  0.9842
F-statistic:  1529 on 2 and 47 DF,  p-value: < 2.2e-16
```

11.4 Was bedeuten die einzelnen Koeffizienten?

Table 11.3: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	1.077	0.066
x1	2.965	0.056
x2	0.708	0.060

Der Unterschied in der abhängigen Variablen, wenn zwei Objekte sich in x_i um eine Einheit unterscheiden und die paarweise gleichen Werte in den verbleibenden $x_j, j \neq i$ annehmen.

11.5 Was bedeuten die Koeffizienten in Kombination?

11.5.1 Full model

Table 11.4: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	1.077	0.066
x1	2.965	0.056
x2	0.708	0.060

11.5.2 um x2 bereinigt

```r
mod_x1_x2 <- lm(x1 ~ x2, df)
res_mod_x1_x2 <- resid(mod_x1_x2)
mod_x1_res <- lm(y ~ res_mod_x1_x2, df)
```

```
             Estimate Std. Error t value
(Intercept)      1.25       0.16    7.61
res_mod_x1_x2    2.97       0.14   20.97
```

11.5.3 um x1 bereinigt

```
mod_x2_x1 <- lm(x2 ~ x1, df)
res_mod_x2_x1 <- resid(mod_x2_x1)
mod_x2_res <- lm(y ~ res_mod_x2_x1, df)
```

```
              Estimate Std. Error t value
(Intercept)       1.25       0.51    2.44
res_mod_x2_x1     0.71       0.47    1.51
```

11.6 Was bedeuten die Koeffizienten in Kombination?

- $\hat{\beta}_1$: Wenn ich x_2 weiß, welche zusätzlichen Informationen bekomme ich durch x_1
- $\hat{\beta}_2$: Wenn ich x_1 weiß, welche zusätzlichen Informationen bekomme ich durch x_2

In Beispiel nicht problematisch, weil nach Konstruktion x_1 und x_2 unabhängig voneinander sind:

```
round(cor(df),3)
```

```
      x1    x2     y
x1 1.000 0.078 0.969
x2 0.078 1.000 0.289
y  0.969 0.289 1.000
```

11.7 Added-variable plots

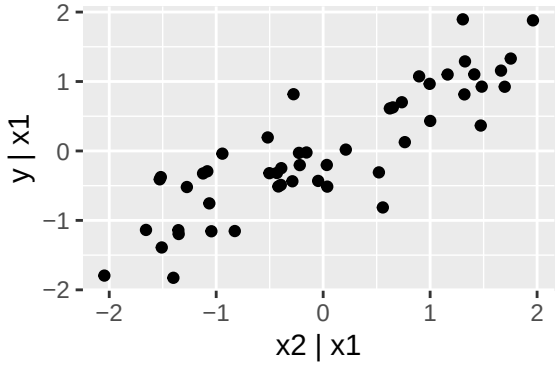
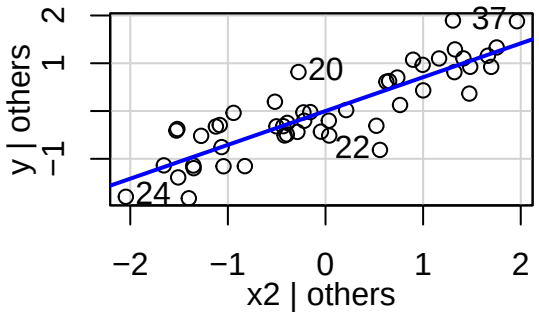


Figure 11.5: Zusammenhang zwischen y und x2 bereinigt um den Einfluß von x1.

11.8 Added-variable plots mit `car::avPlots()`

```
car::avPlots(mod, ~x2)
```



11.9 Was passiert wenn ich einen Prädiktor weg lasse?

Table 11.5: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	1.077	0.066
x1	2.965	0.056
x2	0.708	0.060

```
coef(lm(y ~ x1, df))
```

```
(Intercept)          x1
   1.007466    3.017589
```

```
coef(lm(y ~ x2, df))
```

```
(Intercept)          x2
  1.3377771   0.9555316
```

In unserem Beispiel wieder nicht viel, da die Variablen unabhängig (orthogonal) voneinander sind.

11.10 Was passiert wenn Prädiktoren stark miteinander korrelieren?

Table 11.6: Ausschnitt von Körperfettdaten

triceps	thigh	midarm	body_fat
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7

[1]

11.11 Was passiert wenn Prädiktoren stark miteinander korrelieren?

```
GGally::ggpairs(bodyfat) + theme(text = element_text(size = 10))
```

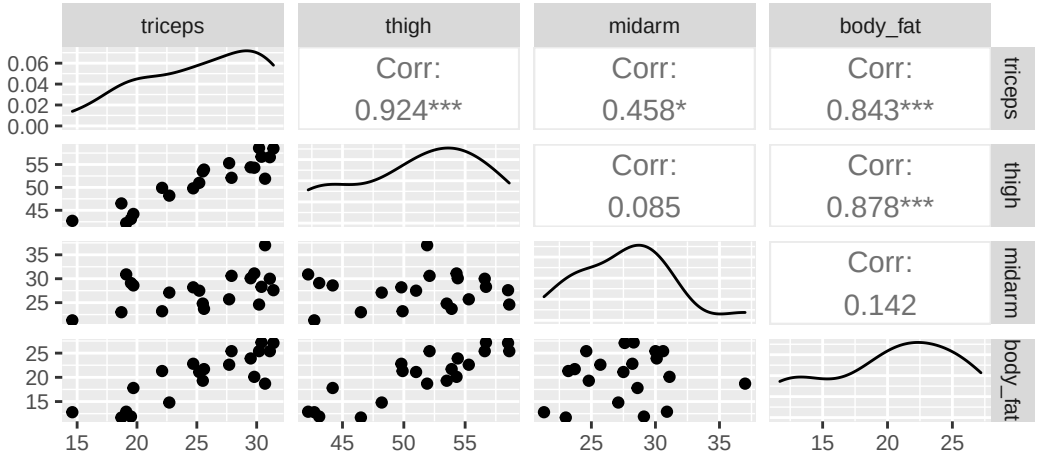


Figure 11.6: Korrelationsmatrize

11.12 Was passiert wenn Prädiktoren stark miteinander korrelieren?

```
# Alle drei Prädiktoren
mod_full <- lm(body_fat ~ triceps + thigh + midarm, bodyfat)
# ohne Arm
mod_wo_midarm <- lm(body_fat ~ triceps + thigh, bodyfat)
```

[1] Beispiel nach Kutner et al. (2005)

```r
# Ohne Oberschenkel
mod_wo_thigh <- lm(body_fat ~ triceps + midarm, bodyfat)
# Ohne Triceps
mod_wo_triceps <- lm(body_fat ~ thigh + midarm, bodyfat)
```

11.13 Was passiert wenn Prädiktoren stark miteinander korrelieren?

Table 11.7: full model

	$\hat{\beta}$	s_e
(Intercept)	117.085	99.782
triceps	4.334	3.016
thigh	-2.857	2.582
midarm	-2.186	1.595

Table 11.8: w/o midarm

	$\hat{\beta}$	s_e
(Intercept)	-19.174	8.361
triceps	0.222	0.303
thigh	0.659	0.291

Table 11.9: w/o thigh

	$\hat{\beta}$	s_e
(Intercept)	6.792	4.488
triceps	1.001	0.128
midarm	-0.431	0.177

Table 11.10: w/o triceps

	$\hat{\beta}$	s_e
(Intercept)	-25.997	6.997
thigh	0.851	0.112
midarm	0.096	0.161

11.14 Multikollinearität[2]

- Große Änderungen in den Koeffizienten wenn Prädiktoren ausgelassen/eingefügt werden
- Koeffizienten haben eine andere Richtung als erwartet
- Hohe (einfache) Korrelationen zwischen Prädiktoren

[2] informell nach Kutner et al. (2005, 407)

- Breite Konfidenzintervalle für "wichtige" Prädiktoren b_j

$$\widehat{\text{Var}}(b_j) = \frac{\hat{\sigma}^2}{(n-1)s_j^2} \frac{1}{1-R_j^2}$$

R_j^2 = Multipler Korrelationskoeffizient der Prädiktoren auf Prädiktorvariable j.

11.15 Variance Inflation Factor (VIF)

$$\text{VIF}_j = \frac{1}{1-R_j^2}$$

> **Tip**
>
> Wenn VIF > 10 ist, dann deutet dies auf hohe Multikollinearität hin.

[3]

11.16 Variance Inflation Factor (VIF)

```
car::vif(mod_full)
```

```
 triceps    thigh   midarm
708.8429 564.3434 104.6060
```

[4]

Üblicherweise wird der größte Wert betrachtet um die Multikollinearität zu bewerten.

11.17 Wenn Prädiktoren sich gegenseitig maskieren[5]

11.18 Wenn Prädiktoren sich gegenseitig maskieren

[3] Manchmal wird auch Tolerance $= \frac{1}{VIF}$ betrachtet.
[4] car::vif berechnet generalized variance inflation factor wenn Prädiktoren Faktoren oder Polynome sind (Fox 2011.)
[5] adaptiert nach McElreath (2016)

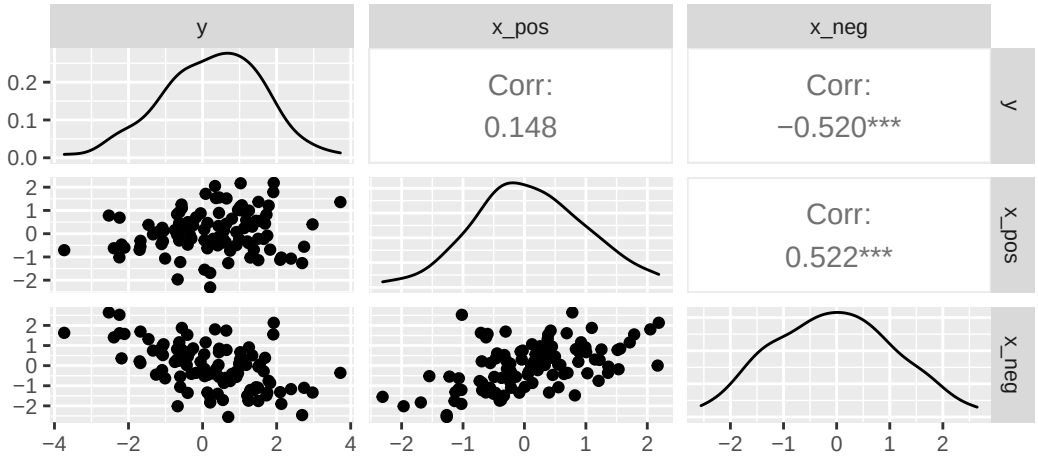


Figure 11.7: x_pos maskiert den Einfluss von x_neg

Table 11.11: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	0.235	0.135
x_pos	0.218	0.147

Table 11.12: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	0.228	0.116
x_neg	-0.618	0.103

Table 11.13: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	0.135	0.096
x_pos	0.850	0.123
x_neg	-0.976	0.099

11.19 Multiple Regression

Aus der einfachen Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

wird

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i$$

mit K Prädiktorvariablen und Multikollinearität.

11.20 Zum Nacharbeiten

N. Altman and Krzywinski (2015a)
Kutner et al. (2005, 278–88)
Fox (2011, 325–27)

12 Interaktionseffekte

12.1 Beispieldaten[1]

Table 12.1: Beispieldaten (synthetisch)

Velocity[m/s]	body mass[kg]	arm span[cm]
185.42	68.71	20.14
184.08	73.85	21.29
200.74	89.43	27.57
170.34	84.97	19.88
176.89	82.40	20.51
200.68	91.57	29.22

12.2 Beispieldaten - Deskriptiv

Table 12.2: Deskriptive Statistik der Handballdaten

	Mean	Std.Dev	Min	Max
arm_span	184.3	7.7	169.4	200.7
body_mass	77.5	10.3	58.0	101.1
vel	21.9	2.3	18.5	29.2

12.3 Beispieldaten

12.4 Beispieldaten - Startmodell

$$Y_i = \beta_0 + \beta_1 \times \text{bm}_i + \beta_2 \times \text{as}_i + \epsilon_i$$

```
mod_1 <- lm(vel ~ body_mass + arm_span, handball)
```

[1] Debanne and Laffaye (2011)

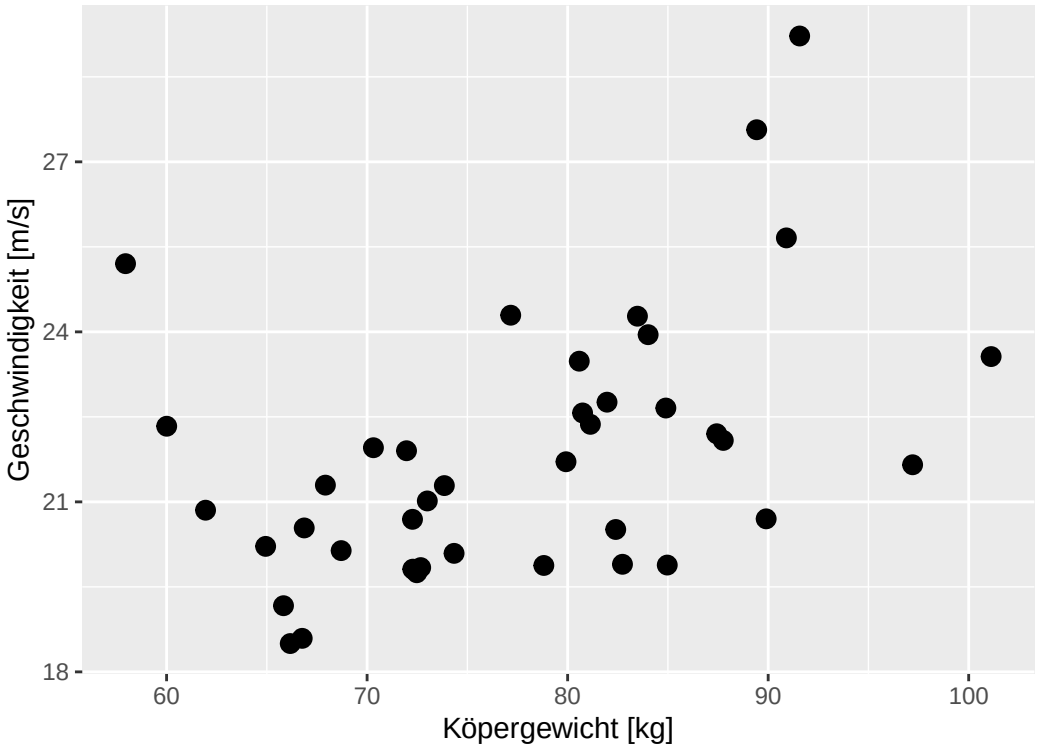


Figure 12.1: Geschwindigkeit gegen Körpergewicht

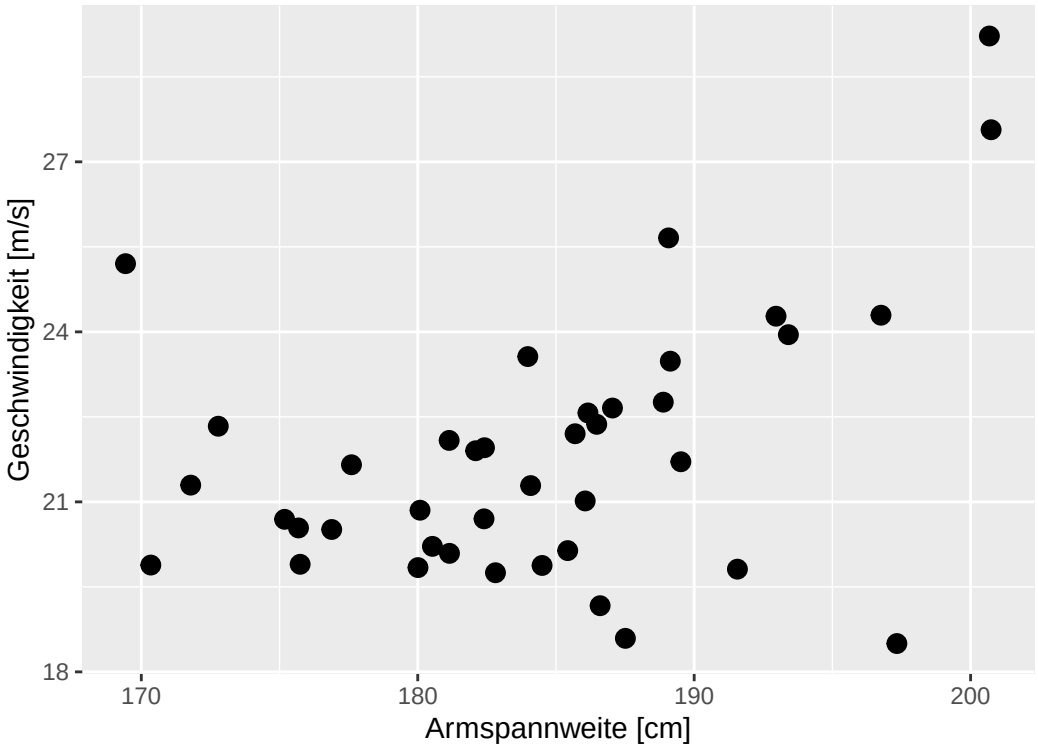


Figure 12.2: Geschwindigkeit gegen Armspannweite

Table 12.3: Modell 1

	$\hat{\beta}$	s_e	t	p
(Intercept)	-1.768	7.632	-0.232	0.818
body_mass	0.077	0.033	2.359	0.024
arm_span	0.096	0.044	2.192	0.035
$\hat{\sigma}$	1.996			

12.5 Modellfit

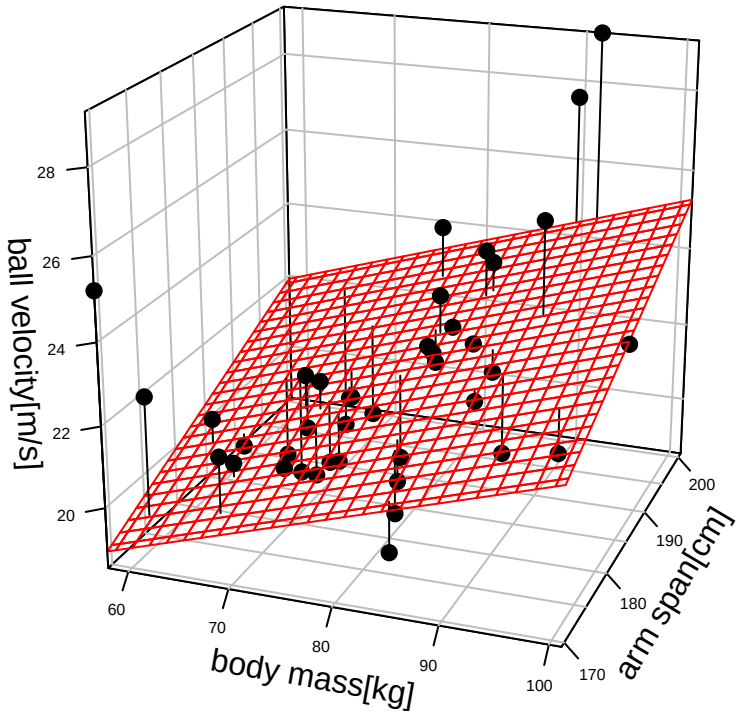


Figure 12.3: 3D Streudiagramm

12.6 Zentrierung

```
handball <- dplyr::mutate(handball,
                          body_mass_c = body_mass - mean(body_mass),
                          arm_span_c = arm_span - mean(arm_span))
```

Table 12.4: Deskriptive Statistik

	Mean	Std.Dev
arm_span	184.29	7.72
arm_span_c	0.00	7.72
body_mass	77.46	10.26
body_mass_c	0.00	10.26
vel	21.85	2.31

12.7 Modell mit zentrierten Variablen

```
mod_2 <- lm(vel ~ body_mass_c + arm_span_c, handball)
```

Table 12.5: Modell 2

	$\hat{\beta}$	s_e	t	p
(Intercept)	21.852	0.316	69.247	<0.001
body_mass_c	0.077	0.033	2.359	0.024
arm_span_c	0.096	0.044	2.192	0.035
$\hat{\sigma}$	1.996			

12.8 Residuen im zentrierten, additiven Modell

12.9 Added-variable plot

12.10 Was passiert wenn die Effekte nicht mehr nur additiv sind?

12.11 Was passiert wenn die Effekte nicht mehr nur additiv sind?

12.11.1 Neues Modell mit Interaktionen:

$$Y_i = \beta_0 + \beta_1 \times \text{bm}_i + \beta_2 \times \text{as}_i + \beta_3 \times \text{bm}_i \times \text{as}_i + \epsilon_i$$

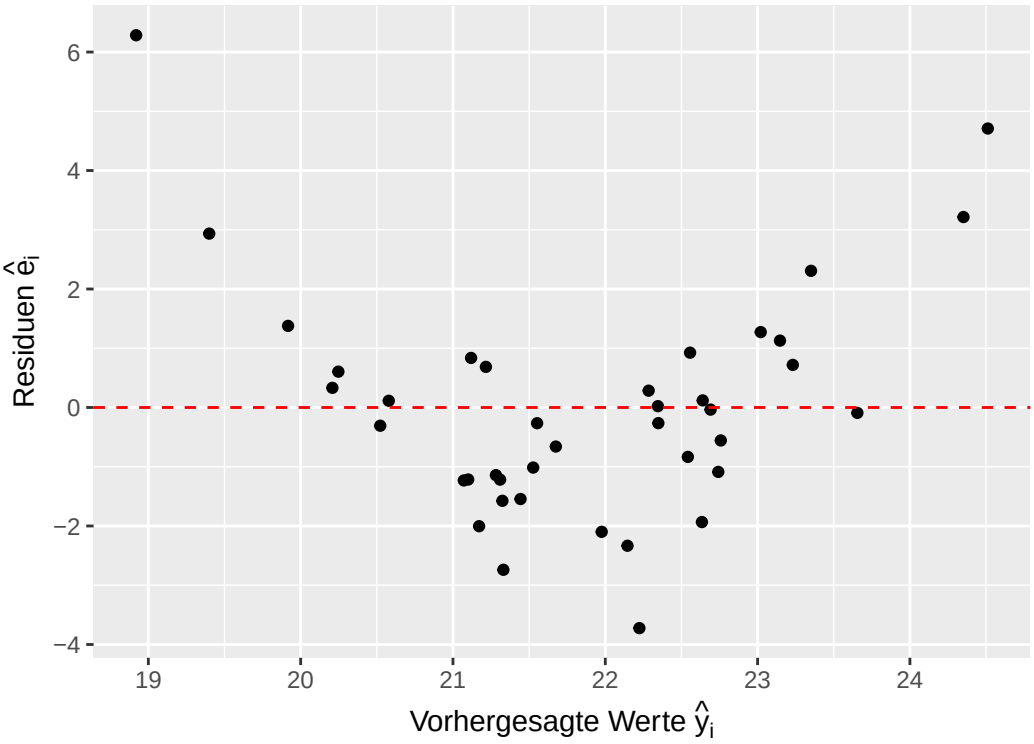


Figure 12.4: Residuenplot

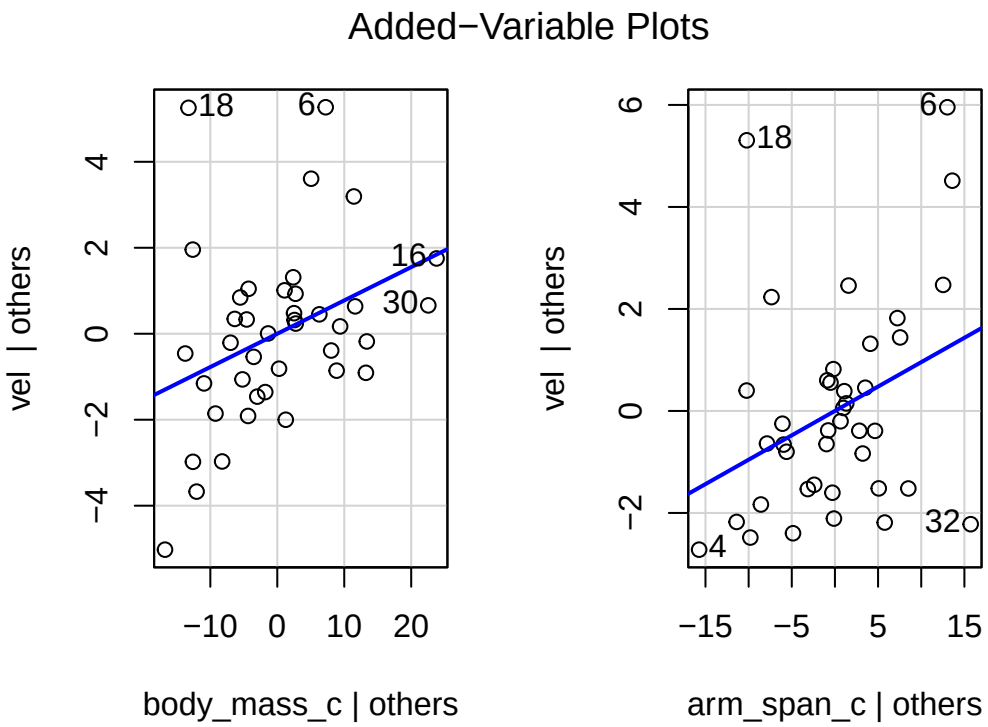


Figure 12.5: Added-variable Graph mit `car::avPlots()`

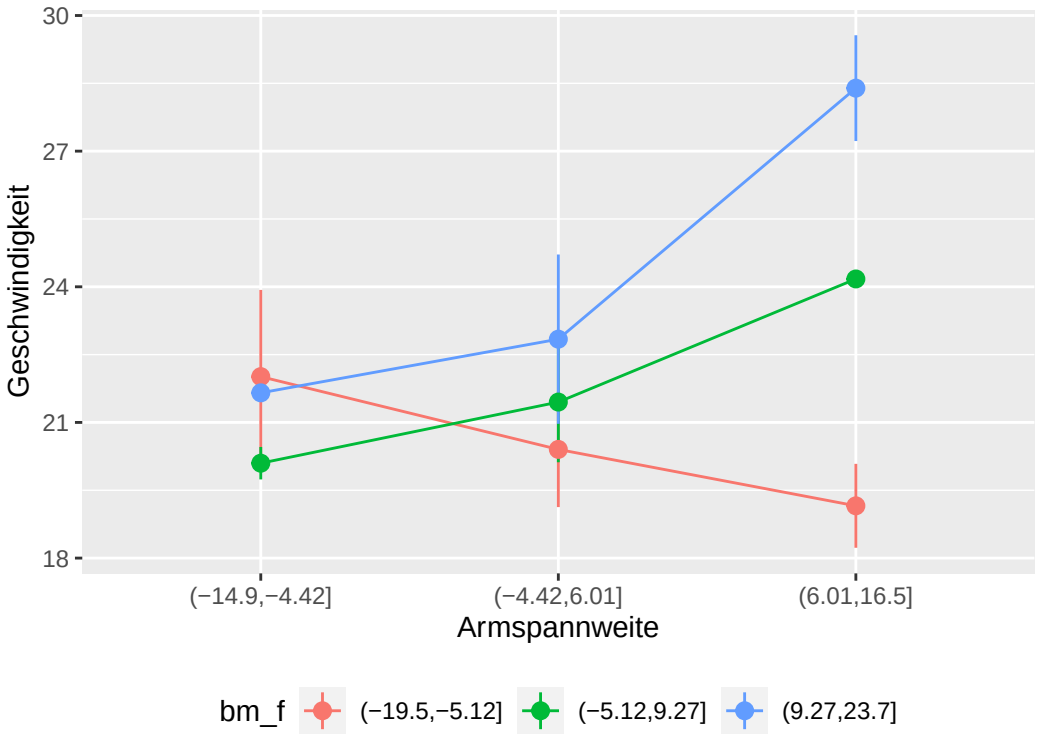


Figure 12.6: Unterteilung von Körpergewicht und Armspannweite in Kategorien

12.12 Modellierung

```
mod_3 <- lm(vel ~ body_mass_c * arm_span_c, handball)
```

Table 12.6: Modell 3

	$\hat{\beta}$	s_e	t	p
(Intercept)	21.346	0.143	149.296	<0.001
body_mass_c	0.119	0.015	8.133	<0.001
arm_span_c	0.083	0.019	4.380	<0.001
body_mass_c:arm_span_c	0.021	0.002	12.633	<0.001
$\hat{\sigma}$	0.868			

[2]

12.13 Einfache Steigungen in Vergleich

12.14 Interaktionen sind symmetrisch

12.15 Warum das Model Sinn macht

Table 12.7: Einfache Steigungen

arm span\centered	β_0	β_1
10	22.18	0.33
0	21.35	0.12
-10	20.51	-0.09

12.16 Warum das Modell Sinn macht

Table 12.8: Einfache Steigungen

arm span\centered	β_0	β_1
10	22.18	0.33
0	21.35	0.12
-10	20.51	-0.09

[2] `A*B wird von R ausmultipliziert in A + B + A:B`. Hätte auch `lm(vel ~ body_mass_c + arm_span_c + body_mass_c:arm_span_c)` verwenden können.

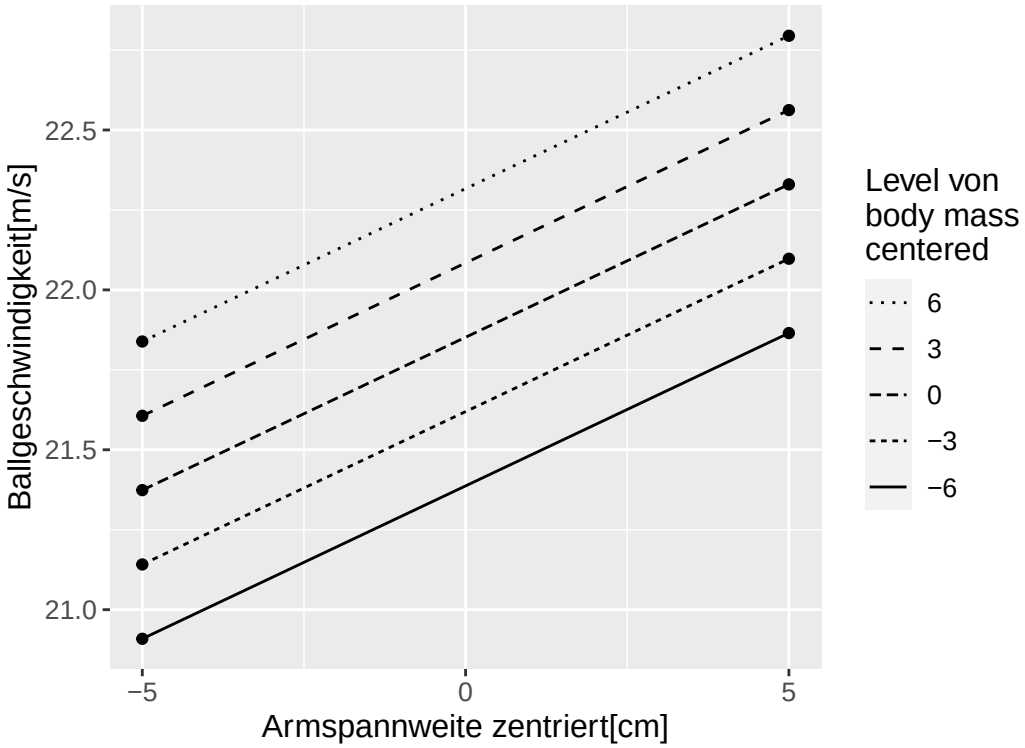


Figure 12.7: Modell ohne Interaktionen

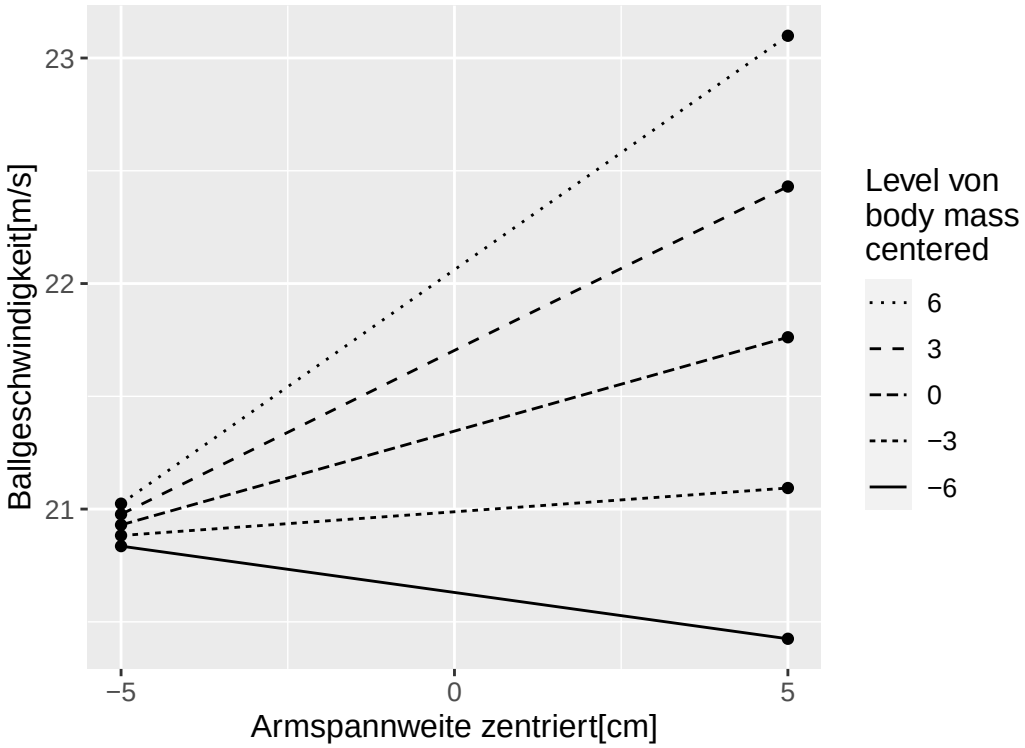


Figure 12.8: Modell mit Interaktionen

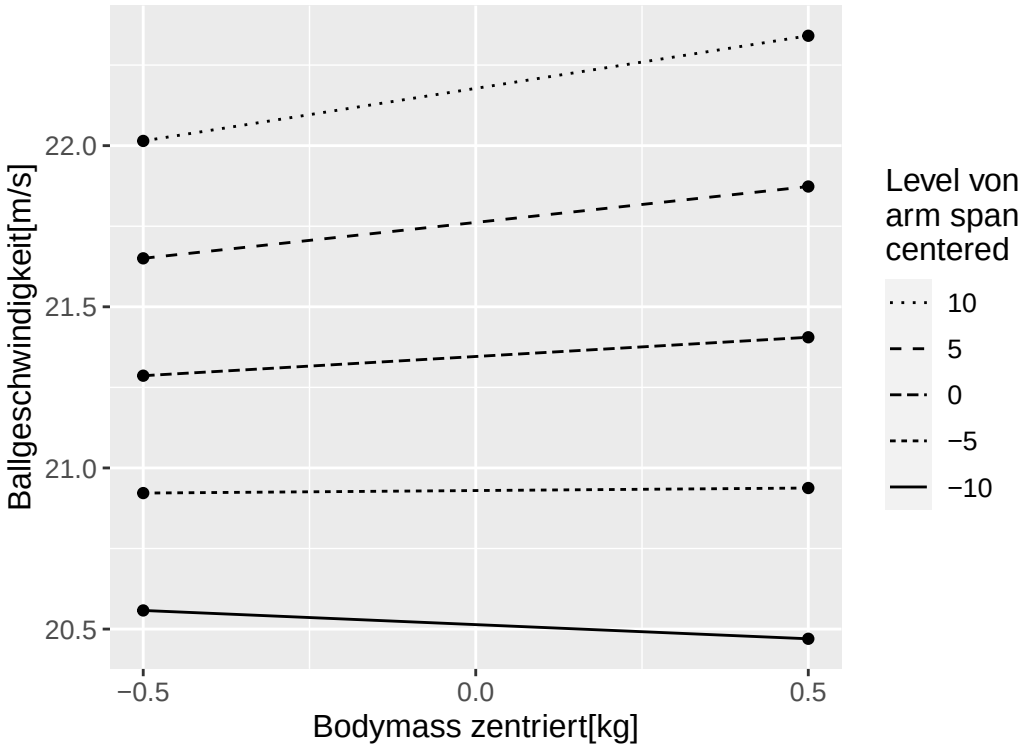


Figure 12.9: Veränderung mit der Körpergewicht

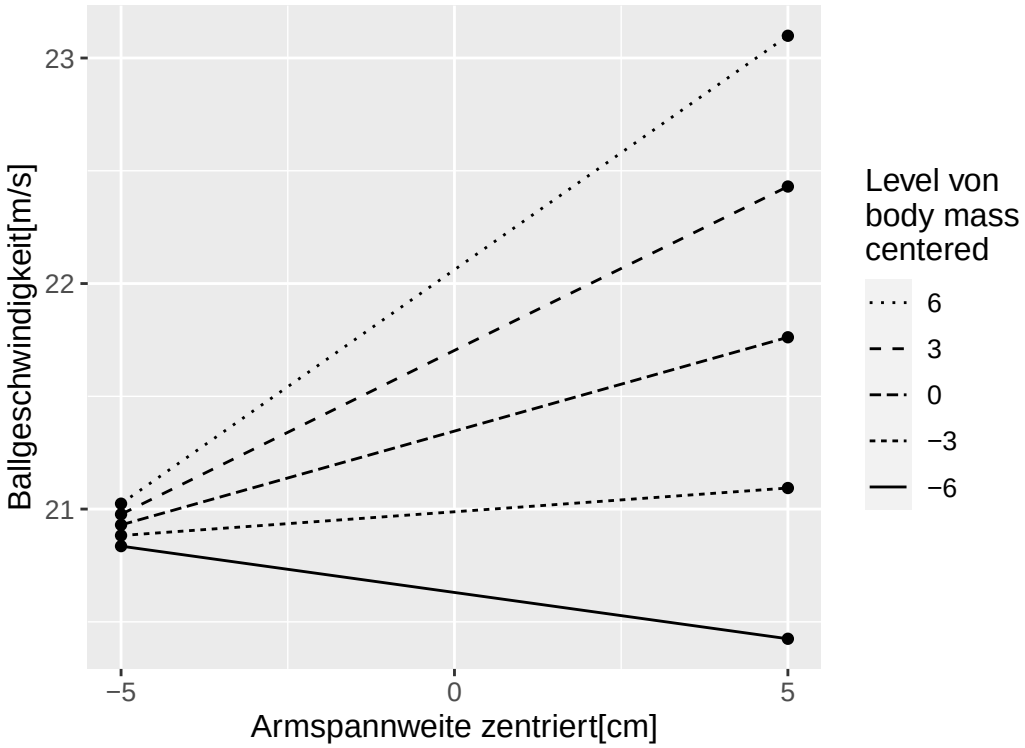


Figure 12.10: Veränderung mit dem Armspannweite

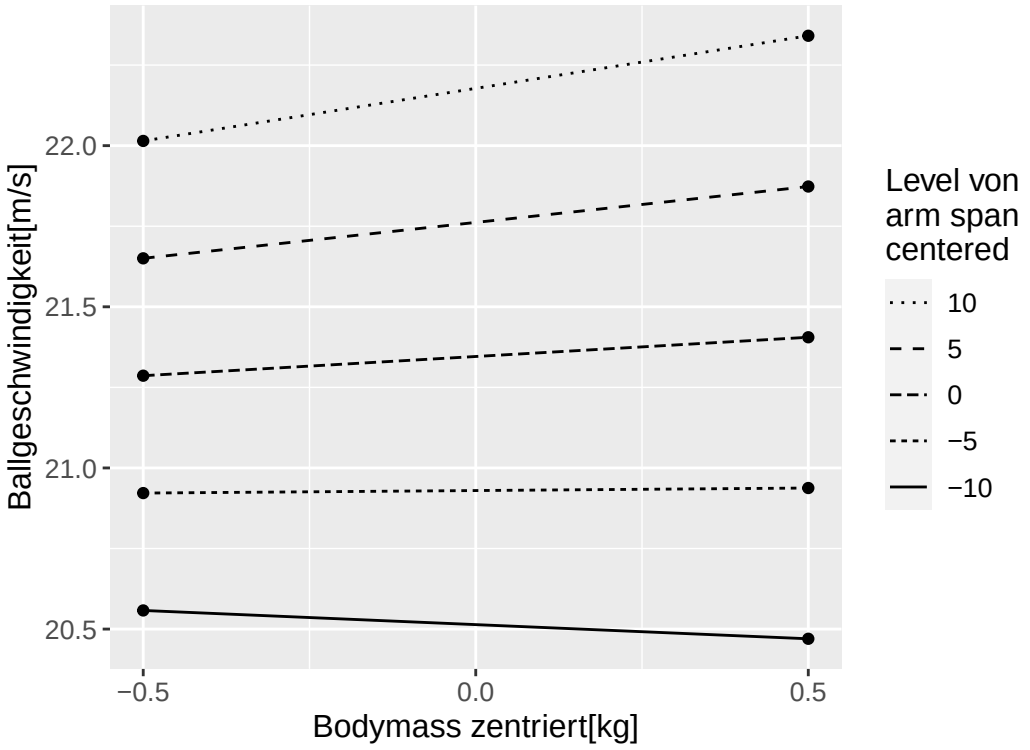


Figure 12.11: Veränderung mit dem Körpergewicht

Table 12.9: Modellkoeffizienten

	betas
b0	21.35
bm_c	0.12
as_c	0.08
bm_c:as_c	0.02

12.17 Interpretation der Koeffizienten

$$Y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_1 \cdot x_2 + \epsilon_i$$

- b_0: (y-Achsenabschnitt) der Wert von $\hat{Y}$ wenn $x_1 = 0$ und $x_2 = 0$ gilt.
- b_1: Der Unterschied in $\hat{Y}$ wenn zwei Objekte sich in x_1 um eine Einheit unterscheiden und $x_2 = 0$ ist.
- b_2: Der Unterschied in $\hat{Y}$ wenn zwei Objekte sich in x_2 um eine Einheit unterscheiden und $x_1 = 0$ ist.
- b_3: (Interaktionskoeffizient) Die Veränderung des Effekts von x_1 auf $\hat{Y}$ wenn x_2 um eine Einheit größer wird bzw. genau andersherum für x_2.

12.18 Aus der Ebene wird eine gekrümmte Fläche

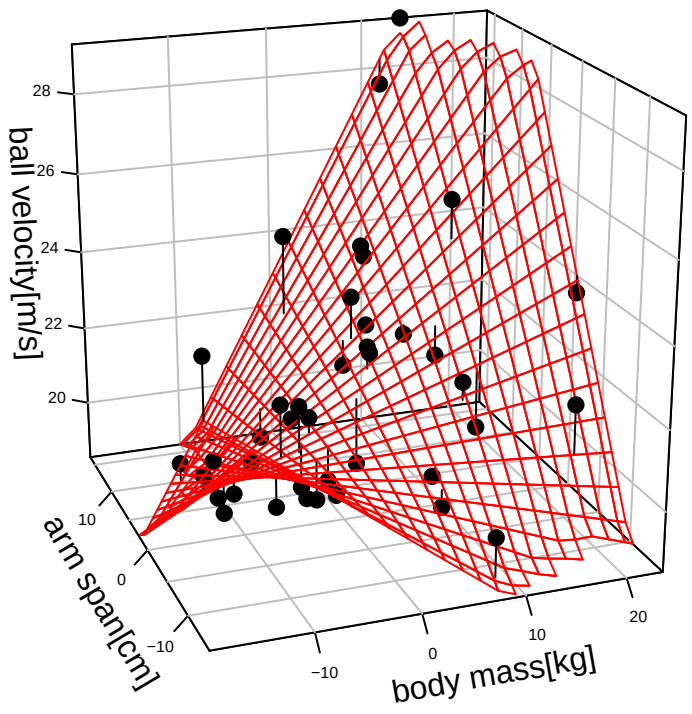


Figure 12.12: 3D Streudiagramm des Interaktionsmodells

12.19 Residuenvergleich

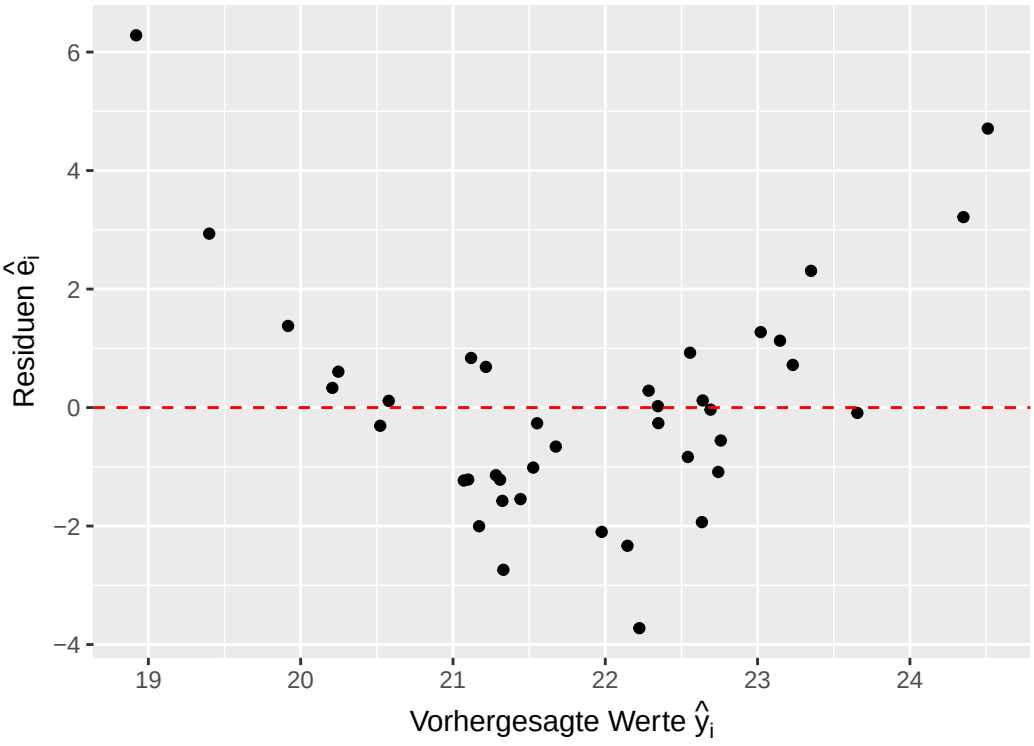


Figure 12.13: Residuen im additiven Modell

12.20 Residuenvergleich - qq-Plot

12.21 Take-away

Interaktionsmodell

- Erhöht die Flexibilität des linearen Modells.
- Bei Interaktionen hängt der Einfluss der einzelnen Variablen immer von den Werten der anderen Variablen ab.
- Achtung: Interpretation der einfachen Haupteffekte nicht mehr möglich bzw. sinnvoll!

12.22 Zuschlag

Was passiert im Interaktionsmodell mit den Koeffizienten wenn die x_{ki}s zentriert werden?

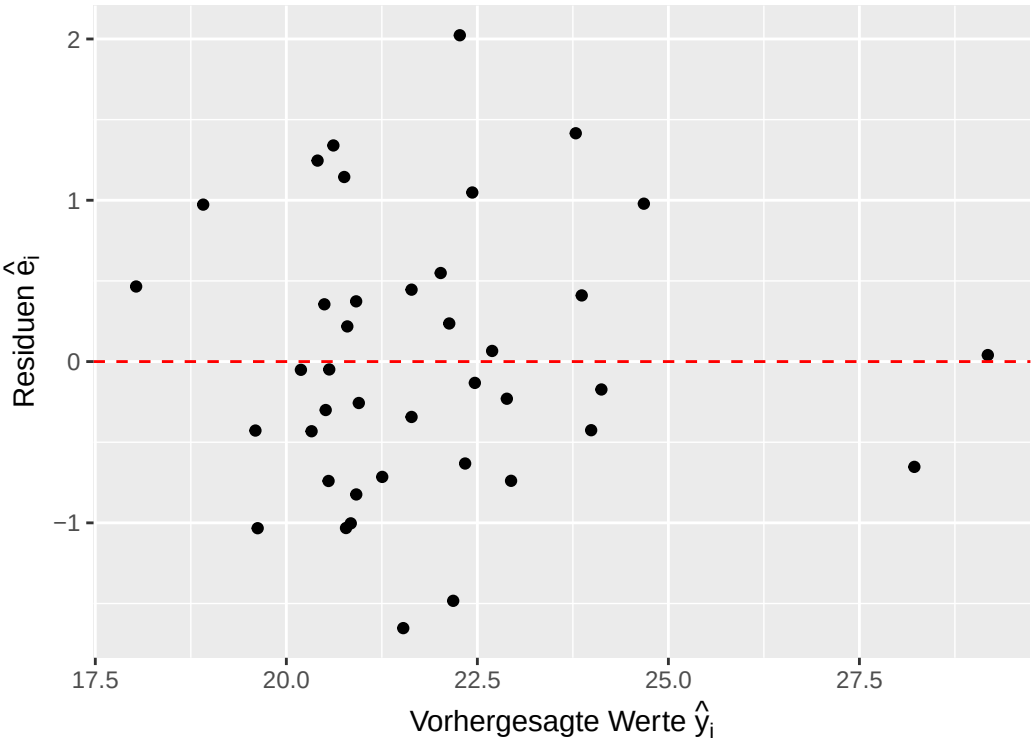


Figure 12.14: Residuen im Interaktionsmodell

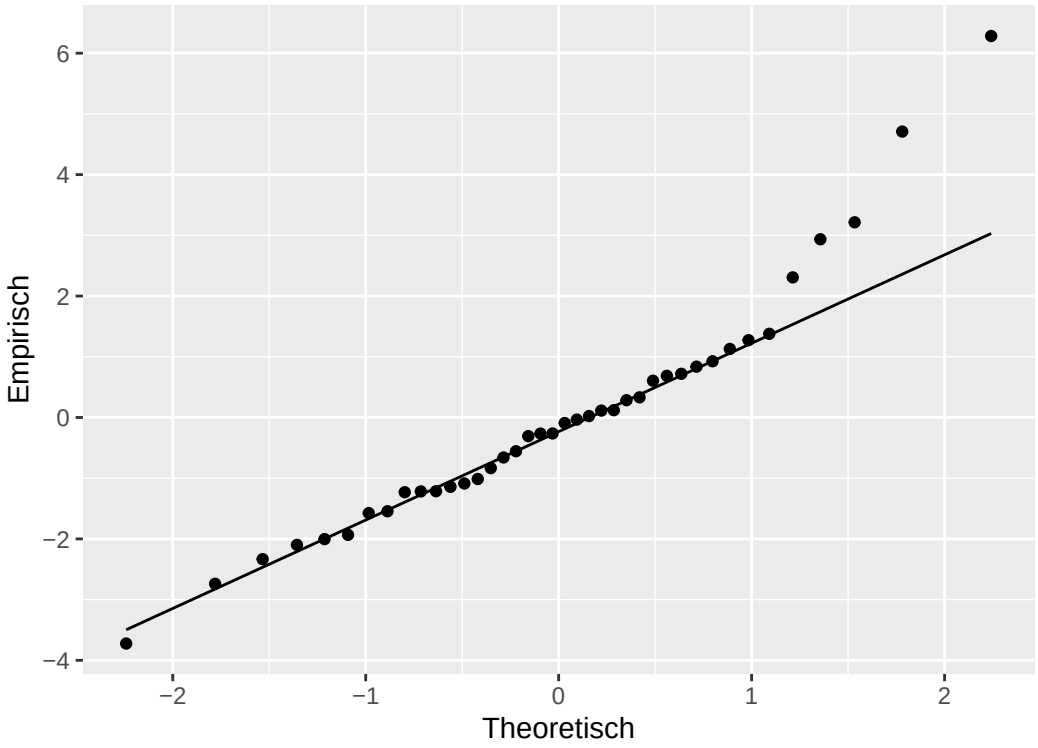


Figure 12.15: additives Modell

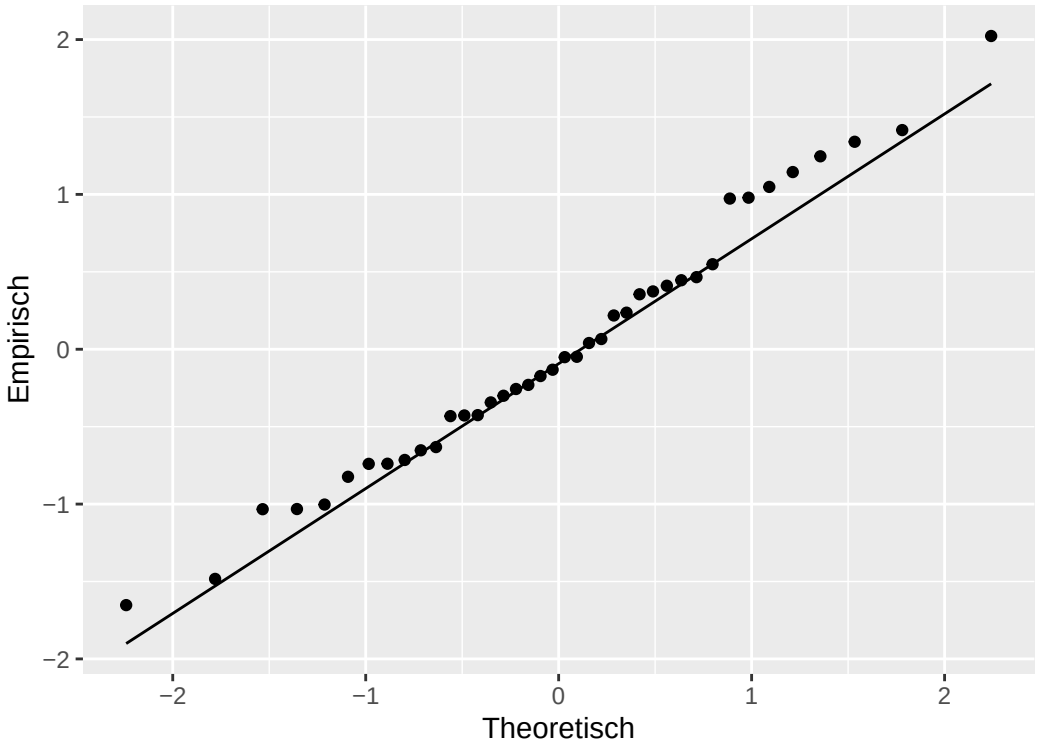


Figure 12.16: Interaktionsmodell

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \\
&= \beta_0 + \beta_1 x_{1i} - \beta_1\bar{x}_1 + \beta_2 x_{2i} - \beta_2\bar{x}_2 + \beta_3 x_{1i}x_{2i} - \beta_3 x_{1i}\bar{x}_2 - \beta_3\bar{x}_1 x_{2i} + \beta_3\bar{x}_1\bar{x}_2 \\
&= \beta_0 - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 + \beta_3\bar{x}_1\bar{x}_2 + \beta_1 x_{1i} - \beta_3\bar{x}_2 x_{1i} + \beta_2 x_{2i} - \beta_3\bar{x}_1 x_{2i} + \beta_3 x_{1i}x_{2i} \\
&= \underbrace{\beta_0 - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 + \beta_3\bar{x}_1\bar{x}_2}_{\beta_0} + \underbrace{(\beta_1 - \beta_3\bar{x}_2)x_{1i}}_{\beta_1 x_{1i}} + \underbrace{(\beta_2 - \beta_3\bar{x}_1)x_{2i}}_{\beta_2 x_{2i}} + \beta_3 x_{1i}x_{2i}
\end{aligned}
$$

12.23 Zum Nacharbeiten

Kutner et al. (2005, 306–13)

13 Integration von nominale Variablen

Bisher haben wir nur kontinuierliche, beziehungsweise metrische, Variablen in unsere lineares Modell aufgenommen. Im Folgenden werden wir sehen, dass wir mit einem kleinem Trick genause nominale Variablen in das Modell integrieren können, ohne dass wir fundamental etwas neues lernen müssen.

13.1 Vergleich von zwei Gruppen

Beginnen wir mit einem einfachen Beispiel. Wir wollen die Unterschiede zwischen Männern und Frauen in Bezug auf die Körpergröße untersuchen. In Figure 13.1 ist ein hypothetischer Datensatz von Körpergrößen von Frauen und Männern abgebildet. Wenig überraschen, da der Datensatz so erstellt wurde, sind Männer im Mittel größer als Frauen.

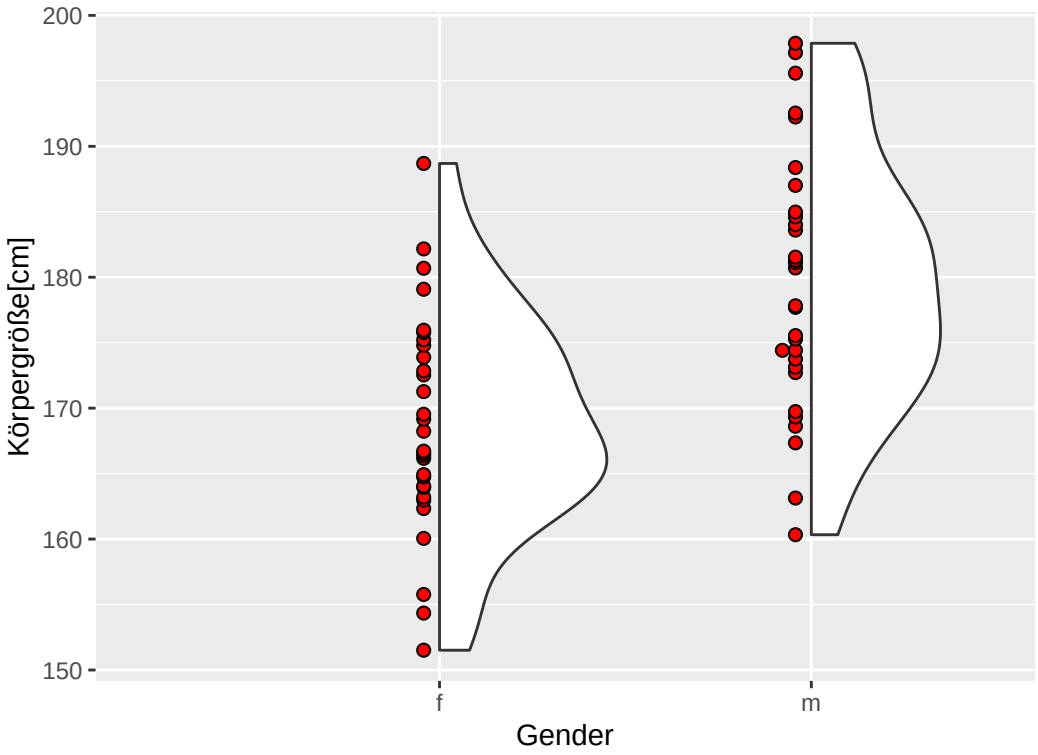


Figure 13.1: Simulierte Daten: Verteilung von Körpergrößen nach Geschlecht für Männer (m) und Frauen (f)

In Table 13.1 ist eine Ausschnit aus den Daten tabellarisch dargestellt. Wenig überraschend, haben wir zwei Datenspalten. In der ersten Spalte stehen die Körpergrößen, während in der zweiten Spalte eine Indikatorvariable steht die entweder den Wert m für Männer oder f für Frauen annimmt.

Table 13.1: Ausschnitt aus den Daten.

cm	gender
174.4	m
177.7	m
195.6	m
171.3	f
164.0	f
176.0	f

In Table 13.2 sind dann auch noch einmal die deskriptiven Statistiken der Körpergrößendaten abgebildet die auch noch einmal den Eindruck aus Figure 13.1 bestätigen.

Table 13.2: Deskriptive Statistiken der Körpergrößendaten.

gender	m	sd
f	168.8	8.4
m	179.5	9.8

Wir müssen jetzt allerdings erst einmal eine kurze Detour nehmen und verstehen, wie nominale Werte in `R` repräsentiert werden.

13.2 Nominale Variablen in `R` (detour)

Nominale Variablen werden in `R` mit einen speziellen Typ repräsentiert dem sogenannten `factor`. Erstellt werden kann ein Faktor mit der `factor()`-Funktion. Die Funktion hat drei wichtige Parameter. Der erste Parameter bezeichnet die Werte, der zweite die möglichen Faktorstufen (`levels`) und der dritte Parameter die dazugehörigen Bezeichnungen (`labels`). Ein einfaches Beispiel sieht dann so aus:

```
gender <- factor(c(0,0,1,1),
                 levels = c(0,1),
                 labels = c('m','f'))
gender
```

```
[1] m m f f
Levels: m f
```

D.h. wir haben einen Datenvektor mit den Elemente $(0, 0, 1, 1)$. Wir spezifizieren die `levels` dementsprechend mit 0 und 1 und definieren die dazugehörigen `labels` mit m und f. Dabei sind jeweils Vektoren übergeben worden (siehe `c()`). Wenn wir die neue Variable `gender` aufrufen erhalten wird der Datenvektor mit den entsprechenden `labels` ausgegeben und zusätzlich gibt `R` die möglichen `labels` an.

Wenn wir den Parameter `levels` nicht angegeben hätten, dann extrahiert `factor()` die eineindeutigen Werte selbst und führt die Abbidlung auf die `labels` entsprechend der Sortierung aus.

```
gender <- factor(c(0,0,1,1),
                 labels = c('m','f'))
gender
```

```
[1] m m f f
Levels: m f
```

```
str(gender)
```

```
 Factor w/ 2 levels "m","f": 1 1 2 2
```

Dabei muss darauf geachtet werden, dass die Abbildung auch tatsächlich diejenige ist, die gewünscht ist.

```
gender <- factor(c(0,0,1,1),
                 labels = c('f','m'))
gender
```

```
[1] f f m m
Levels: f m
```

Daher ist es fast immer sinnvoll `labels` und `levels` immer zusammen zu nehmen. Wenn die Parameter nicht angegeben werden, dann führt `factor` die Abbildung wiederum automatisch durch und für die `labels` werden die Datenwerte übernommen.

```
gender <- factor(c(0,0,1,1))
gender
```

```
[1] 0 0 1 1
Levels: 0 1
```

> **Warning**
>
> Achtung, die Variable `gender` sieht zwar aus wie ein numerischer Vektor, sie ist es aber nicht.
>
> ```
> is.numeric(gender)
> ```
>
> ```
> [1] FALSE
> ```
>
> Intern wird eine Faktorvariable von `R` zwar als ein numerischer Vektor abgelegt. Aber die "sichtbaren" Werte sind nun die Zeichenketten der `labels`, die daher auch angezeigt werden. Die interne numerische Repräsentation muss auch nicht mehr den ursprünglichen Datenwerten entsprechen.
>
> ```
> as.numeric(gender)
> ```
>
> ```
> [1] 1 1 2 2
> ```
>
> Die Datenwerte waren ursprünglich $(0, 1)$ und sind jetzt auf $(1, 2)$ abgebildet worden. Erinnert euch an die Eigenschaft von nominalen Variablen. Nominale Variablen sind einfach voneinander unterscheidbare Werte die jedoch in keiner Ordnung stehen.

Die automatische Konvertierung von `factor()` funktioniert am intuitivsten mit Zeichenkettenvektoren.

```
gender <- factor(c('m','f','m','f'))
gender
```

```
[1] m f m f
Levels: f m
```

```
str(gender)
```

```
 Factor w/ 2 levels "f","m": 2 1 2 1
```

`factor()` ermittelt zunächst die eineindeutigen Werte und sortiert diese dann entsprechend des Typen. In diesem Fall wird die Zeichenkette alphabetisch sortiert. Dann erfolgt die Abbildung der Werte auf die `labels`. Dies führt in diesem Fall dazu, dass die Werte `m` intern den Wert 2 zugeordnet bekommen, obwohl der erste Wert in den Daten `m` ist. Diese Sortierung der Daten wird später noch einmal von Bedeutung werden.

Die Abfolge der `levels` kann durch eine explizite Angabe der Reihenfolge selbst bestimmt werden.

```
gender <- factor(c('m','f','m','f'),
                 levels = c('m','f'))
```

```
  gender
```

```
[1] m f m f
Levels: m f
```

```
  str(gender)
```

```
 Factor w/ 2 levels "m","f": 1 2 1 2
```

```
  gender <- factor(c('m','f','m','f'),
                   levels = c('f','m'))
  gender
```

```
[1] m f m f
Levels: f m
```

```
  str(gender)
```

```
 Factor w/ 2 levels "f","m": 2 1 2 1
```

> **Tip**
>
> Im package `forcats` sind eine Reihe von Funktionen hinterlegt, mit denen die Eigenschaften von `factor`-Variablen einfach manipuliert werden können. Zum Beispiel, wenn die Reihenfolge von Faktorstufen geändert werden soll kann die Funktion `fct_relevel()` verwendet.
>
> ```
> [1] m f m f
> Levels: f m
>
> [1] m f m f
> Levels: m f
> ```
>
> Schaut euch die ausführliche Dokumentation der Funktionen und die Beispiel an, wenn ihr auf Probleme mit `factor`-Variablen stoßt.

> **Tip**
>
> Viele Funktionen in `R`, wie z.B. `lm()`, transformieren Vektoren mit Zeichenketten automatisch in einen `factor()` um. Wird in `lm()` in der Formel beispielsweise `y ~ gender` benutzt und

`gender` ist eine Datenspalte die aus den Zeichenketten `c('m','m','f','f')` besteht, dann ruft `lm()` intern die Funktion `factor()` für diese Daten auf und führt dann die Berechnung mit dem Faktor durch.
Dies erleichtert natürlich oft den Umgang mit den Daten, hat aber den Nachteil das immer klar sein muss, dass die automatische Konvertierung auch tatsächlich diejenige ist, dich auch gewünscht ist.

13.3 Vergleich von zwei Gruppen (continued)

Kommen wir zurück zum Körpergrößenvergleich. Normalerweise würden wir die Unterschiede zwischen den beiden Gruppen mit einem t-Test für unabhängige Stichproben untersuchen. In `R` können wird dies mit der `t.test()`-Funktion durchführen.

```
t.test(cm ~ gender, data=height, var.equal=T)
```

```
    Two Sample t-test

data:  cm by gender
t = -4.5683, df = 58, p-value = 2.617e-05
alternative hypothesis: true difference in means between group f and group m is not equal to 0
95 percent confidence interval:
 -15.45403  -6.03713
sample estimates:
mean in group f mean in group m
       168.7834        179.5290
```

Wenig überraschend finden wir ein statisch signifikantes Ergebnis.

13.4 Modellformulierung beim t-Test $(n_w = n_m)$

$$Y_{if} = \mu_f + \epsilon_{if}, \quad \epsilon_{if} \sim \mathcal{N}(0, \sigma^2)$$
$$Y_{im} = \mu_m + \epsilon_{im}, \quad \epsilon_{im} \sim \mathcal{N}(0, \sigma^2)$$

13.4.1 Hypothesen

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0$$

13.4.2 Teststatistik

$$t = \frac{\bar{y}_m - \bar{y}_w}{\sqrt{\frac{s_m^2 + s_w^2}{2}}\sqrt{\frac{2}{n}}}$$

13.4.3 Referenzverteilung

$$t \sim t_{df=2n-2}$$

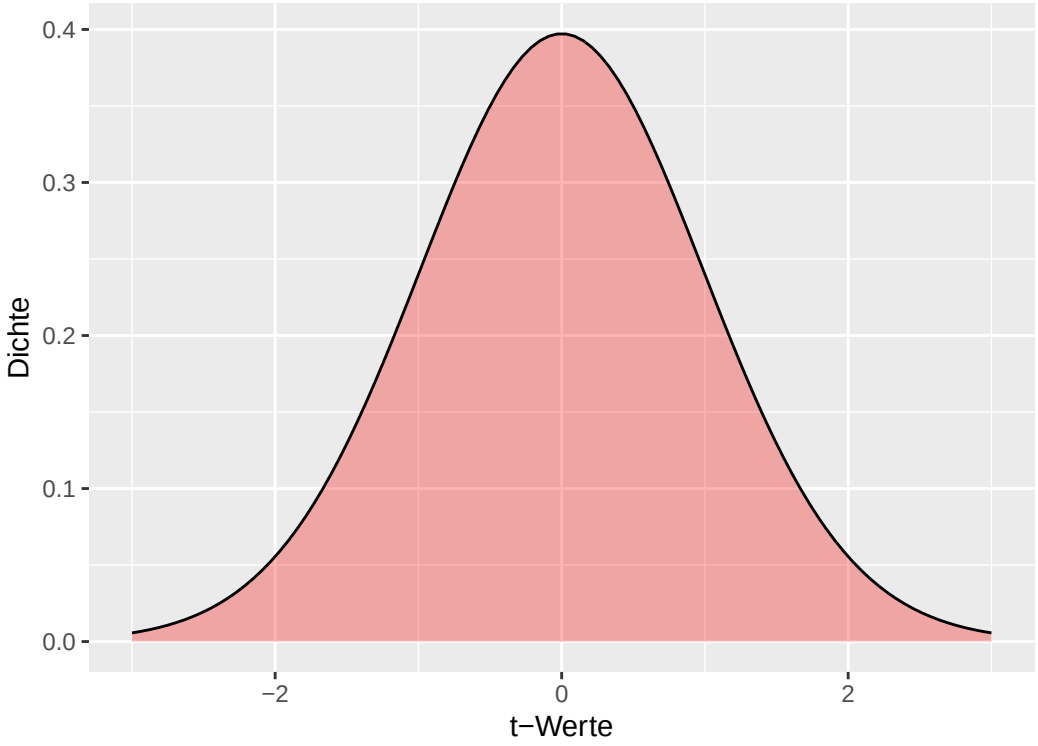


Figure 13.2: t-Verteilung mit $df = 58$

13.5 Kann ich aus dem t-Test ein lineares Modell machen?

13.5.1 t-Test

$$
\begin{aligned}
Y_{if} &= \mu_f + \epsilon_{if}, \quad \epsilon_{if} \sim \mathcal{N}(0, \sigma^2) \\
Y_{im} &= \mu_m + \epsilon_{im}, \quad \epsilon_{im} \sim \mathcal{N}(0, \sigma^2) \\
t &= \frac{\bar{y}_m - \bar{y}_w}{\sqrt{\frac{s_m^2 + s_w^2}{2}}\sqrt{\frac{2}{n}}} \\
t &\sim t_{df=2n-2}
\end{aligned}
$$

13.5.2 Lineares Modell

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 \times x_i + \epsilon_i \\
\Delta_m &= \mu_m - \mu_f \\
Y_i &= \beta_0 + \beta_1 \times x_{??} + \epsilon_i \\
Y_i &= \mu_f + \Delta_m \times x_{??} + \epsilon_i
\end{aligned}
$$

13.6 Dummy- oder Indikatorkodierung

$$
\begin{aligned}
Y_i &= \mu_f + \Delta_m \times x_{1i} + \epsilon_i \\
\Delta_m &= \mu_m - \mu_f \\
x_1 &= \begin{cases} 0 \text{ wenn weiblich} \\ 1 \text{ wenn männlich} \end{cases}
\end{aligned}
$$

Für eine nominale Variable wird eine Indikatorvariablen (Dummyvariable) definiert. Über diese Indikatorvariable kann die Zugehörigkeit eines Messwerts Y_i zu einer Faktorstufe k bestimmt werden. Eine Faktorstufe ist dabei immer die Referenzstufe bei der die Indikatorvariable gleich 0 ist.

13.7 Einfach mal stumpf in `lm()` eingeben

```
mod <- lm(cm ~ gender, height)
```

Table 13.3: Modellfit

	$\hat{\beta}$	s_e	t	p
(Intercept)	168.783	1.663	101.477	<0.001
genderm	10.746	2.352	4.568	<0.001

[1]

13.8 Vergleich der Konfidenzintervalle

13.8.1 Lineares Modell

```
confint(mod)
```

```
               2.5 %    97.5 %
(Intercept) 165.45401 172.11276
genderm       6.03713  15.45403
```

13.8.2 t-Test

```
t.test(cm ~ gender,
       data = height,
       var.equal=T)$conf
```

```
[1] -15.45403  -6.03713
attr(,"conf.level")
[1] 0.95
```

[2]

13.9 Auf welchen Werten wird ein lineares Modell gerechnet???

[1] `R` gibt die Faktorstufe nach dem Namen des Faktors an. Im Beispiel steht `genderm` für Stufe `m` im Faktor `gender`.

[2] Mit `t.test()$conf.int` kann auf das berechnete Konfidenzintervall zugegriffen werden.

Table 13.4: Repräsentation der Faktorvariablen

cm	gender	x_1
174.40	m	1
177.70	m	1
195.59	m	1
160.05	f	0
164.92	f	0
154.35	f	0

13.10 Residuen

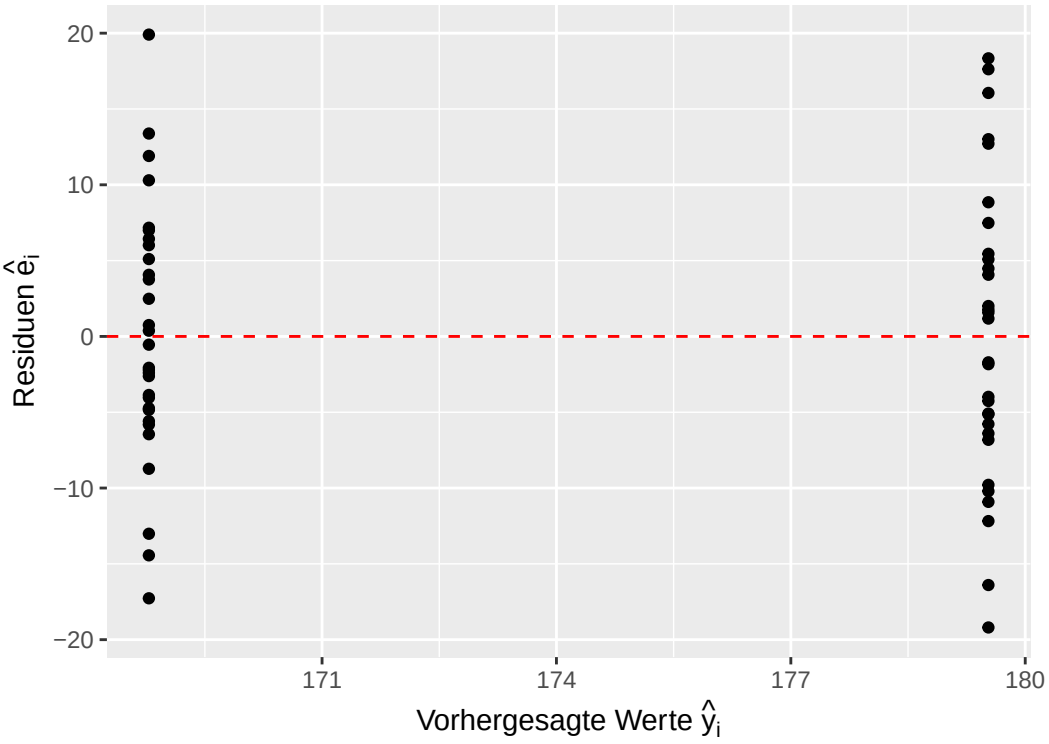


Figure 13.3: Residuen

13.11 Wen's interessiert - t-Wert

Seien beide Gruppen gleich groß (n) mit $N = n_m + n_w = 2 \times n$. Der t-Wert für β_1 berechnet sich aus $t = \frac{b_1}{s_b}$ mit:

$$s_b = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-2} \frac{1}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

Dadurch, das die x_i entweder gleich 0 oder 1 sind, ist $\bar{x} = 0.5$ und die Abweichungsquadrate im zweiten Term sind alle gleich $\frac{1}{4}$.

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N}\left(x_i - \frac{1}{2}\right)^2 = \sum_{i=1}^{N}\frac{1}{4} = \frac{N}{4} = \frac{2n}{4} = \frac{n}{2}$$

Der ersten Term kann mit etwas Algebra und der Definition für die Stichprobenvarianz s^2 auf die gewünschte Form gebracht werden.

$$\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N-2} = \frac{\sum_{i=1}^{n}(\overbrace{y_{im} - \bar{y}_m}^{Mnner})^2 + \sum_{i=1}^{n}(\overbrace{y_{iw} - \bar{y}_w}^{Frauen})^2}{2(n-1)} = \frac{(n-1)s_m^2 + (n-1)s_w^2}{2(n-1)} = \frac{s_m^2 + s_w^2}{2}$$

13.12 Wen's interessiert - $\beta_1 = \mu_w - \mu_m$

Mit $s_x^2 = \frac{N\frac{1}{4}}{N-1} = \frac{N}{4(N-1)}$

$$\begin{aligned}
b_1 &= \frac{cov(x,y)}{s_x^2} \\
&= \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{N-1}\frac{4(N-1)}{N} \\
&= 4\frac{\sum_{i=1}^{n}(y_{im} - \bar{y})\frac{-1}{2} + \sum(y_{iw} - \bar{y})\frac{1}{2}}{N} \\
&= \frac{4}{2}\frac{\sum_{i=1}^{n}(y_{iw} - \bar{y}) - \sum_{i=1}^{n}(y_{im} - \bar{y})}{2n} \\
&= \frac{\sum_{i=1}^{n} y_{iw}}{n} - \frac{n\bar{y}}{n} - \frac{\sum_{i=1}^{n} y_{im}}{n} + \frac{n\bar{y}}{n} \\
&= \bar{y}_w - \bar{y}_m = \Delta
\end{aligned}$$

13.13 Wen's interessiert - $\beta_0 = \mu_m$

Mit $b_1 = \Delta = \bar{y}_w - \bar{y}_m$:

$$\begin{aligned}
b_0 &= \bar{y} - \Delta \times \bar{x} \\
&= \frac{\sum_{i=1}^{N} y_i}{N} - \Delta \times \frac{1}{2} \\
&= \frac{\sum_{i=1}^{n} y_{im} + \sum_{i=1}^{n} y_{iw}}{2n} - \frac{1}{2}(\bar{y}_w - \bar{y}_m) \\
&= \frac{1}{2}\frac{\sum_{i=1}^{n} y_{im}}{n} + \frac{1}{2}\frac{\sum_{i=1}^{n} y_{iw}}{n} - \frac{1}{2}\bar{y}_w + \frac{1}{2}\bar{y}_m \\
&= \frac{1}{2}\bar{y}_m + \frac{1}{2}\bar{y}_w - \frac{1}{2}\bar{y}_w + \frac{1}{2}\bar{y}_m \\
&= \bar{y}_m
\end{aligned}$$

13.14 Können auch mehr als zwei Stufen verwendet werden?

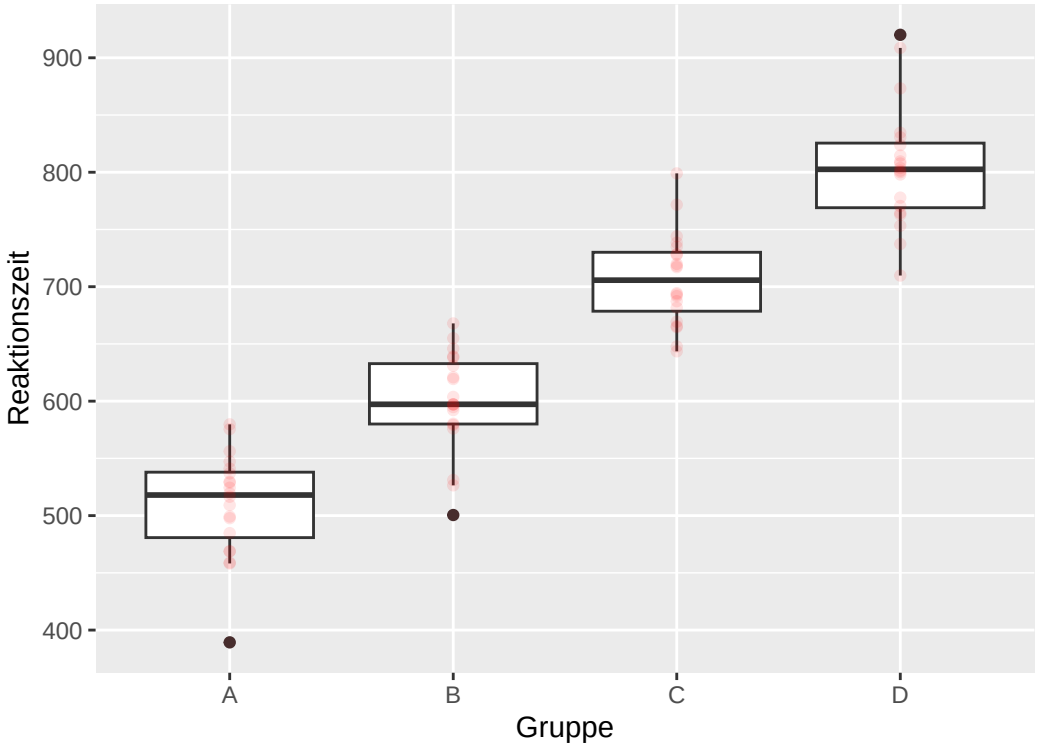


Figure 13.4: Ein Reaktionszeitexperiment mit vier Stufen A, B, C und D

Table 13.5: Gruppenmittelwerte, Standardabweichung und Unterschiede zu Stufe A

Gruppe	$\bar{y}_j$	s_j	Δ_{j-A}
A	509.53	45.66	
B	599.68	43.57	90.15
C	706.94	40.49	197.41
D	805.09	52.51	295.56

	x1	x2	x3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

13.15 Deskriptive Daten

13.16 Reaktionszeitexperiment als lineares Modell

13.16.1 Modell

$$y_i = \mu_A + \Delta_{B-A} x_1 + \Delta_{C-A} x_2 + \Delta_{D-A} x_3 + \epsilon_i$$

13.16.2 Dummyvariablen

13.17 Nochmal allgemeiner

Mit K Faktorstufen werden (K-1) Dummyvariablen $x_1, x_2, \ldots, x_{K-1}$ benötigt. Eine Stufe wird als Referenz definiert. Die x_1 bis x_{K-1} kodieren die Abweichungen der anderen Stufen von dieser Stufe.[3]

[3]Diese Art der Kodierung wird auch als treatment Kodierung bezeichnet.

	x_1	x_2	...	x_{K-1}
Referenz ($j = 1$)	0	0		0
$j = 2$	1	0	...	0
$j = 3$	0	1	...	0
$j = K$	0	0	...	1

Table 13.6: Modellfit

	$\hat{\beta}$	s_e	t	p
(Intercept)	509.526	10.235	49.784	<0.001
groupB	90.150	14.474	6.228	<0.001
groupC	197.414	14.474	13.639	<0.001
groupD	295.561	14.474	20.420	<0.001

Table 13.7: ANOVA-Tabelle

	Df	SSQ	MSQ	F	p
group	3	988935.1	329645.04	157.35	<0.001
Residuals	76	159221.0	2095.01		

13.18 Reaktionszeitexperiment mit `lm()`

```
mod <- lm(rt ~ group, data)
```

13.19 Ausblick

```
anova(mod)
```

13.20 Kombination von kontinuierlichen und nominalen Variablen

13.21 Modellansatz

- Aus gender (K = 2) wird eine **Dummyvariable**
- Frauen werden (zufällig) als Referenz genommen

$$Y_i = \beta_{ta=0,x_1=0} + \Delta_m \times x_1 + \beta_{ta} \times ta + \epsilon_i$$

$$x_1 = \begin{cases} 0 \text{ wenn weiblich} \\ 1 \text{ wenn männlich} \end{cases}$$

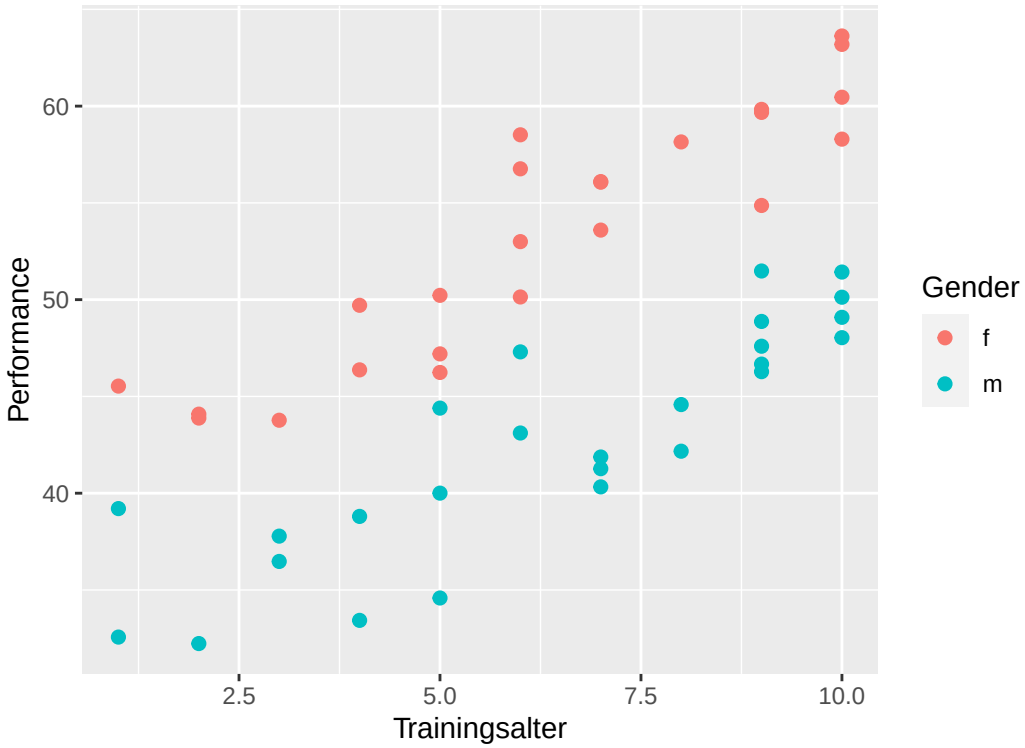


Figure 13.5: Hypothetische Leistungsentwicklung in Abhängigkeit vom Alter und Gender

Table 13.8: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	41.181	1.083
gender_fm	-10.877	0.805
ta	1.927	0.145
$\hat{\sigma}$	2.845	

13.22 Modellieren mit `lm()`

```
mod <- lm(perf ~ gender_f + ta, lew)
```

13.23 Die resultierenden Graden

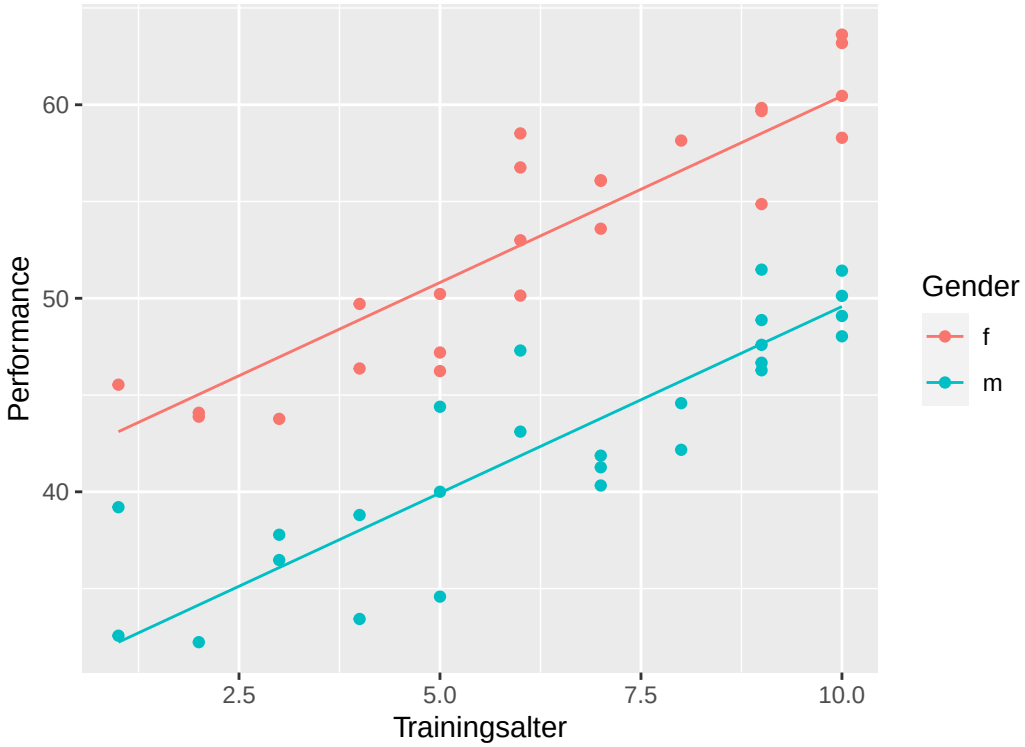


Figure 13.6: Leistungsentwicklung in Abhängigkeit vom Alter und Gender

13.24 Interaktion zwischen kontinuierlichen und nominalen Variablen

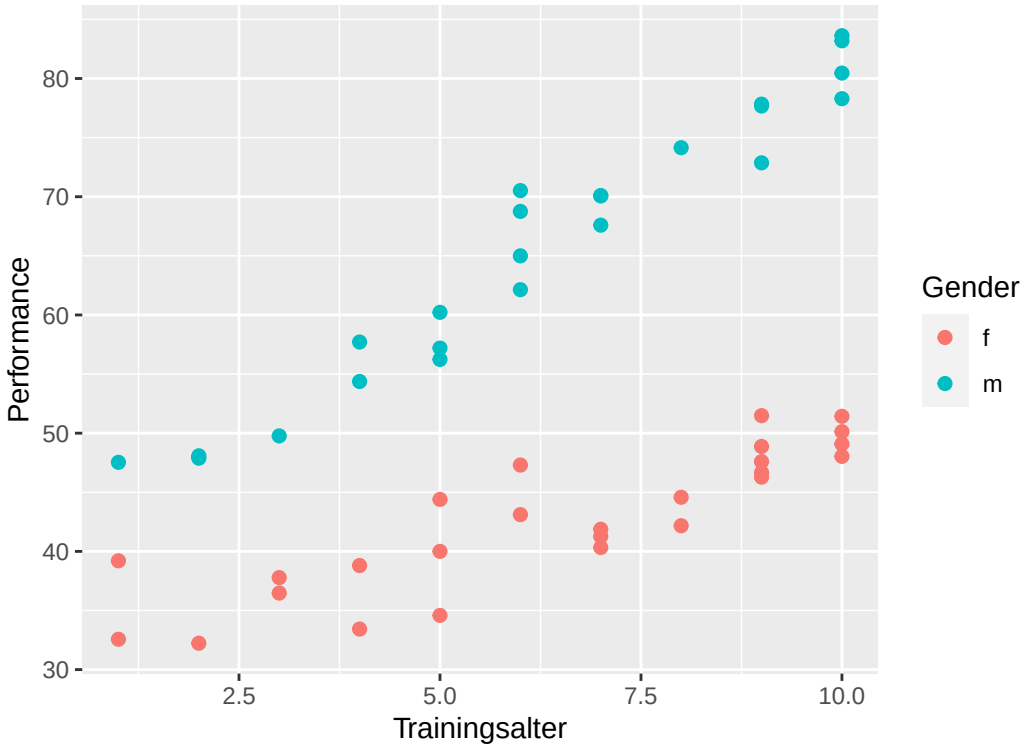


Figure 13.7: Leistungsentwicklung in Abhängigkeit vom Alter und Gender

13.25 Ansatz für ein Interaktionsmodell

Das vorhergehendes Modell wird um einen Interaktionsterm erweitert.

$$y_i = \beta_{ta=0,x_1=0} + \Delta_m \times x_1 + \beta_{ta} \times ta + \beta_{ta \times gender} \times x_1 \times ta + \epsilon_i$$

13.26 Interaktionsmodell mit `lm()`

```
mod <- lm(perf ~ gender_f * ta, lew)
```

Table 13.9: Modellfit

	$\hat{\beta}$	s_e
(Intercept)	31.354	1.370
gender_fm	8.575	2.010
ta	1.763	0.195
gender_fm:ta	2.362	0.290
$\hat{\sigma}$	2.828	

13.27 Regressionsgeraden

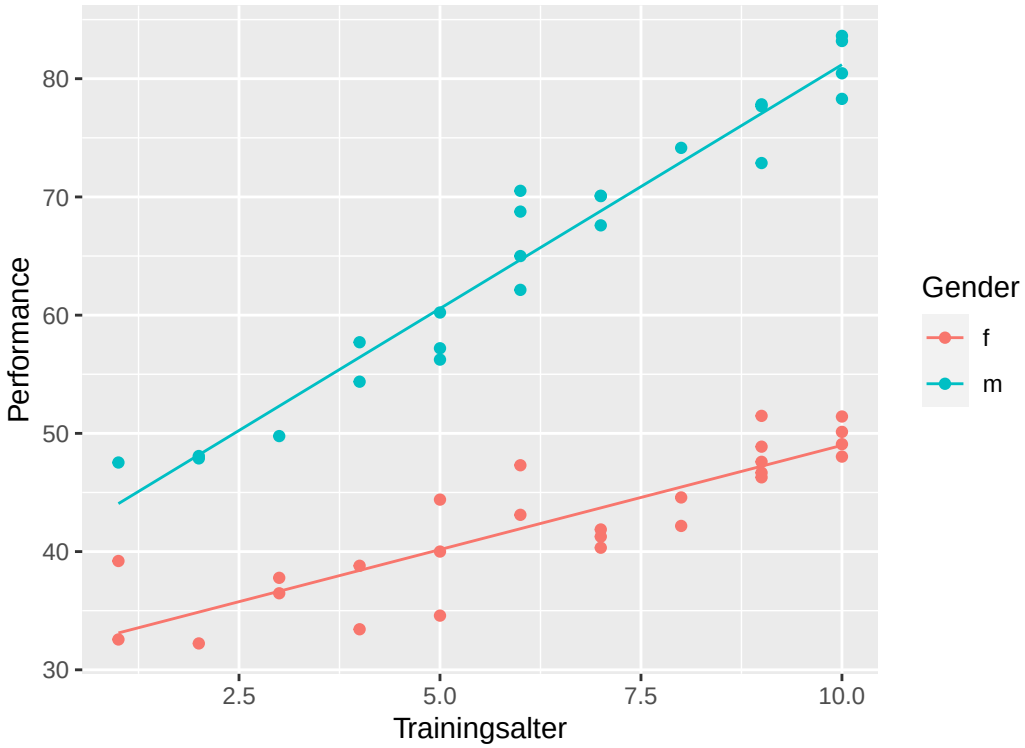


Figure 13.8: Leistungsentwicklung in Abhängigkeit vom Alter und Gender

13.28 Zum Nacharbeiten

Kutner et al. (2005, 313–19)

14 Modellhierarchien

Bisher waren wir damit beschäftigt unsere Modelle immer komplizierter zu machen, angefangen haben dem einfachen linearen Modell, das dann zum multiplen linearen Modell wurde mit mehreren x-Variablen. Die konnten im nächsten Schritt miteinander interagieren und dann im letzten Schritt wurde die Anforderung aufgehoben, das die x-Variablen kontinuierlich sein mussten. Letztendlich konnte aber immer das einfache Modell, die Punkt-Steigungsform, aus der Schule beibehalten werden. Nun werden wir noch eine direkte Verbindung zwischen dem Regressionsmodell und der Varianzanalyse aufzeigen.

14.1 Einfaches Modell

Wir beginnen mit einem einfachen Modell, das wiederum nur eine x und eine y-Variable hat.

```
mod0 <- lm(y ~ x, simple)
summary(mod0)
```

```
Call:
lm(formula = y ~ x, data = simple)

Residuals:
      1       2       3       4
-0.5817  0.9898 -0.2345 -0.1736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8414     0.7008   2.628    0.119
x             0.4574     0.3746   1.221    0.346

Residual standard error: 0.8376 on 2 degrees of freedom
Multiple R-squared:  0.4271,    Adjusted R-squared:  0.1406
F-statistic: 1.491 on 1 and 2 DF,  p-value: 0.3465
```

Hier ist jetzt zunächst einmal nichts Neues. Setzen wir uns aber noch einmal genauer mit den Abweichungen mit den Residuen auseinander. In der Besprechung des Determinationskoeffizienten R^2 haben wir schon die Unterteilung der Quadratsummen kennengelernt. Hier hatten wird auf die Aufteilung der Varianz von Y SSTO in die beiden Komponenten Regressionsvarianz SSR und Fehlervarianz SSE. Es gilt.

$$SSTO = SSR + SSE \tag{14.1}$$

Die Fehlerquadratsumme SSE ist definiert nach:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{14.2}$$

Als die Summe der quadrierten Abweichungen der vorhergesagten Werte $\hat{y}_i$ von den tatsächlich beobachteten Werten y_i.

Um die Werte $\hat{y}_i$ berechnen zu können, benötigen wir unser Modell. Dieses Modell hat zwei Parameter, die beiden Koeffizienten β_0 und β_1. Die Anzahl der Parameter in dem Modell wird meistens per Konvention mit dem Symbol p bezeichnet. In unserem Fall ist gilt daher $p = 2$. Die Anzahl der Parameter p ist verknüpft mit den sogenannten Freiheitsgraden df (degrees of freedom). Die Freiheitsgrade von SSE berechnen sich nach $N - p$, wobei N die Anzahl der Beobachtungen, der Datenpunkte ist. In unseren Fall also $N - 2$

$$df_E \coloneqq n - p \tag{14.3}$$

Die Frheitsgerade bestimmen die *effektive* Anzahl der Beobachtungen die zur Verfügung stehen um die Varianz σ^2 des Modells abzuschätzen. Dadurch, dass zwei Parameter anhand der Daten für das Modell bestimt werden, fallen zwei Datenpunkt raus als unabhängige Informationsquellen. Anders ausgedrückt, wenn ich die beiden Modellparameter, in unseren Fall β_0 und β_1 kenne, dann sind nur noch $N - 2$ Datenpunkt frei variierbar. Sobald ich die Werte von $N - 2$ Datenpunkten und die beiden Parameter kenne, kann ich die verbleibenden beiden Werte berechnen. Daher die Begriff der Freiheitsgrade.

Wir nun SSE durch die Anzahl der Freiheitsgerade teilen, dann lässt sich zeigen, das dieser Wert ein erwartungstreuer Schätzer für die Residualvarianz σ^2 in Bezug auf die Verteilungsannahme zu den $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ der Daten ist. Das Verhältnis von SSE zu df wird als Mean squared error MSE bezeichnet.

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p} = \hat{\sigma}^2 \tag{14.4}$$

Im ersten Moment erscheint diese Begründung etwas undurchsichtig, aber tatsächlich ist diese Formel schon eine alte Bekannte die uns in Form der Stichprobenvarianz begegnet ist.

Wenn wir eine Stichprobe der Größe N mit Werten y_i haben, dann haben wir gelernt die Varianz mittels der Formel:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{2}(y_i - \bar{y})^2 \tag{14.5}$$

zu berechnen. Was bei dieser Formel schon immer etwas quer gesessen hat, ist der Nenner mit $N - 1$ anstatt einfach N wie wir es vom Mittelwert kennen. Aber, um die Varianz zu berechnen benötigen wir einen Parameter, eben den Mittelwert. Dies führt wiederum dazu, dass nur $N - 1$ Werte frei variiert

werden können und wir sobald wir, neben dem Mittelwert $\bar{y}$, $N-1$ Werte kennen, den verbleibenden Wert berechnen können.

14.2 Genereller Linearer Modell Testansatz[1]

14.2.1 Idee

Wir bauen uns eine Teststatistik die die Verbesserung in der Vorhersage (= Reduktion der Fehlervarianz) als Metrik verwendet. Modelle werden in eine Hierarchie gesetzt mit einfacheren Modellen untergeordnet zu komplexeren Modellen.

14.2.2 Leitfrage:

Bringt mir die Aufnahme zusätzlicher Modellparameter eine Verbesserung in der Vorhersage von Y bzw. bezüglich der Aufklärung der Varianz in Y?

14.3 Genereller Linearer Modell Testansatz - Full model

Beispiel einfache lineare Regression

14.3.1 Volles Modell

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

14.3.2 Residualvarianz SSE(F)

$$\text{SSE(F)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

mit $p = 2, dfE(F) = n - 2$

14.4 Genereller Linearer Modell Testansatz - Reduced model

14.4.1 Reduziertes Modell

$$Y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

[1] Kutner et al. (2005), p.72

14.4.2 Residualvarianz SSE(R)

$$\text{SSE(R)} = \sum_{i=1}^{n}(y_i - \beta_0)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \text{SSTO}$$

mit $p = 1, dfE(R) = n - 1$

Im Allgemeinen gilt: $SSE(F) \leq SSE(R)$

14.5 Link: Reduziertes Modell und Stichprobenvarianz

$$\begin{aligned}
SSE &= \sum_{i=1}^{n}(y_i - \beta_0)^2 = \sum_{i=1}^{n}(y_i^2 - 2y_i\beta_0 + \beta_0^2) \\
0 &= \frac{\mathrm{d}}{\mathrm{d}\beta_0}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta_0 + \beta_0^2) \\
0 &= \sum_{i=1}^{n}(-2y_i + 2\beta_0) = -2\sum_{i=1}^{n}y_i + 2\sum_{i=1}^{n}\beta_0 \\
n\beta_0 &= \sum_{i=1}^{n}y_i \\
\beta_0 &= \frac{\sum_{i=1}^{n}y_i}{n} = \bar{y} \rightarrow \frac{SSE}{n-1} = \hat{\sigma}^2 = s^2
\end{aligned}$$

14.6 Genereller Linearer Modell Testansatz

Annahme: Das reduzierte Modell ist korrekt. Dann sollte

$$\text{SSE(R)} - \text{SSE(F)}$$

eher klein sein (Beide Modelle haben einen gleich guten fit).

Annahme: Das reduzierte Modell ist falsch: Dann sollte

$$\text{SSE(R)} - \text{SSE(F)}$$

eher groß sein (Das reduzierte Modell kann die Daten nicht so gut fitten wie das komplizierte Modell)

14.7 Genereller Linearer Modell Testansatz - Teststatistik

Wenn das reduzierte Modell korrekt ist, dann lässt sich zeigen, dass:

$$MS_{\text{test}} = \frac{\text{SSE(R)} - \text{SSE(F)}}{\text{dfE(R)} - \text{dfE(F)}}$$

ein Schätzer für die Varianz σ^2 ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$) ist.

Wenn das reduzierte Modell korrekt ist, dann ist auch das volle Modell korrekt. Daher ist dann:

$$\text{MSE(F)} = \frac{\text{SSE(F)}}{\text{dfE(F)}}$$

auch ein Schätzer für σ^2

14.8 F-Wert als Teststatistik

$$F = \frac{MS_{\text{test}}}{MSE(F)} = \frac{\frac{\text{SSE(R)} - \text{SSE(F)}}{\text{dfE(R)} - \text{dfE(F)}}}{\frac{\text{SSE(F)}}{\text{dfE(F)}}}$$

14.9 Verteilung der F-Statistik

$$F = \frac{MS_{\text{test}}}{MSE(F)} \sim F(\text{dfE(R)} - \text{dfE(F)}, \text{dfE(F)})$$

14.10 Hypothesentest mit F-Wert

[2]

14.11 Teilziel

- Durch den Vergleich von Modellen kann die Verbesserung/Verschlechterung der Modellvorhersage statistisch Überprüft werden
- Alternativ: Brauchen ich zusätzliche Parameter oder reicht mir das einfache Modell?

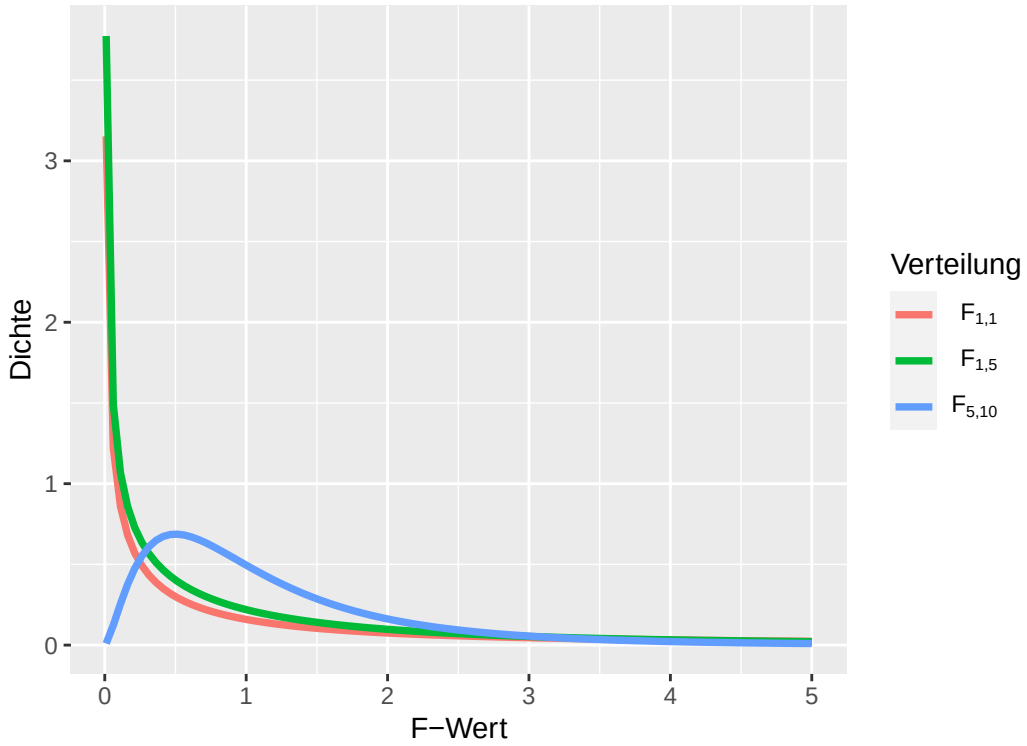


Figure 14.1: Beispiele für die F-Verteilung mit verschiedenen Freiheitsgraden df_1, df_2

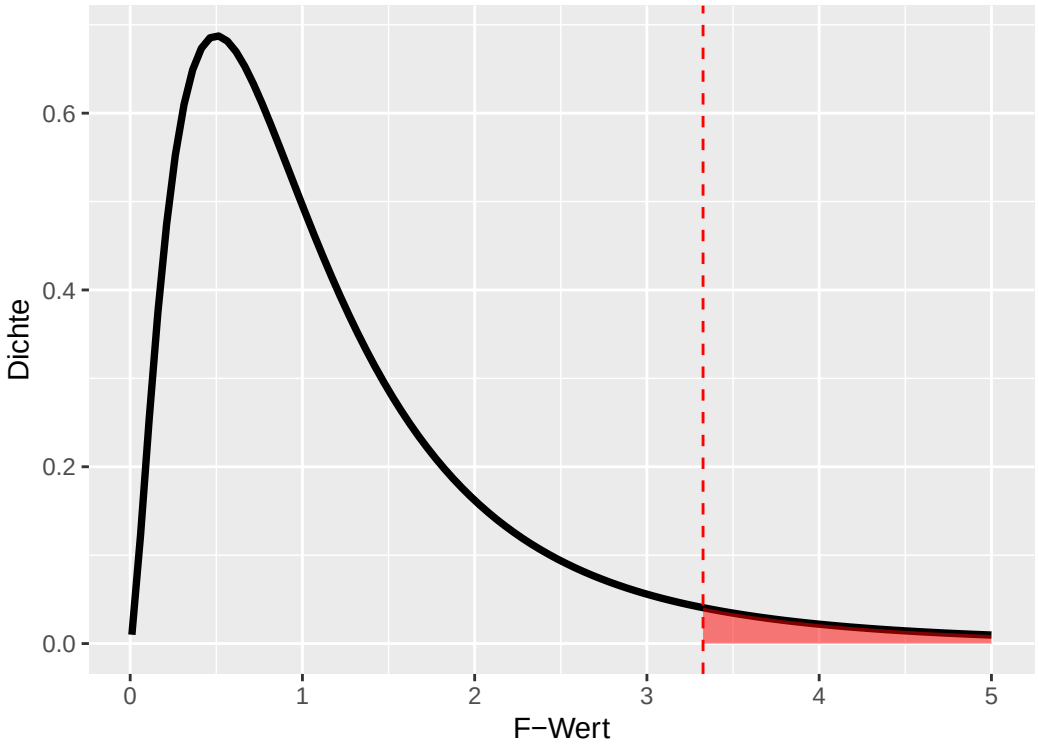


Figure 14.2: F-Verteilung mit $df_1 = 5, df_2 = 10$ und kritischem Wert bei $\alpha = 0.05$

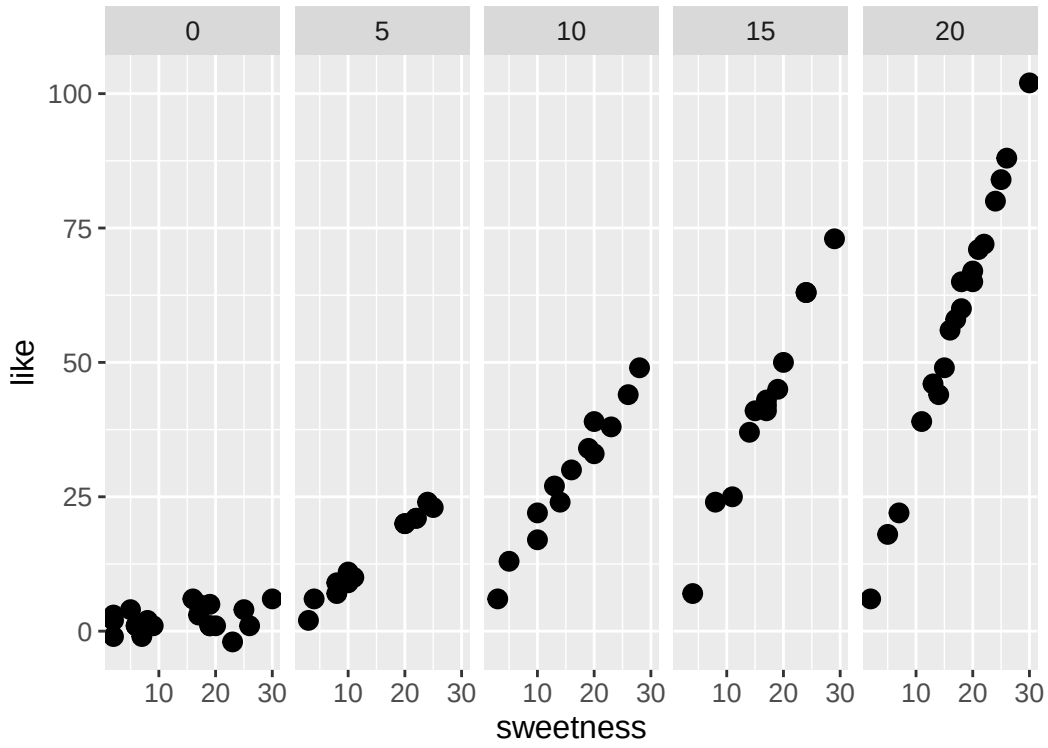


Figure 14.3: Zusammenhang zwischen der Präferenz für ein Bonbon und dem Süßgrad für verschiedene Weichheitsgrade

14.12 Beispiel: Candy-Problem

14.13 Modelle als Hierarchien auffassen

14.13.1 Full model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

14.13.2 Hierarchie

$$\begin{aligned}
m_0 &: y_i = \beta_0 + \epsilon_i \\
m_1 &: y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \\
m_2 &: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \\
m_3 &: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i
\end{aligned}$$

Es gilt: $m_0 \subseteq m_1 \subseteq m_2 \subseteq m_3$

14.14 Modelle als Hierarchien auffassen in R

In R:

```
mod_0 <- lm(like ~ 1, candy)
mod_1 <- lm(like ~ sweetness, candy)
mod_2 <- lm(like ~ sweetness + moisture, candy)
mod_3 <- lm(like ~ sweetness * moisture, candy)
```

14.15 Vergleich m_0 gegen m_1

$$\begin{aligned}
m_0 &: y_i = \beta_0 + \epsilon_i \\
m_1 &: y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i
\end{aligned}$$

```
anova(mod_0, mod_1)
```

[2] In R: `df()`, `pf()`, `qf()`, `rf()`

Table 14.1: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ 1	77				
Model 2: like ~ sweetness	76	1	16685.87	35.44	0

14.16 Vergleich m_1 gegen m_2

$$m_1 : y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$
$$m_2 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

```
anova(mod_1, mod_2)
```

Table 14.2: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ sweetness	76				
Model 2: like ~ sweetness + moisture	75	1	28168.9	277.59	0

14.17 Vergleich m_2 gegen full model m_3

$$m_2 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$
$$m_3 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

```
anova(mod_2, mod_3)
```

Table 14.3: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ sweetness + moisture	75				
Model 2: like ~ sweetness * moisture	74	1	7290.05	1682.27	0

14.18 Vergleich full model m_3 gegen minmales Modell m_0

$$m_0 : y_i = \beta_0 + \epsilon_i$$
$$m_3 : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

```
anova(mod_0, mod_3)
```

Table 14.4: Vergleich der Modellfits

Model	ResDF	DF	SS	F	p-val
Model 1: like ~ 1	77				
Model 2: like ~ sweetness * moisture	74	3	52144.81	4011.04	0

14.19 In `summary()` m_3 gegen m_0

```
Call:
lm(formula = like ~ sweetness * moisture, data = candy)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0332 -1.5255  0.0252  1.4298  4.3622

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.388063   0.765646   1.813   0.0739 .
sweetness           0.071526   0.046340   1.543   0.1270
moisture           -0.053223   0.070232  -0.758   0.4510
sweetness:moisture  0.163963   0.003998  41.016   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.082 on 74 degrees of freedom
Multiple R-squared:  0.9939,    Adjusted R-squared:  0.9936
F-statistic:  4011 on 3 and 74 DF,  p-value: < 2.2e-16
```

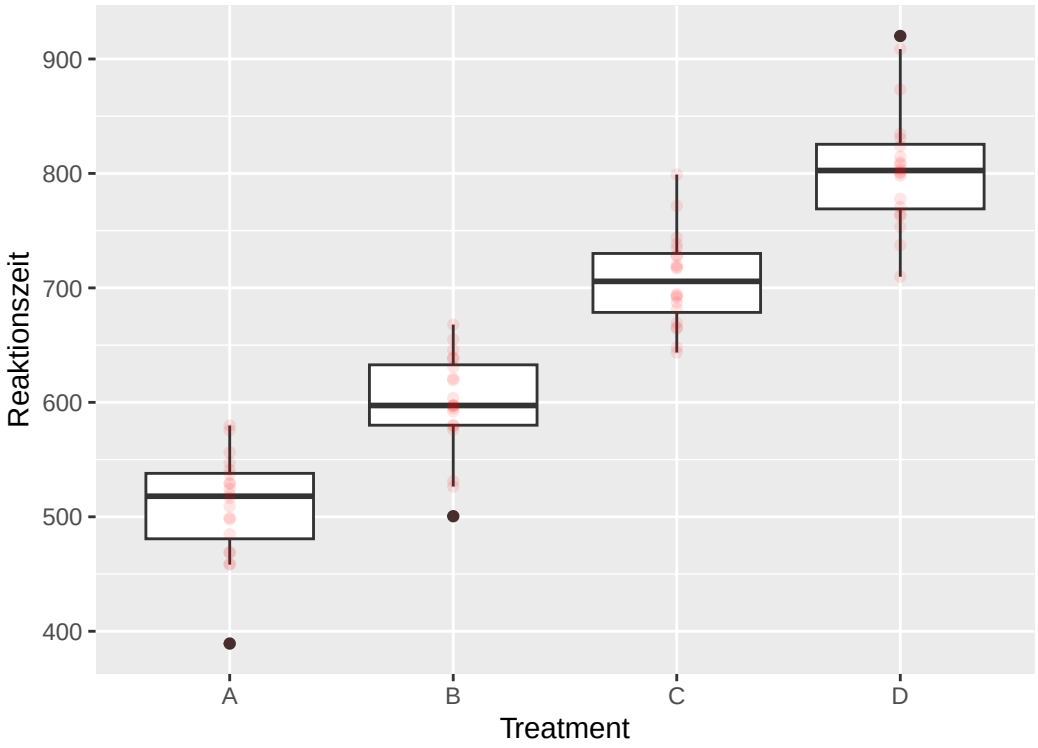


Figure 14.4: Ein Reaktionszeitexperiment mit vier Stufen A, B, C und D

14.20 Eine nominale Variable mit vier Stufen

14.21 Früher - Analysis of Variance (ANOVA bzw. AOV)

$$s^2_{zwischen} = \frac{1}{K-1}\sum_{j=1}^{K} N_j(\bar{x}_{j.} - \bar{x})^2$$

$$s^2_{innerhalb} = \frac{1}{N-K}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(x_{ji} - \bar{x}_{j.})^2 = \frac{1}{N-K}\sum_{j=1}^{K}(N_j - 1)s^2_j$$

$$F = \frac{\hat{\sigma}^2_{zwischen}}{\hat{\sigma}^2_{innerhalb}} \sim F(K-1, N-K)$$

14.22 ANOVA in R

```
mod_aov <- aov(rt ~ group, rt_tbl)
summary(mod_aov)
```

Table 14.5: Ausgabe mit aov()

term	df	sumsq	meansq	statistic	p.value
group	3	988935.1	329645	157.3	0
Residuals	76	159221.0	2095		

14.23 Ansatz mittels Modellhierarchien

14.23.1 Full model

$$y_i = \beta_0 + \beta_{\Delta_{B-A}} x_1 + \beta_{\Delta_{C-A}} x_2 + \beta_{\Delta_{D-A}} x_3 + \epsilon_i$$

14.23.2 Reduced model

$$y_i = \beta_0 + \epsilon_i$$

Wenn das reduced model die Daten gleich gut fittet wie das full model ⇒ Information über das Treatment verbessert meine Vorhersage von y_i nicht.

14.24 Model fit - Full model

```
mod <- lm(rt ~ group, rt_tbl)
```

Table 14.6: Modellfit

	$\hat{\beta}$	s_e	t	p
(Intercept)	509.526	10.235	49.784	<0.001
groupB	90.150	14.474	6.228	<0.001
groupC	197.414	14.474	13.639	<0.001
groupD	295.561	14.474	20.420	<0.001

14.25 `anova()` mit nur einem Modell

```
anova(mod)
```

Table 14.7: Äquivalent zum Vergleich full gegen reduced model

term	df	sumsq	meansq	statistic	p.value
group	3	988935.1	329645	157.3	0
Residuals	76	159221.0	2095		

14.26 Zum Nacharbeiten

Christensen (2018, 57–64)

Part IV

Das allgemeine lineare Modell

15 Synthese

Literatur

Altman, Douglas G, and J Martin Bland. 1995. "Statistics Notes: Absence of Evidence Is Not Evidence of Absence." *Bmj* 311 (7003): 485.

Altman, Naomi, and Martin Krzywinski. 2015a. "Points of Significance: Multiple Linear Regression." *Nature Methods* 12 (12): 1103–4.

———. 2015b. "Points of Significance: Simple Linear Regression." *Nature Methods* 12 (11).

———. 2016a. "Points of Significance: Analyzing Outliers: Influential or Nuisance." *Nature Methods* 13 (4): 281–82.

———. 2016b. "Points of Significance: Regression Diagnostics." *Nature Methods* 13 (5): 385–86.

Christensen, Ronald. 2018. *Analysis of Variance, Design, and Regression: Linear Modeling for Unbalanced Data*. CRC Press.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge.

Cumming, Geoff. 2013. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge.

Debanne, Thierry, and Guillaume Laffaye. 2011. "Predicting the Throwing Velocity of the Ball in Handball with Anthropometric Variables and Isotonic Tests." *Journal of Sports Sciences* 29 (7): 705–13.

Fox, John. 2011. *An r Companion to Applied Regression*. 2nd ed. SAGE Publication Inc., Thousand Oaks.

Gross, Benedict, Joe Harris, and Emily Riehl. 2019. *Fat Chance: Probability from 0 to 1*. Cambridge University Press.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models*. 5th ed. McGraw-Hill Irwin New York.

McElreath, Richard. 2016. *Statistical Rethinking, a Bayesian Course with Examples in r and Stan*. 1st ed. Boca Raton: CRC Press.

Spiegelhalter, David. 2019. *The Art of Statistics: Learning from Data*. Penguin UK.

Wasserstein, Ronald L, and Nicole A Lazar. 2016. "The ASA Statement on p-Values: Context, Process, and Purpose." Taylor & Francis.

Wild, Christopher J, and Georg AF Seber. 2000. *Chance Encounters: A First Course in Data Analysis and Inference*. Wiley Press.

Wilkinson, G. N., and C. E. Rogers. 1973. "Symbolic Description of Factorial Models for Analysis of Variance." *Applied Statistics* 22 (3): 392–99.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134 (2): 557–98.

Index