
S T A T I S T I K 1

Sebastian Sauer

Sebastian Sauer
CC - BY - NC - ND - 4.0
ISBN: 9798343798951
Independently published

Willkommen bei Statistik!

Dieses Buch führt Sie in die Grundlagen des Schwerpunkt auf Vorhersagen. Es ist ein Anfänger. Anders gesagt: Sie lernen, Daten an einfacher Modelle Vorhersagen abzuleiten.

Dieses Buch führt in die Statistik ein; es soll und hat nur ein Thema: Vorhersagen mittels Methoden. Alle Inhalte dieses Buches erklären die Vorhersage-Modellierung. Es wendet Vorkenntnisse in Statistik. Viele Statistiker der Welt, braucht es dann noch eines? Ja, es gibt aber (meines Wissens) in deutscher Sprache Lernen vermittelt, sich auf statistische zentriert und moderne Werkzeuge einsetzt. Buchschließen. Freude am Lernen, bei Angst das gehen? Viele Verständnisschwierigkeiten sind kompliziert geschrieben. Solche Überlegung zugrunde, dass die Konzepte erläutert sein müssten. Meiner Ansicht nach Weg verwechselt: Am Anfang darf eine Erklärung sein. Überblickend die Leser und Leserinnen können sie sich im nächsten Schritt mit dem was Präzision und Tiefe verlangt. Darüber hinaus eine lockere Sprache für einen entspannten Komfort beim Lesen wurde gesorgt: Lernziele, Übungen, Hinweise, Fehlerquellen, Tipps, Codes zu externen Medien und mehr werden in den Erklärbildern nicht gespart.

Vorwort

Der Inhalt des Buches ist ganz auf statistische ausgerichtet. „Statistische Modelle“ ist ein sagt nur, dass es darum geht, fachliche Fragen Bausteine zu gießen. Ein Beispiel: Studentin die Prüfung besteht, wenn Sie 42 Stunden büffeln, dass motivierte Studis ammeisten vom Lernen prägen. Carla ist hingegen überzeugt, dass Lernen nicht die Intelligenz allein für den Prüfungserfolg haben wird drei (noch recht unpräzise) wissenschaften Statistik hat nun die Aufgabe, möglichst präzise Forschungsfrage zu liefern; dafür sind Zahlentypen Bert und Carla ihre Überlegungen fachlich schätzen. Sprache übersetzen, können sie mit Antworten rechnen, manchmal sogar mit präzisen. Was nimmt die Antworten immer richtig oder nützlich sind. Vielleicht.

Mit Blick auf den Spagat zwischen Theorie und Buch (bzw. sein Autor) zugunsten der Seite der Lieber befähigen, praktische Probleme zulösen. Einblick zu vermitteln. Meine Hoffnung ist, dass es befügt, sich im nächsten Schritt tiefer mit dem Thema beschäftigen. Ist es nicht auch so im Alltag? Was Freude machen einstellen, dort vertiefen wir uns gerne weiter.

Da sich das Buch auf ein Thema, Modellierung, andere Themen außen vor, vor allem in der Statistik die eine oder der andere, von diesem Thema trennen. Ich denke, dass Modellierung für die Forschung zentraler Gedanke ist; für zwei große Themen im Buch zu eng. Wenn Sie Fragen oder Feedback haben, Hinweise dankbar. Stellen Sie sie gerne hier ein: sebastiansauer/statistik1/issues.

Die Online-Version dieses Buches ist frei verfügbar unter der BY-NC-ND-4.0-Lizenz publiziert.

Dieses Buch ist meinen Kindern Laurenz und Marlene sowie allen anderen Menschen, die noch viel lernen wollen.

Vorwort

Studierenden sind der Grund für Vieles, was
bin ich dankbar.

Ich wünsche Ihnen viel Freude und Erfolg bei
Ihr

Sebastian Sauer

/

.

ž

1.	1.	E s g e h t u m l h r e n L e r n e r f o l g
1.	2.	L e r n h i l f e n
1.	3.	S o f t w a r e
1.	4.	B e n ö t i g t e D a t e n

/

.

#

2.	1.	E i n s t i e g
2.	2.	W a s i s t S t a t i s t i k u n d w o z u i s t s i e g u t ?
2.	3.	W a s i s t d a s Z i e l I h r e r A n a l y s e ?
2.	4.	W a s s i n d D a t e n ?
2.	5.	A r t e n v o n V a r i a b l e n
2.	6.	M o d e l l e
2.	7.	P r a x i s b e z u g
2.	8.	W i e m a n n i t S t a t i s t i k l ü g t
2.	9.	F a z i t
2.	10.	A u f g a b e n
2.	11.	V e r t i e f u n g
2.	12.	L i t e r a t u r h i n w e i s e

.

3.	1.	E i n s t i e g
3.	2.	E r r r s t k o n t a c t
3.	3.	I n s t a l l a t i o n v o n R u n d R S t u d i o
3.	4.	R - P a k e t e

I n h a l t s v e r z e i c h n i s

3. 5.	M i t R a b e i t e n
3. 6.	E r r i s c h f ü r E i n s t e i g e r
3. 7.	M i t D a t e n a b e r e i t e n
3. 8.	L o g i k p r ü f u n g
3. 9.	P r a x i s b e z u g
3. 10.	A u f g a b e n
3. 11.	V e r t i e f u n g
3. 12.	L i t e r a t u r h i n w e i s e
.	.
4. 1.	E i n s t i e g
4. 2.	D i e V e r b e n d e s D a t e n j u d o s
4. 3.	D i e P f e i f e
4. 4.	B e i s p i e l e f ü r F o r s c h u n g s f r a g e n
4. 5.	P r a x i s b e z u g
4. 6.	W i e m a n n m i t S t a t i s t i k l ü g t
4. 7.	F a l l s t u d i e n
4. 8.	A u f g a b e n
4. 9.	V e r t i e f u n g
4. 10.	L i t e r a t u r h i n w e i s e
//	^
.	.
5. 1.	E i n s t i e g
5. 2.	E i n D i n o s a g t m e h r a l s 1 0 0 0 W o r t e
5. 3.	N o m e n k l a t u r v o n D a t e n d i a g r a m m e n
5. 4.	V e r t e i l u n g e n v e r b i l d l i c h e n
5. 5.	Z u s a m m e n h ä n g e v e r b i l d l i c h e n
5. 6.	U n t e r s c h i e d e v e r b i l d l i c h e n
5. 7.	S o l ü g t m a n n m i t S t a t i s t i k
5. 8.	P r a x i s b e z u g
5. 9.	V e r t i e f u n g
5. 10.	A u f g a b e n
5. 11.	L i t e r a t u r h i n w e i s e

Inhaltsverzeichni

!	
6. 1.	Einstieg
6. 2.	Mittelwert als Modell
6. 3.	Der Median als Modell
6. 4.	Quantile
6. 5.	Lagemaße
6. 6.	Wie man mit Statistik läugt
6. 7.	Vertiefung
6. 8.	Aufgaben
6. 9.	Literaturhinweise
!	
7. 1.	Einstieg
7. 2.	Woran erkennt man ein gutes Modell ?
7. 3.	Streuungsmaße
7. 4.	Streuung als Modellfehler
7. 5.	Die z-Transformation
7. 6.	Fazit
7. 7.	Aufgaben
7. 8.	Literaturhinweise
!	
8. 1.	Einstieg
8. 2.	Zusammenfassen zum Zusammenhang
8. 3.	Abweichungsrechtecke
8. 4.	Korrelation
8. 5.	Wie man mit Statistik läugt
8. 6.	Fallbeispiel
8. 7.	Aufgaben
8. 8.	Quiz zur deskriptiven Statistik
8. 9.	Literaturhinweise
///	
9. 1.	Einstieg
9. 2.	Vorhersagen

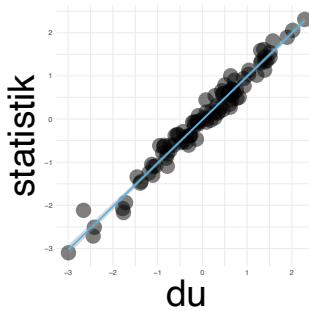
I n h a l t s v e r z e i c h n i s

9. 3.	G e r a d e n m o d e l l e
9. 4.	R - Q u a d r a t a l s M a ß d e r M o d e l l g ü t e
9. 5.	I n t e r p r e t a t i o n e i n e s R e g r e s s i o n s m o d e l l e s
9. 6.	W i e m a n m i t S t a t i s t i k l ü g t
9. 7.	F a l l b e i s p i e l M a r i o k a r t
9. 8.	F a l l s t u d i e I m m o b i l i e n p r e i s e
9. 9.	A u f g a b e n
9. 10.	L i t e r a t u r h i n w e i s e
10. 1.	E i n s t i e g
10. 2.	F o r s c h u n g s b e z u g : G l ä s e r n e K u n d e n
10. 3.	W e t t e r i n D e u t s c h l a n d
10. 4.	M o d e l l e m i t v i e l e n U V
10. 5.	F a l l b e i s p i e l z u r P r o g n o s e
10. 6.	V e r t i e f u n g : D a s A u f t e i l e n i h r e r D a t e n
10. 7.	P r a x i s b e z u g
10. 8.	W i e m a n m i t S t a t i s t i k l ü g t
10. 9.	F a z i t
10. 10.	A u f g a b e n
10. 11.	L i t e r a t u r h i n w e i s e

Ł

◦ f

Ž



A b b i l d u n g 1 . 1 . : S t a t i s t i k u n d D u : P a s

"

/

Ł

M e i s t e r Y o d a r ä t : L e s e n S i e d i e f o l g e n d e n H i n w

Ł

- Die Studenti s s i n d m i t w e s e n t l i c h e n M e t h o d D a t e n a n a l y s e v e r t r a u t u n d k ö n n e n d i e s e s e l
- Die Studenti s k ö n n e n g ä n g i g e F o r s c h u n g s f r d e l l e ü b e r s e t z e n , d i e s e a u f e c h t e D a t e n s ä t E r g e b n i s s e i n t e r p r e t i e r e n .

1. Organisatorisches

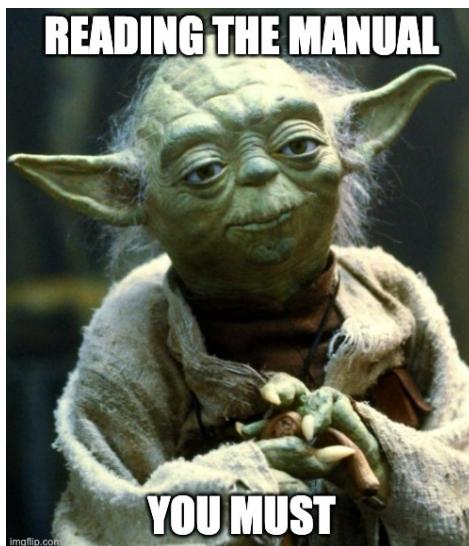


Abbildung 1.2.: Lesen die Inhalte du mus

(

Was lerne ich hier?

Sie lernen ~~den~~ das Werk der Datennahme ~~und~~ Schwerpunkt Vorhersage. Anders gestaltet auf ~~zu~~ ~~bestimmt~~ Daten ~~vorher~~ abgeleiteten. Zum Beispiel: Kommen Ihnen und sagt „Ich habe 42 Stunden für die Note kann ich in der Klausur erwarten?“. Daraus ist meine Daten und meines Modells müsstest. Außerdem lernen Sie, wie man die Güte einer Tätigkeit prüft. Denn Vorhersagen kann man ja beim Wahrsager bekommen. Wir wollen aber beide und wollen zumindest wissen, wie gut die Vorhersage ist.

Warum ist das wichtig?

Wir wollen nicht auf Leuten vertrauen, die bei uns gut ist. Wir wollen selber die Fakten.

Wozu brauche ich das im Job?

1. 1. Es geht um Ihre Lern

Datenanalyse spielt bereits heute in vielen Bereichen stark zunehmend.

Wozu brauche ich das im Studium?

In Forschungsarbeiten (wie ein empirischen Forschungsaufgabe oder Abschlussarbeit) ist es üblich, statistisch quantitativ zu analysieren.

Ist Statistik nicht sehr abstrakt?

Der Schwerpunkt dieses Kurses liegt auf Anwendung der Erkenntnisse eines Handwerks. Theorien und Abstraktionen nur am Rand.

Gibt es auch gute Jobs, wenn man sich mit Daten beschäftigt?

Das World Economic Forum (2020) berichtet zu den *increasing and decreasing demand across industries*:

1. Data Analysts und Scientists
2. AI and Machine Learning Specialists
3. Big Data Specialists

() „“

Das Lesen einer Schwimmfibel nutzt wenig, wenn Sie werden wollen. Es hilft nichts: Rein in die Flucht! Nicht tiefer ist und man jederzeit im Trockenen steht! Ihrem Fortschritt beim Lernen nichts im Vergleich ist nicht gerade subtil. Aber es ist (Lovett & Greenhouse, 2000). Dieses Buch bietet Gelegenheit. Nutzen Sie sie. Jedes Kapitel führt Aufgaben auf, alle mit Lösungen. So können Sie testen. Dass Schwierigkeiten auftreten, wenn Sie normal sind, ist normal. Das geht fast allen so. Ihren Lernerfolg gefährden: Wenn Sie aufgeben. Bleiben Sie dran und siche einstellen! Abbildung 116 S. 46 © engt iDsa, t@nevo

1. Organisatorisches

zeigen, dass regelmäßiges Üben und Dranbleiben (Sauer, 2017).

Dranbleiben ist der Schlüssel zum Erfolg. Üben Sie bei Schwierigkeiten nicht auf.

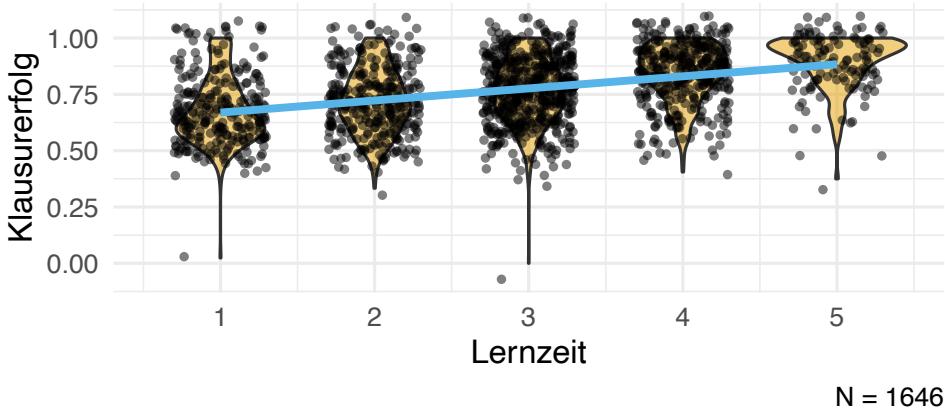


Abbildung 1: Zusammenhang von Lernzeit (1: geringe Lernzeit) mit Klausurerfolg

Motivation nötig? Dann schaue dich das Video an! Es handelt sich dabei um eine Ansprache zur Motivationsförderung.



Um von diesem Kurs am besten zu profitieren, mitbringen:

- Bereitschaft, Neues zulernen
- Bereitschaft, bei Schwierigkeiten nicht aufzugeben
- Kenntnis grundlegender Methodenwissen

Was Sie benötigen, sind besondere Mathe-Vorkenntnisse.

G

Abb. Abbildung 1. 4 gibt einen Überblick über den Inhalt des Buches. Das Diagramm hilft Ihnen, zu verstehen, wie im Gesamt zusammenhängt.

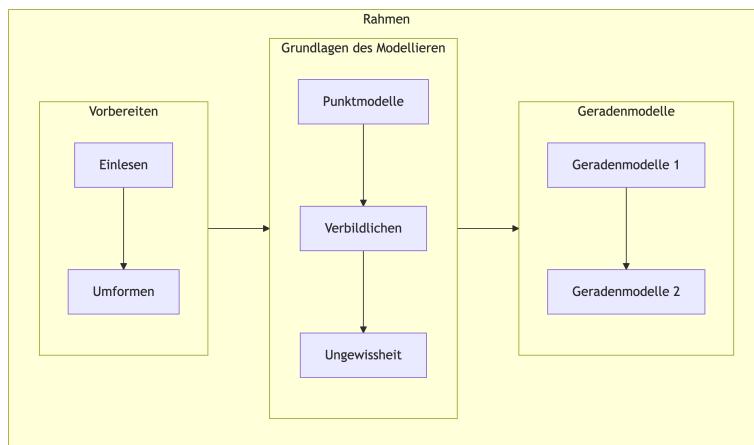


Abbildung 1. 4.: Überblick über den Inhalt

Das Diagramm zeigt auch den Ablauf einer typischen Lösung. Natürlich kann man sich auch andere sinnvolle Ablaufsvorstellungen machen.

E

Auf der Webseite „Datenwerk“ wird eine große Übersicht bereitgestellt.

Am Ende jedes Kapitels dieses Buches finden Sie eine Liste mit den gewählten Aufgabennamen, die Sie auf der Webseite „Datenwerk“ finden können. Beachten Sie die Hinweise zu den einzelnen Aufgaben.



<https://sebastiansauer.github.io/Datenwerk/>

1. Organisatorisches

Außerdem sind ~~an~~ ~~ke~~ ~~n~~ ~~auf~~ auf jedes Kapitel ~~se~~ Übungsschienenen Stellen auf, sodass Sie den jeweiligen Verständnis prüfen können. Im Buch finden Sie ~~Hervorhebung~~ Beispiele, Fehlerquellen Hinweise (und verlinkt), sodass Sie schnell suchen. Das Buch verweist auf ~~MaRöhrlivio~~ ist der gesamte R-Code für dieses Buch auf ~~Buches~~ zu finden: <https://github.com/sebastiansauer/statistik1/chapters>

Schauen Sie sich mal das ~~basti~~ ~~Statistik~~ ~~alstat~~ an! Dort die Playlist ~~alstat~~ ~~dort~~ ~~für~~ Videos zum Thema dieses Buches.



\$

#

Sie benötigen R, RStudio und einige R-Pakete. Das Buch enthält „mittel“ viel R. Auf fortgeschrittenen Anfänger gerichtet mag es dennoch als „viel Code“ erscheinen gewesen, auf R zu verzichten und statt dessen zu verwenden. JASP oder Jamovi sind Beispiele dieser Kategorie. Ich glaube aber, der Verzicht (R) wäre einschlechter Dienst an den Studierenden. „High-Tech-Zukunft“ sollte man zumindest nicht vertraut sein. Auf Computercode zu verzichten ist fahrlässig für die „Zukunftsfestigkeit“ des Lesers.

Sie finden den R-Code für jedes Kapitel im GitHub des Buches.²

²<https://github.com/sebastiansauer/statistik1/chapters>

1.4. Benötigte Dateien

!

In den meisten Kapiteln dieses Buches benötigen R-Pakete die ~~wenige~~ Systemdateien.

```
library(tidyverse)
library(rssy)
```

Weitere Hinweise zu R finden Sie in Kapitel 3.

.. .

In den meisten Kapiteln dieses Buches analysiert ist das der Datensatz Auktionen zu diesem Computer in einigen Merkmalen aufgeführt sind. Sie können folgende Art importieren, s. Listing 1.1.

Listing 1.1: Importieren von Mariokart-Datensätzen

```
mariookpaarst <- 0
" https://vincentarelbundock.github.io/Rdata/
" csv/openintro/mariokart.csv"

mariookrdata(mariookart_path)
```

Ein Data-Dictionary (Codebook) finden Sie in An-

% /

.

#

"

Abbildung 1. 4 zeigt den Standort dieses Kapitels mit einen Überblick über das Thema dieses Kapitels. Abbildung 2. 1 zeigt, dass unser Vorgehen gleich: Schritt für Schritt, in der Anfang bis Ende, erarbeiten wir unser „Datenprojekt“.

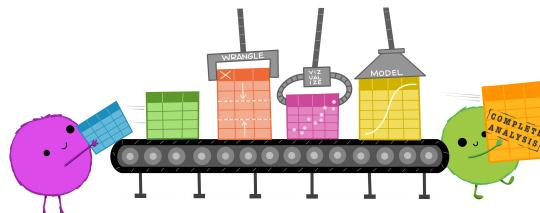


Abbildung 2. 1 zeigt die Abfolge am Fließbandanalyse als seine Analyse eine Abfolge am Fließbandanalyse.

Ł

- Sie können eine Definition von Statistik wie

2. Rahmen

- Sie können eine Definition von Daten wiederholen
 - Sie können den Begriff Tidy-Daten erläutern
 - Sie können Beispiele für verschiedene Strukturen nennen

Drei Faktoren beeinflussen Ihren Lernerfolg:
Mitarbeit im Unterricht und 3) Ihre Eigen-Nachbereitung des Unterrichts), s. Abbildung



Abbildung 2 Lernerfolg besteht aus drei Komponenten, Ihrer Mitarbeit im Unterrichtsstudium, d. h. Ihrer Vor- bzw. Nachbereitung.

Eine gute Lehrkraft ist wie der Funke, der zündet. Aber es braucht Brennmaterial, eine Konzentration im Unterricht. Ein Holzschein braucht mehr Brennmaterial, um das Feuer am Eigenstudium, das stetige, selbständige Be-

(\$)

In diesem Buch werden Sie sich mit dem Thema **Wasserwirtschaft** beschäftigen. Es ist ein Lehrbuch für Studierende der Wasserwirtschaft und Umweltwissenschaften.

Übungsaufgabe 2: von Strategie) einen Charakter (wie Chat GPT), was der Nutzen von Statistik Antwort.

2. 2. Was ist Statistik und wozu

Fragen wir doch einen Chatbot wie ChatGPT, was versteht:

💡 Was ist Statistik? Bitte definieren verstndlich.

🌐 Statistik ist eine Wissenschaft, die sich mit der Analyse, Interpretation, Präsentation und Daten beschäftigt. Ziel ist es, Muster und Daten zu erkennen und verlässliche Schlussfolgerungen zu ziehen. Eine größere Gruppe (z. B. Bevölkerung) auf einer Stichprobe (Teilmenge) zu treffen. Die Statistik ist der Beantwortung von Fragen, indem sie uns die Unsicherheit in den Daten zu messen und zu ver-

Es gibt mehrere Definitionen von Statistik. Hier ist eine:

Definition: Statistik fasst Werte zusammen, deren Unterschiedlichkeit beschreibt die Schlüsse (Kaplan, 2009; Poldrack, 2023).

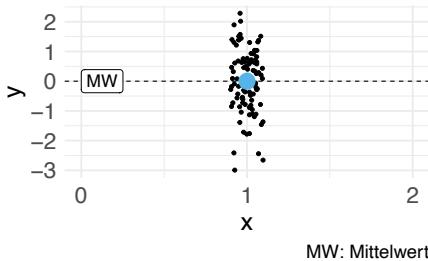
Betrachten wir die drei Bestimmungsstücke einer Statistik genauer: 1. Daten zusammenfassen, 2. Unter- zieren und 3. Ungewissheit beschreiben.

.

Abbildung 2.3 verdeutlicht das Prinzip des Zusammensetzung: Eine Menge von Zahlen wird als „zusammengedampft“. Eine einzelne Zahl ist verstanden als eine große Menge von Zahlen. Bei manchen Überblick verlieren.

¹Release 2023-Jan

2. Rahmen



(a) zusammengefasst zu einem Punkt
(b) zusammengefasst zu einer Linie

Abbildung 2 zeigt zusammenfassen. (a) Zusammenfassen einer Variable zu einem Punktwert, hier zusammenfassen zweier Variablen zu einem Punkt.

&

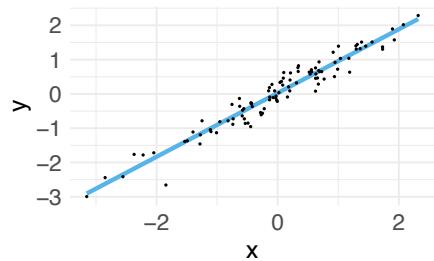
f

Eine allgemeinwährtige Tatsache ist, dass die Menschen, etwa, dass Exemplare einer Gattung sind nicht alle Menschen gleich groß, nicht nicht alle Tage gleich warm.

Ein zentrales Vorgehen bei statistischen Untersuchungen besteht darin, pro Variablen zu quantifizieren, wie das Beispiel

Bei den Basketballspielergruppen ist es in der Körpergröße sind groß, ähnlich groß. Bei den Schachspielern hingegen: Einige Personen sind groß, andere klein.

Ein Abweichungsmaß ist genannt, zeigt hier die Differenz zwischen dem Mittelwert und dem Wert der Körpergröße beobachtet. Nehmen wir an, wir sprechen allgemein von Merkmalen, die man als "Körpergröße" bezeichnet. Dann können wir kategorisieren: Personen mit einer Größe unterhalb des Mittelwerts sind klein, darüber groß.



2. 2. Was ist Statistik und wozu

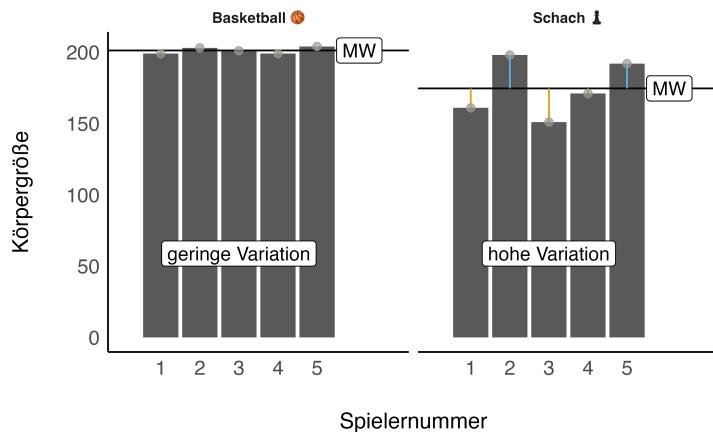


Abbildung 2. 4.: Wenig Variation in der KörpergröÙe lern. Alles lange Kerle. Viel Variation lern: Manches sind klein, andere groß.

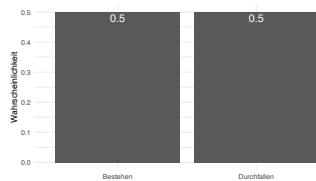
Definition (Residualen): Residuum des Merkmals Beobachtung ist definiert als die Differenz von Referenzwert, etwa dem Mittelwert (

&

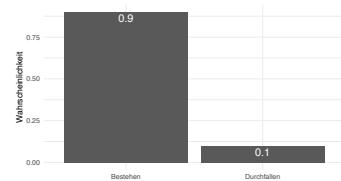
Beispiel: Eine Mathe-Statistik-Klausur geschafft eine Ahnung, ob sie bestehen wird. Beruhend auf der Annahme, dass sie bestanden hat. Die beiden Studentinnen sind also stark in der Ungewissheit hinsichtlich Klausurerfolg, s. Abbildung 2. 5.

Beispiel: Sie haben sich mit einem Zwillingsbruder auf ein Glücksspiel eingelassen: Er wirft einen gewinnt er, bei Zahl sie. Nun hat der Statistiker gewonnen und sieht sich sicher, dass dieser Typ Sie überschlagen hat. Allerdings sind Sie nicht ganz sicher, können Sie es leicht bestimmen?

2. Rahmen



(a) Was Anna denkt



(b) Was Bert a denkt

Abbildung 2.5 zeigt die gewissheit, die wir Ereignisse variieren

sicher weiß, dass er Sie über den Tisch gezogen seine Münze gezinkt ist.

(+ /

+

Statistische Analysen können drei Arten von
dung 2.6.



Abbildung 2.6.: Zielarten einer Dat

Beispiel 2.3.

- Beschreibe wie groß ist der Gender-Paygap in Zeitraum Y?
- Vorhersage nach 100 Stunden auf die Stati welche Note kann ich dann erwarten?

2. 3. Was ist das Ziel Ihrer A

- Erkläre mir viel bringt mir das Lernen auf die
□

Übungsaufgabe 2:nen Sie Beispiele für die die von Datenanalysen!

Eine Forschungsfrage ist die Leitfrage ihrer Arbeit. Sie herausfinden wollen. Häufig fragen Forschungskreise (kausalen) Einfluss auf Y?"

Eine Forschungsfrage weist häufig folgende Struktur auf:

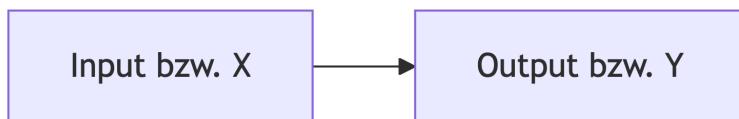


Abbildung 2.7.: Struktur einer Forschung

Beispiele für geschungene Fragen

Hat Lernen (X) einen Einfluss auf den Prüfungserfolg?

Verringert Joggen (X) die Menge des Hüftgelenks

Umwelten Betrag erhöht sich der Umsatz (Y) um 1000 € mehr für Werbung ausgeben? (X)

Verringert intensive Handynutzung die Konzerte?

Beispiel(e) frage: Produktmerkmale und Nachdem Studium haben Sie bei einem großen On angeheuert. Da Sie angeben, es ist sicherheitswichtig, ist

2. Rahmen

beschäftigt zu haben, hat man Sie in die Abteilung Entwicklung (F&E) gesteckt. Heute ist es eine Spielekonsole wie zu analysieren, genauer Mariokart. Ihre Forschungsfrage lautet:

Welche Produktmerkmale stehen mit einem Erlös in Zusammenhang?

\$

Ward et al. (2017) untersuchten die Forschungen in der Gegenwart eines Handys (z.B. wenn es vor Ihnen dazuführt, dass man abgelenkt wird und daher Leistungen zeigt).

Die Autoren formulieren ihre Hypothese ebenfalls implizit aus dem Text herauslesen.

First, smartphones may redirect the orientation of attention away from the focal task and to other behaviors associated with one's phone. Perhaps ample evidence that ... this digital distraction affects both performance ... and enjoyment.

Später präzisieren sie ihre Hypothese (S. 111):

In two experiments, we test the hypothesis that the presence of one's own smartphone reduces cognitive capacity.

Die Ergebnisse unterstützen Ihre Hypothese: Die kognitive Leistung (Y-Achse) ist sowohl im Gedächtnisses als auch in der fluiden Intelligenz auf dem Schreibtisch liegt, als wenn es nicht das Handy ist die kognitive Leistung, wenn das Handy

2.3. Was ist das Ziel Ihrer A

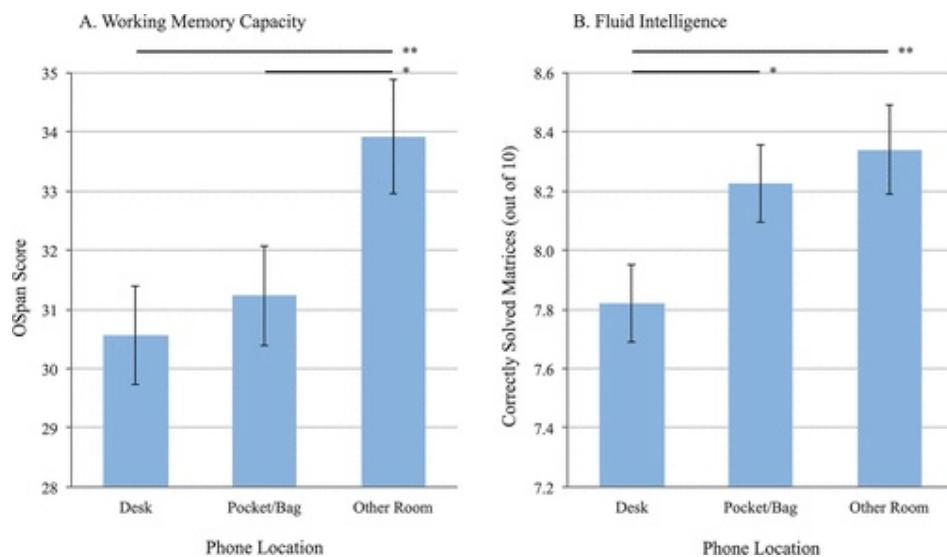


Abbildung 2.3.1 zeigt, dass die kognitiven Leistungen im Vergleich zu den anderen beiden Orten am höchsten sind. (Ward et al., 2017, S. 145)

Übungsaufgaben sind in Sichtweite verringert die kognitiven Leistungen, Ward et al. (2017), S. 145

Datenanalyse ist eine Art des Problemlösens. Aber es nicht zum Spaß (jedenfalls nicht alle von uns) zu erreichen, also ein Problem zu lösen. Daher gleich zu Anfang will drauf los. Zunächst 1) und das Ziel. Dann 2) plant man das Vorgehen, z. erheben möchte. Als nächstes 3) erhebt man die sie auf. Schließlich kann man sie 4) endlich ansprechen nicht für sich, man muss sie 5) interagieren ausziehen. Dazu gehört auch, dass man die Sozialanalyse kritisch beleuchtet, vgl. Abbildung 2.3.1. Auch das PPDAC-Modell (MacKay & Oldford, 2000)

2. Rahmen

- PP.rob (Problem und Ziel und Sachgegenstand)
- PP.la (Vorgehen planen)
- DDat(aDaten erheben und aufbereiten)
- AAanal y (Daten analysieren)
- CConcl u ss (Schlussfolgerungen ziehen)

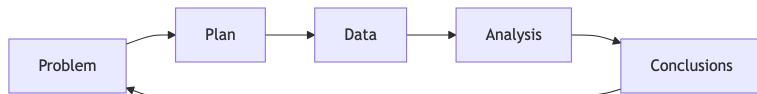


Abbildung 2.9.: Datenanalyse als Prozess

(

Definiti o Da2t. Da2t en sind eine geordnete Folge

Tabellen sind oft das geeignete Format für Daten. Tabelle 2.1 zeigt ein Beispiel für nur eine laufende Nummer. Sie dient dazu, Personen (hier Studenten) zu identifizieren und keine Information. Beispiele für ID-Variablen: Personalausweisnummern oder Bestellnummern.

Tabelle 2.1.: So sehen Daten in Form einer Tabelle aus:

id	name	note
1	Anna	1.3
2	Berta	2.3
3	Carla	3.0

Beispiel Da2t. e6n zur Forschungsfrage 2 Auszug
Daten zur Mabitgliederung Tabelle 2.2.

2.4. Was sind Daten?

Tabelle 2.2.: Auszug aus der Tabelle mar

n_bids	start_pr	total_pr	wheels
20	0.99	52	1
13	0.99	37	1
16	0.99	46	1
18	0.99	44	1
20	0.01	71	2
19	0.99	45	0

Eine Erklärung (Data-Dictionary) aller Variablen findet sich hier.

Definition (a4-Dictionary): was die Variablen (Spalten) einer Datentabelle beschreiben, nennt man Dictionary.

In den Spalten einer Tabelle stehen Merkmale (Variablen), die untersucht werden, z.B. Patienten, Kinder. Die untersuchten Dinge nennt man Beobachtungsmerkmale. Die untersuchten Einheiten stehen in den Zeilen einer Tabelle.

Definition (a5): Variable ist ein Platzhalter für einen Wert, das verschiedene Ausprägungen annehmen kann.

Man kann sich eine Variable wie einen Behälter mit einem Stift geschrieben vorstellen, was für eine Abbildung 2.10.

Definition (a6): Beobachtungseinheiten sind Dinge, die wir untersuchen (beobachten). Beobachtungen sind Träger von Variablen.

²<https://www.openintro.org/data/index.php?data=mar>

2. Rahmen

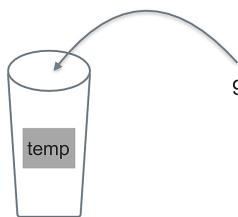


Abbildung 2.1 die Variable „temp“

Tabelle 2.1 enthält Namen von Variablen (links) und die Beobachtungswerte (rechts). Die Beobachtungswerte werden aus den Beobachtungslisten berechnet.

Definition (Wert): Der Inhalt einer Variablen.

In Abbildung 2.10 ist die Variable 2.1 nimmt die Werte Anna, Berta und Carla an.

Definition (Ausprägung): Prägegebenen man die verschiedenen Werte einer Variablen.

Beispiel: In einer Studie wurden zehn Probanden nach Geschlecht und Alter untersucht.

```
geschlecht
"Fr, ä, Män, n, Ma, n, nd, i, v, e, Fr, s, d, u"
geschlecht
# [1] "Mann"   "Frau"    "Frau"    "Frau"    "Mann"    "
# [9] "divers"  "Frau"
```

Die Variable enthält drei Ausprägungen: di

L T i p p

Gerade haben Sie etwas Computer-Syntax gesagt, Befehle aus der Programmiersprache Ihre diese Befehle nicht kennengelernt. Sie versuchen (nicht ganz). Ignorieren Sie diese Befehle einfach.

o

Definition (Thi2d.y9.DuantaE)r dy D(attiad y data, „Normal form“ versteht man eine Tabelle, in der jede Zeile einen darstellt, jede Spalte eine Variable und jede Zeile eine „Name der Variablen“ stehen.)

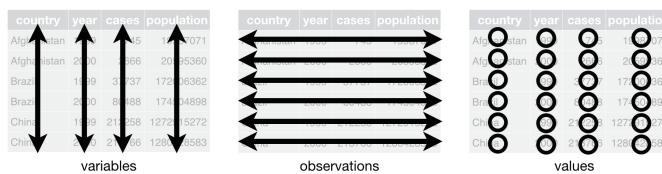


Abbildung 2.11.: Tidy-Data-Sinnbild (Wi

Tabelle 2.1 ist ein Beispiel für Tidy-Data. Abbildung 2.1 zeigt die Tabelle als Abbildung für Tidy-Data (Wickham & Golemud, 2018). Analyse ist es oft sinnvoll, dass die Daten im Tidy-Format sind. Vorteil des Tidy-Formats ist es, dass man weiß, was die Spalten enthalten. Außerdem können Statistikprogramme ohne Umgehen mit den Daten umgehen, s. Abbildung 2.12.

Beispielhre 8 i rma produziert zwei Produkte: Hämofol und Tidex. In Folgendem sind zwei Tabellen dargestellt, die die Umsätze der Firma für beide Produkte darstellen: den Umsatz Ihrer Firma für Hämofol und den Umsatz Ihrer Firma für Tidex. Dazu eine Nicht-Tidy- Tabelle (Tabelle 2.3; Breitformat) und eine Tidy- Tabelle (Tabelle 2.4; Langformat) verwendet.

2. Rahmen

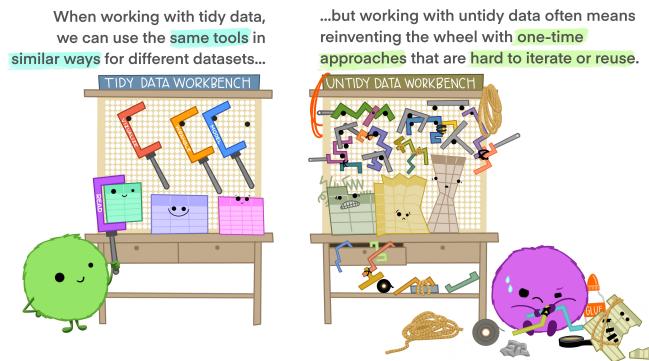


Abbildung 2.12.: Immer schön Ordnung halten

Tabelle 2.3.: Beispiel für eine NICHT-Tidy

Produkt	Umsatz_2021	Umsatz_2022	Umsatz_2023
Hämmer	10	11	12
Nägel	15	10	5

Tabelle 2.4.: Beispiel für eine Tidy - Ta

Produkt	Jahr	Umsatz
Hämmer	2021	10
Hämmer	2022	11
Hämmer	2023	12
Nägel	2021	15
Nägel	2022	10
Nägel	2023	5

Übungsaufgabe 4: Schreiben Sie ein Beispiel für eine K-Tabelle im Lang- vs. Breitformat.

💡 Wo zu braucht man Tidy Data?

2. 4. Was sind Daten?

💡 In vielen Software-Programmen der Datenmanz. B. der X- oder Y-Variablen eine Spalte e. Möchte man etwa die Veränderung des Umsatzes der Jahre visualisieren oder analysieren, Spalten „Jahr“ und „Umsatz“, also ein Tidy-F bzw. Tabelle 2. 4.

Abbildung 2. 13 stellt auf Basis einer „Tidy-Tagrammdar“. Ohne Tidy-Daten wäre dieses Diagramm zu erstellenengewesen.

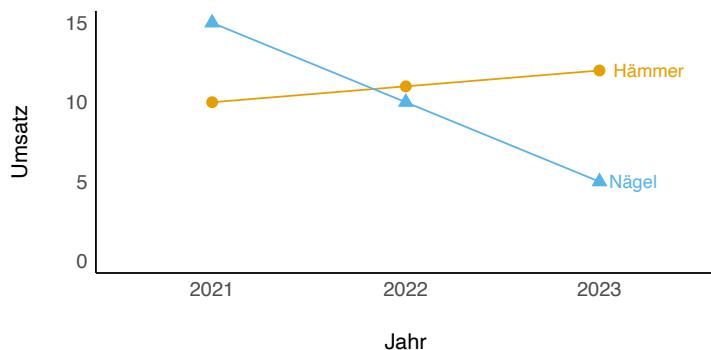


Abbildung 2. 13 ist für eine Visualisierung auf Tabelle, vgl. Tabelle 2. 4

fi

Was Daten betrifft, könnte man behaupten: „V mehr, desto besser“. Natürlich un³Veire\$onst gl Datenmüll ist natürlich nicht besser als sein p Fakten!

Beispiel: Um Ihr eigenes Lehraktivität zu organisches ein Bild machen, wieviel Ihre Nebensitz im Hörsaal solernen. Sie blicken nach links und

³Ceteris paribus, auf Latein, hört sich gleichviel schl

2. Rahmen

dusso?". Sie blicken nach recht und wiederholen Kommissionen, der rechts neben Ihnen die zwei Zahlen (unter der Annahme, dass Sie haben), und teilen durch zwei, um den Mittel-

Einkritischer Geist könnte anmerken, dass Sie nicht gemacht hätten (auch wenn Sie, vielleicht statistische Untersuchung angestellt haben). Personen ist die Ungenauigkeit ihrer Schätzungen bei Studierenden einfach zu hoch.

Abbildung 2.14 veranschaulicht, dass man schätzen kann, wenn man auf eine größere Stichprobe geht. Das Teilbild links zeigt den Mittelwert einer Stichprobe von $n=20$ Beobachtungen. Das Teilbild rechts zeigt den Mittelwert einer Stichprobe von $n=200$ Beobachtungen (jeweils aus der gleichen Population). Wie man sieht, ist im linken Teilbild als im rechten Teilbild:

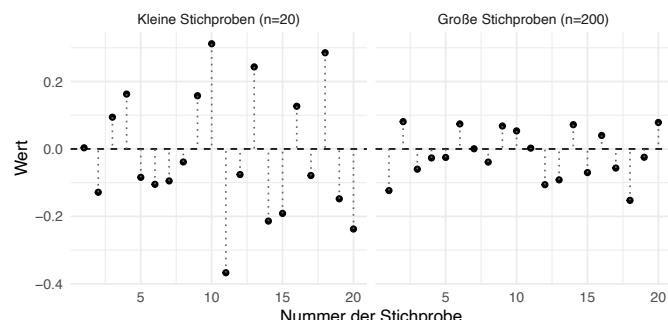


Abbildung 2.14 zeigt die Ungenauigkeit als Funktion der Stichprobengröße. Jeder Punkt stellt eine Stichprobe mit $n=20$ (links) oder $n=200$ (rechts) dar. Kleine Stichproben (links) haben im Schnitt eine Abweichung vom wahren Mittelwert als große Stichproben (rechts).

E W i c h t i g

Mehr Daten = genauere Ergebnisse (unter sonst den)

Übungsaufgabe 2: Experiment zum Effekt der Stichprobengröße
In diesem Live-Experiment untersuchen wir die Streuung des Mittelwerts bei der Münzwürfeln und die Ergebnisse mehrerer Stichproben auf.

In diesem Experiment werfen Sie (in kleinen Gruppen) eine Münze und notieren das Ergebnis. Interessiert dabei die Frage, ob die Ergebnisse (= Münzwürfe) anders streuen als in großen Stichproben.

Sie brauchen nur experimentieren (zwei Partnergruppen mit 2-4 Personen), eine Münze und kann's los gehen! Scannen Sie den QR-Code, um mit dem Experiment zu starten.

Die Daten aller Versuche können Sie hier einsehen.⁴

Beispiel Dorfschulen machen die schlechtesten Schulsportleistungen? Pressemitteilung sei zulesen, dass die bestens schulenden Schulen zufinden seien. (Das ist eine effektive Recherche finden Sie heraus, dass diese Aussage beruht: Tatsächlich sind die Note durchschnittlich deutlich besser als in den großen Schulen. Stimmt die Behauptung der Pressemitteilung? Das Hirnwachsen? Sie recherchiieren noch etwas. Dann fällt Ihnen auf, dass es kommen auch aus den Dorfschulen! Eine statistische Erklärung bischule gibt es nur wenig Kinder und vergleichbar)

⁴<https://tinyurl.com/3w8ke2n2>

2. Rahmen

- die Stichproben sind also klein. Bei kleinen Variationen um den Mittelwert herum, s. Abbildung oben (guter Notenschnitt) und nach unten (schlechter Notenschnitt).

Zum Beispiel:

Angenommen, Ihre Forschungsfrage lautet:

Hat Lernen einen Einfluss auf den Prüfungserfolg?

In dem Fall gilt:

Lernen ist die Input-Variablen, X-Variablen, UV (Befragung ist folgend Output-Variablen, Y-Variablen abhängige Variable (AV))

Abbildung 2.15 stellt diese beiden „Positionen“ dar. Die linke Position ist vor dem Pfeil (X). Die rechte Position ist hinter dem Pfeil (Y).

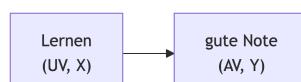


Abbildung 2.15 zeigt zwei synonyme Bezeichnungen für Output-Variablen einer Forschung.

Übungsaufgabe: Legen Sie sich eine Forschungsfrage mit einer UV und einer AV. Nennen Sie eine angeschwungene Frage und fragen Sie, was die UV und die AV belohnen. Sie großzügig.

2.5. Arten von Variablen

Ž

\$

Definition von Skalen. Die reale Klassifikation wird verwendet, um die Art und Menge der Information, die es gibt, zu benennen. Diese Klassifikation basiert auf den Daten und den mathematischen Operationen, die auf den Daten angewendet werden können.

Abbildung 2.16 gibt einen Überblick über typische Skalenniveaus.

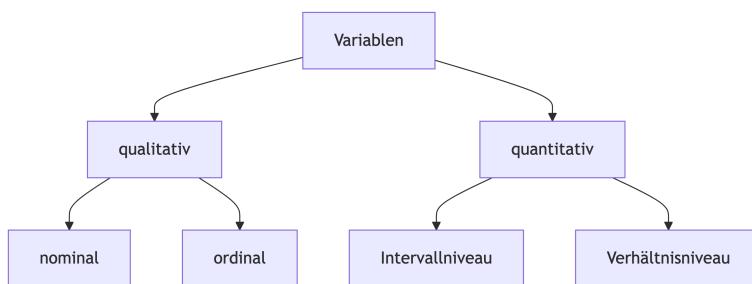


Abbildung 2.16.: Skalenniveaus

\$

Beispiele zu den Skalenniveaus sind in Tabelle 2.5. dargestellt.

Tabelle 2.5.: Beispiele für Skalenniveaus

Variablen	Skalenniveau
Haarfarbe	Nominalskala
Augenfarbe	Nominalskala
Geschlecht	Nominalskala
Automarke	Nominalskala
Partei	Nominalskala
Lieblingsessen	Ordinalskala

2. Rahmen

Tabelle 2.5.: Beispiele für Skalen

Variab le	Skalenniveau
Medaillen beim 100-Meter-Lauf	Ordinal
Uniranking	Ordinalskala
IQ	Intervallskala
Extraversion	Intervallskala
Temperatur in Celsius	Interval
Temperatur in Fahrenheit	Interval
Temperatur in Kelvin	Verhältnis
Körpergröße	Verhältnisskala
Geschwindigkeit	Verhältnis
Länge	Verhältnisskala

Jenachdem, über welches Skalenniveau eine
schi edenen Rechenoperationen erlaubt, s.

Tabelle 2.6: obte Rechenoperationen nach Skalenniveau; nein, nicht erlaubt:

SkalennQuantita - veau tiv	+	×
Nominalnivea u	x	x
Ordinalnivea u	x	x
Intervallnivea u		x
Verhältnisnivea u		

Was soll das bedeuten, „Rechenoperationen“
Skalenniveauein „Rechenbeispiel“ an.

Nominal: siehe Wageschlecht ist nominal skaliert. Da
dass ihre Ausprägungen. B. nicht (sinnvoll)

2.5. Arten von Variablen

oder sonst wie „verrechnet“ werden können. Man tippen zuerst in einer Fernwechse mit. Damit darf man aber nicht rechnen! Nicht addieren etc. Es macht keinen Sinn zu sagen: „Ich habe Mann in meiner Tabelle, das ist im Schnitt ein der Mittelwert von 1 Frau und 1 Mann“! „Geschlecht“ kann man auf der Nominalskala machen. Ein Mann kann feststellen, ob ein Objekt gleich zu unterschiedlich. Also ob zwei Personen das gleiche Geschlecht sind. Aber von unterschiedlichem Geschlecht sind.

- FRAU MANN
- FRAUFRAU
- MANN MANN

Ordinal Skala stellt einer Rangordnung nach ist etwa die geordnete Abfolge ihres Leibes (Schädel, 3. Schnitzel). Etwas „formaler“ ausgedrückt:

- PIZZA → SPAGHETTI

Das komische Zeug dagegen Ben: „Ist auf meiner Liste richten weiter oben, mag ich milchabgeern“, „Mach kann mag aber Pizza um 42% mehr als die Spaghetti und mehr als ein Schnitzel!“. Zum mindest kann man dort Informationen und Annahmen. Es gibt also Dinge (Leicht eine Rangordnung bringen kann, aber der Größe der Unterschiede bemessen kann. Das Die Ordinalskala erlaubt also, Objekte zu ordnen (Merkmals). Die Abstände zwischen den Objekten quantifziert werden.

Intervallass ist vielleicht eine Überraschung, heute 10° Celsius und morgen 5° . Das heißt, es ist so warm wie morgen. Ja, 10 ist das Doppelte von 5 , nicht doppelt so warm wie 20° Celsius. Wenn Sie das normal, so geht es vielen Leuten, wenn sie das Den Grund, dass es nicht erlaubt ist, Verhältnisse soviel etc.) auf der Celsius-Skala zu bilden.

2. Rahmen

der Skala, 0°C , keine echter physikalischer eben nicht Null Wärmeenergie vor. Statt der Wärmemengen gewählt, die für uns Menschenfällig ist: der Gefrierpunkt von Wasserlaubt ist, ist das Addieren (und Subtrahieren) das ist ein Unterschied von 5°C . Oder: Im Sommer ist genau in der Mitte von 5 und 10°C . Abbildung zeigt die Intervallskala.

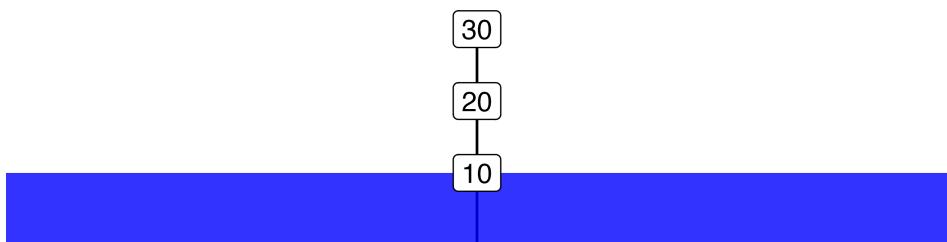
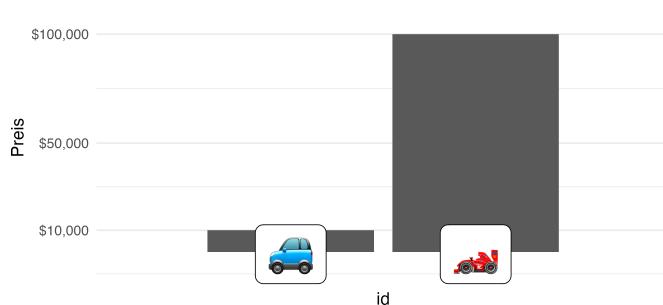


Abbildung zeigt, dass Metermaßstecktmäßig im Wassermaß können wir die aufgedruckten Maße ablesen. Aber wir wissen nicht, ob der Metermaßstecktmäßig steht. Wir wissen demnach nicht, ob die angegebene Nullpunkt der wahre Nullpunkt (auf dem Boden) ist.

Verhältnisskala ist das, was unter einer metrische Variable vorstellt: alle Rechenoperationen sind erlaubt. Zuzüglich anderer, „niedrigeren“ Skalenniveaus erlaubt von Verhältnissen - Multiplizieren (und das Bildung 2.18).

Außerdem können quantitative Variablen wie:

- stetige Variablen, das sind Variablen, bei denen Ausprägungen immer noch eine weitere quantitative Größe zwischen diesen Werten zwischen 1.601 m und 1.602 m, etc.



A b b i l d u n g 2 . 1 1 D 8 e r : r o t e F l i t z e r i s t 1 0 M a l s o t e u
Möhre. Kohlen zusammenkratzen.

I n d i e s e m V i d e o g i b t e s n o c h a u - l i c h e r e E
z u m T h e m a S k a l e n n i v e a u s .



- d i s k r e t e Variablen, dass sind metrische Variablen
te Ausprägungen haben, häufig sind das die
1, 2, E.i.n. Beispiel wäre die Anzahl der Kinder

F r a g e n n a c h S k a l e n n i v e a u s g e h ö r e n z u d e n L i e b
i n d i e s e m T h e m e n b e r e i c h . S i e s i n d g u t b e r a t e n ,
F r a g e i n t e n s i v e r z u b e s c h ä f t i g e n . A u c h i n t h e
F ä c h e r n w i r d i m m e r w i e d e r d i e F r a g e n a c h d e m S
w o r f e n . D a s z e i g t n a t ü r l i c h a u c h d i e h o h e R e l e

Ü b u n g s a u f g ü b b e r 2 . 1 2 S i e s i c h f ü r e i n i g e V a r i a b
n i v e a u s u n d b e f r a g e n S i e d a n n i n t e r e s s i e r t e M

{

W o r a n d e n k e n S i e b e i m W o r t „ M o d e l l “ ? V i e l l e i c h
A b b i l d u n g 2 . 1 9 .

2. Rahmen



Abbildung Ma2ch9.o: - Autos sind Modelle für
(2017)

Definition Modellle sind ein vereinfach
Realitätspausen (Kapitel am 2009).

Beispiel Bei spielle für Modelle sind
Landkarten für Landstriche und das Atommodell
Modell für Atome.

Auch in der Statistik nutzen wir Modelle. Er
blickt nicht durch, s. Beispiel 7.1. Gerade
Stunden seine Studenten auf die Prüfung. Die
Zahlen kann er nicht umgehen ... Geben Sie ihm
Sie ihm, wie lang die Studiostypischerweise
einfach Mittetneidwertenzeiten, 9.6 Stunden).

Beispiel Prof. Weiss - Ois blickt nicht durch)

💡 Vorher: 12, 8, 10, 11, 10, 9, 13, 9, 14, 9,
12, 9, 6, 9, 12, 13, 9, 9, 6, 10... Oh je, soviel
nix! Wieviel lernen denn jetzt meine Studenten?

💡 Ah, 9.6 Stunden! Yeah, jetzt weiß ich,
so typischerweise lernen. Viel zu wenig

Prof. I. Ch. Weiss - Ois hat den Mittelwert von

2.7. Praxisbezug

Der Nutzen von Modellen ist, dass sie komplexe
chen und damit oft überhaupt erst dem Verständ-
suchung zugänglich machen: Modelle ermöglichen
Datenanalyse bzw. Statistik (die beiden Begriffe
synonym gebraucht) fassen Modelle oft viele Da-
z. B. zu einer einzelnen Kennzahl. Das Verrückt-
man Informationen weg wirft, um eine (andere, h-
Information zu bekommen (Stigler, 2016). Wenig

!

Wir leben im Datenzeitalter; Daten durchdringen
beruflichen, gesellschaftlichen und privaten
Raum. Das hat sich in den letzten Jahren massiv verändert.
Methoden einen regelrechten Boom erlebt haben.
ist durchaus auch kritisch zu betrachten; viele
die Entwicklung im Datenzeitalter - Stichworte
mit Sorge. Egal ob man Daten als Segen oder Fluch
Fällen ist es wichtig, mit Daten umgehen zu können.
Bedeutung von Daten wächst in gleichem Maße wie
Datenanalyse. Denn Daten ohne Sinn sind nutzlos.
kann man sagen, dass Datenanalyse (und damit eine
spezielle Art von Datenanalyse) zu stark nachge-

(\$

Dasilie - Draweir Spiegelbemerkungen tolles Experiment durch
viel Arbeit, viel Stress, endlich geschafft,
(als AV, s. Kapitel 2.5), die Sie untersucht hat.
interessanten Effekt, leider. 1 von 20, das hört es
es dann nicht „elegant“, die 19 Variablen ohne sie
der Schublade liegen zu lassen bis zum Sankt-Nikolaus.
können Sie statt dessen als Ergebnis nur die einigen
Ergebnisse präsentieren, ganz ohne widersprechende

2. Rahmen

Dieser Versuchung nicht zu erliegen, kann gefährlich, missliebige Ergebnisse zuver schen bekommen dann ein falsches Bild der Er von Publikationsbias (Marks-Anglin, Ariel Ergebnisse verschweigt, verzerrt die ges 2014) - ein Fall von wissenschaftlichem Feh

Die Aufgabe von Statistik ist es, durch Zu Modelle zu bilden, die es uns einfacher macht zu verstehen. Zentral ist dabei die Daten. Daten kommen in verschiedenen Vari a in Tabellenform, möglichst im tidy - Format

Die Webseite datenwerk.netlify.app stellt Übungsaufgaben bereit. Sie können die Such nutzen, um die Aufgaben mit den folgenden N

1. variation01
2. Def - Statistik01
3. tidy1
4. Skalenniveau1a
5. Ziele - Statistik
6. variation02
7. Skalenniveau1b
8. tidydata1

“ fl

In vielen Organisationen werden Exceltabellen der Datenverarbeitung verwendet. Excel und ähnliche haben bestimmte Stärken und Vorteile, aber auch Schwächen; das liegt z.T. daran, dass Excel für beide und für andere weniger gut geeignet ist. Wenn man wiederholen sicherfahrungsgemäß immer wieder bzw. kommt es wiederholt zu einer suboptimalen Aufbau einer Exceltabelle.

Dieser Artikel von Brozman & Woo (2018) zeigt abschließend Tipps, wie man Exceltabellen so aufbaut werden.

Übungsaufgabe 8 Sie den Artikel von Brozman zusammenlassen SWesdenst (umchaur das Wesentliche) Artikel zusammen.

\$ \$ / |

Sind Sie süchtig nach Ihrem Handy? Passen Sie eine kleine Studie dazu (ggf. live im Klassenzimmer) durch! In Sie diese Umfrage zum Thema Smartphonese aus (anonym und kein Muss). 

Kernstück der Umfrage ist die Smartphone-Sucht (2013). Eine Studie fand, dass ca. ein Siebtel nach ihrem Smartphone ist (Haug et al., 2015). Das Thema eine hohe Bedeutsamkeit zu kommen.

2. Rahmen

↳

Einen Einblick in die Fundamente statisti (2016). Cetinkaya-Rundel & Hardin (2021) st der Analyse von Daten im Kapitel 1, „Hello illustriert statistische Überraschungen vor allem: sofataugliche Art.

◦

"

A b b . A b b i l d u n g 1 . 4 d e n S t a n d o r t d i e s e s K a p i t e l s , d a m i t e i n e n Ü b e r b l i c k ü b e r d a s T h e m a d i e s e s K a p i t e l .

Ł

- Sie können R und R Studio starten.
- Sie können R-Pakete installieren und starten.
- Sie können Variablen in R zuweisen und auslesen.
- Sie können Daten in R importieren.
- Sie können dRempBreogdruizfidebarkeeh.

fl \$ #

B i t t e s t e l l e n S i e s i c h e r , d a s s S i e R r e c h t z e i t i g u n t e n i n d i e s e m K a p i t e l f n d e n S i e I n s t a l l a t i o n . F a l l s S i e d i e s e s K a p i t e l z u m e r s t e n M a l b z w . a u s k e n n e n , w e r d e n S i e v i e l l e i c h t e i n i g e n I n h a u t e n n i c h t g l e i c h v e r s t e h e n . K e i n e S o r g e , d a Ú b u n g w i r d l i h n e n b a l d a l l e s s c h n e l l v o n d e r H a n d .

3. Daten einlesen

"

(#

Gründe, die für den Einsatz von R sprechen:

1. R ist kostenlos, andere Softwarepaketefrei.
2. R und R-Befehle sind quell offen, d. h. man liest den Computerbefehle anschauen. Jemand kann vernünftig gearbeitet. Alle können beitragen.
3. R hat die neuesten Methoden.
4. R hat eine große Community.
5. R ist maßgeschneidert für Datenanalyse.

Allerdings gibt es auch abweichende Meinungen

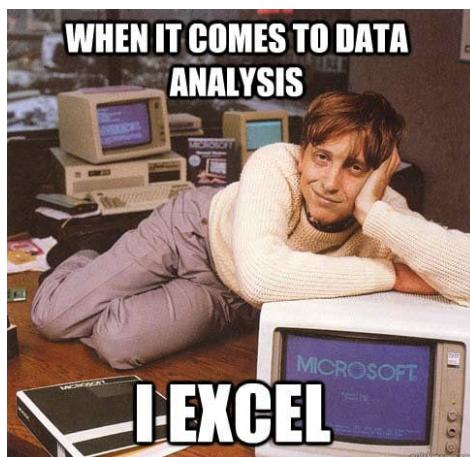


Abbildung 3c. Manche finden Excel cooler als R, nicht (imgflip, 2024a)

#

Definition (Originaldefinition): Ein brauchbares wissenschaftliches Ergebnis ist reproduzierbar, wenn andere Personen mit dem gleichen Ergebnis (wie in der ursprünglichen Arbeit, Plesse et al., 2018).

Definition 3.1 ist, etwas überspitzt, in Abbildung 2 dargestellt.



Abbildung 2: Syntax + genaue Beschreibung des Ergebnisses = reproduzierbar

Beispiel aus der Forschung: Reproduzierbarkeit von Befunden

💡 Wie steht es um die Reproduzierbarkeit in der Psychologie? Sind die Befunde zuverlässig?

Obel et al. (2020) haben die Reproduzierbarkeit von Befunden untersucht. Sie berichten folgendes Ergebnis:

We examined data and code sharing for Registered Reports published in the psychological literature and attempted to independently compute the main results in each article. Of the 62 articles meeting our inclusion criteria, 41 had data and code for replication scripts available. Both data and code for 31 articles were shared. We could run the scripts for 21 articles and reproduced the main results for 21 articles.

\$

Wenn wir sagen, „wir arbeiten mit R“, dann heißt das, wir arbeiten mit R und mit RStudio.

3. Daten einlesen

Ismay & Kim (2020) zeigen in einer schönen Abschließung, dass RStudio macht, s. Abbildung 3.3. (Statt RStudio für die Datenanalyse irrelevant Sie werden es nicht missen wollen.)

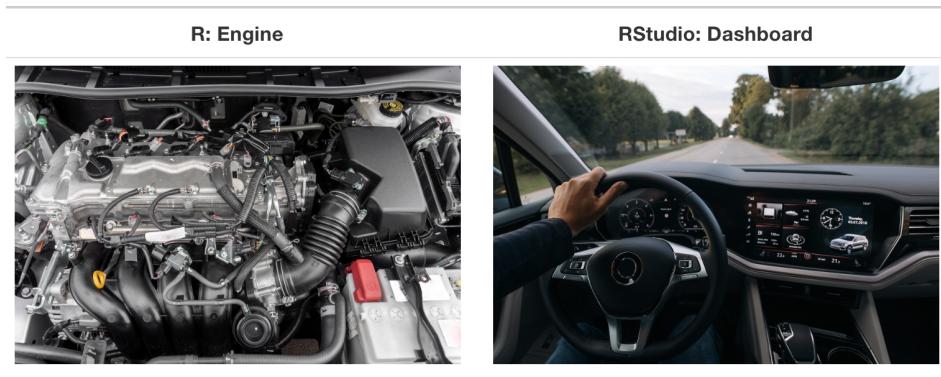


Abbildung 3.3. RStudio: R macht die Arbeit, RStudio: Übersicht (Ismay & Kim, 2020).

Kurz gesagt: Das eigentlich Arbeiten besorgt die Schönheit ist RStudio zuständig. Auch ein

/ # # \$

/ #

R ist ein Softwarepaket für statistische Beobachtungen. Ihr Betriebssystem herunter unter <https://www.r-project.org>

Wenn Sie beim Herunterladen gefragt werden, auswählen sollen, heißt das, Sie sollen eine von dem Sie R herunterladen. Der sollte möglicherweise stehen, dann spart es vielleicht etwas Zeit.

3.3. Installation von R und

Wenn Sie die Installationsdatei heruntergeladen haben (Doppelklick) und Sie werden durch die benötigten Admin-Rechte auf Ihrem Computer.)

/ # \$

RStudio ist eine grafische Benutzeroberfläche (GUI) für R, plus ein paar Geometriesettingen für Entwicklungsumgebungen und Development environments. Es kann für Windows, Mac OS X und Linux heruntergeladen werden. Sie können die Installationsdatei heruntergeladen haben (Doppelklick) und Sie werden durch die benötigten Admin-Rechte auf Ihrem Computer.

! # \$

Posit Cloud bzw. RStudio Cloud (<https://rstudio.cloud>) ist ein Dienst von Posit (zum Teil über das Unternehmen mit R arbeiten). Sie können es als Installation von RStudio auf Ihrem Computer verwenden. RStudio Cloud ist, dass es sich um eine Installation handelt, die auf dem Desktop stattfindet. Ein Nachteil ist, dass es etwas langsam ist und kostet es etwas. Sie müssen einen Anbieter anlegen, um den Dienst nutzen zu können.

Wenn Ihnen jemand (z.B. eine Lehrkraft) einen Projektordner bzw. einen Link dazu bereitstellt, dann kann die Lehrkraft dann schon Pakete installieren und andere Nettigkeiten vorbereiten kann für Sie. Sie den Projektordner in Ihrem eigenen Konto an und speichern möchten, da Sie vermutlich keinen

<https://posit.co/download/rstudio-desktop/>

3. Daten einlesen

Projektordnerr dieser nettern Person (Ihre Sie dazu auf „Save a permanent copy“, s. Abb



Abbildung 5: Ein Projektordner im eigenen Kont Schreibrecht zu haben

Sie können auch von der Cloud exportieren, unterladen. Klicken Sie dann auf „Export“ „Fill“

Hinweis

RStudio starten, nicht R.

Wir verwenden beide Programme (R und RStudio). RStudio findet selbständig R und öffnen Sie nicht noch extra R (sonst wäre R zweimal läuft). Prem Computer/Desktop) kann RStudio (online-Version) starten

!

(# !

Typisch für R ist sein modularer Aufbau: Man erweiterungen (Pakete) hinzufügen, alles los. In RPaketen „wohnen“ R-Befehle, also sozusagen. Außerdem können in R-Paketen auftreten. Damit man die Inhalte eines R-Pakets zuerst installieren und dann verfügbar machen sich daher ein R-Paket vorstellen wie ein hat, dann kennt es die Inhalte. Diese Inhalte

3.4. R-Pakete

Formeln, also Berechnungen sein. Es könnte aber auch für ein schönes Diagramm sein. Ist ein spezieller Computer installiert, so können Sie diese Funktionen benutzen.

Die Anzahl der R-Pakete ist groß; alleine auf dem CRAN (nennt sich „CRAN“) von R gibt es ca. 20,000 Pakete. Und es kommen immer mehr dazu.

!

Wie jede Software muss man Pakete (Erweiterungen) installieren, bevor man sie verwenden kann. Üblicherweise reicht das Installieren und dann den Namen des zu installierenden Pakets eingeben.

💡 Welche R-Pakete sind denn schon installiert?

Im Rechner können Sie nachschauen, welche Pakete auf dem Computer schon installiert sind.

Alternativ können Sie zum Installieren von Paketen die `install.packages()` Funktion benutzen. Als Option kann man `tidyverse` angeben, um das Paket `tidyverse` zu installieren.

💡 Ja, aber welche R-Pakete „soll“ ich denn eigentlich brauchen?

Im Moment sollten Sie die folgenden Pakete installieren:

- `tidyverse`
- `easystats`

Wenn Sie die noch nicht installiert haben sollten, kann man die `install.packages()` Funktion benutzen, die nur dafür da sind, mehrere Pakete zusammen zu installieren. Die Paketdateien verbinden die Pakete und damit wird nicht alle Pakete einzeln installieren und s

3. Daten einlesen

das Paketdry vdeemskeomfort, alle die Pakete diese einmal zu starten. Praktisch.)

F Vorsicht

Bevor Sie ein R-Paket (oder überhaupt irg
stallieren/update, sollten Sie das entsp
ßen/beenden. Sonst schrauben Sie sozusag
schen Gerät herum, das noch unter Stromste
einfachste Art, alle Pakete zu brennen
klicken (in RStudio).

!

Wenn Sie ein Softwareprogramm installiert
startSeine erkennen leicht, ob ein Paket bere
wenn Sie ein Häkchen vor dem Namen des Paket
(ReiPtaerk å gæshen. Ein bestimmt es ReiPmaelt imus
install.IAbeerremann rjuesdse esMai n e uwstharmtaennR (bzw.
RStudio) startet.

{ #

! #

EiProjienRStudio ist letztlich ein Ordner,
Reihe von zusammengehörigen Dateien verwendet
nennen Ihr Projekt_RStudio legt uns diesen Ordner
von uns gewählten Platz auf unserem Computer
tisch, weil man dann sagen kann „Hey R, niemand
ohne einen Pfad anzugeben. Vorausgesetzt, der
jektorcdmodr_).tRStudio-Projekte kann angelegt
Icon, das einen Quader mit dem Buchstaben R
Sie RStudio-Projekte, das macht Ihr Leben le

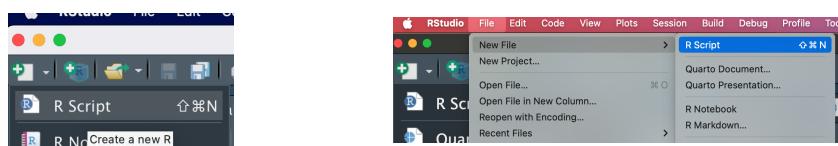
3.5. Mit R arbeiten

nutzen ist viel sicherer als das Arbeitsverzeichnis oder mit Pfaden herumzubasteln.

\$

Die R-Befehle („Syntax“) schreiben Sie am besten vorgesehene Textdatei in RStudio. Eine Sammlung von Befehlen nennt man Skriptdatei. Spricht man bei Dateien Syntax, enthält Skriptdateien

Um eine neue R-Skriptdatei zu erstellen, gibt es: klicken Sie auf das Icon, das ein weißes Pluszeichen zeigt, s. Abbildung 3.5.



(a) Klicken auf Icon (b) Auswahl im Menü

Abbildung gibt verschiedene Wege, um eine neue Skriptdatei in RStudio zu öffnen. (a) Per Klick auf Menüfile > R Script

Vergessen Sie nicht, wenn Sie ein tolles Skript geschrieben haben. Dafür gibt es mehrere Möglichkeiten:

1. Tastaturskürzel
2. Maus > Save
3. Klicken auf das Icon mit der Diskette, s. Abbildung 3.5

Eine existierende Skriptdatei können Sie mit

1. Strg + O
2. Klicken auf das Icon mit der Akte und dem grünen Auge, s. Abbildung 3.5
3. Maus > Open File...

3. Daten einlesen

Quarto ist ein Programm zum Erstellen von Texten. Syntax einfügen kann. Die Ausgabe der R-Befehle im Dokument eingebunden. Quarto ist in RStudio eine komfortable und leistungsfähige Methode, Syntax zu schreiben. Sie sind aber nicht verbindlich. Stattdessen können Sie Ihre Syntax auch in Schriftarten ändern.

Wenn Sie Quarto nutzen möchten, müssen Sie es herunterladen. Dann können Sie in RStudio ein neues Quarto-Dokument öffnen. Sie gehen > New File > Quarto Document

In jeder Programmiersprache kann man Variablen definieren:

```
richtige_Antwort =  
falsche_Antwort =  
typ("Antwort")  
ist_korrekt =
```

Alternativ zum Gleichheitstest `==` ist `!=` für den Vergleich von Werten bestimmt. Beide führen zum gleichen Ergebnis. Allerdings ist der Zuweisungspfeil `=` prägnanter.

Die Zuweisung `=` weist das Gleichheitstest `==` einem neuen Variablenwert überschreibt den Inhalt, wenn es existiert).

<https://quarto.org/>

```
r i c h t i g e _ 4 a n t w o r t <-
f a l s c h e _ a n t w o r t <-
t y p " A - n t w o r t "
i s t _ k o T R U K t <-
```

Sie können sich eine Variable wie einen Becher oder bestimmte Werte enthalten, z. B. den Wert „9° C“ steht der Name des Bechers geschrieben, z. B. „können Sie die Werte aus dem Becher entfernen ersetzen (vgl. Abbildung 3.6).

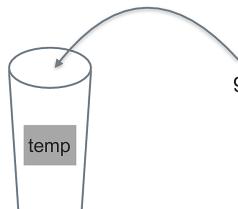


Abbildung 3.6: Beispiel ablen zuweisen: Der Vatremple mit zuweisen wir den Wert

Kann übrigens auch rechnen. Probiieren Sie es aus!

```
d i e _ s u m m e <- f a l s c h e _ t à g e w a n t w o r t
```

Aber was ist jetzt der Wert, die „stammelt“ der Variable? Um den Wert, d. h. den Inhalt dieses neuen Variablen in einfacher den Namen des Objekts sein:

```
d i e _ s u m m e
# # [ 1 ] 85
```

Was passiert wohl die_wertermittlung folgt defnieren

3. Dateneinlesen

```
d i e _ s u m m e <- f a l+s c h e t ä g t +W a n t w o r t
```

Wer hätte's geahnt:

```
d i e _ s u m m e  
# # [ 1 ] 86
```

Variablen können auch „leer“ sein:

```
a l t e N r A # N A wie "not available", nicht vorhanden  
a l t e r  
# # [ 1 ] N A
```

NAs steht für unavailable und ist verfügbar und macht deutlich, dass ein Wert fehlt.

💡 Wo zu brauche ich bitte fehlende Werte?

Fehlende Werte sind ein häufiges Problem in R, wenn sich die befragte Person geweigert, ihr Alter anzugeben oder als Sie die Daten in Ihren Computer eingibt. Katze über die Tastatur gelauen und alles.

Das, was R kann, ist in „Funktionen“ hinter „Befehl“ an eine Funktion.

Definition (funktional) Eine Funktion ist eine Regel, die einen Wert (auch Argument genannt) in einen Ausdruck umwandelt. Abbildung 3.

3.6. Errisch für Einsteiger

Ein Beispiel für eine solche Funktion könnte sein „Bestwert dieser Datenreihe“ (schauen wir uns gleich nach).

Antwort(4243)

Der Befehl `c` (combine) fügt mehrere Werte zusammen zu einer „Liste“ (einem Vektor). (Streng genommen sollte die Liste sprechen, da es in R noch eine Insel der Objekte gibt, und eine verallgemeinerte Form meines Vektors ist.)

Definition (Vektor): Ein Vektor (Datenreihe) bezeichnet eine geordnete Folge von Werten. In R kann man sie erstellen. Die Werte eines Vektors bezeichnet.

Mit dem Zuweisungspfeil geben wir diesem Vektor einen Namen. Dieser Vektor besteht aus 3 Werten, kommt 3

Beispiele (Beispiele für Vektoren): (praktisch) liebig lang sein, z.B. drei Elemente.

```
x <- c(1, 2, 3)
y <- c(2, 1, 3) # x und y sind ungleich (Reihenfolge der Werte)
z <- c(3, 1, 2, 4) # 1
name <- "Anzahl Bälle im Kasten" # Titel - Vektor
```

Zwei wichtige Typen von Vektoren sind numerische Zahlen; in R numerisch (z.B. 1, 2, 3) und Textvektoren in R aus Strängen (z.B. "Anzahl Bälle im Kasten").

Beispiel: Weist Größe Beispiel für Funktionen sind:

- „Erstelle eine Liste (Vektor) von Werten.“
- „Lade dieses R-Paket.“
- „Gib den größten Wert dieser Datenreihe aus.“

3. Daten einlesen

&

Jetzt wird's ernst. Jetzt kommt die Statistik
erste statistische Funktion: Den Mittelwert

```
mean( Antworten )  
## [1] 42
```

Sie hätten zwei aufeinanderfolgende Zahlen eingetragen können, so wie Sie ja schließlich die Variable gerade definiert haben.
Sie kann nun die Variable mit der Befehlszeile
mean(Antworten)

```
summary(c(42, 43))
```

Schriftart 2:

Abbildung 3.7 stellt eine Funktions schematische Darstellung 3.7 dar. Eine Funktion hat einen Parameter x , der Argumente, die sind Daten oder Verarbeitungshinweise, die eingegeben werden. Eine Funktion liefert im Ergebnis y (Output), in der das Ergebnis einer Funktion abgelegt ist.

Definition Argumente einer Funktion „einer“ Funktion, in denen man die Eingabewerte implementiert

So hat die Funktion folgende Argumente, s. L

- x : das ist der Vektor, für den der Mittelwert berechnet werden soll
- trim=0: Sollen die extremen Werte abgeschnitten werden, also nicht in die Berechnung des Mittelwerts einbezogen werden?

3.6. Errisch für Einste

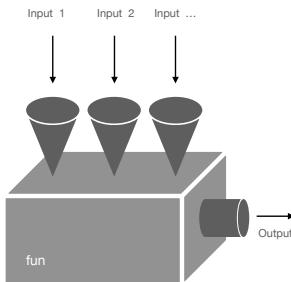


Abbildung 3.7.: Schema einer Funktion

Liste Dig & Argumente der MaFunktion

me a(nxt,ri on=a,rFnA±S E...)

- na. r m = F A M i S e E s o l l mi t f e h l e n d e n W e r t e n w e r d e n ? I m S t a n d a r t u n d f e i r e l l e a n d e r e a r i t h m e F u n k t i o n a l a z u m R ü c k . R s c h w e n k t s o z u s a g e n d i e r u m z u s i g n a l i s i e r e n : A c h t u n g , M e n s c h , h i e i n O r d n u n g . S e t z a t . m a n n , d a s e i n e n t f e r n t (r e m o v e r m) R d i e f e h l e n d e n W e r t e u n d b e r e c h n e t d e n w e i t e r e H i n w e i s e z u d e n f e h l e n d e n W e r t e n .
- . hei ßt „sonstiges Zeugs, das manchmal eine R darum kümmern wir uns jetzt nicht.“

E i n i g e A r g u m e n t e s h a b e d a r i z w e n d i o r e i n s t e l l u n g (e n d e f à u l s t o w i r d b e i d e r e a f m a k t a i n d a r d n i c h t g e t r i m t m i t (m) = u n d f e h l e n d e W e r t e w e r d e n n r m i c h t e n F A L S E)

3. Dateneinlesen

Wenn ein R-Befehl ein Argument mit Voreinstellung für dieses Argument aufzufüllen. In dem Fall wird die Voreinstellung zurückgegriffen. Argumente, die keinem Namen entsprechen müssen Sie aber auf jeden Fall mit einem Namen wenden, da es meist so angefordert ist:

Bei jedem R-Befehl haben die Argumente eine Voreinstellung, etwa `mean(x, trim = 0, na.rm = FALSE, ...)`

(Nur) wenn man die Argumente in ihrer vorgegebenen Anzahl anspricht, muss man den Namen des Arguments anführen:

`mean(Antworten, 0, FALSE)`

Hält man sich aber nicht an die vorgegebene Regel, was zu tun ist und führt es sich in einer Fehlermeldung:

```
mean(nAntworten) # FALSCH, DON'T DO IT
## Error in mean.default(Antworten, FALSE, 0)
length one
```

Wenn man die Namen der Argumente anspricht, ist es egal:

```
mean(mar = 5, Antwort) # ok
mean(trim = Antwort) # ok
```

Übrigens: Leerzeichen sind R fast immer erlaubt, sichtlichkeit sollte man aber Leerzeichen nicht erlauben:

- < -
- <= etc.
- Variablennamen

3.6. Errisch für Einsteiger

'

(

Sagen wir, wir haben einen fehlenden Wert in un-

```
Antwort(4 24 3NA)
```

```
Antworten
```

```
## [1] 42 43 NA
```

Wenn wir jetzt den Mittelwert berechnen wollen, einemschön.³ Das ist natürlich ein Hinweis, dass die Werte fehlen.

```
mean(Antworten)
```

```
## [1] NA
```

R meint es gut³ mit Ihnen. Sie sich vor, dass R Sie Probleme aufmerksam machen möchte:

⚠ Achtung, NAs, fehlende Werte, lieber Herr du hast nicht mehr alle Latten am Zaun, will sie im Vektor!

(Danke, R.)

Möchten Sie aber lieber R dieses Verhalten aus dem Argument mit dem Wert `na.rm=TRUE` entfernen, also fehlenden Werte).

```
mean(Antworten, na.rm=TRUE)
```

```
## [1] 42
```

³⚠ Naja, manchmal.

3. Dateneinlesen

#

Definition (Vektorielles Rechnen mit Vektoren bezeichnet man Vektorrechnen)

Vektorielles Rechnen ist ein praktische folgende Dinge einfache Rausrechnen.

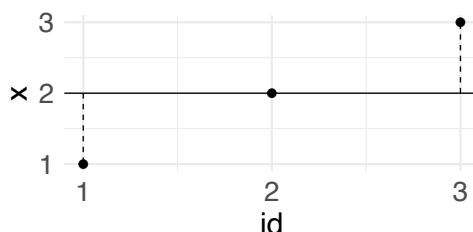
Gegeben sei Vektor $\vec{v} = (1, 0, 1)$. Dann können wir die Differenz („Minus-Rechnen“) jedes Elementes von \vec{v} vom ersten abrechnen:

```
x - me a(nx )  
# # [ 1 ] - 1 0 1
```

Etwas eleganter ausgedrückt: Wir haben die Differenz („Minus-Rechnen“) aufgewandelt. Einzelnen haben wir also folgenden drei Differenzen:

```
1 - 2  
2 - 2  
3 - 2
```

Diese drei Rechenschritte sind symbolisch dargestellt.



Abbildungskörper des vektoriellen Rechnens: auf jedes Element eines Vektors anzuwenden: $1 - 1 = 0$; $2 - 2 = 0$; $3 - 2 = 1$

3.6. Errisch für Einsteiger

"

Ihre R-Muskeln sind gestählt?
Macht nichts! Trainieren Sie sich mit dem R-Qui-

Datenwerk - Webseite!



/ # |

- Wo finde ich Hilfe zu einer bestimmten Funktion? Sie dazu folgende R-Pakete herunterladen: ich geben Ihnen den Namen der Funktion in R Studio Help Suchfeld ein.
- Wenn ich ein R-Paket installiere, fragt mich auch Pakete installieren, will die „kompile ich das machen?“ das ist zumeist nicht nötig; ein.
- In welchem Paket wohnt? Sie suchen häufiger Funktion auf der Webseitaktion
- Ich weiß nicht, wie der R-eBook hilfreich ist oder Ihr Problem auch, und meistens hat irgendschon gelöst. Am besten suchen Sie mal auf fow.com.
- Ich muss mal grundlegend verstehen, wozu ein gut ist. Wässturz? Sie die Dokumentation („Vignette“) des R-Pakets durch. Sie so einen Überblick über die verfügbaren Vignette erhalten. Sie können die entsprechenden (Könnwissel) sich dann die wünschte Vignette auswählen lassen.
- Oh nein, ich sehe rot, das heißt, R zeigt mir Schrift an. Ist ja kein Zeichen ausgelöscht in seiner nicht sparsam mit roter Farbe. Solange es nicht scheint, ist es meist kein Problem.

⁴<https://www.rdocumentation.org/>

3. Dateneinlesen

- R hat sich aufgehängt oder bringt einen sonst alles für bibirementShætauf jeden F AEG- Prinzip (Aus- Ein- Gut): sprich, R ne
- Ich sucheschon seit einer Stunde einen F Ich habeschon verschiedene Gegenstände geworfen. Was MæchænSitetne? ne Pause. Doc ernst gemeint. Meine Erfahrung: Mit etwas klarer und man findet das Problemviel einf ist ein nem das Problem danach schlicht we
- Ir gendwie reagiert R komisch, vielleicht Starten Sie R neu. SKelisci kœnn>SRœsatuafr t R
- Ich muss mal klar Schiff machen und alle Löschen. Wie werden dich dæsi ZmeNuguvsiteadretvlo werden alle Objekte (Variablen) gelöscht Sie selektiv löschen mit ~~læmniotl ðæmnafr éðoaksar t~~ Objektnamen in Skart

F Vorsicht

R ist penibel am Grundsaime medwei verschiedene Vari für R.

Groß- und Kleinschreibung wird von R streng Regal, ob Sie zur besseren Übersichtlichkeit tippen. Ausnahmen sind spezielle Operatoren

Eine gute Nachricht: Warnhinweis warnung) sa können Sie das zumeist Wairgmüönsigt ekreám EFEHööRe (und meistens nicht gravierend oder nicht trotzdem durch. Im Zweifel ist Google eine Erssopricht, ist es auch ein Fehler und Ihre S

{ .

(.

Damit Sie eine Datendatei importieren können, die Datei ist. Schauen wir uns zwei Möglichkeiten an.

1. Irgendwo im Internet
2. Irgendwo auf Ihrem Computer, z.B. in Ihren

In beiden Fällen wird der „Aufenthaltsort“ der Datei definiert (Der Pfad einer Datei gibt an, in welchem Ordner (Unter-Unterordner) die gesuchte Datei liegt). Der Pfad kann aussehen: „/Users/sebastiansaueruser/github/mariokart.csv“ mit dem Namen der Datei definiert.

Hinweis

Wir werden in diesem Kurs häufig direkt den Dateipfad angeben; Sie finden ihn hier.

Daten werden in verschiedenen Formaten im Computer gespeichert. Häufig als

- Excel - Datei
- CSV - Datei

⁵Z.B. hier: <https://vincentarelbundock.github.io/mariokart.csv>

⁶Auf dieser Webseite <https://vincentarelbundock.github.io/mariokart.html> finden Sie eine große Zahl an Datensätzen. Nur für die Übung ist es erlaubt.

3. Daten einlesen

In der Datenanalyse ist das gebräuchlichste Format die CSV-Datei. Für uns Endverbraucher tut diese CSV-Datei beherbergt einfach nichts. In diesem Buch werden wir mit dem CSV-Format arbeiten.

Übungsaufgabe (CSV-Datei öffnen)

Öffnen Sie die CSV-Datei (sie ist direkt mit Word und auch nicht mit Excel). Schauen Sie sich die Datenstruktur an.

Lösung

Eine CSV-Datei repräsentiert eine Datentabelle mittels eines Kommas dargestellt (man wählen, um Spalten voneinander abzugrenzen).

Sie können Daten aus verschiedenen Quellen in einem R-Paket, von einer Webseite oder von einer anderen Quelle einlesen. Es ist egal, ob Sie die Desktop- oder die Cloudnutzen.

Ist Ihr Datensatz schon in einem R-Paket gespeichert, kann man es aus diesem R-Paket starten. Das ist die bequeme „Welt“ der **Matematik R-Paket**.

L Tipp

Häufig wird vergessen, dass ein R-Paket vor dem Starten geladen werden muss.

Auf der anderen Seite muss man ein R-Paket (wie es in der Matematik R-Paket Welt) neu installieren - das Paket muss neu installiert werden.

```
da("mariokart.csv")#nParko'dt muss installiert sei
```

Eine Data-Dictionary findet sich in Anhang B.

Der Befehl `bi` bietet eine Möglichkeit, Daten (in Form von einer Webseite (URL)) in R zu importieren, s

Liste mit Mariokart-Datensätzen downloaden (mit

```
mariookartdata()
" https://vincentarelbundock.github.io/Rdatasets/csv/openintro/mariokart.csv"
```

Es liegt bei Ihnen, welchen Namen Sie der Tabelle wähle oft die Mängel dieses ein kurzer Namen, muss nicht so viel tippen. Auf die Häufigkeit der Spalten ist es

Werfen Sie einen Blick auf die entsprechende engl.

glim(pds)

Online findet sich eine Erklärung (Data-Dictionarie). Sie können auch von Ihrem Computer aus Daten importieren.

Gehen wir davon aus, dass sich die Datei in die R-Daten (Datei) befindet, in der Sie den Importieren schreiben. Dann können Sie die Daten:

```
d <read("mariokart.csv")
```

⁷<https://vincentarelbundock.github.io/Rdatasets/datasets/do>

3. Daten einlesen

Dieses Video erklärt die Schritte des Importierens von Datendateien von Ihrem Computer.



Das Importieren von Daten aus dem Internet ist einfacher als das Importieren von Dateien auf Ihrem Computer zu RStudio. Zum Importieren von Dateien auf Ihrem Computer in RStudio müssen Sie die Datei vorab hochladen müssen, sie kann nicht direkt von der Cloud importiert werden. Ich downloade die Datei und wähle den entsprechenden Ordner als Ziel, indem ich die Datei zum Importieren schreibe, befindet sich in dem Ordner.

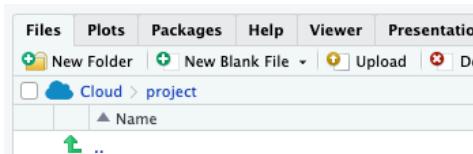


Abbildung 3.9.

Es gibt verschiedene Formate, in denen Daten auf dem Computer abgespeichert werden. Die gebräuchlichsten sind XLSX und CSV. Es gibt auch mehrere R-Befehle, um Daten aus verschiedenen Quellen zu importieren. Bevor Sie die entsprechenden Befehle verwenden, müssen Sie sicherstellen, dass Sie die entsprechenden Pakete installiert haben. Ein Beispiel für die Installation eines Pakets ist:

Die GUI (Benutzeroberfläche) von RStudio erleichtert das Importieren von Dateien erheblich. Sie können die Datei per Klick auswählen und sie wird automatisch in das Projekt geladen. Dies ist besonders praktisch, wenn Sie häufig mit Excel-Dateien arbeiten. Um eine CSV-Datei zu importieren, klicken Sie auf das entsprechende Icon im Menü und wählen Sie die Datei aus. RStudio erkennt das Format und importiert die Daten automatisch.

In der sich öffnenden Maske können Sie nun die Importierende Datendatei auswählen und mit dem Importieren beginnen.

3.7. Mit Datenarbeit

schließlich in R importiert.

Eine in R importierte Tabelle (mit bestimmte Datafr~~ame~~ Dataframes sind in der Datenanalyse von Tabelle 2.2 ist die Tabelle mit den Mario kart gesprochen ein Dataframe). In der Normalform ist Tabelle in Normalform (Tidy - Format), vgl. Definition 2.

Definition (Dataframe): Ein Dataframe (engl. data frame; le) genannt; von „tbl“ wie Table) ist ein Datenobjekt von Tabellen. Dataframes bestehen aus einer oder mehreren Spalten, die einen Namen, sozusagen einen „Spaltennamen“. Spalten müssen die gleiche Länge haben; anschaulich Tabelle (in R) rechteckig. Jede Spalte einzeln aufgefasst werden.

Geben Sie den Namen eines Dataframes ein, um zu zeigen zu lassen. Beachten Sie, dass Sie die Daten anschauen, nicht ändern können.

% #

Wenn Sie in R z.B. Mario Kart spielen, betrachten wollen, klicken Sie am besten auf das Environment-Dropdown neben dem Namen Abbildung 3.10.

mariokart 143 obs. of 12 variables

Abbildung 3.10: Klickt auf das Tabelle-Icon können Ihnen eine Ansicht Ihrer Tabelle

Alternativ öffnet wird eine Blöckeicht (che Ansicht).

3. Daten einlesen

7

💡 Wer will schon wieder wen prüfen?!

In diesem Abschnitt lernen wir, wie man einen Wert überprüfen, ob eine Variable einen bestimmten Wert als Referenzwert hat.

Defnieren wir zuerst eine Variable,

```
x <4-2
```

Dann fragen wir R, ob diese Variable den Wert

```
x = -4.2  
# # [ 1 ] TRUE
```

👋 Hallo, Mensch. Ja, diese Variable hat den Wert (-4.2).

Möchte man mit R prüfen, ob eine Variable den Wert -4.2 hat, so schreibt man:

```
x == -4.2
```

Man beachte die doppelte Gleichheitszeichen. Zur Prüfung der Gleichheit muss man das doppelte Gleichheitszeichen verwenden.

F Vorsicht

Ein beliebter Fehler ist es, bei der Prüfung einer Gleichheitszeichen `x == 7` Minuszeichen zu vergessen. Gleichheitszeichen müssen nicht übereinstimmen, sondern definiert die Variable oder bestimmt einen Wert. Siehe Kapitel 3.6.1.

Tabelle 3.1 gibt einen Überblick über wichtige das Zeichen für das häufigste Maer Tastatur zuer drückt Omaer + 7

Tabelle 3.1.: Logische Prüfungen in R

Prüfung auf	R-Syntax
Gleichheit	$x == \text{Wert}$
Ungleichheit	$x != \text{Wert}$
Größer als Wert	$x > \text{Wert}$
Größer oder gleich Wert	$x >= \text{Wert}$
Kleiner als Wert	$x < \text{Wert}$
Kleiner oder gleich Wert	$x <= \text{Wert}$
Logisches UND	$(x < \text{Wert}1) \& (x > \text{Wert}2)$
Logisches ODER	$(x < \text{Wert}1) (x > \text{Wert}2)$

!

■ R in der Praxis wirklich genutzt? Oder ist R von (vielleicht verwirrten) Profis im Elfen

Schauen wir uns dazu die Suchanfragen bei stack größten FAQ-Forum für Software-Entwicklung anfragen mit fragen [mehr] anfragen [mehr] SPSSnStat ist eine an Hochschulen verbreitete Statistik-S sind in Abbildung Abbildung 3.11 dargestellt

Das ist grob gerechnet ein Faktor von 200 (der SPSS). Dieses Ergebnis lässt darauf schließen mehr als SPSS gebraucht wird.

■ Aber ist R wirklich ein Werkzeug, das mir i

⁸Die Daten wurden am 2022-02-24, 17:21 CET, abgerufen.

3. Daten einlesen

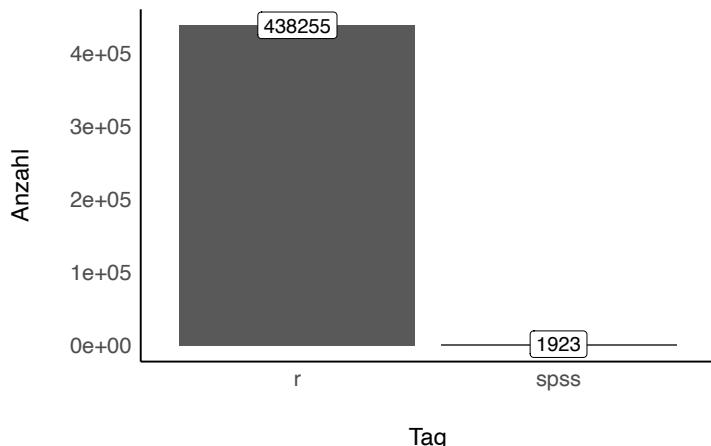


Abbildung 3.11.: Suchanfragen nach R bzw SPSS

💡 Viele Firmen weltweit nutzen R zur Datenanalyse.

💡 R ist der Place-to-be für die Datenanalyse.

💡 Aber ist Datenanalyse wirklich etwas, wobei ein guter Job bekommen?

💡 Berufe mit Bezug zu Daten, Datenanalysten, Künstlicher Intelligenz (artificial intelligence) zu den stark wachsenden Berufen:

Artificial intelligence (AI) continues to grow in our Emerging Jobs lists, which is no surprise given the rise in AI in fields such as data science and because it's a relatively new field that may demand more knowledge of AI than you might expect, real estate and business development (2019).

⁹wie diese Liste zeigt: <https://www.quora.com/What-is-the-best-data-science-tool>

Übungsaufgaben Bi Statistik und Data Science eines schönen Menschen zum Thema Statistik, Datenanalyse und Data Sci

Die Webseite datenwerk.netlify.app stellt eine Übungsaufgaben bereit. Sie können die Suchfunktion nutzen, um die Aufgaben mit den folgenden Namen zu finden:

1. Typ - Fehler - R - 01
2. Typ - Fehler - R - 02
3. Typ - Fehler - R - 03
4. Typ - Fehler - R - 04
5. Typ - Fehler - R - 06a
6. Typ - Fehler - R - 07
7. Typ - Fehler - R - 08 - name - clash
8. Logikprüfung 1
9. Logikprüfung 2
10. there - is - no - package
11. Wert berechnen 2
12. Wert zuweisen _mc
13. argumente
14. import - mtcars
15. Wert zuweisen
16. Wert prüfen
17. wrangle 1
18. repro1 - sessioninfo
19. mw - berechnen

Prüfen Sie Ihr Wissen zu ¹⁰Nominalnominierung? Checken Sie alle Aufgaben mit dem Tag R auf dem Datenwerk.

¹<https://sebastiansauer.github.io/Datenwerk/posts/>
²<https://sebastiansauer.github.io/Datenwerk/#category>

3. Daten einlesen

Hinweis

Die Webseite Datenwerk stellt eine Reihe von Statistik bereit.

Jeder Aufgabe sind im Datenwerk ein oder mehrere Schlagwörter zugeordnet. Wenn Sie auf eines dieser Schlagwörter klicken, werden die Aufgaben mit diesem Schlagwort sortiert. Es kann sein, dass einige Aufgaben nicht gelöst werden können, da Wissen Sie (noch) nicht haben. Lassen Sie sich davon nicht entmutigen! Ignorieren Sie solche Aufgaben fürs Erste.

```
read.csv
```

Eine weitere Möglichkeit, um Daten von einer Webseite im Internet oder auf Ihrem Computer zu erhalten, ist die Funktion `read_csv`:

```
library(easyrstats)
#<data_(präzise)
" https://vincentarelbundock.github.io/Rdcsv/openintro/mariokart.csv"
```

Der Unterschied ist, dass `read_csv` eine Vielzahl an Formaten (XLSX, CSV, SPSS, ...) verarbeitet und angibt, welche er einlesen kann.

Schauen wir uns die letzte R-Syntax im Detail an.

Hey R,

hol das "Buch" easyrstats aus der Bücherei und definiere es als "d" die Tabelle, die du unter der angegebenen URL findest.

3.11. Vertiefung

In R gibt es oft viele Möglichkeiten, ein Ziel zu haben wir hier daher Bveefraudndet, um Daten zu importieren. Andere, gebräuchliche Befehle, die bei Breend (aus dem Standard-R, kein Exetad-Paket cs (aus dem Metta-dPyake) ts e

/ " %

Mit der Funktion `Importieren` kann man viele verschiedene Datenformate importieren, auch Excel-Tabelle

Als Beispiel betrachte [extravers dam & Pakatz](#) {prada} diesen Datensatz werden die Ergebnisse der Umfrage zu den Korrelationen von Extraversionsbedarf zugrundeliegenden Studien finden.¹³ Siehe hier: [hier](#). Laden Sie die Excel-Datei herunter. Angenommen Sie die Excel-Datei in einem Microsoft Word Dokument ein. Dann können Sie die Daten so in einem Projektordner speichern. Dann können Sie die Daten so in einem Projektordner speichern.

```
l i b(r e a r s y s t a t s )  
e x t r d a <- a _ ("r e a t d a / e) x t r a . x l s "
```

Allerdings kann man Dateien aus dem Internet in praktisch wären. Stattdessen muss die Datei liegen.

Wenn Sie allerdings „remote“, also aus dem Internet importieren möchten, sind Import und Synchronisierung nicht möglich:

```
I i b(r a i r y)  
e x t r a _ p a s t e o  
" h t t p s : / / g i t h u b . c o m / s e b a s t i a n s a u e r / s t a t i s t i
```

¹²<https://github.com/sebastiansauer/pradadata>

<https://gitlab.com/sebastiansauder/produdatu>
Ein Daten-Dictionary findet sich hier: <https://gitstatistik1/raw/main/data/extradictonary.md>.

3. Daten einlesen

```
"raw/main/data/extraxls"  
extraimportpath)
```

CSV-Dateien werden auf vielen Computern als Excel öffnen kann und das auch tut, wenn man den Link klickt. Dennoch ist das CSV-Format kein spezielles Dateiformat, sondern eine einfache Text-Datei, die auch mit einem Texteditor geöffnet und bearbeitet werden kann. Alternativ kann man sie auch mit Excel oder anderen Importierern.

• • •

ž

In Definition 3.5 hatten wir Vektoren definiert, sozusagen freivom Rahmen (Schauen Sie mal dort). Die Spalten einer Tabelle sind aber auch Vektoren. Um einen Vektor direkt ansprechen, so kann man die Operatoren anwenden. Angenommen, Sie möchten sich den Vektor `totalausgaben` der Tabelle ausziehen, dann könne man dies mit dem Dollar-Operator tun:

```
mari@katastrophal:~  
## [1] 523746447145
```

Der Dollar-Operator trennt den Namen der Spalte.

Natürlich können Sie mit dem resultierenden Vektor weiterarbeiten, etwa indem Sie ihn in einanderen Vektoren spezifizieren:

```
verkaufsprei$teo<-matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10), nrow = 5, ncol = 2)  
mea(verkaufspreise)  
mea(mari$lokat#_spyrnonyzmur obigen Zeile)
```

```
# # [ 1 ] 50
# # [ 1 ] 50
```

.

Das Kombinieren von Funktionen kann kompliziert sein.

Listing 3.8 SchachtelteFunktionen

```
x <c(1,2,3)
sum(a+b(means(x)))
# # [ 1 ] 2
```

Die Funktion `sum` gibt den (Absolut)zusummenfänger an und für jeden Wert das Vorzeichen.

Verschachtelte Ausdrücke lesen sich von innen nach außen (in dieser Reihenfolge abgearbeitet). Für uns ergibt sich:

1. Berechne den Mittelwert von
2. Ziehe vom Mittelwert jeweils die Elemente ab
3. Nimm vom Ergebnis jeweils den Absolutwert
4. Summiere diese Werte

Kurz gesagt: Hier haben wir die mittlere Abstande zu dem Mittelwert ausgerechnet.

.

Irgendwann werden wir mit unserer Version von R zufrieden sein. Installieren Sie dann einfach die neue Version wie oben beschrieben, s. Kapitel 3.3.

So updaten Sie Ihre R-Pakete (z.B. `RS` im Repertoire) auf die neueste Version und die Anzahl der zu aktualisierenden Pakete erhält.

3. Dateneinlesen

ist, dann besser nicht alle auswählen, son
nächsten paar Pakete usw. Denken Sie daran,
RStudio, R-Paket), die Sie updaten/install
darf.

Sie benötigen in diesem Kapitel¹ die Daten
online oder über R eingespielt werden kann.

Import via Download:

```
mario_kraerat@osx:0
" https://vincentarelbundock.github.io/Rd
" csv/openintro/mariokart.csv"
```

Import via R-Paket:

```
# Das Paket 'openintro' muss installiert sein
da(mario_kraerat@gopen) ntrro"
```

7

„Warum R? Warum, R?“ heißt ein Kapitel in Zum Round Contravon Rausführt. Kapitel 3 hält Hinweise zum Starten von R. Kapitel 4 von „Errisch“. Kapitel 5 führt in die Daten anspruchsvoller). Alternativ bietet Kapitel einen guten und anwenderfreundlichen Über den Vorteil, dass es komplett frei online

¹Über diese Internetadresse: <https://vincentarelbundock.github.io/Rd/csv/openintro/mariokart.csv>

3.12. Literaturhinwe

dazu ist Cetinkaya-Rundel & Hardin (2021), vieler; auf jeden Fall genau das richtige Niveau angewandten nicht-technischen Studiengänger

◦

"

Ł

- Sie können folgende Verben deaktivieren:
f i tseerl, s a t m a r g i r s o e u p m u b t y a t e
- Sie können R-Befehle mit der „Pfeife“ verwenden:

Übungsaufgabe klar! In diesem Kapitel 3.7.3. probieren Sie alle des Datenjudos aus, zunächst, indem Sie einen und selber ausführen.



Erkläre einem Fünfjährigen, warum Statistik wichtig ist.

👉 Okay, hör zu. Stell dir vor, du hast viele und du willst herausfinden, welches Auto spielt wird. Du kannst jeden Tag zählen, wie oft du es benutzt hast. Aber das wäre sehr mühsam! Schnell durcheinander gebracht werden. Deshalb gibt es eine einfache Methode zum Beispiel eine Liste machen und jede

4.1. Einstieg

bestimmen Farben markieren. Dann kannst du jachzählen, wie oft du jede Farbe verwendet hast und du herausfindest, welches Auto am meisten wurde, indem du siehst, welche Farbe am meisten gewählt wurde. Das ist ein Beispiel dafür, wie Stellenkann, Muster in großen Datensätzen Entscheidungen zu treffen, ohne alles einz

Definitionen und Methoden Prozess des Aufbereitens, Umformens oder Zusammenfassens für einzelne Beobachtungen (Zeilen einer Datei (Spalten einer Datentabelle) oder einer ganzen

! . \$

Denkt man an Data Science, stellt man sich wohl Francisco oder Berlin), die an abgefahrenen und komplexen statistischen Modellen für gigantische Daten. Laut Harvard Business Review, verbringen Wissenschaftler „80%“ ihrer Zeit damit (Bowne-Anderson, 2018). Ja: mit uncool en Tätigkeiten wie daten entfernen oder die Daten überhaupt nicht zu machen.

Das zeigt zumindest, dass das Aufbereiten von Daten nicht so einfach ist. Und es zeigt auch, dass man alleine damit schon weit kommen kann. Einricht ist (vielleicht), dass das Aufbereiten von Mathematik verlangt, statt dessen muss man einige Kniffe kennen. Daher ist es wichtig, dass man gut um das Datenjudo.

Beispiel für typische Tätigkeiten des Datenjudos sind:

¹Quelle: ChatGPT 3.5, 2023-02-09

4. Daten umformen

- Zeilfiltern (z.B. nur Studenten des Studiengangs)
- Zeilen sortieren (z.B. Studenten mit guten Noten zuerst)
- Spalten wählen (z.B. 100 weitere Produkte ausbauen)
- Spalten zusammenfassen (Notenschnitt Klausur)
- Tabellenkupplung (z.B. Analyse getrennt nach Studiengang)
- Werte aus einer Spalte neue Spalten (z.B. Punkte in Prozent - Richtige umrechnen).
- .□

{

Es gibt einen (einfachen) Trick, wie man umfangreiche Analysen eleganter und schneller kriegt.

Der Trick besteht darin, komplexe Operationen in Teilschritte zu zerlegen. (In gewisser Weise ist die Analyse eben darin: die Zerlegung eines Gegenstands.) Man könnte vom „Lego-Prinzip“ sprechen: Der linke Teil von Abbildung 4.1 sieht man ein: Er zerlegt man es aber in seine Einzelteile, so dass geometrische Objekte wie Dreiecke oder Quadratdiagramme entstehen. Damit Sie es selbst machen können, erläutere ich das Prinzip im nächsten Kapitel 3.7.3.

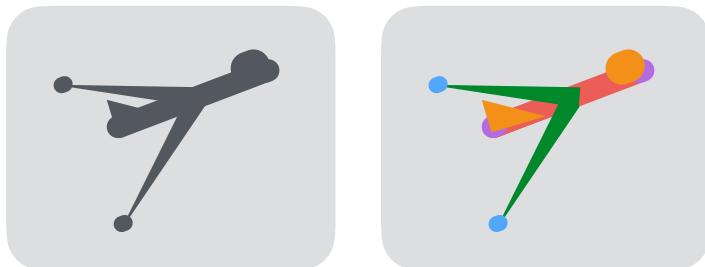


Abbildung 4.1.: Das Lego-Prinzip (S. 10)

4.2. Die Verben des Datenjudo

Beispiel(e) für Datenjudo: Sie können immer noch bei dem großen Online-Auktionshaus. Mittlerweile des „Datenjudo“ erworben. Vielleicht, weil Sie Scienceset zu 80% Datenjudo, das hat irgendwie Naja, jedenfalls müssen Sie jetzt mal zeigen, Sprüche draufhaben, sondern auch die Daten erkennen. Sie analysieren dafür im Folgenden den Na, dann los.

• • •

Im R-Paket das wiederum Teil dieser R-Pakete ist eine Reihe von R-Befehlen, die das Datenjudo facher Verben herunterbrechen (Funktionen). Sie sind nicht installiert haben sollten, wäre jetzt ein wichtigsten Verben des Datenjudo schauen wir uns an.

Wir betrachten dazu im Folgenden einen einfachen Datensatz, an dem wir zunächst die Verben des Datenjudo Tabelle 4.1.

Tabelle 4.1.: Eine einfache Datensatz von

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

Die Verben des Datenjudo werden ebenfalls in Pakete unterteilt, wird, wenn Sie vergessen tidyzussteuern, dann funktionieren diese Befehle.

4. Daten umformen

% arrange

SortideerreZeilen ist eine einfache, aber häufige Judos, s. Abbildung 4.2.

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

arrange →

id	name	gruppe	note
3	Charli	B	1.7
1	Anni	A	2.7
2	Berti	A	2.7

Abbildung 4.2.: Sinnbild für das Sortieren

Beispiele Wa4s. Sind die höchsten Werte in der Note angefangen. Daher stellen Sie sich folgende Frage: Lässt sich die höchste Note in der Spalte auf? Wenn ja, kann die Spalte sortiert werden? (s. Datei `maxtotal.xls`). Excel kann die Spalte, nach belieben sortieren möchte, einfach anklicken und ziehen.

Die Funktion `arrange` macht es uns ziemlich einfach, s. Abbildung 4.2.

```
arrange(mariorokart, total_pr)
```

Tabelle 4.2.: Die Daten im Datensatz `total_pr`, sortiert nach der Note

29	0.99
30	0.01
31	0.99
31	1.99
31	30.00
31	0.01

Übersetzen wir die R-Syntax ins Deutsche:

4.2. Die Verben des Daten

Hey R,
arrangiere (sortiere) `mariokart` nach der Spalte

Garnichtsschwer.

Übrigens war die Voreinstellung auf steigende Werte, man ein Minus vor der zu sortierenden Spalte, absteigend geändert:

```
mario_s oarrtriemter<`-at koatratl ,_ pr )
```

Übungsaufgabe 4.2: Zeigen Sie die Mariokart-Daten an der Anzahl der bei gelagerten Lenkräder.

+ filter

Zeichen bedeutet, dass statt numerische Beobachtungen behalte, die restlichen Zeilen brauchen ihnen. Wir haben also ein Filterkriterium in die Tabelle flern, s. Abbildung 4.3.

The diagram illustrates a filtering operation. On the left, there is a table with four rows and four columns labeled 'id', 'name', 'gruppe', and 'note'. The rows contain the following data: Row 1: id=1, name='Anni', gruppe='A', note=2.7; Row 2: id=2, name='Berti', gruppe='A', note=2.7; Row 3: id=3, name='Charli', gruppe='B', note=1.7. An arrow points from this table to the right, labeled 'filter'. On the right, there is another table with the same structure and data, but only the first two rows are highlighted in yellow, representing the filtered result.

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

filter

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7

Abbildung 4.3.: Sinnbild für das Filtern einer Tabelle

Beispielabbildung Foto für den Verkaufsspannungsunterschied ist es, ob eine Ware einen höheren Verkaufspreis erzielen als

```
mario kart_foto(mario kart, =stypes "photo")
```

4. Daten umformen

Sie filtern also die Bildbearbeitungsaufgaben im Datenbleiben, welche mind. ein Photo haben, mit `(Beobachtungen) stock_pho >= TRUE`

Angestachelt von Ihren Erfolgen möchten Sie diese prüfen: Erzielt eine Stockfotoplattform `Photo` einen höheren Preis als die übrigen Anbieter?

Anders gesagt haben Sie zwei Filtern, die `Photo` auf `stock_pho` erfüllen, möchten Sie näher untersuchen:

```
mario_filter1 <-
  filter(mario[k!="the"] & stock_pho == "new")
```

Hm. Was ist mit den Auskunftsberichten (niedrige) ein Preisvergleich angeschaut, oder beides (Filtern mit ODER)?

```
mario_filter2 <-
  filter(mario[k!="the"] | stock_pho == "new")
```

Zur Erinnerung: Logische Operatoren sind:

Übungsaufgabe 4k: Kann man noch viele interessante Fotos prüfen, denken Sie sich und tun das auch.

Übungsaufgabe 4r4: Sie die Spiele mit nur einer Versandkosten.

Übungsaufgabe 4r5: Sie die Spiele mit nur einem überdurchschnittlichen Verkaufspreis der Fundkommission bei uns zu kaufen und werte einen Datensatz zu erfahren (diese Fausy ist falsch).

4.2. Die Verben des Daten

\$ select

Eine Tabelle mit vielen Spalten kann schnell
Daher lohnt es sich, eine alte goldene Regel zu bea-
so einfach wie möglich, aber nicht studiafacher.
Spalten aus, die uns interessieren und entfernen.
Abbildung 4.4 als Beispiel.

The diagram illustrates a selection query. On the left, there is a table with four rows and four columns labeled 'id', 'name', 'gruppe', and 'note'. The first two rows have their 'name' column highlighted in yellow. A large black arrow points from this table to the right, where a smaller table shows only the 'id' and 'note' columns for the same two rows, indicating the result of the selection.

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

select

id	note
1	2.7
2	2.7

Abbildung 4.4.: Sinnbild für das Auswählen von Spalten

Beispiel(e) auf nur zwei Spalten) auchte Sie deutlich geringere Preise erzielen im Vergleich. Sie entschließen sich, mal ein Stündchen auf den Starren. Dafür wählt sie entweder Spalten aus.

```
mario_select(mariokart, cond, total_pr)
```

Der Befehl `select` wartet als Input eine Tabelle und eine Tabelle zurück - genau wie die meisten anderen Befehle. Auch wenn Sie nur eine Spalte auswählen, eben eine Tabelle mit nur einer Spalte.

`select` achtet Komfort; Sie können Spalten auf mehreren:

```
select(mariokart, 1, 2) # Spalten 1 und 2
select(mariokart, 2, 5) # Spalten 2 bis 5
select(mariokart, *) # Alle Spalten außer 1
```

Übungsaufgabe 4.6 Sie ziehen mit der Wählmaschine die zweite Spalte aus der Tabelle

```
select(mariokart, total_pr, cond, 2)
```

4. Daten umformen

Vertiefte Informationen zum Auseinsvöhlleenn von sich hier.

\$ + summarise

Bei spielt Wa4s. i6st der mittlerer MiVseurnkmaur fss.prei Listing 4. 1, können wir den mittleren Verkauf berechnen (50).

So eine lange Spalte mit Zahlen - mal ehrlich durch? Machen wir uns das Leben leichter, indem wir mit Zahlen zu einer einzigen Zahl zusammenfassen. Studierende - Anni, Berti, Charli - haben einen Durchschnittswert. Die Noten waren 2.7, 2.7 und 1.7. Damit erhalten wir den Mittelwert bei 2.4; s. Abbildung 4. 5.

				summarise	
id	name	gruppe	note		note_mw
1	Anni	A	2.7		
2	Berti	A	2.7		
3	Charli	B	1.7		2.4



Abbildung 4. 5 zeigt den zu einer einzelnen Zahl zusammengefassten Mittelwert.

Fassen wir als Nächstes die Preise zusammen, zwar zum Mittelwert. Dann wissen wir, für wieviel Durchschnitt verkauft wird, s. Listing 4. 1.

$$\begin{array}{r} \text{preis_mw} \\ \hline 50 \end{array}$$

Aha! Etwa 50 € erzielt so eine Auktion im Schuhgeschäft ein bisschen abstrakte Ergebnisse. Schließlich fassen wir (einzeln) zusammen, s. Abbildung 4. 6.

³https://tidyverse.org/reference/tidy_trans.html

4.2. Die Verben des Datenfassens

Liste **Digital Funktionen** summarise fasst einen Vektor zusammen

```
mariokart_mittelneu(mariokart,  
                     preis_mean(total_pr))  
mariokart_mittelwert
```



Abbildung 4maßfasst eine Spalte (oder mehrere) zellen Zahl zusammen

Auf welche **Spaltenmenge** fassst werden soll, z.B. anstatt werts oder Maximalwerts, muss noch zusätzliche **summa** riansgäbe geben werden.

Übungsaufgabe 4.1 Fazieren Sie den höchsten Kauftyp Mariokart-Spiels!

Übungsaufgabe 4.2 Fazieren Sie den Mittelwert der Tempausfälle!

%

Es ist ja gut und schön, zu wissen, was so ein Saisonverlauf interessanter wäre es doch, denken Sie

⁴summarise(mariokart, hoechster_preis = max(total_pr))

⁵summarise(mariokart, mw_versand = mean(total_pr))

4. Daten umformen

neuen Spiele im Schnitt mehr kosten als die ausrechnen kann?

👋 Hallo R, kannst du mir die mittleren Alten und neuen Spielen ausrechnen?

🤖 Ich tue fast alles für dich.

Also gut, R, dann gruppiere die Tabelle, s.

The diagram illustrates the process of grouping data. On the left, there is a table with three rows:

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

In the center, a large black arrow points to the right, indicating the transformation. To the right of the arrow, the grouped data is shown in a second table:

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7

id	name	gruppe	note
3	Charli	B	1.7

Abbildung 4.7.: Gruppierung von Daten

Durch das Gruppieren wird die Tabelle in „Teile der Gruppen“ aufgeteilt. Das sieht man derlich an. Aber alle nachfolgenden Beobachtungen beziehen sich auf eine Gruppe.

Bei spielerischer Preisgruppenberechnung
Ihnen zwei (oder mehrere) Teiltabellen, eine
gebrauchte Spielergruppen Werte
gruppieren anhand einer Variable

```
mario_kart_grouping(mario_kart, cond)
```

Wenn Sie die neue Tabelle betrachten, sehen Sie nur einen Hinweis, dass die Tabelle gruppiert ist. Die Teiltabelle hat weitere Berechnungen zur Berechnung des mittleren Verkaufspreises.

4.2. Die Verben des Datenj

```
summa(maeiokartgruppierungspreisfotototal_pr))
```

cond	preis_mw
new	54
used	47

Ah, die neuen Spiele sind teurer, wer hätt' s gesagt? Sie sich wie ein Datenchecker ...

Übungsaufgabe 4.9.

Aufgabe: Berechnen Sie den mittleren und maximalen Preis getrennt für Spiele mit und ohne Foto!

Lösung

```
mariokart_gruppierungspreisfotototal_photo)
```

```
mariokart_verkaufspreis_foto <-
  summa(maeiokart_gruppiert_foto,
         total_pme_avg(total_pr),
         total_pma(total_pr))
```

```
mariokart_verkaufspreis_foto
```

stock_photo	total_pr_avg	total_pr_m
no	54	327
yes	48	75

Bei Auktionen mit Foto wird im Schnitt ein höherer Preis bezahlt.

4. Daten umformen

\$ mutate

Immer mal wieder ~~Spaltenmaßnahmen~~ werden Werte umrechnen, s. Abbildung 4.8.

id	name	gruppe	note	
1	Anni	A	2.7	
2	Berti	A	2.7	
3	Charli	B	1.7	

id	name	gruppe	note	punkte
1	Anni	A	2.7	73
2	Berti	A	2.7	72
3	Charli	B	1.7	89

Abbildung 4.8.: Spaltenveränderungen/neubilden

Beispiel 4 Hersteller des Mario Kart in Japan; daher erscheint es Ihnen opportunum mit dem Hersteller die Verkaufspreise von umzurechnen. Nach etwas Googlenfnden Sie eine von 1: 133.

```
mario_kart_yen <-  
  mutate(mario_kart_yen, total_pr_yen = 1/133 * pr)  
mario_kart_yen #<data-frame: mario_kart_yen, total_pr_yen, total_pr>  
mario_kart_yen #<data-frame: mario_kart_yen, total_pr_yen, total_pr>
```

total_pr_yen	total_pr
6856	52
4926	37
6052	46
5852	44
9443	71
5985	45

Sicherlich werden Sie Ihre Geschäftspartner mit abrechnen Sie (in einer Tabelle) neu.

Die Dimensionen werden für jede Zeile verwendet.

4.2. Die Verben des Daten

Beispiele für Funktionen, die sich, z. B. um Spalten zu addieren, zu multiplizieren, formieren (z. B. den Logarithmus anwenden oder Spalte von jeder Zeile abziehen).

+ count

Arbeitet man mit nominal skalierten Daten, ist kann, die entsprechenden Zeilen der Tabelle zu fragen, wie viele neue und wie viele alte Spiele mariokart vorhanden sind.

Beispiel Nach der letzten Präsentation Ihrer Chefngestöhnt: „Oh nein, alles so kompliziert! Kann man das nicht einfacher machen?“ Anstelle komplizierten Berechnungen (Mittelwert?) möchten Treffen nur zeigen, wie viele Computerspiele gebraucht sind (in Ihrem Datensatz). Schließlich ist Ihre Chefnnicht wieder überfordert.

```
mariocart_count(mariocart, cond)  
mariocart_counted
```

cond	n
new	59
used	84

Aha! Es gibt mehr gebrauchte als neue Spiele.

Jetzt könnte man die Grenzen: Welcher Anteil der 143 Spiele ist neu, welcher gebraucht?

4. Daten umformen

```
mut(a(mariokart, cond))
```

cond	n	Anteil
new	59	0.41
used	84	0.59

Übungsaufgabe 14en Sie, wie viele Auktionen haben.

Übungsaufgabe 14en Sie, wie viele Auktionen innerhalb der gebrauchten Spiele sind. Anders gesagt: Teilen Sie den Datensatz sowohl nach Foto auf und zählen Sie jeweils, wie viele jeweilige Gruppen gehören.

Die Befehle („Verben“) des tidyverse sind jedoch aufgaben des Datenaufbereitens („Datenjungen“) wie sie erwarten diese Befehle sein werden. Tabellen aus Output zurück, s. Abbildung 4.9. werden beim „tidydatatutor“ anschaulich dargestellt!

Das ist keine Pfeife, wie René Magritte 1927 schrieb, s. Abbildung 4.10.

⁶count(mariokart, stock_photo)

⁷count(mariokart, stock_photo, cond)

⁸<https://tidydatatutor.com>



A b b i l d u n g d i y v 9 e r s e - B e f e h l e e r w a r t e n n o r m a l e r w
(t i b b l e) a l s I n p u t u n d g e b e n a u c h e i n e T
O u t p u t



A b b i l d u n S g o 4 s i 1 e 0 h t d i e P f e i f e i n R a u s (J a j a , d a
s o n d e r n e i n S y m b o l e i n e r B P E ð h ð e ...) . L i n
P f e i f e (M 7 , 2 0 0 4) . M i t t e u n d R e c h t s : D i
f ü r e i n e „ P f e i f e ” (p i p e) .

4. Daten umformen

!

Computerbefehle und im Speziellen R-Befehl - oder viel mehr: ineinander-stapeln, so eine Puppe (vgl. Kapitel 3.6.3). Schauen wir uns das an. Dazu definieren wir zuerst die Variable `x`:

```
x <- c(1, 2, 3)
```

Und dann kommt unser verschachtelter Befehl:

```
sum(mxmean(x))  
# [1] 0
```

Wie schon erwähnt, arbeitet R so ein wenig, versteckt in den nachstehen-

den Zeilen:

```
sum(x - mean(x))
```

```
sum(x - 2)
```

```
sum((x - 2) ^ 2)
```

Schritt für Schritt. Ganz schön kompliziert!

Soweit kann man noch einigermaßen folgen. Kann man noch extremer machen, dann wird's mal folgende (Pseudo-) Syntax an:

F Vorsicht

Eine beliebter Fehler ist es übrigens, nach schließenden Klammern häufig (nicht übersichtlich) mehrere FALSCHEN ZAHLEN KLAFFERN.

Listing 4.1 zeigt die verschachtelte Sequenz von R-Befehlen:

```
fasse_zusammen
gruppieren
wähle_sepa(ten
filter(meinhedaten))))
```

◦ ! #

Listing 4.2 ist schon harter Tobak, was für einen einfacheren, man könnte Listing 4.2 wie folgt schreiben:

```
Nimm"meine_daten" * und dann *
filter gewünschte Zeilen * und dann *
wähle gewünschte Spalten * und dann *
teile in Subgruppen * und dann *
fasse sie zusammen.
```

Definition (oPh4ej) ~~dann~~ dann "heißt ~~andere~~ ~~Ergebnis~~ anonym
| .> Man nennt diesen Befehl „Pfeife“ (engl.

I Hinweis

Der Befehl kann empfohlen werden. Der Shortcut für diesen Befehl ist Strg - Shift - M. Wählen Pfeife im Flyoutmenü.

Mittlerweile (Seit R 4.1) ist auch im Standard-Syntaxeditor eingebaute Pfeife funktioniert anders, aber den Vorteil des Befehls mussen. Da der Befehl sowieso praktisch immer stark bringt es uns keinen Vorteil, die neuere Pfeife verwenden. Aber auch kein Nachteil. Optimal.

⁹Genauer gesagt wird ein rechteckiges Objekt erstellt. Also nicht um das Sie sich kümmern müssten.

4. Daten umformen

können Sie einstellen, dass der Shortcut **S** Pfeife verwendet.



Abbildung 4.10 Illustration für eine Pfeifensequenz wie am Fließband.

Listing 4.4e Pfeifen-Befehlssequenz (Pseudocode)

```
meine_Waagen
filter_ge_wünschte
wähle_ge_wünschte
gruppiere
fasse_zusammen
```

Und jetzt kommt's: So eine Art von Befehls-Schauensie sich mal Listing 4.3 an. So ein ist ein wie ein Fließband, an dem es mehrere Stationen weitergereicht und an jeder Stelle teilt Ihre „Pfeifen-Sequenz“ für den Mario Kart aus Listing 4.4.

cond	total_pr_mean
new	54
used	43

Die Syntax (mariokart und die Sprungtafel) > filter(total_in�ideen) schd | A) Füge meixner:

4.4. Beispiele für Forschung

Liste Mariokart am Fließband: Die „Pfeifen-Sy“

```
# Hey R, nimm die Tabelle "mariokart":  
mariokart  
# filter nur die günstigen Spiele:  
filter(stock_percent <= 10)  
# wähle die zweien Spalten:  
select(second, total_price)  
# gruppieren die Tabellen nach Zustand des Spiels:  
group_by()  
# fasse bei den Gruppen nach dem mittleren Preis zu  
summa(price_mean(mtotal_price))
```

Übungsaufgabe 4. Siehe die Lösungen der folgenden Lesen, versuchen Sie die Aufgaben selbst zu lernen, nicht gleich auf die Lösungen zuschauen.

Sie arbeiten als strategischer Assistent der Kunden für Faktenchecks und andere Daten-Aufgaben. Sie sollen Ihnen zeigen, was Sie können (Schluck).

Übungsaufgabe 4. Welches ist das erste Spiel?

💡 Ich würde von Ihnen gerne wissen, was das ist, aber jeweils für neue und gebrauchte Spiele, die mit Foto verkauft wurden!

Lösung

```
mariokart  
filter(stock_percent <= 10)  
group_by()  
summa(price_mean(mtotal_price))
```

4. Daten umformen

cond	total	_pr_max
new	75	
used	62	

Die Funktion `max` liefert den größten Wert eines Vektors.

```
x <- c(1, 2, 10)
max(x)
# [1] 10
```

Das teuerste Spiel ist 10 Euro wenn es gebraucht ist.

Übungsaufgabe 4 mit einer Versandpauschale

Ich würde gerne die mittlere Versandpauschale getrennt nach Anzahl der Lenkräder beigefügt sind. Und ich will nur Gruppen aus mindestens 10 Spielen bestehen!

Lösung

Wenn wir die Anzahl der Spiele zählen in Abhängigkeit von Lenkrädern kommen wir eine Tabelle mit wheelsanzahl, wie viele Spiele (Zeilen) in „Teiltabellen“ davon

```
matrix(wheels)
count(wheels)
```

wheels	n
0	37
1	52
2	51

4.4. Beispiele für Forschung

wheels	n
3	2
4	1

Aus dieser Tabelle sehen wir, dass 3 oder 4 Lenzen bzw. 1 Mal) bei gelegt wurden und wir solche Spiegel bevor wir den Mittelwert der Versandkosten ausrechnen.

marie okamoto
 f1 (wheels < 3) %>%
 group_by(wheels %>%)
 summa(mittelere_verseandkosten ~ .) =,
 anzahl_s(p)ielle =

wheels	mittlere_verseandkosten	anzahl
0	2.7	37
1	3.6	52
2	2.9	51

Die Funktion bindet die Anzahl der Zeilen pro Teiltafel.

Die mittleren Versandkosten bewegen sich also zwischen 2.7 und 3.6 Dollar, je nach Anzahl der bei gelegten Lenzen.

Übungsaufgabe (Von 4.000 Stück spr. Preis in Yen?)

Ich würde gerne den Verkaufspreis in Yen weiterverhandeln. Dann rechne mal den mittleren Verkaufspreis und ziehe 10% ab, die wir als Provision unserer Zahlen müssen.

Lösung

4. Daten umformen

mar i o%>a% t	
sel (etc dt d%>_%p r)	
mut ftoet al _ ptro_tyad1n3)3%>	
summa f i se	
preis_yem_a(mtwo=t al _ pr _yen) ,	
preis_yen_mw_ mpneis 1 0 0 n*0 m ovi s_yen_mw)	
<hr/>	
preis_yen_mw	preis_yen_mw_mi nus_1
6 6 3 4	5 9 7 1
<hr/>	

Wie man sieht kann man nämlich mehr als eine Berechnung einstellen. In diesem Fall haben wir zweimal den Mittelwert und einmal den Mittelwertswerts).

Übungsaufgabe 14t. (Kostenkosten) Sie sich selber als Forschungsfragen aus. Stellen Sie diese einmal in Minuten bzw. einem vertrauenswürdigen Kommentar. Sie, ob Ihre Aufgabe richtig gelöst wird. Praktisch!

!

Die Covid19-Epidemie hatte weltweit massive psychologische Auswirkungen auf die Menschen. Lukomet al. (2020) berichteten eine Studie über die Auswirkungen untersucht; die Studie ist bei der Open Science Foundation (OSF), <https://osf.io/4b9p2/> die R-Syntax zusehen. Die Daten wurden mit R ausgewertet und bereit für die Analyse vorbereitet. Eine gute Analyse ist in diesem Kapitel beschrieben, ein schönes Beispiel, wie Forschung und Praxis zusammenführen können: Angewandte Forschung als Beitrag zum Problems, der Corona-Pandemie.

4.6. Wiemann mit Statistik

(\$)

Ein (leider) immer mal wieder zu beobachtend
zu frisieren ist, nur die Daten zu berichten,
passen.

Beispiel 4: Ein Analytiker möchte zeigen, dass der Verkauf von Mario Kart-Spielen „viel zu niedrig“ ist. Er kommt zu dem Ergebnis, dass die Analystin den Preis von 50 Euro zu niedrig schätzt.

👉 Kann man den Wert nicht „kreativ verbessern“?
Statistik-Tricks anwenden?

Um dieses Ziel zu erreichen, teilt die Analysten nach Anzahl der dem Spiel ~~wie~~ ^{an} ~~gle~~ ^{ge} Dagen Len wird der Mittelwert pro Gruppe berechnet.

```
m a r i o k a r t _ w h e e l s <-
m a r i o%>a% t
g r o u p(w b y %> %)
s u m m a(p i s_e_m e a n t_o t a l _ p r ) ,
c o u n t n( n#)` n` g i b t d i e A n z a h l d e r Z e i l e n p r o G r u
```

whe e l s	p r _ m e a n	c o u n t _ n
0	4 1	3 7
1	4 4	5 2
2	6 1	5 1
3	7 0	2
4	6 5	1

Schließlich berechnet **ungewichtete Mittel** die Werte über diese 5 Gruppen:

4. Daten umformen

```
mario kar %> wheels  
summa(mesa(np_r_mean))
```

$$\frac{\text{mean(pr_mean)}}{56}$$

Und das Ergebnis lautet: 56 Euro! Das ist doch 50 Euro.

Natürlich ist dies irreführend, hier einen ungewichteten Mittelwert zu berechnen. Der gewichtete Mittel korrekten Ergebnis \square 50 Euro, führen.

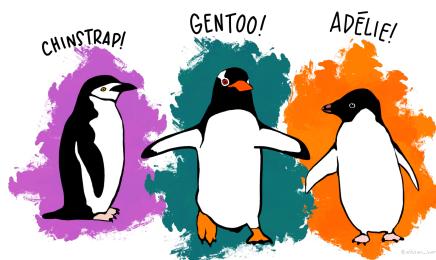


Abbildung 4.12.: Possierlich: Die Pinguine

Übungsaufgabe 1: Sie sich zunächst mit dem Datensatz vertraut. Sie sind geübt in der Praktikation von palmer penguins (siehe auf gewohnte Art in Kapitel 3.4.2); im Internet findet man den Datensatz. Fokussieren Sie Ihre Analyse auf die Zielvariable

Forschungsfragen:

1. Was ist das mediane Gewicht von Pinguinen, gesiezt und nach Gewicht?
2. Wie viele Pinguine gibt es pro Spezies?
3. Wieviel wiegt der schwerste und der leichteste Zies?

· ž ' / ° \$ % # " \$ \$



Abbildung 4.13.: Studie COVIDISTRESS (Lie

Lesen Sie die Beschreibung der Studie COVIDISTRESS 2022). Hier ist ein Abstract:

The COVIDISTRESS global survey is an international collaborative undertaking for data gathering on behaviors, behavior and attitudes during the COVID-19 pandemic. In particular, the survey focuses on psychological compliance with behavioral guidelines to prevent the spread of Coronavirus, and trusting governmental institutions and their preventive measures, but multiple fear scales are included for descriptive statistics and comparative mapping between participants. Round one data collection was concluded in March 2020. To gather comparable data swiftly from across societies and individuals, the collaboration was constructed as an urgent collaborative project involving contributors and groups in the COVIDISTRESS network.

4. Daten umformen

below) conducted translations to each language. Online links by their own best means in each language.

Die Daten stehen unter <https://osf.io/z39uq/>. Sie können diese echten Dateneigenständig anfordern.

L Chat GPT

Nutzen Sie einen Chat-Bot wie ChatGPT, um R-Syntax geben zu lassen.

Die Webseite datenwerk.netlify.app stellt Übungsaufgaben bereit. Sie können die Suche nutzen, um die Aufgaben mit den folgenden Namen zu finden:

1. wrangle3
2. wrangle4
3. wrangle5
4. wrangle7
5. wrangle9
6. wrangle10
7. tidydata1
8. affairs-dplyr
9. dplyr-uebersetzen
10. haefgkeit01
11. mariokart-mean1
12. mariokart-mean2
13. mariokart-mean3
14. mariokart-mean4
15. mariokart-max1
16. mariokart-max2
17. filter01
18. affairs-dplyr

19. summarise01
 20. summarise02
 21. mutate01
 22. wrangle3

"

#

| Hinweis

In weiterführendem Material werden Sie immer halte treffen, die Sie noch nicht kennen, die Unterricht behandelt wurden. Seien Sie unbesorgt, können Sie diese Inhalte einfach auslassen, zu verlieren. Einfach ignorieren.

Häufig ist es nützlich, die Werte einer Variable „weiblich“ in „Female“ umzuwandeln. Möglichkeit, dies imzen, bietet der `mutate`-Befehl. Befehl wohnt¹ in `tidyverse`. Datenwerk finden Sie dazu Übungen, etwa `mutate`.

" "

R will nicht, so wie Sie wollen? Sie haben das störrisch den Dienst, vermutlich reinaus Boshärgern? Ausführliches Googeln und ChatGPT befriedigt? Kurz, Sie brauchen die Hilfe eines Sies sollten Ihren Hilfeschrei so artikulieren, sondern auch verstanden wird und einen anderen undermöglichst Ihnen zu helfen.

¹https://www.statology.org/dplyr-case_when/

4. Daten umformen

Also: Sie müssen Ihr Problem nachvollziehbar machen. Das nennt man **Reproducibility**, reproduzierbares Problem mit (R-)Syntax:

- einfache einfachste Syntax, die Ihr Problemstellung produziert. Es bietet sich an, einen bekannten Datensatz zu verwenden, etwa
- reproduzierbar (z.B. als Textdatei oder indem die Fehlermeldung entstehen lässt)

Beispiel: Bei Spiel für Preibill Einbierer steht es nicht, eine Fehlermeldung kommt.

Ziel: Ich möchte die Automatik autos filtern (am besten)

Was ich schon verschub habe hablegende Post gelöst ohne Erfolg.

Erbie

```
data(mtcars)
library(tidyverse), denn "dplyr" reicht
mtcars %>%
  filter(# den kürzesten Code, der Ihren Fehler
sessionInfo()# gibt Infos zur R-Version etc. aus
```

Mit dem Präfix %>% kann sich R-Syntax schön formatieren lassen. Das ist perfekt, um den Code in einer Mail einzustellen. Dafür müssen Sie nur die Tastenkombination drücken und drücken und drücken: auf : aus für **ctrl+M** und **ctrl+L**. Dann können Sie die schön formatierte Syntax (sowie die Ausgabe) dann irgendwo hinstellen.

L T i p p

Posten Sie Ihr Erbe bei <http://erbe.rist.github.com/publicgist>. Hier ist ein Beispiel.



Er

Sauer (2019), Kap. 7, gibt eine Einführung in (mit Hilfe von R), ähnlich zu den Inhalten die Tiefe des „Datenjudo“ führen; der Autor Hadley Community sehr bekannt. Er ist einer der Hauptautoren des R-Pakets `dplyr` (Wickham & Grolemund (2018) behandelt (etwas ausführlicher) die Themen di

% / /

~

{

◦

"

Ł

- Sie können erläutern, wann und wo zu das V i s c h e r I n h a l t e s i n n v o l l i s t .
- Sie kennen typische Arten von Datendi agra n
- Sie kennen typische Datendi agramme mit R v
- Sie können zentrale Ergebnisse aus Datendi s e n .

-

!

-

Neben den üblich e d y P u a c k r d s e s y n b a n t i g e n S i e i n
d i e s e m K a p i D t a e t l a f o p h d r o e p t g o p u a b o g s t a t s p l o t

```
l i b(r t a i r d y v e r s e )  
l i b(r e a s y s t a t s )  
l i b(r D a r t a E x p l o r e n t )v e r g e s s e n z u i n s t a l l i e r e n  
l i b(r g g p u # b o r p t i o n a l , D a t e n v i s u a l i s i e r u n g  
l i b(r g g s y t a # s o p l t o i t o n a l , D a t e n v i s u a l i s i e r u n g
```

W i r a r b e i t e n w i e d e r m i n d e r n a K a p i s t a k 3 . 7 . 3 .

5. Datenverbildlichen

(

„Wir müssen die Galaxis retten, Kermit!“

„Schlock“

“ . . . ”

(

Es heißt, ein Bild sage mehr als 1000 Worte. Dinosaurier sagt auch mehr als 1000 Worte, dagegen 5.1 sieht man verschiedene „Bilder“, Dinos und einmal einen Kreis. Obwohl die Bilder sind die zentralen statistischen Kennwerte, selbe Bresche schlägt „Anscombe Quartett“ vier Datensätze, in denen die zentralen Statistiken also Mittelwerte, Streuungen, Korrelationen und Abweichungen sind grundverschieden. Anscombes Beispiel visualisiert enthüllt, was der Statistik: Statistische Diagramme können Einblicke geben, grundlegenden Statistiken (Kennwerten) im Cortex ist sehr leistungsfähig. Wir können zahlreiche Informationen aufnehmen und parallel Grund sind Datendiagramme eine effektive und Erkenntnisse zu ziehen. Nutzen Sie Datendiagramme sind einfach zu verstehen und doch sehr mächtig.

Definition: Ein Diagramm ist ein Diagramm, das Daten und Statistik Erkenntnisse daraus zu ziehen.

Beispiel(eAuswertung der Forschung: Ein aufwändiges Datendiagramm)

Auf Basis des Korruptionsindex von Transparency International Wilke (2019/2024) erstellt ein Diagramm

5. 2. Ein Dino sagt mehr als 100

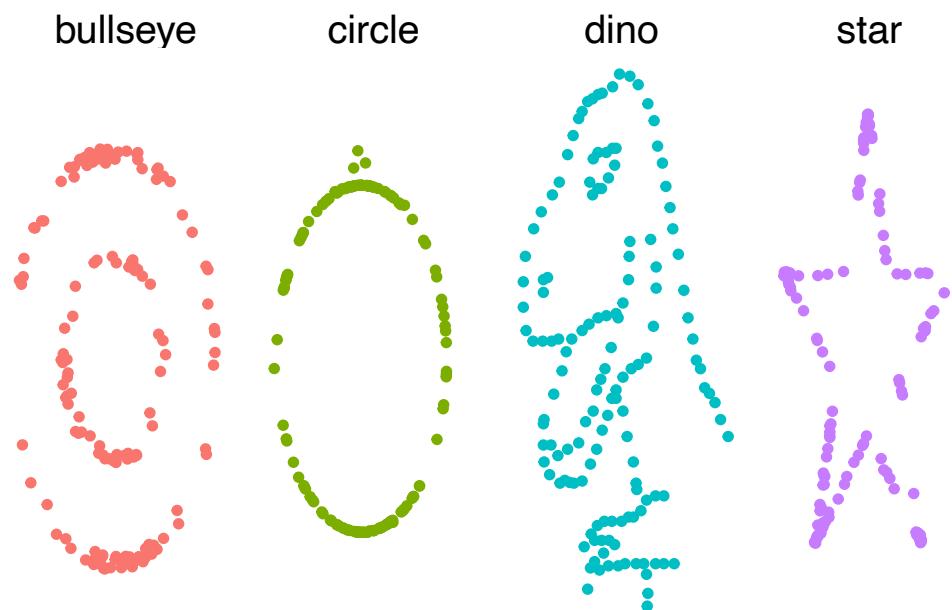


Abbildung 5.15e. Diagramme haben gleiche statistische wie Mittelwert und Streuung und Korrelationen. Datengrundlagen sind komplett verschieden.

5. Datenverbildlichen

Entwicklungsindex (Lebenserwartung, Bill et al. (2015)) und Korruption, jeweils auf L

Es finden sich in der Literatur (im Internet handwerklich meisterhaft erstellt) Datendi mit erstellt werden (vgl. Scherer et al.,

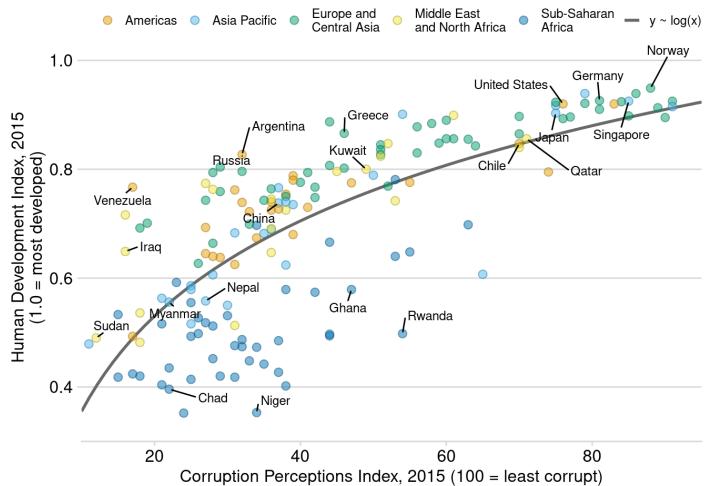


Abbildung 5.3 zeigt ein Bild mit mehreren (nur „Dimension“ entsprechen. Wie man (nicht) unübersichtlich. Offenbar kann man in einem Variablen sinnvoll reinquetschen. Die „Dimension“ hat ihre Grenzen, vielleicht bei vier bis den Zusammenhang von vielen Variablen, z. hen, kommen wir mit Bildern nicht weiter. Da Werkzeuge: Statistik, komm zu Hilfe. Bei kleinen Variablen braucht man keine (aufwändige) Datendiagramm, ist dann oft ausreichendes Statistik nur deshalb gibt, weil unser A sechs Variablen nicht gleichzeitig umgehe

5.3. Nomenklatur von Datendiagrammen

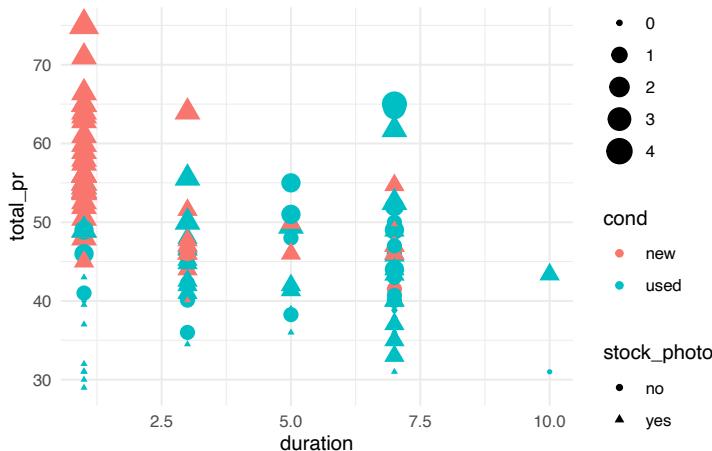


Abbildung 5.5: Das Diagramm kann nur eine begrenzte Anzahl von Daten zeigen. Wenn Sie dieses Bild nicht genau darstellen soll das Bild zeigen.

Übungsaufgabe 5.1: Welche Variablen sind in Abbildung 5.1 dargestellt?

Ž

Tabelle 5.1 zeigt eine - sehr kurze Nomenklatur. Weitere Nomenklaturen sind möglich, aber wir einfache. Wer an Vertiefung interessiert ist, für einen Überblick über verschiedene Typen an Diagrammen einer systematischen Nomenklatur: <https://www>

5. Datenverbildung

Tabelle 5.1.: Ein (sehr kurze) Nomenklatur		
Erkenntnisziel	qualitativ	quantitativ
Verteilung	Balkendiagramm	Histogramm
Zusammengefülltes hang	Balkendiagramm	Dichtediagramm Streudiagramm
Unterschied	gefüllt Balkendiagramm	Boxplot

Definition (Wörterbuch): Häufigkeitsverteilung beschreibt die Häufigkeit der Ausprägungen, die in einem Datensatz vorkommen.

Beispiel (Abbildung 5.2): Es zeigt die Häufigkeitsverteilung, also der Zustand des Artikels, neuer oder gebrauchter. Die Variable hat 5 Ausprägungen: neu, gebraucht, mal vor.

Tabelle 5.2: Häufigkeitsverteilung des Datensatzes mario kart

cond	n
new	59
used	84

Zugegeben, das Datei mario kart im Vorsprung aufregt. Abbildung 5.4. Wie man sieht, besteht es aus einer

5.4. Verteilungen verbil

daher heißt es das Diagramm: Säulendiagramm). Man kann es ein Diagramm um 90° drehen; keine Ausrichtung ist besser als die andere.

Definition (horizontal bar chart): Ein Balkendiagramm ist eine graphische Darstellung von Werten, zumeist für die Häufigkeiten von Kategorien, also Ausprägungen nominaler Variablen. Es verwendet eckige Balken, und die Länge eines Balkens ist zur dargestellten Häufigkeit.

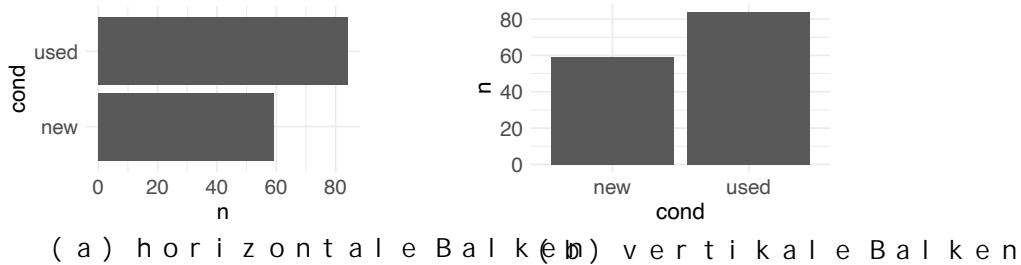


Abbildung 5.4.: Häufigkeitsverteilung der

Es gibt viele Methoden, sich mit R einzufinden. Eine einfache, komfortable und praktische ist die Abbildung 5.4; wir betrachten gleich die Syntax. Wir rufen die Daten, s. Listing 3.2. Außerdem müssen wir den Paket 'data.table' installiert haben, bevor Sie es nutzen werden. Listing 5.1 zeigt die Syntax, um es erstellen. Auf der Hilfeseite der Funktion `barplot` sind die Funktionen erläutert.

5. Datenverbildung

Listing 5.1 Syntax zur Erstellung eines Balkendiagramms

```
library(Dartray Explorer)
```

```
matrix %>%  
  select(cond)  
  plot_(bar)
```

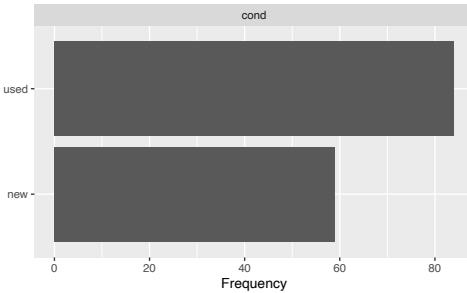


Abbildung 5.5.: Ein Balkendiagramm.

Die Syntax ist in Listing 5.1 abgedruckt. Zu die „Pfeife und lässt sich als“ und dann“ übersetzen wir die Syntax ins Deutsche:

Nimm den Datensatz `mariokart` * und dann * wähle die Spalte cond * und dann * zeichne ein Balkendiagramm. Fertig!

Übungsaufgabe 15: In welchen Spalten wählen für das Balkendiagramm wir andere Spalten ausgewählt, so würde das die Verteilung jener Variablen anzeigen. Ja, Sie können einmal auswählen. Probieren Sie das doch.

Übungsaufgabe 16: Basieren Sie die entsprechende Erstellen Sie ein geeignetes Diagramm, um die Verteilung von cond (Datenmaut) darzustellen.

Lösung

```
m a r i o | k a r t
s e l ( e s c t o c k _ | p h o t o )
p l o t ( _ ) b a r
```

Mit plot kann man Balkendiagramme darstellen.

Bei einer quantitativen Variablen mit vielen Ausprägungen ist ein Balkendiagramm nicht so aussagekräftig, s. Abb. 5.6.

Die Lösung: Wir reduzieren die Anzahl der Ausprägungen auf ganze Dollar runden. Oder, um noch weniger Ausprägungen zu erhalten, können wir einfache Gruppen definieren:

- Gruppe 1: 0 - 5 Dollar
- Gruppe 2: 6 - 10 Dollar
- Gruppe 3: 11 - 15 Dollar
- ...

In Abbildung 5.6 (rechts) sind z.B. die Ausprägungen (total) in Gruppen der Breite von 5 Dollar aufgetragen. Die Balken sind noch die einzelnen Werte als schwarz

Definitiōn: Histogramm istogramm ist ein Diagramm zur Darstellung der Häufigkeitsverteilung einer quantitativen Variablen. Die Daten werden in Gruppen (Klassen) eingeteilt und die Häufigkeit der Daten in dieser Gruppe (in diesem Klassenzahlbereich) wird durch einen Balken (pro Klasse) dargestellt werden. Die Höhe des Balkens entspricht der Häufigkeit der Daten in dieser Klasse (in diesem Klassenzahlbereich).

5. Datenverbildung

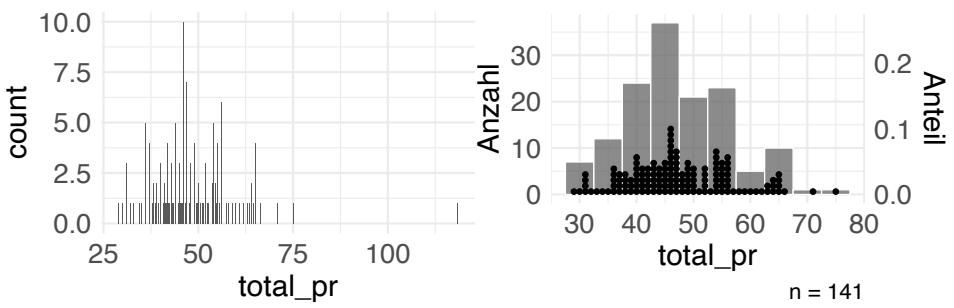


Abbildung 5.7 Balkendiagramm vs. Histogramm für (total_pr)

Es gibt keine klare Regel, in wie viele Balken sein sollte. Nur: Es sollten werden sehr viele Balken (links) bzw. ein Histogramm für (total_pr) können Sie die Syntax Listing 5.2 rechts.

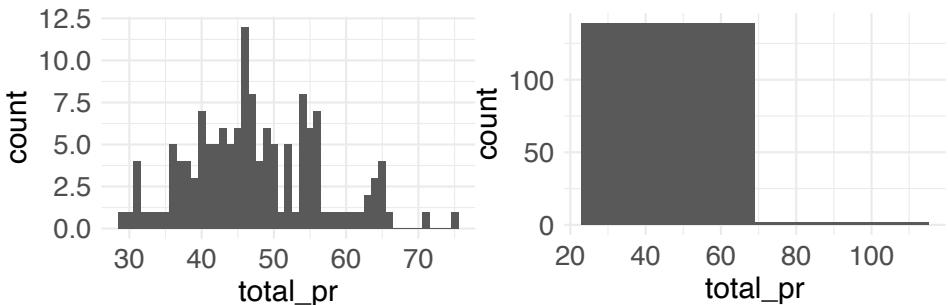


Abbildung 5.8 Histogramme mit zu wenigen und nicht zu vielen Balken

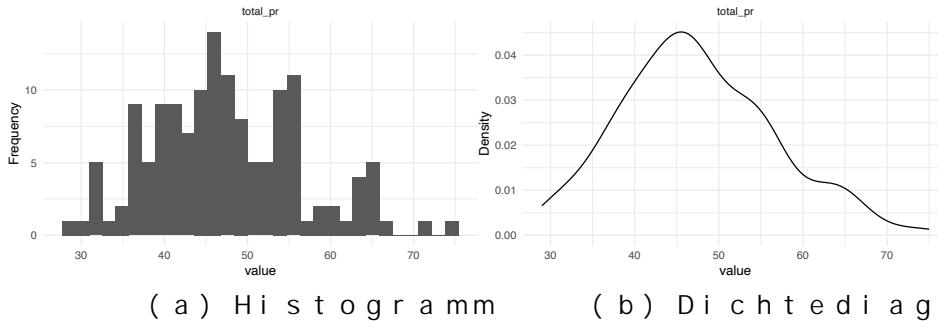
Abbildung 5.9 fügt zu einem Histogramm durch gezogene Linie). Ein Dichtediagramm ähnelt "Histogramm".

Definition (Histogramm): Histogramm ist eine Verteilung einer stetigen Variablen. Im Geg

5.4. Verteilungen verbill

Listing 5.2x zur Erstellung eines Histogramms

```
library(ggplot2)
set.seed(12345)
filter(total_pr >= 0) # ohne Extremwerte
plot_histogram
```



(a) Histogramm (b) Dichtediagramm

Abbildung 5.8.: Eine stetige Verteilung

der Verlauf der Kurve geglättet, so kann Rauschen beseitigt werden. Die Wahrscheinlichkeit für die Anzahl Beobachtungen pro Einheit der Variablen auf der

Übungsaufgabe 5.4. Sie das Diagramm Abbildung 5.8. Teil 2. Diagramm!

Verteilungen unterscheiden sich z. B. einer **symmetrischen** oder „mittleren“ Wert (vgl. Kapitel 6.5) und einer **asymmetrischen** (negativ oder positiv) Verteilung oder **schief** (nicht symmetrisch) sein, s. Abbildung 5.11.

Abbildung 5.11 zeigt verschiedene Formen von Verteilungen, die man „zweigipflig“ und „multimodal“ entsprechen.

²Grob genug ist es %>% plot_density()

³Quelle: ifes/FOMHochschule, https://github.com/FOMHochschule/Statistik_in_der_Psychologie

5. Datenverbilldlichen

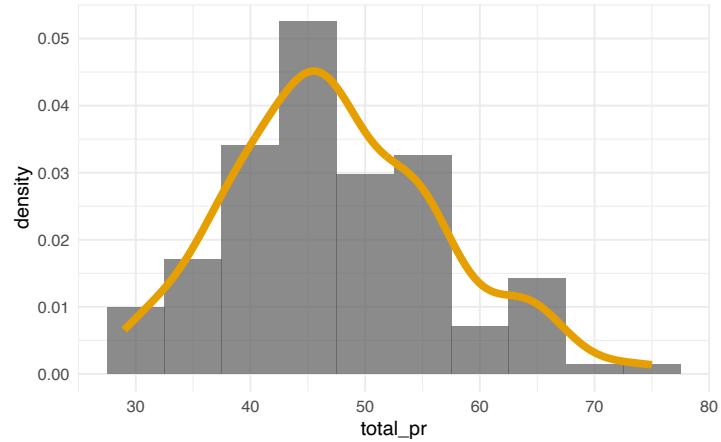


Abbildung 5.9: Histogramm (graue Balken) und Dichtefunktion (orange)

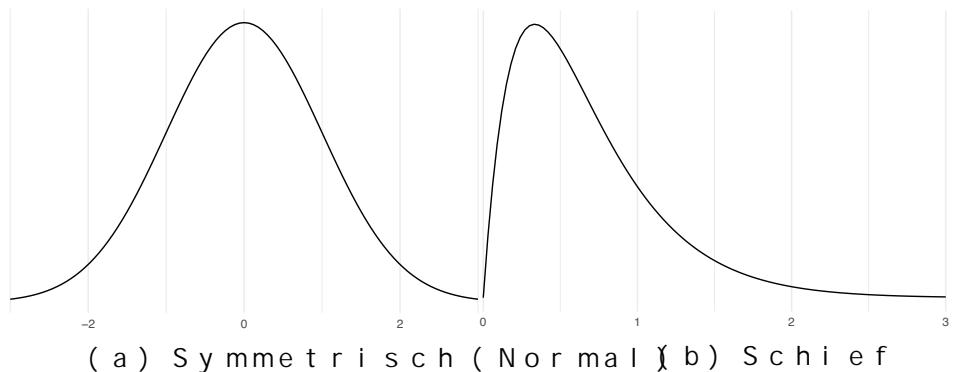


Abbildung 5.10.: Symmetrische vs. schief

5.4. Verteilungen verbill

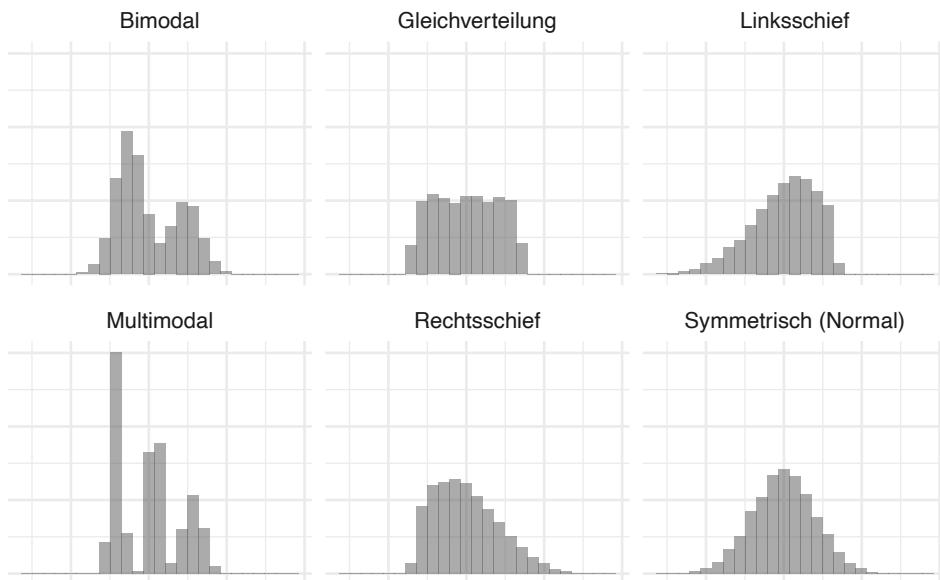


Abbildung 5.11.: Verschiedene Verteilungen

Übungsaufgabe Beispielfragen
Benennen Sie die am besten passende Verteilungstyp für die Variable

Lösung

Die Verteilung ist rechtsschief.

$\$$

\check{Z}

Eine Normalverteilung ist eine bestimmte Art quantitativen Variablen. Aber sie ist besonders hier besonders hervorgehoben. Eine Normalverteilung ist eine Verteilung, die sich unter (rechte Häufigkeiten) befindet, das sie sich unter (rechte Häufigkeiten) befindet. Sie hat u.a. folgende Eigenschaften:

- symmetrisch
- glockenförmig

5. Datenverbildung

- stetig
- eingipflig (unimodal)
- Mittelwert, Median und Modus sind identisch

Beispiele für normalverteilte Variablen: Lebensdauer von Männern oder Frauen, IQ-Werte, Prüfungsbewerber, Lebensdauer von Glühbirnen, Gewichte von Brustumfangschottischer Schafe, Brustumfang von Kühen, Brustumfang schottischer Schafe.

Definition: Eine Normalverteilung ist die Art von Verteilung einer quantitativen, metrischen, glockenförmigen, stetigen, unimodalen und modusidentischen Variable. Sie lässt sich durch zwei Parameter beschreiben: Mittelwert und Standardabweichung.

Definition: Wenn die Summe mehrerer, unabhängiger, eingesetzter Variablen ein Summanden, dann kann man erwarten, dass sie tendenziell normalverteilt ist.

Dieses Phänomen kann man gut am Beispiel des Gitters eines Bretts veranschaulichen.



Eine Normalverteilung lässt sich exakt beschreiben: ihres zentralen Wertes (Mittelwert) und seiner Standardabweichung.

Kennt man diese beiden Parameter, so kann man den Anteil der Fläche unter der Kurve bestimmen (Abbildung 5.12).

Davon leitet sich die „68-95-99.7-Prozent-Regel“ ab:

- 68 % der Werte im Bereich $\mu \pm \sigma$
- 95 % der Werte im Bereich $\mu \pm 2\sigma$
- 99,7 % der Werte im Bereich $\mu \pm 3\sigma$

5 . 5 . Z u s a m m e n h ä n g e v e r b i l d

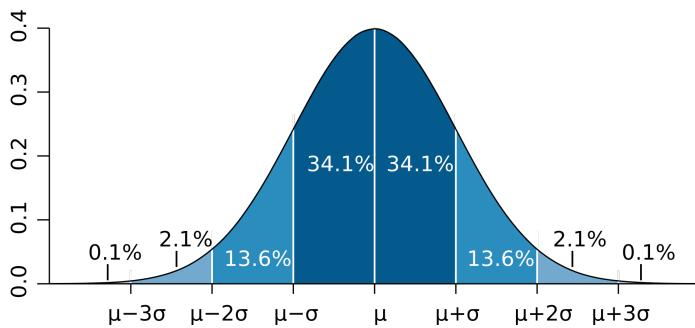


Abbildung 5.5. Flächeninhalte (Wahrscheinlichkeiten) einer Normalverteilung in Abhängigkeit der Standardabweichung (Ainali, 2007)

+

+

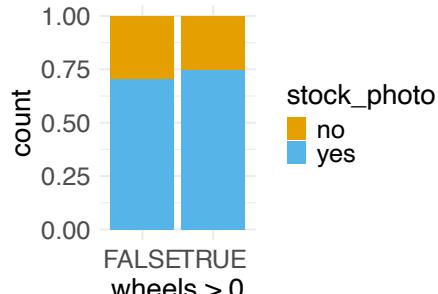
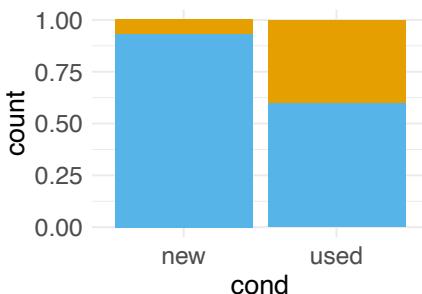
Beispiele für Zusammenhänge bei nominalen Variablen:

- Hängt Berufserfolg (Führungskraft ja/nein) zusammen?
- Hängt der Beruf des Vaters mit dem Schulabschluss (Abitur, Realschule, Mittelschule) zusammen?
- Gibt es einen Zusammenhang zwischen Automarke und Präferenz einer Partei?

Sagen wir, Sie arbeiten immer noch bei McDonald's und fragen sich, ob ein Produktfoto wohl primär bei Neukunden beiliegt, aber nicht bei gebrauchten? Dazu benötigen wir die Daten, s. Abbildung 5.13.

Tatsächlich: Es findet sich ein Zusammenhang zu ob dem versteigerten Produkt ein Foto beigelegt war (Abbildung 5.13, links). Bei neu-

5. Datenverbildung



(a) starker Zusammenhang, (b) schwacher Zusammenhang

Abbildung zeigt einen starken Zusammenhang zwischen nominal skalierten verbildunglichen. (a) Es findet sich ein Foto und Zustand in den Daten. (b) Es gibt kein Zusammenhang mit dem Foto in den Daten.

(ca. 90%) ein Foto dabei; bei gebrauchten Spuren Hälfte der Fälle.

Anders sieht es aus für die Frage, ob ein (oder keine) Spielbeilagen (oder nicht) in Zusammenhang steht. Es fast keinen Unterschied zwischen neuen und gebrauchten Produkten. Die Frage nach „Foto des Produkts dabei“ betraf ca. 70% der gebrauchten Produkte. Der Anteil betrug jeweils ca. 70%. Das zeigt, dass es keinen Unterschied zwischen den beiden Gruppen gibt.

Bildlich gesprochen: Unterschieden sich die Gruppen in Bezug auf die Variable „Foto dabei“, so unterscheiden sich die Gruppen in Bezug auf die Variable „Foto dabei“. Dies ist eine Aussage über die Variablen „Foto dabei“ und „Foto dabei“.

So können Sie sich in Reingefülltes Balkendiagramm z. B. `print(ggplot(mario %>% filter(cond == "new") %>% group_by(stock_photo), aes(stock_photo)) + geom_bar()` darstellen. Eine Darstellung eignet sich, um Zusammenhänge zwischen nominal skalierten Variablen zu verbinden. Die Füllfarbe wird in den Säulen angegeben, wie in Abbildung 5.3.

5.5. Zusammenhänge verbild

Listing Syntax für eingefülltes Balkendiagramm

```
mariko%> t  
sel(cbind, stok%photo)  
plot(byl=co) #däus dem Paket DataExplorer
```

Gefüllte Balkendiagramme zur Analyse eines Zusammenhangs zwischen nominalskalierten Variablen. bei den Variablen nur zwei Ausprägungen aufweisen. Zusammenhänge nicht mehr so gut zu erkennen. Auf Balken auf gleiche Länge (100%) ausgerichtet.

Übungsaufgabe 5: Zusammenhang von Mariko und seinen Variablen

```
mariko%> t  
# Mache aus einer metrischen eine nominale Variable  
mut(whee|e If a e(twhee|el>s))  
sel(cbind, wheels)  
plot(by=co) d "
```

Lösung

Wheels ist als metrisch nicht kategorial (h. Ganzzahl) für im Datensatz. Wir müssen Sie zunächst als Fakt umformatisieren, damit sie als nominal skaliert werden.

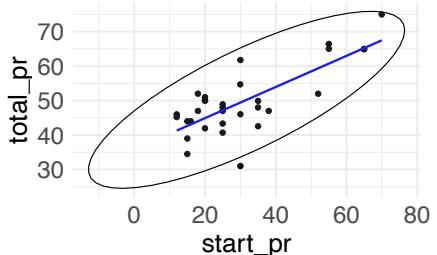
+

,

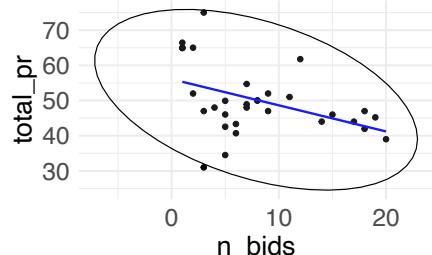
Den (etwaigen) Zusammenhang zweier metrischer Werte mit einem Balkendiagramm visualisieren (engl. scatterplot). Abbildung 5.14 links untersucht den Zusammenhang des Abschlusspreises (X-Achse) und Abschlusspreises (Y-Achse) der Versteigerungen des Computerspiels Mario Kart.

5. Datenverbildung

eine „Trendgerade“ (Regressionsgerade), hängt besser zu verdeutlichen. Die Trendgerade nach rechts). Daraus kann man schließen: Ein gleichsinniger Zusammenhang: Je höher der Abschlusspreis, zumindest tendenziell verläuft „mittig“ in den Daten (wir definiere eine Trendgerade gibt Aufschluss über „typisch“ ist „typisch“ für einen bestimmten X-Wert untersucht den Zusammenhang zwischen Anzahl und Abschlusspreises (Y-Achse). Es handelt sich um einen positiven Zusammenhang: Je mehr Gebote gibt es, desto höher der Abschlusspreis. Das erkennt man an der sinkenden Trendgeraden, wie eng die Daten um die Trendgerade herum liegen, wie stark der Absolutwert der Abweichung ist.



(apositive, mittelstark) Zusammenhang



(negativ, schwach) Zusammenhang

Abbildung 5.16: Beispiele zur Darstellung eines zweier metrischer Variablen

Definition einer Zulässigkeit: die Beziehung zwischen zwei Variablen mit einer Geraden man von einem linearen Zusammenhang. Änderungen einer Variablen um einen bestimmten Wert (z. B. 1) um einen proportionalen Wert (z. B. 0.5).

Natürlich könnte man auch nicht-lineare Zusammenhänge einfache halber konzentrierte

5.5. Zusammenhänge verbild

Beispiele für nicht-lineare Zusammenhänge sehen.

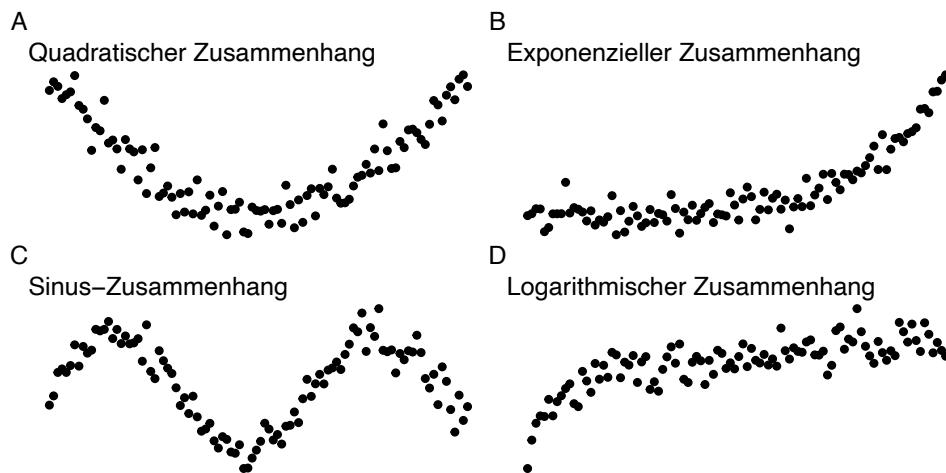


Abbildung 5.15.: Beispiele nichtlinearer

Definition: **Stärke** eines positiven (negative) Zusammenhangs ist die Größe der Abhängigkeit, die die Stärke des Zusammenhangs anzeigt. Die Stärke kann durch den Korrelationskoeffizienten bestimmt werden.

Starke Zusammenhänge erkennt man an schmalen Ellipsen, schwache Zusammenhänge an breiten Ellipsen („Taschen“). Eine starke positive Abhängigkeit ist ein starkes positive Vorzeichen, eine schwache positive Abhängigkeit ist ein schwaches positives Vorzeichen. Eine starke negative Abhängigkeit ist ein starkes negatives Vorzeichen, eine schwache negative Abhängigkeit ist ein schwaches negatives Vorzeichen. Ein Korrelationskoeffizient von +1 oder -1 bedeutet einen perfekten linearen Zusammenhang, ein Wert von 0 bedeutet keinen linearen Zusammenhang.

- ± 0 : Kein Zusammenhang

⁴Quelle: Aufbauend auf FOM/fifes, Autor: Norman Markgraf

5. Datenverbillchen

- ± 0.1 : schwacher Zusammenhang
- ± 0.3 : mittlerer Zusammenhang
- ± 0.5 : starker Zusammenhang
- ± 1 : perfekter Zusammenhang

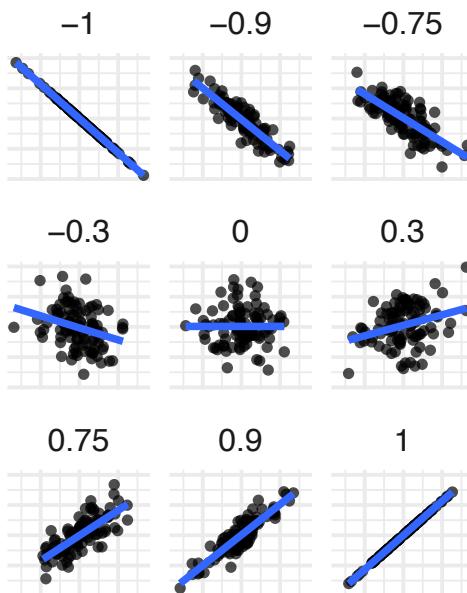


Abbildung 5.16: Zusammenhänge verschieden
Richtung

Abbildung 5.17 hat die gleiche Aussage wie Abbildung 5.16, nur dass die Abhängigkeiten hier nicht linear sind. Die Beziehungen sind hier nicht linear, sondern exponentiell, logarithmisch oder polynomisch.

Man sieht in Abbildung 5.16 und Abbildung 5.17 dass die Abhängigkeiten nicht linear sind. Die Beziehungen sind hier nicht linear, sondern exponentiell, logarithmisch oder polynomisch.

5.5. Zusammenhänge verbinden

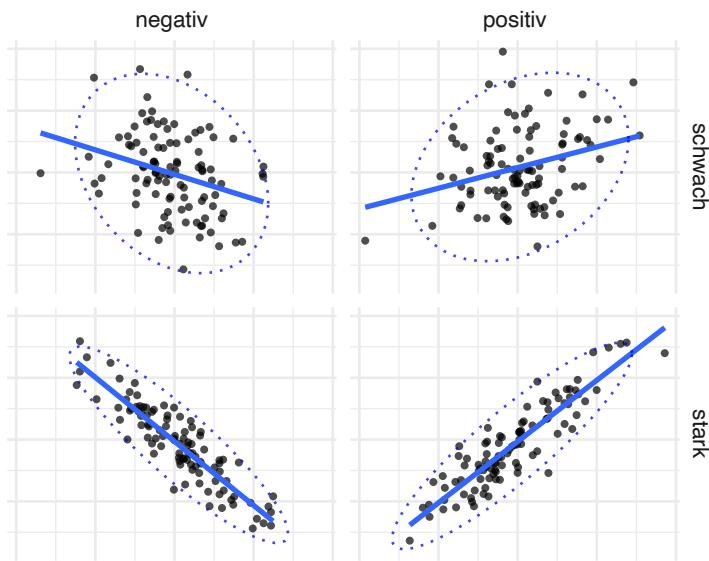


Abbildung 5.5. Übersicht über starke vs. schwache bzw. negative Zusammenhänge

Beispiel 5.5 Arbeiten nach wie vor bei einem Online und manchmal gehört Datenanalyse zu Ihren Aufgaben. Fassiert Sie, ob welche Variable mit dem Absatz zusammenhängen. Sie verbindlichen R, und zwar nutzen Sie das `plot()`-Kommando. Dazu müssen wir noch die Daten importieren, falls noch nicht gemacht.

So, jetzt kann die eigentliche Arbeit losgehen: Die Variablen konzentrieren werden, indem die übrigen Variablen ausgesondert werden. Dann weisen Sie R an, mit welchen Variablen es sich handelt. Das Ergebnis sieht man hier:

```
marie%>%t
sel(enctbids, start_percent, total_percent)
plot_sc(byt="etoptlandrt_opt")
```

5. Datenverbillchen

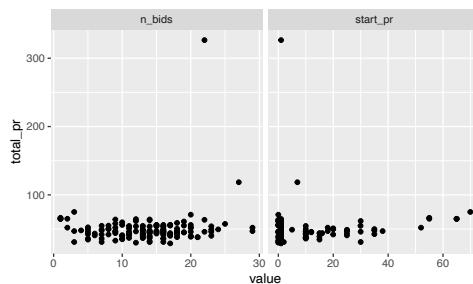


Abbildung 5. Z1u8s. ammenhang einiger metrischer Abschlusspreis

Aha... Was sagt uns das Bild? Hm. Es scheint etwas geben, die dafür sorgen, dass der Rest der Datenquetsche auf dem Bild erscheint. Vielleicht kann man diese Werte lieber entfernen? Sie entscheiden den Abschlusspreis von weniger als 1000000000. Listing 5.4.

Listing 5.4: Kart ohne Extremwerte

```
mariokart_no_extreme <-  
  mariokart %>%  
  filter(total_pr <
```

Ohne Extremwerte schätzt sich ein deutlicher (starker) Zusammenhang zwischen Anzahl der Bids und Abschlusspreis zusammen. Das ist in der Boxplot_scatterplot_Roatn, als Y-Variablen ablesbar. Alle übrigen Variablen kommen jeweils einzeln.

Übungsaufgabe: Melden Sie einen Zusammenhang der beiden metrischen Variablen. Verwenden Sie den Datensatz ohne Extremwerte.

Lösung

5.6. Unterschiede verbinden

```
mario_kart_ho_extreme  
sel(esctart_pr>total_pr)  
plot_sct(pyt="trop1)d_t_pr"
```

Zuerst wählt man die Spielart und dann visualisiert man diese in einem Plot.

&

&

Gute Nachrichten: Für nominale Variablen bietet es sowohl zur Darstellung von Zusammenhängen als den an. Genaugenommen zeigt ja Abbildung 5.13 (zischen neuen und gebrauchten Spielen hinsichtlich Fotos beiliegen. Und wie man in Abbildung 5.19 der Spiele mit Foto bei den neuen Spielen höheren Anteil an.

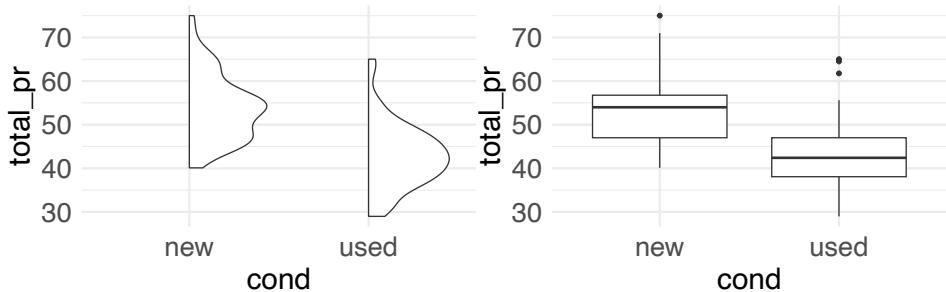
&

Eine typische Analysefrage ist, ob sich zwei Gruppen unterscheiden. Wenn man z.B. oft, ob sich die Mittelwerte der Zielvariablen deutlich unterscheiden. Das bestätigen wir schauen uns einige Beispiele an, s.

Das linke Teildiagramm von Abbildung 5.19 zeigt, dass der Anteil von gebrauchten Spielen für neue und gebrauchte Spiele unterschiedlich ist. Das rechte Teildiagramm zeigt die gleichen Werte einer vereinfachten, groben Balkendiagramms.

⁵Übrigens: Freunde lassen Freunde nicht Balkendiagramme darstellen: <https://github.com/cxli233/Friendship-lets-friends-make-bar-plots-for-means-separately>

5. Datenverbildung



(a) Histogramm pro Gruppe Boxplot pro Gruppe

Abbildung 5.1 zeigt Unterschiede zwischen zwei Gruppen
Variable, nominale X-Variablen

„deutlicher“ (substanzieller, bedeutsamer
Zusammenhang ist, ist keine statistische
die man mit Sachverstand zum Forschungsgesetzen
muss.

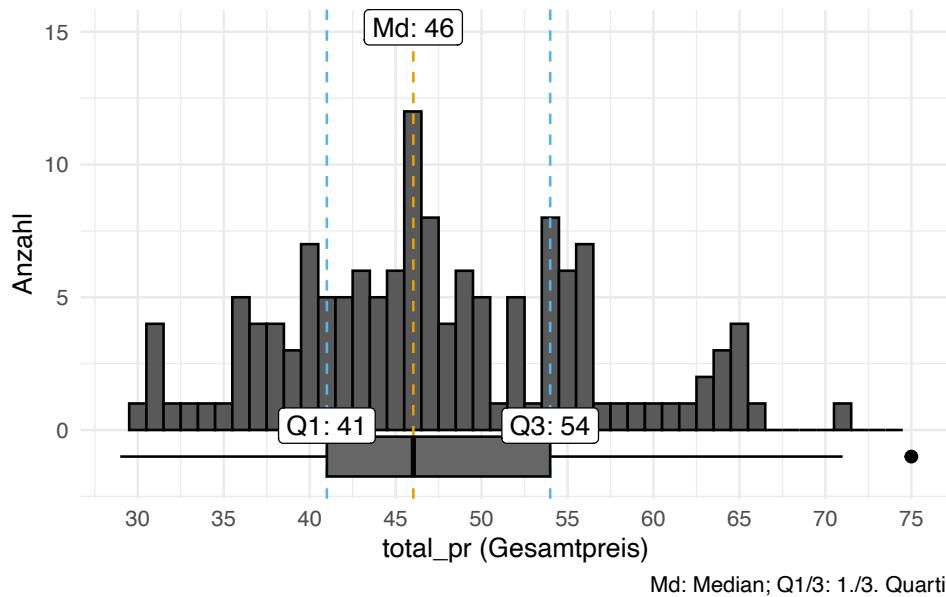
Definition: Der Boxplot ist eine Vereinfachung
Zusammenfassung eines Histogramms. Damit ist die
Verteilung (einer metrischen Variablen) dargestellt.

In Abbildung 5.20 sieht man die „Übersetzung“
zueinem Boxplot (unten). Ob der Boxplot horizontal
ist, ist Ihrem Geschmack überlassen.

Schauen wir uns die „Anatomie“ des Boxplots an:

1. Die dicke Strichbox zeigt den Median (Kapitel 6.3).
2. Die Enden der Box sind das 1. Quartil (41) bzw. Das 3. Quartil (49). Damit zeigt die Breite der Box die Streuung genauer gesagt die Streuung der inneren 50% der Werte. Je breiter die Box, desto größer die Streuung. Man nennt man auch die Quartilsstandartabweichung.
3. Die Antennen des Boxplots zeigen die Streuung der 25% der Werte (linke Antenne) bzw. die Streuung der 75% der Werte (rechte Antenne).

5 . 6 . U n t e r s c h i e d e v e r b i l



A b b i l d u n g 5 . 2 0 . : Ü b e r s e t z u n g e i n e s H i s t o g r a

der Werte (rechte Antennen). Je länger die Antennen, desto größer ist die Streuung.

4.Falls es sich um eine symmetrische Verteilung handelt, so sollte die Box die gesamte Breite des Wertebereichs ausfüllen. Da die Antennenlängen auf die 1,5-fache Länge der Box verteilt sind, werden die Boxlängen von Q1 bis Q3 entsprechen. Die Boxlängen von Q1 bis Q3 entsprechen der Boxlänge von Q3 entfernt sind.

5.Liegt der Median-Strich in der Mitte der Box, so ist die Verteilung symmetrisch. Liegt der Median-Strich nicht in der Mitte der Box, so ist die Verteilung asymmetrisch. Dies gilt für die Antennenlängen: Sind die Antennen gleich lang, so ist die Verteilung symmetrisch, andernfalls ist sie asymmetrisch.

Beispiele in [5.1](#) der vorherigen Analyse haben Sie gesehen, dass es einen engen Bezug zwischen Abschlusspreis und der Anzahl der Lebewesen gibt. Jetzt möchten Sie eine sehr ähnliche Frage stellen:

5. Datenverbildung

unterschied ~~h~~ die Verkaufspreise je nach Anzahl Lenkräder? Flink erstellen Sie die dazugehörigen Links. Es zeigt die Verteilung des Abschlusses Anzahl Lenkräder "durchschnitts"

Aber ganz glücklich sind Sie mit dem Diagramm wheelos nicht aufgeteilt. Es wäre eigentlich Gruppen von 2, 3 und aufteilen würde. Aber schaut Y-Achse (im linken Teil des Diagramms von Abbildung 5.21) man, dass es eine stetige Zahl betrachtet und nicht gruppiert. Vielleicht so, dass in jeder Gruppe 1000 Räder wären. Aber wir möchten jeden einzelnen Wert, der in einer Gruppe stehen. Mit anderen Wörtern müssen wir eine Variable definieren. Das kann man nicht erreichen (siehe Abbildung 5.21 rechts).

```
mariokart_no_extreme
ggplot(x ~ total_wheels)
geom_boxplot
```

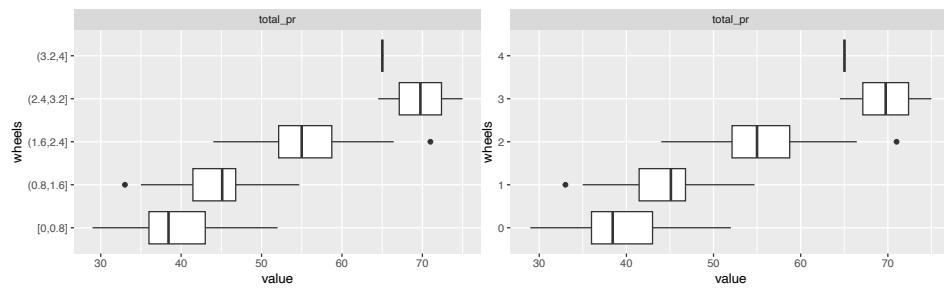
```
mariokart_no_extreme
sel(etcatal_p%>%wheels)
plot_by_wheel
```

```
mariokart_no_extreme
sel(etcatal_p%>%wheels)
mut_if_satisfies()
plot_by_wheel
```

Sie schließen aus dem Bild, dass Lenkräder anhängen. Allerdings schwanken sie weniger. Das prüfen Sie nach:

```
mariokart_no_extreme
cou(nwheels)
```

5.6. Unterschiede verbill



(a) he e allss metrische V(a)lue ballss nominale Variab le

Abbildung 5.21: Preis nach Anzahl von Beigedern

wheels	n
0	36
1	52
2	50
3	2
4	1

Tatsächlich mag ich kein Lenkrad mehr. Lenkräder schaffen es nicht, so viele Daten zu speichern. Wenn wir die Belastbarkeit dieser Betrachtungen untersuchen, müssen wir die Belastbarkeit der Daten untersuchen. Übrigens bezeichnet sie „Datengott“.

Übungsaufgabe 5.21: Basieren Sie den Unterschied zwischen gebrauchten und gebrauchten Gebrauchtwagenpreisen auf den Diagrammtypen, die sich anbieten; mehrere Lösungen sind möglich.

Lösung

```
mario_kart_photographie
sel(exand, total_pr)
plot_boxplot("
```

5. Datenverbildung

Boxplots sind eine gute Möglichkeit, die Variablen, aufgebrochen auf mehrere Gruppen

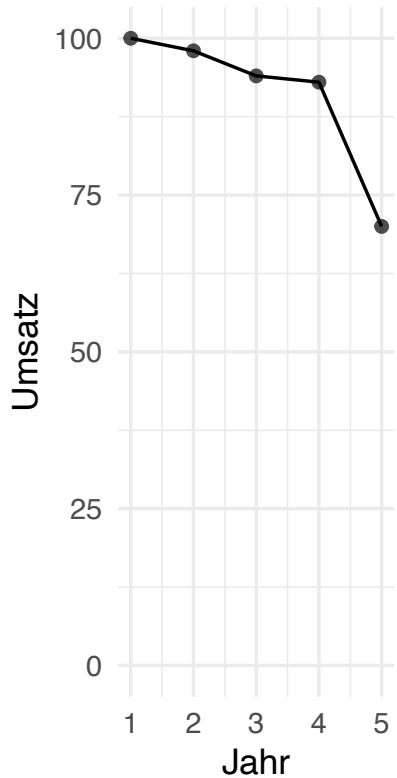
Übungsaufgabe 5a: Preisunterschiede im Verkauf von Mario Kart 8. Interpretation
Unterschied im Verkauf von Mario Kart 8. Eine metrische Gruppierungsvariable. Interpretation
Lösung

```
mario_kart_no_extreme  
sel(eshti_p_pr | total_pr)  
plot_b(y_p_bit)p_pr"
```

plot_b(y_p_bit)p_pr": automatisch in fünf Gruppen (mit gleichen Ranges). Wir müssen die metrische Gruppierungsvariable umzuwandeln. Es sieht so aus, als würden die Gruppen leicht steigen, mit Ausnahme der mittleren.

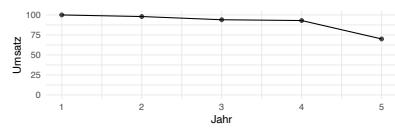
\$ \$

Diagramme werden mitunter eingesetzt, um „schenken“. Hier folgen einige gebräuchliche Techniken:
Achsen zu stauchen, ist eine einfache Trick, kann man auch durch „Abschneiden“ der Y-Achse einen Effekt erzielen, s. Abbildung 5.23. So kann man z.B. einen kausalen Effekt zu verkaufen, ist einmal wieder beobachten kann. Ein Beispiel: von einem Zusammenhang von Schokoladenkonsum und Nobelpreisen (Beobachtungseinheit: Länder) ist doch ganz klar: Schokofutter macht schlau. Leider ist hier ein Kausalitätszweifel: Auch wenn bei den Vätern der Kinder ein hohes Alter vorkommt, kann es nicht gesagt werden, dass die Variable die Ursache der Wirkung sein muss. So könnte auch eine Drittwirkung eine Rolle spielen.



(aO)hnein, dramatischer Einbruch des Umsatzes!

Abbildung 5.2.2: Ein und Stauchen der Y-Achse, um zu lügen



(b) kaum der Redewert, ist nur ein bisschen Schwankung!

5. Datenverbildung

Umsatz

die g
se in
wick
mehn
Eutsw.

Abbi

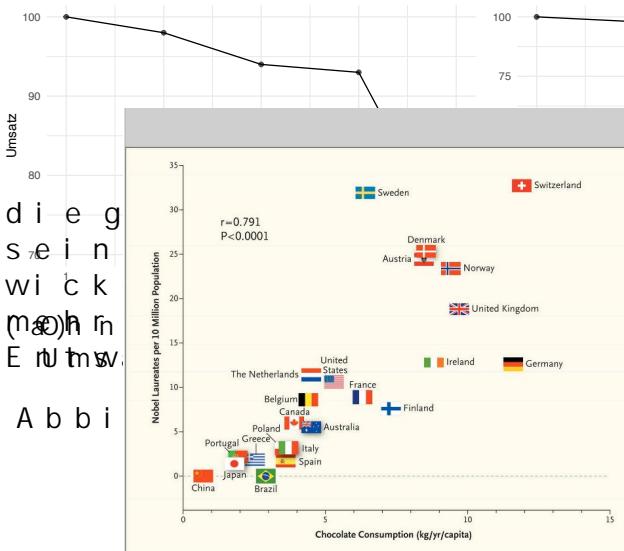


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

chokolade
stamdhöher
okoladekorn
vergleichsztu
ng!

er Y-Achse,

Abbildung 5.24.: Schokoladenkonsumun

!

Ein, wie ich finde, schlagendes Beispiel zur Stärke des Einflusses der Impfung ist Abbildung 5.25. Das Diagramm zeigt die Häufigkeit von Mäusen vor und nach der Einführung der Impfung. Die Datenvisualisierung gehen auf van Panhuis et al. (2015) zurück.

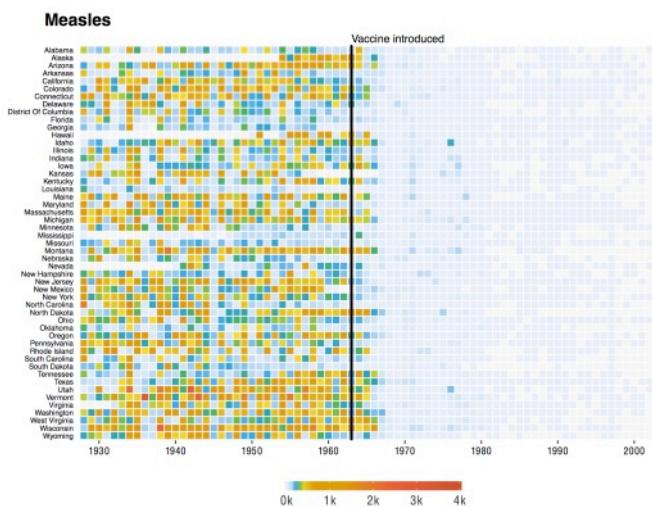


Abbildung 5.25: Häufigkeit von Mäusen und Impfung in den USA (van Panhuis et al., 2015)

In der „freien Wildbahn“ findet man häufig sog. Zwar sind sie beliebt, doch von ihrer Verwendung bei Tortenstücken ist es schwer, die Größe

"

\$

"

Ein Teil der Diagramme dieses Kapitels wurden mit dem R-Paket treemap erstellt. Mit diesem Paket lassen sich einfache

5. Datenverbildung

gramme erstellen.

```
library(ggplot2)
beginning_installations nicht vergess
mario%>%
filter(totalspr %
ggbox(x = "cond", y = "total_pr")
```

Möchte man Mittelwerte vergleichen, so sind diese ja nicht den Mittelwerten ausreichend Abhilfe (also eine Darstellung des Mittelwurnd der Fugktionsn Abbildung 5. 26).

```
ggviolin(mario %>%
filter(total_pr ~ cond)
```

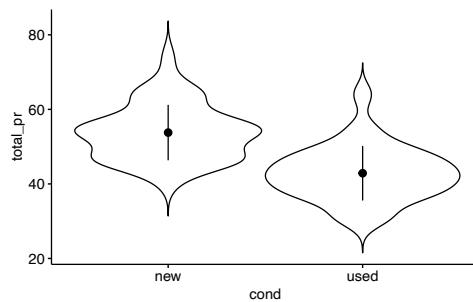


Abbildung 5.26: Gleich der Verteilungen zweier Mittelwert und Standardabweichung präzisiert

Ein „Violinenplot“ hat die gleiche Aussage. Je breiter die „Violine“, desto mehr Beobachtungen liegen an einer Stelle.

Übrigens sind Modelle- und Diagramme sind Vereinfachung, lassen also Informationen wichtige Informationen.

Einige Überlegungen zur Farbwahl findet sich bei Ichihara. Die Farbpalette von Okabe und Ito ist (vgl. Ichihara) ansprechend, da sie auch bei Schwarz-Weiß-Druck und die Farben noch recht gut unterscheiden lässt.

```
mari o%>a% t
f l ( t o t a l 0 ) 0 % 
g g b o x ( p l o c t o n y d = " t o t a f l l _ " p e r o " ) + d "
s c a l e _ f l l ( ) o k a b e i t o
```

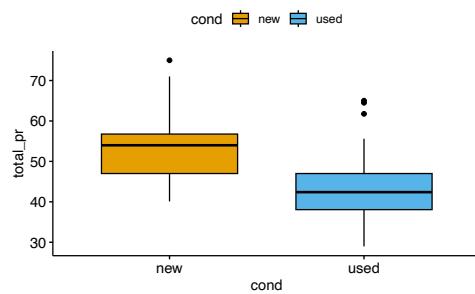


Abbildung 5.27. Boxplot von Okabe und Ito: Geeignete Schwächen und für Schwarz-Weiß-Druck net anzuschauen.

Mit `cond = new` erkennt man, dass die Färbung gegebenen Wert wird: Jeder `color` (Von `new`) bekommt eine eigene Werte. Welche das ist, hängt vom verwendeten Farbschema ab. Das Farbschema von Okabe und Ito verwendet (Ichihara)

Die Webseite `datenwerk.netlify.app` stellt eine Übungsaufgaben bereit. Sie können die Suchfunktion nutzen, um die Aufgaben mit den folgenden Namen zu finden:

5. Datenverbildlichen

1. boxhist
2. max-corr1
3. max-corr2
4. Histogramm-in-Boxplot
5. Diamonds-Histogramm-Vergleich2
6. Boxplot-Aussagen
7. boxplots-de1a
8. movies-vis1
9. movies-vis2
10. vis-gapminder
11. boxplots-de1a
12. diamonds-histogramm-vergleich
13. wozu-balkendiagramm
14. diamonds-histogram
15. n-vars-diagram

Weitere Aufgaben zum Thema Datenvisualisierung unter dem Tagvis.

7

So wohl präzise als auch Explanatorisch sind viele andere R-Pakete auf dem R-Grafikum-Plotting-Platz eines der am weitesten arbeiteten Softwarepakete zur Erstellung Buchs zur Software (vgvom Autorenempfehlenswertham, 2016). Eine neuere, gute Einführung findet sich bei Wilke (2019). Beide Bücher sind kostengünstig (2019) gibt einen hervorragenden Überblick über die Datenvisualisierung; gut geeignet, wenn man eine Richtung geht Fisher & Meyer (2018).

!

"

In diesem Kapitel benötigen Sie die üblicheeasy start Dantaen Økast Kapitel 3.7.3.

Ł

- Sie können gängige Arten von Lagemaße definieren.
- Sie können erläutern, inwiefern man einen Lagernehmen kann.
- Sie können Lagemaße mit R berechnen.

{

}

Der „klassische“ Mittelwert (das arithmetische Beispiel für ein Modell in der Statistik

Übungsaufgabe 1 Vorstellung haben Sie, wenn dass der „typische deutsche Mann“ 1,80m groß (2013) ?

- a) Die Hälfte der Männer ist größer als 1,80 m kleiner.

6. Punktmodelle 1

- b) Das arithmetische Mittel der Männer beträgt 1,78 m.
- c) Die meisten Männer sind 1,80 m groß.
- d) Etwas anderes.
- e) Keine Ahnung!

Übungsaufgabe 6: In der Statistischen Bundesamtsstudie „Gesundheit in Deutschland“¹ beträgt der Wert der mittleren Größe deutscher Männer 1,78 m, während er bei Frauen 1,65 m beträgt. Ist das viel?

Ist das viel?

- a) ja
- b) nein
- c) kommt drauf an
- d) weiß nicht

Antwort

Auf dieser Frage gibt es keine Antwort, zumal es sich um eine geschätzte Größe handelt. Annahmen. So könnte man z.B. sagen, „mehr als 10 cm mehr ist aber kein statistisch relevanter Unterschied.“

Beispiel: Ein Kurs bestehend aus drei Studentinnen: Anna, Berta und Carla. Sie haben die Klausur erfahren. Anna hat eine 1, Berta eine 2 und Carla einen Durchschnitt (das arithmetische Mittel),

QUESTION MARK emoji: Zueasty!

QUESTION MARK emoji: Schongut! Chill mal. Wir dgleich spannen uns auf.

Die Rechenregel zum Mittelwert lautet:

1. Addiere alle Werte
2. Teile durch die Anzahl der Werte
3. Fertig!

¹https://en.wikipedia.org/wiki/Average_height_in_Germany

6.2. Mittelwert als Modell

Etwas abstrakter kann man Beispiel 6.1 in folgenden Schritten darstellen, s. Abbildung 6.1.

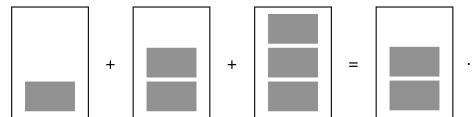


Abbildung 6.1.: Visualisierung von Beispiel 6.1.

Das Beispiel zeigt uns: Der Mittelwert ist eine Zähleinheit, die mehrfach multipliziert, gleichzeitig sein Mittelwertsumme. Der Nutzen des Mittelwerts liegt darin, dass das Modell ist!) für die „typische Note“ im Statistik-

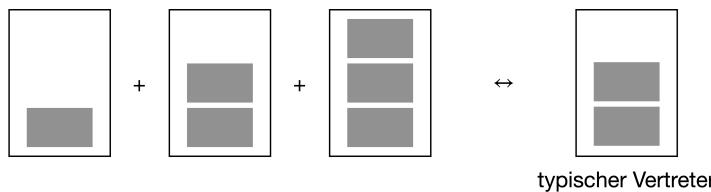


Abbildung 6.2.: Der Mittelwert als „typischer Vertreter“.

Der Nutzen des Mittelwerts liegt darin, dass er eine „typische Note“ („typische Vertreter“) zugeordnet werden kann, als dass die Werte aller Merkmale zusammengefasst werden. Er gibt uns eine (mögliche) Vorstellung, wie wir uns die Werte der Datenreihe vorstellen können. Eine anschauung zum Mittelwert ist die Vorstellung eines ausbalancierten Wippes, s. Abbildung 6.3. In „Maßen des Mittelwerts“ geht es darum, mit dem Mittelwert die Rechenregeln der Gleichung 6.1.

$$\frac{1}{3} = \frac{1 + 2 + 3}{3} = 1 \quad (6.1)$$

Definition (Mittelwert): Der Mittelwert (MW, arithmetisches Mittel) ist die Summe der Werte aller Merkmale, die die Summe der Werte aller Merkmale darstellt.

6. Punktmodelle 1

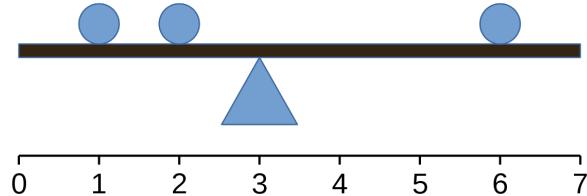


Abbildung Mittelwert als ausbalancierte Wippe
(Maphry, 2009)

der Elemente verteil ist durch den Mittelpunkt, heißt Wert und bezeichnet man auch mit

Bei spielt gekommen, wir haben eine Reihe von Mittelwert der Noten bestimmt ($1 + 2 + 3 = 6 / 3 = 2$)

Da der Mittelwert eine zentrale Rolle spielt, ihn uns noch etwas genauer anschauen. Ins. Alle Noten von (dieses Mal) vier Studentinnen. Die Linie zeigt den Mittelwert der vier Noten. die Daten, in dem Fall die einzelnen Noten. Die Abweichungen der Noten zum Mittelwert.

Bezeichnen wir die Abweichung - auch als „Fehlern“ bezeichnet - y (negative Fehler) und die Note, x , dann können wir mit Gleichung 6.

$$y = x - \bar{x} \quad (6.2)$$

6.2. Mittelwert als Modell

Anders ausgedrückt (s. Gleichung 6.3):

$$\text{Daten} = \text{Modell} + \text{Rest} \quad (6.3)$$

Der Mittelwert ist hier unser Modell der Daten. ist eine vereinfachte (zusammengefasste) Beschreibung. Um Modelle darzustellen, wird in der Datenanalyse Art von Modellgleichung verwendet, s. Gleichung

$$x = m + e \quad (6.4)$$

Lies: „Der Modell ist eine Funktion der Variablen“ „~“ soll also hier heißen „... ist eine Funktion die „Welle“ ~ nennt man auch „Tilde“.

Mit ist die vorhergesagte Variable AV, Output-Variablen, Zielvariable y bezeichnet. Das „vorhergesagter Y-Wert“ oder „Y-Wert laut dem Modell“, beobachtete Wert zusammen aus dem Modell plus einem Fehler Gleichung 6.5.

$$= m + e \quad (6.5)$$

Anstelle kommt man mit „y-tilde“). In diesem Fall Modell einfache gleich dem Mittelwert (und nicht des Mittelwerts), sodass wir mit Gleichung 6.

$$= + \quad (6.6)$$

Die Zielvariablen sind also durch ihren eigenen Mittelwert abgesehen wir von einem Fehler Modell vorhersagen. Nobody is perfect. In späteren Kapiteln werden heranziehen, um die Zielvariable zu erklären wollen, dass die Funktionen ein markantes Verhältnis würden wir schreiben (s. Gleichung 6.7):

6. Punktmodelle

x

(6.7)

Da wir im Moment aber keine andere Variable erklären, schreibt man mit Gleichung 6.8 auf

1

(6.8)

Diese Schreibweise sieht anfangs verwirrend aus, zeigt, dass wir keine andere Variable abhängt, daher steht hier kein Buchstabe des obigen Modells. Die mathematische Hintergrund liegt in der Art, wie man

Beispiel(e) seien, Mittelwert. Abweichungen von Anna, Berta, Carl, Dani - haben ihre Statistik (Schlüssel). Die Noten sehen Sie in Abbildung 6.8 aus. Außerdem ist der Mittelwert (nicht die Abweichungen) bezeichnet) der einzelnen Noten vom Mittel.

Schauen Sie sich die Abweichungsbalken (im Bild) näher an. Jetzt stellen Sie sich vor, Sie verzerrten die Balkenlängen aneinander längere Striche legen. Können Sie sich das vorstellen? Die Abweichungsbalken liegen ineinander (die durch gezogenen Linien). Wer viel Phantasie hat, kann die Gesamtlänge der „nach oben ragenden Balken“ (die nach oben ragenden Balken) und die „unten ragenden Balken“, voneinander ausgedrückt und ohne Phantasie zu strapazieren.

$$(\quad - \quad) = - \quad = \quad - \quad (6.9) \neq 0$$
$$= 1 \quad = 1 \quad = 1$$

Wie man in Gleichung 6.9 sieht, ist die Summe der Abweichungen vom Mittelwert Null.

6.2. Mittelwert als Modus

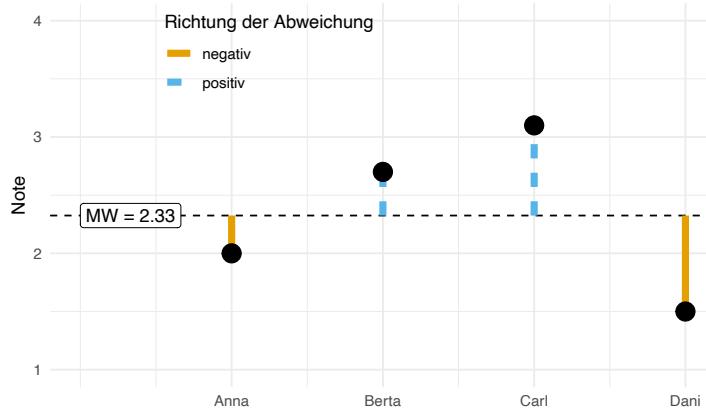


Abbildung 6.2. Der Mittelwert als horizontale (gestrichelte Linie) und vertikale Linien zeigen die Abweichungen der Werte zum Mittelwert. Die Abweichungen sind sich zu Null auf.

Übungsaufgabe 6c: h3ä.tzen Sie, wie hoch das mittlere Einkommen (arithmetisches Mittel) der Haushalte in Deutschland im Jahr 2021 auf Basis einer Umfrage (Bundesbank) war:

- a) 50.000 Euro
- b) 100.000 Euro
- c) 150.000 Euro
- d) 200.000 Euro
- e) 300.000 Euro

Beispiel: Der wertvollste Fußballspieler der Welt ist Kylian Mbappé. Kommt der wertvollste Fußballspieler der Welt in die Bundesliga, wieviel kostet er? Es ist Kylian Mbappé (Transfermarkt, 2024) liegt bei ca. 120 Millionen Euro (Arad, 2023).

Hey Leute, wie geht's denn so! Wieviel kostet Kylian Mbappé eigentlich?

2316 Tsd Euro

6. Punktmodelle

Äh, wir studieren und verdienen fast ni

Die 100 Studis im Hörsaal schauen verdatter das für eine komische Frage! ? Aber zumindest Autogramme .

Übungsaufgabe 6: Einkommen im Hörsaal ,
Schätzen Sie - im Kopf - das mittlere Vermögen Sie davon aus, dass alle der 100 Studierende verdienen.

In R kann man das mittlere Einkommen (präzise Mittel des Einkommens) wie folgt berechnen der Syntax , z. B. `mean` & `length` haben Bedeutung

Liste mit 100 multiplizieren Einkommen von 100 Studis

```
set.(42# Zufallszahlen festlegen, hier nicht  
einkommen_rset(p=1000000000# "rep" wie "repeat":  
wiederhole 1000 USD 100-mal  
einkommen<-einkommen[1:200]# 100 Studis mit 1000,  
Mb appé mit 120 Mi o  
einkommen_nmem(einkommen)  
einkommen_mw  
## [1] 1189109
```

Hinweis

1 Million hat 6 Nullen hinter der führenden Taschenrechner- oder Computereskalade ist selbst zu lesen als „1 Mal 1000000000“ mit

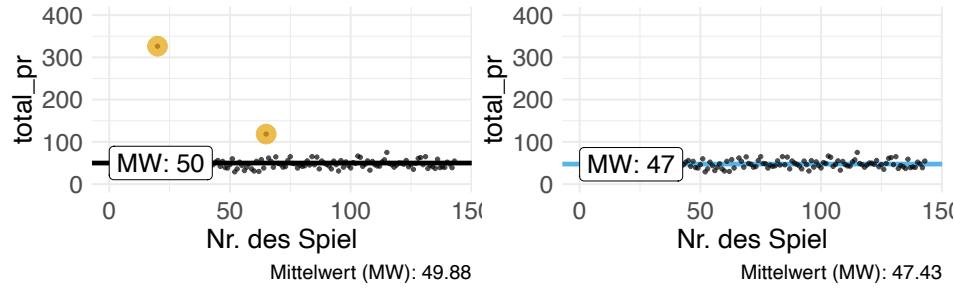
Der Mittelwert im Hörsaal beträgt also 1,1 eine Million. Ist das ein gutes Modell für Hörsaal ?

6.2. Mittelwert als Modell

Man kann den Mittelwert als Gerade einzeichnen bzw. als Gerade begreifen. Insofern kann man von linearen Modellen.

Definition: Ein einfaches Modell kann die Daten durch eine Gerade beschreiben.

□



(a) Mit Extremwerten ohne Extremwerte (< 100 €)

Abbildung 6.5 zeigt die Preis von Mario Kart-Spielen. Die Gerade eingezeichnet; einmal mit einmal ohne (b).

Abbildung 6.5 zeigt den Mittelwert des Verkaufspreises pro Spiel mit (farbig markierten) Extremwerten (a) und ohne Extremwerte (b).

Definition: Ein einfaches Modell kann die Daten durch eine Gerade beschreiben. Eine Beobachtung, deren Wert deutlich vom Großteil der Beobachtungen im Datensatz abweicht, z.B. viel geringer ist, wird als Ausreißer bezeichnet.

Berechnen wir mal den Mittelwert mit dem Befehl `lm`

6. Punktmodelle 1

```
Lineinkommen m wie "lineares Modell" oder engl  
# #  
## Call:  
## lm(formula = einkommen ~ 1)  
##  
## Coefficients:  
## (Intercept)  
## 1189109
```

Der Befehl hier mit `coefficients` gibt den Mittelwert der einzelnen Werte zurück und wird **Mittelwert** auch **Listing 6.1**. Dieser Wert wird verständlich, wenn man weiß, dass die Gerade (des Mittelwerts) genau an schneidet. Die Syntax ist `fit = lm(einkommen ~ 1)`. Ignorieren Sie das fürs Erste, wir besprechen es ausführlich später. Übrigens für „lineares Modell“

• { }

Hey, der Mittelwert ist doch Quatsch! Das ist ein schlechter Wert für die Menschen im Hörsaal. Weißt du noch für uns Studis!

Ja, da habt ihr Recht.

Die Welt ist schon ungerecht!

Abbildung 6.6 stellt die Verteilung des Einkommens dar. Erinnerung: 4.04 € 017 € 4000 € 0000 € 4 gefolgt von 7 Nullen.

Der Mittelwert ist Hörsaal ist nicht typisch für einen Hörsaal: Weder für Mbappé, noch für die Studis. Mittelwert in diesem Fall ziemlich nutzlos. Extremwerte: Gibt es einen Extremwert in einem Hörsaal?

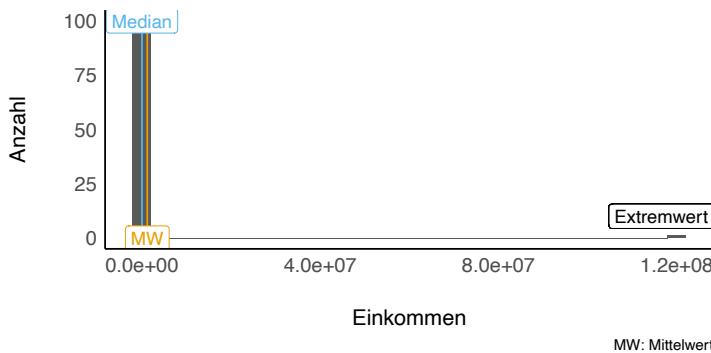


Abbildung 6.6.: Die Einkommensverteilung

der Mittelwert stark diesen Wert wider und wenige gemäßigten Werte. Man sagt, der Mittelwert ist über den Mittelwert ist Externwerten).

E wichtig

Bei (sehr) schießen Verteilungen (s. Abbildung 6.6.) ist der Mittelwert über den Mittelwert ist Externwerten).

Beispiel: Das Median-Einkommen einer Flügelfrässtudentin nennen sich über ihr Einkommen aus. Es handelt sich um eine schiefe Verteilung.

Wir könnten jetzt behaupten, dass Carla das typische Einkommen dieser Datenreihe aufweist, da es genauso viele mehr verdienen, wie solche, die weniger verdienen.

Definition: Median Merkmalssausprägung, die bei (sortierten) Beobachtungen in der Menge liegt,

Der Median ist gegenüber Extremwerten: Fügt man Extreme zu einer Verteilung hinzu, ändert sich der Median.

6. Punktmodelle 1

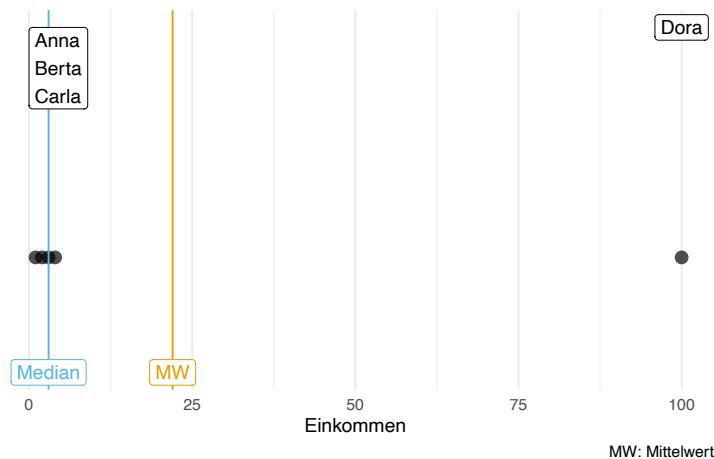


Abbildung 6.7 Einkommen einiger Studentinnen
wert (MW) ihres Einkommens

weniger als der Mittelwert. Abbildung 6.8 stellt dar.

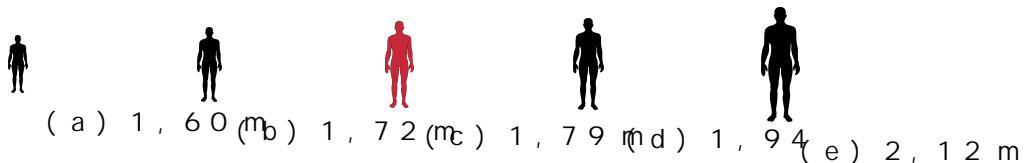


Abbildung 6.8 Median als der Wert des „mittleren“ Wertes, die Objekte aufsteigend sortiert viele Objekte mit kleinerem Wert als der Median. In dieser Abbildung farbig markiert.

Bei gerader Anzahl sind die beiden mittleren Werte bilden arithmetische Mittel aus diesen beiden Werten.

Beispiel: Bei der Messreihe 1, 2, 3, 4, 5, 6, 8, 4. Es.

6.3. Der Median als Modus

Übungsaufgabe 5: Würde die Einkommen eines Patienten Emma steinreich. Ihr Einkommen erhöht sich um 100000. Wie verändert sich der Median?

Übungsaufgabe 6: Sämtliche „mittel“? Median oder Mittelwert?

💡 Das arithmetische Mittel sollte Mittelwert sein. Die Mitten des Abstands zweier Zahlen widerspricht von 1 und 10 ist die Mitten 5,5 - also genau bei 5.

⚠️ Moment! Der Median und nur der Median zeigt den Mittelwert! Links und rechts sind gleich verteilt, wenn man die Werte der Größen nach sortiert. Also Median genau in der Mitte!

Nehmen Sie Stellung zu dieser Diskussion!

Beispiel (Eltern „mittlerer“ Preis für Mario Kart) und der Median (Mittelwert) das Streckenlängen sind nicht gleich, der Median ist größer als der Mittelwert.

```
mario%> t  
summary(price_mean(start_pr),  
        price_md(as$start_pr))
```

price_mw	price_md
8.8	1

Wie man sieht, ist der Mittelwert größer als der Median.

³ Er bleibt gleich, verändert sich kaum, wenn Extremwerte vorliegen.

6. Punktmodelle 1

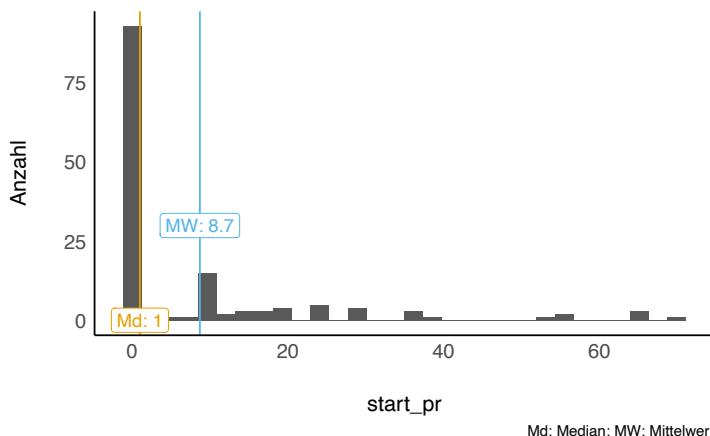


Abbildung 6.5: Startgebot bei Mario Kart - Spiel
Median und Mittelwerts sind unterschiedlich

Klaffen Mittelwert und Median auseinander, wenn die Verteilung rechtsschief ist. Ist der Mittelwert größer als der Median, so ist die Verteilung rechtsschief. Bei schiefen Verteilungen ist der Median ein besseres Modell für den „typischen Wert“.

Übungsaufgabe: Mario Kart ohne Extremwerte
mario kart hat einige wenige Spiele, die für einen hohen Preis verkauft wurden. Diese Extremen verdecken die tatsächlichen Verkaufspreise möglicherweise übereinstimmend. Entfernen Sie diese Werte und berechnen Sie den Median erneut. Vergleichen Sie die Ergebnisse.

Lösung

```
mario_kart_no_extreme <-  
mario_kart %>%  
filter(totals == 0)  
  
# ohne Extremwerte:  
mario_kart_no_extreme  
summa(totals ~ pr_mean(pr))
```

6.3. Der Median als Modus

```
total_pr_medi(ation total_pr))  
# mit Extremwerten:  
matrix  
summa(totale_pr_mittelwert),  
total_pr_medi(ation total_pr))
```

$$\begin{array}{rcc} & \text{total_pr_mittelwert} & \text{total_pr_mediana} \\ \hline & 47 & 46 \\ \hline \end{array}$$

$$\begin{array}{rcc} & \text{total_pr_mittelwert} & \text{total_pr_mediana} \\ \hline & 50 & 46 \\ \hline \end{array}$$

Wie man sieht, verändert sich der Mittelwert, wenn entfernt. Für den Median trifft das nicht □

Übungsaufgabe 6c: In welcher Woche haben die Haushalte in Deutschland im Jahr 2021 im Durchschnitt ⁴ ?

- a) 50 Tsd Euro
- b) 100 Tsd Euro
- c) 150 Tsd Euro
- d) 200 Tsd Euro
- e) 300 Tsd Euro

⁴ca. 84 Tsd Euro

6. Punktmodelle 1

"

Der Median teilt eine Verteilung in eine untere und obere Hälfte (die „Mediananteile“). Er markiert sozusagen eine „50-Prozent-Marke“ (die „Median“) in einer geordneten Reihe von Beobachtungen. Betrachten wir eine Liste von 100 Euroverkäufen (aufgelistet nach Größe). Abbildung 6.10. 50% dieser Spiele wurden für weniger als 46 Euro verkauft und 50% für mehr als 46 Euro.

Jetzt könnten wir nur die günstigere Hälfte des Medianfragments (bis zu 46 €) oder die „Mediananteile“ der unteren Hälfte“ grenzen damit das insgesamt günstigste Verkaufsangebot ab. In unserem Datensatz liegt dies bei 54 Euro. Entsprechend kann man nach dem Wert für das „Viertel vom Rest“ der Verkaufsangebote ab trennen.

Definition (Quartile): Man teilt die Daten aufsteigend in vier gleiche Teile. Der Median trennt die untere Hälfte vom oberen Viertel (25% bis 75%).

Beispiel (Quartile des Verkaufsangebots): Zeigt die Quartile für das Verkaufsangebot.

Jetzt könnte man sagen, hey, warum nur in 25% aufteilen? Warum nicht in 10%-Schritten?

Definition (Decile, Quantile): Eine Verteilung in 10 gleich große Teile unter „Gleich groß“ heißt, dass in jedem Dezil gleichviel liegt.

6.4. Quantile

Oder vielleicht in 1%-Schritten oder in sons Quartile in 25%-Schritten? Eine aufteilt Schritte auf.

Definition: Ein Quantil ist der Wert, der von der Werte nicht überschritten wird. Ein Quantile, Dezile etc.

Abbildung 6.10 zeigt das 1. (Q1), das 2. (Median) und das 3. Quartil 2

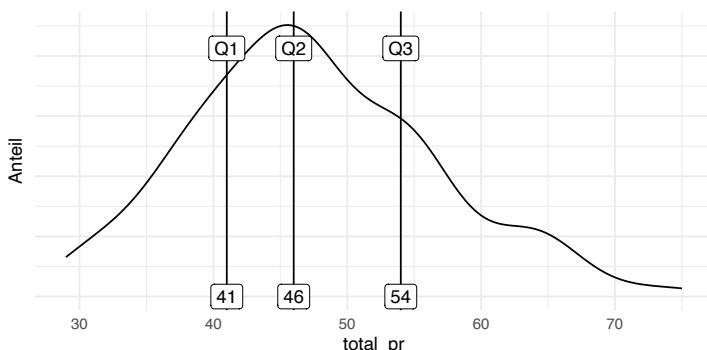


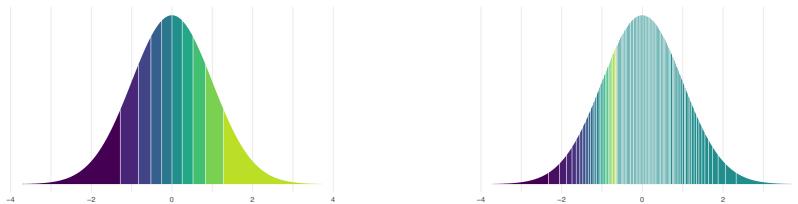
Abbildung 6.16. Q1 und Q3 für das Schlussgebot (nur weniger als 100 Euro) in einem Dichted

Quantile man in R mit `quantile()` holen:

```
mario_quantile <-  
mario %>% t  
filter(total <= 100)  
summa(q1 = n(ttidete215_pr, .),  
q50 = n(ttidete510_pr, .),  
q75 = n(ttidete715_pr, .))
```

Abbildung 6.11 visualisiert verschiedene Qualle Regionen gleichgröße Flächen (d.h. Wahrscheinlichkeiten).

6. Punktmodelle 1



(a) 10%-Schritte: D(ebz) i 1%-Schritte: Perzen

Abbildung 6.11.: Verschiedene Quantile

?

💡 Was ist der Oberbegriff für Median, Mittlerer?

💡 Gute Frage! Wie würden Sie ihn nennen?

Definition (Online-Einheiten): Maß der zentralen Tendenz für eine Verteilung gibt einen Verteilungstypisch, normal, erwartbar ansehen sollten.

Gebräuchliche Lagemaße sind:

- Mittelwert (arithmetisches Mittel)
- Median
- Quantile wie z. B. Quartile
- Minimum (kleinster Wert)
- Maximum (größter Wert)
- Modus (häufigster Wert)

Berechnen wir Lagemaße für den Mario Kart beschriebenen (tmaing okar. tis ist über Sie die Variablen benennen und Stichprobenmeandere in_krasser_vaarli laabs erklaire

Häufig möchte man Statistiken wie Lagemaße für z. B. Mittlere Körpergröße von Frauen vs.

6. 5. L a g e m a ß e

L i s t i n g 6. 2 x z u r B e r e c h n u n g v o n L a g e m a ß e n

```
desciibe_d imatri idk ari tio n
```

```
# E i n i g e S p a l t e n i n t e r e s s i e r e n u n s h i e r n i c h t :  
sel(Sektoren & Kustens, ins_mis sing)
```

Männer n - berechnen und dann vergleichen. Dies steht Efndres chungärmtaeg bauten: „Unterscheidet sich der Mittelwert der Körpergröße von Frauen und Männer.“ Hängt das Geschlecht mit der Körpergröße zusammen gedrückt: Körpergröße ist Funktion des Geschlechtsmodells. Kommt es als solaugruppierte Lagemäße also in Rz. B. so berechnen, s. Listing 6. 3.

L i s t i n g 6. 3 p p i e r t e L a g e m a ß e

```
mariokart_lagemaße_gruppiert <-  
mariokart
```

```
group(wby = %>%) # neue Zeile, der Rest ist gleich!  
summa(mws=mean(otal_pr))
```

Abbildung 6. 13 zeigt ein Beispiel für ungruppierte (rechts) Mittelwerte; vgl. Abbildung 6. 5. Wie sieht, k **re** a **s** i d **a** s **u** m **w** e **d** a **e** r bei einer Gruppierung Vergleich zu einem ungruppierten, „globalen“ der Gruppe ohne Lenkräder und innerhalb der Gruppe sind die Abweichungen zu ihrem Gruppen-Mittel im Vergleich zu den Abweichungen der Preise im Mittelwert.

Definiert man **Modell**, welches für alle Belege eine und denselben Wert annimmt (modell sagt Anders gesagt fasst ein Punktmodell eine Werte eine Tabellenpalte) zu einer einzelnen Zahl in diesem Sinne, s. Abbildung 6. 12.

6. Punktmodelle

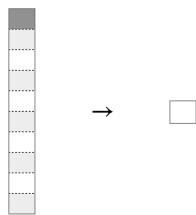
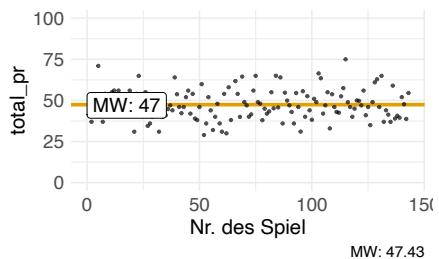
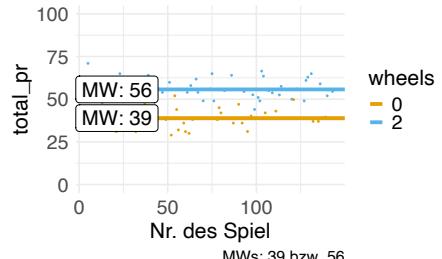


Abbildung 6.12: Kriptive fasst eine Spalte zusammen.

Mittelwert, Median und Quartile sind Beispiele für eine Verteilung zu einzelnen Werten eines „Bildes“ der Daten, machen sie uns verhältnismäßig ein Modell.



(a) ungruppiert



(b) gruppiert

Abbildung 6.13: Die Preis von Mario Kart - Spiele Gerade eingezeichnet. (a) ungruppiert nach Anzahl der Lenkräder.

(\$

Es heißt, mit Statistik könne man vortrefflich Der Grund ist, dass die Statistik Freiheit einen richtigen Weg, um eine statistische Art Wege führen nach Rom (aber nicht alle). Um abzuwehren oder einfache Fehler und Unsicherheiten zu verhindern.

6.7. Vertiefung

aufzudecken, gibt es ein Praktikum mit sensollen transparent sein: Das Vorgehen und Entscheidungen offenlegen. Hier ist eine (nicht vollständige) Liste, was Sie nachprüfen sollten, um die Belegbarkeit zu sichern (Wichert et al. (2016)):

1. Wurde die Art und die Zeitdauer der Datenerhebung legt und berichtet?
2. Wurden ausreichend Daten gesammelt (z. B. mindestens pro Gruppe)?
3. Wurden alle untersuchten Variablen berichtet?
4. Wurden alle durchgeführten Interventionen dokumentiert?
5. Wurden Daten aus der Analyse entfernt? Wenn ja, warum? (stichhaltige) Begründung?

Stellen Sie hohe Anforderungen an die Transparenz und die Dokumentation der Analyse. Nur durch Nachprüfbarkeit können Sie die Gütekriterien der Ergebnisse und deren Interpretationen bestimmen.

Beispiel (Süd. v.aEli-nTei Später): Ein Student in einem Bachelorkurs „Angewandte Medien- und Wirtschaftsprüfung“ hat seine Bachelorarbeit „Surveillance und Data Science in der Survival-Tipps“ für den Bachelorabschluss eingereicht.

1. Wenn man mal nicht weiterkommt, hilft es auch, Abstand von R und Statistik zu nehmen.
2. Es hilft, sich während des Semesters neue Begegnungen zu schreiben.
3. Gut ist auch, sich mit Kommilitonen aus höheren Semestern nach Tipps zu fragen.

💡 Ir g end wie kann ich mir R-Code so schlecht machen?

💡 Frag doch mal Chat GPT, oder einen anderen AI, dort bekommt man auch R-Code ausgegeben.

6. Punktmodelle 1

Übungsaufgabenfragen vom & -Bot in einem Chat-Bot wie ChatGPT nach Übungsaufgaben. genden Prompt orientieren. Empfehlenswerte Prompts zu experimentieren.

Ich bin ein Student in einem Bachelor-Psychologie. Gerade bereite ich mich auf das Fach „Grundlagen der Statistik“ vor. Bitte geben, die mir helfen, mich auf die Prüfung vorzubereiten. Die Fragen sollten folgende Themenbeinhalten: zentralen Tendenz, Grundlagen von R, Skalen (Nominalskala vs. Intervallskala), Verteilungen, z-Werte. Bitte schreibe die richtigen-Falsch-Aufgaben. Schreibe ca.

Ein Teil der folgenden Aufgaben kann Stoff nicht kennen, aber später kennenzulernen. Ich bin (teile) mit (noch) unbekanntem Stoff.

Die Webseite datenwerk.netlify.app stellt Übungsaufgaben bereit. Sie können die Suche nutzen, um die Aufgaben mit den folgenden Namen zu finden:

1. Kennwert - robust
2. mw - berechnen
3. mariokart - max 2
4. nasa01
5. mariokart - mean 1
6. wrangle10
7. summarise01
8. mariokart - max 1
9. Schiefe1
10. mariokart - mean 2
11. summarise03

- 12. mariokart - mean 4
- 13. mariokart - mean 3
- 14. summarise 02

L T i p p

Schauen Sie sich auch mal auf datenwerk.net li zu z. B. dem Tag EDA an.

Ł

Es gibt viele Lehrbücher zu den Grundlagen dieses Kapitels gehören zu den Grundlagen des am einfachsten, wenn Sie einfach in Ihrer Bibliothek ein typisches Lehrbuch schauen. Beispiele: Mittag & Schüller (2020) oder Oestreich & Romberg (2010). Ein Fokus auf Englisch nicht zurückgeschreckt, ist mit Cetinka oder Poldrack (2023) gut beraten. Bei den Büchern: Tipp: Mit dem Browser einfach auf Deutsch über

{

"

In diesem Kapitel benötigen Sie die **Übersichtsleiste** des easy standards Datacenter Økasst Kapitel 3.7.3.

Ł

- Sie kennen gängige Maße der Streuung einer Verteilung.
- Sie können gängige Maße der Streuung einfach berechnen.
- Sie können die Bedeutung von Streuung für die Statistik erläutern.

◦

\$

!

(

Prof. Weiss-Oiss hat eine Erfindung gemacht (faticon, 2024).

Übungsaufgabe 1: Sie die Pille von Prof. Weiss-Oiss nehmen?

- a) ja, ich zahle 1000 Euro
- b) ja



(a) Was er sag^t habe eiⁿ m^as^er N I C H T A l a g e r d i n g s
S c h l a n k h e i t s p i l l e e n t s w i r k e l u e h n d d e W e r t e d e r G e w i c h t e r
p r o E i n n a h m e d a s G e w i c h t ä n n e r u n g u m 1 0 k g u m d e r
S c h n i t t u m 1 k g r e d u z i e M i t t l e r w e r t h e r u m . "

- c) nein
- d) Nur wenn ich 100 Euro bekomme
- e) Okay, für ☐ 000 Euro

Wie sehr die Werte eines Modells streuen, ist bei Prof. Weiss-Oiss' Pillen kaum zu übersehen, dass Sie die Pillen einnehmen.

(fl) .

Treffensich zwei Bauern, Fritz Furchenzieher. Fritz will mit seiner Kuh einen Fluss überqueren, schwimmen (ob es Fritz kann, ist nicht überlie-

👉(Fritz): Sag mal, Karla, ist der Fluss tiefer?

👉(Karla): Nö, im Schnitt nur einen Meter.

Also führt Fritz seine Kuh durch den Fluss, leider an anderer Uferan, da im Flößers offen, s. Abbildung.

👉(Karla): Übrigens: Lagemäße sagen nicht alles.

👉(Fritz): Läuft die Kuh durch den Fluss, kann sie nicht schwimmen.

7. Modellgüte

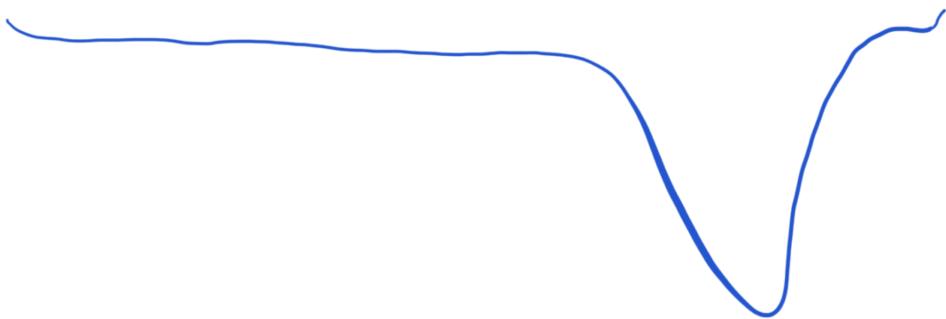


Abbildung 7.2 zeigt, dass im Schnitt nur einen Modell ist die Kuhersoffen.

E Wichtig

Die Streuung ihrer Daten zu kennen, ist eine wichtige Information.

() { }

Abbildung 7.3 zeigt ein einfaches Modell (links) vs. ein einfaches Modell mit Variation (rechts). Die Streuung der Daten ist kleiner als die Streuung der Daten mit den Schätzfehlern. Die vertikalen Abstandsbalken in Abbildung 7.3 kennzeichnen den Abstand eines Datenpunkts zum Mittelwert (horizontal) und die vertikalen Abstandsbalken insgesamt. Die X-Achse zeigt die Versuchspersonen auf.

Bei einem Modell-Symetrie liegt die tatsächliche Abweichung an den Modellwerten (vorhergesagten Werten); die Abweichung endet so gering (der Modellfehler ist klein). Bei einem Modellsymmetrie liegt der Modellfehler groß.

7. 2. Woran erkennt man eine gute

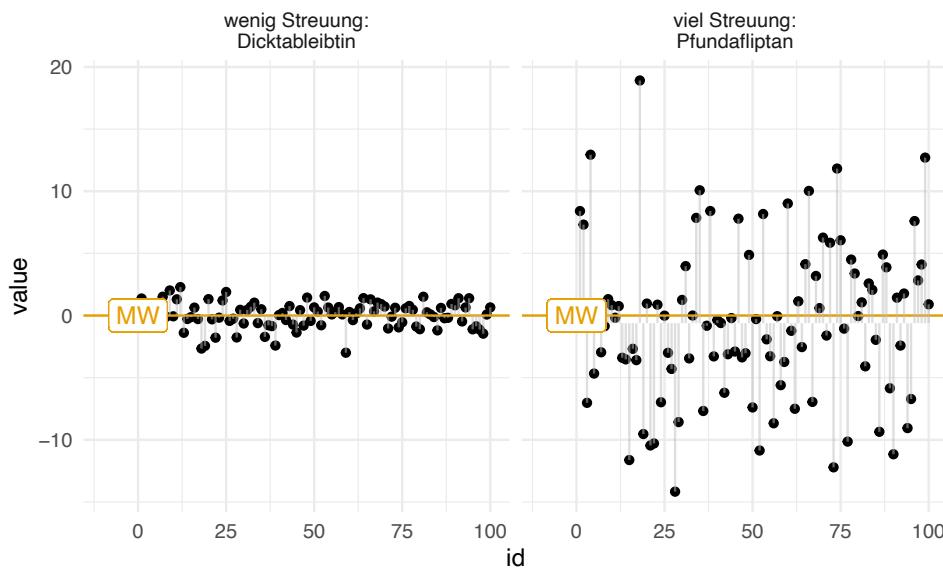


Abbildung 7. 3.: Wenig (links) vs. viel Streuung (rechts)

Bei spielerisch zur Schrankheitskurve von Prof. Weiß-Oiss. Wiedergabe 7. 3 sind die Daten zu der Gewichtsveränderung von „Schrankheitspillen“ zweier verschiedener unterscheidet sich die typische (vorhergesagte) Veränderung zwischen den beiden Präparaten kaum, obwohl es schon Längsschon. Links sieht man die Gewichtsveränderung des Präparats „Dickableibtin extramild“ (ca. von Prof. Weiss-Oiss), „Pfundafliptan Forte“. Wie Sie lieber einnehmen?

Wir wollen ein präzises Modell, also kurze Fehler die Daten gut erklären, also wenig vom tatsächlichen Modell abweichen. Jedes Modell sollte Informationen über die Prädiktoren und der Modellwerte (Vorhersagen) angeben. Ein gutes Modell ist daher der Prädiktor der Schätzungen wenig nützt.

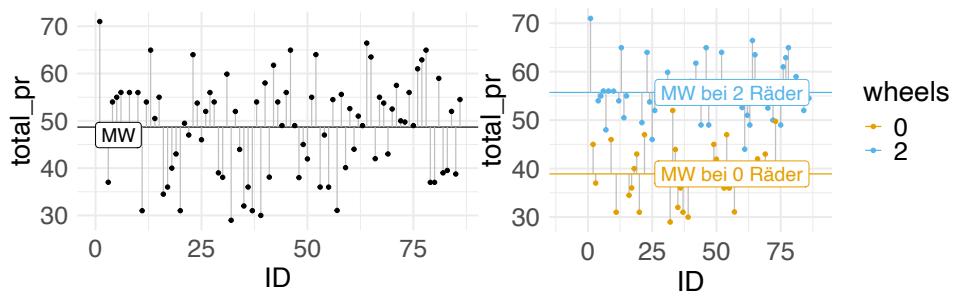
💡 Ich frage mich, ob man so ein Modell nicht kann?

7. Modellgüte

💡 Die Frage ist, was wir mit „verbessern“

💡 Naja, kürzere Fehlerbalken, ist doch k

Im Beispiel von Mario Kart: Da die Anzahl Verkaufspreis zusammenhängt, könnte es vi Lenkräder-Anzahl da irgendwie nutzen zu sollten wir ausprobieren. Abbildung 7.4 zu kürzen werden, wenn wir ein (sinnvolles) kompl Innerhalb jeder der beiden Gruppen (mit 2 Lenkrädern) sind die Fehlerbalken jeweili (rechtes Teildiagramm) als im Modell ohne Teildiagramm). Aus Gründen der Übersichtlichkeit mit Verkaufsgebot von weniger als 100 Euro Spielen mit 0 oder mit 2 Lenkrädern.



(a) Einfaches Modell (b) Komplexeres Modell

Abbildung 7.4 Fehlerbalken in einem einfachen und komplexen Modell. (a) Fehlerbalken in einem einfachen Modell; viel Streuung insgesamt, $y \sim 1$. (b) Fehlerbalken in einem komplexen Modell. Zwei Mittelwerte in jeder Gruppe, $y \sim G$. Das erkennt die vertikalen, grauen Abstandsbalken als einfache Modell.

E Wichtig

Durch sinnvolle, komplexere Modelle sinkt die Variabilität eines Modells.

\$

Definition der Streuungsmaße) Streuungsmaß quantifiziert die Variabilität (Unterschiedlichkeit, Streuung)

Definition (Spannweite) Die Spannweite ist die Differenz von größtem und kleinstem Wert einer Variable X:

Beispiel angenommen, wir haben einen Datensatz „Alter“ mit den Werten 1, 23, 42, 100. Dann beträgt die Spannweite 99, was bedeutet, dass die Werte des Merkmals über 99 Einheiten (Jahre in diesem Fall) verteilen.

Die Spannweite ist aber nicht robust (gegenüber Ausreißern), sollte daher nur mit Einschränkung verwendet werden.

.

💡 Wir müssen jetzt mal präziser werden! Wie berechnen die Streuung?

💡 Gute Frage! Am einfachsten ist es, wenn wir die Länge eines Abweichungsbalkens ausrechnen.

Legen wir (gedanklich) alle Abweichungen ab, so erhalten wir den „Abweichungsbalken“, dessen Längen könnten. Diesen Koeffizienten nennen wir **Absolutabweichung**. Siehe Absolutabweichung (Abs.). Gleichung 7.1.

7. Modellgüte

Definition: Die absolute Abweichung (MAA, MAE) ist definiert als die Summe der Differenzen eines Messwerts zum Mittelwert der Messwerte. (Wenn man solche Sätze liest, einfacher an.)

$$MAE = \frac{|y_1 - \bar{y}| + |y_2 - \bar{y}| + \dots + |y_n - \bar{y}|}{n} = \frac{\sum_{i=1}^n |y_i - \bar{y}|}{n} = \square. \quad (7.1)$$

Beispiel Abbildung 7.5 visualisiert eine einfache Rechnung, in der wir den MAE für das Beispiel von Abbildung 7.4 berechnen.

$$MAE = \frac{|1 - 1| + |3 - 1| + |6 - 1| + |1 - 1|}{4} = \frac{1 + 2 + 5 + 0}{4} = 1.5$$

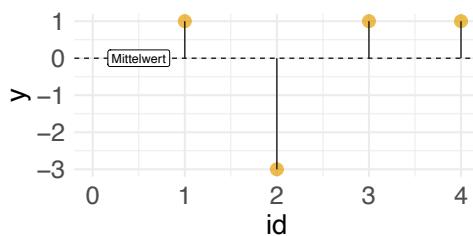


Abbildung 7.5.: Abweichungsbalkenurkunde

Natürlich können wir auch die Rechenarbeit ausarbeiten:

👉 Loving it!!

Schauen Sie: Den Mittelwert (s. Abbildung 7.4) und Rechtlinearität einer Modellgeraden, betrachten. Schließlich kann die Gerade (die parallel verläuft). In R gibt es einen Maßbefehl `abmd` hnen er heißt `Die Syntax` `abmd(y ~ 1, data = meine_da`

In Worten:

Hey R, berechne mir ein lineares Modell zu `Aber verwenden keine andere Variable zur Sonderung` nimmt den Mittelwert von Y.

7.3. Streuungsmaße

```
Im1 |<(y 1, d a t d )=
```

Den MAE können wir uns jetzt so ausgeben lassen

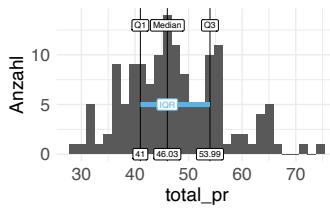
```
ma@Im# aus dem Paket easystats  
## [1] 1.5
```

Der Interquartilsabstand (IQA; engl. interquartile range) ist ein Maß der Streuung, das nicht auf dem Mittelwert aber auch nicht auf der Varianz oder der Standardabweichung basiert. Es stellt den Abstand zwischen den ersten und dritten Quartilen dar.

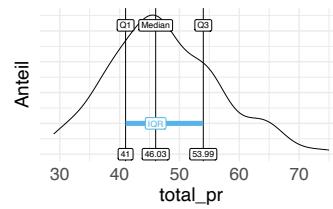
Der Interquartilsabstand (IQA; engl. interquartile range) ist ein Maß der Streuung, das nicht auf dem Mittelwert aber auch nicht auf der Varianz oder der Standardabweichung basiert. Es stellt den Abstand zwischen den ersten und dritten Quartilen dar.

Definiert als der die (absolute) Differenz vom 3. - 1. Quartil:

Bei spielsweise im Hörsaalheim Statistikkurs beträgt der Körpergröße: Q1: 1.65 m, Q2 (Median): 1.75 m, Q3: 1.80 m. Der IQR beträgt dann $Q_3 - Q_1 = 1.80 - 1.65 = 0.15$ m, d.h. 15 cm.



(a) Histogramm



(b) Dichtediagramm

Abbildung Q1; Q2 und Q3 für das Schlussgebot weniger als 100 Euro)

7. Modellgüte

\$ \check{Z}

Normalverteilungen sind recht häufig anzutreffen in der Datenanalyse. Daher lohnt es sich, zu überlappenden Verteilungen kompakt zusammenfassend. Man kann zentralen Wert und Streuung einer Verteilung sich kompakt beschreiben (Lyon, 2014). Außerdem ist eine Verteilung normalverteilt, wenn man eine Konstante zu (bzw. subtrahiert), „verschiebt“ man den Bereich der Verteilung, s. Abbildung 7.7.

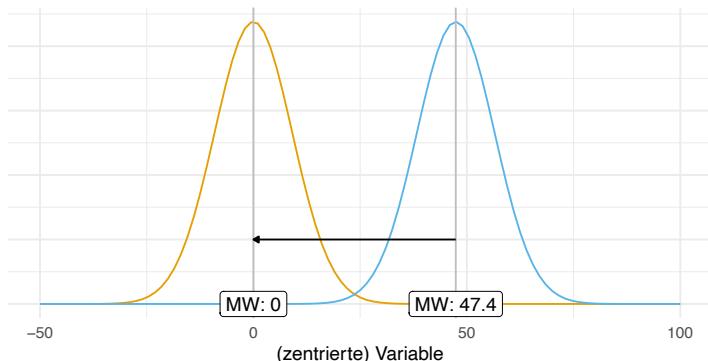


Abbildung 7.7 zeigt, dass Abweichungen zum Mittelwert (Mittelwerte) von zentrierten Variablen sind selber eine unzentrierte Verteilung; links: zentriert.

Hat man normalverteilte Variablen abgesehen von möglichen Ausreißern, so kann man die Standardabweichung s als Maßeinheit verwenden, um die Streuung (Abweichungen) der Normalverteilung gut beschreiben.

💡 Aber wie berechnet man jetzt diese Standardabweichung?

💡 Moment, noch ein kurzer Exkurs zur Variabilität (siehe unten)

Die Varianz einer Variable (z. B. Verkaufspreise) ist der mittlere quadrierte Abstand jedes Verkaufspreises von dem Mittelwert.

Abbildung 7.10 illustriert den "mittleren Quadratfehler".

1. Man geht von der Häufung der Daten aus.

2. Betrachtet man die Daten auf einer Wippe, so sind die Punkte der Wippe der

3. Man zeichnet für jeden Abstand ein Quadrat mit einem die seineinem Abstand entspricht.

4. Diese Quadrate werden Rechtecke umgeformt (oder Fläche) und dass sie ein Rechteck im Quadrat bilden.

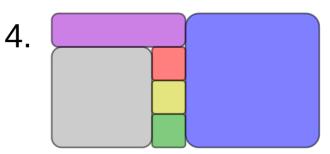
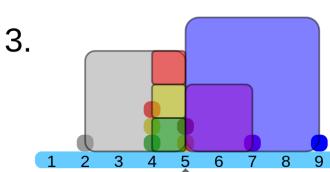
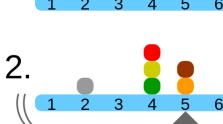


Abbildung 7.9 visualisiert die Abweichungsquadrate, rechteckweise abgebildet, als Abweichungsquadrat (Caglee, 2015)

Abbildung 7.9 visualisiert die Abweichungsquadrate, rechteckweise abgebildet, als Abweichungsquadrat.

¹Die Abweichungsquadrate wirken optisch nicht quadratisch skaliert dargestellt ist als die Y-Achse. Trotzdem sind wenn Sie wissen, was ich meine ...

7. Modellgüte

I Hinweis

Die Varianz ist also ein Maß, das die typischen Beobachtungen vom Mittelwert in einer Zahl

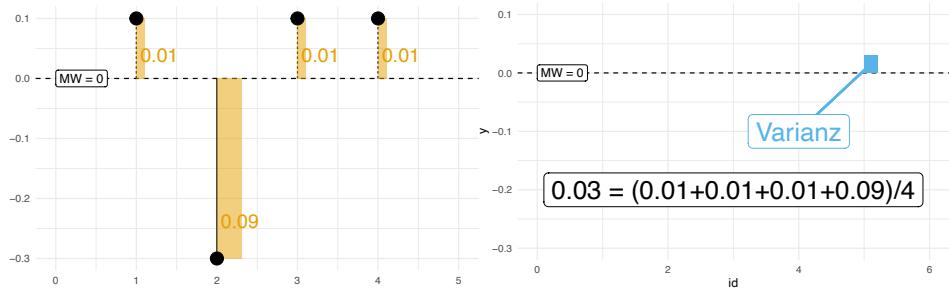


Abbildung 7.9.: Sinnbild zur Varianz als

Bei spielsarbeiten immer noch bei einem Online und untersuchenden Verkauf von Videospielen, dass Ihre Firma mehr von dem Zeug verkaufen. Sie die Streuung in den Verkaufspreisen, s.

Tabelle 7.1.: Kennwerte der Streuung für d

pr_mw	pr_iqr	pr_maa	pr_var	pr_sds
47	13	7.2	83	9.1

Statistiken sind ja schön ... aber Bilder sind Datendiagramme eignen sich gut, um (grob) zu erfassen.

```
mario%>%t
mario %>%
  select(-extremeExtremwerte)
plot_density() # oder plot_violin
```

Listing 7.1: Berechnung der Streuung des Verkaufspreises für die Modellgüte des Mittelwerts

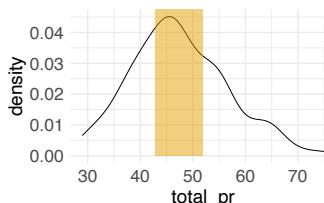
```
mariokart_no_extreme <-
  mariokart %>%
  filter(total_pr >= 10) # rohne Extremwerte

m_summ <-
  mariokart %>% extreme
  summa(ise
    pr_mean(total_pr),
    pr_ilqr(total_pr),
    pr_mad(ambient_anomaly(total_pr))),
    pr_var(total_pr),
    pr_ssd(total_pr))
```

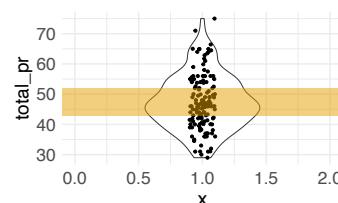
Wer sich die Berechnung von Maßzahlen für die Modellgüte des Mittelwerts (s. Listing 7.1), kann Median und Standardabweichung ausrechnen.

Um die Standardabweichung zu berechnen, berechnet man die Varianz. Hier ist ein „Kochrezept“ Berechnung der Varianz:

- Für alle Datenpunkte berechnet man die Abweichungen vom Mittelwert,



(a) Dichtediagramm



(b) Violindiagramm

Abbildung 7.1: Verteilung des Verkaufspreises. Spielen mit $MW \pm SD$ farblich markiert

7. Modellgüte

2. Quadriere diese Werte.
3. Summiere dann auf.
4. Teile durch die Anzahl der Werte.

Als Formel ausgedrückt lautet die Definition der Gwößfolgt, s. Gleichung 7.2. (Hingegen ist es die Stichprobenvarianz; um Varianz der zugehörigen Population zu schätzen, sondern durch)

$$\text{Var}^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2 = \frac{1}{n-1} \sum y_i^2 - \bar{y}^2. \quad (7.2)$$

Die Varianz ist definiert als der Mittelwert der quadrierten Abweichungen (vom Mittelwert).

Die Varianz steht im engen Verhältnis zur Korrelation. Die Varianz kann auch vom Mittelwert abweichen (Mean Squared Error, MSE) eines Modells, s. Gleichung 7.3.

$$\text{Var}^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2. \quad (7.3)$$

Im Fall eines Punktmodells ist der Mittelwert eines Modells.

$$\bar{y} = \frac{1}{n} \sum y_i$$

Die Standardabweichung ist definiert als die Quadratwurzel der Varianz.

$$s = \sqrt{\frac{1}{n-1} \sum (y_i - \bar{y})^2} \quad (7.4)$$

7.3. Streuungsmaße

Kennt man die Varianz, so lässt sich die Standardabweichung leicht aus der Größe der Varianz berechnen. Durch das Quadratwurzelziehen erhält man eine Größe, die die Daten (im Gegensatz zur Varianz, die durch die Summe der Abweichungen von der Mittelgröße addiert werden kann). Aus einem Modellierungsblatt definiert man als die Wurzel von MSE den sogenannten Squared Error (MSE):

| Hinweis

Die SD ist nicht direkt zur MAE, aber (fast) gleich groß. Entsprechend ist die Varianz (fast) gleich zu

Beispiel: Arbeiten weiter an Ihrem Mario Kart und schätzen Sie heute keine Lust auf viel Tippen heraus, trauten Sie sich mit der Fehlerstreuung der tatsächlichen Werte? Tabelle 7.2.

| library(sytest)

```
library(sytest)
sel(etc@t@l>_%pr)
describe_distribution
```

Tabelle 7.2.: Ausgabe des Fehlerstreuungstests

Variable	Mean	SD	IQR	n
total_pr	50	26	13	143

Wow! Das war einfach. Reicht auch mal für heute.

Beispiel 7.2: Als Datenanalyst ist es anstrengend interessant. So war auch der heutige Tag. Bevor möchten Sie noch eine Sache anschauen. In einer Abbildung 7.4 fanden Sie heraus, dass die Fehler wenn man ein geschickteres und komplexeres Modell

7. Modellgüte

Sie natürlich prüfen. Sie überlegen: „Okay, in dem der Mittelwert das Modell des Verkaufspreises ist.“ Das spezifizieren Sie so:

```
l m1<(tota1,doptmariokart)
ma€l m1)Modellgüte bzw. Modellfehler
# # [ 1 ] 10
```

Im nächsten Schritt spezifizieren Sie ein Modell, das die Abhängigkeit des Verkaufspreises von der Anzahl der Lenkräder darstellt (7.4):

```
l m2<(tota1,whpedast,mariokart)
ma€l m2)
# # [ 1 ] 7. 4
```

Ah! Sehr schön, Sie haben mit diesem Modell also den Mittelwert gefunden. Ab nach Hause!

\$

{

Wenn wir den Mittelwert als Punktmodell darstellen, so kann man die verschiedenen Kennwerte des Modells auf verschiedene Weise ermitteln. Definieren wir zunächst als Punktmodell auf:

```
l m_mari(tota1,doptmariokart)
```

Zur Erinnerung: Wir modellieren die Preärdeinkommen, so als Punktmodell, und zwar schätzen wir den Modellfehler.

Das (Metamodell) **predict komfortable Befehle, gute zu berechnen:**

7.5. Die z-Transformation

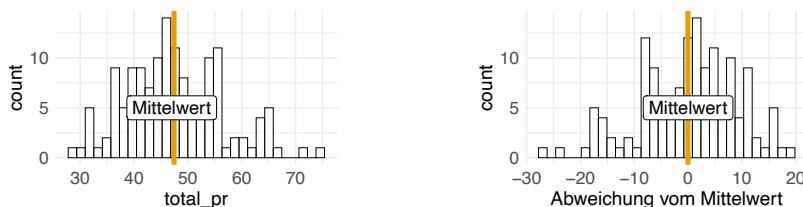
```
ma <- lm_ma # iMö 1) n absolute error
ms <- lm_ma # iMö 1) n squared error
rms <- lm_ma # iRö b) t mean squared error
## [1] 10
## [1] 655
## [1] 26
```

◦ %

Sie arbeiten immer noch als Data Miner bei einem kleinen Online-Auktionshaus. Heute untersuchen Sie, wie preise mit einer einzigen Zahl, dem mittleren benlassen. Einige Widersprüche haben die Datensätze verweise. Schon mächtig interessant sind extreme

```
mario_kart_no_extreme <-
mario %>% t
filter(totals == 0)
```

Abbildung 7.11 (links) zeigt, dass es einige Scherum gibt. Abbildung 7.11 (rechts) zeigt die reidaten.



(am) eine Abdrängung sichtbar bei weiten Abweichungen vom Mittelwert: preise um ihren Mittelwert Daten

Abbildung 7.11. mario rotkärrl um vertreeme

7. Modellgüte

Tja, das ist doch etwas Streuung um den Mittelwert.

E wichtig

Je weniger Streuung um den Mittelwert (ca. besser eignet sich der Mittelwert als Modell), desto höher ist die Modellgüte.

Ja, es ist ~~sie~~ Streuung, aber wie viel? Kann man das ausrechnen? Sie überlegen ... und überlegen. Da! Ein Index!

Man könnte vielleicht angeben, wie viel Einfluss der Wert entfernt ist. Je größer diese Abweichung, desto schlechter die Modellgüte! Also rechnen Sie diese Abweichung aus.

Liste mit Zielen einer Variablen

```
mariokart_no_extreme <-  
mariokart_no_extreme  
mut atbew4=7 -t4otal_pr)
```

Anders gesagt: Wir haben unterschiedliche Aufpreise für verschiedene Modelle.

Definition: Zentrale Merkmale sind bedeutet, von jeder Verteilung abweichen. Der Mittelwert zu subtrahieren. Der Mittelwert (der zentrierte Verteilung) gilt.

Aber irgendwie sind Sie noch nicht am Ziel Ihres Betriebs gewesen, ob 10 Euro oder 20 Euro „viel“ Abweichung ist? Man müsste die Abweichung eines Verkaufs in Bezug setzen. Wieder! Ein Geistesblitz! Die Abweichung in Bezug auf den absoluten Mittelwert (MAA)! Ein alternativer, ähnlicher Kennwert hat gehörig, dass die SD gebräuchlicher ist. Checker zu präsentieren, berechnen Sie also Koeffizienten und ja ähnlich.

7.5. Die z-Transformation

Also: Wenn ein Spiel 10 Euro vom Mittelwert ab Euro betragen sollte, dann hätten wir eine „st manchmaß ~~mit~~ Abweichung von 1, weil $10/10 = 1$. Bege Geistesblitze machen Sie sich ans Werk.

```
mariokart_no_extreme <-
mariokart_no%extreme
mut atbew_sabws(abw)s,t d wie "standardisiert"
abw_sabwameamb(sa b w ) )
```

Zufrieden betrachten Sie Ihr Werk, s. Abbildung. Sieht man oben die Rohwerte und unten die transformierten, die wir hier erst als „Standardisierten“ darstellen, da wir sie in Bezug auf die typischen Abweichungen“, der SD, gesetzt haben.

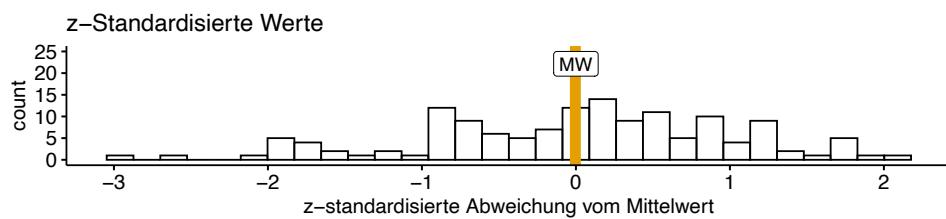
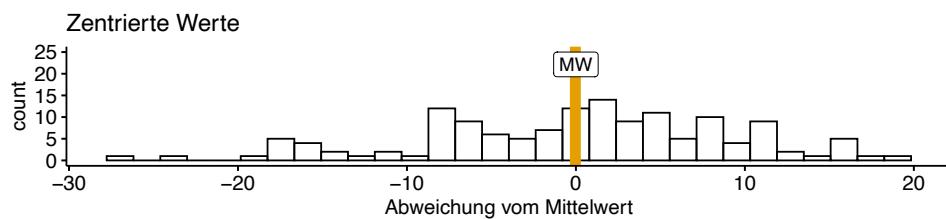


Abbildung Standardisierung von Abweichungswertverteilung; der vertikale Balken zeigt

Wir fassen die Schritte unserer Umrechnung („Trumen“ wie in einem Kochrezept:

7. Modellgüte

1. Nimm die Verteilung der Verkaufspreise
2. Berechne die Abweichungen vom mittleren
Mittelwert und jeweiliger Verkaufssumme
3. Teile die Abweichungen (Schritt 2) durch

Diese Art von Transformation basiert im Wesentlichen auf der resultierenden Werte als

Definitionswerte sind das Resultat der zentralen Tendenz. Für die Variablen kann sich der zentrale Wert so bestimmen:

z-Werte sind nützlich, weil sie die „relative Beobachtungen vom Mittelwert“ anzeigen. Nach man von extremen Abweichungen (Extremwerte).
2.(Shimizu, 2022).

Der „gesunde Menschenverstand“ würde spontan den Abstand (MAA oder MAE) der Varianz (oder SD) vorziehen. Das ist vernünftig, denn die Varianz ist nützlicher als die Varianz und die Standardabweichung. Das ist nützlich, denn die Varianz ist nur um deskriptive Statistik geht, braucht nicht unbedingt. Gründe, warum Sie die Varianz nutzen sollten, sind:

- Die SD ist nützlich zur Beschreibung der Varianz.
- Die Varianz wird häufig verwendet bzw. im Berichterstattung, also müssen Sie die Varianz kennen.

Liegen Extremwerte vor, kann es vorteilhaft gegenüber Mittelwert basierten Streuungsmaßen sein.

7.7. Aufgaben

Die Webseite datenwerk.netlify.app stellt eine Übungsaufgaben bereit. Sie können die Suchfunktion nutzen, um die Aufgaben mit den folgenden Namen zu finden:

- mariokart-sd2
- mariokart-sd3
- Kennwert-robust
- summarise04
- summarise05
- vis-mariokart-variab
- sd-vergleich
- nasa01
- Streuung-Histogramm
- mariokart-sd1
- summarise06
- mariokart-desk01

Ł

Allan Downey (2023) stellt in seinem vergnüglichen und kurzweiligen Einführung in die Statistik vor; dabei einen Auftritt. Wer mehr „Lehrbuch-Fakten“ will, kann Cetinkaya-Rundel & Hardin (2021) fündig (dafür verfügbar). Es ist kein Geheimnis, dass Streuung und Correlation die wichtigsten Themen in der Statistik sind. Aber hey, Oldie ist Oldie und es geht's nicht. Jedenfalls werden Sie einiges an Lernmaterial finden, das Sie in der Bib (oder sonst wo) aus dem Regal zu diesem Thema. Die Bücher unterscheiden sich in ihrem Anspruch bzw. der didaktischen Aufmachung; für

!

"

In diesem Kapitel benötigen Sie die **übliche** easy **Statistik** Dantaen Økaszt Kapitel 3.7.3.

Ł

- Sie können die Begriffe Kovarianz und Korrelationen Zuhren Zusammenhang erläutern.
- Sie können die Stärke einer Korrelation

+ "

Übungsaufgabe 8.1.

1 Suchen Sie sich eine vertrauenswürdige vertrauenswürdigen Partner. Im Zweifel neben Ihnen sitzt.

2. Nennen Sie je zwei Variablen, die wie folgt zusammenhängen:

- gleichsinnig (Viel von dem einen, viel von dem anderen)
- gegensinnig (viel von dem einen, wenig von dem anderen)
- Scheinzusammenhang (hängt zusammen, ist kausal)

8. 2. Zusammenfassen zum Zusammenfassen

+

+

In Kapitel 6 haben wir gelernt, dass das Wesen als Zusammenhang (eines Vektors) zu einer Zahl, zu einem „Punkt“ sozusagen, zusammengefasst wird. In diesem Kapitel zw. Spalten zusammenfassen, wiederum zu s. Abbildung 8. 1. Während wir in Kapitel 6 eine Lagemäßigkeit beschrieben (bzw. dargestellt, visualisiert) haben, tun wir hier das Gleiche für zwei metrische Variablen, so geht es um die Frage, ob sie miteinander zusammenhängen: Wie die beiden Variablen miteinander (in welcher Form) zusammenhängen bzw. welche Abhängigkeiten bestehen.

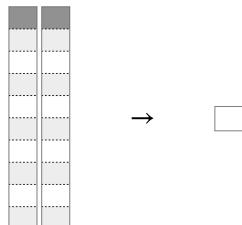


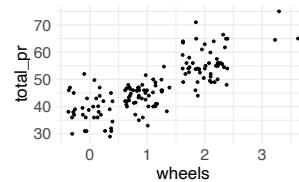
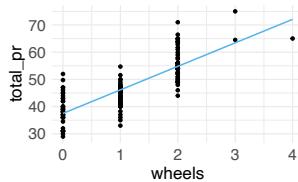
Abbildung 8. 1.: Zwei Spalten werden zu einer Z

Beispiele für Zusammenhänge)

- Lernzeit und Klausurerfolg
- Körpergröße und Schuhgröße
- Verbrauchtes Benzin und zurückgelegte Strecke
- Produktionsmenge und Produktionskosten
- Bildschirmzeit und Schlafqualität
- Umweltschutz und Biodiversivität

Die Verbilldlichung (Visualisierung) zweier metrischer Variablen haben wir bereits in Kapitel 5. 5. 2 kennengelernt. Zusammenhang zweier metrischer Variablen ausspielen kann man Abbildung 8. 2.

8. Punktmodelle 2



(a) Streudiagramm mit Trennwackeltes Streudiagramm

Abbildungsschlüssierung des Zusammenhangs von total_pr. (a) Streudiagramm mit Trennwackel zur Verdeckung. (b), Verwackeln und die einzelnen Punkte besser zu erkennen.

Die Stärke des linearen Zusammenhangs zweier Variablen kann man gut mithilfe von Abweichungsrechnung ablesen!

z

Beispiel (Wieder Statistikkenntner), Carl und Daniel ihre Statistikklasse zurückkämpfen. Die erreichten Punktzahlen je mehr desto besser) zugehören nicht so schlecht ausgefallen wie gestern. □

Tabelle Punktzahlen in der Statistikklasse (x, 0 - 100)

i	d	y	x
1	72	70	
2	44	40	
3	39	35	

¹Typisches Lehrerbeispiel!!

8. 3. Abweichungsrechtecke

T a b e l l e P e u n k t e i n d e r S t a t i s t i k k l a u s u r (x , 0 - 1 0 0)

i	d	y	x
4	5	0	67

Zeichnen wir uns die Daten als Streudiagramm, zeichnen wir Abweichungsserienrechtecke

Definition Abweichungsrechtecke imdimensionalen Flächenraum spannt sich ein Abweichungsrechteck vom Mittelpunkt aus und genausowürde es eine Messwertdistanz (Abweichung) vom Mittelpunkt der Messwerte betrachten also = $-$. Die Fläche des Abweichungsrechtecks ist Produkt der Abweichungen:

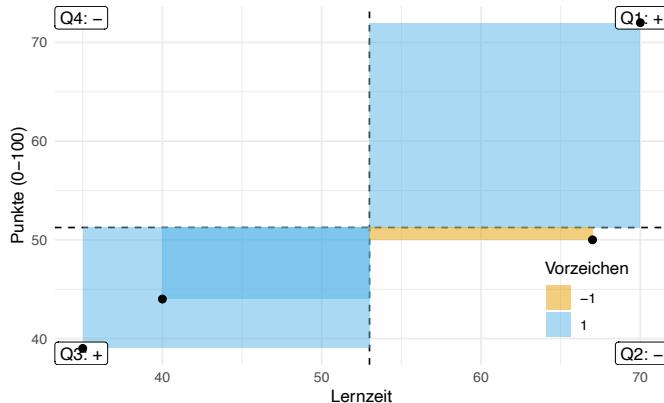
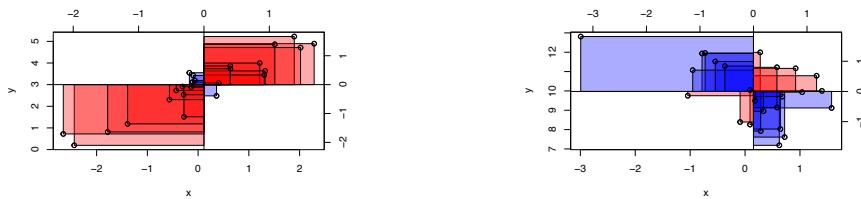


Abbildung zeigt Kovarianz als mittleres Abweichung der vier Quadranten (Q_1, Q_2, Q_3, Q_4) der Abweichungsrechtecke dargestellt. Der Abweichungsrechteckespiegelnd der.

Stellen Sie sich vor, wir legen alle Rechtecke 8. 3. Nennen wir das resultierende Rechteck "Ja, ich weiß, ich strapaziere mal wieder".

8. Punktmodelle 2

kommt's: Je größer die Fläche des Summenrechtecks (lineare) Zusammenhang. Beachten Sie, welche der vier Quadranten sie stehen. Die verdeutlichen dies, s. Abbildung 8.4. Links eine positive Summe, rechts eine negative Summe. Man sieht im rechten Diagramm, dass das umgekehrte Vorzeichen im rechten Diagramm ist.



(a) Positive Vorzeichen (positive Vorzeichen (Quadranten 1 und 3) überwiegen, was in einer positiven Kovarianz resultiert - Negative Vorzeichen (Quadranten 2 und 4) übertreffen die negativen, was in einer negativen Kovarianz resultiert)

Abbildung 8.4.: Positive und negative Kovarianz, positive Summe, einmal eine negative Fläche der Abweichungsrechtecke

Wir können das Summenrechteck noch durch Punkte teilen, das ändert nichts an der Aussage gegenüber der Summe den Vorteil, dass er Aussage von der Anzahl der eingegangenen Datenrechtecke nicht hat. Wir können für den Zusammenhang von Lernzeit und Klausurfläche des mittleren Abweichungsrechtecks:

8. 3. Abweichungsrechnung

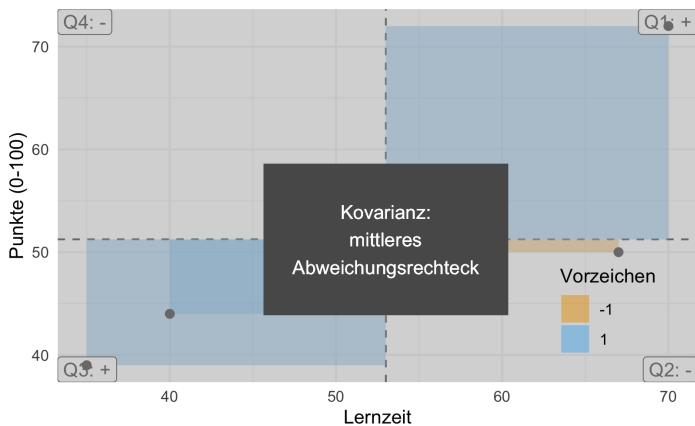


Abbildung 8.5 Kovarianz als mittleres Abweichungsrechteck entspricht dem Wert der Kovarianz.

fl

Definition (Kno8.2): Kovarianz ist definiert als die mittleren Abweichungsrechtecke. Sie ist eine Richtung des linearen Zusammenhangs zweier merkmale.

💡 Zu viele Bilder! Ich brauch Zahlen.

💡 Kommen gleich!

Tabelle 8.2 zeigt beispielhaft, wie sich die Kovarianz wirkt. Nächstes das mittlere Abweichungsrechteck für die Noten und Lernzeit der vier Studierenden beträgt 162.

8. Punktmodelle 2

Tabelle Wertzeile der Abweichungsrechtecke. av (Mittelwert), cov_sign: Vorzeichen der Kovarianz, Wert auf der entsprechenden Achse (Produkt von x_delta und y_delta)

i	d	y	x	x_avg	y_avg	x_delta	y_delta	c
1	72	70	53	51	17	20.	8	1
2	44	40	53	51	-13	-7.	2	1
3	39	35	53	51	-18	-12.	2	1
4	50	67	53	51	14	-1.	2	-1

d % > %

summa (kioska r imena(nx=y_area))

$$\frac{\text{kovarianz}}{162}$$

Die Formel der Kovarianz lautet, s. Gleichung 8.1:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (8.1)$$

Gleichung 8.1 in Worten ausgedrückt:

1. Rechne für jede Abweichung vom Mittelwert
2. Rechne für jede Abweichung vom Mittelwert
3. Multipliziere, auf alle die Abweichungsrechtecke zu erhalten.
4. Addiere die Flächen der Abweichungsrechtecke
5. Teile durch die Anzahl der Beobachtung

Beispiel (Variablen mit positiver Kovarianz)

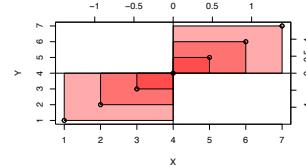
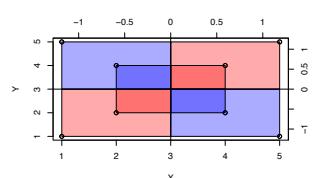
8. 3. Abweichungsrechtecke

- Größe und Gewicht
- Lernzeit und Klausurerfolg
- Distanz zum Ziel und Reisezeit
- Temperatur und Eisverkauf

Beispiele (Variablen mit negativer Kovarianz)

- Lernzeit und Freizeit
- Alter und Restlebenszeit
- Temperatur und Schneemenge
- Lebenszufriedenheit und Depressivität

Zwei Extrembeispiele für Kovarianz - Werte sind stellt.



(a) kein Zusammenhang (b) perfekter (positiver) Zusammenhang

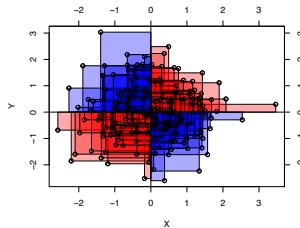
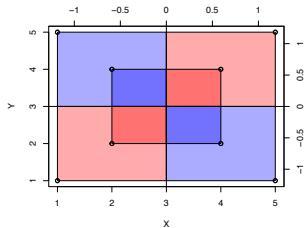
Abbildung 8.6.: Verschiedene Werte der Kovarianz

Bei einer Kovarianz von (ungefähr) Null ist die Abweichungsrechtecke, wenn man es symmetrisch gemacht hat, ungefähr gleich groß, s. Abbildung 8.7. Zur Erinnerung: Quadrate; bei der Kovarianz sind es jetzt Rechtecke (unter Beachtung der Vorzeichen); blau = negativ), so beträgt die Summe in etwa (ca. 10).

Damit ist die Kovarianz in diesem Fall etwa (betragsmäßig) Null. Wenn die Summe der Abweichungsrechtecke ist auch ihr Mittelwert (MW) Null. Damit ist die Kovarianz Null.

8. Punktmodelle 2

$$\begin{aligned}
 & () = 0 \\
 \text{MW}() & = 0 \quad (8.2) \\
 \text{co}(v,) & = 0
 \end{aligned}$$



(a) 4 Abweichungsrechtecke 200 Abweichungsrechtecke

Abbildung zeigt, dass die Kovarianz gleichensinnig rechtecke aufweist; d.h., ihre Flächeninhalt ist proportional zu den Abweichungen.

fl

Die Kovarianz hat den Nachteil, dass sie abhängt von den Einheiten der Variablen. So steigt die Kovarianz z.B. um den Faktor 100, wenn ein Einkommen von 100 Euro auf 1000 Euro steigt. Der Zusammenhang zwischen Einkommen und Lebenszufriedenheit ist unabhängig von den Einheiten. Außerdem ist es schwierig, den Maximalwert, der einen perfekten Zusammenhang darstellt, zu interpretieren. Die Kovarianz ist daher nur wenig verwendete Anwendung.

fl

fl

!

Der Korrelationskoeffizient Pearson I löst das Problem der Kovarianz schwer interpretierbar ist. Der Wert von -1 (perfekte negative lineare Korrelation) und +1 (lineare Korrelation). Er deutet den Zusammenhang.

Die Korrelation berechnet sich wie folgt:

1. Teilt man die Standardabweichung,
2. Teilt man die Standardabweichung,
3. Berechne mit diesen Werten die Kovarianz

Teilt man die Standardabweichung, führt man zu einer z-Transformation durch. Daraus erhält man den Korrelationskoeffizienten wie in Definition 8.

Der Korrelationskoeffizient (nach Pearson) ist definiert als das mittlere Produkt der Kovarianz zweier metrischer Variablen. [-1; +1] bei keinem Zusammenhang, 0 bei einem starken Zusammenhang.

$$= \frac{1}{n} \quad (8.3)$$

Man beachte, dass eine Korrelation (genauso wie für metrische Variablen definiert ist). Aus dem Koeffizienten kann man zwei Informationen ableiten:

1. Vorzeichen: Ein positives Vorzeichen bedeutet positiver linearer Zusammenhang (und umgekehrt), ein negatives, also negativer linearer Zusammenhang.

8. Punktmodelle 2

2. Absolutere Korrelation: Der Absolutwert der Korrelationskoeffizienten gibt die Stärke des Zusammenhangs an. Je näher der Wert bei 1 liegt, desto starker ist der Zusammenhang.

- $= 0$: Kein linearer Zusammenhang
- $= 1$: Perfekter linearer Zusammenhang

Eine Zuordnung des Korrelationskoeffizientengramms zeigt Abbildung 8.8.

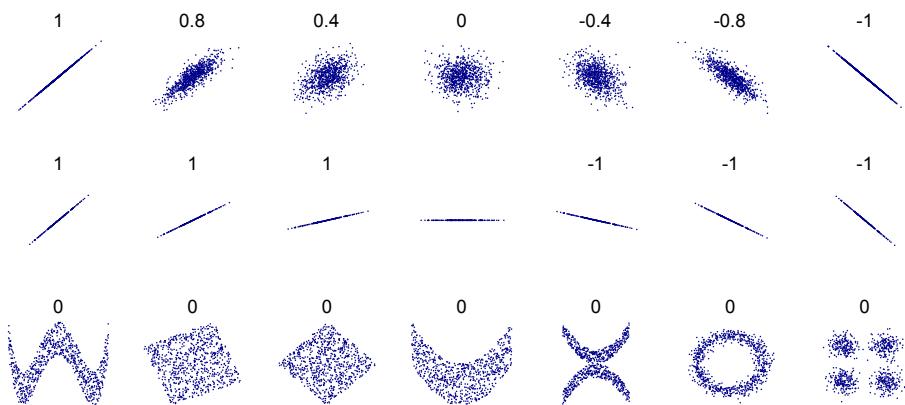


Abbildung 8.8: Sieben Streudiagramme, die verschiedene Korrelationskoeffizienten unterscheiden.

Die untere Zeile von Abbildung 8.8 zeigt Beispiele für schwache Zusammenhänge. Wie man sieht, liegt in diesen Fällen kein eindeutiger Zusammenhang bestehen.

fl

#

Ob der Verkauf von Insekticiden (mit der Dauerdrucktaktik oder mit der Anzahl der Besuchte zusammenhängt) wissen wir nach! Diese Funktion ist ohne mehr als 5 Minuten Rechnen für uns, s. Tabelle 8.4.

```
mario|kcart
sel(etc@tal_pr, duration, n_bids)
correl)@#oanu$ `easystats'
summary
```

Tabelle 1 Korrelation berechnen mit `easystats`

Parameter	n_bids	duration
total_pr	0.13	-0.04
duration	-0.12	

Sie können auch auf die letzten Zeilen zurückgreifen. Dann ist die Ausgabe ausführlicher.

fl fl

Eine Studie fand eine starke Korrelation zwischen dem Konsum eines Landes und (Anzahl der) Messer (Messerli, 2012), s. Abbildung 8.9.

Korrelation (bzw. Zusammenhang) ist ungleichbedeutend, dass eine Kausation vorliegt, sein, dass Kausation vorliegt. Liegt Korrelation vor, spricht man von einer Scheinkorrelation.

fl +

Beispiel einer Korrelation: Es ist Mythen, dass Babys in diesem Landkreis (vgl. Matthews, 2000)

Eine mögliche Erklärung für dieses (nur scheinbare) „Naturbelassenheit“ des Landkreises liegt

8. Punktmodelle 2

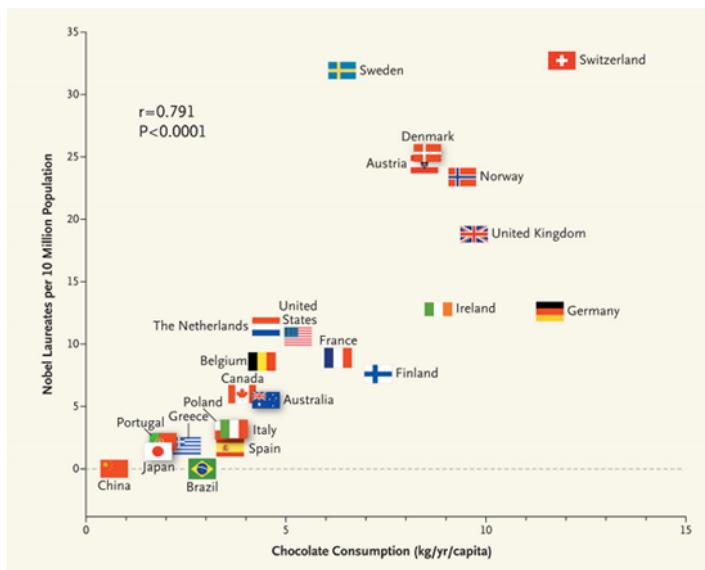


Abbildung 8.9.: Schokofutter macht schlau

Störche ist (Störche lieben Natur) und für B bei hoher Naturbelassenheit eine höhere Z begünstigen). Wir müssen die Erklärung kein Beispiel nur konkreter machen. Uns geht es h von Scheinkorrelation.

Bei spielsweise machen Männer aufgepasst eine Glatze krank? Männer mit Glatze bek (Goren et al., 2020): „Bald men at higher risk research finds“. Eine alternative Erklärung hat auf Glatze (je älter ein Mann, desto wa er eine Glatze hat) und auf die Schwere des Menschen haben deutlich schwere Coronar-

(\$

" \$

Durch (nicht-randomisierte) Einschränkung weite einer (oder beider) Variablen sinkt die einer Korrelation, vgl. Cohen et al. (2003) un-

Erstellen wir wiederum Datensätze mit je zwei Variablen und mit Umfang 10. Der der beiden Datensätze schränkt der Spannweite nun den Wertespektrum ab - verteilt \neq (Mittelwert) Standardabweichung); d.h. Datensätze in Listing 8.1. Man kann nun `schranken()` auf rufen. Wir schränken dann die Werte bis auf 0-5. Datensätze ersehen Listing 8.1.

Listing 8.1: Relation mit eingeschränkter Spannweite

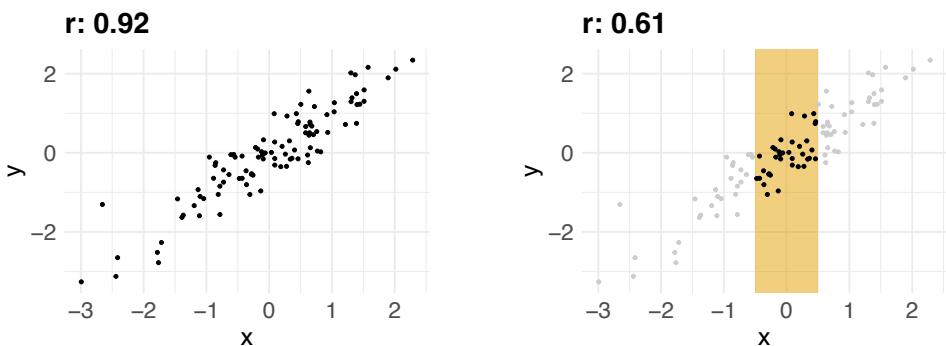
```
set.seed(42)
n < 1e2
d <- tibble(x = runif(n, min = 0, sd = 1),
            e = runif(n, min = 0, sd = .5),
            y = x + e)

x_min < 5
x_max < 5

d_filter#> Rücksicht - Einschränkung:
d %>% filter(x_min <= x & x <= x_max)
```

Übungsaufgabe Schätzen Sie die Koeffizienten prüfen Sie nach! Berechnen Sie die Koeffizienten satz und filtern

8. Punktmodelle 2



(a) ohne Einschränkung des MR an
Starke Korrelation Schwächere Korrelation

Abbildung 8.2: r sinkt, wenn der Range einer Variable verkleinert wird, so sinkt die Stärke der Korrel

In Ihrer Arbeit beim Online-Auktionshaus variablen mit dem Verkaufspreis von Computers. Falls der Datensatz auf Ihrem Computer (am Beispiel in RStudio) abgelegt ist, können Sie gewohnter Manier importieren, s. Listing

Listing 8.2: mario.kart importieren, wenn die CSV Projektordner liegt.

```
mario.kart<-read.csv("mario.kart.csv")
```

Falls der Datensatz im Unterordner mit Name liegt, so würden Sie folgenden Pfad eingeben:

Man beachte, dass solche somit eine Unterordnerstruktur, die relativ zu Ihrem Arbeitsverzeichnis, R-Studio nicht benötigt, nem Schrägstrich (Slash).

Falls Sie die Daten nicht auf Ihrem Computer

8.6. Fallbeispiele

Listing 8. Datenatz Mariokart importieren, wie im Unterdossier unter Daten unter Mario Kart.

```
mario_kart<-read_csv("Mein_Unterordner/mariokart.csv")
```

bequem von z.B. der Webseite von Vincent Arel - den:

Den Pfad hatten wir in Listing 1.1 definiert.

```
mario_kart<-mariokart_path)
```

Sie wählen die Variablen, die in diesem Fall interessieren – natürlich nur die metrischen – und lernen Korrelation aller Variablen untereinander auf.

```
mario<-t  
display(cstart_pr, ship_pr, total_pr)  
cor(2) # Runden auf zwei Dezimalen  
## start_pr ship_pr total_pr  
## start_pr 1.00 0.03 0.07  
## ship_pr 0.03 1.00 0.54  
## total_pr 0.07 0.54 1.00
```

Achtung, Namensverwechslung! Es kann vorkommen, dass R-Pakete geladen haben, in denen es jeweils zu Namenskonflikten kommt. Wir sind in dem Fall diejenigen Funktionen aus dem Paket Sie als letztes gestartet haben. Dies resultiert in einer verwirrenden Fehlermeldung „Hey Mensch, du hast Argumente in der Funktion garnicht verwenden darfst, da es `select` nicht gibt“.
`select(., duration, n_bids, start_pr, ship_pr, arguments(duration, n_bids, start_pr, ship_pr))`

8. Punktmodelle 2

Eine einfache Abhilfe ist es, R zu sagen: „aus dem Punkt „ort“ wohnt“ ein „Ort“ auf „Erichs sprach sich darüber: select(...)

Etwas schöner sieht die Ausgabe aus, s. Tabelle 8.5.

```
library(mariokart)
display(select_start_pr, ship_pr, total_pr)
correlation
summary()
```

Tabelle 8.5.: Korrelationsstabelle (tidy)

Parameter	total_pr	ship_pr
start_pr	0.07	0.03
ship_pr	0.54***	

Die Sternchen in Tabelle 8.5 geben die sog. Korrelation an; ein Thema, das wir einfach man nur einzelne Korrelationskoeffizienten die Idee des Zusammenfassens, s. Abbildung

```
library(mariokart)
summary(corse_super_cörtcohttailig_pr, wheels))
```

cor_super_wichtig
0.33

Im Falle von fehlenden Werten muss es in geschüchternen Vorsicht befreien undermutig einen Korrelationskoeffizienten auszugeben nutze = "complete.obs"

```
matrix %>% t
  summa %>% select -c(0:(tail(gapless_employees)-1)) %>% observe
```

$$\frac{\text{cor_super_wichting}}{0.33}$$

⚠️ Immer so viele Zahlen! Ich brauch Bilder.

Mit dem `plotly_express` kann man eine ansehnliche Heatmap zur Verdeutlichung der Werte, s. Abbildung 8.11.

```
library(plotly)
```

```
matrix %>% t
  plotly(extract_start_pr, ship_percent_pr, total_percent)
  plot_colrelation
```

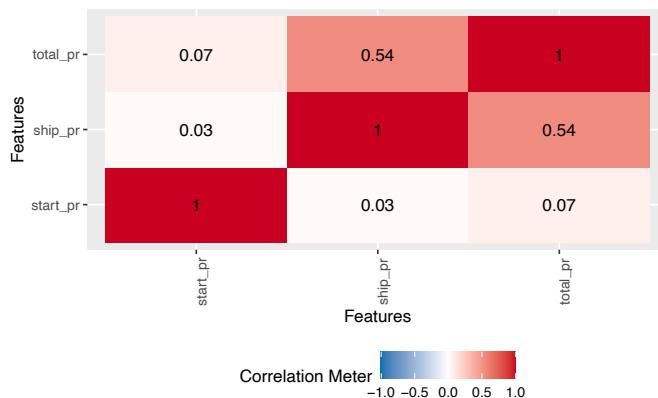


Abbildung 8.11 Heatmap zu den Korrelationen im Datensatz.

8. Punktmodelle 2

Schauen Sie sich auch mal auf [unseren Webseiten](#) zu dem Tag [Association](#) an.

1. nasa02
2. mariokart-korr1
3. mariokart-korr2
4. mariokart-korr3
5. mariokart-korr4
6. korr01
7. korr02

"

\$

Testen Sie Ihr Wissen mit einer  zur [Statistik](#) (Maße der zentralen Tendenz, Varianz, Verteilungsformen, Normalverteilung, Korrelation).

Ł

Auch die Korrelation ist ein Allzeit-Favorit, schenkt wird Ihnen jedes typische Statistik-Tutorium. Schauen Sie doch mal, was Ihre Bibliothek oder eine unorthodoxe (geometrische!) Herangehensweise an Korrelation (und Regression) sucht, darf sich auch bei Kaplan (2009) freuen. Einschönes, modernes Psychologie-Prof Russell Poldrack (2013) ist online verfügbar. Tipp: Nutzen Sie die Übersetzungswörterbücher, wenn Sie das Buch nicht in Englisch lesen.

²<https://sebastiansauer.github.io/Datenwerk/>

8. 9. Literaturhinweise

wenn auch nicht mehr ganz frisch, ist Cohen et al.
sehr empfehlenswert, aber etwas höheren Anspruch.
korrelation und was ist „echte“ Korrelation?
für die Wissenschaft zentral ist – wird von Pearson
entspannte Art erläutert; nebenbei lernt man
der Wissenschaft.

% / / /

~

„
In diesem Kapitel benötigen Sie die üblichen
easysoft-Dataträger für Kapitel 3.7.3.

“

- Sie können ein Punktmödell von einem Geraden unterscheiden.
- Sie können die Bestandteile eines Geradenmodells erläutern.
- Sie können die Güte eines Geradenmodells anlen bestimmen.
- Sie können Geradenmodelle sowie ihre Modelle benennen.

‘

Vorhersagen sind eine nützliche Sache, unter Voraussetzungen:

1. Sie sind präzise.
2. Wir wissen, wie präzise.

9. Geradenmodelle 1

3. Jemand interessiert sich für die Vorhersage
Die Methode des Vorhersagens, die wir hier auch als Regressions

!“

Beispiel Nach intensiver Beschäftigung mit Statistik ist mir als Checker bekannt. Viele jüngere Studierende kommen eine Studentin zu Ihnen, Toni, und fragen: „Wie kann ich in der Klausur erwarten?“ Sie erwidern: „Sie haben es sicherlich gelernt.“ Die Antwort: „Sagen Sie ich kann Ihnen nicht geben Sie den Notenschnitt der letzten Klausur der Person. Dazu rechnen Sie schnell den Notenschnitt.“

Ihre Antwort lautet: „Im Schnitt habe ich die Note ungefähr 91.12 erzielt. Diesen Wert kann ich Ihnen nicht genauer nennen, da ich keine genauen Infos habe, z.B. wie viel die anderen Studierenden keine genauere Vorhersage machen. Sorry!“

Ohne Kenntnis eines Prädiktors (UV) (wie z.B. Mittelwert eingeeigneter Vorhersagewerte für die Klausur 9.1. Wir nutzen den Mittelwert als Punktmittelmodell).

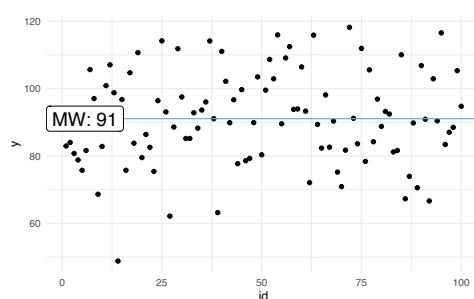


Abbildung Mittelwert als Vorhersagewert, bspw. Punktmodell

9.2. Vorhersagen

Definition Nullmodell (Punktmodell) Prädiktor
Punktmodell also, kann man das auch Prädiktoren hat, nennt man es auch manchmal „Nullmodell“

Auf Englisch kann man dieses Nullmodell so spez

```
lm0<-lm(y~1,data=noten2)
lm0
## 
## Call:
## lm(formula = y ~ 1, data = noten2)
## 
## Coefficients:
## (Intercept)
##      91.1
```

Es steht für „linearer Modells“ es keine Prädiktoren im Fall wird der Mittelwert, 91, als Gerade vom Modell bestimmt und dieser einzige Punkt ist ein Punktmodell ist, sagt es für alle Beobachtungen gleichen Wert vorher, nämlich 91.

!

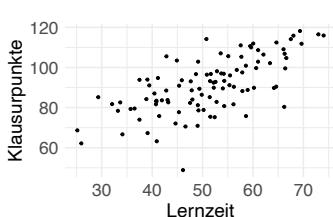
Beispiel: Toni verrät Ihnen mit Sicherheit, wie sich dann noch die Lernzeit zuverraten: „Okay, also ich gelernt, insgesamt.“ Jetzt müssen Sie erst mal Klausurpunkte angeben vorher, wenn Toni 42 St. Sie visualisieren sich zur Hilfe die vorliegen (a).

```
library(DaRty Explorer)
note<-2
plot_sc(byt="y", pY=note)
variable muss angegeben werden
```

9. Geradenmodelle 1

Auf dieser Basis antworten Sie Toni: „Bei 40 Lernstunden so 83 Punkte bekommen. Könnte mit dem Betrieb der Schule nicht begeistert von Ihrer Prognose.“

Der „Trend“ (im Sinne eines Linealzusammenhangs) ist deutlich zu erkennen: Je mehr Lernzeit, desto mehr Klausurpunkte. Mit einem Lineal könnte man eine Gerade in das Streudiagramm einzeichnen, s.



(a) Streudiagramm (b) Streudiagramm mit „Trend“ (blau)

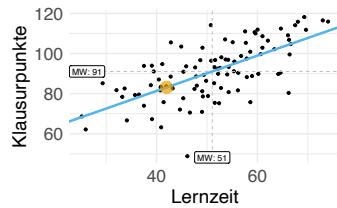


Abbildung 9.1 zeigt und Lernzeit: Rohdaten (a) und Mittelwerte sind mit gestrichelten Linien verdeckt. Toni's Vorhersage ist mit einem Punkt markiert.

Wir verwenden eine Gerade als Modell für die Beziehung. Anders gesagt: Wir modellieren den Zusammenhang mit einer **Geradenmodellierung**. Eine Verallgemeinerung des Punktmodells: Ein Punktmodell sagt für einen gegebenen Wert vorher, was bei diesem Wert passieren wird. Ein Geradenmodell gegenübersetzen.

In einem Geradenmodell wird nicht mehr (notwendig) die gleiche Voraussicht (wie das bei einem Punktmodell der Fall ist).

Definition: Eine Gerade ist das, was man bekommt, wenn man eine lineare Funktion in ein Koordinatensystem einträgt. Sie kann durch zwei Punkte bestimmt werden: Achsenabschnitt (y-Achse) und Steigung (Steigung). Maglich mal wird (z.B. Schulunterricht) der Achsenabschnitt bei einer Geraden mit β_0 bezeichnet:

$$(\beta_0 + \beta_1 x) = y = \beta_0 + \beta_1 x.$$

In der Statistik wird folgende Nomenklatur benutzt: $\beta_0 + \beta_1 x$ oder $(\beta_0 + \beta_1 x) = y$.

Die Nomenklatur hat mehrere Vorteile, dass man das Modell erweitern kann. Anstelle β_0 kann man auch σ_0 schreiben, um zu einer Aussage über eine Population, nicht nur über einen Individuum machen möchte.

Das „Dach“ überträgt aus, dass es sich um den gesuchten Wert vom Modell vorhergesagten („modellierten“) Wert handelt. Es skizziert die Elemente einer Regression.

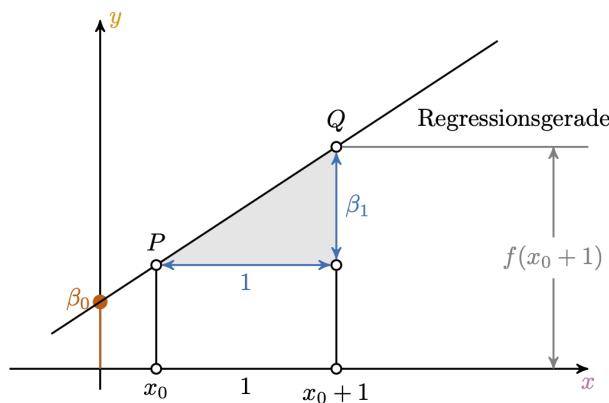


Abbildung 9.3.1: Schaubild einer Regressionsgeraden (Menk, 2014)

9. Geradenmodelle 1

Definition: Ein einfaches lineares Modell beschreibt den Wert einer abhängigen Variable als lineare Funktion von einer (oder mehreren) plus einem Fehlerterm. Gleichung 9.1.

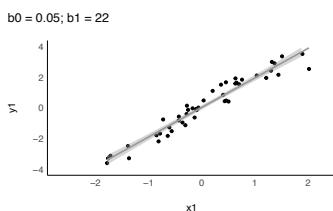
$$= (\textcolor{violet}{b}_0 + \textcolor{blue}{b}_1 x_1) + \textcolor{brown}{\epsilon} \quad (9.1)$$

Mit:

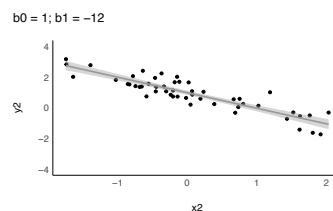
- $\textcolor{brown}{b}_0$: geschätzter y-Achsenabschnitt (dargestellt)
- $\textcolor{blue}{b}_1$: geschätzte Steigung (Regressionsgerade)
- $\textcolor{brown}{\epsilon}$: Fehler des Modells

In Gleichung 9.1 sind die tatsächlichen beobachteten Messwerte von vorhergesagtem Modell für beide darstellen.

Jenach Datenlage können sich Regressionsgeraden unterscheiden, s. Abbildungen



(a) Datensatz 1



(b) Datensatz 2

Abbildung 9.1: Regressionsanalysen mit verschiedener Güte

Bei Beispiel 9.1 will es gesehen werden, ob der Konsument zwei Fragen. Erstens: Wie viele Punkte benötigt er, um nicht zu lernen? Zweitens, wie viele Punkte benötigt er, um zu lernen?

9.3. Geradenmodelle

Stunde? Ist immerhin meine Lebenszeit, krieg sind gute Fragen. Wer da ausurpungt, der Achsenabschnitt zurück. Schnell skizzieren Abbildung 9.5. Puh, die Antwort wird Toni nich

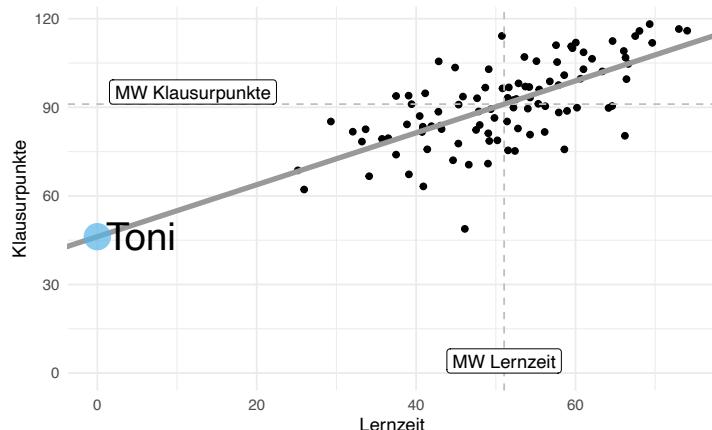


Abbildung 9.5 Achsenabschnitt: Wie viele Punkte kann bei 0 Lernstunden? (roter Punkt be

Anstelle auf Abbildung 9.5 zuschauen, können SKlausurerfolg vorhersagen (top predict) lassen

Hey R, predicte mir mal auf Basis vom Modell Lernerfolg für Toni, wenn der x = 0 Stunden le

Okay, ich predicte mit Modell „lm1“ und ne Datentabelle Tonis Lernzeit (x = 0)!

```
tonis_lernzeit <- tibble` erstellt einen Datafra
```

```
predict(lm1, newdata = tonis_lernzeit)
## 1
## 46
```

predict startet für das Argument Dataframe. In diesem Beispiel ist lernzeit

9. Geradenmodelle 1

$$\$ \quad f$$

Ein Geradenmodell kann man im einfachsten Gleichung 9.3 dargestellt.

$$x \quad (9.2)$$

Lies: „Laut meinem Modell ist gneidne (gesFaunäktz von“).

Wir erinnern uns, und das sie ist:

$$(9.3)$$

Wir werden als Funktion nur Geraden verwenden der Gerade lassen wir uns vom Computer ausrechnen Sie so ins Errische übersetzen:

```
I m(y ~ x dat meine_daten)
```

steht für „lineares Modell“, also eine Gerade nennt man Rauschfleck (nsg amdeine Stelle in der unscharf als „Trendgerade“ bezeichnet).

Beispiel (Zahlen für Tonton ist nicht zufrieden mehr sagen): „Jetzt hör mal auf mit deinem Lied Ich will es genau wissen, sage mir präzise Z

```
I m1 I m(y ~ x dat a oten 2)
I m1
# #
## Call :
## I m( formula = y ~ x , data = noten 2)
# #
## Coeff cients :
```

```
# # (Intercept)      x
# # 46.191    0.879
```

R gibt Ihnen die beiden Koeffizienten für die Gleichung des Objekts. Können Sie folgende Ausssetzung der Regressionsgleichung $y_{\text{a_up}} = 46.191 + 0.879 \cdot x$ herstellen?

8. ist der Achsenabschnitt zweidimensional. Dieser Wert von 46.191 ist das Regressionsgewicht, d. h. die Steigung der Lernzeit steigt der vorher gesagte Punkt.

Mit Kenntnis der beiden Koeffizienten kann man ausrechnen, gegebenenfalls jemand zum Beispiel eine Stunde gelernt, würden wir folgendes Klausur-

```
lernzeit <-
y_pred <- 46.191 + 0.879 * lernzeit
y_pred
# [1] 55
```

Bei spielsweise für Klausurerfolg gemäß ersten beiden, noch etwas Gutes für Toni zutun. R hilft Ihnen die erwartete Punktzahl aus, wenn Toni 73 Stunden aber auch selber rechnen, wenn Ihnen das lieber geht.

```
tonis_lernzeit <-
```

tonis_lernzeit ist eine Tabelle mit einer Zeile und einer Spalte.

```
tonis_lernzeit
```

9. Geradenmodelle 1

$$\begin{array}{r} \overline{x} \\ \hline 73 \end{array}$$

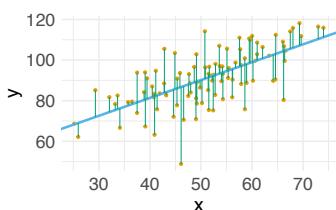
```
predict(mdl, newdata = table_mi_t_pradi  
# # 1  
# # 110
```

Die Syntax erfordert:

```
predict(modell, newdata = table_mi_t_pradi
```

Die Funktion `predict` liefert eine Vorhersagelinie für einen und für einen bestimmten Dataframe (der die muss).

Die Differenz zwischen vorhergesagtem Wert und dem tatsächlichen Wert nennt man (errordreßiduum -).



(a) Geradenmodell (lmb) Punktmodell (lmo)

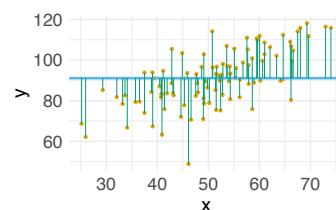


Abbildung 9: Vorhersagefehler als Abweichung geradenmodell, sind die Vorhersagefehler (Kürzer) als in (b)

9.3. Geradenmodell

Wie ist es mit den Vorhersagefehlern von beiden Modellen? Lassen wir uns von R die Streuung (Residuen) und die Absolutabweichung (MAE) aussagen? (aus dem Paket `PerformanceAnalytics`)

```
ma €I m0)
ma €I m1)
# # [ 1] 11
# # [ 1] 8
```

Vergleichen wir MAE im Nullmodell mit MAE in der Regressionslinie.

```
verhaeltnis_mafeehler(m1) - ma€I m0)
verhaeltnis_fehler_mae
# # [ 1] 0.71
```

Ah! Das Geradenmodell ist ~~schwach~~ schlechter als das Nullmodell. Die Absolutlänge des Fehlerbalkens auf 71 Prozent ist sehr schlecht!

Definition: Fehlerstreuung bezeichnet die Gesamtheit der Abweichungen vom beobachteten vorhergesagten Wert.

Zur Berechnung der Fehlerstreuung gibt es mehrere Methoden: MAE oder MSE. Ein Geradenmodell ist immer besser (im Hinblick auf die Verringerung der Fehlerstreuung) korreliert.

Natürlich können wir – in Analogie zur Varianz – die Quadratfehlerbalken (Mean Squared Error, MSE) berechnen. Das mag, kann den MSE auch von `ma€I m0` berechnen:

$$\text{mean}(\text{noten}2\$y))^2$$

```
ms €I m0)
ms €I m1)
# # [ 1] 193
# # [ 1] 106
```

9. Geradenmodelle 1

```
verhaeltnis_mfseeHimis_Eins@) <-  
verhaeltnis_fehler_ms  
## [1] 0.55
```

Betrachtet man die MSE, so kann man eine Ver 0.55 feststellen.

$$\hat{y} = f(x)$$

Aber wie legt man die Regressionsgerade in dasgesprochen?

Die Regressionskoeffizienten (hier $\hat{\beta}_0$ und $\hat{\beta}_1$) wählt man so, dass die Residuen $\hat{y}_i - y_i$ möglichst genau gesagt werden. Die Summe der quadrierten Residuen minimiert:

$$\min \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (9.4)$$

Es gibt verschiedene Methoden, um die Koeffizienten zu schätzen (die aber nicht in diesem Buch zu finden sind). Dazu findet sich bei Kaplan (2009).

$$\# " \quad \hat{y} = f(x)$$

Das Modell ist noch 0.55 der Fehlerstreuung, das ist auf 0.45 verbessert.

```
1-verhaeltnis_fehler_ms  
## [1] 0.45
```

9.4. R-Quadrat als Maß der Modellqualität

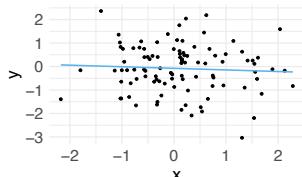
Definition: Der Anteil der Verringerung (Abbildung 9.6) der Fehlerstreuung der Zielvariable im untersuchten Modell ist das R-Quadrat (eines Modells). Definiert als die Verringerung der Varianz des Modells im Nullmodell $\frac{MS_E}{MS_{E_0}}$.

R-Quadrat ist Maß der Modellqualität, das besser ist als Vorhersage. Da es ein Anteilsmaß ist, liegt der Wert zwischen 0 und 1. Im Nullmodell beträgt R-Quadrat per Definition 0. Modellen des Typs $y = \beta_0 + \beta_1 x$. □

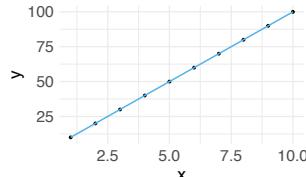
Einfach gesagt an, wie gut (zu welchem Anteil) Zielvariable vorhergesagt wird. Wir können nur R-Quadrat berechnen, so ausgeben lassen:

```
r(21 m1)
## # R2 for Linear Regression
##     R2: 0.448
## adj. R2: 0.442
```

Bei einer perfekten Korrelation ist das auch vgl. Abbildung 9.7.



(a) Keine Korrelation



(b) Perfekte Korrelation

Abbildung 9.7 zeigt Fälle von R-Quadrat: 0 und 1. (a) Keine Korrelation, $r = 0$ und $R^2 = 0$. Prognose durch Mittelwerte. Regressionsgerade ist (ungefähr) parallel zur x-Achse. (b) Perfekte Korrelation: Die Prognose ist gleich dem beobachteten Wert.

Bei einer perfekten Korrelation liegen alle Punkte auf einer Geraden. Im Gegenteil liegen die Punkte auf einer Geraden.

9. Geradenmodelle 1

genauso gut, wie wenn man für jede Vorhersage gen würde. Je größer R-Quadrat, desto besser die Daten; desto besser „erklärt“ das Modell die „Fit“, sagt man).

/ #

{

Die Residuen (Vorhersagefehler) bestimmen die Güte des Modells. Sind die Residuen im Schnitt groß, so ist die Modellgüte gekehrt. Verschiedenen Koeffizienten stehen (als Korrelation von Autoregressionsmodellen R-MSE, MAE, ...)

f1 f

Die Modellkoeffizienten, obwohl sie eigentlich sind nur eingeschränkt zu interpretieren, genden kausalen Abhängigkeiten nicht keinen statistischen Zusammenhang darf man keine annehmen. Ohne eine zugrundeliegende Theorie kann man kein Argumentieren. Oder sie der Modellgüte und den Vorhersagen begnügt ist.

Im Modell liegt der Achsenabschnitt b_0 hinzu. Bei Betrachtungen mit $=$ könnten also die Werte linear abhängen, es ist aber häufig so, dass Prädiktorwerte von 0 in der Praxis vorkommen, sodass der Achsenabschnitt dann wenig nützt.

Beispiel (Regression Größe Nutz Gewicht Körpergröße) per gr und das Gewicht von Menschen vorherzusagen, von Körpergröße wenig nützlich, da es keine 0. □

9.6. Wiemann mit Statistik

So interpretiert man die Germania-Mittelwerte im Vergleich, der Regressionskoeffizienten, die den zeitlichen Verlauf einer Stunde unterschließen und um den Wert t_1 von

F V o r s i c h t

Häufig liest man, der „Effekt des Prädiktors“ also „Effekt“ ist aber ein Wort, das man kausal. Ohne weitere Absicherung kann man aber Regressen nicht kausal verstehen. Daher sollte man mit Vorsicht genießen. Manches sprechen daher „statistischen Effekt“.

(\$)

Der Unterschied im Modellgut \hat{e} z. um \hat{d} sch. e. \hat{a} , sagen
 ist ein kleiner zweiter Schätzfehler = . \hat{d} ist ein (lineares
 Maß der Modellgut \hat{e} un \hat{d} aufen Unterschiede in
 auf die gleiche Weise interpretiert werden
 Abbildung 9. 8 zeigt den Zusammenhang von

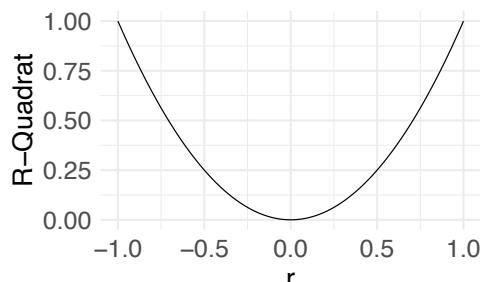


Abbildung 9 Zusammenhang von r und R-Quadrat ist
ar.

9. Geradenmodelle 1

F Vorsicht

Unterschiede zwischen Korrelationsdifferenzen interpretiert werden.

Als mittlerweile anerkannter Extrem-Datensatz eines Auktionshauses, indem Sie arbeiten, haben Sie möchten eine genaue Vorhersage von Verkäufen von diesem Planberichteten, leuchtet ein. Genaue Vorhersagen sind von hoher betriebswirtschaftlicher Bedeutung. Mariokart-Daten laden, am besten ohne Extraktionsfehler, und los geht's (und die üblichen Pakete starten).

```
I m2<#> tota$tmpdatmariokart  
r(m2)  
## # R2 for Linear Regression  
## R2: 0.005  
## adj. R2: -0.002
```

Ohnein! Unterirdisch schlecht. Anstelle überlegen Sie und schauen dann nach, welche Korrelation bestimmt.

```
I m3<#> tota$hippdaptmariokart  
parametem3$
```

9.7. Fallbeispiel Mario

Tabelle 9.2.: Modellparameter von lm3

Parameter	Coefficient	SE	95% CI	t (141)
(Intercept)	36.25	2.54	(31.23, 41.26)	
ship_pr	4.34	0.57	(3.22, 5.46)	7.6

Der Achsenabschnitt liegt bei ca. 36 Dollar, während ein Spiel, das mit null Dollarn vertrieben wird, kalkulatorisch kein Verkauf spricht. Da der Warratne (Versandkosten) steigt, der zuerwartende finale Verkaufspreis um 95% gibt einen Schätzbereich für den jeweiligen An, denn es handelt sich bei den Koeffizienten um wahre Werte in der Population ist unbekannt. Wurde nur eine Stichprobe der Regressionsgleichung vorgenommen und damit $\hat{y}_{pred} = 36 + 4 * ship_pr$

In Worten:

Der vorhergesagte Gesamtpreis eines Spiels („Sockelbetrag“ plus 4 mal die Versandkosten)

Man kann sich die erwarteten Werte („expected prices“ in Abhängigkeit vom Wehtschaft („to estimate“) lassen es nun dar, war es möglich statt (Tabelle 9.3.)

Tabelle 9.3.: Die vorhergesagten (predicted) Werte und die vom vorhergesagten Wert (Residuals)

Tabelle 9.3.: Model-based Expectation

ship_pr	Predicted	SE	95% CI	Residual
4.00	53.59	1.87	(49.89, 57.30)	-
3.99	53.55	1.87	(49.85, 57.25)	-

9. Geradenmodelle 1

ship_pr	Predicted	SE	95% CI	Resid.
3.50	51.43	1.82	(47.82, 55.03)	
0.00	36.25	2.54	(31.23, 41.26)	
0.00	36.25	2.54	(31.23, 41.26)	
4.00	53.59	1.87	(49.89, 57.30)	

Variablenpredicted: total_pr

„Ah, bei 4 Dollar Versandkosten ist laut dem Verkaufspreis zu erwarten“, fassen Sie sich

👉 Das sieht man im Predicted. Schon steht der vorher sagte Wert auf für jeden bestimmten Wert ship_pr

👉 Kann ich das nutzen? Ich würde gerne Verkaufspreis wissen, wenn die Versandkosten 4 Dollar liegen.

👉 Ja, klar!

```
neue_daten(bike
ship_c(1,4) # zwei Werte zum Vorhersagen
```

```
predict(wdane_daten)
# # 1 2
# # 41 54
```

Aber nützlich wären noch, das Modell (bzw. die Daten) als Diagramm zu bekommen. Das Abbildung 10.7.

```
estimate(xlpm3) total(tot)
```

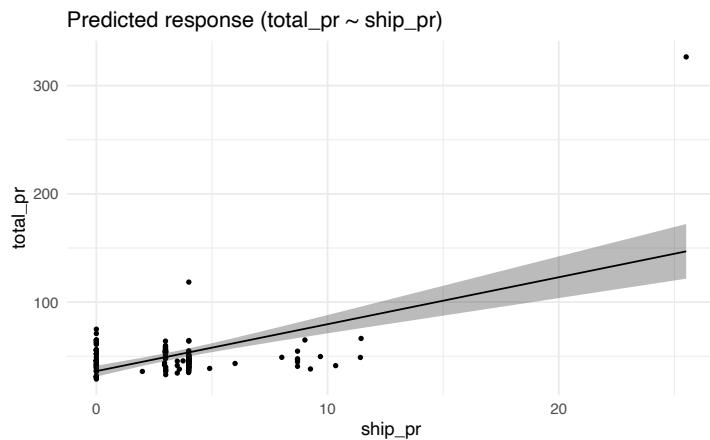


Abbildung 9.9.: Verbilldlichung der erwarteten

estimaten \rightarrow bestätigt, sogenannt „schätzenden zuerst den Wert“. Kurz gesagt: Wir wollen eine Vorhersage

Am wichtigsten ist Ihnen aber im Moment die F-Statistik. Das Modell ist, spricht wie lang oder kurz die (absoluten) Balken sind:

```
maef1m3)
## [1] 13
```

Das Modell erklärt einen Anteil von ca. 0.29 des

```
r2f1m3)
## # R2 for Linear Regression
##      R2: 0.294
## adj. R2: 0.289
```

```
maef1m3)
## [1] 13
```

9. Geradenmodelle 1

Im nächsten Meeting erzählen Sie Ihrem Chef preis von Mario Kart - Spielen im Schnitt aufgen! ". Hört sich gut an. Allerdings hätte ich man da noch was machen?

.

1

In dieser Fallstudie geht es darum, die Preisanalysen. Kurz gesagt: Sagen Sie die Hauspreise Ihre Vorhersagen als CSV bei kaggle.com ein, die Prognosewettbewerbe veranstaltet. Ich Sie teil an der Kaggle-Competition „House Price Prediction Techniques“, die Sie auf der Kaggle-Webseite auch eine nähere Beschreibung, das Ziel des Wettbewerbs.

.

Sie können die Daten von www.kaggle.com herunterladen. Sie müssen Sie folgende Dateien herunterladen:

- Data_descriptions.html, d.h. Beschreibung im Datensatz
- train.csv von Häusern, die Sie nutzen, stellen
- test.csv von Häusern, von denen Sie den Preis vorhersagen sollen
- sample_submission.csv für die Prognosedaten, mit der Sie Ihre Vorhersagen einreichen.

Sie können auch über das [Gutenberg Project](#) die Daten zugreifen:

```
d_train_path_online <-
  "https://raw.githubusercontent.com/sebastian-
  refs/heads/main/data/kaggle-train.csv"

d_test_path_online <-
  "https://raw.githubusercontent.com/sebastian-
  refs/heads/main/data/kaggle-test.csv"

d_train <- d_train_path_online
d_test <- d_test_path_online
```

Laden Sie diese Daten am besten herunter und sie in ein passendes Unterverzeichnis (Ihres Projekt ab).

Importieren wir die Daten von der Festplatte datenR (davon ausgehend, dass der Unterordner Ihres aktuellen R-Projekts ist):

```
d_train <- dattah/kaggle-train.csv"
d_test <- dattah/kaggle-test.csv"
d_train <- d_train_path
d_test <- d_test_path
```

Wenn das Importieren von der Festplatte nicht hilfreich, wenn man Daten von der eigenen Festplatte fürs Erste können Sie die Daten auch von oben Pfad importieren.

!

Die Prognosедatei ist die Datei, die Ihr Vorhältnis. Sie soll prinzipiell so aussehen wie in Tabelle 9.1.

9. Geradenmodelle 1

Tabelle 9.4.: Beispiel für den Aufbau der

i d	SalePrice
1461	169277
1462	187758
1463	183584

Die Prognosedatei besteht also aus drei Spalten, welche die Hauspreise angeben, welches Haus eine Zeile ihrer Prognosedatei gemeint ist - für einen Kaufpreis vorhergesagt. Die Variable `id` steht für den Kaufpreis des Hauses mit der ID, die steht. Insgesamt soll die Prognosedatei gerader Test-Datensatz, also die Tabelle, die angeibt. Alles klar? Los geht's!

Schauen Sie sich zu Beginn einmal die Verteilungen, z.B. `disttrain` (`d_train`)

Eine einfache Antwort auf die Frage, welche sage eignen, ist, die Korrelation aller PrVariablen berechnen, s. Listing 9.1.

Aha! Ein Mängelnformation... Wenn Sie Teile nicht verstehen: Im Zweifel einfach ignorieren: Führen Sie die Syntaxstruktur ein und das Ergebnis erneut. Es führt wieder die Ausgabe betrachten. Listing 9.1 auf geführten Variablen sind ein

¹die vorherzusagende Variable, auch Ziel- oder Outcome Variable genannt.

Liste Woche Variablen korrelieren stärker als

```
d_t r%>i%n
sel(d)>%
correl()%>%# berechne Korrelationen
f l { Parame t e s a 2 e }%>i %c # "aber nur, wo die zweite Var
" SalesPrice" ist
arr a(a b(er)%>%# sortiere absteigend nach der Höhe
Korrelationskoeffizienten r
f l faeb(sr>3) # nur |r| > . 3
```

Zielvariable Preise beeinflusst. Nutzen wir also diese Variable und Ganzen durch den Zusammenhang am besten. Diese Variable ist am stärksten mit dem Preis korreliert und daher ein guter Kandidat für das Modell.

Im ersten Modell gehen wir davon aus, dass der Wohnungspreis und die Größe durch den Zusammenhang am besten erklärt werden kann. Diese Variable ist am stärksten mit dem Preis korreliert und daher ein guter Kandidat für das Modell.

```
m1 <- lm(SalePrice ~ OverallQual + train)
paramet m a s u e s y s t a t s
```

Tabelle 9.5.: Parameter von

Tabelle 9.5.: Fixed Effects	
Parameter	Coefficient
(Intercept)	-96206.08
Overall Qual	45435.80

Wie gut ist das Modell?

9. Geradenmodelle 1

```
r ms(m1) aus easystats  
## [1] 48589
```

Im Schnitt liegt 86% Molar daneben. Ob das weniger ist, wird sich im Vergleich mit anderen R-Quadrat liefern einen anderen Blick auf die Daten.

```
r Qm1) aus easystats  
## # R2 for Linear Regression  
## R2: 0.626  
## adj. R2: 0.625
```

Man kann mehrere UV (Prädiktorvariablen) aufnehmen. Dazu trennt man sie in einem Modell.

```
mein_mod <- lm(uv~1+uv2.+uv_dna, t meine_daten)
```

Dabei ist das Pluszeichen kein arithmetisches „+, als UV nimmt UV1 und UV2 und ...“. Berechnet ein Modell mit mehreren UV,

```
m2 <- lm(uv~1+uv2.+uv_dna, t meine_daten)  
par(mfrow=c(1,2))
```

Tabelle 9.6 zeigt die Koeffizienten von

Tabelle 9.6.: Modellparameter von

Parameter	Coef f c i e n t
(Intercept)	- 98832.49
Overall Qual	27104.83

Parameter	Coefficient
GrLivArea	50.67
GarageCars	21298.96

Wie gut sind die Vorhersagewerte dieses Modells von train?

```
rmse(m2)
## [1] 40566
```

Im Schnitt liegen unter 406100\$ abgrenzbar. Ist das gut? Betrachten wir noch

```
r2(m2)
## # R2 for Linear Regression
##   R2: 0.739
## adj. R2: 0.739
```

Ob die Modellgüte (R-Quadrat, RMSE, etc.) „gut“ wortet man am besten im Vergleich zu anderen Modellen.

Zum Vergleich berechnen wir das maximal einfache Modell. Man nennt es das Nullmodell. In diesem jedes Haus einfach den mittleren Preis aller Häuser.

```
mo <- SalePrice %>% train()
```

Wie gut ist die Vorhersage des Nullmodells?

```
rmse(mo)
## [1] 79415
```

Beim Nullmodell liegen wir ca. 80 Tausend Dollar über dem tatsächlichen Preis. Das R-Quadrat des Nullmodells ist per Definition 0.

9. Geradenmodelle 1

```
r Qm0)
# # # R2 for Linear Regression
# #   R2: 0.000
# # adj. R2: 0.000
```

% m2

Wir haben jetzt unser eAnl CheHöpfifomung ruht auf Modell. Ob die Vorhersagen im Test-Sample pr

himmel weit daneben? Enttäusche uns nicht!

Hier sind die Vorhersagen:

```
m2_pred <- predict(m2, newdata = test)
head(m2_pred)
## 1 2 3 4 5 6
## 103395 152441 161838 187676 225467 190260
```

(1)
(2)

- Erstelle eine Vorhersage anhand der Regressionswerte anhand der Daten aus
- Zeige den „Kopf“ dem Projekt und sag eine (erste) Vorhersagen.

Die Vorhersagen fügen wir jetzt dem Test-Sample hinzu:

```
d_test <-
d_ts %>
mutate(Stael_e_Prim2C_Pred)
```

"

So, wir haben unsere Vorhersagen! Jetzt reichen sie aber noch nicht. Für die Prognosedatei (submission file) müssen wir die Preise

```
m2_subm<-
d_t@%
sel(elcd, SalePrice)
```

Kaggle möchte keine fehlenden Werten in den Vorwir das mal:

```
m2_subm%
drop(%)#①
nr@%#②
## [1] 1458
```

- Lass alle Zeilen mit NAs (fehlenden Werten) fallen, filtere diese Zeilen also raus
- zähle die Anzahl der Zeilen (die noch verbleiben)

Die Anzahl der Zeilen, die wir hier erhalten, ist Zeile `nr@%` gibt also keine fehlenden Werte.

```
nr@%_test)
## [1] 1459
```

Diesen Tibbles speichern wir als CSV-Datei angelegt sie wihiatnez ucvs er wenn dve int akt sovma tisch (ungefragt noch eine lfd-Spalte ohne Name einfügt (mit den mag aber Kaggle nicht). Kaggle erwartet exakt mit den Namen `SalePrice`

```
write(m2_subm, "ames-kaggle/m1-subm.csv")
```

Und dann laden Sie `load` die seben Deutschen Kaggle hoch und hoff auf einen Hauptgewinn. Das Modell. 655261 teilte ein

9. Geradenmodelle 1

Diese Fallstudie hat eine einfache Prognose. Es gibt es viele Ansätze, dieses Modell zu verbessern. Fragen, die Sie sich dazustellen können:

- Welche Prädiktoren sollte ich in das Modell einfügen?
- Wie gehe ich mit fehlenden Werten um?
- Wenn ein Prädiktor schief ist, sollte er transformiert werden?
- Vielleicht sollte man manche Prädiktoren weglassen?
- Wie gehe ich mit nominalskalierten Variablen um?
- ...

Viel Spielraum für Ihre Kreativität!

Die Webseite datenwerk.netlify.app stellt zahlreiche Übungsaufgaben bereit. Sie können die Suche nutzen, um die Aufgaben mit den folgenden Namen zu finden:

- Aussagen-einfache-Regr
- interpret-koeff-lm
- korrelations-regr
- Linearitaet1a
- lm1
- mtcars-regr01
- nichtlineare-regr1
- penguins-regr02
- regression1
- regression1b
- Regression3
- Regression4
- Regression5

9.10. Literaturhinweise

- Regressionen
- Ames-Kaggle

Schauen Sie sich auch weitere Aufgaben des Datensatzes aus der Sammlung passend zum Stoff des Tages an.

Nicht alle Aufgaben dieser Sammlung passen zum Stoff des Tages; vielleicht können Sie einige Aufgaben nun einfach diese Aufgaben.

7

Gelman et al. (2021) liefert eine deutliche Einführung in die Regressionsanalyse als dieses Kapitel orientierte Einführung in Statistik inklusive eines Aha-Potenzials. Cohen et al. (2003).

Ł

- Sie können Regressionsmodelle für Försch
nominaler und metrischer UVerläutern u
- Sie können Interaktionseffekte in Regre
und in Ranwenden.
- Sie können den Anwendungszweck von Ze
Transformationen zur besseren Interpret
modellen erläutern und in Ranwenden.
- Sie können Modelle nutzen, um Vorhersage
zu erstellen.

!

Neben den üblichem dyPunkt des sysbantigen Sie i
diesem Kapitael automatisch p t g opnab r

```
I i b(rtairy v e r s e )  
I i b(ryaaryd #tfiücrk)Modell Güte im Test - Sample  
I i b(rera sys t a t s )  
I i b(rggpy #bDra)t en visualisieren, optional
```

10.2. Forschungsbezug: Gläse

Dieses Mal arbeiten wir nicht nur mit den Marien, auch mit Klima- bzw. Wetterdaten.

```
wetter_ppasta(the@-
```

```
"https://raw.githubusercontent.com/sebastiaan  
statistik1/main/data/wetter-dwd/precipitation  
wetter@(wetter_path)
```

Die Wetterdaten stammen vom DWD und können Sie hier herunterladen.

fl

Lineare Modelle (synonym: Regressionsanalyse) sind ein mächtiges Werkzeug. Sie gehören immer noch zum Repertoire moderner Analystinnen und Analysten. Von linearen Modellen zeigt sich (leider?!) im Allgemeinen:

Beispiel: Wie gut kann man Ihre Persönlichkeit aus Facebook-Media-Posts vorhersagen? Mit viel Medienanalyse untersuchten Kosinski et al. (2013), wie durch Facebook-Daten (Likes etc.) vorhergesagt werden kann. Autoren resümieren im Abstract:

We show that easily accessible digital records, such as Facebook Likes, can be used to automatically predict a range of highly sensitive personal traits, including: sexual orientation, ethnicity, political views, personality traits, intelligence, addictive substances, parental separation and more.

¹Lizenzhinweis: Datenbasis: Deutscher Wetterdienst, ehemalige Statistikabteilung des DWD

²<https://raw.githubusercontent.com/sebastiansauer/statistik1/main/data/wetter-dwd-data-dict.md>

10. Geradenmodelle 2

Die Autoren berichten über einzelne Methoden tatsächlichen persönlichen Attribut und Ihres Modells, s. Abbildung 10.1. Das eingeschaltete Modell beruht auf einem linearen Modell, also ähnlichen vorgestellten Methoden. Nebender analytischen Analyse zeigt das Beispiel auch, wie die internen sind.

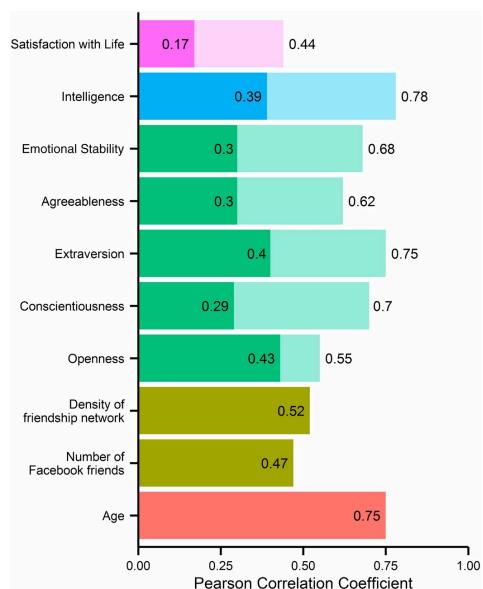


Abbildung 10.1: Pearson correlation accuracy of regression attributes and traits expressed by the Pearson coefficient between predicted and actual values

(

◦

Beispiel: Wer ist der Nächste? Wenn Sie eine Zeit als bei dem Online-Auktionshaus gearbeitet haben, dann hat was Abwechslung. Viel Geld verdienen

10.3. Wetter in Deutschland

dann fallen Ihnen ein, dass Sie ja zu Generation Z geschönen Mammon nicht so hoch schätzen sollte sich, Ihre hochgeschätzten Analyse-Skills für Ihnen sinnvoll erscheint: Die Analyse des Klimas

Beim Deutschen Wetterdienst, DWD haben Sie sie Deutschland heruntergeladen. Nach etwas Date nicht eingehen wollen, resultiert eine schöne analysieren möchten. (Im Datensatz ist die Temperatur in Celsius angegeben; $\text{Wert} \text{ in } \text{N}^{\circ}\text{C}$ lag pro Quadratmeter.) Hervorragend! An die Arbeit!

& '

In diesem Abschnitt untersuchen wir lineare mehreren metrischen UV (und einer metrischen AV). Sie stellen sich nun folgende Forschungsfragen:

💡 Um wieviel ist die Temperatur in Deutschland gestiegen, wenn man die letzten ca. 100 Jahre vergleicht?

Die Modellparametermodell in Tabelle 10.1 zu sehen.

```
lm_wettkampfpreiswetter)
parametern_wetter1)
```

Tabelle 10.1.: Modellparameter von lm_wetter1

Parameter	Coef f ci - t	SE	95 % CI	t (28864)	p
(Intercept)	-14.25	1.85	(-17.87, -10.63)	-7.77	< .001
year	0.0104	9.478e-0132	.300.01	<.001	

10. Geradenmodelle 2

Laut Ihrem Modell wurde es pro Jahr um 0.01 zehnt also 0.1 und pro Jahrhundert 1 Grad.

💡 Das ist sicherlich nicht linear! Vermut tur bis 1950 konstant geblieben und jetzt Decke!

💡 Mit der Ruhe, dass schauen Sie sich später

In Tabelle 10.1 finden sich zwei Arten von Achsenabschnittsmaßzahlen des Regressionsmodells wichtsv

1. Punktschätzungen für die entsprechenden „Best-Guess“ (Punktschätzer) für den entsprechenden Wert in der Population. Das ist sozusagen das Modell festlegen würde, wenn es sonst ni

2. Bereichsschätzungen als Punktschätzungen (Intervallschätzungen): Hinterlicher Wert für den entsprechenden Wert an plausibler Werte Kombinationen auf -dence interval, CI) bezüglich der Werte -ze aus. So können wir ablesen, dass das Rye ärgendwo zwischen praktisch Null (0. geschätzt wird. Je schmäler das Konfidenz wird der Effekt geschätzt (unter sonst gleichen Bedingungen) einfacher. Einatmen, ausatmen.

Definition Konfidenzintervall (CI): Ein Konfidenzintervall (CI) gibt einen Schätzbereich plausionswert an, auf Basis der Schätzung, die

Das Modell kann die Schätzungen zu den erwarteten Werten so ausgeben lassen, s. Abbil

10.3. Wetter in Deutschland

sind das zu viele Datenpunkte. Wir sollten es überprüfen, s. Abbildung 10.2, rechts. Dazu aggregiert man eines Jahres zu jeweils seinem Mittelwert.

```
wetter_summ<-  
  wet %>%  
  group_by(%>%  
    summarise(ties=mean(ntemp),  
              precipitation=precipitation: engl. für Niederschlag)
```

Auf dieser Basis erstellen wir `imw_iwesaline` in einer Tabelle 10.2.

```
im_wetter<-map(.adrat_wetter_summ)  
parametern_wetter1a)  
sel(eParameter, Coeffcient)
```

Tabelle 10.2.: Modellparameter von `im_wetter`

Parameter	Coeffcient
(Intercept)	-14.14
year	0.01

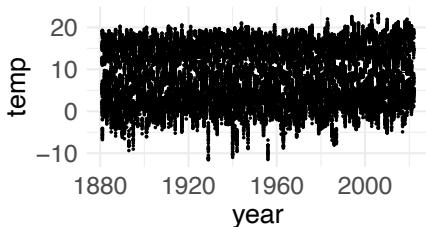
Dann plotten wir also `Mödelt1_mitte_relativtion` (`im_wetter`) das Modell wettersprechend, s. Abbildung 10.2.

💡 Moment mal, der Achsenabschnitt liegt bei 14. Was bedeutet das?

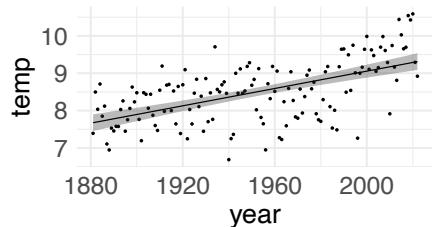
&'

Zur Erinnerung: Der `Avg` hat bei `mittelt` einen Wert, der höher ist als der Wert anderer Städte. Kapitel 9.5.

10. Geradenmodelle 2



(a) Ein Punkt pro Tag



(b) Ein Punkt pro Jahr

Abbildung zeigt Änderung der mittleren Temperatur im Zeitverlauf (Datenquelle: Punkt ist ein Tag (viel Overplotting)). Rechts: Jeder Punkt ist ein Jahr (wobei es sich um die Regressionsgerade handelt).

In den Wetterdaten wäre Jahr = 0 Christi Geburt aufzeichnung gerade mal ca. 150 Jahre in der Zukunft vollkommen vermessen, dass Modell 2 genheit zu extrapolieren, ganz ohne, dass es Sinn macht, einen Referenzwert festzulegen, etwa 1950. Wenn wir dann von alten Daten auf das Jahr 1950 zum neuen Jahr Null. Dann haben wir Daten. Hat man nicht einen bestimmten Referenzwert angegeben, so ist es üblich, z. B. einen Referenzwert zu nehmen. Diese Transformation (Zentrierung, centering) der Daten, s. Definition.

So zentriert man eine Verteilung:

```
wetter <-
  weather %>%
    mutate(average = mean(year)) %>% select(-year)
```

Das mittlere Jahr in unserer Messwertreihe ist:

10.3. Wetter in Deutschland

we t %e %
s u mma(miesa(ny e a r))

mean (year)
1951

Die Steigung (d. h. der Regressionskoeffizient) ändert sich, s.

I m_wetter1_Lz(fretnet myie am dtact wetter)
parametern_wetter1_zentriert)

Tabelle 10.4.: Modellparameter von Imwelt

Parameter	Coefficient	SE	95% CI	t (28864)	p
(Intercept)	-8.49	0.04	(-8.42, -8.57)	219.43	
year c	0.01 0.4	9(4.780 0.01)	-0132.30	<.001	

Jetzt ist die Interpretation des Achsenabschnittes 1951 ($x=0$) lag die mittlere Temperatur in Deutschland ca. 8.5 Grad Celsius. Die Regressionsgleichung lautet $y = 8.49 + 0.01 \cdot x$. Wörteren: Wir sagen eine Temperatur als Summe von 8.49 Grad plus 0.01 mal das Jahr berechnet. Zur Erinnerung: Der Referenzwert der Referenzgruppe entspricht dem Y-Wert bei $x=0$ im

Wie gut erklärt unser Modell die Daten?

10. Geradenmodelle 2

```
r Q1 m_wetter #_azuesn't{r e a s y}st at s } `  
## # R2 for Linear Regression  
## R2: 0.005  
## adj. R2: 0.005
```

Viel Varianz des Wetters gehabt das Modell ungeahnt gleich stark multiert, daher wir Modellgütenicht unterscheiden.). Macht aus Jahreszahl spielt z.B. die Jahreszeit eine Rolle. Das haben wir nicht berücksichtigt.

💡 Wie warm ist es laut unserem Modell dann?

```
pred1 m_wetter n_e wett er h(Year, 100)  
## 1  
## 9.7
```

💡 Moment! Die Vorhersage ist doch Quatsch! 2022 lag die Durchschnittstemperatur bei Wilke, 2013).

💡 Wir brauchen ein besseres Modell! Zum Beispiel ambitionierte Nachwuchs-Wissenschaftler & Co.

Definition: Eine Variable, die als Indikatorvariable dient, hat nur zwei Ausprägungen.

Beispiel: Die Variable `is_dry` zeigt, ob es im vergangenen Jahr früh oder spät geregnet hat.

Beispiel: Interessiert Sie folgende Forschungsthemen?

💡 Ob es in der zweiten Hälfte des 20. Jahrhunderts wärmer war, im Durchschnitt, als vorher?

Aber wie erstellen Sie ein Modell, das die Hälfte des 20. Jahrhunderts (und danach) zu fassen? Nachdem Sie auf die Idee, das vektorisierte Rechnen auszunutzen:

```
year <- 1950
after_1950 <- year > 1950
after_1950
## [1] FALSE FALSE TRUE
```

Die ersten zwei Jahre sind größer als 1950, schon.

Ja, so könnte das klappen! Diese Symbole übertragen Daten:

```
wetter <-
wet %>%
mut_datfet <- year > 1950
file <- file("Daten für Gesamt - Deutschland")
```

Scheint zu klappen! Jetzt ein lineares Modell für die 10.5.

```
lm_wetter <- lm(wetter ~ after_1950)
```

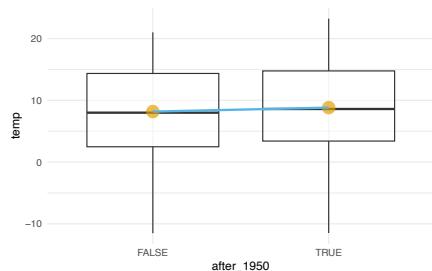
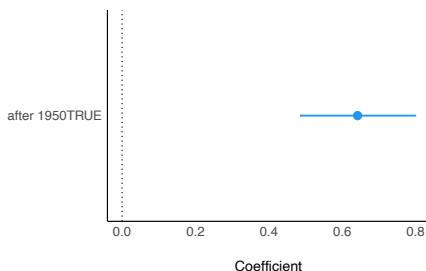
10. Geradenmodelle 2

Tabelle 10.5.: IPm_rwaehteteer_biom_uv

Tabelle 10.5.: Fixed Effects

Parameter	Coefficient	CI
(Intercept)	8.18	(8.06, 8.29)
after_1950TRUE	0.64	(0.48, 0.80)

Die Parameter des Modells lassen darauf schließen, dass es wärmer war nach 1950, und zwar im Schnitt von 0.64 Grad, s. Abbildung 10.3.



(a) Mittelwertsunterschiede der Regressionsparameter
b) Boxplot zur Unterscheidung der Gruppenvergleich

Abbildung 10.3: (a) Schätzbereich des Parameters $\beta_{\text{after}} = 0.64$. Der Unterschied ist sehr groß.
(b) Der Unterschied sieht in dieser Größe aus.

Leider zeigt ein Blick zum Ergebnis die Werte für den geschätzten Parameter des Modells zu wünschen übrig. Das kann auf verschiedene Weise geschehen.

Um die Koeffizienten eines linearen Modells man eine metrische X- und eine metrische Y-Variable mit der Variablen X umzuwandeln, gibt es für uns ohne viel Ankündigung durchführt: mehrere Variablen, s. Definition 10.2.

Hat eine nomiz ~~wesitewan~~en, so überführt (synonym mit ~~da~~) diese Variable in eine binäre Variable. Da eine binär wie eine metrische angesehen werden kann, kann gewohnter Weise durchgeführt werden. Wenn Sie Parameter betrachten, so sehen Sie die neu erst Tabelle 10.5). Man beachte, dass der ursprüngl ändert wird, nur während ~~die~~ ~~die~~ Änderung der Variable durchgeführt.

In unserem Fall ist die logische Variable mit den Werten TRUE und FALSE werden von R automatisch bzw. auslesen standen. Also: Eine logische Variable.

💡 Ein kannst duals „Ja! Richtig!“ verstehen „Nein! Falsch!“

E wichtig

Ein lineares Modell mit binärer UV zeigt nicht die Differenz der Gruppenmittelwerte.

```
wet %>%
  group_by(after %>% 50)
  summarise(temp_mean))
```

	after_1950 temp_mean
FALSE	8.2
TRUE	8.8

Die Interpretation eines linearen Modells mit Licht Abbildung 10.4: Der Achsenabschnitt ist der Mittelwert der 1. Gruppe. Der Mittelwert der 2. Gruppe aus Achsenabschnitt und dem Koeffizienten der Zeile 10.4 zeigt nur die Daten für den Monat Juli der Einfachheit und Übersichtlichkeit halber.

10. Geradenmodelle 2

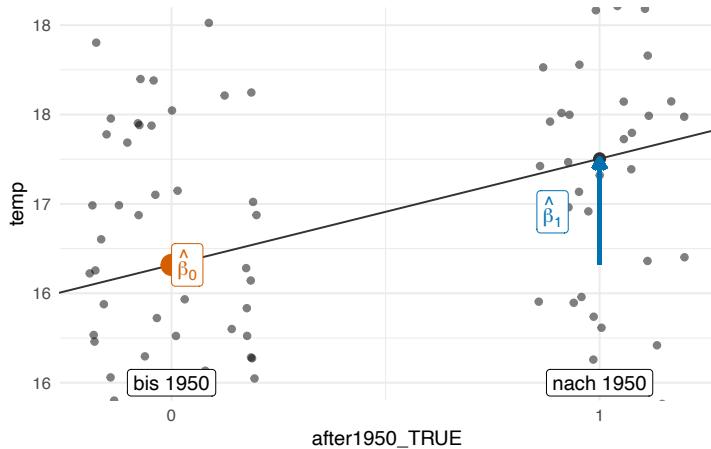


Abbildung 10.24: Zur Interpretation eines binären UV (reingezoomt, um den Unterschied hervorzuheben)

Fassen wir die Interpretation der Koeffizienten des UV zusammen:

1. Mittelwert der 1. Gruppe ($\hat{\beta}_0$ bis 1950): Achse
 2. Mittelwert der 2. Gruppe ($\hat{\beta}_1$ nach 1950): Achse
- Summe der Regressionsgeraden ($\hat{\beta}_0 + \hat{\beta}_1$)

Für die Modellwerte seien:

- Temperatur laut Modell $\hat{y} = 16.7 + 1.3 \cdot 0 = 16.7$ bis 1950:
- Temperatur laut Modell $\hat{y} = 16.7 + 1.3 \cdot 1 = 18.0$ nach 1950

Bemerkung: Wenn auch beide Variablen kategorialen Schaltestvariablen sind, kann man die Schaltvariablen zusammenfassen.

³Ich danke Karsten Lübke für diese Idee.

ž &

In diesem Abschnitt betrachten wir ein linear
onym: Regressionsmodell) mit einer mehrstufigen
UV. So ein Modell ist von den Ergebnissen her
ein Varianz amatleyisne einziges UV.

Beispiel abeo. Wohl substantielle Temperaturunterschieden
schen den Bundesländern gibt?

Befragen wir dazu ein lineares Modell; in Tab.
Parameter der Punktschätzer (Koeffizient) und
bereich (Konfidenzintervall) mit Ober- und Unt.

Im_wetter_Liniegmpgida, wetter)

Tabelle 10.7.: Modell_1wetter metegifoür

Parameter	Coeffcient	CI_hi	CI_lo
(Intercept)	8.25	8.56	7.94
regionBayern	-0.63	-0.19	-0.97
regionBrandenburg	0.57	1.14	0.00
regionBrandenburg/Berlin	0.58	0.86	0.30
regionHessen	0.11	0.56	-0.44
regionMecklenburg-Vorpommern	0.08	0.26	-0.18
regionNiedersachsen	0.52	0.80	0.24
regionNiedersachsen/Hamburg/Bremen	0.51	0.79	0.23
regionNordrhein-Westfalen	0.80	1.18	0.42
regionRheinland-Pfalz	0.46	0.74	0.18
regionSaarland	0.71	1.11	0.31
regionSachsen	-0.04	0.40	-0.48
regionSachsen-Anhalt	0.55	1.03	0.07
regionSchleswig-Holstein	0.17	0.45	-0.11
regionThueringen	-0.48	-0.76	-0.20
regionThueringen/Sachsen-Anhalt	0.11	0.59	-0.48

10. Geradenmodelle 2

Hat die nominal skalierte UV mehr als zwei Stufen, so ist sie in mehr als eine Indikatorvariable umgewandelt. Eine Indikatorvariable weniger als es Stufen gibt, gilt nicht. Allgemein gilt: Wenn wir diese Variable abhängige Variablen umgewandelt.

Betrachten wir ein einfaches Beispiel, etwa das Gründen der Einfachkredit hier. Bundesländermarke mit den kategorialen Indikatorvariablen umgewandelt.

i d	B u n d e s l a n d	i d	B L _ B a y e r n	B L _ B r a
1	BaWü	1	0	0
2	Bayern	2	1	0
3	Brandenburg	3	0	1

Auch im Fall mehrerer Ausprägungen einer nominalen Logik der Interpretation wie bei bis:

1. Mittelwert der 1. Gruppe: Achsenabschneide der Regressionsgeraden (\hat{y}_1)
2. Mittelwert der 2. Gruppe: Achsenabschneide der Regressionsgeraden (\hat{y}_2)
3. Mittelwert der 3. Gruppe: Achsenabschneide der Regressionsgeraden (\hat{y}_3)
4. usw.

Es kann vorkommen, dass das Beauftragte nicht den Wert für einen linearen Modell aus. Beispielsweise war die Reihenfolge bereits festgelegt, vgl. Kapitel 10. Die Gruppe entspricht dem Achsenabschnitt.

Beispiel: Achsenabschnitt Danach steht der Wert 2, da das alphabetische erste Bundesland ist, wird

ausgewählt, dessen Mittelwert als Achsenabschnitt genommen wird.

Ameinfachstenverfahren weicht nicht mit einem Diagramm, s. Abbildung 10.5.

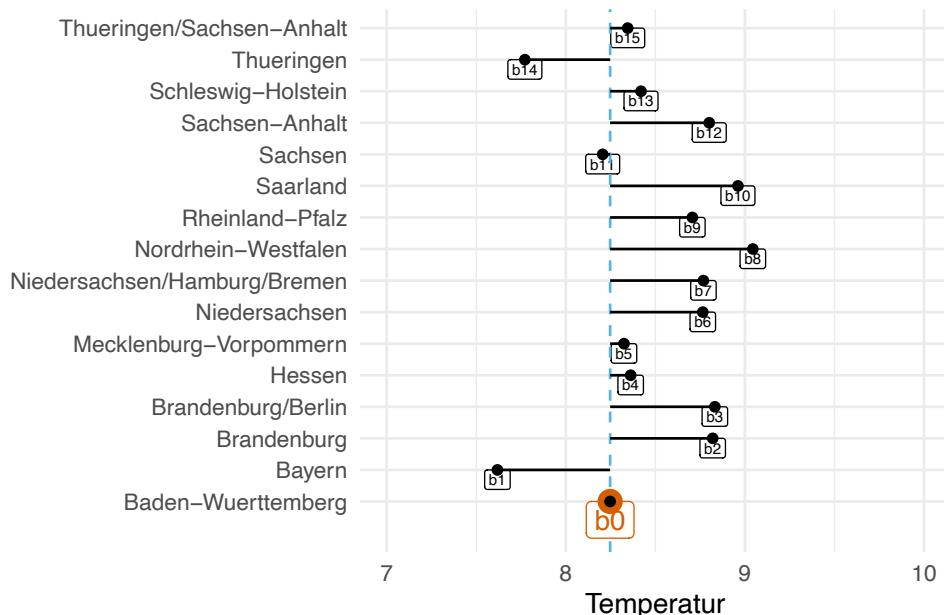


Abbildung 10.5: Zur Interpretation eines linearen nominalen UV (reingezoomt, um den Mittelschied hervorzuheben).

Beispiel Ni1: Oder schlagsmenge im Viergleewecher Monatere Forschungsfrage, die Sie nicht außer acht Fragen nach den jahreszeitlichen Unterschieden precipitation) Los R, rechne!

Endlich geht's weiter!

```
Im_wetter_L(montagsdienstagwetter)
parametern_wetter_month)
```

10. Geradenmodelle 2

Ja, das scheint es deutliche Unterschiede im brauchen ein Diagramm zur Verdeutlichung (plot(estimate_expect) at iOmn(elim:we bte trandte als numerische Variable! Aber „Monat“ bzw. sein.

⌚ Abemontst als Zahl in der Tabelle hinterliche Maschine verarbeitete eine Zahl als

🟡 Okay, R, wir möchten eine nominale Variable formieren. Wie geht das?

⌚ Dazu kannst du ~~die~~ umschmehl. Damit wandelt du eine numerische Variable in eine nominale (Faktorvariable) um. Faktisch heißt Zahl als Text gesehen wird.

Bei spiellra0s8. Form mit man, so wir 4d2a" us der Zahl wird ein Text. Alle metrischen Eing die Variable ist jetzt auf nominalen Niveau

```
wetter <-  
wet %>%  
mut intoen t h _ f a c t(bmoonth )
```

Jetzt berechnen wir mit der faktorisierten s. Tabelle 10. 10.

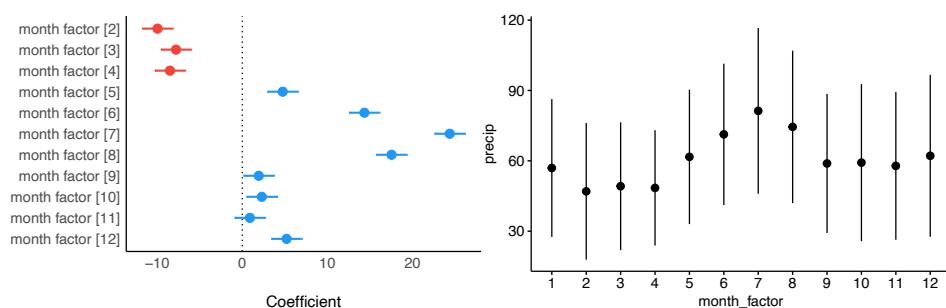
```
lm_wetter_mohnth(prfamnotrh_fdaact,ter)  
parameters wetter_month_factor  
sel(eRatmeter, Coeffcient)
```

Tabelle 10.10.: Modellparameter von lm_wetter

Parameter	Coefficient
(Intercept)	56.95
month_factor 2	-9.95
month_factor 3	-7.78
month_factor 4	-8.49
month_factor 5	4.74
month_factor 6	14.34
month_factor 7	24.36
month_factor 8	17.52
month_factor 9	1.93
month_factor 10	2.29
month_factor 11	0.89
month_factor 12	5.20

Sehr schön! Jetzt haben wir eine Referenzgruppe und 11 Unterschiede zum Januar, s. Abbildung 1

```
parameters_wetter_mop %>% 
  ggplot(aes(x=month_factor, y=precip)) +
  geom_point() +
  geom_errorbar()
```



(a) Regressionsgewicht Mittelwerte plus SD

Abbildung 10.6.: Niederschlagsmengen nach Monat

10. Geradenmodelle 2

Möchte man die Referenzgruppe eines Faktor mit einbeziehen:

```
wetter <-
  wet %>%
  mutate(month_factor,
```

Sie sieht dann die geänderte Reihenfolge aus

```
lev(wetter$month_factor)
## [1] "7"  "1"  "2"  "3"  "4"  "5"  "6"  "8"  "9"  "10"
```

&

In diesem Abschnitt untersuchen wir ein linearisiert (übergären) Umpeleturis und die (in einem Modell kann auch Klassifizierungsmöglichkeitssanalyse oder covariance analysis of covariaten bezeichnet werden.)

Beispiel 1: Ob die Niederschlagsmenge zwischen den Monaten entwickelt hat in den ersten Einfachheitshalber greifen Sie sich nur und Juli).

```
wetter_month_1_7 <-
  wet %>%
  filter(month == 7)
```

Ich muss mal kurz auf eine Sache hinweisen

Eine Faktorvariable ist einer der beiden Datumskaliierten Variablen, nämlich **Faktorenvariablen**. Ein wichtiger Unterschied ist, dass prägungen („Faktorstufen“) bei einer Fakt

⁴Zum Dollar-Operators. Kapitel 3.11.3

werden, bei der Text-Variablen nicht. Das kann bei einer Faktorvariable immer klar, welche Variable möglich sind.

Beispiel(B1e0.s1p0) für eine Faktorvariable

```
geschlecht<--  
geschlecht_af<-as.factor(geschlecht)  
geschlecht_factor  
## [1] f f m  
## Levels: f m
```

Beispiel(IF10.t11n verändert die Werte von Geschlecht der Faktor ⁵gaersichtlich ist. Ein "m" steht für weiblich, so dass z. B. nur die ersten beiden Elemente übrig. Ausprägungen markieren sich R trotz ~~wie~~ Faktor, dass es keinen gibt ("fun d n").

Genauso ist es, wenn man die automatisch herausgefilterten Faktorstufen „Löschen“, s. Faktorvariable neu definieren mit

```
wetter_month_1_7 <-  
wet %>%  
f1(month=1|month=7) %>%  
# Faktor (und damit die Faktorstufen) neu berechnen  
mutant(month_factor(month))
```

Okay. Wie spezifiziert man jetzt das lineare Modell?

Hat man mehrere („multiple“) X-Variablen (Prädiktoren), so trennt man sich mit einem Plus-Zeichen $+ \text{Bavar}_c + \text{month}$

⁵Synonym: nominal skalierte Variable

10. Geradenmodelle 2

Definition und Beispiele Regressionsmodelle mit mehr als einer X-Variablen. Die Modellform:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

Das Pluszeichen hat in der Modellgleichung die additive Wirkung. Es wird nicht im Modellgleichungssatz das Pluszeichen nur „+“ geschrieben.

Die obige Modellgleichung liest sich also so:

Temperatur ist eine Funktion von Zentertemperatur und Monat

```
lm_year_month_factor ~ month_factor + month_1_7)
```

Die Modellparameter sind in Tabelle 10.11 zu sehen.

Tabelle 10.11.: Modellparameter von

Parameter	Estimate	SE	95 % CI	t (4525)
(Intercept)	56.94	0.68	(55.60, 58.62)	83.015
year c	0.03	0.01	(0.05, 0.05)	2.7
month factor	204.37	0.97	(22.48, 26.27)	212.02

Die Modellkoeffizienten sind so zu interpretieren:

1. Achsenabschnitt (Referenzjahr): Regressionsmodell für den Monat Januar mit niedrigstem Niederschlagsmengenquadratmeter.

2. Regressionskoeffizient für Jänner: Für Jänner ist die Niederschlagsmenge im Schnitt um 0.02 mm an-

3. Regressionskoeffizienten für Monat (i) lag die mittlere Niederschlagsmenge (im Reziprozen über dem mittleren Wert des Referenzmonats)

Die Regressionsgleichung lautet $\text{pred} = 56.94 + 0.03 * \text{year_c} + 24.37 * \text{month_factor_7}$

Im Monat Januar hat $\text{month_factor_7} = 0$

⚠️ Puh, kompliziert!

⚠️ Es gibt einen Trick, man kann sich von beliebigen Y-Wert berechnen lassen, s. Bei-

Beispiel (IN1@deutschschlag laut Mohleey R, bneJruel ch@20? uns anhand neuer Daten den laut Modell zu erwarten für Januar im Jahr 2020!)

```
neue_daten<-car_2020-1951
month_factor7<-
predict(month_factor7,
# # 1
# # 59
```

Alle Regressionskoeffizienten basieren auf der Annahme, dass alle übrigen Prädiktoren den Wert auf weisen

Visualisieren wir uns die geschätzten Erwartungswert, s. Abbildung 10.3. $\text{expectation}(\text{Im_year}$

Mit `scale_color_blackwhiteStandard` (Farbpaletten von (Okabe & Ito, 2023) ersetzt (s. Barrett, 2022) nötig, aber robuster bei Schwarz-Weiß-Druck u. vgl. Kapitel 5.9.2. `Dilemnykelaartheorie` von

10. Geradenmodelle 2

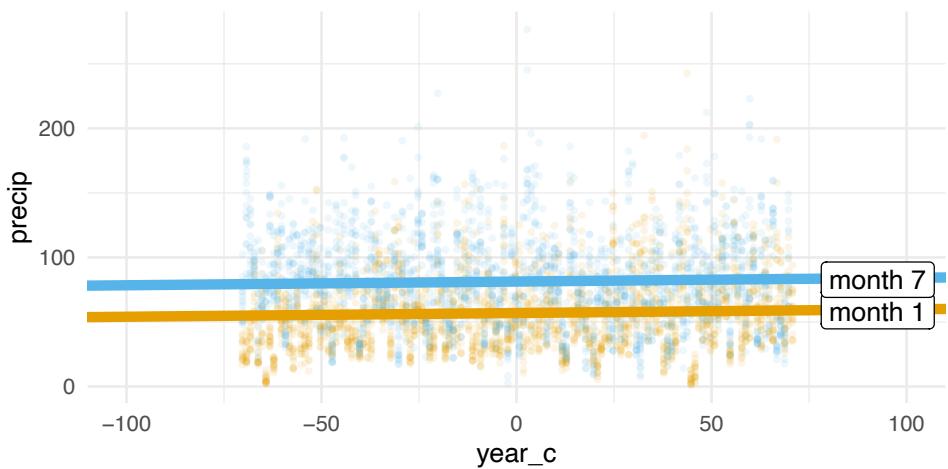


Abbildung zeigt die Regressionsgeraden für Januar (month 1) und Juli (month 7). Der Verlauf der Jahre ist deutlich zu erkennen.

```
r QI m_year_month)
## # R2 for Linear Regression
## # R2: 0.124
## # adj. R2: 0.124
```

/

Eine Modellgleichung kann die Regressionsgeraden dazu, parallel zu verlaufen, besser in die Punktwolken passen, wenn wir nicht parallel verlaufen zu dürfen? Nicht parallel verlaufen wir, indem wir das Regressionsmodell visualisieren, s. Listing 10.1.

Visualisiert ist das Modell in Abbildung 10.1.

List in Egi 10. mit Interaktionsmodell spezifiziert man
 $+ x_2 + x_1 : x_2$

```
lm_year_month_LinInteraktion <-  

  preetyParMonth_factor, yearMonth_factor,  

  data = weather_month_1_7)
```

```
plot(estimate(xlpm_cyeteatri_omonth_interaktion))  

scale_color(r#sohneistbarschema
```

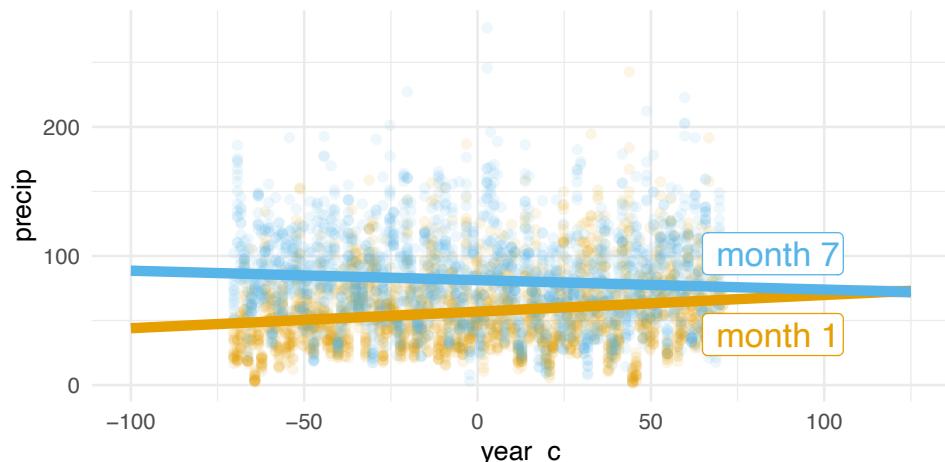


Abbildung zeigt die scheinbare Veränderung im Jahresverlauf und Month mit Interaktionseffekt: Die Veränderung über Jahre ist unterschiedlich für die Monate. Die beiden Regressionslinien unterscheiden sich deutlich.

Der Doppelpunkt () markiert die Interaktion zwischen Jahr und Monat. Der Regressionsgleichungsterm $x_1 : x_2$ beschreibt die Interaktion zwischen Jahr und Monat ($month_factor$)

$precip \sim year_c + month_factor + year_c : month_factor$

10. Geradenmodelle 2

Definition: Der Interaktionseffekt ist der Effekt, der die Abhängigkeit zwischen den Variablen x_1 und x_2 darstellt.

$$y \sim x_1 + x_2 + x_1 \cdot x_2$$

In Worten:

y wird modelliert als eine Funktion von x_1 und x_2 . Der Interaktionseffekt von x_1 mit x_2 .

Wie man in Abbildung 10.8 sieht, sind die beiden Regressionsgeraden von zweiten nicht parallel, es besteht ein Interaktionseffekt.

Beispiel: Der Interaktionseffekt von Niederschlag ist die Veränderung der Niederschlagsmenge in Jahren (X -Achse) auf darauf folgenden Monat (y -Achse). Der Effekt deutet Zeitschritte hin: Der Niederschlag ab im Juli ist höher als im Januar.

Liegt ein Interaktionseffekt vor, kann man mehrere statistischen Effekte eines Prädiktors (auf die Y-Achse) unterscheiden: Je nach Gruppe (z.B. der Effekt des Jahres auf die Niederschlagsmenge) ist der Effekt statistisch, nie kausal gemeint.)

Betrachten wir die Parameterwerte des Modells 10.12.

Tabelle 10.12.: Modellparameter von lm_y

Parameter	Coefficient
(Intercept)	56.91
year_c	0.13
month_factor 7	24.37
year_c:month_factor 7	-0.20

Neu bei der Ausgabe zu diesem Modell ist die `year_c`. Sie gibt die Stärke des Interaktionseffekts unterschiedlich sich die Niederschlagsmenge in den Monaten im Verlauf des Jahres wieder. Der Effekt ist im Jahr 2000 um 0.20 mm geringer. Die Regressionsgerade für Juli deutlicher als im Referenzmonat Januar, als Null ist. Die Regressionsgeraden für die Monate Juli bis Dezember liegen um 24.37 mm höher als im Januar.

E Wichtig

Der Achsenabschnitt gibt den Wert für Yan und dass alle Prädiktoren den Wert Null aufweisen.

Wenn eine Beobachtung in allen X-Variablen den Achsenabschnitt den Niederschlag für den Januar. Die Regressionskoeffizienten geben die Zunahme der jeweiligen Prädiktorwerte um 1 steigt, die übrigen Werte aufweisen.

Das R^2 von `year_month` ist trotz der hohen Anzahl der Prädiktoren nur geringer als im Modell ohne Interaktion:

```
r2lm_year_month #_aiinst`e{reakstyisotn}ts } `  
## # R2 for Linear Regression  
## R2: 0.139  
## adj. R2: 0.138
```

{ & '

+ & '

Ein Modell mit zwei metrischen UV kann man sich visualisieren, s. Abbildung 10.11, oder im 2D-Raum oder im 3D-Raum wird die Regressionsfläche dargestellt.

10. Geradenmodelle 2

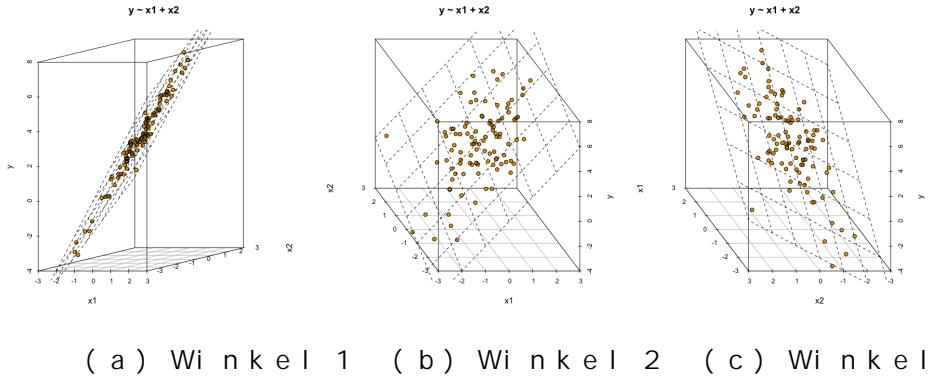


Abbildung 10.10.: 3D-Diagramm für ein Regressionsmodell mit zwei Prädiktoren im 3D-Raum.

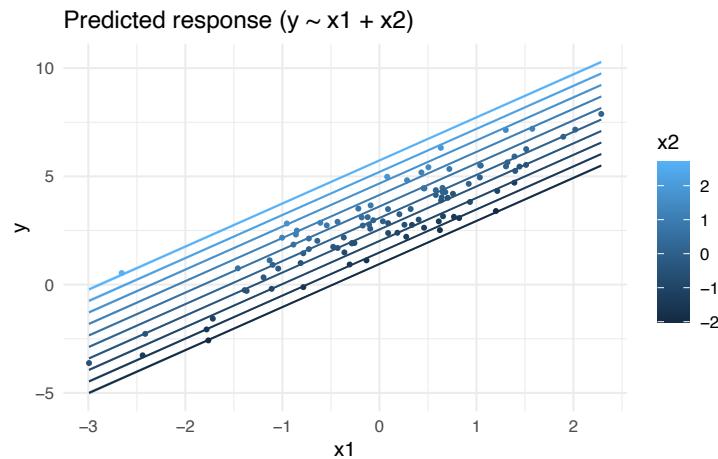


Abbildung 10.10.: 2D-Diagramm für ein Regressionsmodell mit zwei Prädiktoren im 3D-Raum.

10.4. Modelle mit vielen

Grundsätzlich kann man viele Prädiktoren nehmen. Betrachten wir z. B. folgendes lineare Schen UV.

I m_ m a r i o l _ f 2 u o t a \$ t p r + s _ h p i r p d _ a t r m a r i o % a % t
f l (t o t a l 0) 0 r

Abbildung 10.11 visualisiert das Modell
Diagramm (betrachtet aus verschiedenen Winkeln)

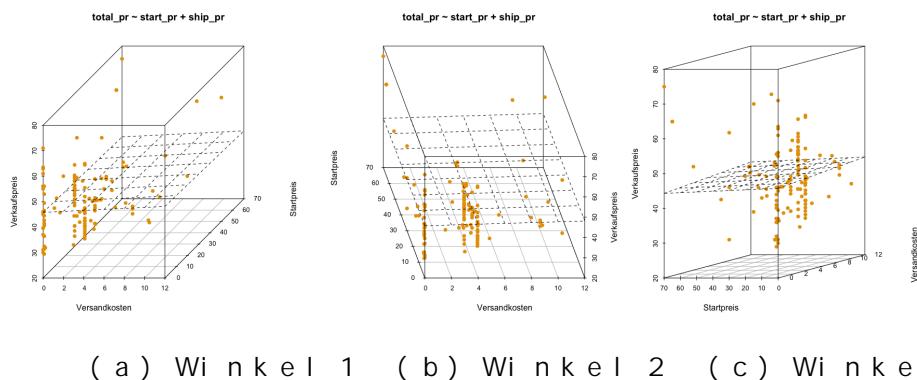


Abbildung 10.11 visualisiert das Modell mit mehreren metrischen UV (und 1 nominalen AV) als 3D-Diagramm

' & '

Wir könnten im Prinzip alle Variablen unserer Torenn in das Regressionsmodell aufnehmen. Das Sinn?

Hier sind einige Richtlinien, die helfen, welche viele) man in ein Modell aufnehmen sollte (Ge 199:

1. Man sollte alle Prädiktoren aufnehmen, von ist, dass Sie Ursachen für die Zielvariable

10. Geradenmodelle 2

2. Bei Prädiktoren mit starken (absoluten) Interaktionseffekten kann man nicht auf die einzelnen Prädiktoren schließen.
3. Prädiktoren mit kleinen und schwächeren Interaktionseffekten können im Modell belassen werden, da sie die Modellgenauigkeit erhöhen.

! .

Beispiel: Präsentieren Sie eine Analyse des Verkaufsmarkts von Business-Welt und ihren Gratifikationen und wissenschaftlichen Ambitionen. Sie haben den Verkaufspreis von Mario Kart-Spielen möglich gemacht. Also gut, das Honorar ist phantastisch, Sie haben Geld.

{

Um die Güte Ihrer Vorhersagen zu prüfen, teilen Sie die Daten in zwei zufällige Teile.

Ich teile dem Data-Exploratriz in zwei Teile. Der ersten Teil kannst du nutzen, um Modelle („trainieren“) und ihre Güte zu prüfen. „Trainingssample“, hört sich cool an, oder? Das ist ein Anteil (von 70% der Daten, okay?) lichen Daten behalte ich. Wenn du eingekommen bist du und wir berechnen die Güte der Modelle in dem verbleibenden Teil, die übrigens diesen Teil nennen wir Test-Sample, alle.

Wenn die Daten auf Ihrer Festplatte liegen, dann können Sie sie von dort importieren:

```
mariokart@datacamp/mariokart_train.csv"
```

10.5. Fallbeispiel zur P

Alternativ können Sie sie auch von dieser Empfehlung der Cloud herunterladen:

```
mariokart_tprasi(neOpPath <-
  "https://raw.githubusercontent.com/sebastia
  "stastik1/main/data/mariokart_train.csv"

mariokartTest(mariokart_trainPath)
```

Dann importieren wir auf gleiche Weise Test-Sa

```
mariokart_pastePath <-
  "https://raw.githubusercontent.com/sebastia
  "stastik1/main/data/mariokart_test.csv"

mariokartTest(mariokart_testPath)
```

Also los. Sie probieren mal die „All-in-Strategie“ das Modell. Viel hilft viel, oder nicht?

```
library(mariokart_train)
#& m_a#& n)easystats
## # R2 for Linear Regression
##     R2: 0.994
## adj. R2: 0.979
```

Der Punktationswert „alle Variablen in der Tabelle.“.

👉 Hey! Das ist ja fast perfekte Modellgütigkeit!

⚠️ Vorsicht: Wenn ein Angebot aussieht wie „true“, dann ist es meist auch too good to be true.

10. Geradenmodelle 2

Der Grund für den fast perfekten Modell ist, dass der Titel die Auktionen aufsagen bzw. das Verkaufsgebot präzise beschreibt. Der nützliche Titel der Auktionen ist jedoch nicht in den Auktionen selbst enthalten. Im Test-Sample werden unsere Vorhersagen schlecht sein, wenn wir uns auf die Titel der Auktionen stützen. Merke: Höchstidiografische Informationen etc. sind nicht nützlich, um allgemeine Musterrichtungen zu erkennen. Exakte Prognosen zu erstellen.

Probieren wir also die Vorhersage im Test-Sample:

```
pred1 <- predict(mariokart_test)
## Error in eval(predvars, data, env): object
```

Oh nein! Was ist los? Eine Fehlermeldung!

Nominalskalierte Prädiktorenvariablen mit % sind problematisch. Kommt es zu Ausprägungen vor, die nicht im Train-Sample gab, so resultiert eine Häufung, die sinnvoll, auf diese Variablen oft zu Overfitting führen.

{ %

Okay, also auf die Titelspalte sollten wir Nächster Versuch.

```
mariokart_train2 <-
mariokart %>% train
  sel(exclude, V1, id))
```

Wir entfernen außerdem Spalten ebenfalls keine Titeln bergen.

10.5. Fallbeispiel zur P

```
l m _ a l l i n _ l (n o t _ o f f i a l _ b e p t _ m a r i o k a r t _ t r a i n 2 )  
r ( 1 m _ a l l i n _ n o _ t i t l e )  
# # # R2 for Linear Regression  
# # R2: 0.521  
# # adj. R2: 0.441
```

Das R-Quadrat ist ja durchaus ordentlich. Schrms(edie SD der Vorhersagefehler) an; der Bef
performativität des Metasystems

👏 Gut gemacht!

```
perf o r m a n s ( d e m _ a l l i n _ n o _ t i t l e )  
# # [ 1 ] 20
```

Sie rennen zu ihrem Chef, der jetzt die Güte ih
restlDa bestimmen soll.

👉 Da wir dein Modell in diesem Teil des K
Datensatzes nennen wir diesen Teil das „T
Sample“.

Ihr Chef schaut sich die Verkaufspreise im Te

```
m a r i o k a % & % t e s t  
s e l ( e i c h , t o t a l _ p r )  
h e a d
```

i d	t o t a l _ p r
1 . 2 e + 11	37
2 . 9 e + 11	55
1 . 8 e + 11	56
1 . 8 e + 11	56
3 . 5 e + 11	65

10. Geradenmodelle 2

i d	t o t a l _ p r
1 . 1 e + 1 1	4 6

😊 Okay, hier sind die ersten paar echten
Jetzt mach mal deine Vorhersagen auf Basis
Modells!

Hier sind Ihre Vorhersagen (engl. predictions):

```
m_allin_predictions
```

Hier sind Ihre ersten paar Vorhersagen:

```
head(m_allin_predictions)
## 1 2 3 4 5 6
## 29 54 53 54 42 47
```

Diese Vorhersagen fügen wir noch der Ordnung mit den Test-Daten:

```
mariokart_test <-
marioka%>%test
matrialline_predictions %>%
  mariokart_test))
```

😊 Okay, was ist jetzt der mittlere Vorhersa-

Um die Vorhersagegenauigkeit im Test-Sample auszurüsten, nutzen wir die Funktion `mean`. Sie vielleicht noch in einem anderen Paket gefunden haben. Sie vielleicht noch in einer anderen Funktion gefunden haben.

```
library(yardstick)
```

```
yardstimate(akt_mariokart_test,
            truth_table(epchter Verkaufspreis
                        esti_mameal_lin_p#leidirections)sage
yardstimate(akt_mariokart_test,
            truth_table(epchter Verkaufspreis
                        esti_mameal_lin_p#leidirections)sage
```

. metric . estimator . estimate	
mae standard	10

. metric . estimator . estimate	
rms e standard	13

Ihr mittlerer Vorhersagefehler (RMSE) liegt
dagegen haben wir ~~die~~ ~~die~~ : r mse geschrägt und nun
da es sowohl im Paket ~~die~~ als auch des ~~die~~
auch im ~~die~~ Paket ~~die~~ des ~~die~~ Befehl des
Name ~~die~~ gebt. Name - Clash - Alarm! Risiken daher die
meinen als Sie, was garantiert zu Verwirrung
oder bei Ihnen.)

👉 Ganz okay.

Wie ist es um das R-Quadrat Ihrer Vorhersagen?

```
# `rsq` ist auch aus dem Paket yardstick:
rsq(akt_mariokart_test,
    truth_table(epchter Verkaufspreis
                esti_mameal_lin_p#leidirections)sage
```

10. Geradenmodelle 2

. metric	. estimator	. estimate
r sq	standard	0.17

👉 17%, nicht berauschend, aber immerhin

Wie das Beispiel zeigt, ist die Modellgüte geringer im Train-Sample. Die Modellgüte im unter überrmäßig optimistisch. Dieses Phänomen nennt man Vorhersagen eines Modells sich, die Modellgüte in einem neuen Datensatz zu überprüfen.

Wenn Sie eine robuste Schätzung der Güte möchten, bietet sich folgendes Vorgehen an:

- 1.Teilen Sie Ihren Datensatz (das Train-Set, sog. Validation-Sample und das sog. Assessment-Set)
- 2.Berechnen Sie Ihr Modell im ersten Teil (Validation-Sample)
- 3.Prüfen Sie die Modellgüte im zweiten Teil (Assessment-Sample)

Diese Aufteilung Ihres Datensatzes ist auch Validierung (validation split); Sie können sie selbst vornehmen:

```
library(rasample)
mario_kraeratd<-read_csv("mario.csv") # Wenn die CSV-Datei in
# Unterordner mit Namen "daten" liegt
meine_aufteilung<-sample_split(mario_kraeratd, p = 0.7)
```

10.6. Vertiefung: Das Aufteilen

initial_split⁶ teilt jede Zeile (Beobachtung) zufällig in die Zeile in das Analyse- oder in das Assessment-Teil. Im Standard werden 75% der Daten in das Analyse-Assessment-Sample gesetzt, es ist jedoch sinnvoll die Aufteilung des Samples zu ändern. Argumentation: Es wäre nämlich absurd für Ihr Sample z.B. nur die teuren, und im Test-Sample nur die günstigen Landen würden. Anderes Beispiel: In den ersten Zehn Zeilen aus Land A und in den unteren Zeilen nur aus Land B würde sich Ihr Modell unnötig schwer tun.

Im nächsten Schritt können Sie mit `initial_split` die gewünschte Aufteilung der Daten festlegen. Nur, welche Zeile in welche der beiden Stichprobenteile fällt? Die eigentliche Aufteilung wird aber noch nicht durchgeführt.

```
mariokart_train <-  
  train(mengne_aufteilung) Sample  
mariokart_test <-  
  test(mengne_aufteilung) assessment - Sample
```

Ich persönlich nenne die Tabelle `initial_split`. Das ist kürzer zu tippen und einheitlich. Auch ein eigenes Namens-Schema nutzen; was auf die Konsistenz in der Benamung, außerdem Kürze und Namen.

% % \$

Das Train-Sample stellt die bekannten Daten dar, die wir lernen, d.h. unser Modell berechnen. Das Problem der wirklichen Welt dar: Neue Beobachtungen (noch) nicht weiß, was der Wert der AV ist. Der Zettel unterscheidet verschiedene, aber zusammengehörige Arten von Abildung 10.12 dargestellt.

⁶vgihelp(initial_split)

10. Geradenmodelle 2

Definiert man einen Datensatz, für die Sie als Auskönder haben, nennt man Train-Sample.

Definiert es ein Datensatz, auf dem Sie der UV, aber nicht zu der AV vorliegen haben
□

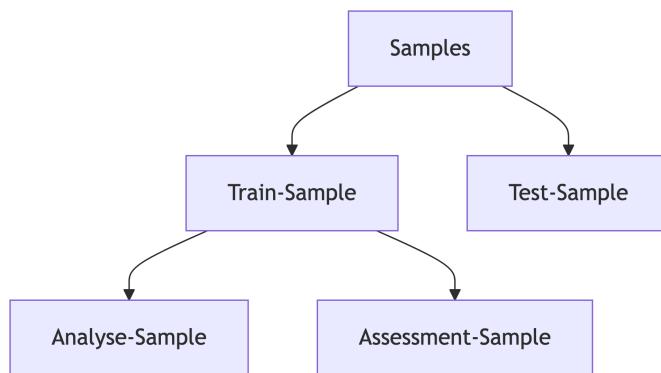


Abbildung zeigt die Arten von Zusammengesetzten Probenarten im Rahmen einer Prognose!

!

Ein Anwendungsbezug von moderner Datenanalyse, welche Kunden „abwanderunggefährdet“ Zukunft bald nicht mehr unsere Kundensind gibt eine ganze Reihe von Untersuchungen et al. (2022). Die Forschis versuchen anhand der linearen Regression vorherzusagen, wie sein werden. Die Autoren berichten von einem in Ihrem (besten) Vorhersagemodell.

10.8. Wie man mit Statistik

(\$

!

Ein Forscher-Team untersucht Pinguine von der Arktis. Das Team ist am Zusammenhang zwischen Schnabellänge und Schnabbedeckung interessiert, s. Abbildung 10.13.

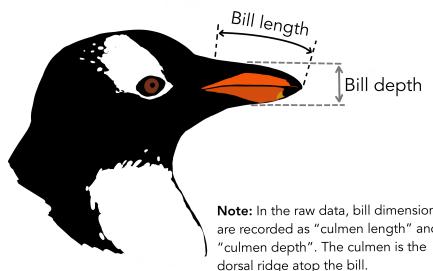


Abbildung 10.13.: Schnabellänge und Schnabbedeckung

Das Team beschreibt die Eiszeiten 34 400 Jahre und vergleicht die Tiere vermessenen bei antarktischen Temperaturen

```
penguins_path <-  
  "https://vincentarelbundock.github.io/"  
  "Rdatasets/csv/palmerpenguins/penguins.csv"  
  
penguins <- read_csv(penguins_path)
```

Man untersucht, rechnet und überlegt. Ah! Jetzt ist es ein Fall! Eine Eingangsfrage: Gibt es einen Zusammenhang von Schnabellänge und Schnabbedeckung? Das ist bestimmt einer der ersten Schritte, um die Ergebnisse zu publizieren!

10. Geradenmodelle 2

gg scatterplot(x="bill_length_mm", y="bill_depth_mm"
+ add="reg") # ase "ggpubr"

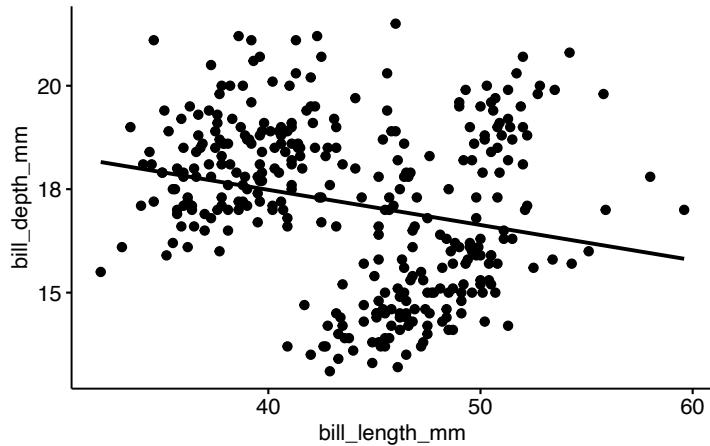


Abbildung 10.14 veranschaulicht den negativen Zusammenhang von Schnabeltiefe und Schnabellänge.

Hier sind die statistischen Details, s. Tabelle 10.1.

Im Bild (10.14) ist der negative Zusammenhang von Schnabeltiefe und Schnabellänge.

Tabelle 10.1: Koeffizienten des Modells 1: Negativer Zusammenhang von Schnabeltiefe und Schnabellänge.

Parameter	Koeffizient
(Intercept)	20.89
bill_length_mm	-0.09

Kurz darauf veröffentlichte eine verfeindete Satzung gleichen Themas. Gegenetbeitlinge An-

10.8. Wie man mit Statistik

gebnis sieber Raus untersuchten) Pingui nengilt positiv Zusammenhang von Schnabel länge und Schnabellung 10.15.

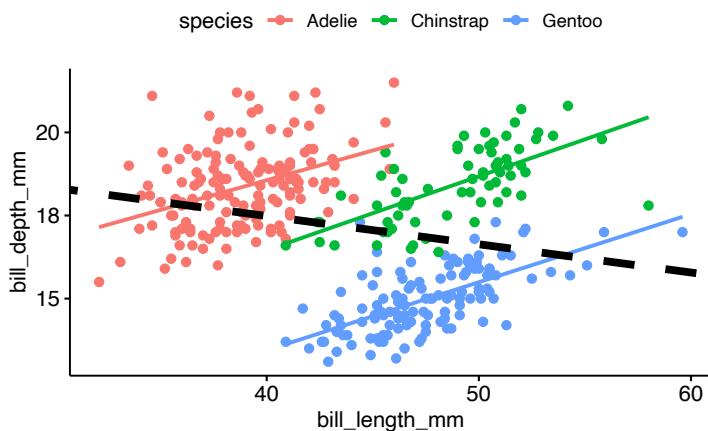


Abbildung 10.15 Zusammenhang von Schnabel länge bei Tiefe pro Gruppe von Pinguinen: Die Gruppe pro Gruppe steigt. Hingegen si sionsgerade ohne Beachtung der Gruppe gestrichelte Linie)

Ohnein! Was ist hier nur los? Datenlügen nicht Hier sind die statistischen Details der zweiten Lmz weiter mit dem gleichen Prädiktor neu in (zusätzlich zur Schnabel länge).

I m_pinguin(bill_length_mm ~ bill_depth_mm)

Tabelle 10.15: Koeffizienten des Modells 2: Positiver Verlength_mm

Parameter	Coefficient
(Intercept)	10.6
bill_length_mm	0.2

10. Geradenmodelle 2

Tabelle 1: Koeffizienten des Modells 2: Position
Length_mm

Parameter	Coefficient
species Chinstrap	-1.9
species Gentoo	-5.1

Ohne Hintergrundwissen oder ohne ausreichende Unterschiede werden, welche Analyse - Gesamt oder dierichtigeist. Nicht-experimentelle Studien führen, je nachdem ob Pinguin zugetragen oder weggenommen werden.

f /

E Wichtig

Interpretieren die Modellkoeffizienten kann man so:

Nur wenn man die Ursache-Wirkungs-Beziehungen kennt, macht es Sinn, die Modellkoeffizienten zu interpretieren. Andernfalls lässt man besser die Finge Modellkoeffizienten und begnügt sich mit der Güte und mit Vorhersage (synonym: Prognose). „Der betrachtete Abbildung 10.16, im linken Bild auf und interpretiert man, X hat einen starken Effekt auf Y!“.

In der nächsten Studie nimmt der Forscher die Gruppe (B. Geschlecht) in das Modell ein. Jetzt ist auf einmal nicht mehr stark positiv, sondern in jeder Gruppe, s. Abbildung 10.16, re

Dieses Umschwenken der Regressionslinie ist nicht kausal, ist „nicht echte“, also nicht-kausale Zusammenhänge also, so können sich die Modellkoeffizienten verändern (sogar das Vorzeichen kann wechselt). Simpsons Paradoxie kann das Modell verändert, also hinzufügt oder aus dem Modell entfernt.

Wenn man die kausalen Abhängigkeiten nicht kennt, ob die Zusammenhänge kausal oder nicht-kausal sind, ob die Modellkoeffizienten belastbar oder nicht.

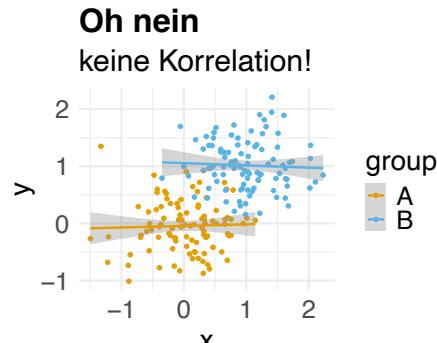
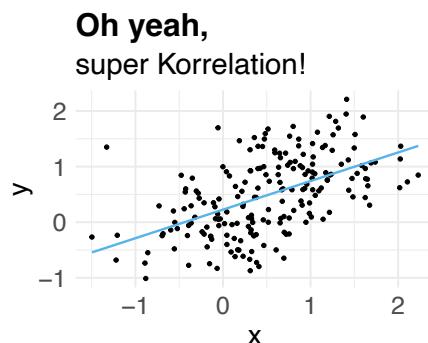


Abbildung zeigt man in ein Modell eine Variable, die sich die Koeffizienten massiv ändern. In Abbildung (a) wurden die gleichen Daten verwendet, um einen starken positiven Zusammenhang, (b) keinen Zusammenhang in beiden Gruppen zu erhalten.

Man könnte höchstens sagen, dass man (wenn man nicht kennt) die Modellkoeffizienten verändert, z. „Dort wo es viele Störche gibt“⁷ Lgeibdte es auch vielerlei auf kausale Zusammenhänge geprägt. Zusammenhänge nicht kausal zu interpretieren.

⁷Das Störche-Babys-Beispiel passt auch zu Abbildung 1

10. Geradenmodelle 2

tive Befunde immer wieder unzulässig kausal und Wissenschaftlern auch.

In diesem Kapitel haben Sie lineare Modelleche Modelle gelernt. Dazu gehören modelle, das sind Modelle mit mehr als einer UV (Praktionsmodelle). Außerdem haben Sie sich mit gesamtgesellschaftlichen Nutzen beschäftigt. Spiel zum Schluss war vielleicht erhellend. Modell im Train-Sample nicht (notwendig) Test-Sample führt.

Wenn Sie dranbleiben an der Statistik, wir s. Abbildung 10.17.



Abbildung 10.17. Sie und Party: Gestern gen. Wenn Sie dranbleiben, wird die Freund (imgflip, 2024 b).

Die Webseite datenwerk.netlify.app stellt eine Übungsaufgaben bereit. Sie können die Suchfunktion nutzen, um die Aufgaben mit den folgenden Namen zu finden:

- interpret-koeff-lm
- Aussagen-einfache-Regr
- interpret-koeff
- regression1b
- mtcars-regro1
- regression1a
- lm1
- Regression5
- Regression6
- lm-mario1
- lm-mario2
- lm-mario3
- ausreisser1
- mario-compare-models

→

Ein empfehlenswertes Buch für Regressionsanalyse ist von Andrew Gelman und Jennifer Hill (Gelman et al., 2021). Sein Buch ist für Sozialwissenschaftler nicht typisch, hat aber deutlich mehr Kapitel.

7

- Ainali. Standard Deviations [Arbeitsp]Mitrop
commonswiki media.org/w/index.php?cur
Anscombe, F. J. (1973). Graphs in Statistical Sit
tis t i2@i1@n 17-21.
- Arad, C. (201241, JaunnNb5a)pp: Gehalt und Vermög
(2024)td. de. https://www.ftd.de/vermo
vermögen/
- Barrett, Mg (2024)to: 'Okabe-Ito' Scales fo
[Manual]. https://CRAN.R-project.org/p
Berger, G. (2019,T Dezebsorff 1TQ)morrow: Link
Emerging JobsRteppso:r/t/www.linkedin.com/b
career/the-jobs-of-tomorrow-linkedin
report
- Bortz, J., & Schüßtaetris, C. (2010)uman- Und So
schaf Springer. https://doi.org/10.100
Bowne-Anderson, H. (2018). What Data S
According to 35 DataScientists Business Re
https://hbr.org/2018/08/what-data-s
according-to-35-data-scientists
- Bromann, K. W., & Woo, K. H. (2018). Data Orga
The American S 7@1)s t2i-c1iOahn https://doi.
00031305.2017.1375989
- Bundesamt, S. (2023-0OK3ö-r2p7e2r0m2a3B-e0@fie 27A)l.
gruppen und Gesellschaftsschafes Bundesamt.
destatis.de/DE/The men/Gesell schaft-U
Gesundheitszustand-Relevantes-Verhal
koerpermasse.html
- Bundesbank, D. ou(s2e0h2o3I)d Wealth and Financ
many: Results of the 2021 Haubehb\$dhWea

Literatur

- Bundesbank. <https://www.bundesbank.de/908924/3ef9d9a4eaeae8a8779cc3ac464970vermoegensbefragung-data.pdf>
- Cetinkaya-Rundel, M. J. & Haordduint, I. B. n (2020) *Statisti*. <https://openintro-im.netlify.app/>
- Cetinkaya-Rundel, M., Diez, D., Bray, A., Kim, Paterno, N., & BaperenCn (2020) *data sets and Support Functions from 'OpenIntro' Textbook CRAN R-L project.org/package=openintro*
- Cmglee. *Geometric Visualisation of Example Distribution (2, 4, 4, 4, 5, 5, 7, 9)* <https://commons.wikimedia.org/w/index.php?curid=2911558>
- Cohen, J. (1992). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.
- Dennis Boigelot (2011) *Redesign File: Correlation Using Vector Graphics*. <https://commons.wikimedia.org/w/index.php?curid=30211>
- Downey, A. P. (2020) *Overthinking It: How to Use Questions, Avoid Statistical Trap, and Make Decisions*. University of Chicago Press.
- Fisher, D., & Meyer, M. (2018) *Practical Data Visualization: A Practical Guide to Seeing Through Data*. Flatiron.
- Gelman, A., Hill, J., & Deveaux, R. (2017) *Statistics and Data Analysis*. Cambridge University Press.
- Goren, A., Vaño-Galván, S., Wambier, C. G., McAllister, A., Moreno-Arribes, O. M., Shapiro, J., Sineiro, Kovacevic, M., Mesinkovska, N. A., Goldust, M. (2020) *A Preliminary Observation: Male Pattern Hair Loss in COVID-19 Patients in Spain - A Potential Androgenetic COVID-19 Risk Factor*. *Selvøffic Toys met,ic Dermatology*, 1545–1547. <https://doi.org/10.1111/j.1365-2760.1971.tb01261.x>
- Haug, S., Castro, R. P., Kwon, M., Filler, A., Kell, S. (2015). *Smartphone Use and Smartphone Addiction*.

Literatur

- People in Switzerland. *Behavioral & Social Sciences*, 40(4), 2001-2017. <https://doi.org/10.1556/2006.4.2015.0>
- Hornik, K., Ligges, U., & Zeileis, A. (2020). *Journal*, 295-296.
- Horst, A.T. (2020) [rat work]. <https://allisonhorst.com/>
- Horst, A.S. (2024) [Art work]. <https://allisonhorst.com/>
- Hou, J., Walsh, P. P., & Zhang, J. (2015). The development of the Social Cognitive Model of Social Colors for All Colors. *Journal of Universal Design*, 9(3), 2014-2017. <https://doi.org/10.1016/j.soscij.2014.07.001>
- Ichihara, Y. G., Okabe, M., Iga, K., Tanaka, T., & Yamamoto, S. (2014). Color Universal Design: The Selection of Functional Colors for All Colors. *Journal of Universal Design*, 9(3), 2014-2017. <https://doi.org/10.1016/j.soscij.2014.07.001>
- Hardcopy, and A680D7206f208s <https://doi.org/10.1016/j.soscij.2014.07.001>
- imgflip. (2024) [Cartoon] <https://imgflip.com>
- imgflip. (2024) [Meme] <https://imgflip.com>
- imgflip. (2024) [Meme] <https://imgflip.com>
- International Transparency Index. https://www.transparency.org/corruption_perceptions_index_2016
- Ismay, C., & Kim, S. (2021) [Conference via Modern Dive into Random Forests with Trasylor & Friends Group]. <https://modern-dive.com/>
- Kaplan, D. (2020) [Modeling Attribution Models]. <https://dtkaplan.github.io/SM2-book/docs/>
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). How Predictable are Digital Tributes? *Proceedings of the National Academy of Sciences*, 110(45), 18580-18585. <https://doi.org/10.1073/pnas.1218772110>
- Kwon, M., Kim, D.-J., Cho, H., & Yang, S. (2017). The Perception Scale: Development and Validation of a Self-Perception Scale. *Journal of Clinical Psychology in Practice*, 13(2), e83558. <https://doi.org/10.1037/cpr.0083558>
- Lalwani, P., Mishra, M. K., Chadha, J. S., & Churn Prediction System: A Machine Learning Approach. <https://doi.org/10.1037/cpr.0083558>

Literatur

- 10(42), 271–294. <https://doi.org/10.1007/s00120-017-0505-1>
- Lieberoth, A., Rasmussen, J., Stoeckli, S., T. H., Lin, S.-Y., Tuominen, J., Travagliano, G. COVIDI STRESS GI. <https://doi.org/10.17605/Z39US>
- Lieberoth, A., Rasmussen, J., Stoeckli, S., T. H., Lin, S.-Y., Tuominen, J., Travagliano, G. COVIDI STRESS GI. <https://doi.org/10.17605/Z39US>
- Lovett, M. C., & Greenhouse, J. B. (2000). Appendix to Statistical Methods in Medicine. 1961–2006. <https://doi.org/10.1080/00031305.2000.10481653>
- Lyon, A. (2014). Why Are Normal Distribution Journal for the Philosopphy 201f–S4i9e nhctet ps: / <https://doi.org/10.1093/bjps/axs046>
- M7. (2000a4v) inelli's Italiyah \$mek'ing oRimpos. https://www.wikimedia.org/wiki/File:Pipa_savinelli.jpg
- Mackay, R. J., & Oldford, R. W. (2000). Scientific Method and the Special Case of Statistics. <https://doi.org/10.1214/ss/1009212817>
- Maprhy. (2000e0s9a) w with AMetawork]. <https://www.wikimedia.org/w/index.php?curid=79390659>
- Marks-Anglin, Arielle and Chen, Yong. (2020) PublicatResBaresh Synth(\$6)s, Me25h-07d4s2. <https://doi.org/10.1002/jrsm.1452>
- Matthews, R. (2000). Storks Die in the Basalt P. tti, 22, 36–38. <https://doi.org/10.1111/1467-9531.00222>
- Menk. (2014) i huelair Rg. the \$pisor / t example. net / examples/linear-regression/
- Messerli, F. H. (2012). Chocolate Consumption Nobel Lancet England Journ, 36(71f6Me d156i2n-e1 564. <https://doi.org/10.1056/NEJMMon1211064>
- Mittag, H. - J., & Schältl, K.: E2020Einführung aktiven El. eSpenni negner. <https://doi.org/10.161912-4>
- Moore, B. (2015, April 9) the Vaccination Health Economics. <https://benjamini-moore.wordpress.com/>

L i t e r a t u r

- recreating-the-vaccination-heatmaps-i
Mulukom, V. van, Muzzolini, B., Rutjens, E.
M. (2020) chological Impact of hCtOtpSD:-/9dPoi
org/10.17605/OSF.IO/TSJNB
- Obels, P., Lakens, D., Coles, N. A., Gottf
Analysis of Open Data and Computational
gistered Reports Aidn Basyskoilnology Methods and
ces in Psychol p36128128012872 de <https://doi.org/10.2515245920918872>
- Oestreich, M., & Rothkötter, P. n (2014) Statist
und Spaß im Horrorfach nicht espmii rscher.
<https://doi.org/10.1007/978-3-658-046>
- Okabe, M., & Itoqikor (2020/23)e)ral Design (CUD)
Barri er hFtrypes: // j f ly. uni - koeln. de / col o
- Pearl, J., & Mackenzie (2018) The New Sc
Cause and Ef ffect Books.
- Plessner, H. E. (2018). Reproducibility vs
of a Confused T erraindeleargsy. n Neut, b76 form
<https://doi.org/10.3389/fninf.2017.00>
- Poldrack, R. S tAat(i2s0t23c)al Thinking: Analyzi
certain WorlIndet on University Press. ht
github. io / statsthinking21 - core - site /
- Roser, M., Appel, C., & Ritchie Our Wo(r2013n).
Data <https://ourworldindata.org/human->
- Rothstein, H. R. (2014) Wi IPeub\$t@\$R@n:B\$@st
ference. Ond hm Wiley. <https://doi.org/10.>
stat07071
- Sauer, S. D 2008) : Predictors of PerDfadraman o
set]. Open Science Framework. <https://doi.org/10.>
SJHUY
- Sauer, S. M 2019) . Datenanalyse mit R: Daten
ten, visualisieren und modell hitepen // www
com/de/book/9783658215866
- Scherer, C., Radchuk, V., Staubach, C., Müll
& Kramer Schadt, S. (2019). Seasonal Hos
Fuel Disease Dynamics at Dijfouem@nto\$paAntiim
Ecol, 80(1), 1812-1824. <https://doi.org/10.>

Literatur

- Shimizu, Y. (2022). Multiple Desirable Methods: Univariate Data With Respect to Research Questions. *Psychology & Psychopathology*, 13(1), 81–98. <https://doi.org/10.3389/fpsyg.2022.819854>.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). *Psychology: Undisclosed Flexibility in Data Allows Presenting Any t*. *Psychological Science in the Public Interest*, 12(1), 135–136. <https://doi.org/10.1177/1524837211426111>.
- Spurzem, L. V. (2021, March 7). von Wikinghtnp: 87 de. [https://en.wikipedia.org/wiki/Modellautomobil#/media/File:Modell_VW_1303_\(um_1975\).JPG](https://en.wikipedia.org/wiki/Modellautomobil#/media/File:Modell_VW_1303_(um_1975).JPG)
- Stigler, S. T. (2009). *Pillars of Statistical Wisdom*. Cambridge University Press.
- TransfermarktDie (2024). https://www.transfermarkt.de/spieler-statistik/marktwertetop/spielerposition_id/8/page/1
- van Panhuis, W. G., Grefenstette, J., Jung, S., Eng, H., Lee, B. Y., Zadorozhny, V., Brown, S., & D. S. (2013). Contagious Diseases in the United Kingdom. *Journal of Medical Microbiology*, 62, 1521–1528. <https://doi.org/10.1056/JMM.1215400>
- Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2014). The Mere Presence of One's Own Smartphone Reduces Cognitive Control of the Association of Consumption. *Journal of Consumer Psychology*, 24, 140–154. <https://doi.org/10.1080/10648024.2014.891462>
- Wichert, J. M., Veldkamp, C. L. S., Augustijn, R. C. M. van, & Assen, M. A. L. M. van. (2016). Data Planning, Running, Analyzing, and Reporting: A Checklist to Avoid Poor Handling of Data. <https://doi.org/10.3389/fpsyg.2016.01832>
- Wickham, H. (2022). *Elegant Graphing with Data* (2nd edition). Springer.
- Wickham, H. (2023). *Data Structures*. <https://r4ds.had.co.nz/data-tidy.html#fig-tidy-structure>
- Wickham, H., & Grolemund, D. (2018). *Daten mit R bearbeiten, bereinigen, umformen, mitschreiben und visualisieren* (1. Auflage). O'Reilly. <https://rforcats.had.co.nz/>
- Wilke, C. R. (2019). *Fundamentals of Data Visualization: A Practical Guide*. Springer.

L i t e r a t u r

- Informativ e and Compoé IRleii n g y i gnttøss : // c
com/ dataviz /
Wi l k e , C.W(2 0 2 4)a.b / P.r Wi c k e a l a ñ g h t t p s : / / g i t
wi l k e l a b / p r a c t i c a l g g (O r i g i n a l w o r k p u b
Wi l k e , S . (2 0 T 1 r 3 e n J d u s n d e 2 6 L) u f t B e m p e h a t u t U m w e
b u n d e s a m t ; U m w e l t b u n d e s a m t . h t t p s : / / w v
d e / d a t e n / k l i m a / t r e n d s - d e r - l u f t t e m p e r a
Wor l d E c o n o m i c F o r u m F u (t 2 u 0 r 2 e 0) f . J o b s W R e p o r t 2
E c o n o m i c F o r u m . h t t p s : / / w w w 3 . w e f o r u
F u t u r e _ o f _ J o b s _ 2 0 2 0 . p d f

◦ f

- Abweichungsrechnik on 8. 1, s. S. 188
- Argument einer Deutung 3. 4, s. S. 55
- Auspräg De g n i t i o n 2. 8, s. S. 25
- Balkendiag raf m i t i o n 5. 3, s. S. 114
- Beobachtung s D e e f m i t i o n 2. 6, s. S. 24
- Binäre Variablen 10. 2, s. S. 243
- Boxplot D e f n i t i o n 5. 10, s. S. 131
- Data-Dict De f a r y i o n 2. 4, s. S. 24
- Dataframe D e f n i t i o n 3. 6, s. S. 66
- Datei D e f n i t i o n 2. 3, s. S. 23
- Datendiag raf m i t i o n 5. 1, s. S. 109
- Datenj W e f n i t i o n 4. 1, s. S. 78
- Dezimaldefinition 6. 6, s. S. 157
- Dichtediag raf m i t i o n 5. 5, s. S. 117
- Das einfache linidefektiv Modell 9. 3, s. S. 211
- Extremwertdefinition 6. 3, s. S. 150
- Fehlerstreuung 9. 4, s. S. 216
- Funktion D e f n i t i o n 3. 2, s. S. 53
- Geradlinigkeit 9. 2, s. S. 210
- Histogrammdefinition 5. 4, s. S. 116
- Interaktion D e f n i t i o n 10. 4, s. S. 258
- Interquartile D e f n i t i o n 7. 4, s. S. 172
- Konfidenzintervall 10. 1, s. S. 239
- Kovariation D e f n i t i o n 8. 2, s. S. 190
- Lagemitteldefinition 6. 8, s. S. 159
- Linearer Zusammenhang 5. 8, s. S. 125
- Lineares Modell 6. 2, s. S. 150
- Mittlere Absolute Abweichung 3, s. S. 171
- Median D e f n i t i o n 6. 4, s. S. 152

A. Definitionen

- Model Definition 2.11, s. S. 37
- Multiple Regression 10.3, s. S. 255
- Mittelwertdefinition 6.1, s. S. 144
- Normalverteilung 5.6, s. S. 121
- Entstehung einer Normalverteilung s. S.
- Nullmodell (Punktmodell) 9.1, s. S. 208
- Pfeildefinition 4.2, s. S. 94
- Punktmodelldefinition 6.9, s. S. 160
- Quantiledefinition 6.7, s. S. 158
- Quartiledefinition 6.5, s. S. 157
- Korrelationskoeffizient 8.3, s. S. 194
- R-Quadratdefinition 9.5, s. S. 218
- Spannwertdefinition 7.2, s. S. 170
- Reproduzierbarkeit 3.1, s. S. 44
- Residualdefinition 2.2, s. S. 17
- Standardabweichung 7.6, s. S. 177
- Skalenniveau 2.10, s. S. 32
- Statistikdefinition 2.1, s. S. 16
- Streuungsschaetzung 7.1, s. S. 170
- Test-Sampledefinition 10.6, s. S. 271
- Tidy Datadefinition 2.9, s. S. 26
- Train-Sampledefinition 10.5, s. S. 271
- Variablendefinition 7.5, s. S. 177
- Variablendefinition 7.5, s. S. 177
- Vektorielles Rechnen 3.5, s. S. 59
- Vektordefinition 3.3, s. S. 54
- Wertdefinition 5.2, s. S. 113
- Wertdefinition 2.7, s. S. 25
- z-Wertdefinition 7.8, s. S. 183
- Zentriierung 7.7, s. S. 181
- Richtung und Stärke einer Beziehung 126

l i b(r k a n i y t r)
l i b(r k a a b l e E x t r a)

In diesem Datensatz werden Auktionen in den USA im Oktober 2009 gesammelt. Es handelt sich um eine Beobachtung (Auktionen) und 12 Spalten (Variablen) in US-Dollar angegeben.

Die Quelle des Datensatzes ist (Ceterdossen Ray Park Rund et al., 2024).

Tabelle B. 1 zeigt das Data-Dictionary.

Mit Hilfe (oder mit der Hilfe) einer Hilfeseite zum Datensatz muss das Projektteam richtig gestellt sein).

B. Data-Dictionary für Mariokart

Tabelle B.1.: Data-Dictionary für M

V a r i a b l e	E r k l ä r u n g
i d	I D der Auktion
d u r a t i o n	Dauer der Auktion in Tagen
n _ b i d s	Anzahl der Gebote
c o n d	Zustand (new/used)
s t a r t _ p r	Anfangspreis bei der Auktion
s h i p _ p r	Versandgebühr
t o t a l _ p r	Gesamtpreis (inkl. Versand)
s h i p _ s p	Versandmethode bzw. -geschwindigkeit
s e l l e r _ r a t e	Bewertung des Verkäufers; zwischen positiven und negativen Bewertungen.
s t o c k _ p h o t o	Lag der Auktionen in "stock photo". Foto in vielen Auktionen benutzt und "stock photo" genannt.
w h e e l s	Anzahl der enthaltenen Wi i - Räder
t i t l e	Name der Auktion