

eviangelion



Statistik 1

für Biologie, Gesundheitswissenschaften und
Pharmazeutische Wissenschaften or not.

i've figured it out, yeeahh

Markus Kalisch, Peter Bühlmann und Hansruedi Künsch
Seminar für Statistik
ETH Zürich ✓

Februar 2017

Inhaltsverzeichnis

1	Einführung (Stahel, Kap. 1)	1
1.1	Was ist Statistik?	1
1.2	Kann ich Statistik überhaupt brauchen?	3
1.3	Was ist der Inhalt dieses Kurses?	5
1.4	Software	6
1.5	Literatur	6
2	Modelle für Zähldaten	7
2.1	Wahrscheinlichkeitsmodelle (Stahel, Kap. 4.1, 4.2)	7
2.1.1	Wahrscheinlichkeiten berechnen	9
2.2	Der Begriff der Unabhängigkeit (Stahel, Kap. 4.6)	10
2.3	Bedingte Wahrscheinlichkeit (Stahel, Kap. 4.7)	11
2.4	Zufallsvariable (Stahel, Kap. 4.3, 4.4)	14
2.5	Binomialverteilung (Stahel Kap. 5.1)	16
2.6	Kennzahlen einer Verteilung (Stahel Kap. 5.3)	18
2.6.1	Kumulative Verteilungsfunktion	20
2.7	Weitere Verteilungen (Stahel Kap. 5.2)	21
2.7.1	Poissonverteilung	21
2.7.2	Hypergeometrische Verteilung	23
2.8	Software	24
3	Statistik für Zähldaten	25
3.1	Drei Grundfragestellungen der Statistik (Stahel Kap. 7.1)	25
3.2	Schätzung, statistischer Test und Vertrauensintervall bei Binomial- Verteilung (Stahel Kap. 7.2, 8.2, 9.1, 9.2)	26
3.2.1	(Punkt-)Schätzung	26
3.2.2	Statistischer Test	28
3.2.3	Vertrauensintervall	37
3.3	Software	40
4	Modelle und Statistik für Messdaten	43
4.1	Lernziele	43
4.2	Einleitung	44
4.3	Deskriptive Statistik (Stahel, Kap. 2 und 3.1, 3.2)	45
4.3.1	Kennzahlen	45
4.3.2	Grafische Methoden	48

4.3.3	Analogien zwischen Modellen und Daten	50
4.4	Stetige Zufallsvariablen und Wahrscheinlichkeitsverteilungen (Stahel, Kap. 6.1 – 6.4, 11.2)	51
4.4.1	(Wahrscheinlichkeits-)Dichte	52
4.4.2	Kennzahlen von stetigen Verteilungen	53
4.5	Wichtige stetige Verteilungen (Stahel, Kap. 6.2, 6.4, 6.5, 11.2)	53
4.5.1	Uniforme Verteilung	53
4.5.2	Exponential-Verteilung	54
4.5.3	Normal-Verteilung (Gauss-Verteilung)	55
4.5.4	Funktionen einer Zufallsvariable	57
4.5.5	Überprüfen der Normalverteilungs-Annahme	60
4.6	Funktionen von mehreren Zufallsvariablen (Stahel, Kap. 6.8 – 6.11)	62
4.6.1	Unabhängigkeit und i.i.d. Annahme	63
4.6.2	Kennzahlen und Verteilung von \bar{X}_n	63
4.6.3	Verletzung der Unabhängigkeit	67
4.7	Statistik für eine Stichprobe (Stahel, Kap. 8.3 – 8.5, 9.3)	68
4.7.1	(Punkt-) Schätzungen	68
4.7.2	Tests für μ	69
4.7.3	Vertrauensintervall für μ	75
4.7.4	Tests für μ bei nicht-normalverteilten Daten	76
4.8	Tests bei zwei Stichproben (Stahel, Kap. 8.8)	78
4.8.1	Gepaarte Stichprobe	78
4.8.2	Ungepaarte Stichproben	79
4.8.3	Weitere Zwei-Stichproben-Tests bei ungepaarten Stichproben	83
4.9	Versuchsplanung (Stahel, Kap. 14.1 - 14.2)	83
4.10	Software	85
4.10.1	Verschiedenes	85
4.10.2	Zwei-Stichproben t-Test für ungepaarte Stichproben	86
4.10.3	Zwei-Stichproben t-Test für gepaarte Stichproben	87
4.10.4	t-Test für eine Stichprobe	88
5	Regression	89
5.1	Lernziele	89
5.2	Einfache lineare Regression	89
5.2.1	Das Modell der einfachen linearen Regression	91
5.2.2	Parameterschätzungen	93
5.2.3	Tests und Konfidenzintervalle	94
5.2.4	Das Bestimmtheitsmass R^2	96
5.2.5	Allgemeines Vorgehen bei einfacher linearer Regression	97
5.2.6	Residuenanalyse	97
5.3	Multiple lineare Regression	101
5.3.1	Das Modell der multiplen linearen Regression	102
5.3.2	Parameterschätzungen und t-Tests	103
5.3.3	Der F-Test	104
5.3.4	Das Bestimmtheitsmass R^2	104

5.3.5	Residuenanalyse	105
5.4	Software	105
5.4.1	Einfache lineare Regression	105
5.4.2	Multiple Lineare Regression	107
6	Tabellen	111
7	Anhang	115
7.1	Normalapproximation des Binomialtests: Zwischenrechnung . . .	115

Vorbemerkungen

Die Vorlesung behandelt zuerst die Wahrscheinlichkeitsrechnung und Statistik für diskrete Variablen, welche Werte zum Beispiel in $\{0, 1\}$, in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ oder in $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$ annehmen.

Danach werden die erarbeiteten Konzepte auf stetige Variablen übertragen, mit Wertebereichen zum Beispiel in \mathbb{R} oder $[0, 1]$. Deshalb ist der Aufbau leicht repetitiv, was sich aber in vorigen Jahren gut bewährt hat.

Schlussendlich wird auf komplexere Modellierung anhand der Regressions-Analyse eingegangen.

Für weitere Erläuterungen verweisen wir jeweils auf das folgende Buch:

Werner A. Stahel, Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler. 4. Aufl. (2002), Vieweg, Braunschweig.

Manchmal erklären wir die grundlegenden Begriffe auch an Hand von Glücksspielen, obwohl wir wissen, dass Sie nicht speziell an solchen interessiert sind. Es gibt dort einfach weniger Verständnis- und Interpretationsprobleme als bei anderen Beispielen. Wir hoffen auf Ihr Verständnis.

Prüfung

Achtung: Die rechtlich verbindlichen und stets aktuellen Informationen finden Sie im Vorlesungsverzeichnis! Unter Umständen sind folgende Informationen veraltet.

Es wird eine schriftliche Prüfung von 120 Minuten Dauer geben, bei der eine Zusammenfassung und der Taschenrechner erlaubt sind. Die Prüfung wird vier Bereiche abdecken und aus 80 Multiple-Choice Fragen (richtig/falsch) bestehen:

- Binomialverteilung und -test (Kapitel 2 und 3)
- t-Test (Kapitel 4.6 und 4.7)
- Lineare Regression (Kapitel 5)
- Gemischte Fragen (Rest des Skripts)

Die beste Prüfungsvorbereitung besteht aus folgenden Punkten:

- Besuchen Sie die Vorlesung regelmässig und arbeiten Sie aktiv mit (z.B. eduApp-Fragen)
- Lösen Sie die Quizzes termingerecht; in den Semesterferien ist es zu spät alles nachzuarbeiten; falls etwas nicht klar ist, scheuen Sie sich nicht, Fragen zu stellen!

Wer diese Punkte befolgt, sollte für die Prüfung gut gerüstet sein. Bei Fragen können Sie sich gerne an mich oder die Übungsassistenten wenden.

Kapitel 1

Einführung (Stahel, Kap. 1)

1.1 Was ist Statistik?

Statistik ist die Lehre vom Umgang mit quantitativen Daten. Oft verschleiern Zufall oder Unsicherheiten einen Sachverhalt. Ein Ziel der Statistik ist, in diesen Fällen klare Aussagen zu machen.

Betrachten wir ein Beispiel. Vor der Fussball Weltmeisterschaft 2010 wurden Panini-Sammelalben für 661 verschiedene Aufkleber von Fussballstars und -stadien verkauft. Jedes Bild hat eine Nummer von 1 bis 661 auf der Rückseite. Die Aufkleber gab es entweder in einzelnen 5er Packs oder in ganzen Kisten mit hundert 5er Packs zu kaufen. Gemäss Panini sind alle Bilder gleich häufig und werden wohl zufällig auf die Tüten verteilt.

Vor dieser WM hatten Kollegen von mir den Eindruck, dass man überraschend wenige doppelte Bilder hat, wenn man nicht einzelne 5er Packs sondern eine ganze Kiste mit hundert 5er Packs kauft. Sind die Bilder in einer Kiste vielleicht nicht zufällig verteilt worden? Bei der Fussball Weltmeisterschaft 2010 wollte ich diese Vermutung prüfen. Ich habe eine Kiste mit hundert 5er Packs gekaufte (500 Bilder) und alle in ein leeres Album geklebt. Ich konnte 477 Bilder einkleben, der Rest waren doppelte Bilder. Sind das nun "überraschend wenige doppelte" oder nicht? Ist es also billiger eine Kiste statt einzelne 5er Packs zu kaufen? Mit der Intuition ist diese Frage praktisch nicht zu beantworten, aber die Statistik liefert eine Lösung.

Angenommen, die Bilder werden tatsächlich zufällig auf die Tüten verteilt. Ist es plausibel, dass man unter 500 Bildern 477 oder mehr Bilder hat, die nur einmal vorkommen? Anders gefragt: Wie wahrscheinlich ist es, dass man mit zufälligem Eintüten 477 oder noch mehr aus 500 Bildern in ein leeres Album einkleben kann? Wenn diese Wahrscheinlichkeit sehr klein ist (kleiner als eine von uns festgelegte Grenze, z.B. eins zu einer Million = 10^{-6}), dann sind die Bilder wohl nicht zufällig eingetütet worden.

Nun, probieren wir es einfach aus! Mit dem Computer habe ich 500 Zahlen zufällig aus der Menge 1,...,661 mit Zurücklegen gezogen (jede Zahl steht für

ein Panini-Bild; jedes Bild kann mehrfach gezogen werden). Dabei hatte ich 350 Zahlen, die nur einmal vorkommen. In diesem zufällig gefüllten fiktiven Album hätte ich also nur 350 einfache Bilder einkleben können. Das ist weniger als die beobachteten 477 Bilder, aber vielleicht könnte ich ja bei dem nächsten zufällig gefüllten Album viel mehr Bilder einkleben.

Um herauszufinden, ob 477 einfache Bilder durch Zufall entstehen können, muss ich also noch mehr fiktive Alben füllen. Mit dem Computer habe ich eine Million fiktive Alben zufällig gefüllt. Die grösste Anzahl einfacher Bilder war dabei 387. Falls die Bilder wirklich zufällig eingetütet werden, ist die Wahrscheinlichkeit mit einer Kiste mehr als 387 eingeklebte Bilder zu haben also kleiner als 10^{-6} .

Das lässt folgenden Schluss zu: Wenn ich ein Album mit 500 zufällig ausgewählten Bildern füllen würde, könnte ich nicht mal in einer Million Alben so viele Bilder einkleben, wie ich in Realität beobachtet habe. Anders gesagt: In einer Welt, in der die Panini-Bilder zufällig verteilt werden, ist die Wahrscheinlichkeit 477 oder mehr einfache Bilder in einer Kiste zu finden also kleiner als eine Million. Die Hypothese "Die 500 Bilder werden zufällig verteilt" und die Beobachtung "Ich konnte 477 Bilder einkleben" passen also nicht zusammen. Die Hypothese wird deshalb verworfen: Die Panini-Bilder werden nicht zufällig eingetütet und es scheint vorteilhaft zu sein eine ganze Kiste zu kaufen.

In der Statistik ist das eben geschilderte Vorgehen fundamental und wird **Hypothesentest** genannt. Ein Hypothesentest besteht aus sechs Schritten:

1. Man stellt ein **Modell** auf, das erklärt, wie die Daten entstehen. (Wir ziehen 500 Bilder mit Zurücklegen aus einer Menge von 661 Bildern.)
2. Man stellt eine **Hypothese** (häufig auch **Nullhypothese**¹ genannt) auf. ("Panini Bilder in einer Kiste werden *zufällig* aus 661 Bildern mit Zurücklegen gezogen und eingetütet".) Wenn die Nullhypothese nicht zutrifft, muss die sog. **Alternative** zutreffen. ("Panini Bilder werden *nicht zufällig*, sondern nach irgendeinem Muster gezogen und eingetütet".)
3. Man überlegt sich eine **Teststatistik** (Anzahl Bilder, die man in ein leeres Album einkleben kann, wenn man 500 Bilder hat) und die **Wahrscheinlichkeitsverteilung der Teststatistik** unter der Annahme, dass die Nullhypothese stimmt. (Welche Anzahlen von einfachen Bildern sind plausibel, wenn zufällig 500 Bilder aus 661 mit Zurücklegen gezogen werden? Ich habe die Frage mit dem Computer beantwortet; oft kann man analytische Resultate finden.)
4. Man legt ein **Signifikanzniveau** fest. (Festgelegte Grenze von 10^{-6})
5. Man bestimmt einen **Verwerfungsbereich** für die Teststatistik zum gegebenen Signifikanzniveau. (Angenommen, die Bilder werden zufällig ein-

¹Die Namensgebung ist eher Konvention und macht für Sie im Augenblick wohl wenig Sinn. Oft wird ein Medikament getestet. Die Hypothese ist, dass das Medikament gar nicht wirkt. Der Effekt des Medikaments ist also *Null*. Daher wird die Hypothese dann "Nullhypothese" genannt. Dieser Sprachgebrauch hat sich auch in anderen Bereichen durchgesetzt.

getütet. In der Computersimulation haben wir gesehen, dass die Wahrscheinlichkeit mehr als 387 eingeklebte Bilder zu haben kleiner 10^{-6} ist. Der Verwerfungsbereich ist also $\{388, 389, \dots, 500\}$.²

6. Man **beobachtet einen Wert der Teststatistik** (477 eingeklebte Bilder) und fällt den **Testentscheid**: Verwerfe die Nullhypothese, falls der beobachtete Wert in dem Verwerfungsbereich liegt. (477 ist in der Menge $\{388, 389, \dots, 500\}$. Daher wird die Nullhypothese verworfen.)

Manchmal wird eine kleine Variante des obigen Verfahrens verwendet. Man berechnet den **P-Wert**: Angenommen die Nullhypothese stimmt. Wie wahrscheinlich ist es eine Teststatistik zu beobachten, die mindestens so extrem ist wie die beobachtete Teststatistik? Die Antwort liefert der P-Wert. (Wie wahrscheinlich ist es, in meiner Computersimulation 477 oder mehr einfache Bilder zu erhalten? Ich habe es in einer Million Simulationen kein einziges Mal beobachtet. Also ist die Wahrscheinlichkeit sicher³ kleiner als $\frac{1}{1000000} = 10^{-6}$. Also ist $p < 10^{-6}$.) Der P-Wert muss nicht immer so gross sein wie das Signifikanzniveau. Ich habe mein Beispiel so gewählt, dass beide Zahlen gleich sind. Üblicherweise wählt man als Signifikanzniveau $\alpha = 0.05$.

Wenn man eine Nullhypothese nach obiger Methode verwirft, kann man sich natürlich irren. Es könnte ja sein, dass die Bilder zufällig eingetütet werden und ich hatte einfach unglaubliches Glück und konnte 477 Bilder einkleben. Dann würde ich die Nullhypothese fälschlicherweise verwerfen. Dieses Ereignis tritt aber mit einer Wahrscheinlichkeit ein, die höchstens so gross ist wie das Signifikanzniveau bzw. der P-Wert. D.h., die Wahrscheinlichkeit, die Nullhypothese fälschlicherweise zu verwerfen, ist höchstens so gross wie das Signifikanzniveau bzw. der P-Wert. Man nennt diesen Fehler auch **Fehler 1. Art** ⁴.

Versuchen Sie, dieses Beispiel zu verstehen. Es ist die Grundlage zu einem grossen Teil des Stoffs in diesem Kurs. Wir werden diese Art von Test noch etliche Male antreffen. Wenn Sie bei einem neuen Test verwirrt sind, blättern Sie zu diesem Beispiel zurück und versuchen Sie Gemeinsamkeiten zu finden. Dann klärt sich das Problem vielleicht. Falls nicht, fragen Sie!

1.2 Kann ich Statistik überhaupt brauchen?

An der ETH haben wir einen Statistischen Beratungsdienst ⁵. Hier kommen wir jede Woche mit vielen praktischen Statistikproblemen aus Forschung und Wirtschaft in Kontakt.

²Ich habe mich hier für einen einseitigen Test entschieden; es gibt auch zweiseitige Tests. Später erfahren Sie mehr darüber.

³Für diejenigen, die es ganz genau wissen wollen: Da wir die Wahrscheinlichkeit mit einer Computersimulation ermittelt haben, stimmt das Ergebnis nur approximativ. In unserem Fall ist der Unterschied aber so deutlich, dass das keine praktische Rolle spielt.

⁴Es gibt auch noch den **Fehler 2. Art**. Er tritt dann auf, wenn man die Nullhypothese beibehält, obwohl sie eigentlich falsch ist. Dazu später mehr.

⁵<https://www.math.ethz.ch/sfs/consulting.html>

Gerade aus den Bereichen Biologie, Pharmazie und Bewegungswissenschaften treffen sehr viele Anfragen bei uns ein. Das kann von kleinen Problemen in der Bachelor-Arbeit über Schwierigkeiten in der Doktorarbeit bis hin zu Fragestellungen in einer grossen Firma gehen. Hier sind ein paar Beispiele:

BWS Mitarbeiter der BASPO Magglingen: Die (totale) Hämoglobinmasse ist entscheidend für die Leistungsfähigkeit im Ausdauersport (stark korreliert mit V_{O2max}). Es ist fraglich, ob sie während der Wachstumsphase durch Training gesteigert werden kann. Man hat Spitzenathleten und Nichtsportler in drei Altersgruppen (16, 21-23, 23-30) mit je 16 Personen untersucht. Über ein Jahr haben alle Personen Trainingstagebuch geführt (Gesamtzeit des Ausdauertrainings). Zu drei Zeitpunkten im Jahr (Anfang Jahr, Mitte und Ende) wurde die HbMasse bestimmt. Ist die Entwicklung der HbMasse bei 16 jährigen Spitzenathleten signifikant anders als bei den übrigen Gruppen?

BWS Masterarbeit BWS: Eine neue Methode zum Dehnen von Muskeln wurde untersucht. Probanden, die mit der neuen Methode trainiert haben, haben leicht bessere Ergebnisse gezeigt als diejenigen, die mit der alten Methode trainiert haben. Könnte dieser Unterschied Zufall sein oder ist die neue Methode mit grosser Wahrscheinlichkeit wirklich besser?

BWS Bachelorarbeit BWS: Man beobachtet, dass eine gewisse Phase des Skischwungs bei Elite-Skifahrern deutlich kürzer ist als bei anderen Skifahrern. Kann es sein, dass Fahrkönnen und Länge der speziellen Skischwungphase nichts miteinander zu tun haben und die Personen mit dem kürzeren Schwung per Zufall alle in einer Gruppe (Elite-Gruppe) gelandet sind?

Pharm Doktorand in Pharmazie: Eine bahnbrechende Methode wurde an der ETH entwickelt um festzustellen, welche Moleküle aus einer Bibliothek von ca. 1 Mio Molekülen mit einem Zielmolekül reagieren. Oft reagieren alle Moleküle mit dem Zielmolekül ein bisschen, aber manche besonders stark. Aussichtsreiche Moleküle werden in einem zweiten, viel aufwändigeren Experiment bestätigt. Das zweite Experiment möchte man nur dann machen, wenn man sich bzgl. einem Kandidaten ziemlich sicher ist. Wie stark muss ein Molekül reagieren, damit es ziemlich sicher kein Zufall war und es sich um einen echten Kandidaten handelt?

Pharm Häufige Frage bei Tier und Mensch: Eine zufällig ausgewählte Gruppe von Patienten hat ein neues Medikament in Form einer Tablette erhalten, eine andere zufällig ausgewählte Gruppe von Patienten ein Placebo⁶. Die Medikamentengruppe hat nach zwei Wochen eine deutliche Verbesserung der Symptome gezeigt. Kann es sein, dass das Medikament gar nicht wirkt und die Personen mit der Verbesserung zufällig alle in der Medikamentengruppe sind?

⁶Tablette mit gleichem Aussehen und Geschmack wie das Medikament aber ohne Wirkstoff.

Biol Doktorand in Biologie: In einem Experiment hat man beobachtet, dass gewisse Formen einer Pflanze den Polleneinfang fördern oder hemmen können. Könnte dieses Ergebnis auch durch Zufall entstehen, wenn die Form der Pflanze gar keinen Einfluss auf den Polleneinfang hat?

Biol Häufige Frage: Man beobachtet, dass in einer Gruppe von Krebspatienten gewisse Gene stärker aktiv sind als in einer Kontrollgruppe. Könnte es Zufall sein, dass alle Personen, bei denen diese Gene aktiver sind in derselben Gruppe gelandet sind?

Biol Genome Wide Association Studies (GWAS): Wie kann man die genetische Information (z.B. SNPs) mit den Merkmalen eines Lebewesens in Verbindung bringen?

Statistik ist in den Life-Sciences einfach nicht mehr wegzudenken. Egal ob Sie selber Auswertungen machen wollen oder ein Paper in einem Journal verstehen wollen: Sie müssen wissen, was ein P-Wert ist, wie ein statistischer Test funktioniert und was die grundlegenden Techniken in der Statistik sind.

Diese Vorlesungen soll Ihnen helfen, ein Fundament zu legen und die Grundbegriffe in der Statistik verstehen und anwenden zu können. Wenn Sie Interesse haben, gibt es eine Vielzahl von Vorlesungen, mit denen Sie Ihr Wissen vertiefen können.

1.3 Was ist der Inhalt dieses Kurses?

In Kapitel 2 geht es zunächst darum, den Begriff "Zufall" mathematisch genau zu definieren. Wenn das geschehen ist, kann man mit Wahrscheinlichkeiten einfach rechnen. Wir werden zunächst nur Zählraten behandeln, also Situationen, in denen nur ganze Zahlen auftreten (z.B. die Anzahl Gewinne, wenn man 20 Lose kauft). Wir werden die Begriffe "Unabhängigkeit", "bedingte Wahrscheinlichkeit", "Zufallsvariable", "Erwartungswert" und "Standardabweichung" kennenlernen. Ausserdem werden wir verschiedene Arten von Zufall gruppieren und dabei den Begriff der Verteilung kennenlernen.

In Kapitel 3 lernen wir die drei Grundfragen der Statistik - Punktschätzung, Hypothesentest, Vertrauensintervall - anhand von Zählraten kennen. Damit können wir folgende generische Frage beantworten: Angenommen wir haben ein Medikament an 100 Testpersonen ausprobiert. 67 Personen wurden gesund. Bei welchem Anteil der Gesamtbevölkerung wirkt das Medikament? In welchem Bereich liegt wohl der wahre Anteil mit grosser Wahrscheinlichkeit (z.B. mit 95%-Wahrscheinlichkeit zwischen 60% und 70%)? Der **Binomialtest** wird uns hier eine Antwort liefern.

In Kapitel 4 erweitern wir die bisherige Methodik auf kontinuierliche, reelle Daten (z.B. Grösse, Gewicht, etc.). Angenommen, wir haben neue Augentropfen entwickelt, die den Augeninnendruck senken sollen. Wir wählen zufällig 20 Testpersonen und teilen sie zufällig in zwei Gruppen mit je 10 Personen auf.

Gruppe N wird mit den neuen Augentropfen behandelt, die andere Gruppe A mit herkömmlichen Augentropfen. In Gruppe N scheint der Augeninnendruck stärker zu sinken als in der Gruppe A. Die zentrale Frage, die wir in diesem Kapitel beantworten, ist folgende: Könnte es sein, dass beide Medikamente gleich gut wirken, aber die Personen, die besonders gut auf Augentropfen ansprechen, zufällig in Gruppe N zusammengefasst wurden? Der **t-Test** wird uns hier eine Antwort liefern.

In Kapitel 5 werden wir Zusammenhänge zwischen kontinuierlichen Daten untersuchen. Angenommen, wir wollen die Ausdauerfähigkeit (z.B. VO_2max) durch andere Faktoren wie Alter, Trainingsaufwand pro Woche und Zeit für einen 3-km Lauf bestimmen. Wie modelliert man den Zusammenhang zwischen einer kontinuierlichen Grösse und mehreren erklärenden Variablen? Die **lineare Regression** wird uns hier Antworten liefern.

1.4 Software

In der Vorlesung *Statistik 1* werden wir uns auf Konzepte konzentrieren und den Computer kaum verwenden. In der Praxis gibt es allerdings sehr mächtige Statistikprogramme, mit denen praktisch alle Methoden aus diesem Skript in nur einer Zeile ausgeführt werden können. Um Statistik in der Praxis anwenden zu können, müssen Sie sich mit Software auskennen. Ich empfehle dazu unser Online-Lernprogramm "etutoR"⁷. Es ist genau auf unsere Vorlesung zugeschnitten und zeigt Ihnen, wie Sie die Methoden aus der Vorlesung in der Praxis einfach anwenden können.

Für die Prüfung ist die Statistiksoftware "R" allerdings *nur* für die Interpretation von Regressions-Output relevant.

1.5 Literatur

Ein gutes Begleitbuch in deutscher Sprache ist "Statistische Datenanalyse" von W. Stahel im Vieweg Verlag. Ein ausgezeichnetes Buch in englischer Sprache ist "Mathematical Statistics and Data Analysis" von John A. Rice. Inzwischen sind auch die Einträge in Wikipedia brauchbar und vor allem meist gut verständlich. Wenn Sie Mühe haben eine Methode zu verstehen, schlage ich Ihnen vor, den Begriff in Wikipedia nachzulesen.

⁷<https://moodle-app2.let.ethz.ch/course/index.php?categoryid=34>

Kapitel 2

Modelle für Zähldaten

Lernziele

- Sie kennen die drei Axiome der Wahrscheinlichkeitsrechnung.
- Sie kennen den Begriff der Unabhängigkeit und können einfache Rechenaufgaben lösen.
- Sie kennen den Begriff der bedingten Wahrscheinlichkeit und können damit umgehen. Insbesondere kennen Sie den Satz von Bayes und den Satz der totalen Wahrscheinlichkeit. Sie wissen, wie man $P(A|B)$ und $P(B|A)$ mit Formeln in einen Zusammenhang bringen kann und können einfache Aufgaben damit lösen.
- Sie kennen die Begriffe odds, log-odds und odds-Ratio. Sie können sie interpretieren und einfache Rechenaufgaben mit ihnen lösen.
- Sie kennen den Begriff der Zufallsvariable, der Wahrscheinlichkeitsverteilung und kumulativen Verteilungsfunktion.
- Sie kennen die Binomial- und die Poissonverteilung und den Poissonprozess.
- Sie kennen die Begriffe Erwartungswert, Varianz und Standardabweichung. Sie können diese Werte für einfache Verteilungen berechnen.

2.1 Wahrscheinlichkeitsmodelle (Stahel, Kap. 4.1, 4.2)

Wir betrachten **Zufallsexperimente**, bei denen der Ausgang nicht exakt vorhersagbar ist. Ein **Wahrscheinlichkeitsmodell** beschreibt, welche Ereignisse in einem solchen Experiment möglich sind und welche Chancen die verschiedenen Ergebnisse haben. Ein Wahrscheinlichkeitsmodell erlaubt mittels Simulation mögliche Ergebnisse zu erzeugen und so eine Vorstellung der zufälligen Variabilität zu gewinnen.

Ein Wahrscheinlichkeitsmodell hat die folgenden Komponenten:

- **Grundraum** Ω , bestehend aus den **Elementarereignissen** ω ,
- **Ereignisse** A, B, C, \dots ,
- **Wahrscheinlichkeiten** P .

Elementarereignisse sind mögliche Ergebnisse oder Ausgänge des Experiments, die zusammen den Grundraum bilden:

$$\Omega = \underbrace{\{\text{mögliche Elementarereignisse } \omega\}}_{\text{mögliche Ausgänge/Resultate}}$$

Bei der Durchführung des Experiments wird ein Elementarereignis zufällig gewählt.

Beispiel: 2-maliges Werfen einer Münze

$\Omega = \{KK, KZ, ZK, ZZ\}$ wobei $K = \text{“Kopf”}$ und $Z = \text{“Zahl”}$ bezeichnet. Ein Elementarereignis ist zum Beispiel $\omega = KZ$.

Unter einem **Ereignis** A versteht man eine Teilmenge von Ω :

$$\text{Ereignis } A \subset \Omega$$

„Ein Ereignis A tritt ein“ bedeutet, dass das Ergebnis ω des Experiments zu A gehört.

Beispiel (Forts.): $A = \{\text{genau 1-mal Kopf}\} = \{KZ, ZK\}$.

Die Operationen der Mengenlehre (Komplement, Vereinigung, Durchschnitt) werden für Ereignisse verwendet:

$$\begin{aligned} A \cup B &\Leftrightarrow A \text{ **oder** } B, \text{ wobei das “oder” nicht-exklusiv ist (”oder/und”)} \\ A \cap B &\Leftrightarrow A \text{ **und** } B \\ A^c &\Leftrightarrow \text{**nicht** } A \end{aligned}$$

Beispiel: $A = \text{morgen scheint die Sonne}$, $B = \text{morgen regnet es}$.

$A \cup B$ bedeutet: morgen scheint die Sonne oder morgen regnet es (und dies kann auch bedeuten, dass morgen die Sonne scheint und dass es morgen regnet); $A \cap B$ bedeutet: morgen scheint die Sonne und morgen regnet es; A^c bedeutet: morgen scheint die Sonne nicht.

Jedem Ereignis A wird schliesslich eine **Wahrscheinlichkeit** $P(A)$ zugeordnet. Dabei müssen die folgenden drei grundlegenden Regeln (**Axiome von Wahrscheinlichkeitsrechnung**) erfüllt sein:

1. Die Wahrscheinlichkeiten sind immer nicht-negativ: $P(A) \geq 0$
2. Das Ereignis Ω hat Wahrscheinlichkeit eins: $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ falls $A \cap B = \emptyset$, d.h. für alle Ereignisse, die sich gegenseitig ausschliessen.

Beispiel (Forts.) Beim Wurf zweier Münzen ist es plausibel, dass alle 4 Elemente von Ω gleich wahrscheinlich sind. Wegen $P(\Omega) = 1$ müssen sich die Wahrscheinlichkeiten zu Eins addieren:

$$P(KK) = P(KZ) = P(ZK) = P(ZZ) = \frac{1}{4}.$$

Weitere Regeln können daraus abgeleitet werden. *Beispiel:*

$$P(A^c) = 1 - P(A),$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{Additionssatz}).$$

Im Wesentlichen werden in der Wahrscheinlichkeitstheorie die Wahrscheinlichkeiten gewisser Ereignisse A festgelegt (auf Grund von Plausibilitäten, Symmetriüberlegungen, wissenschaftlichen Theorien, Fachwissen und Daten) und daraus die Wahrscheinlichkeiten von gewissen anderen Ereignissen B aus den obigen Gesetzen hergeleitet.

Die Statistik geht umgekehrt vor: aus Daten, d.h. aus der Information, dass gewisse Ereignisse eingetreten sind, versucht man Rückschlüsse auf ein unbekanntes Wahrscheinlichkeitsmodell (unbekannte Wahrscheinlichkeiten) zu machen.

Interpretation von Wahrscheinlichkeiten:

- **Frequentistisch:** Idealisierung der relativen Häufigkeiten bei vielen unabhängigen Wiederholungen
- **Bayes'sch:** Mass für den Glauben, dass ein Ereignis eintreten wird

2.1.1 Wahrscheinlichkeiten berechnen

Summe der Elementarereignisse

Wir behandeln in diesem Kapitel diskrete Wahrscheinlichkeitsmodelle, bei denen der Grundraum endlich oder "abzählbar" ist (d.h. man kann die Elementarereignisse durchnummerieren). Zum Beispiel ist $\Omega = \{0, 1, \dots, 10\}$ endlich und deshalb diskret; $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$ ist zwar unendlich, aber noch abzählbar und daher trotzdem diskret; $\Omega = \mathbb{R}$ ist nicht abzählbar.

Im diskreten Fall ist die Wahrscheinlichkeit eines Ereignisses durch die Wahrscheinlichkeiten der zugehörigen Elementarereignisse $P(\{\omega\})$ festgelegt:

$$P(A) = \sum_{\omega \in A} P(\{\omega\}).$$

Beispiel (Forts.) Für $A = \text{genau einmal Kopf} = \{KZ, ZK\}$ hat man also $P(A) = P(KZ) + P(ZK) = 1/4 + 1/4 = 1/2$.

Laplace-Modell: Günstig durch Möglich

In vielen Fällen ist es plausibel anzunehmen, dass jedes Elementarereignis die

gleiche Wahrscheinlichkeit hat. In diesen Fällen gibt es eine besonders einfache Möglichkeit, die Wahrscheinlichkeit eines Ereignisses E , dass aus verschiedenen Elementarereignissen besteht ($E = \{\omega_1, \omega_2, \dots, \omega_g\}$; Grundraum $\Omega = \{\omega_1, \omega_2, \dots, \omega_m\}$) zu berechnen.

$$P(E) = \frac{g}{m}$$

Man teilt die Anzahl für das Ereignis "günstigen" Elementarereignisse durch die Anzahl der "möglichen" Elementarereignisse.

Beispiel: Es werden zwei Würfel geworfen. Was ist die Wahrscheinlichkeit, dass die Augensumme 7 ergibt? Ein Elementarereignis beschreibt die Augenzahlen auf beiden Würfeln, also z.B. (1, 4), wenn der eine Würfel eine 1 und der andere eine 4 zeigt. Es sind insgesamt 36 Elementarereignisse möglich: $\{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}$. Es gibt davon 6 Elementarereignisse, bei denen die Augensumme 7 ist: $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$. Da alle Elementarereignisse gleich wahrscheinlich sind, ist die Wahrscheinlichkeit für das Ereignis E : $P(E) = \frac{6}{36} = \frac{1}{6}$.

Mengenoperationen / Venn-Diagramm

Wie oben gesehen (z.B. Additionssatz) gibt es auch allgemeine Regeln aus der Mengenlehre, die Umformungen von Wahrscheinlichkeiten ermöglichen. Dazu ist oft die Verwendung von einem Venn-Diagramm hilfreich.

Beispiel: Gegenereignis A^c :

$$P(A^c) = 1 - P(A)$$

2.2 Der Begriff der Unabhängigkeit (Stahel, Kap. 4.6)

Wenn man die Wahrscheinlichkeiten $P(A)$ und $P(B)$ kennt, lässt sich daraus $P(A \cap B)$ im Allgemeinen nicht berechnen: Es sind alle Werte zwischen 0 und dem Minimum von $P(A)$ und $P(B)$ möglich. Ein wichtiger Spezialfall liegt vor, wenn folgende Produktformel gilt

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

Dann heissen A und B **stochastisch unabhängig**.

Beispiel (Forts.): Es sei $A = K$ im 1. Wurf und $B = K$ im 2. Wurf. Dann gilt $P(A) = P(KK) + P(KZ) = \frac{1}{2}$ und analog $P(B) = \frac{1}{2}$. D.h., $P(A)P(B) = \frac{1}{4}$. Wegen $P(A \cap B) = P(KK) = \frac{1}{4}$, sind also A und B unabhängig.

Viel wichtiger ist jedoch der umgekehrte Schluss. Wenn zwischen den Ereignissen A und B kein kausaler Zusammenhang besteht (d.h. es gibt keine gemeinsamen Ursachen oder Ausschlüsse), dann **postuliert** man stochastische Unabhängigkeit und nimmt damit an, dass obige Produktformel gilt. In diesem Fall kann also $P(A \cap B)$ aus $P(A)$ und $P(B)$ berechnet werden.

Beispiel (Forts.): Es ist plausibel, dass es keinen kausalen Zusammenhang zwischen dem Ergebnis des ersten und des zweiten Wurfs gibt. Die Ereignisse A

und B sind also unabhängig. Deshalb kann man $P(A \cap B)$ wie folgt ausrechnen: $P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

Beispiel: Angenommen, wir werfen zwei sechseckige Würfel. Sei E das Ereignis, dass die Augensumme 6 ist. Sei F das Ereignis, dass der erste Würfel die Augenzahl 4 zeigt. Dann gilt

$$P(E \cap F) = P(\{4, 2\}) = \frac{1}{36}$$

aber

$$P(E) \cdot P(F) = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{216}$$

Deshalb sind E und F nicht unabhängig. Das Ergebnis kann man intuitiv einfach verstehen. Wenn wir daran interessiert sind mit beiden Würfeln die Augensumme 6 zu würfeln, sollten wir recht froh sein, wenn der erste Wurf eine 4 zeigt (oder irgendeine Zahl aus 1,2,3,4,5), denn dann haben wir noch eine Chance nach dem zweiten Wurf auf Augensumme 6 zu kommen. Falls aber der erste Wurf schon eine 6 zeigt, gibt es keine Chance mehr auf die Augensumme 6 zu kommen. D.h., die Chance die Augensumme 6 zu erzielen hängt von der Augenzahl des ersten Wurfs ab. Daher können E und F nicht unabhängig sein.

Bei mehr als zwei Ereignissen A_1, \dots, A_n bedeutet Unabhängigkeit, dass die entsprechende Produktformel für alle k -Tupel von Ereignissen gilt:

$$P(A_{i_1} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$$

für jedes $k = 2, 3, \dots, n$ und jedes $1 \leq i_1 < \dots < i_k \leq n$. Insbesondere muss also gelten

$$P(A_1 \cap A_2) = P(A_1)P(A_2), \quad P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3), \text{ etc.}$$

2.3 Bedingte Wahrscheinlichkeit (Stahel, Kap. 4.7)

Was ist die Wahrscheinlichkeit, dass ein sechseckiger Würfel die Zahl 3 zeigt, wenn wir schon wissen, dass eine ungerade Zahl geworfen wurde? Es gibt drei ungerade Zahlen auf dem Würfel (1,3,5). Die 3 ist eine davon. Wenn man davon ausgehen kann, dass jede Zahl auf dem Würfel gleich wahrscheinlich ist, ist das Ergebnis $\frac{1}{3}$. Fragen von dieser Art kann man auch in komplizierteren Sachverhalten mit dem Begriff der **bedingten Wahrscheinlichkeit** leicht beantworten.

Die bedingte Wahrscheinlichkeit von Ereignis A gegeben B , d.h., die Wahrscheinlichkeit von A , wenn wir wissen, dass B eingetreten ist, wird mit $P(A|B)$ bezeichnet und so definiert:

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

Beispiel: Wir rechnen nochmals das Beispiel aus dem Anfang dieses Unterkapitels. Sei $A = \text{“Die 3 wird geworfen”}$ und $B = \text{“Eine ungerade Zahl wird geworfen”}$. Für die bedingte Wahrscheinlichkeit $P(A|B)$ gilt also

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Im allgemeinen ist $P(A|B)$ **nicht** das gleiche wie $P(B|A)$. Dieser Fehler wird vor Gericht (und nicht nur da) so oft gemacht, dass er schon einen eigenen Namen hat: Prosecutor’s fallacy¹.

Beispiel: Ein medizinischer Test für eine Krankheit (D falls vorhanden, D^c falls nicht vorhanden) kann positiv (+) oder negativ (−) sein. Die Wahrscheinlichkeiten sind in folgender Tabelle gegeben. Z.B. ist die Wahrscheinlichkeit, dass die Krankheit vorhanden ist und der Test positiv ausfällt $P(D \cap +) = 0.009$.

	D	D^c
+	0.009	0.099
−	0.001	0.891

Daraus folgt

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

und

$$P(-|D^c) = \frac{P(- \cap D^c)}{P(D^c)} = \frac{0.891}{0.891 + 0.099} = 0.9.$$

Offensichtlich ist dieser Test recht genau. Kranke Personen werden zu 90% als positiv eingestuft, und gesunde Personen werden zu 90% als negativ eingestuft. Angenommen, Sie gehen zu einem Test und werden als positiv eingestuft. Wie gross ist die Wahrscheinlichkeit, dass Sie die Krankheit wirklich haben? Die meisten Leute würden 0.9 antworten. Die richtige Antwort ist aber

$$P(D|+) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.08.$$

Die Lektion ist: Bei bedingten Wahrscheinlichkeiten darf man seiner Intuition nicht trauen, sondern muss das Ergebnis ausrechnen.

Das **Bayes Theorem** liefert einen oft nützlichen Zusammenhang zwischen $P(A|B)$ und $P(B|A)$:

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

¹Suchen Sie in Wikipedia danach!

Beispiel (Forts.): Das Bayes Theorem liefert die gleiche Lösung wie unsere obige Rechnung:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.9 \cdot (0.009 + 0.001)}{0.009 + 0.099} = \frac{0.009}{0.009 + 0.099} = 0.08.$$

Eine zweite nützliche Formel ist das **Gesetz der totalen Wahrscheinlichkeit**: Betrachte Ereignisse A_1, \dots, A_k , die miteinander keine Schnittmenge haben und zusammen alle möglichen Fälle erfassen². Dann gilt für jedes beliebige Ereignis B :

$$P(B) = \sum_{i=1}^k P(B|A_i)P(A_i)$$

Beispiel (fiktiv): Angenommen, ich teile meine emails in drei Kategorien ein: A_1 = "spam", A_2 = "niedrige Priorität" und A_3 = "hohe Priorität". Von früheren Beobachtungen weiss ich, dass $P(A_1) = 0.7$, $P(A_2) = 0.2$ und $P(A_3) = 0.1$ ³. Sei B das Ereignis, dass das Wort "free" in der email auftaucht. Von früheren Beobachtungen weiss ich, dass $P(B|A_1) = 0.9$, $P(B|A_2) = 0.01$ und $P(B|A_3) = 0.01$ ⁴. Angenommen, ich erhalte eine email, die das Wort "free" enthält. Wie gross ist die Wahrscheinlichkeit, dass es sich um spam handelt? Das Bayes Theorem zusammen mit dem Gesetz der totalen Wahrscheinlichkeit liefern die Lösung:

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} = \\ &= \frac{0.9 \cdot 0.7}{(0.9 \cdot 0.7) + (0.01 \cdot 0.2) + (0.01 \cdot 0.1)} = 0.995 \end{aligned}$$

Viele Spamfilter basieren tatsächlich auf diesem Prinzip.

Um zu beschreiben, wie wahrscheinlich ein Ereignis E ist, kann man die Wahrscheinlichkeit $P(E)$ angeben. In den Life Sciences ist eine zweite Beschreibung der Wahrscheinlichkeit sehr verbreitet: Die **odds** des Ereignisses E , $odds(E)$. Die odds von einem Ereignis und die Wahrscheinlichkeit von einem Ereignis können immer ineinander umgerechnet werden. Sie drücken also ein- und denselben Sachverhalt auf zwei verschiedene Arten aus. Die $odds(E)$ geben an, wieviel mal wahrscheinlicher das Eintreten von E als das Eintreten von E^c ("nicht E ") ist.

$$odds(E) = \frac{P(E)}{1 - P(E)}$$

²Eine solche Aufteilung nennt man eine **Partitionierung**. Z.B. bei einem Münzwurf: A_1 = Kopf, A_2 = Zahl; bei einem Würfelwurf: A_1 = gerade, A_2 = ungerade; oder $A_1 = \{1\}$, $A_2 = \{2,4\}$, $A_3 = \{3,5,6\}$

³Beachte, dass $P(A_1) + P(A_2) + P(A_3) = 1$, wie es bei einer Partitionierung auch sein sollte.

⁴Beachte, dass hier die Summe nicht 1 ergibt.

Beispiel: Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person Krankheit K hat, ist $P(K) = 0.1$. Die odds, dass eine zufällig ausgewählte Person Krankheit K hat, sind also $odds(K) = 0.1/0.9 = \frac{1}{9}$. Es ist also neunmal so wahrscheinlich, dass eine Person Krankheit K nicht hat als dass sie K hat.

Beispiel: Angenommen, die odds eines Ereignisses sind $odds(E) = 3$. Dann ist $\frac{P(E)}{1-P(E)} = 3$. Auflösen nach $P(E)$ liefert: $P(E) = \frac{odds(E)}{1+odds(E)} = \frac{3}{4}$.

Die **log-odds** ist einfach der Logarithmus der odds. Dabei wird üblicherweise der natürliche Logarithmus verwendet (also zur Basis e und nicht zur Basis 10).

Beispiel: Angenommen, die odds eines Ereignisses sind $odds(E) = 3$. Dann gilt für die log-odds des Ereignisses: $\log\text{-odds}(E) = \log(odds(E)) = \log(3) \approx 1.099$.

Angenommen, wir haben zwei Gruppen $G = 1$ und $G = 2$. Die odds für das Ereignis E in Gruppe 1 sind

$$odds(E|G = 1) = \frac{P(E|G = 1)}{1 - P(E|G = 1)}.$$

Analog sind die odds für das Ereignis E in Gruppe 2

$$odds(E|G = 2) = \frac{P(E|G = 2)}{1 - P(E|G = 2)}.$$

Das Verhältnis dieser odds wird **odds-Ratio** genannt (OR).⁵

Beispiel: Personen können gegen eine Krankheit K geimpft werden. Gegeben man ist geimpft ($I = 1$), ist die Wahrscheinlichkeit Krankheit K zu bekommen nur 0.0001, d.h., $P(K|I = 1) = 0.0001$. Die odds für die Krankheit gegeben man ist geimpft sind also $odds(K|I = 1) = \frac{P(K|I=1)}{1-P(K|I=1)} = \frac{0.0001}{0.9999} \approx 0.0001$. Wenn man nicht geimpft ist ($I = 0$), ist die Wahrscheinlichkeit Krankheit K zu bekommen 0.2, d.h., $P(K|I = 0) = 0.2$. Die odds sind in diesem Fall also $odds(K|I = 0) = \frac{0.2}{0.8} = 0.25$. Das odds-Ratio OR ist dann:

$$OR = \frac{odds(K|I = 0)}{odds(K|I = 1)} = \frac{0.25}{0.0001} = 2500$$

(Ob das odds-Ratio als $\frac{odds(K|I=0)}{odds(K|I=1)}$ oder als $\frac{odds(K|I=1)}{odds(K|I=0)}$ definiert wird, muss angegeben werden; beides ist möglich, solange die Definition klar gemacht wird.). Die odds zu erkranken werden durch die Impfung um dem Faktor 2500 reduziert. Die Impfung ist also sehr wirksam.

2.4 Zufallsvariable (Stahel, Kap. 4.3, 4.4)

Bemerkung: In diesem Kapitel erklären wir das Konzept der Zufallsvariable auf eine exakte aber etwas kompliziertere Art. Falls Sie das verwirrt, merken Sie

⁵Falls Sie es später mal brauchen: Mit der Funktion `fisher.test` können Sie in **R** odds-Ratio und Vertrauensintervalle dazu berechnen. Das werden wir in dieser Vorlesung aber nicht behandeln.

sich als grobe Faustregel: Eine Zufallsvariable ist ein Zufallsexperiment, das als Ergebnis eine Zahl ⁶ hat. Nun versuchen wir aber, diese Idee noch etwas präziser zu definieren.

Oft sind mit einem Zufallsexperiment Zahlenwerte verknüpft, d.h. zu jedem Elementarereignis ω gehört ein Zahlenwert $X(\omega) = x$.

Beispiel: Wert einer gezogenen Jass-Karte.

$$\begin{aligned}\omega = \text{As} &\mapsto X(\omega) = 11 \\ \omega = \text{König} &\mapsto X(\omega) = 4 \\ &\vdots \\ \omega = \text{Sechs} &\mapsto X(\omega) = 0\end{aligned}$$

Wie man in diesem Beispiel sieht, ist X eine Funktion auf dem Grundraum Ω . Wir halten fest:

Eine **Zufallsvariable** X ist eine **Funktion**:

$$\begin{aligned}X : \quad \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega)\end{aligned}$$

Die Notation X (oder auch Y, Z, \dots) ist eher ungewohnt für die Bezeichnung einer Funktion, ist aber üblich in der Wahrscheinlichkeitsrechnung. Sie haben hoffentlich in der Analysis gesehen, dass man Funktionen wie Zahlen addieren oder multiplizieren kann (man addiert oder multipliziert einfach die Funktionswerte). Summen und Produkte von Zufallsvariablen sind also wieder Zufallsvariablen.

Konvention: Eine Zufallsvariable wird durch einen **Grossbuchstaben** (z.B. X) dargestellt. Der gleiche **Kleinbuchstabe** (z.B. x) stellt einen konkreten Wert dar, den die Zufallsvariable annehmen kann. Das Ereignis, dass die Zufallsvariable X den Wert x annimmt, können wir dann so schreiben: $X = x$.

Bei einer Zufallsvariable ist nicht die Funktion $X(\cdot)$ zufällig, sondern nur das Argument ω : Je nach Ausgang des Zufallsexperiments (d.h. von ω) erhalten wir einen anderen Wert $x = X(\omega)$, x ist **eine Realisierung** der Zufallsvariablen X . Wenn wir aber das Experiment zweimal durchführen und zwei Mal das gleiche Ergebnis ω herauskommt, dann sind auch die realisierten Werte von X gleich.

Wenn der Grundraum Ω diskret ist, dann muss auch der Wertebereich $W = W_X$ (Menge der möglichen Werte von X) diskret sein, d.h. endlich oder abzählbar. Wir werden in diesem Kapitel bloss diskrete Zufallsvariablen genauer diskutieren. Insbesondere sind Anzahlen stets diskret, während Messungen besser als kontinuierlich, d.h. mit dem Wertebereich \mathbb{R} modelliert werden (obwohl man praktisch nur mit endlicher Genauigkeit messen kann).

⁶Im Gegensatz zu einem Buchstaben (z.B. ein zufälliger Buchstabe aus dem Alphabet), einer Person (z.B. eine zufällige Person aus unserer Vorlesung) oder ähnlichem

Wahrscheinlichkeitsverteilung einer Zufallsvariablen

Die Werte einer Zufallsvariablen X (die möglichen Realisationen von X) treten mit gewissen Wahrscheinlichkeiten auf. Die Wahrscheinlichkeit, dass X den Wert x annimmt, berechnet sich wie folgt:

$$P(X = x) = P(\{\omega; X(\omega) = x\}) = \sum_{\omega; X(\omega)=x} P(\omega).$$

Beispiel (Forts): X = Wert einer gezogenen Jass-Karte.

$$\begin{aligned} & \text{Wahrscheinlichkeit für Zahl 4} = P(X = 4) \\ &= P(\{\omega; \omega = \text{ein König}\}) \\ &= P(\text{Eicheln-König}) + P(\text{Rosen-König}) + P(\text{Schellen-König}) + P(\text{Schilten-König}) \\ &= 4/36 = 1/9. \end{aligned}$$

Die "Liste" von $P(X = x)$ für alle möglichen Werte x heisst (diskrete) **(Wahrscheinlichkeits-) Verteilung** der (diskreten) Zufallsvariablen X . Dabei gilt immer

$$\sum_{\text{alle möglichen } x} P(X = x) = 1.$$

Beispiel (Forts): X = Wert einer gezogenen Jass-Karte.

Die Wahrscheinlichkeitsverteilung von X ist

x	0	2	3	4	10	11
$P(X = x)$	4/9	1/9	1/9	1/9	1/9	1/9

Umgekehrt ist jede Liste von nichtnegativen Zahlen, die sich zu eins addieren, die Verteilung einer gewissen Zufallsvariablen. Wenn man nur an der Zufallsvariablen X interessiert ist, kann man den zu Grunde liegenden Raum Ω vergessen, man braucht nur die Verteilung von X . Zufallsvariablen sind einfach Zufallsexperimente, bei denen die Ergebnisse Zahlen sind.

2.5 Binomialverteilung (Stahel Kap. 5.1)

Wir betrachten die Situation wo es um das Zählen der Anzahl Erfolge (oder Misserfolge) geht. Solche Anwendungen treten z.B. auf bei der Qualitätskontrolle, Erfolg/Misserfolg bei Behandlungen (medizinisch, biologisch) oder auch bei Glücksspielen.

Die Verteilung einer Zufallsvariable X mit Werten in $W = \{0, 1\}$ kann durch einen einzelnen Parameter π beschrieben werden:

$$P(X = 1) = \pi, \quad P(X = 0) = 1 - \pi, \quad 0 \leq \pi \leq 1.$$

Diese Verteilung heisst **Bernoulli(π)-Verteilung**. Sie beschreibt einfach das Eintreffen oder Nicht-Eintreffen eines bestimmten Ereignisses, z.B. das Ergebnis "Kopf" beim Werfen einer Münze. Falls die Münze fair ist, so ist $\pi = 1/2$.

Etwas interessanter wird es, wenn wir das Experiment n Mal wiederholen, also z.B. die Münze n -mal werfen. Der Grundraum Ω besteht dann aus allen "Wörtern" der Länge n , welche man mit den Buchstaben K (für "Kopf") und Z (für "Zahl") schreiben kann. Ω hat also 2^n Elemente. Wir betrachten die Zufallsvariablen

$$X_i = \begin{cases} 1 & \text{falls K im } i\text{-ten Wurf} \\ 0 & \text{falls Z im } i\text{-ten Wurf.} \end{cases}$$

$$X = \sum_{i=1}^n X_i = \text{Gesamtzahl von Würfeln mit K}$$

Um die Verteilung von X bestimmen zu können, müssen wir eine Wahrscheinlichkeit auf Ω festlegen. Wir postulieren, dass die Ereignisse $X_i = 1$ (also "K im i -ten Wurf") alle die Wahrscheinlichkeit π haben und unabhängig sind. Dann gilt zum Beispiel:

$$\begin{aligned} P(X = 0) &= P(X_1 = \dots = X_n = 0) = (1 - \pi)^n, \\ P(X = 1) &= P(\text{ein } X_i = 1 \text{ und alle anderen } X_j = 0) = \\ &= P(X_1 = 1, X_2 = 0, \dots, X_n = 0) + P(X_1 = 0, X_2 = 1, X_3 = 0, \dots, X_n = 0) + \dots \\ &= n\pi(1 - \pi)^{n-1}. \end{aligned}$$

Um $P(X = x)$ zu berechnen, muss man offenbar bestimmen, auf wieviele Arten man x Einer auf n Plätze anordnen kann. Die Antwort ist gegeben durch den Binomialkoeffizienten

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

So kommt man auf die Binomial-Verteilung.

Binomial(n, π)-Verteilung:

Eine Zufallsvariable X mit Werten in $W = \{0, 1, \dots, n\}$ heisst Binomial(n, π)-verteilt, falls

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, \dots, n.$$

Dabei ist $0 \leq \pi \leq 1$ der Erfolgsparameter der Verteilung.

Wie in obigem Beispiel motiviert, ist die Binomialverteilung angebracht für die Zufallsvariable "Anzahl Erfolge/Misserfolge" (Eintreten eines bestimmten Ereignis) bei n **unabhängigen** Versuchen. Das Prädikat "unabhängig" ist wesentlich für die Korrektheit der Binomialverteilung.

Konvention: Wenn man notieren will, dass die Zufallsvariable X einer gewissen Wahrscheinlichkeitsverteilung F folgt, schreibt man abgekürzt: $X \sim F$. Dabei kann F von Parametern abhängen, also z.B. $X \sim F(\theta)$. Wenn also X einer

Binomial-Verteilung mit Parametern n und π folgt, schreibt man abgekürzt $X \sim \text{Binomial}(n, \pi)$ oder einfach nur $X \sim \text{Bin}(n, \pi)$.

Beispiel Bei einer Losbude steht: "Jedes 5. Los gewinnt!", d.h., die Gewinnwahrscheinlichkeit ist bei jedem Los $\pi = 0.2$. Nehmen wir weiter an, dass das Ziehen von einem Los keinen Einfluss auf das Ziehen des nächsten Loses hat (z.B. gibt es eine riesige Anzahl Lose und die Lostrommel wird nach jedem Verkauf eines Loses gut gemischt). Wir kaufen 100 Lose und bezeichnen mit X die Anzahl Gewinne unter den 100 Losen. Dann ist X Binomial($n = 100, \pi = 0.2$) verteilt. Abgekürzt: $X \sim \text{Binomial}(100, 0.2)$.

Beispiel: Spermasexing (Tages-Anzeiger 6.12.2000)

Geschlechts-Beeinflussung von Kuhkälbern mit einer Methode, die Spermasexing genannt wird. Ziel ist es, ein weibliches Kalb zu züchten. In einem Testlauf wurden zwölf Kühe mit Spermien besamt, die optisch nach dem Y-Chromosom sortiert wurden (d.h. mit der Spermasexing-Methode). Da die Methode nicht hundertprozentig sicher ist, können wir das als Zufallsexperiment auffassen. Sei $X = \text{Anzahl weiblicher gezüchteter Kuhkälber}$. Eine vernünftiges Modell ist dann:

$$X \sim \text{Binomial}(12, \pi),$$

wobei π unbekannt ist. Effektiv beobachtet wurden $x = 11$ weiblich gezüchtete Kuhkälber: d.h. $X = x = 11$ wurde tatsächlich **realisiert**. Später mehr dazu.

Eigenschaften der Binomialverteilung (siehe Abb. 2.1): $P(X = x)$ ist maximal wenn x gleich dem ganzzahligen Teil von $(n+1)\pi$ ist, und auf beiden Seiten von diesem Wert nehmen die Wahrscheinlichkeiten monoton ab. Wenn $n\pi(1-\pi)$ nicht allzu klein ist, ist die Verteilung praktisch symmetrisch und hat die Form einer Glocke. Wenn n gross ist, sind die meisten Wahrscheinlichkeiten $P(X = x)$ verschwindend klein, d.h. grosse Abweichungen von $(n+1)\pi$ sind extrem unwahrscheinlich.

2.6 Kennzahlen einer Verteilung (Stahel Kap. 5.3)

Eine beliebige (diskrete) Verteilung kann vereinfachend zusammengefasst werden durch 2 Kennzahlen, den **Erwartungswert** $\mathcal{E}(X)$ und die **Standardabweichung** $\sigma(X)$.

Der Erwartungswert beschreibt die mittlere Lage der Verteilung und ist wie folgt definiert:

$$\mathcal{E}(X) = \sum_{x \in W_X} xP(X = x), \quad W_X = \text{Wertebereich von } X.$$

Die Standardabweichung beschreibt die Streuung der Verteilung. Rechnerisch

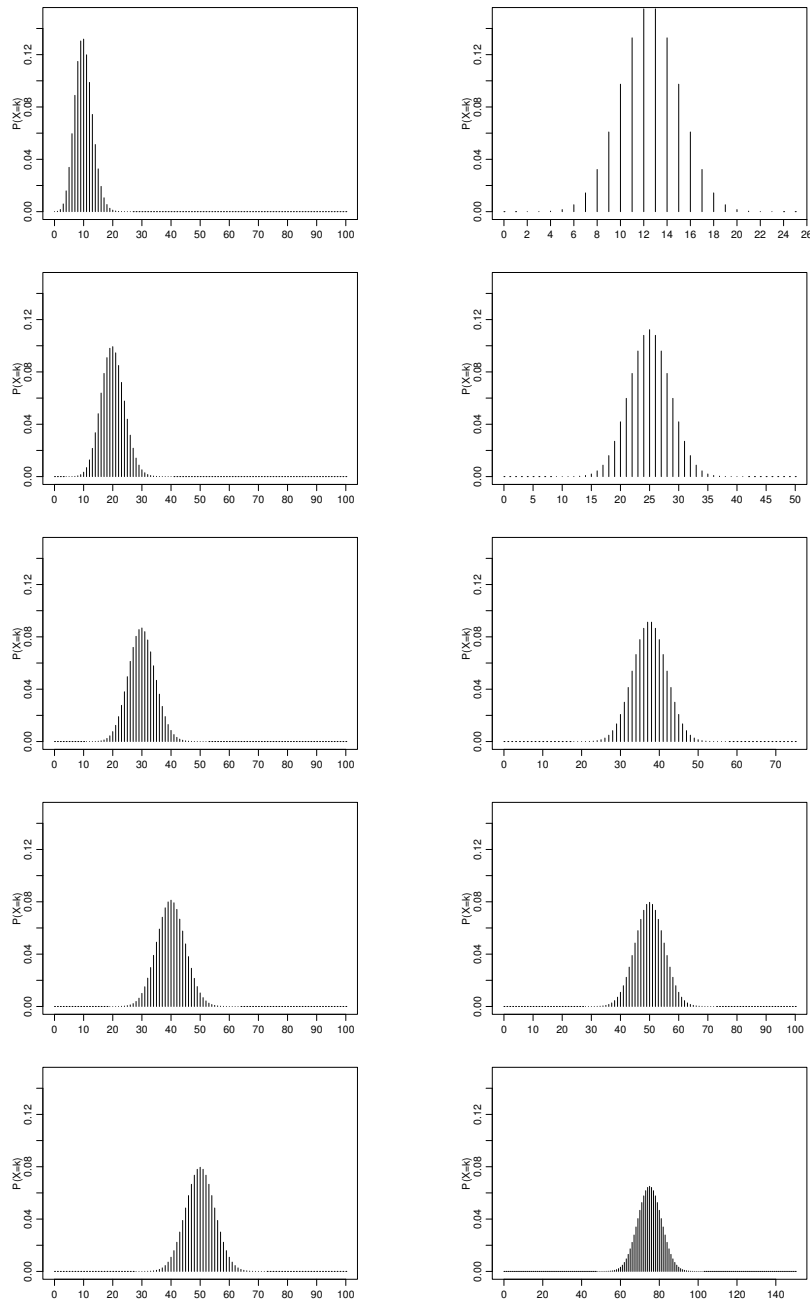


Abbildung 2.1: Die Binomialwahrscheinlichkeiten $P(X = x)$ als Funktion von x für verschiedene n 's und π 's. Links ist $n = 100$ und $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ (von oben nach unten) und rechts ist $\pi = 0.5$ und $n = 25, 50, 75, 100, 150$ (von oben nach unten).

ist das Quadrat der Standardabweichung, die sogenannte **Varianz** bequemer:

$$\text{Var}(X) = \sum_{x \in W_X} (x - \mathcal{E}(X))^2 P(X = x)$$

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

Die Standardabweichung hat dieselbe Einheit wie X , während die Einheit der Varianz deren Quadrat ist: Wird z.B. X in Metern (m) gemessen, so besitzt $\text{Var}(X)$ die Dimension Quadratmeter (m^2) und $\sigma(X)$ wiederum die Dimension Meter (m).

Beispiel: Sei $X \sim \text{Bernoulli}(\pi)$.

Dann:

$$\begin{aligned}\mathcal{E}(X) &= 0 \cdot P(X=0) + 1 \cdot P(X=1) = \pi, \\ \text{Var}(X) &= (0 - \mathcal{E}(X))^2 P(X=0) + (1 - \mathcal{E}(X))^2 P(X=1) = (0 - \pi)^2(1 - \pi) + (1 - \pi)^2 \pi \\ &= \pi(1 - \pi), \\ \sigma(X) &= \sqrt{\pi(1 - \pi)}.\end{aligned}$$

Für die Binomial-Verteilung erhält man mit einigen Rechnungen

$$X \sim \text{Binomial}(n, \pi) \Rightarrow \mathcal{E}(X) = n\pi, \quad \text{Var}(X) = n\pi(1 - \pi), \quad \sigma(X) = \sqrt{n\pi(1 - \pi)}.$$

(Weil $\text{Bernoulli}(\pi) = \text{Binomial}(1, \pi)$, stimmt das mit obigen Formeln überein). Die Kennzahlen fassen also sehr gut zusammen, was wir in der Abbildung 2.1 gesehen haben: Die Verteilung ist um den Erwartungswert konzentriert, die Streuung wächst mit n , aber langsamer als n . Für festes n ist die Streuung maximal, wenn $\pi = 1/2$.

Beispiel (Forts.) Wir sind wieder bei der Losbude, bei der wir (nach dem vierten Bier) 100 Lose gekauft hatten. Um die Freundin zu beeindrucken, kramen wir unser Statistikwissen hervor und berechnen im Kopf den Erwartungswert und die Standardabweichung der Anzahl Gewinne unter 100 Losen.

$$\begin{aligned}\mathcal{E}(X) &= n \cdot \pi = 100 \cdot 0.2 = 20 \\ \sigma(X) &= \sqrt{n\pi(1 - \pi)} = \sqrt{100 \cdot 0.2 \cdot 0.8} = 4\end{aligned}$$

Wir erinnern uns, dass Beobachtungen typischerweise ein bis zwei Standardabweichungen vom Erwartungswert entfernt liegen und prophezeien der Freundin mit Stolz geschwellter Brust, dass wir wohl zwischen 16 und 24 Gewinnen zu erwarten haben. Sie solle sich schonmal einen Teddybär aussuchen. Hundert nervenaufreibende Öffnungsversuche später stehen wir mit nur 8 Gewinnen und hängenden Schultern da. Kann das Zufall sein? Ja, aber wir beobachten ein sehr unwahrscheinliches Ereignis (später werden wir solche Überlegungen mit einem statistischen Test präzise formulieren können). Wir beschliessen, den hühnerhaften und grimmig drein schauenden Losbudenbesitzer nicht mit unserer Erkenntnis zu verärgern und denken uns: "Pech im Spiel, Glück in der Liebe!"

2.6.1 Kumulative Verteilungsfunktion

Manchmal ist es für Rechnungen nützlicher, statt der "Liste" $P(X = x)$ (für alle x) die sukzessiven Summen

$$\sum_{y \in W_X; y \leq x} P(X = y) = P(X \leq x)$$

anzugeben. Dabei läuft x ebenfalls über den Wertebereich W_X von X . Man kann in dieser Definition aber auch beliebige reelle Werte x betrachten und erhält dann eine Funktion

$$F(x) = P(X \leq x) = \sum_{y \in W_X; y \leq x} P(X = y),$$

die sogenannte **kumulative Verteilungsfunktion**. Diese springt an den Stellen, die zum Wertebereich gehören, und ist dazwischen konstant. Siehe auch Abbildung 2.2.

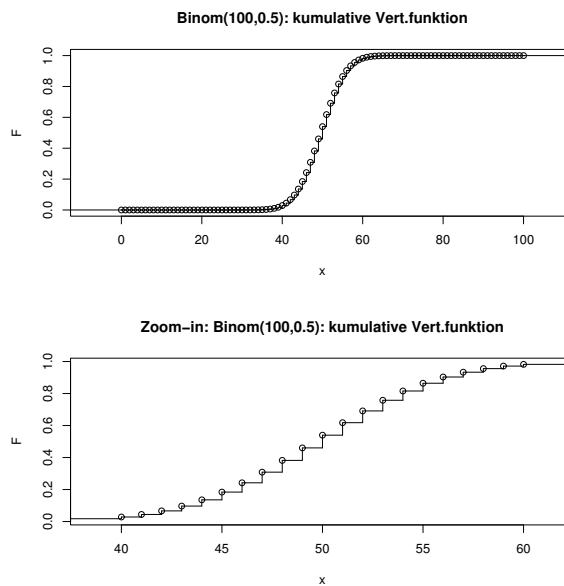


Abbildung 2.2: Kumulative Verteilungsfunktion $F(\cdot)$ für $X \sim \text{Binomial}(100, 0.5)$. Unten: zoom-in für die Werte $x \in [40, 60]$. Die Kreise zeigen an, dass an den Sprungstellen der obere Wert gilt.

Aus der kumulativen Verteilungsfunktion kann man die "Liste" $P(X = x)$ zurückgewinnen: $P(X = x)$ ist einfach die Höhe des Sprungs an der Stelle x . Insbesondere gilt für X mit Wertebereich in den ganzen Zahlen und ganzzahliges x

$$P(X = x) = F(x) - F(x - 1), \quad P(X \geq x) = 1 - P(X \leq x - 1) = 1 - F(x - 1).$$

2.7 Weitere Verteilungen (Stahel Kap. 5.2)

2.7.1 Poissonverteilung

Der Wertebereich der Binomial(n, π)-Verteilung ist $W = \{0, 1, \dots, n\}$. Falls eine Zufallsvariable nicht im vornherein einen beschränkten Wertebereich hat,

so bietet sich für Zähldaten die Poisson-Verteilung an.

Eine Zufallsvariable X mit Werten in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ heisst
Poisson(λ)-verteilt, falls

$$P(X = x) = \exp(-\lambda) \frac{\lambda^x}{x!} \quad (x = 0, 1, 2, \dots)$$

wobei $\lambda > 0$ ein Parameter der Verteilung ist.

Die Poisson-Verteilung ist die Standardverteilung für unbeschränkte **Zähl-**
daten.

Beispiele: Die Poisson(λ)-Verteilung kann bei folgenden Anwendungen als Modell gebraucht werden:

Anzahl Schadenmeldungen eines Versicherten pro Jahr,

Anzahl spontaner Ereignisse in einer Nervenzelle während einer Sekunde via Transmitterfreisetzung an einer Synapse.

Die Kennzahlen sind wie folgt: für $X \sim \text{Poisson}(\lambda)$:

$$\mathcal{E}(X) = \lambda, \quad \text{Var}(X) = \lambda, \quad \sigma(X) = \sqrt{\lambda}.$$

Poisson-Approximation der Binomial-Verteilung

Betrachte $X \sim \text{Binomial}(n, \pi)$ und $Y \sim \text{Poisson}(\lambda)$. Falls n gross und π klein mit $\lambda = n\pi$, dann:

$$P(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \approx P(Y = x) = \exp(-\lambda) \frac{\lambda^x}{x!} \quad (x = 0, 1, \dots, n).$$

Das heisst: für grosse n und kleine π : $\text{Binomial}(n, \pi) \approx \text{Poisson}(\lambda)$ für $\lambda = n\pi$. Mit anderen Worten: die Poisson-Verteilung kann interpretiert werden als Verteilung für **seltenere Ereignisse bei vielen unabhängigen Versuchen** (selten für einen einzelnen Fall, die Gesamt-Anzahl kann trotzdem gross sein).

Summen von Poisson-verteilten Zufallsvariablen

Die Poisson-Verteilung hat die folgende Additionseigenschaft: Wenn $X \sim \text{Poisson}(\lambda_X)$ und $Y \sim \text{Poisson}(\lambda_Y)$ **unabhängig** sind, dann ist $X + Y \sim \text{Poisson}(\lambda_X + \lambda_Y)$. Wenn also zum Beispiel die Anzahlen spontaner Ereignisse in einer Nervenzelle in zwei disjunkten Zeitintervallen Poisson-verteilt und unabhängig sind, dann ist auch das Total wieder Poisson-verteilt. Wir erhalten also eine Poisson-Verteilung für alle Intervalle. Weil sich bei der Addition der Zufallsvariablen die Parameter der Poisson-Verteilung addieren, ist üblicherweise λ proportional zur Länge des betrachteten Zeitintervalls.

Oft entstehen Poisson verteilte Daten von einem sog. **Poisson-Prozess** auf einem Set S (z.B. Zeit, Fläche oder Raum) mit Parameter k . Das Modell sagt folgendes aus: Wenn S_1, S_2, \dots, S_n nicht überlappende Subsets von S sind (z.B. Zeitintervalle, Teile der Gesamtfläche oder Teile vom Raum), dann sind die Anzahl Ereignisse N_1, N_2, \dots, N_n in jedem Subset unabhängige Zufallsvariablen, die jeweils einer Poisson-Verteilung mit den Parametern $k \cdot |S_1|, k \cdot |S_2|, \dots, k \cdot |S_n|$ folgen. Dabei ist $|S_i|$ die "Grösse" von dem Set S_i (z.B. Zeitdauer in Sekunden, Fläche in m^2 oder Volumen in m^3).

Beispiel Angenommen ein Büro erhält Telefonanrufe als ein Poisson-Prozess mit $k = 0.5$ pro Minute⁷. Die Anzahl Anrufe in einem 5 Minuten Intervall folgt dann einer Poisson-Verteilung mit Parameter $\lambda = 5 \cdot k = 2.5$. Die Wahrscheinlichkeit, dass es in einem 5 Minuten Intervall keine Anrufe gibt, ist daher $\exp(-2.5) = 0.082$. Die Wahrscheinlichkeit, dass es genau einen Anruf gibt ist $2.5 \cdot \exp(-2.5)$.

2.7.2 Hypergeometrische Verteilung

Die Binomialverteilung beschreibt die Anzahl Gewinne bei n Losen, wobei jedes Los (unabhängig von den anderen Losen) die gleiche Gewinnwahrscheinlichkeit p hat. In der Praxis könnte man sich dazu eine Losbude vorstellen, bei der die n Lose aus einer grossen Lostrommel mit vielen Losen gezogen werden. Angenommen, es gibt in der ganzen Lostrommel 100 Lose und darunter sind 10 Gewinne. Wenn wir zufällig ein Los ziehen, ist die Wahrscheinlichkeit für einen Gewinn $10/100 = 0.1$ (mit dem Laplace-Modell). Jetzt hat die Lostrommel also nur noch 99 Lose mit 9 Gewinnen und 90 Nieten. Wenn wir nochmals zufällig ein Los ziehen, ist die Wahrscheinlichkeit für einen Gewinn $9/99 \approx 0.091$. Die Gewinnwahrscheinlichkeit für die Lose ist also **nicht konstant**, sondern hängt davon ab, wie viele Gewinne oder Nieten vorher schon gezogen wurden.

Wenn es sich um eine grosse Lostrommel mit sehr vielen Losen handelt und wir nur wenige Lose ziehen, ist dieser Effekt völlig vernachlässigbar. Dann ist die Binomialverteilung eine sehr gute Approximation. Wenn die Lostrommel aber klein ist (z.B. nur 10 Lose und wir ziehen 8 davon), ist die Binomialverteilung eine schlechte Approximation. Für solche Fälle gibt es die Hypergeometrische Verteilung.

Angenommen, wir haben eine Urne mit N Kugeln. Davon sind m Kugeln markiert (= Erfolg) und wir ziehen n Kugeln (ohne Zurücklegen; die Erfolgswahrscheinlichkeit ändert sich also bei jedem Zug) aus der Urne. Die Zufallsvariable X beschreibt die Anzahl markierter Kugeln unter den n gezogenen Kugeln. Dann gilt X ist hypergeometrisch verteilt: $X \sim \text{Hyper}(N, n, m)$. Die Wahrscheinlichkeitsverteilung von X lässt sich mit dem Laplace Modell berechnen:

$$P(X = x) = \frac{\binom{m}{x} \binom{N-m}{n-x}}{\binom{N}{n}}$$

⁷Beachte, dass hier "pro Minute" steht. Der Parameter des Poisson-Prozesses misst immer eine Anzahl geteilt durch etwas (z.B. Zeitdauer, Fläche, Volumen, etc.)

Weiter gilt:

$$\begin{aligned}\mathcal{E}(X) &= \frac{nm}{N} \\ \text{Var}(X) &= \frac{nm(N-m)(N-n)}{N^2(N-1)}\end{aligned}$$

2.8 Software

In der Statistiksoftware R sind sehr viele Wahrscheinlichkeitsverteilungen schon vorprogrammiert. Wenn man eine Zufallsvariable X hat, die einer Verteilung namens "xxx" (xxx steht für eine Abkürzung des Namens der Verteilung, z.B. "binom" oder "pois") folgt, kann man folgende drei Dinge bequem erledigen:

dxxx Berechnet $P[X = x]$

pxxx Berechnet $P[X \leq x]$

rxxx Liefert eine Zufallszahl gemäss der Verteilung von X

Konkrete Beispiele sind: **dbinom**, **pbinom**, **rbinom** für die Binomialverteilung und **dpois**, **ppois**, **rpois** für die Poissonverteilung. Schauen Sie sich die Hilfe-seite dieser Funktionen an, in dem Sie vor den Befehl ein "?" tippen. Also z.B. **?dbinom**.

In R gibt es noch weitere Modelle für Zähldaten: Geometrische Verteilung (Abkürzung **geom**), Negative Binomialverteilung (Abkürzung **nbinom**), Hypergeometrische Verteilung (Abkürzung **hyper**), usw. Die genannten Verteilungen werden in der Praxis oft gebraucht. Wenn es Sie interessiert, schauen Sie doch mal auf Wikipedia nach, in welchen Fällen diese drei Verteilungen Verwendung finden.

Kapitel 3

Statistik für Zählraten

Lernziele

- Sie kennen die Momentenmethode und die Maximum-Likelihood-Methode zum Schätzen von Parametern und können Sie auf einfache Beispiele anwenden.
- Sie können die sechsstufige Struktur eines statistischen Tests aufschreiben und haben die Bestandteile verstanden. Insbesondere kennen Sie die Begriffe Fehler 1. Art, Fehler 2. Art und Macht eines Tests.
- Sie können testen, ob eine vorgegebene Gewinnwahrscheinlichkeit bei einer Binomialverteilung plausibel ist (Binomialtest) und können den Test auf dem Papier durchführen (einseitig und zweiseitig). Sie können entscheiden, ob ein einseitiger oder ein zweiseitiger Test angebracht ist.
- Sie haben zwei Definitionen des P-Wertes verstanden und können ihn in kleinen Beispielen auf dem Papier berechnen.
- Sie haben begriffen, was ein Vertrauensintervall ist. Sie können ein approximatives Vertrauensintervall für eine Gewinnwahrscheinlichkeit berechnen.
- Sie können einen Binomialtest mit R durchführen und auswerten.

3.1 Drei Grundfragestellungen der Statistik (Stahel Kap. 7.1)

Die Statistik fasst Beobachtungen als **Realisierung** von Zufallsvariablen auf. Sie untersucht dann, was für Schlüsse man aus den Beobachtungen über die zu Grunde liegende Verteilung ziehen kann. Meist zieht man nur Verteilungen in Betracht, die bis auf einen (oder wenige) Parameter bestimmt sind und möchte

dann Fragen über diesen Parameter beantworten. Man unterscheidet die folgenden drei Grundfragestellungen:

1. Grundfragestellung: Welches ist der zu den Beobachtungen plausibelste Parameterwert? Die Antwort auf diese 1. Grundfrage heisst **(Punkt-)Schätzung**.

2. Grundfragestellung: Sind die Beobachtungen kompatibel (statistisch vereinbar) mit einem vorgegebenen Parameterwert? Die Antwort auf diese 2. Grundfrage heisst **statistischer Test**.

3. Grundfragestellung: Welche Parameterwerte sind mit den Beobachtungen kompatibel (statistisch vereinbar)? Die Antwort auf diese 3. Grundfrage heisst **Konfidenzintervall** oder **Vertrauensintervall**. Das Konfidenzintervall ist allgemeiner und informativer als ein statistischer Test.

Beispiel (Forts.): Sei $x = 11$ die effektive Anzahl weiblicher gezüchteter Kuhkälber beim Spermasexing (vgl. Kapitel 2.5). Wir fassen $x = 11$ als **Realisierung** einer Zufallsvariablen X auf, und nehmen an, dass X $\text{Binom}(12, \pi)$ -verteilt ist. Wir möchten jetzt von der Beobachtung $x = 11$ auf den unbekannten Parameter π Schlüsse ziehen. Die Grundfragen lauten hier:

1. Welches ist der plausibelste Wert π (zu der Beobachtung $x = 11$)?
2. Ist die Beobachtung $x = 11$ kompatibel mit $\pi = 0.5$ (reiner Zufall) oder mit $\pi = 0.7$ (ökonomisch zu wenig interessant) ?
3. Welcher Bereich (Intervall) für den Parameter π ist mit der Beobachtung $x = 11$ kompatibel?

3.2 Schätzung, statistischer Test und Vertrauensintervall bei Binomial-Verteilung (Stahel Kap. 7.2, 8.2, 9.1, 9.2)

Wir betrachten folgende Situation: Gegeben ist eine Beobachtung x , welche als Realisierung von $X \sim \text{Binomial}(n, \pi)$ aufgefasst wird. Wir möchten Schlüsse ziehen über den unbekannten Parameter π .

3.2.1 (Punkt-)Schätzung

Konvention: Um einen Schätzwert eines Wertes zu kennzeichnen, wir ein Hut ($\hat{\cdot}$) auf die Variable gesetzt. Z.B. ist $\hat{\pi}$ ein Schätzwert des Parameters π ; $\widehat{\mathcal{E}(X)}$ ist ein Schätzwert des wahren Erwartungswertes $\mathcal{E}(X)$; \hat{y} ist ein Schätzwert der Variable y .

Es gibt zwei verbreitete Methoden um Parameter einer Wahrscheinlichkeitsverteilung zu schätzen.

Momentenmethode

Eine Schätzung für π kann pragmatisch hergeleitet werden. Da $\mathcal{E}(X) = n\pi$ (siehe Kapitel 2.6) gilt: $\pi = \mathcal{E}(X)/n$. Der Wert n (Anzahl unabhängiger Versuche) ist als bekannt vorausgesetzt: die einzige Unbekannte ist dann $\mathcal{E}(X)$. Eine pragmatisch motivierte Schätzung ist dann: $\widehat{\mathcal{E}(X)} = x$ (= beobachtete Anzahl Gewinne), d.h. man ersetzt den Erwartungswert durch die Beobachtung. Somit ergibt sich die relative Häufigkeit

$$\hat{\pi} = x/n$$

als Schätzung der Erfolgswahrscheinlichkeit.

Beispiel (Münzwurf) Wir haben eine Münze, bei der wir uns fragen, ob sie fair ist oder systematisch eher Kopf ergibt. Um dies zu beantworten, wird die Münze 100-mal geworfen, und wir erhalten 58 mal Kopf.

Betrachte X = Anzahl Kopf (K) bei 100 Würfeln. Es ist vernünftig, das Modell $X \sim \text{Binomial}(100, \pi)$ zu benutzen. Beobachtet (realisiert) wurde $x = 58$. Die Wahrscheinlichkeit, dass die Münze bei einem Wurf Kopf zeigt, ist gemäss der Momentenmethode also $P[\text{Kopf}] = \frac{58}{100} = 0.58$.

Beispiel (Forts. Spermasexing) Im Beispiel ist also die geschätzte Erfolgswahrscheinlichkeit für das Verfahren gleich $\frac{11}{12} = 0.917$. Es leuchtet unmittelbar ein, dass bei regelmässiger Anwendung des Verfahrens langfristig nicht genau 11 von 12 Kälbern das gewünschte Geschlecht haben werden. Mit andern Worten: Die wahre Erfolgswahrscheinlichkeit π ist nicht das Gleiche wie die geschätzte Erfolgswahrscheinlichkeit $\hat{\pi}$. Idealerweise gibt man einen Bereich an, in dem die wahre Erfolgswahrscheinlichkeit mit grosser Sicherheit liegt (ein Vertrauensintervall). Dazu später mehr.

Maximum-Likelihood

Am einfachsten lässt sich diese Methode an einem Beispiel erklären:

Beispiel (Forts. Münzwurf) Wir haben angenommen, dass die Anzahl "Kopf" bei n Münzwürfen folgendermassen verteilt ist: $X \sim \text{Binomial}(n, \pi)$. In unserem Beispiel ist $n = 100$ und die Zufallsvariable X hat den Wert 58 angenommen. Die Aufgabe besteht nun darin, einen Wert für π zu finden, der möglichst gut zu unserer Beobachtung passt. Welches Kriterium könnte man verwenden um zu zeigen, dass ein Wert π_1 besser zu der Beobachtung passt als π_2 ? Eine Möglichkeit ist die Folgende: Wir berechnen die Wahrscheinlichkeit, genau 58 mal Kopf bei 100 Münzwürfen zu erzielen. Einmal verwenden wir dabei π_1 und das andere mal π_2 . Anschliessend wählen wir das π , dass zu der grösseren Wahrscheinlichkeit für 58 mal Kopf führt. In der Praxis möchte man natürlich nicht nur zwei Werte von π vergleichen, sondern am besten alle, die denkbar sind. Wir wollen also den Ausdruck $P[X = x] = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$ (wobei $n = 100$ und $x = 58$) bzgl. π maximieren. Dieses Problem lässt sich leicht mit der Mathematik, die Sie im Basisjahr gelernt haben (Ableitung gleich null setzen und nach π auflösen), lösen. Oft ist der Ausdruck, den man ableiten muss aber recht

kompliziert (wie auch in unserem Beispiel). In vielen Fällen kann man dann einen Trick anwenden: Jedes Extremum der Funktion $f(x)$ ist auch ein Extremum der Funktion $\log(f(x))$ und umgekehrt. D.h., anstatt $P[X = x]$ bzgl. π zu maximieren, können wir auch $\log(P[X = x])$ bzgl. π maximieren, falls das leichter geht. Das Ergebnis ist völlig identisch. In unserem Fall ist die zweite Variante tatsächlich etwas einfacher:

$$\begin{aligned}\log(P[X = x]) &= \log\left(\binom{n}{x} \pi^x (1 - \pi)^{n-x}\right) = \\ &= \log\left(\binom{n}{x}\right) + \log(\pi^x) + \log((1 - \pi)^{n-x}) = \\ &= \log\left(\binom{n}{x}\right) + x \cdot \log(\pi) + (n - x) \cdot \log(1 - \pi)\end{aligned}$$

Durch Ableiten nach π und Nullsetzen erhalten wir¹:

$$\begin{aligned}\frac{d}{d\pi}(\log\left(\binom{n}{x}\right) + x \cdot \log(\pi) + (n - x) \cdot \log(1 - \pi)) &= 0 \\ \frac{x}{\pi} + (n - x) \cdot \frac{1}{1 - \pi} \cdot (-1) &= 0\end{aligned}$$

Wenn wir diese Gleichung nach π auflösen, erhalten wir:

$$\begin{aligned}\frac{x}{\pi} - (n - x) \cdot \frac{1}{1 - \pi} &= 0 \\ \frac{x}{\pi} &= \frac{n - x}{1 - \pi} \\ x - \pi x &= \pi n - \pi x \\ \pi &= \frac{x}{n}\end{aligned}$$

Mit den Zahlen in unserem Beispiel erhalten wir $\pi = \frac{58}{100} = 0.58$. In diesem Beispiel ist das Ergebnis identisch mit dem Ergebnis aus der Momentenmethode. Im Allgemeinen ist das aber nicht der Fall.

Mit obiger Methode wählen wir also das π , mit dem die Beobachtung am wahrscheinlichsten ist. Daher nennt man diese Methode auf englisch "Maximum-Likelihood-Methode". Sie ist die mit Abstand gebräuchlichste Methode, um Parameter zu schätzen und oft der Momentenmethode überlegen.

3.2.2 Statistischer Test

Die Grundidee eines statistischen Tests ist bestechend einfach. Wenn Sie sie nicht mehr im Kopf haben, lesen Sie nochmal das Beispiel in Kapitel 1.1 nach.

Beispiel (Forts. Münzwurf): Ist es plausibel bei einer fairen Münze 58 mal oder öfter Kopf zu werfen? Die zentrale Frage ist, ob $\pi = 1/2$ oder $\pi > 1/2$.

¹Beachte dabei, dass $\log\left(\binom{n}{x}\right)$ nicht von π abhängt und deshalb beim Ableiten verschwindet. Ausserdem gilt: $\frac{d}{dx} \log(x) = \frac{1}{x}$

Motivation

Im obigen Beispiel stellen wir die folgende Überlegung an. Wir nehmen einmal an, dass die Münze fair ist, d.h. dass $\pi = 1/2$, und berechnen die Wahrscheinlichkeiten für Ereignisse von der Form $\{X \geq c\}$ für "grosse" Werte c , weil grosse Werte eher für $\pi > 1/2$ sprechen. Wir wollen damit herausfinden, wie plausibel die beobachtete Anzahl $x = 58$ bei einer fairen Münze ist. Die folgende Tabelle liefert die Zahlen für $X \sim \text{Binomial}(100, 1/2)$.

	$c = 53$	$c = 54$	$c = 55$	$c = 56$	$c = 57$	$c = 58$	$c = 59$	$c = 60$
$P(X \geq c)$	0.309	0.242	0.184	0.136	0.097	0.067	0.044	0.028

Wenn die Münze fair wäre, würde man das Ereignis "58 mal oder öfter Kopf" mit einer Wahrscheinlichkeit von nur ca. 6.7% beobachten. D.h., wenn Sie und weitere 999 Studenten jeweils 100 mal diese Münze werfen, würden etwa 67 Studenten das Ereignis "58 mal oder öfter Kopf" erleben. Diese Zahl (6.7%) wird in der Literatur auch **P-Wert** genannt (später noch mehr dazu). In diesem Fall ist der P-Wert also 6.7%. In der Wissenschaft hat es sich so ergeben, dass man einer Hypothese nicht mehr glaubt, wenn der P-Wert kleiner als 5% ist².

Formales Vorgehen

Ein statistischer Test für den Parameter π im Modell $X \sim \text{Binomial}(n, \pi)$ ist wie folgt aufgebaut.

1. **Modell:** X : Anzahl Treffer bei n Versuchen;
 $X \sim \text{Binomial}(n, \pi)$
2. Spezifiziere die sogenannte **Nullhypothese** H_0 :

$$H_0 : \pi = \pi_0,$$

und (anhand der Problemstellung) eine sogenannte **Alternative** H_A :

$$H_A : \begin{array}{l} \pi \neq \pi_0 \text{ (zwei-seitig)} \\ \pi > \pi_0 \text{ (ein-seitig nach oben)} \\ \pi < \pi_0 \text{ (ein-seitig nach unten).} \end{array}$$

Am häufigsten ist die Nullhypothese $H_0 : \pi = 1/2$ (d.h. $\pi_0 = 1/2$), also "reiner Zufall" oder "kein Effekt". Meist führt man einen Test durch, weil man glaubt, dass die Alternative richtig ist und man auch Skeptiker davon überzeugen möchte.

²Diese Grenze ist natürlich völlig willkürlich. Sie könnte genauso gut 4.7%, 7% oder irgendeine andere kleine Prozentzahl sein. Wichtig ist nur, dass man einer Hypothese keinen glauben schenkt, wenn der P-Wert "klein" ist.

3. **Teststatistik:** T : Anzahl Treffer bei n Versuchen.
Verteilung der Teststatistik unter H_0 : $T \sim \text{Bin}(n, \pi_0)$ ³
4. Lege das sogenannte **Signifikanzniveau** α fest. Typischerweise wählt man $\alpha = 0.05$ (5%) oder auch $\alpha = 0.01$ (1%).
5. Bestimme den sogenannten **Verwerfungsbereich** K für die Teststatistik⁴. Qualitativ zeigt K in Richtung der Alternative:

$$\begin{aligned} K &= [0, c_u] \cup [c_o, n] && \text{falls } H_A : \pi \neq \pi_0, \\ K &= [c, n] && \text{falls } H_A : \pi > \pi_0, \\ K &= [0, c] && \text{falls } H_A : \pi < \pi_0. \end{aligned}$$

Quantitativ wird K so berechnet, dass

$$P_{H_0}(X \in K) = \underbrace{P_{\pi_0}}_{\text{von Binomial}(n, \pi_0)} (X \in K) \stackrel{\approx}{\leq} \alpha.$$

Dabei bedeutet $\stackrel{\approx}{\leq}$, dass die linke Seite kleiner oder gleich der rechten Seite sein soll, aber so nahe wie möglich.

6. **Testentscheid:** Erst jetzt betrachte, ob der beobachtete Wert der Teststatistik in den Verwerfungsbereich K fällt:
 Falls ja: so verwerfe H_0 (H_0 ist dann statistisch widerlegt, die Abweichung von der Nullhypothese ist "signifikant")
 Falls nein: belasse H_0 (was nicht heisst, dass deswegen H_0 statistisch bewiesen ist).
 Diese Art der Test-Entscheidung beruht auf dem Widerspruchs-Prinzip: Hypothesen können nur falsifiziert und nicht verifiziert werden.

Beispiel (Forts.): Für das Beispiel mit den Münzwürfen heisst das konkret:

1. **Modell:** X : Anzahl Würfe, die Kopf zeigen, wenn man 100 mal wirft.
 $X \sim \text{Bin}(100, \pi)$.
2. **Nullhypothese:** $H_0 : \pi = 0.5$
Alternative: $H_A : \pi > 0.5$
3. **Teststatistik:** T : Anzahl Würfe, die Kopf zeigen, wenn man 100 mal wirft.
Verteilung der Teststatistik unter H_0 : $T \sim \text{Bin}(100, 0.5)$

³Es mag verwirrend sein, dass die Teststatistik T und X genau gleich sind. Das ist nur in diesem einfachen Beispiel der Fall und im Allgemeinen nicht so. Später werden wir Beispiele kennenlernen, in denen die Teststatistik T eine Funktion von X ist.

⁴Der Verwerfungsbereich hängt vom Signifikanzniveau α ab. Wie er grösser oder kleiner, wenn α kleiner wird?

4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich:** $K = [59; 100]$ ⁵
6. **Testentscheid:** Der beobachtete Wert der Teststatistik ($t = 58$)⁶ liegt nicht im Verwerfungsbereich ($K = [59; 100]$). Daher kann die Nullhypothese auf dem Signifikanzniveau 5% nicht verworfen werden. Das heisst, es gibt nicht genügend statistische Evidenz (auf dem Signifikanzniveau $\alpha = 0.05$) dafür, dass die Münze zu Gunsten von Kopf (K) gefälscht ist.

Beispiel (Forts.): Beim Spermasexing (vgl. Kapitel 2.5) wurden $x = 11$ Kuhkälber geboren von insgesamt 12 Kälbern. Es scheint ziemlich klar zu sein, dass dies nicht reiner Zufall sein kann. Wir wollen trotzdem noch sehen, was der Test sagt:

1. **Modell:** X : Anzahl weiblicher Kälber;
 $X \sim \text{Binomial}(12, \pi)$
2. **Nullhypothese:** $H_0 : \pi = \pi_0 = 0.5$,
Alternative: $H_A : \pi > \pi_0 = 0.5$.
3. **Teststatistik:** T : Anzahl weiblicher Kälber;
Verteilung der Teststatistik unter H_0 : $T \sim \text{Bin}(12, 0.5)$
4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich:** Mit dem Computer oder Taschenrechner kann man folgende Tabelle leicht berechnen:

	$c = 8$	$c = 9$	$c = 10$	$c = 11$	$c = 12$
$P(X \geq c)$	0.194	0.073	0.019	0.003	0.0002
- Daher ist der Verwerfungsbereich für die Teststatistik: $K = [10, 11, 12]$.
6. **Testentscheid:** Der beobachtete Wert der Teststatistik ($t = 11$) liegt im Verwerfungsbereich der Teststatistik. Daher kann die Nullhypothese auf dem Signifikanzniveau 5% verworfen werden. D.h., mit der Methode "Spermasexing" werden tatsächlich mehr weibliche Kälber geboren.

Für das Signifikanzniveau $\alpha = 0.05$ ist also der Verwerfungsbereich $K = \{10, 11, 12\}$. Für das Niveau $\alpha = 0.01$ ist $K = \{11, 12\}$. Für beide Niveaus wird die Nullhypothese also verworfen, d.h. der Effekt der Methode "Spermasexing" ist statistisch signifikant, sowohl auf dem 5%- als auch auf dem 1%-Niveau.

Wenn jemand nur an einer Methode interessiert ist, deren Erfolgswahrscheinlichkeit grösser als 70% ist (weil die Methode z.B. erst dann ökonomisch interessant wird), dann wird sie in Punkt 2 des Tests folgende Änderung machen:

⁵Aus obiger Tabelle kann man ablesen, dass $c = 59$ die kleinste Zahl ist, bei der $P[X \geq c]$ kleiner als 5% ist.

⁶Beachten Sie, dass ich hier ein kleines t und nicht T verwendet habe, weil ich eine konkrete Realisierung der Zufallsvariable T mit dem Buchstaben t bezeichnen will.

$$H_0 : \pi = \pi_0 = 0.7$$

$$H_A : \pi > \pi_0 = 0.7$$

Mit Hilfe einer Tabelle erhält man aus $P_{\pi=0.7}(X \in K) \approx 0.05$ den Verwerfungsbereich $K = \{12\}$.

Der beobachtete Wert der Teststatistik liegt nun nicht mehr im Verwerfungsbereich. Daher wird H_0 belassen, d.h. eine Erfolgswahrscheinlichkeit von über 70% ist nicht signifikant nachgewiesen.

Bisher haben wir den Verwerfungsbereich K immer exakt berechnet mit Hilfe einer Tabelle der kumulativen Verteilungsfunktion der Binomialverteilung (statistische Software liefert solche Tabellen). Für Überschlagsrechnungen kann man auch die sogenannte **Normalapproximation** benutzen: Für $\alpha = 0.05$ und die Alternative $\pi > \pi_0$ gilt

$$c \approx n\pi_0 + 1.64\sqrt{n\pi_0(1-\pi_0)} \text{ aufgerundet auf die nächstgrössere ganze Zahl}$$

Für $\alpha = 0.05$ und die Alternative $\pi < \pi_0$ gilt

$$c \approx n\pi_0 - 1.64\sqrt{n\pi_0(1-\pi_0)} \text{ abgerundet auf die nächstkleinere ganze Zahl.}$$

Für $\alpha = 0.05$ und die Alternative $\pi \neq \pi_0$ schliesslich erhält man c_o , bzw. c_u , indem man in den obigen Formeln 1.64 durch 1.96 ersetzt.

Beispiel: Wir betrachten wie in obigem Beispiel den Test mit den $n = 100$ Münzwürfen und testen $H_0 : \pi = 0.5$ gegen $H_A : \pi > 0.5$ auf dem $\alpha = 0.05$ Signifikanzniveau. Diesmal berechnen wir den Verwerfungsbereich mit Hilfe der Normalapproximation:

$$c \approx 100 \cdot 0.5 + 1.64 \cdot \sqrt{100 \cdot 0.5 \cdot (1 - 0.5)} = 58.2$$

Aufgerundet zur nächstgrösseren Zahl ist also $c \approx 59$. Daher ist der Verwerfungsbereich mit der Normalapproximation $[59; 100]$. Das Ergebnis entspricht dem Ergebnis in obigem Beispiel, bei dem wir die exakte Binomialverteilung verwendet haben.

Beispiel: Wir berechnen nun noch den Verwerfungsbereich für obiges Beispiel zum Spermasexing. Es gilt:

$$c \approx 12 \cdot 0.5 + 1.64 \cdot \sqrt{12 \cdot 0.5 \cdot (1 - 0.5)} = 8.83$$

Aufgerundet ergibt das $c \approx 9$. Der Verwerfungsbereich mit der Normalapproximation ist also $[9, 10, 11, 12]$. Beachten Sie, dass wir hier **nicht** das gleiche Ergebnis wie mit der Binomialverteilung erhalten (dort hatten wir als Verwerfungsbereich $[10, 11, 12]$). Die Normalapproximation macht also einen kleinen Fehler.

Allgemein kann man sagen, dass die Normalapproximation umso besser funktioniert, je grösser die Anzahl Versuche n ist. Sie funktioniert nicht so gut, wenn π nahe bei 0 oder 1 ist. Als Faustregel kann man sich merken, dass die Normalapproximation gut ist, falls $n\pi > 5$ und $n(1 - \pi) > 5$.

Fehler 1. und 2. Art

Bei einem statistischen Test treten 2 Arten von Fehlern auf.

Fehler 1. Art: Fälschliches Verwerfen von H_0 , obwohl H_0 richtig ist.

Fehler 2. Art: Fälschliches Beibehalten von H_0 , obwohl die Alternative zutrifft.

Per Definition des Fehlers 1. Art gilt

$$P(\text{Fehler 1. Art}) = P_{H_0}(X \in K) \leq \alpha.$$

Nach Konstruktion des Tests ist der Fehler 1. Art höchstens gleich α . Dem Fehler 1. Art wird traditionell mehr Gewicht gegeben als dem Fehler 2. Art⁷. Daher wird er direkt kontrolliert mittels der Konstruktion eines Tests. Über die Wahrscheinlichkeit eines Fehlers 2. Art haben wir hingegen keine solche Kontrolle.

Die beiden Fehlerarten konkurrenzieren sich gegenseitig:

$P(\text{Fehler 2. Art})$ wird grösser falls α kleiner gewählt wird.

Die Wahl von α steuert also einen Kompromiss zwischen Fehler 1. und 2. Art. Weil man aber primär einen Fehler 1. Art vermeiden will, wählt man α klein, z.B. $\alpha = 0.05$.

Beispiel (Forts. Spermasexing): Beim Spermasexing nehmen wir einmal an, dass in Tat und Wahrheit der Parameter $\pi = 0.8 \in H_A$ ist (die Spezifikationen des Tests sind wie oben: $H_0 : \pi = 0.7$, $H_A : \pi > 0.7$ und $\alpha = 0.05$). Wie gross ist die Wahrscheinlichkeit, dass unser Test die Richtigkeit der Alternativhypothese $H_A : \pi > 0.7$ nicht entdeckt? Da der Verwerfungsbereich $K = \{12\}$ ist (siehe oben), gilt:

$$P(\text{Test behält } H_0 \text{ bei, obschon } \pi = 0.8) = P_{\pi=0.8}(X \leq 11) = 1 - P_{\pi=0.8}(X = 12) = 0.93.$$

Das heisst, dass ein Fehler 2. Art (unter der Annahme dass $\pi = 0.8$) mit grosser Wahrscheinlichkeit auftritt. Das ist natürlich enttäuschend, wenn $\pi = 0.8$ ökonomisch interessant wäre. Bei der kleinen Anzahl von 12 Versuchen, kann man einfach nur sehr schlecht zwischen $\pi = 0.7$ und $\pi = 0.8$ entscheiden. Beachte, dass die Wahrscheinlichkeit für einen Fehler 1. Art höchstens 0.05, also klein ist.

Statt der Wahrscheinlichkeit eines Fehlers 2. Art gibt man oft die sogenannte **Macht** an. Diese ist definiert als

$$\text{Macht} = 1 - P(\text{Fehler 2. Art}) = P(\text{Verwerfen von } H_0 \text{ falls } H_A \text{ stimmt}) = P_{H_A}(X \in K).$$

Die Macht gibt also die Wahrscheinlichkeit an, H_A zu entdecken, falls H_A richtig ist.

⁷Wissenschaftler arbeiten genau und haben Angst, einen Humbug zu publizieren, der sich dann als falsch herausstellt. Wenn man einen Effekt beobachtet, möchte man sicher sein, dass es sich nicht bloss um Zufall handelt (Fehler 1. Art soll vermieden werden). Dabei nimmt man in Kauf, dass man manchmal einen wichtigen Effekt verpasst (Fehler 2. Art ist zweitrangig).

Beispiel (Forts. Spermasexing) Im Beispiel ist die Macht $1 - 0.93 = 0.07$.

Wenn man mehrere Alternativen hat, hängen die Macht und die Wahrscheinlichkeit eines Fehlers 2. Art davon ab, welche Alternative man betrachtet.

Einseitige und zweiseitige Tests

Im Beispiel mit dem Münzwurf wollten wir testen, ob es sich um eine faire Münze handelt. Wir hatten die Vermutung, dass Kopf zu oft oben liegt und haben deshalb einen einseitigen Test ($H_A : \pi > \pi_0$ oder $H_A : \pi < \pi_0$) gemacht. Die Münze wäre aber auch unfair, wenn Kopf zu häufig **oder** zu selten oben liegt. In diesem Fall ist ein zweiseitiger Test angebracht ($H_A : \pi \neq \pi_0$). Oft stellt sich die Frage, ob ein einseitiger oder ein zweiseitiger Test für eine Fragestellung besser geeignet wäre. Im Folgenden zeigen wir, dass beide Tests Vor- und Nachteile haben.

Zweiseitige Tests detektieren sowohl zu viele Erfolge als auch zu wenige Erfolge. Wenn der Test einseitig mit $H_A : \pi > \pi_0$ ist, dann detektiert er nur zu viele Erfolge. Wenn der Test einseitig mit $H_A : \pi < \pi_0$ ist, dann detektiert er nur zu wenige Erfolge. Damit scheint der zweiseitige Test überlegen zu sein. Für die "Sehschärfe" in beide Richtungen zahlt man allerdings einen Preis:

- Der zweiseitige Test detektiert zwar Abweichungen in beide Richtungen von H_0 , aber die Abweichung muss sehr deutlich sein, damit er sie erkennt. D.h., die Macht des zweiseitigen Tests ist klein.
- Der einseitige Test detektiert nur Abweichungen in eine Richtung von H_0 , aber die Abweichungen müssen nicht so gross sein, damit er sie erkennt. D.h., die Macht des einseitigen Tests ist gross.

Es ist gängige Praxis, die Entscheidung bzgl. der Form von H_A zu fällen, **bevor** die Beobachtungen gemacht werden.

Beispiel: Wir haben wieder eine Münze und wollen prüfen, ob sie fair ist (Kopf und Zahl mit Wahrscheinlichkeit 50%). Wir werfen 10 Münzen. Zunächst machen wir einen einseitigen Test:

1. **Modell:** X : Anzahl Würfe, die Kopf zeigen, wenn man 10 mal wirft.
 $X \sim \text{Bin}(10, \pi)$.
2. **Nullhypothese:** $H_0 : \pi = 0.5$
Alternative: $H_A : \pi > 0.5$ ⁸
3. **Teststatistik:** T : Anzahl Würfe, die Kopf zeigen, wenn man 10 mal wirft.
Verteilung der Teststatistik unter H_0 : $T \sim \text{Bin}(10, 0.5)$

⁸So detektieren wir nur, wenn Kopf "zu häufig" vorkommt. Um zu testen, ob Kopf "zu selten" vorkommt, hätten wir $H_A : \pi < 0.5$ wählen müssen. Ich habe mich in diesem Beispiel willkürlich für die erste Variante entschieden.

4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich:** $K_1 = \{8, 9, 10\}$
6. **Testentscheid:** Weil ich nicht gesagt habe, wie oft wir Kopf beobachten, können wir den Testentscheid nicht fällen.

Im zweiseitigen Test erhalten wir einen anderen Verwerfungsbereich:

1. **Modell:** X : Anzahl Würfe, die Kopf zeigen, wenn man 10 mal wirft.
 $X \sim \text{Bin}(10, \pi)$.
2. **Nullhypothese:** $H_0 : \pi = 0.5$
Alternative: $H_A : \pi \neq 0.5$
3. **Teststatistik:** T : Anzahl Würfe, die Kopf zeigen, wenn man 10 mal wirft.
Verteilung der Teststatistik unter H_0 : $T \sim \text{Bin}(10, 0.5)$
4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich:** Den zweiseitigen Verwerfungsbereich kann man leicht berechnen: Zunächst bestimmt man die Verwerfungsbereiche ($K_>$ und $K_<$) für die einseitigen Alternativen $H_A : \pi > 0.5$ und $H_A : \pi < 0.5$ mit dem **halben** Signifikanzniveau $\frac{\alpha}{2}$, also hier 0.025. Anschliessend nimmt man die Vereinigung der beiden Verwerfungsbereiche. Mit folgender Tabelle⁹ ergibt sich $K_> = \{9, 10\}$. Für die andere Alternativhypothese ergibt sich $K_< = \{0, 1\}$.

	0	1	2	8	9	10
$P(X \geq c)$				0.055	0.011	0.001
$P(X \leq c)$	0.001	0.011	0.055			

Für den Verwerfungsbereich des zweiseitigen Tests erhalten wir also:

$$K_2 = K_< \cup K_> = \{0, 1\} \cup \{9, 10\}$$

6. **Testentscheid:** Weil ich nicht gesagt habe, wie oft wir Kopf beobachten, können wir den Testentscheid nicht fällen.

Angenommen, in Wahrheit ist die Wahrscheinlichkeit für Kopf $\pi = 0.8$. Die Macht entspricht der Wahrscheinlichkeit, dass das wahre Modell ($\pi = 0.8$) eine Beobachtung erzeugt, die in den Verwerfungsbereich fällt. Diese Wahrscheinlichkeiten können wir leicht ausrechnen. Zunächst berechnen wir die Macht M_1 des einseitigen Tests (wir nehmen an, dass $X \sim \text{Bin}(10, 0.8)$):

$$\begin{aligned} M_1 &= P(X \in K_1) = P(X \in \{8, 9, 10\}) = \\ &= P(X = 8) + P(X = 9) + P(X = 10) = \\ &= \binom{10}{8} 0.8^8 0.2^2 + \binom{10}{9} 0.8^9 0.2^1 + \binom{10}{10} 0.8^{10} = 0.68 \end{aligned}$$

⁹Ich habe in der Tabelle nur die relevanten Zahlen eingetragen. Können Sie sich erklären, warum die Zahlen in der Tabelle in beiden Zeilen gleich sind? Wäre das auch so, wenn wir $H_0 : \pi = 0.7$ testen würden?

Nun berechnen wir die Macht M_2 des zweiseitigen Tests (wir nehmen wieder an, dass $X \sim \text{Bin}(10, 0.8)$):

$$\begin{aligned} M_2 &= P(X \in K_2) = P(X \in \{0, 1\} \cup \{9, 10\}) = \\ &= P(X = 0) + P(X = 1) + P(X = 9) + P(X = 10) = \\ &= \binom{10}{0} 0.2^{10} + \binom{10}{1} 0.8^1 0.2^9 + \binom{10}{9} 0.8^9 0.2^1 + \binom{10}{10} 0.8^{10} = 0.38 \end{aligned}$$

Die Macht des einseitigen Tests ist also viel grösser als die des zweiseitigen Tests. Wenn man sich also sicher ist, dass nur die Abweichung in die eine Richtung von der Nullhypothese relevant ist, ist der einseitige Test wegen der grösseren Macht vorzuziehen.

Der P-Wert

Der P-Wert ist ein Wert zwischen 0 und 1, der angibt, wie gut Nullhypothese und Daten zusammenpassen (0: passt gar nicht; 1: passt sehr gut). Genauer gesagt ist der P-Wert die Wahrscheinlichkeit, unter Gültigkeit der Nullhypothese das erhaltene Ergebnis oder ein extremeres¹⁰ zu erhalten (vgl. Abb. 3.1). Mit dem P-Wert wird also angedeutet, wie extrem das Ergebnis ist: je kleiner der P-Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese. Werte kleiner als eine im voraus festgesetzte Grenze, wie 5%, 1% oder 0.1% sind Anlass, die Nullhypothese abzulehnen.

Es gibt noch eine andere, äquivalente Definition des P-Wertes¹¹: Die Entscheidung eines Tests mit “Verwerfen” oder “Beibehalten” der Nullhypothese H_0 ist abhängig von der etwas willkürlichen Wahl des Signifikanzniveaus α . Mathematisch bedeutet dies, dass der Verwerfungsbereich $K = K(\alpha)$ abhängig von der Wahl von α ist.

Man kann sich einfach überlegen, dass qualitativ Folgendes gilt:

Verwerfungsbereich $K = K(\alpha)$ wird kleiner mit kleiner werdendem α ,

denn α ist ja die Wahrscheinlichkeit für einen Fehler 1. Art, und diese wird natürlich dann klein, wenn wir weniger oft die Nullhypothese H_0 verwerfen. Umgekehrt gilt natürlich auch, dass $K = K(\alpha)$ grösser wird mit wachsendem α . Dies impliziert: es gibt ein Signifikanzniveau, bei dem die Nullhypothese H_0 “gerade noch” verworfen wird.

Der P-Wert ist definiert als das kleinste Signifikanzniveau,
bei dem die Nullhypothese H_0 (gerade noch) verworfen wird

¹⁰“extrem” bezieht sich auf die Alternativhypothese. Wenn $H_A : \pi > \pi_0$, dann sind sehr grosse Anzahlen von Treffern extrem. Wenn $H_A : \pi < \pi_0$, dann sind sehr kleine Anzahlen von Treffern extrem. Wenn $H_A : \pi \neq \pi_0$, dann sind sowohl sehr grosse als auch sehr kleine Anzahlen von Treffern extrem.

¹¹Auf den ersten Blick erscheint diese zweite Definition wohl viel komplizierter. Sie sollten sie aber trotzdem durchdenken, weil Sie damit die Begriffe Signifikanzniveau und Vertrauensintervall besser verinnerlichen. In der Praxis ist die erste Version allerdings besser zu behalten.

Zusammenfassend haben wir also zwei äquivalente Definitionen des P-Wertes:

1. Der P-Wert ist die Wahrscheinlichkeit, unter Gültigkeit der Nullhypothese das beobachtete Ergebnis oder ein extremeres zu erhalten.
2. Der P-Wert ist das kleinste Signifikanzniveau, bei dem die Nullhypothese H_0 (gerade noch) verworfen wird.

Der P-Wert liefert mehr Information als bloss die Test-Entscheidung bei einem vorbestimmten Signifikanzniveau α (z.B. $\alpha = 0.05$). Insbesondere gilt aufgrund der Definition des P-Werts:

verwerfe H_0 falls P-Wert $\leq \alpha$
 belasse H_0 falls P-Wert $> \alpha$.

Zusätzlich zu dieser Entscheidungsregel quantifiziert der P-Wert wie signifikant eine Alternative ist (d.h. wie gross die Evidenz ist für das Verwerfen von H_0). Manchmal werden sprachliche Formeln oder Symbole anstelle der P-Werte angegeben:

P-Wert ≈ 0.05 : schwach signifikant, " . "
 P-Wert ≈ 0.01 : signifikant, " * "
 P-Wert ≈ 0.001 : stark signifikant, " * * "
 P-Wert $\leq 10^{-4}$: äusserst signifikant, " * * * "

Beispiel (Forts.): Beim Spermasexing betrachten wir die Null-Hypothese $\pi = 0.7$ und die Alternative $\pi > 0.7$. Beobachtet wurde $x = 11$, aufgefasst als Realisierung von $X \sim \text{Binomial}(12, \pi)$. Der P-Wert ist dann:

$$P_{\pi=0.7}(X \geq 11) = P_{\pi=0.7}(X = 11) + P_{\pi=0.7}(X = 12) = 0.085.$$

Wie wir bereits früher gesehen haben, liefert dies kein Verwerfen von H_0 auf dem Signifikanzniveau $\alpha = 0.05$ (weil der P-Wert grösser als 0.05 ist). Auf dem Signifikanzniveau $\alpha = 0.1$ könnte man H_0 allerdings verwerfen, weil der P-Wert kleiner als 0.1 ist.

3.2.3 Vertrauensintervall

Informativer als ein statistischer Test ist ein sogenanntes Vertrauensintervall (auch Konfidenzintervall genannt). Es liefert eine Antwort auf die 3. Grundfragestellung von Kapitel 3.1: Welche Werte von π sind mit der Beobachtung x kompatibel (statistisch vereinbar).

Ein Vertrauensintervall I zum Niveau $1 - \alpha$ besteht aus allen Parameterwerten, die im Sinne des statistischen Tests zum Signifikanzniveau α mit der Beobachtung verträglich sind (üblicherweise nimmt man den zweiseitigen Test). Mathematisch heisst dies:

$$I = \{\pi_0; \text{ Nullhypothese } H_0 : \pi = \pi_0 \text{ wird belassen}\}. \quad (3.1)$$

Einseitiger Test mit Alternative $H_A: \pi > \pi_0$

Verteilung von X unter $H_0: \pi = \pi_0$

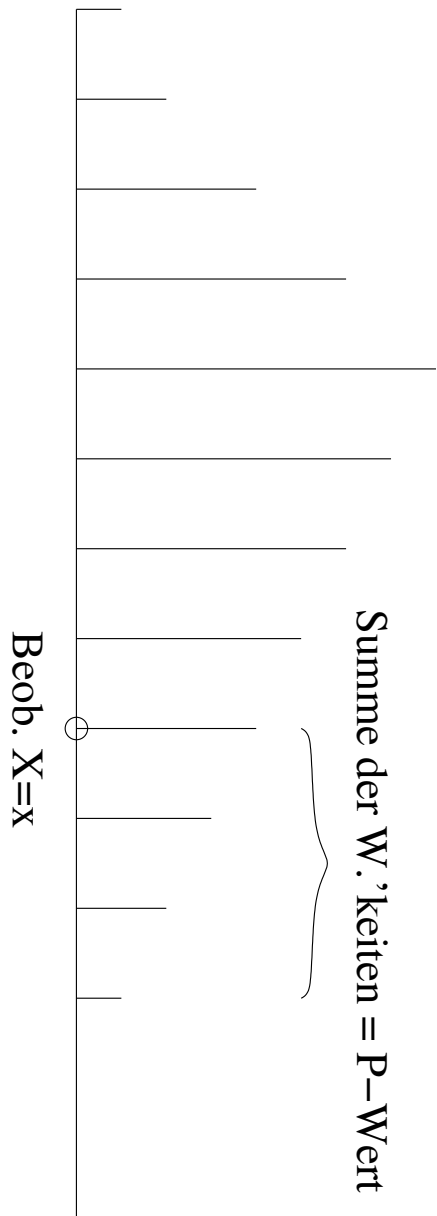


Abbildung 3.1: Schematische Darstellung des P-Werts bei einer einseitigen Alternative $H_A: \pi > \pi_0$.

Diese Beziehung stellt eine Dualität zwischen Tests und Vertrauensintervallen dar.

Die Berechnung erfolgt in der Praxis meist mit dem Computer. Falls n “gross” ist, so kann die sogenannte Normalapproximation (siehe Kap. 4.6) benützt werden. Letztere ergibt folgendes approximatives Konfidenzintervall I zum Niveau $1 - \alpha = 0.95$ für den unbekannten Parameter π :

$$I \approx \frac{x}{n} \pm 1.96 \sqrt{\frac{x}{n} \left(1 - \frac{x}{n}\right) \frac{1}{n}} \quad (3.2)$$

Das Vertrauensintervall $I = I(x)$ hängt von der Beobachtung ab. Wenn man anstelle der Beobachtung die zugehörige Zufallsvariable X einsetzt, so ist $I(X)$ zufällig und hat die Eigenschaft:

$$P_\pi(\pi \in I(X)) \stackrel{\approx}{>} 1 - \alpha \quad \text{für jedes } \pi.$$

Dies kann so interpretiert werden, dass das Konfidenzintervall I den wahren Parameter π mit Wahrscheinlichkeit $1 - \alpha$ mit einschliesst.

Beispiel: Um das zu veranschaulichen, haben wir mit dem Computer folgende Situation simuliert: Hundert Personen gehen auf den Jahrmarkt. Alle kommen an einer Losbude vorbei, bei der die wahre Gewinnwahrscheinlichkeit $\pi = 0.1$ ist. Allerdings ist die Gewinnwahrscheinlichkeit nirgends angeschrieben, sodass die Personen sie nicht kennen. Jede Person beschliesst nun 50 Lose zu kaufen und ein 95%-Vertrauensintervall für die Gewinnwahrscheinlichkeit zu bestimmen. Da ein 95%-Vertrauensintervall den wahren Parameter mit 95% Wahrscheinlichkeit enthält, sollten rund 95 Personen von den 100 den wahren Parameter in ihrem Vertrauensintervall haben¹². In Abb. 3.2 ist das Ergebnis dieser Simulation dargestellt.

Beispiel (Forts.): Beim Spermasexing erhält man für ein zweiseitiges Konfidenzintervall zum Niveau $1 - \alpha = 0.95$ mittels einer Tabelle oder dem Computer für die Berechnung von (3.1):

$$I = (0.615, 0.998)$$

Das heisst, dass der wahre “Zucht”-Parameter π mit einer Wahrscheinlichkeit von 95% in I liegt. Es besteht also auf Grund der kleinen Stichprobe grosse Unsicherheit, wie erfolgreich die Methode bei langfristigem Einsatz tatsächlich sein wird. Die Näherungsformel in (3.2) ist für dieses Beispiel nicht besonders gut, weil $n = 12$ eher klein ist. Man erhält mit (3.2):

$$I \approx (0.760, 1.073)$$

Der rechte Endpunkt ist natürlich zu gross, denn der Parameter π ist ja kleiner oder gleich 1.

¹²Wer es ganz genau wissen will: Es werden wahrscheinlich nicht **genau** 95 Personen sein. Die Anzahl Personen P , die den wahren Parameter in ihrem Vertrauensintervall haben ist binomialverteilt: $P \sim \text{Bin}(100, 0.95)$.

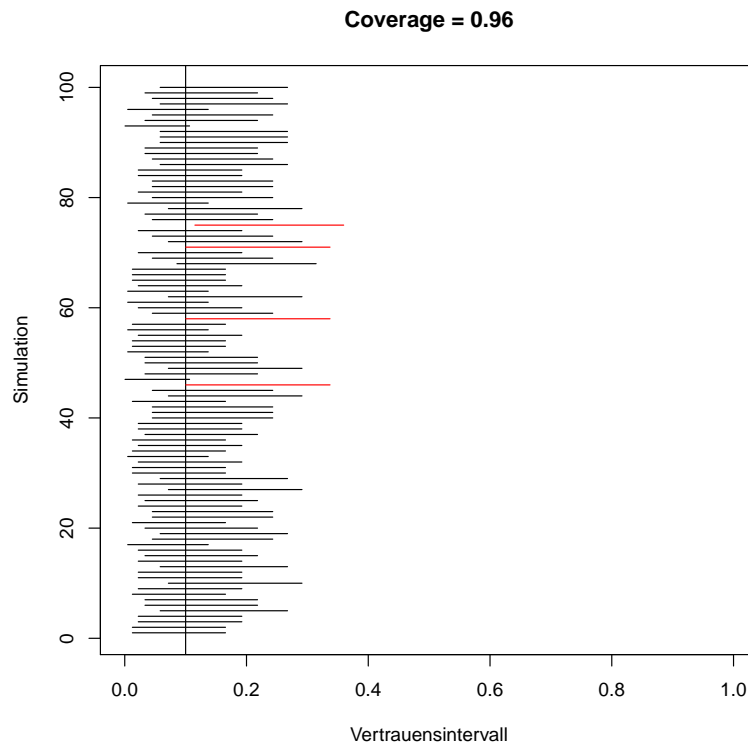


Abbildung 3.2: Vergleich von 100 95%-Vertrauensintervallen. Auf der x-Achse ist die Gewinnwahrscheinlichkeit aufgetragen. Der wahre Wert ist 0.1. Darüber sind 100 horizontale Striche gezeichnet. Jeder Strich repräsentiert das 95%-Vertrauensintervall für die Gewinnwahrscheinlichkeit von einer der 100 Personen. Der vertikale Strich zeigt die Position der wahren Gewinnwahrscheinlichkeit. Man sieht, dass 4 Vertrauensintervall die wahre Gewinnwahrscheinlichkeit nicht enthalten (sie sind rot markiert). Die übrigen 96 Vertrauensintervalle enthalten die wahre Gewinnwahrscheinlichkeit. Die Abbildung zeigt also, dass ein 95%-Vertrauensintervall mit ca. 95% Wahrscheinlichkeit den wahren Parameter enthält.

3.3 Software

In der Statistik-Software **R** ist die Funktion `binom.test` zum Durchführen eines Binomialtests schon vorprogrammiert. Um das Vertrauensintervall zu berechnen wird dort eine genauere (aber auch kompliziertere) Methode als die Normalapproximation verwendet. Sehen Sie sich die Hilfeseite mit `?binom.test` an.

Beispiel: Wir haben eine Münze hundertmal geworfen und 58 mal Kopf gesehen. Wir wollen einen Test mit den Hypothesen $H_0 : \pi = 0.5$ und $H_A : \pi \neq 0.5$ durchführen. Mit **R** erledigt das der Befehl

```
binom.test(x=58, n=100, p=0.5, alternative= "two.sided", conf.level=0.95)
```

Als Computeroutput erhalten wir Folgendes:

```
Exact binomial test
```

```
data: 58 and 100
number of successes = 58, number of trials = 100, p-value = 0.1332
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4771192 0.6780145
sample estimates:
probability of success
          0.58
```

In der ersten Zeile (*data*) werden die Inputdaten erneut aufgelistet. In der zweiten Zeile steht die Interpretation der Inputdaten für den Binomialtest: Wir hatten 58 Erfolge bei 100 Versuchen. Bei der von uns festgelegten Nullhypothese und Alternative entspricht das einem P-Wert von 0.1332. In der dritten Zeile wird die Alternativhypothese in Worten ausformuliert. In der vierten und fünften Zeile steht das 95%-Vertrauensintervall für die Gewinnwahrscheinlichkeit: Mit 95% Wahrscheinlichkeit liegt die Gewinnwahrscheinlichkeit im Bereich [0.477, 0.678]. In den letzten drei Zeilen erscheint noch der Punktschätzer für die Gewinnwahrscheinlichkeit. Wie wir schon vorher gesehen hatten, ist er sowohl für die Momentenmethode also auch für die Maximum-Likelihood-Methode 0.58.

Beispiel (Forts. Spermasexing) Wir berechnen den P-Wert aus obigem Beispiel nun mit dem Computer: Beim Spermasexing betrachten wir die Null-Hypothese $\pi = 0.7$ und die Alternative $\pi > 0.7$. Beobachtet wurde $x = 11$, aufgefasst als Realisierung von $X \sim \text{Binomial}(12, \pi)$. Mit R lässt sich der Test so durchführen: `binom.test(x=11, n=12, p=0.7, alternative= "greater", conf.level=0.95)`

Das Ergebnis ist:

```
Exact binomial test
```

```
data: 11 and 12
number of successes = 11, number of trials = 12, p-value = 0.08503
alternative hypothesis: true probability of success is greater than 0.7
95 percent confidence interval:
 0.6613193 1.0000000
sample estimates:
probability of success
          0.9166667
```

Der Computer liefert den gleichen P-Wert wie in obigem Beispiel.

Kapitel 4

Modelle und Statistik für Messdaten

4.1 Lernziele

- Sie kennen Methoden der deskriptiven Statistik, können sie interpretieren und folgende Größen ausrechnen: arithmetisches Mittel, Standardabweichung, Varianz, Quantil, Median, Korrelation; Histogramm, Boxplot, empirische kumulative Verteilungsfunktion, Normal-Plot, Streudiagramm.
- Sie kennen das Konzept der Wahrscheinlichkeitsdichte und der kumulativen Verteilungsfunktion; Sie kennen die uniforme Verteilung, die Exponentialverteilung und die Normalverteilung. Sie können die Normalverteilung standardisieren und mit der Tabelle für beliebige Normalverteilungen Wahrscheinlichkeiten ausrechnen.
- Sie können den Erwartungswert, Varianz und Quantil für linear transformierte Zufallsvariablen berechnen. Sie können den Erwartungswert und die Varianz für Summen von unabhängigen Zufallsvariablen und von Zufallsvariablen mit vorgegebener Korrelation berechnen.
- Sie kennen das Gesetz der grossen Zahlen und das \sqrt{n} -Gesetz. Sie kennen den Zentralen Grenzwertsatz und können ihn auf einfache Sachverhalte anwenden.
- Sie kennen den Unterschied zwischen Standardabweichung einer Einzelbeobachtung X_i und dem Standardfehler.
- Sie können den z-, t- und Vorzeichen-Test für eine Stichprobe auf Papier und Computer durchführen (einseitig und zweiseitig). Sie können entscheiden, ob ein einseitiger oder ein zweiseitiger Test angebracht ist. Sie können den Computeroutput eines t-Tests interpretieren (bei der Software R).
- Sie können den t-Test für zwei Stichproben auf Papier und Computer durchführen (einseitig, zweiseitig; gepaart, ungepaart). Sie können ent-

scheiden, ob ein einseitiger/zweiseitiger und gepaarter/ungepaarter Test angebracht ist. Sie können Computeroutputs eines t-Tests interpretieren (bei der Software R).

- Sie wissen, dass es den Vorzeichen-Test (evtl. auch den Wilcoxon-Test) als Alternative zum t-Test gibt.

4.2 Einleitung

In vielen Anwendungen hat man es nicht mit Zähl-, sondern mit Messdaten zu tun, bei denen die Werte im Prinzip kontinuierlich sind. Zur Illustration betrachten wir zwei Datensätze. Beim ersten werden zwei Methoden zur Bestimmung der latenten Schmelzwärme von Eis verglichen. Wiederholte Messungen der freigesetzten Wärme beim Übergang von Eis bei -0.72°C zu Wasser bei 0°C ergaben die folgenden Werte (in cal/g):

Methode A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03
Methode A	80.02	80.00	80.02							
Methode B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97		

Obwohl die Messungen mit der grösstmöglichen Sorgfalt durchgeführt und alle Störeinflüsse ausgeschaltet wurden, variieren die Messungen von Fall zu Fall. Wir werden diese Variationen innerhalb der Messreihen als zufällig modellieren, das heisst wir interpretieren diese Werte als Realisierungen von Zufallsvariablen. Wir werden dann die Frage beantworten, ob die Unterschiede zwischen den Methoden ebenfalls als zufällig angesehen werden können, oder ob ein systematischer Unterschied plausibler ist, der auch in der ganzen Population, d.h. in weiteren Messungen, bestehen bleibt. Im letzteren Fall werden wir dann noch zusätzlich angeben, wie gross der systematische Unterschied etwa ist.

Im zweiten Beispiel wurde bei 11 Individuen die Aggregation von Blutplättchen vor und nach dem Rauchen einer Zigarette gemessen. Die folgenden Daten geben den Anteil aggregierter Blutplättchen (in Prozent) nach einer Stimulation an.

Individuum	1	2	3	4	5	6	7	8	9	10	11
Vorher	25	25	27	44	30	67	53	53	52	60	28
Nachher	27	29	37	56	46	82	57	80	61	59	43

Wieder variieren die Werte in einer nicht vorhersehbaren Art. Diesmal handelt es sich jedoch weniger um Messfehler, sondern um Variation zwischen Individuen (vermutlich gäbe es auch noch eine gewisse Variation beim gleichen Individuum, wenn der Test wiederholt würde). Die Aggregation bei diesen 11 Individuen ist meistens, aber nicht immer nach dem Rauchen höher, und die Fragestellung lautet, ob es sich hier um einen zufälligen Effekt handelt, der auf die spezifische Stichprobe beschränkt ist, oder ob man dieses Resultat auf eine grössere Population verallgemeinern kann. Im letzteren Fall möchte man wieder angeben, wie gross die mittlere Zunahme etwa ist.

4.3 Deskriptive Statistik (Stahel, Kap. 2 und 3.1, 3.2)

Bei einer statistischen Analyse ist es wichtig, nicht einfach blind ein Modell anzupassen oder ein statistisches Verfahren anzuwenden. Die Daten sollten immer mit Hilfe von geeigneten grafischen Mitteln dargestellt werden, da man nur auf diese Weise unerwartete Strukturen und Besonderheiten entdecken kann. Kennzahlen können einen Datensatz grob charakterisieren. Im Folgenden werden die Daten mit x_1, \dots, x_n bezeichnet.

4.3.1 Kennzahlen

Häufig will man die Verteilung der Daten numerisch zusammenfassen. Dazu braucht man mindestens zwei Kenngrößen, eine für die Lage und eine für die Streuung. Die bekanntesten solchen Größen sind das *arithmetische Mittel* für die Lage,

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

und die *empirische Standardabweichung* für die Streuung,

$$s_x = \sqrt{\text{var}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

(der Nenner $n-1$, anstelle von n , ist mathematisch begründet und hat die Eigenschaft, dass kein “systematischer” Fehler auftritt).

Alternative Kenngrößen sind der *Median* als Lagemass und die *Quartilsdifferenz* als Streuungsmass. Diese werden mit Hilfe von sogenannten Quantilen definiert.

Quantile

Das *empirische α -Quantil* ist anschaulich gesprochen der Wert, bei dem $\alpha \times 100\%$ der Datenpunkte kleiner und $(1 - \alpha) \times 100\%$ der Punkte grösser sind.

Zur formalen Definition führen wir die geordneten Werte ein:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Das empirische α -Quantil ist dann gleich

$$\begin{aligned} & \frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)}) \quad \text{falls } \alpha \cdot n \text{ eine ganze Zahl ist,} \\ & x_{(k)} \quad \text{wobei } k = \alpha n + \frac{1}{2} \text{ gerundet auf eine ganze Zahl; falls } \alpha \cdot n \text{ keine ganze Zahl ist.} \end{aligned}$$

Der (empirische) Median ist das empirische 50%-Quantil: d.h., es markiert die “mittlere” Beobachtung, wenn die Daten sortiert wurden. Es ist also ein Mass für die Lage der Daten.

Die Quartilsdifferenz ist gleich

$$\text{empirisches 75\%-Quantil} - \text{empirisches 25\%-Quantil}$$

und ist ein Streuungsmass für die Daten.

Median und Quartilsdifferenz haben den Vorteil, dass sie robust sind: das heisst, dass sie viel weniger stark durch extreme Beobachtungen beeinflusst werden können als arithmetisches Mittel und Standardabweichung.

Beispiel: Messung der Schmelzwärme von Eis mit Methode A

Das arithmetische Mittel der $n = 13$ Messungen ist $\bar{x} = 80.02$ und die Standardabweichung ist $s_x = 0.024$. Ferner ist für $n = 13$ $0.25n = 3.25$, $0.5n = 6.5$ und $0.75n = 9.75$. Damit ist das 25%-Quantil gleich $x_{(4)} = 80.02$, der Median gleich $x_{(7)} = 80.03$ und das 75%-Quantil gleich $x_{(10)} = 80.04$. Wenn bei der grössten Beobachtung ($x_9 = 80.05$) ein Tippfehler passiert wäre und $x_9 = 800.5$ eingegeben worden wäre, dann wäre $\bar{x} = 135.44$ und der Median nach wie vor $x_{(7)} = 80.03$. Das arithmetische Mittel wird also durch Veränderung einer Beobachtung sehr stark beeinflusst, während der Median gleich bleibt - er ist “robust”.

Korrelation und empirische Korrelation

Die gemeinsame Verteilung von abhängigen Zufallsvariablen X und Y ist i.A. kompliziert, und man begnügt man sich oft mit einer vereinfachenden Kennzahl zur Beschreibung der Abhängigkeit. Die **Kovarianz** und **Korrelation** zwischen X und Y sind wie folgt definiert:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \quad (\text{Kovarianz})$$

$$\text{Corr}(X, Y) = \rho_{XY} = \text{Cov}(X, Y) / (\sigma_X \sigma_Y) \quad (\text{Korrelation}),$$

wobei $\sigma_X = \sqrt{\text{Var}(X)}$, und analog für σ_Y .

Die Korrelation ρ_{XY} ist eine dimensionslose, normierte Zahl mit Werten $\rho_{XY} \in [-1, 1]$.

Die Korrelation misst Stärke und Richtung der **linearen Abhängigkeit** zwischen X und Y . Es gilt

$$\text{Corr}(X, Y) = +1 \text{ genau dann wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b > 0,$$

$$\text{Corr}(X, Y) = -1 \text{ genau dann wenn } Y = a + bX \text{ für ein } a \in \mathbb{R} \text{ und ein } b < 0.$$

Überdies gilt:

$$X \text{ und } Y \text{ unabhängig} \implies \text{Corr}(X, Y) = 0. \quad (4.1)$$

Die Umkehrung gilt i.A. nicht.

Für die numerische Zusammenfassung der Abhängigkeit ist die **empirische Korrelation** r (oder auch mit $\hat{\rho}$ bezeichnet) am gebräuchlichsten:

$$r = \frac{s_{xy}}{s_x s_y}, \quad s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}.$$

Die empirische Korrelation ist eine dimensionslose Zahl zwischen -1 und +1. Das Vorzeichen von r misst die Richtung und der Betrag die Stärke des linearen Zusammenhangs zwischen den beiden Variablen. Man sollte jedoch nie r berechnen, ohne einen Blick auf das Streudiagramm zu werfen, da ganz verschiedene Strukturen den gleichen Wert von r ergeben können. Siehe dazu Abb. 4.1.

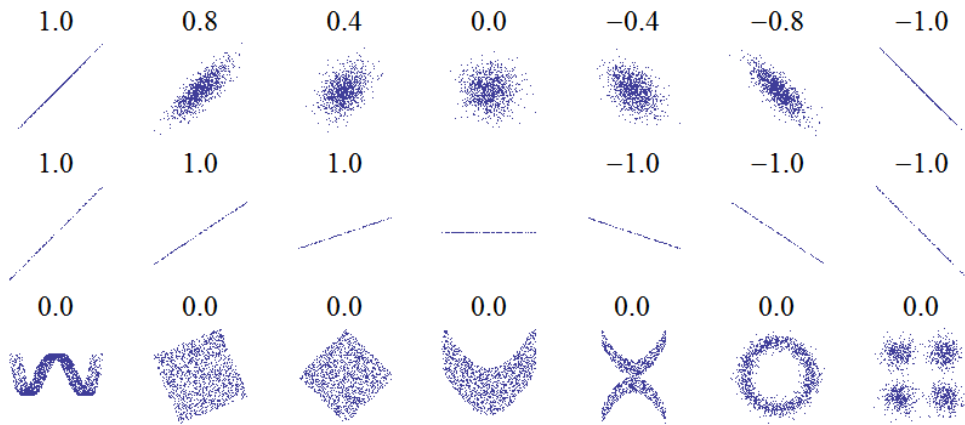


Abbildung 4.1: Es sind 21 verschiedene Datensätze dargestellt, die je aus vielen Beobachtungspaaren (x, y) bestehen. Für jedes Paar wurde ein Punkt gezeichnet. Über jedem Datensatz steht jeweils die zugehörige empirische Korrelation. Bei perfektem linearen Zusammenhang, ist die empirische Korrelation +1, -1 oder 0 (je nachdem ob die Steigung positiv, negativ oder null ist; siehe zweite Zeile). Je mehr die Punkte um den linearen Zusammenhang streuen, desto kleiner wird der Betrag der empirischen Korrelation (siehe erste Zeile). Da die empirische Korrelation nur den *linearen* Zusammenhang misst, kann es einen Zusammenhang zwischen den beiden Variablen x und y geben, auch wenn die empirische Korrelation null ist (siehe unterste Zeile).

Standardisierung

Durch Verschiebung und Skalierung der Werte kann man erreichen, dass zwei oder mehrere Datensätze die gleiche Lage und Streuung haben. Insbesondere kann man einen Datensatz so standardisieren, dass das arithmetische Mittel gleich Null und die Standardabweichung gleich 1 ist. Dies erreicht man mittels der linear transformierten Variablen

$$z_i = \frac{x_i - \bar{x}}{s_x} \quad (i = 1, \dots, n).$$

Alle Aspekte einer Verteilung, die bei einer Verschiebung oder Skalierung unverändert bleiben, machen die Form der Verteilung aus. Dazu gehört insbesondere die Schiefe (Asymmetrie) der Verteilung, für die es auch Kennzahlen gibt.

4.3.2 Grafische Methoden

Einen Überblick über die auftretenden Werte ergibt das *Histogramm*. Es gibt verschiedene Arten von Histogrammen; wir behandeln nur die gebräuchlichste. Um ein Histogramm zu zeichnen, bildet man Klassen konstanter Breite und zählt, wie viele Beobachtungen in jede Klasse fallen. Dann zeichnet man für jede Klasse einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist.¹

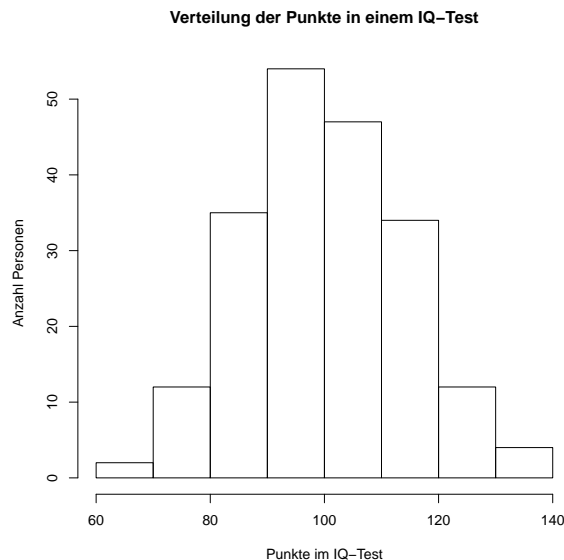


Abbildung 4.2: Histogramm von dem IQ-Test Ergebnis von 200 Personen. Die breite der Klassen wurde als 10 IQ-Punkte festgelegt und ist für jede Klasse gleich. Die Höhe der Balken gibt die Anzahl Personen an, die in diese Klasse fallen. Z.B. fallen ca. 12 Personen in die Klasse zwischen 120 und 130 IQ-Punkten.

Beim *Boxplot* hat man ein Rechteck, das vom empirischen 25%- und vom 75%-Quantil begrenzt ist, und Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten “normalen” Wert gehen (per Definition ist ein normaler Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt). Zusätzlich gibt man noch Ausreisser durch Sterne und den Median durch einen Strich an. Der Boxplot ist vor allem dann geeignet, wenn man die Verteilungen

¹Eine andere, weit verbreitete Art des Histogramms erlaubt Klassen unterschiedlicher Breite. Hier wird die vertikale Achse so gewählt, dass die Fläche des Histogramms eins ist. Die Fläche jedes Balkens ist proportional zur Anzahl Beobachtungen in der zugehörigen Klasse. Auf diese Art kann man eine Wahrscheinlichkeitsdichte (siehe unten) schätzen.

einer Variablen in verschiedenen Gruppen (die im allgemeinen verschiedenen Versuchsbedingungen entsprechen) vergleichen will; siehe Abbildung 4.3.

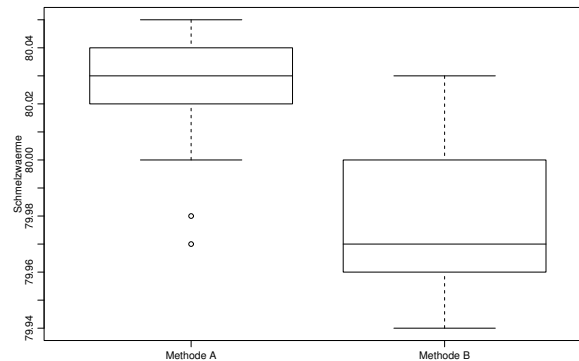


Abbildung 4.3: Boxplots für die zwei Methoden zur Bestimmung der Schmelzwärme von Eis.

Die *empirische kumulative Verteilungsfunktion* $F_n(\cdot)$ ist eine Treppenfunktion, die links von $x_{(1)}$ gleich null ist und bei jedem $x_{(i)}$ einen Sprung der Höhe $\frac{1}{n}$ hat (falls ein Wert mehrmals vorkommt, ist der Sprung ein Vielfaches von $\frac{1}{n}$). In andern Worten:

$$F_n(x) = \frac{1}{n} \text{Anzahl}\{i \mid x_i \leq x\}.$$

Abbildung 4.4 zeigt die empirische kumulative Verteilungsfunktion für die Messungen der Schmelzwärme von Eis mit Methode A.

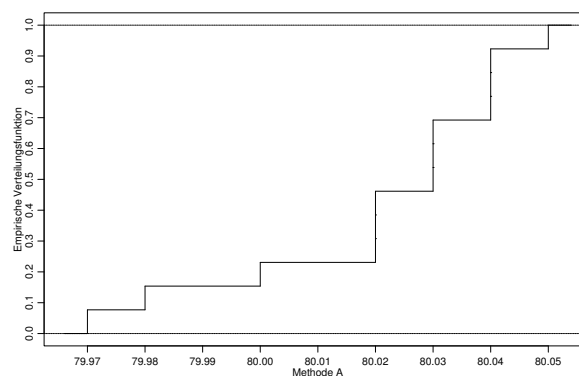


Abbildung 4.4: Empirische kumulative Verteilungsfunktion der Messungen der Schmelzwärme von Eis mit Methode A.

Mehrere Variablen

Wenn wir bei jeder Versuchseinheit zwei verschiedene Größen messen, d.h. wenn die Daten von der Form $(x_1, y_1), \dots, (x_n, y_n)$ sind, interessiert man sich in

erster Linie für die Zusammenhänge und Abhängigkeiten zwischen den Variablen. Diese kann man aus dem *Streudiagramm* erkennen, welches die Daten als Punkte in der Ebene darstellt: Die i -te Beobachtung entspricht dem Punkt mit Koordinaten (x_i, y_i) . Die Abbildung 4.5 zeigt das Streudiagramm für die Werte “vorher” und “nachher” bei der Blutplättchen-Aggregation. Man sieht einen klaren monotonen Zusammenhang, Individuen haben also eine Tendenz zu starker, bzw. schwacher Aggregation, unabhängig vom Rauchen.

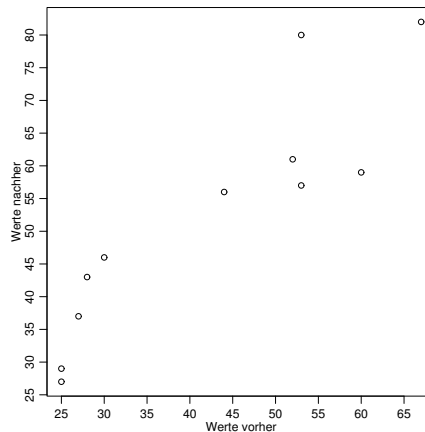


Abbildung 4.5: Streudiagramm der Blutplättchen-Aggregation vor und nach dem Rauchen einer Zigarette.

4.3.3 Analogien zwischen Modellen und Daten

Zufallsvariablen und Verteilungen beschreiben die Population. Daten x_1, \dots, x_n interpretieren wir als Realisierungen von Zufallsvariablen X_1, \dots, X_n (man könnte auch die n Daten als n Realisierungen von einer Zufallsvariablen X interpretieren; die Schreibweise mit mehreren Zufallsvariablen hat jedoch Vorteile, siehe Abschnitt 4.6).

Aus Daten können wir Rückschlüsse auf die zugrunde liegende Verteilung ziehen. Insbesondere haben alle Grössen, die für Zufallsvariablen definiert sind, ein Gegenstück für Datensätze gemäss folgender Tabelle. Die empirischen Grössen sind Schätzungen für die theoretischen Grössen. Diese werden mit wachsendem Stichprobenumfang n immer genauer.

Daten	Population (Modell)
Histogramm	Dichte
empirische kumulative Verteilungsfkt.	theoretische kumulative Verteilungsfkt.
empirische Quantile	theoretische Quantile
Arithmetisches Mittel	Erwartungswert
empirische Standardabweichung	theoretische Standardabweichung.

Es ist jedoch wichtig, die Unterschiede zwischen den empirischen und theoretischen Grössen zu verstehen: Mit Hilfe der Statistik werden wir quantifizieren,

wie gross die Unsicherheit ist, wenn wir die theoretischen Grössen mit Hilfe der empirischen Grössen schätzen.

4.4 Stetige Zufallsvariablen und Wahrscheinlichkeitsverteilungen (Stahel, Kap. 6.1 – 6.4, 11.2)

Eine Zufallsvariable X heisst stetig, wenn deren Wertebereich W_X kontinuierlich ist; z.B. $W_X = \mathbb{R}$, \mathbb{R}^+ oder $[0, 1]$.

Konvention: Um bei einem Intervall zu notieren, ob die Grenzen innerhalb oder ausserhalb des Intervalls sein sollen, verwenden wir eckige und runde Klammern. Bei einer runden Klammer ist der Wert ausserhalb des Intervalls, bei einer eckigen Klammer ist der Wert innerhalb des Intervalls. Das Intervall $(a, b]$ beschreibt also alle Punkte x mit $x > a$ und $x \leq b$.

In Kapitel 2.4 hatten wir gesehen, dass die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen beschrieben werden kann, indem man die “Punkt”-Wahrscheinlichkeiten $P(X = x)$ für alle möglichen x im Wertebereich angibt. Für eine stetige Zufallsvariable X gilt jedoch:

$$P(X = x) = 0 \text{ für alle } x \in W_X.$$

Dies impliziert, dass wir die Wahrscheinlichkeitsverteilung von X nicht mittels der Angaben von “Punkt”-Wahrscheinlichkeiten beschreiben können.

Beispiel: Diesen Sachverhalt kann man intuitiv einfach an einem Beispiel verstehen. Angenommen, wir haben eine Zufallsvariable X_0 , die jeden Wert aus der diskreten Menge $W_0 = \{0, 1, 2, \dots, 8, 9\}$ mit gleicher Wahrscheinlichkeit annimmt. Die Wahrscheinlichkeit, dass die Zufallsvariable X_0 einen konkreten Wert x (denken Sie z.B. an $x = 3$) aus dem Bereich W_0 annimmt, ist also $P[X_0 = x] = \frac{1}{10}$, weil W_0 aus zehn Elementen besteht. Jetzt vergrössern wir die diskrete Menge, indem wir jede Zahl auf eine Nachkommastelle genau angeben: $W_1 = \{0.0, 0.1, 0.2, \dots, 9.8, 9.9\}$. Die Zufallsvariable X_1 nimmt jeden Wert aus W_1 mit gleicher Wahrscheinlichkeit an; also ist $P[X_1 = x] = \frac{1}{100}$, weil W_1 aus hundert Elementen besteht. Konkret ist also auch $P[X_1 = 3.0] = \frac{1}{100}$. Wenn man noch eine Nachkommastelle hinzufügt, erhält man eine Menge aus tausend Elementen und die Wahrscheinlichkeit ein bestimmtes Element zufällig zu ziehen (z.B. die Zahl 3.00) ist nur noch $\frac{1}{1000}$. Wenn wir diese Entwicklung fortsetzen, kommen wir zu folgender Regel: Wenn jede der Zahlen zwischen 0 und kleiner als 10 mit i Nachkommastellen mit gleicher Wahrscheinlichkeit gezogen wird, ist die Wahrscheinlichkeit eine bestimmte Zahl zu ziehen nur $\frac{1}{10^{i+1}}$. Wenn wir unendlich viele Nachkommastellen zulassen, wird aus der diskreten Menge die kontinuierliche Menge $W_\infty = [0, 10]$. Entsprechend ist die Wahrscheinlichkeit ein bestimmtes Element aus dieser Menge, das auf unendlich viele Nachkommastellen genau bestimmt ist, zu ziehen gleich $P[X_\infty = x] = \frac{1}{\infty} = 0$.

Die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen X kann jedoch beschrieben werden, indem man die Wahrscheinlichkeiten für alle Inter-

valle $(a, b]$ ($a < b$) angibt:

$$P(X \in (a, b]) = P(a < X \leq b).$$

Dazu genügt es, die kumulative Verteilungsfunktion $F(x) = P(X \leq x)$ anzugeben, denn es gilt

$$P(a < X \leq b) = F(b) - F(a).$$

Zusammenfassend heisst dies, dass die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen X durch die kumulative Verteilungsfunktion beschrieben werden kann.

Weil für stetige Zufallsvariablen $P(X = a) = P(X = b) = 0$, spielt es keine Rolle, ob wir $<$ oder \leq schreiben. Obige Formeln sind jedoch auch richtig für diskrete Zufallsvariable.

4.4.1 (Wahrscheinlichkeits-)Dichte

Für stetige Zufallsvariablen können wir einen analogen Begriff zur “Punkt”-Wahrscheinlichkeit $P(X = x)$ für diskrete Variablen mit Hilfe der Ableitung gewinnen.

Die (Wahrscheinlichkeits-)Dichte $f(\cdot)$ ist definiert als Ableitung der kumulativen Verteilungsfunktion:

$$f(x) = F'(x).$$

Damit erhalten wir folgende Interpretation:

$$P(x < X \leq x + h) \approx hf(x) \quad \text{falls } h \text{ klein ist.}$$

Die Begründung dafür ist:

$$\frac{P(x < X \leq x + h)}{h} = \frac{F(x + h) - F(x)}{h} \approx f(x)$$

wobei die letzte Approximation aus der Definition einer Ableitung folgt.

Aus der Dichte kann man die kumulative Verteilungsfunktion zurückgewinnen:

$$F(x) = \int_{-\infty}^x f(y) dy$$

(weil F eine Stammfunktion von f ist und $F(-\infty) = 0$). Ausserdem gelten die folgenden Eigenschaften:

1. $f(x) \geq 0$ für alle x (da $F(\cdot)$ monoton wachsend ist)
2. $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$ (Fläche zwischen a und b unter $f(x)$)
3. $\int_{-\infty}^{\infty} f(x) dx = 1$ (wegen 2.)

4.4.2 Kennzahlen von stetigen Verteilungen

Der **Erwartungswert** $\mathcal{E}(X)$ und die **Standardabweichung** σ_X einer stetigen Zufallsvariablen X haben dieselbe Bedeutung wie im diskreten Fall in Kapitel 2.6: Der Erwartungswert beschreibt die mittlere Lage der Verteilung und die Standardabweichung deren Streuung. Die Formeln ergeben sich, indem wir beim diskreten Fall $P(X = x)$ durch $f(x)dx$ und die Summe durch ein Integral ersetzen:

$$\mathcal{E}(X) = \int_{-\infty}^{\infty} x f(x) dx,$$

$$\text{Var}(X) = \mathcal{E}((X - \mathcal{E}(X))^2) = \int_{-\infty}^{\infty} (x - \mathcal{E}(X))^2 f(x) dx, \quad \sigma_X = \sqrt{\text{Var}(X)}.$$

Die **Quantile** $q(\alpha)$ ($0 < \alpha < 1$) einer Zufallsvariablen X , bzw. deren Verteilung, sind wie folgt definiert:

$$P(X \leq q(\alpha)) = \alpha.$$

Das heisst:

$$F(q(\alpha)) = \alpha \Leftrightarrow q(\alpha) = F^{-1}(\alpha).$$

Dies kann auch so interpretiert werden, dass $q(\alpha)$ der Punkt ist, so dass die Fläche von $-\infty$ bis $q(\alpha)$ unter der Dichte $f(\cdot)$ gleich α ist. Siehe auch Abbildung 4.6. Das 50%-Quantil heisst der **Median**.

4.5 Wichtige stetige Verteilungen (Stahel, Kap. 6.2, 6.4, 6.5, 11.2)

Im diskreten Fall haben wir die Binomial- und Poissonverteilung als diskrete Verteilungen kennengelernt. In diesem Kapitel werden wir die wichtigsten stetigen Verteilungen kennenlernen.

Wir haben in Kapitel 4.4 gesehen, dass wir die Wahrscheinlichkeitsverteilung einer stetigen Zufallsvariablen mit der kumulativen Verteilungsfunktion $F(\cdot)$ oder der Dichte $f(\cdot)$ charakterisieren können.

4.5.1 Uniforme Verteilung

Die Uniforme Verteilung tritt als Formalisierung der völligen “Ignoranz” auf.

Eine Zufallsvariable X mit Wertebereich $W_X = [a, b]$ heisst $\text{Uniform}([a, b])$ verteilt, falls

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{falls } a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

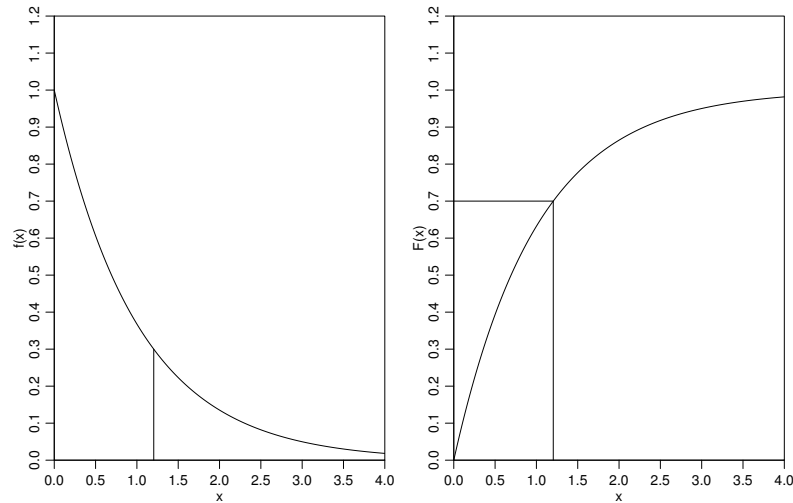


Abbildung 4.6: Illustration des 70%-Quantils $q(0.7)$. Links ist die Dichte gezeichnet (Es handelt sich um eine Exponentialverteilung) (links von $x = 0$ ist die Dichte gleich null). Die vertikale Linie wurde so gewählt, dass links von ihr die Fläche unter der Kurve gleich 0.7 ist. Dort, wo die vertikale Linie die x-Achse trifft, ist per Definition das 70%-Quantil. In diesem Fall ist das 70%-Quantil also etwa 1.2. Auf der rechten Seite ist die kumulative Verteilungsfunktion gezeigt. Auch mit dieser Abbildung kann man sehr leicht das 70%-Quantil bestimmen: Wir suchen den Wert von x , bei dem die kumulative Verteilungsfunktion gleich 0.7 ist. Das Ergebnis ist wie vorhin etwa 1.2.

Die Dichte ist also konstant auf dem Intervall $[a, b]$. Das heisst, dass die gleiche Wahrscheinlichkeit vorliegt auf dem ganzen Wertebereich $W_X = [a, b]$, deshalb der Name “uniform”.

Die zugehörige kumulative Verteilungsfunktion ist

$$F(x) = \begin{cases} 0 & \text{falls } x < a \\ \frac{x-a}{b-a} & \text{falls } a \leq x \leq b \\ 1 & \text{falls } x > b \end{cases}$$

Für $X \sim \text{Uniform}([a, b])$ sind die Kennzahlen wie folgt:

$$\mathcal{E}(X) = \frac{a+b}{2},$$

$$\text{Var}(X) = \frac{(b-a)^2}{12}, \quad \sigma_X = \frac{b-a}{\sqrt{12}}.$$

4.5.2 Exponential-Verteilung

Die Exponential-Verteilung ist das einfachste Modell für Wartezeiten auf Ausfälle.

Beispiel: Ionenkanäle

In Membranen von Muskel- und Nerven-Zellen gibt es viele Kanäle durch die Ionen fließen können, falls der Kanal offen ist. Simple kinetische Modelle motivieren, dass die Offenzeit eines Kanals mit der Exponential-Verteilung modelliert werden kann.

Eine Zufallsvariable X mit Wertebereich $W_X = \mathbb{R}^+ = [0, \infty)$ heisst Exponentialverteilt mit Parameter $\lambda \in \mathbb{R}^+$ ($X \sim \text{Exp}(\lambda)$) falls

$$f(x) = \begin{cases} \lambda \exp(-\lambda x), & \text{falls } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

Die zugehörige kumulative Verteilungsfunktion ist

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{falls } x \geq 0 \\ 0 & \text{falls } x < 0 \end{cases}$$

Die Dichte und kumulative Verteilungsfunktion für $\lambda = 1$ sind in Abbildung 4.6 zu sehen.

Für $X \sim \text{Exp}(\lambda)$ sind die Kennzahlen wie folgt:

$$\begin{aligned} \mathcal{E}(X) &= \frac{1}{\lambda}, \\ \text{Var}(X) &= \frac{1}{\lambda^2}, \quad \sigma_X = \frac{1}{\lambda}. \end{aligned}$$

Es besteht folgender Zusammenhang zwischen der Exponential- und Poisson-Verteilung: Wenn die *Zeiten* zwischen den Ausfällen eines Systems Exponential(λ)-verteilt sind, dann ist die *Anzahl* Ausfälle in einem Intervall der Länge t Poisson(λt)-verteilt.

4.5.3 Normal-Verteilung (Gauss-Verteilung)

Die Normal-Verteilung (manchmal auch Gauss-Verteilung genannt) ist die häufigste Verteilung für Messwerte.

Beispiel: Zufällige Messfehler; Summe von unabhängigen, gleichverteilten Zufallsvariablen (wegen des Zentralen Grenzwertsatzes, siehe Kap. 4.6.2)

Eine Zufallsvariable X mit Wertebereich $W_X = \mathbb{R}$ heisst Normalverteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$ (in Formeln $X \sim \mathcal{N}(\mu, \sigma^2)$) falls

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

Die zugehörige kumulative Verteilungsfunktion $F(x) = \int_{-\infty}^x f(y)dy$ ist nicht explizit darstellbar mit Standardfunktionen wie $\exp(x)$, $\log(x)$ etc. Deshalb werden die Werte tabelliert. Auf den ersten Blick könnte man meinen, dass man

für jede Kombination von (μ, σ^2) eine eigene Tabelle benötigt. Im nächsten Abschnitt werden wir aber sehen, dass es genügt nur eine einzige Tabelle (nämlich die für die Standardnormalverteilung) anzufertigen.

Für $X \sim \mathcal{N}(\mu, \sigma^2)$ sind die Kennzahlen wie folgt:

$$\begin{aligned}\mathcal{E}(X) &= \mu, \\ \text{Var}(X) &= \sigma^2, \quad \sigma_X = \sigma.\end{aligned}$$

Das heisst, dass die Parameter μ und σ^2 eine natürliche Interpretation als Erwartungswert und Varianz der Verteilung haben. Drei Normalverteilungen mit verschiedenen Werten von μ und σ sind in Abbildung 4.7 dargestellt.

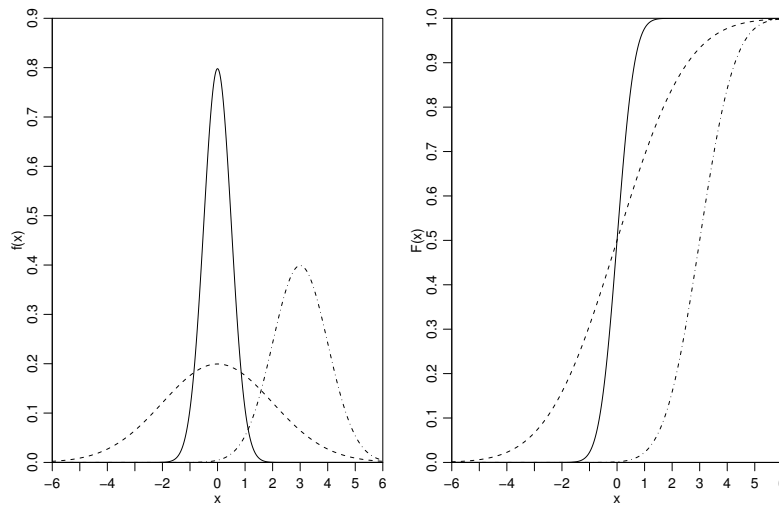


Abbildung 4.7: Dichten (links) und kumulative Verteilungsfunktionen (rechts) der Normalverteilungen mit $\mu = 0, \sigma = 0.5$ (ausgezogen), $\mu = 0, \sigma = 2$ (gestrichelt) und $\mu = 3, \sigma = 1$ (Strich-Punkte).

Die Standard-Normalverteilung

Die Normal-Verteilung mit $\mu = 0$ und $\sigma^2 = 1$ heisst Standard-Normalverteilung. Deren Dichte und kumulative Verteilungsfunktion werden wie folgt bezeichnet:

$$\begin{aligned}\varphi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \\ \Phi(x) &= \int_{-\infty}^x \varphi(y) dy.\end{aligned}$$

Die Werte von $\Phi(\cdot)$ sind in Kapitel 6 tabelliert.

Beispiel: Sei $Z \sim \mathcal{N}(0, 1)$. Mit der Tabelle kann man folgende Fragen beantworten.

- Was ist $P[Z \leq 1.13]$? Wir lesen in der Zeile, in der 1.1 und in der Spalte, in der 0.03 steht: $P[Z \leq 1.13] = \Phi(1.13) = 0.8708$.

- Für welchen Wert von z ist $\Phi(z) = 0.7910$? Anders gefragt: Was ist $\Phi^{-1}(0.7910)$? Wir durchsuchen die Tabelle nach der Zahl 0.7910. Wir finden sie in der Zeile 0.8 und in der Spalte 0.01. Also ist $\Phi^{-1}(0.7910) = 0.81$.
- Was ist $P[Z \leq -0.2]$? In der Tabelle sind keine negativen Werte von z aufgeführt. Weil die Standardnormalverteilung um null herum symmetrisch ist, kann man die Frage trotzdem leicht beantworten: Die Fläche links von -0.2 ist wegen der Symmetrie genau so gross wie die Fläche rechts von 0.2 . D.h., $P[Z \leq -0.2] = P[Z \geq 0.2]$. Die Fläche rechts von 0.2 kann man umschreiben² mit $P[Z \geq 0.2] = 1 - P[Z < 0.2] = 1 - P[Z \leq 0.2]$. In der Tabelle finden wir $P[Z \leq 0.2] = \Phi(0.2) = 0.5793$. Das Ergebnis ist also $P[Z \leq -0.2] = 1 - 0.5793 = 0.4207$.

Wir werden unten sehen, dass eine Normal-Verteilung $\mathcal{N}(\mu, \sigma^2)$ immer in eine Standard-Normalverteilung transformiert werden kann. Deshalb genügen die Werte von $\Phi(\cdot)$, um Wahrscheinlichkeiten und Quantile einer allgemeinen $\mathcal{N}(\mu, \sigma^2)$ -Verteilung zu berechnen und man muss nicht für jede Kombination von (μ, σ^2) eine eigene Tabelle anfertigen.

4.5.4 Funktionen einer Zufallsvariable

Wenn $g : \mathbb{R} \rightarrow \mathbb{R}$ eine Funktion von \mathbb{R} nach \mathbb{R} und X eine Zufallsvariable ist, dann ist die Zusammensetzung

$$Y = g(X),$$

eine neue Zufallsvariable. Die Zusammensetzung bedeutet einfach, dass zu jeder Realisierung x von X die Realisierung $y = g(x)$ von Y gehört. Solche Transformationen treten häufig auf.

Die kumulative Verteilungsfunktion und die Dichte von Y sind im Prinzip durch die Verteilungsfunktion und die Dichte von X bestimmt. Die Berechnung kann aber je nach Eigenschaften von g kompliziert sein. Für den Erwartungswert gilt jedoch stets die folgende Formel

$$\mathcal{E}(Y) = \mathcal{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Im Folgenden werden wir einige wichtige Spezialfälle diskutieren.

Lineare Transformationen

Wir betrachten hier zuerst den Fall einer linearen Transformation

$$g(x) = a + bx \quad (a, b \in \mathbb{R}).$$

²Erinnern Sie sich daran, dass bei einer stetigen Variable X gilt: $P[X = x] = 0$. Daher ist bei einer stetigen Variable $P[X \leq x] = P[X < x]$.

Für $Y = a + bX$ gilt dann

$$\begin{aligned}\mathcal{E}(Y) &= \mathcal{E}(a + bX) = a + b\mathcal{E}(X), \\ \text{Var}(Y) &= \text{Var}(a + bX) = b^2 \text{Var}(X), \quad \sigma_Y = |b|\sigma_X.\end{aligned}\quad (4.2)$$

Die erste Gleichung folgt aus

$$\begin{aligned}\mathcal{E}(Y) &= \mathcal{E}(a + bX) = \int_{-\infty}^{\infty} (a + bx)f_X(x)dx = a \int_{-\infty}^{\infty} f_X(x)dx + b \int_{-\infty}^{\infty} xf_X(x)dx \\ &= a \cdot 1 + b \cdot \mathcal{E}(X) = a + b\mathcal{E}(X).\end{aligned}$$

Die zweite Gleichung folgt aus der Definition der Varianz und aus der ersten Gleichung:

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(a + bX) = \mathcal{E}((a + bX - \mathcal{E}(a + bX))^2) = \\ &= \mathcal{E}((a + bX - (a + b\mathcal{E}(X)))^2) = \mathcal{E}(b^2(X - \mathcal{E}(X))^2) = \\ &= b^2 \mathcal{E}((X - \mathcal{E}(X))^2) = b^2 \text{Var}(X)\end{aligned}$$

Durch Ziehen der Wurzel erhält man für die Standardabweichung $\sigma_Y = b \cdot \sigma_X$. Überdies gelten für $b > 0$:

$$\begin{aligned}\alpha - \text{Quantil von } Y &= q_Y(\alpha) = a + bq_X(\alpha), \\ f_Y(y) &= \frac{1}{b}f_X\left(\frac{y-a}{b}\right).\end{aligned}\quad (4.3)$$

Beispiel: Wir haben eine Temperatur in Grad Celsius gemessen und kennen die Standardabweichung des Messfehlers auf dieser Skala: $\sigma_C = \frac{1}{3}$ Grad Celsius. Für einen Bericht, der im englischsprachigen Raum gelesen werden soll, wollen wir die Temperatur aber nicht in Grad Celsius, sondern in Grad Fahrenheit angeben. Wie gross ist die Standardabweichung σ_F des Messfehlers, wenn wir die Temperatur in Grad Fahrenheit angeben? Die Temperatur T_c in Grad Celsius kann man folgendermassen in die Temperatur T_F in Grad Fahrenheit umrechnen: $T_F = \frac{9}{5} \cdot T_C + 32$. Daher ist die Standardabweichung in Grad Fahrenheit $\sigma_F = \frac{9}{5}\sigma_C = \frac{9}{5} \cdot \frac{1}{3} = \frac{3}{5}$.

Beispiel: Wenn $X \sim \mathcal{N}(\mu, \sigma)$, dann ist $Y = g(X) \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$, denn nach der Regel (4.3) gilt

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma b} \exp\left(-\frac{(y-a-b\mu)^2}{2\sigma^2 b^2}\right).$$

Eine linear transformierte Normalverteilung ist also wiederum eine Normalverteilung. Diese Eigenschaft, dass man mit linearen Transformationen innerhalb der Verteilungs-Klasse bleibt ist eine spezielle Eigenschaft der Normalverteilung und im Allgemeinen nicht richtig.

Standardisieren einer Zufallsvariablen

Wir können X immer linear transformieren, so dass die transformierte Zufallsvariable Erwartungswert = 0 und Varianz = 1 hat. Dies geschieht wie folgt: betrachte die lineare Transformation

$$g(x) = \frac{x - \mathcal{E}(X)}{\sigma_X} = -\frac{\mathcal{E}(X)}{\sigma_X} + \frac{1}{\sigma_X}x = a + bx$$

mit $a = -\frac{\mathcal{E}(X)}{\sigma_X}$ und $b = \frac{1}{\sigma_X}$. Damit bilden wir die transformierte Zufallsvariable

$$Z = g(X) = \frac{X - \mathcal{E}(X)}{\sigma_X}.$$

Mit Hilfe der Regeln in (4.2) gilt dann: $\mathcal{E}(Z) = 0$, $\text{Var}(Z) = 1$.

Falls $X \sim \mathcal{N}(\mu, \sigma^2)$, so ist die standardisierte Zufallsvariable wieder normalverteilt, wie wir im obigen Beispiel gesehen haben, hat nun aber Erwartungswert null und Varianz eins. Man erhält also die Standard-Normalverteilung:

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

Damit lassen sich Wahrscheinlichkeiten für beliebige Normalverteilungen mit Hilfe der Standard-Normalverteilung berechnen. Es reicht daher, wenn man nur die Werte der Standard-Normalverteilung tabelliert.

Beispiel 1: Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ mit $\mu = 2$ und $\sigma^2 = 4$. Berechne $P(X \leq 5)$!

$$P(X \leq 5) = P\left(\frac{X - \mu}{\sigma} \leq \frac{5 - \mu}{\sigma}\right) = P\left(Z \leq \frac{5 - 2}{2}\right) = P(Z \leq 1.5) = \Phi(1.5) = 0.93$$

Beispiel 2: Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ mit $\mu = 2$ und $\sigma^2 = 4$. Berechne $P(X > 5)$!

$$P(X > 5) = 1 - P(X \leq 5) = 1 - \Phi(1.5) = 1 - 0.93 = 0.07$$

Beispiel 3: Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ mit $\mu = 2$ und $\sigma^2 = 4$. Wie gross ist das 90% Quantil γ von X ?

$$\begin{aligned} P(X \leq \gamma) &= 0.9 \\ \Rightarrow P\left(Z \leq \frac{\gamma - \mu}{\sigma}\right) &= 0.9 \\ \Rightarrow \Phi\left(\frac{\gamma - \mu}{\sigma}\right) &= 0.9 \end{aligned}$$

Aus der Tabelle entnehmen wir, dass $\frac{\gamma - \mu}{\sigma} = 1.28$. Auflösen nach γ ergibt:
 $\gamma = \mu + 1.28\sigma = 2 + 1.28 \cdot 2 = 4.56$

Beispiel 4: Sei $X \sim \mathcal{N}(\mu, \sigma^2)$ mit $\mu = 2$ und $\sigma^2 = 4$. Berechne $P(|X| \leq 2)$!

$$\begin{aligned} P(|X| \leq 2) &= P(-2 \leq X \leq 2) = P(X \leq 2) - P(X \leq -2) = \\ &= P\left(Z \leq \frac{2-2}{2}\right) - P\left(Z \leq \frac{-2-2}{2}\right) = P(Z \leq 0) - P(Z \leq -2) = \\ &= \Phi(0) - \Phi(-2) = \Phi(0) - (1 - \Phi(2)) = 0.5 - (1 - 0.97) = \\ &= 0.5 - 0.03 = 0.47. \end{aligned}$$

Die Lognormal-Verteilung

Für positive Zufallsvariablen X wird oft die Logarithmus-Transformation verwendet. Führt diese Transformation auf eine Normalverteilung, dann heisst X **Lognormal**-verteilt. In andern Worten, wenn $Y \sim \mathcal{N}(\mu, \sigma^2)$, dann heisst $X = \exp(Y)$ Lognormal-verteilt mit Parametern $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$. Die Lognormal-Verteilung ist nicht mehr symmetrisch und es gilt: $\mathcal{E}(X) = \exp(\mu + \sigma^2/2) > \exp(\mathcal{E}(Y))$.

4.5.5 Überprüfen der Normalverteilungs-Annahme

Oft wollen wir überprüfen ob eine Verteilung ein brauchbares Modell für einen Datensatz darstellt. Das heisst, wir wollen überprüfen, ob ein Datensatz x_1, \dots, x_n als Realisierungen von einer Zufallsvariablen X mit einer Modell-Verteilung (z.B. mit einer kumulativen Verteilungsfunktion $F(\cdot)$) aufgefasst werden kann.

Im Prinzip kann man das Histogramm der empirischen Daten mit der Dichte der Modell-Verteilung vergleichen. Oft sieht man aber die Abweichungen oder Übereinstimmungen besser, wenn man stattdessen die Quantile benutzt.

Q-Q Plot

Die Idee des Q-Q-Plot (Quantil-Quantil Plot) ist die empirischen Quantile gegen die theoretischen Quantile der Modell-Verteilung zu plotten. Konkret: plote für $\alpha = 0.5/n, 1.5/n, \dots, (n - 0.5)/n$ die theoretischen Quantile der Modell-Verteilung $q(\alpha)$ auf der x-Achse gegen die empirischen Quantile, welche den geordneten Beobachtungen $x_{[1]} < x_{[2]} < \dots < x_{[n]}$ entsprechen, auf der y-Achse. Wenn die Beobachtungen gemäss der Modell-Verteilung erzeugt wurden, sollten diese Punkte ungefähr auf der Winkelhalbierenden $y = x$ liegen.

Normal-Plot

Meist will man nicht eine spezifische Verteilung, sondern eine ganze Klasse von Verteilungen prüfen, also zum Beispiel die Klasse der Normalverteilungen mit beliebigem μ und σ .

Ein Q-Q Plot, bei dem die Modell-Verteilung gleich der Standard-Normalverteilung $\mathcal{N}(0, 1)$ ist heisst Normal-Plot.

Falls die Daten Realisierungen von $X \sim \mathcal{N}(\mu, \sigma^2)$ sind, so gilt für die Quantile von X :

$$q(\alpha) = \mu + \sigma \Phi^{-1}(\alpha),$$

siehe (4.3). Wenn man also einen Normal-Plot macht, so sollten die Punkte im Normal-Plot ungefähr auf der Geraden mit Achsenabschnitt μ und Steigung σ liegen.

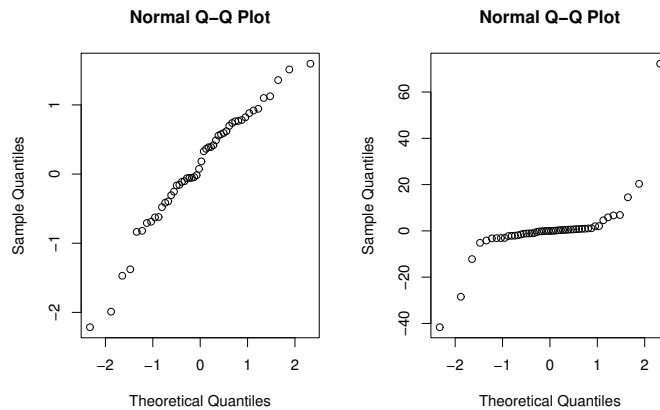


Abbildung 4.8: Links: Normal-Plot für 50 Realisierungen von $\mathcal{N}(0, 1)$. Die Punkte liegen auf einer Geraden. Rechts: Normal-Plot für 50 Realisierungen von Cauchy-Verteilung (sehr langschwänzig). Die Punkte liegen nicht auf einer Geraden.

Abbildung 4.8 zeigt zwei Normal-Plots: einmal ist die Daten-generierende Verteilung eine Normal-Verteilung. Die Punkte liegen in etwa auf einer Geraden. Das andere mal sind die Daten von einer sehr langschwänzigen Verteilung erzeugt worden. Hier liegen die Punkte gar nicht auf einer Geraden.

Natürlich liegen die Punkte in einem Normal-Plot nicht perfekt auf einer Geraden. Damit Sie ein Gefühl dafür bekommen, was eine akzeptable Abweichung von der Geraden ist, sind in Abb. 4.9 die Normal-Plots von neun Datensätzen (je 50 Beobachtungen) gezeigt, die alle von einer Standard-Normalverteilung simuliert wurden. Abweichungen in dem Mass wie sie dort zu sehen sind, kann man also erwarten, wenn die Daten tatsächlich von einer Normalverteilung stammen.

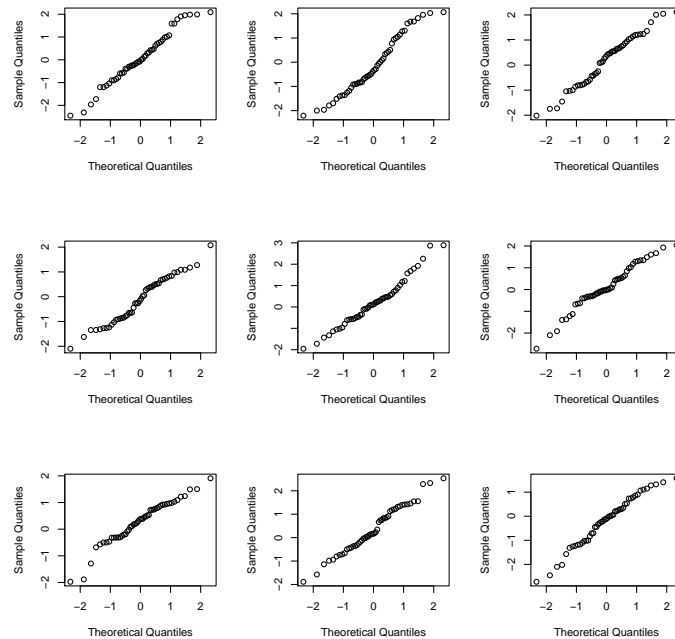


Abbildung 4.9: Normal-Plots von neun Datensätzen (je 50 Beobachtungen), die alle von einer Standard-Normalverteilung simuliert wurden. Abweichungen von einer Geraden in diesem Ausmass sind also zu erwarten, wenn die Daten wirklich normalverteilt sind. Falls die Abweichung von einer Geraden deutlich grösser ist, sind die Daten wohl nicht normalverteilt.

4.6 Funktionen von mehreren Zufallsvariablen

(Stahel, Kap. 6.8 – 6.11)

In Kapitel 4.5.4 haben wir untersucht, wie die Funktion einer Zufallsvariable verteilt ist. In den meisten Anwendungen hat man es nicht mit einer, sondern mit mehreren Zufallsvariablen zu tun. Üblicherweise misst man die gleiche Grösse mehrmals (man hat mehrere Individuen, oder man wiederholt die Messungen). In diesem Kapitel untersuchen wir, wie eine Funktion mehrerer Zufallsvariablen verteilt ist.

Die Messungen x_1, x_2, \dots, x_n fassen wir als Realisierungen der Zufallsvariablen X_1, \dots, X_n auf. Diese Auffassung ist oft bequemer als die Interpretation, dass die Messungen n unabhängige Realisierungen einer Zufallsvariablen X sind. Oft sind wir an Funktionen der Messwerte x_1, x_2, \dots, x_n interessiert: $y = g(x_1, \dots, x_n)$ wobei $g : \mathbb{R}^n \rightarrow \mathbb{R}$. Wenn x_1, x_2, \dots, x_n Realisierungen der Zufallsvariablen X_1, \dots, X_n sind, dann ist y eine Realisierung der Zufallsvaria-

ble

$$Y = g(X_1, \dots, X_n).$$

Wir betrachten hier vor allem die Funktion "arithmetisches Mittel"

$$g(x_1, \dots, x_n) = \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i.$$

Das arithmetische Mittel der Daten \bar{x}_n ist also eine Realisierung der Zufallsvariablen \bar{X}_n .

Wir sind an der Verteilung der Zufallsvariablen \bar{X}_n interessiert: Die Kenntnis dieser Verteilung wird uns erlauben, Statistik aufgrund von arithmetischen Mitteln von Daten zu machen.

4.6.1 Unabhängigkeit und i.i.d. Annahme

Oft treffen wir die Annahme, dass die Zufallsvariablen X_1, \dots, X_n **unabhängig** voneinander sind. Anschaulich heisst das, es gibt keine gemeinsamen Faktoren, die den Ausgang der verschiedenen Messungen beeinflussen, und keine "carry over" Phänomene von einer Messung zur nächsten. Die mathematische Definition lautet: X_1, \dots, X_n sind unabhängig, falls die Ereignisse $\{X_1 \leq b_1\}, \dots, \{X_n \leq b_n\}$ unabhängig sind für beliebige $b_1, \dots, b_n \in \mathbb{R}$.

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

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

Die Abkürzung i.i.d. steht für: independent, identically distributed. Wir werden meistens mit dieser i.i.d. Annahme arbeiten. Welche Verteilung die X_i 's haben, lassen wir offen. Es kann, aber muss nicht eine Normalverteilung sein.

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

$$\mathcal{E}(X_1 + X_2) = \mathcal{E}(X_1) + \mathcal{E}(X_2)$$

gilt immer,

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

jedoch nur, wenn X_1 und X_2 unabhängig sind.

4.6.2 Kennzahlen und Verteilung von \bar{X}_n

Wir nehmen in diesem Abschnitt an, dass

$$X_1, \dots, X_n \text{ i.i.d.} \sim \text{kumulative Verteilungsfkt. } F.$$

Wegen dem zweiten “i” in i.i.d. hat jedes X_i dieselbe Verteilung und dieselben Kennzahlen: $\mathcal{E}(X_i) = \mu$, $\text{Var}(X_i) = \sigma_X^2$.

Die Kennzahlen von \bar{X}_n folgen dann aus den allgemeinen Regeln für Erwartungswert und Varianz von Summen:

$$\begin{aligned}\mathcal{E}(\bar{X}_n) &= \mu, \\ \text{Var}(\bar{X}_n) &= \frac{\sigma_X^2}{n}, \quad \sigma(\bar{X}_n) = \frac{\sigma_X}{\sqrt{n}}.\end{aligned}$$

Die Standardabweichung von \bar{X}_n heisst auch der **Standard-Fehler** des arithmetischen Mittels.

Der Erwartungswert von \bar{X}_n ist also gleich demjenigen einer einzelnen Zufallsvariablen X_i , die *Streuung nimmt jedoch ab mit wachsendem n* . Für $n \rightarrow \infty$ geht die Streuung gegen null. Es gilt das **Gesetz der grossen Zahlen**: Falls X_1, \dots, X_n i.i.d., dann

$$\bar{X}_n \rightarrow \mu \quad (n \rightarrow \infty).$$

Die Streuung des arithmetischen Mittels ist jedoch nicht proportional zu $1/n$, sondern nur zu $1/\sqrt{n}$. Das bedeutet, dass man für eine doppelte Genauigkeit nicht doppelt so viele, sondern vier Mal so viele Messungen braucht. Diesen Sachverhalt nennt man auch **\sqrt{n} -Gesetz**.

Die Verteilung von \bar{X}_n ist im allgemeinen schwierig anzugeben. Ein Spezialfall ist der folgende:

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right) \quad \text{falls } X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu, \sigma_X^2).$$

Falls die einzelnen X_i 's nicht normal-verteilt sind, so gilt erstaunlicherweise die obige Verteilungs-Formel immer noch approximativ. Dies liefert der folgende berühmte Satz. (Manchmal ist man an der Summe vieler Zufallsvariablen $S_n = \sum_{i=1}^n X_i$ interessiert, deshalb erwähne ich die Formel für S_n auch.)

Zentraler Grenzwertsatz: falls X_1, \dots, X_n i.i.d., dann

$$\begin{aligned}\bar{X}_n &\approx \mathcal{N}(\mu, \sigma_X^2/n), \\ S_n &\approx \mathcal{N}(n\mu, n\sigma_X^2).\end{aligned}$$

wobei die Approximation im Allgemeinen besser wird mit grösserem n . Überdies ist auch die Approximation besser, je näher die Verteilung von X_i bei der Normal-Verteilung $\mathcal{N}(\mu, \sigma_X^2)$ ist.

Beispiel: Wir ziehen $n = 10$ Zufallszahlen X_i . Die zehn Zufallsvariablen sind unabhängig und es gilt $X_i \sim \text{Uniform}([0, 1])$ für jedes i . Wie gross ist die Wahrscheinlichkeit, dass die Summe der Zufallszahlen $S_{10} = \sum_{i=1}^{10} X_i$ grösser als Sechs ist? D.h., wir suchen $P[S_{10} > 6]$. Aus Kapitel 4.5.1 wissen wir, wie man Erwartungswert und Varianz von jedem X_i berechnet: $\mathcal{E}(X_i) = 0.5$ und $\text{Var}(X_i) = \frac{1}{12}$. Aus dem Zentralen Grenzwertsatz folgt:

$$S_n \approx \mathcal{N}(n\mathcal{E}(X_i), n\text{Var}(X_i)) = \mathcal{N}\left(5, \frac{10}{12}\right) = \mathcal{N}(5, 0.83)$$

Damit kommen wir zu folgender Lösung:

$$\begin{aligned} P(S_n > 6) &= 1 - P(S_n \leq 6) = 1 - P\left(\frac{S_n - 5}{\sqrt{0.83}} \leq \frac{6 - 5}{\sqrt{0.83}}\right) = 1 - P(Z \leq 1.1) \\ &= 1 - \Phi(1.1) = 1 - 0.86 = 0.14 \end{aligned}$$

Für eine exakte Formulierung des Zentralen Grenzwertsatzes betrachtet man die standardisierte Zufallsvariable

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma_X}.$$

Diese ist ungefähr $\mathcal{N}(0, 1)$ verteilt, was bedeutet, dass für alle x gilt

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x).$$

Der Zentrale Grenzwertsatz gilt auch für diskrete Zufallsvariablen X_i . Im Folgenden leiten wir die Normalapproximation für den Binomialtest her, die wir in Kapitel 3.2 schon mehrfach verwendet haben. Im Fall von n unabhängigen Wiederholungen eines binären Experiments mit den beiden möglichen Ergebnissen "Erfolg", bzw. "Misserfolg", setzen wir $X_i = 1$, falls die i -te Wiederholung ein Erfolg ist, und $X_i = 0$ falls die i -te Wiederholung ein Misserfolg ist. Diese X_i sind i.i.d. Bernoulli(π)-verteilt, wobei π die Wahrscheinlichkeit für einen Erfolg bei einer Durchführung ist (siehe Abschnitt 2.5). Erwartungswert und Varianz von X_i lassen sich somit leicht berechnen: $E(X_i) = \pi$, $\text{Var}(X_i) = \pi(1 - \pi)$. Das arithmetische Mittel \bar{X}_n ist dann nichts anderes als die relative Häufigkeit der Erfolge. Gemäss dem Zentralen Grenzwertsatz gilt $\bar{X}_n \approx \mathcal{N}(\mu, \sigma_X^2/n)$ mit $\mu = \pi$ und $\sigma_X^2 = \pi(1 - \pi)$. Damit können wir die Wahrscheinlichkeit berechnen, dass der Schätzer \bar{X}_n in der Nähe des wahren Parameters π liegt (die Rechnung ist sehr ähnlich wie Beispiel 4 in Kapitel 4.5.4; versuchen Sie es nachzurechnen! Eine Auflösung ist in Kap. 7.1):

$$P(|\bar{X}_n - \pi| > \varepsilon) \approx 2 - 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma_X}\right) \quad (4.4)$$

Somit können wir die **Normalapproximation des Binomialtests** durchführen:

1. **Modell:** X : Anzahl Erfolge bei n Versuchen.
 $X \sim \text{Bin}(n, \pi)$.
2. **Nullhypothese:** $H_0 : \pi = \pi_0$
Alternative: $H_A : \pi \neq \pi_0$
3. **Teststatistik:** \bar{X}_n : Anzahl Erfolge geteilt durch Anzahl Versuche.
Verteilung der Teststatistik unter H_0 : $\bar{X}_n \approx \mathcal{N}(\mu, \sigma_X^2/n)$ mit $\mu = \pi_0$ und $\sigma_X^2 = \pi_0(1 - \pi_0)$
4. **Signifikanzniveau:** α

5. **Verwerfungsbereich:** Wir müssen einen Wert für ε finden, sodass $P(|\bar{X}_n - \pi_0| > \varepsilon) = \alpha$ ist. Mit Gleichung (4.4) erhalten wir:

$$\begin{aligned} P(|\bar{X}_n - \pi_0| > \varepsilon) &= \alpha \\ \Leftrightarrow 2 - 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma_X}\right) &= \alpha \\ \Leftrightarrow \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma_X}\right) &= 1 - \frac{\alpha}{2} \\ \Leftrightarrow \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) &= \frac{\varepsilon\sqrt{n}}{\sigma_X} \\ \Rightarrow \varepsilon &= \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \end{aligned}$$

Der Verwerfungsbereich sind dann alle Werte \bar{x}_n der Teststatistik \bar{X}_n , die in den Bereich

$$|\bar{x}_n - \pi_0| > \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

fallen.

6. **Testentscheid:** Prüfe, ob der beobachtete Wert \bar{x}_n der Teststatistik \bar{X}_n in den Verwerfungsbereich fällt.

Beispiel: Wir haben eine Münze $n = 100$ mal geworfen und 58 mal Kopf beobachtet. Ist es plausibel, dass es sich um eine faire Münze handelt oder ist eine von beiden Seiten bevorzugt? Wir könnten nun einen Binomialtest durchführen; zur Übung, werden wir aber die Normalapproximation des Binomialtests anwenden:

1. **Modell:** X : Anzahl Erfolge bei $n = 100$ Versuchen.
 $X \sim \text{Bin}(n, \pi)$.
2. **Nullhypothese:** $H_0 : \pi = \pi_0 = 0.5$
Alternative: $H_A : \pi \neq \pi_0$
3. **Teststatistik:** \bar{X}_n : Anzahl Erfolge geteilt durch Anzahl Versuche.
Verteilung der Teststatistik unter H_0 : $\bar{X}_n \approx \mathcal{N}(\mu, \sigma_X^2/n)$ mit $\mu = \pi_0 = 0.5$ und $\sigma_X^2 = \pi_0(1 - \pi_0) = 0.5 \cdot 0.5 = 0.25$
4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich:** Aus der Tabelle sehen wir, dass $\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \Phi^{-1}(0.975) = 1.96$. Damit ergibt sich mit $\pi_0 = 0.5$:

$$|\bar{x}_n - \pi_0| > \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 0.05 \cdot 1.96 \approx 0.098$$

Der Verwerfungsbereich ist also:

$$K = (-\infty, \pi_0 - 0.098] \cup [\pi_0 + 0.098, \infty) = (-\infty, 0.402] \cup [0.598, \infty)$$

6. **Testentscheid:** Der beobachtete Wert der Teststatistik ist $\bar{x}_n = \frac{58}{100} = 0.58$. Da dieser Wert nicht im Verwerfungsbereich der Teststatistik $K = (-\infty, 0.402] \cup [0.598, \infty)$ liegt, kann die Nullhypothese auf dem 5% Signifikanzniveau nicht verworfen werden.³.

Sie fragen sich nun vermutlich, warum man eine Normalapproximation des Binomialtests braucht, wenn man doch schon den eigentlichen Binomialtest hat. Ist die Normalapproximation nicht überflüssig? Auf diese Frage gibt es zwei Antworten. Erstens ist die Normalapproximation des Binomialtests nur ein Anwendungsbeispiel des Zentralen Grenzwertsatzes. Man kann die Normalapproximation auf sehr viele andere Tests/Sachverhalte anwenden, bei denen ein Mittelwert oder eine Summe von Zufallszahlen vorkommen. Ich habe den Binomialtest als Beispiel gewählt, weil Sie schon die exakte Version des Tests aus Kapitel 3.2 kennen und somit überprüfen können, dass beide Tests wirklich etwa das gleiche Ergebnis liefern⁴. Es gibt aber noch einen zweiten Grund. Bei dem exakten Binomialtest muss man Wahrscheinlichkeiten der Form $P(X \leq k)$ (das ist die kumulative Verteilungsfunktion) ausrechnen, wobei $X \sim \text{Bin}(n, \pi)$. Leider gibt es für die kumulative Verteilungsfunktion keine einfache Formel, sodass man $P(X \leq k) = \sum_{i=1}^k P(X = i)$ verwendet. Für grosse n kann das schnell zu sehr vielen Summanden führen und die Berechnung wird sehr aufwändig (Versuchen Sie mal von Hand $P(X \leq 250)$ mit $X \sim \text{Bin}(500, 0.02)$ auszurechnen; da sind Sie schnell mal einen Tag lang beschäftigt.). Jetzt können Sie sagen, dass in der Praxis der Computer diese Berechnung ja ausführen kann. Da haben Sie natürlich recht, aber bei grossen n bekommt selbst der Computer Probleme: Erinnern wir uns, dass $P(X = k) = \binom{n}{k} \cdot \pi^k (1 - \pi)^{n-k}$. Wenn n gross und π klein ist, wird der erste Faktor riesig und der zweite Faktor winzig. Das Produkt ergibt wieder eine Zahl zwischen 0 und 1. Weil Computer Zahlen nur bis zu einer gewissen Grösse (etwa 10^{18}) abspeichern können, führt eine naive Berechnung von $P(X = k)$ auch mit dem Computer zu Problemen. Hier wird die Normalapproximation (oder auch die Poissonapproximation) zur guten Lösung. Wer bis hierher gelesen hat, soll auch belohnt werden: In der Prüfung haben wir natürlich keine Zeit, Sie einen exakten Binomialtest mit grossen Zahlen durchführen zu lassen. Entweder, die Zahlen sind klein oder die Zahlen sind gross und wir wollen eine "geeignete Approximation" sehen, was dann die Normalapproximation wäre.

4.6.3 Verletzung der Unabhängigkeit

Was ändert sich, wenn die Zufallsvariablen X_1, \dots, X_n nicht mehr unabhängig sind? Die Annahme der identischen Verteilung behalten wir bei, weil diese Zufallsvariablen die Wiederholung von Messungen unter identischen Bedingungen

³Wir erhalten das gleiche Ergebnis wie mit dem Binomialtest in Kapitel 3.2

⁴Ganz genau gleich wird das Ergebnis natürlich nicht sein, denn schliesslich verwenden wir die Normalverteilung ja nur als *Approximation* der wahren Verteilung der Teststatistik.

modellieren sollen. Dann gilt

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

$$\text{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n} \left(1 + \frac{2}{n} \sum_{1 \leq i < j \leq n} \rho(X_i, X_j) \right).$$

Dabei ist $\rho(X_i, X_j)$ die Korrelation zwischen den Zufallsvariablen X_i und X_j , das theoretische Gegenstück zur empirischen Korrelation in Abschnitt 4.3.2, siehe auch Abschnitt 4.3.1:

$$\rho(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i} \sigma_{X_j}}, \quad \text{Cov}(X_i, X_j) = \mathcal{E}((X_i - \mathcal{E}(X_i))(X_j - \mathcal{E}(X_j))).$$

Dies bedeutet, dass Abhängigkeit zwischen den Messungen die Genauigkeit des arithmetischen Mittels beeinflusst. Bei positiver Korrelation nimmt die Genauigkeit ab, bei negativer Korrelation kann sie auch zunehmen. Es ist daher wichtig, die Messungen so anzulegen, dass sie als unabhängig angenommen werden können, wenn man die in den folgenden Kapiteln beschriebenen statistischen Methoden anwenden will.

4.7 Statistik für eine Stichprobe (Stahel, Kap. 8.3 – 8.5, 9.3)

Wir betrachten Daten x_1, \dots, x_n welche als Realisierungen von X_1, \dots, X_n i.i.d. aufgefasst werden. Zwei Kennzahlen der Zufallsvariablen X_i sind: $\mathcal{E}(X_i) = \mu$ und $\text{Var}(X_i) = \sigma_X^2$. Typischerweise sind diese (und andere) Kennzahlen unbekannt, und man möchte Rückschlüsse aus den Daten darüber machen.

Beispiel: Blutplättchen-Aggregation (siehe Abschnitt 4.2)

Die Blutplättchen-Aggregation ist ein Beispiel eines sogenannten *gepaarten Vergleichs*, wo man bei jeder Versuchseinheit eine Grösse unter zwei verschiedenen Bedingungen misst. Von Interesse ist, ob ein systematischer Unterschied bezüglich der Aggregation vor und nach dem Rauchen einer Zigarette besteht. Um dies zu untersuchen bilden wir die *Differenzen*: $x_i = \text{Aggregation "nachher"} - \text{Aggregation "vorher"}$ ($i = 1, \dots, 11$), und wir haben somit eine (uns interessierende) Stichprobe.

4.7.1 (Punkt-) Schätzungen

Die (Punkt-) Schätzungen für den Erwartungswert und die Varianz sind:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i,$$

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Beachte dass die Schätzer hier als Funktionen der Zufallsvariablen X_1, \dots, X_n geschrieben sind: insbesondere sind $\hat{\mu}$ und $\hat{\sigma}_X^2$ selbst wieder Zufallsvariablen (die Verteilungs-Eigenschaften von $\hat{\mu}$ wurden in Abschnitt 4.6 diskutiert). Mit der Interpretation, dass die Daten x_i Realisierungen der Zufallsvariablen X_i sind, sind die realisierten Schätzer gleich dem arithmetischen Mittel und der empirischen Varianz der Daten.

4.7.2 Tests für μ

Beispiel: Blutplättchen-Aggregation (Forts.)

Wir wollen testen, ob ein systematischer Unterschied zwischen Aggregation “nachher” und Aggregation “vorher” besteht. Da x_i gerade die Differenz der Aggregationen zwischen “nachher” und “vorher” ist, betrachten wir das folgende Test-Problem:

$$H_0 : \mu = 0, \quad H_A : \mu > 0.$$

Wir verwenden einen einseitigen Test, weil es bei der Planung der Studie bereits klar war, dass man nur an einer Erhöhung der Aggregation durch Nikotin interessiert ist.

Um auf den Parameter μ zu testen, machen wir vorerst einmal die Annahme, dass

$$X_1, \dots, X_n \text{ i.i.d. } \mathcal{N}(\mu, \sigma_X^2), \sigma_X \text{ bekannt} \quad (4.5)$$

Eine Abschwächung dieser Annahme wir später diskutiert.

Der z-Test: σ_X bekannt

Wir nehmen an, dass die Daten x_1, \dots, x_n Realisierungen von (4.5) sind. Überdies machen wir die Annahme, dass σ_X^2 bekannt ist.

Der z-Test für den Parameter μ ist dann wie folgt.

1. **Modell:** X_i ist eine kontinuierliche Messgrösse;
 X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2), \sigma_X$ bekannt
2. **Nullhypothese:** $H_0 : \mu = \mu_0$,
Alternative: $H_A : \mu \neq \mu_0$ (oder “<” oder “>”)

3. Teststatistik:

$$Z = \frac{(\bar{X}_n - \mu_0)}{\sigma_{\bar{X}_n}} = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{Standardfehler}}.$$

Verteilung der Teststatistik unter H_0 : $Z \sim \mathcal{N}(0, 1)$

4. Signifikanzniveau: α

5. **Verwerfungsbereich für die Teststatistik:**

$$\begin{aligned} K &= (-\infty, -\Phi^{-1}(1 - \frac{\alpha}{2})] \cup [\Phi^{-1}(1 - \frac{\alpha}{2}), \infty) \text{ bei } H_A : \mu \neq \mu_0, \\ K &= (-\infty, -\Phi^{-1}(1 - \alpha)] \text{ bei } H_A : \mu < \mu_0, \\ K &= [\Phi^{-1}(1 - \alpha), \infty) \text{ bei } H_A : \mu > \mu_0. \end{aligned}$$

6. **Testentscheid:** Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt.

Im Unterschied zu den Tests in Kapitel 3.2.2 basiert der z-Test auf *mehreren* Beobachtungen. Diese werden aber mittels einer realisierten Teststatistik z zusammengefasst (eine Funktion der Daten). Ansonsten sind die Konzepte genau gleich wie in Kapitel 3.2.2.

Beispiel (Forts. Schmelzwärme): Bei Methode A (vgl. Kapitel 4.2) scheint die Schmelzwärme grösser als 80.00 zu sein. Angenommen, wir wissen aus vorhergehenden Studien, dass die Standardabweichung unseres Messinstruments $\sigma_X = 0.01$ ist. Ist es plausibel, dass die Schmelzwärme genau $80.00 \frac{\text{cal}}{\text{g}}$ ist? Wir machen einen z-Test:

1. **Modell:** X_i ist eine kontinuierliche Messgrösse;
 X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2)$, $\sigma_X = 0.01$ bekannt, $n = 13$
2. **Nullhypothese:** $H_0 : \mu = \mu_0 = 80.00$,
Alternative: $H_A : \mu \neq \mu_0$
3. **Teststatistik:**

$$Z = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\sigma_X}$$

Verteilung der Teststatistik unter H_0 : $Z \sim \mathcal{N}(0, 1)$

4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich für die Teststatistik:**

$$K = (-\infty, -\Phi^{-1}(1 - \frac{\alpha}{2})] \cup [\Phi^{-1}(1 - \frac{\alpha}{2}), \infty) \text{ bei } H_A : \mu \neq \mu_0,$$

Aus der Tabelle entnehmen wir mit $\alpha = 0.05$: $\Phi^{-1}(1 - \frac{\alpha}{2}) = \Phi^{-1}(0.975) = 1.96$. Damit ergibt sich für den Verwerfungsbereich der Teststatistik $K = (-\infty, -1.96] \cup [1.96, \infty)$.

6. **Testentscheid:** Aus den $n = 13$ Daten errechnen wir $\bar{X}_n = 80.02$. Damit ergibt sich als Wert für die Teststatistik $z = \frac{\sqrt{13}(80.02 - 80.00)}{0.01} = 7.21$. Der beobachtete Wert liegt im Verwerfungsbereich der Teststatistik. Daher wird die Nullhypothese auf dem 5% Signifikanzniveau verworfen.

Die Begriffe Fehler 1. Art, Fehler 2. Art und Macht, die wir in Kapitel 3.2.2 beim Binomialtest kennengelernt haben, werden auch für alle anderen Tests verwendet. Wir wiederholen hier nochmals die Definitionen:

Fehler 1. Art = fälschliches Verwerfen von H_0 , obschon H_0 richtig ist,

und

Fehler 2. Art(μ) = (fälschliches) Beibehalten von H_0 falls $\mu \in H_A$ richtig ist.

Die Wahrscheinlichkeit für einen Fehler 1. Art ist gerade gleich α ; beim Fehler 2. Art (μ) betrachtet man oft die Macht:

Macht(μ) = $1 - P(\text{Fehler 2. Art } (\mu)) = P(\text{Verwerfen von } H_0 \text{ falls } \mu \in H_A \text{ stimmt}).$

Für $\mu \in H_A$ kann man die Macht(μ) als die Chance interpretieren, dass man richtigerweise H_A entdeckt falls $\mu \in H_A$ stimmt. Für eine Teststatistik T und einen dazugehörigen Verwerfungsbereich K gilt dann:

$$\begin{aligned} P_{\mu_0}(T \in K) &= \alpha, \\ P_{\mu}(T \in K) &= \text{Macht}(\mu). \end{aligned}$$

Der t-Test: σ_X unbekannt

Wie vorhin nehmen wir an, dass die Daten Realisierungen von (4.5) sind. In der Praxis ist die Annahme, dass σ_X bekannt ist, oftmals unrealistisch. In solchen Fällen kann die Teststatistik z nicht berechnet werden, weil sie auf σ_X basiert. Allerdings können wir stattdessen die Schätzung

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

benutzen. Dies führt aber zu einer zusätzlichen Unsicherheit, was zur Folge hat, dass sich die Verteilung der Teststatistik ändert.

Die Teststatistik beim t-Test ist

$$T = \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\hat{\sigma}_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}.$$

Deren Verteilung unter der Nullhypothese $H_0 : \mu = \mu_0$ ist eine sogenannte t-Verteilung mit $n - 1$ Freiheitsgraden, die wir mit t_{n-1} bezeichnen.

Die t_n -Verteilung ist eine symmetrische Verteilung um 0, welche langschwänziger ist als die Standard-Normalverteilung $\mathcal{N}(0, 1)$. Für $T \sim t_n$ gilt:

$$\begin{aligned} \mathcal{E}(T) &= 0 \\ \text{Var}(T) &= \frac{n}{n-2}. \end{aligned}$$

Für grosse n ist t_n ähnlich zu $\mathcal{N}(0, 1)$: insbesondere strebt die t_n -Verteilung gegen die Standard-Normalverteilung $\mathcal{N}(0, 1)$ falls $n \rightarrow \infty$.

Den Verwerfungsbereich beim Test erhalten wir, indem wir einen Bereich der Wahrscheinlichkeit α bei der t_{n-1} -Verteilung abschneiden (je nach Alternative auf einer Seite, oder je die Hälfte auf beiden Seiten). Dazu brauchen wir die Quantile $t_{n,\alpha}$, welche tabelliert sind oder mittels Computer berechnet werden können.

Eine Tabelle von t-Verteilungen finden Sie im Kapitel 6.

Beispiel: Aus der Tabelle kann man folgende Informationen ablesen:

- Sei $T \sim t_{13}$. Für welches t gilt $P[T \leq t] = 0.95$? Wir sehen in der Tabelle in der Zeile 13 und in der Spalte $t_{0.95}$: $t_{13;0.95} = 1.771$
- Sei $T \sim t_7$. Was ist $P[T \leq 2.998]$? Wir suchen in der Zeile 7 den Zahlenwert 2.998 und finden ihn in der Spalte $t_{0.99}$. Daher gilt also $P[T \leq 2.998] = 0.99$.

Zusammenfassend ist der t -Test wie folgt:

1. **Modell:** X_i ist eine kontinuierliche Messgrösse;
 X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2)$, σ_X wird durch $\widehat{\sigma}_X$ geschätzt
2. **Nullhypothese:** $H_0 : \mu = \mu_0$,
Alternative: $H_A : \mu \neq \mu_0$ (oder “<” oder “>”)
3. **Teststatistik:**

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\widehat{\sigma}_X} = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}}.$$

Verteilung der Teststatistik unter H_0 : $T \sim t_{n-1}$

4. **Signifikanzniveau:** α
5. **Verwerfungsbereich für die Teststatistik:**

$$\begin{aligned} K &= (-\infty, -t_{n-1;1-\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty) \text{ bei } H_A : \mu \neq \mu_0, \\ K &= (-\infty, -t_{n-1;1-\alpha}] \text{ bei } H_A : \mu < \mu_0, \\ K &= [t_{n-1;1-\alpha}, \infty) \text{ bei } H_A : \mu > \mu_0. \end{aligned}$$

6. **Testentscheid:** Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt.

Da das Quantil der t -Verteilung grösser ist als das Quantil der Normalverteilung, erhält man einen etwas kleineren Verwerfungsbereich als beim z -Test. Für grosse n ist der Unterschied allerdings minim (da $t_{n-1} \approx \mathcal{N}(0, 1)$ falls n gross).

Beispiel (Forts. Schmelzwärme): Wir berechnen nochmals das Beispiel aus dem z -Test. Diesmal schätzen wir allerdings die Standardabweichung σ_X aus den Daten. Wir machen also einen t -Test auf dem 5% Signifikanzniveau:

1. **Modell:** X_i ist eine kontinuierliche Messgrösse;
 X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2)$, σ_X wird geschätzt
 $\widehat{\sigma}_X = 0.024$
2. **Nullhypothese:** $H_0 : \mu = \mu_0 = 80.00$,
Alternative: $H_A : \mu \neq \mu_0$
3. **Teststatistik:**

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\widehat{\sigma}_X}$$

Verteilung der Teststatistik unter H_0 : $T \sim t_{n-1}$

4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich für die Teststatistik:**

$$K = (-\infty, -t_{n-1; 1-\frac{\alpha}{2}}] \cup [t_{n-1; 1-\frac{\alpha}{2}}, \infty) \text{ bei } H_A : \mu \neq \mu_0,$$

Aus der Tabelle entnehmen wir mit $\alpha = 0.05$ und $n = 13$: $t_{n-1; 1-\frac{\alpha}{2}} = t_{12; 0.975} = 2.179$. Der Verwerfungsbereich der Teststatistik ist also:

$$K = (-\infty, -2.179] \cup [2.179, \infty)$$

6. **Testentscheid:** Aus den $n = 13$ Daten haben wir $\bar{x} = 80.02$ und $\hat{\sigma}_X = 0.024$ errechnet. Damit ergibt sich für die Teststatistik der Wert $t = \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\hat{\sigma}_X} = \frac{\sqrt{13}(80.02 - 80.00)}{0.024} = 3.00$. Der beobachtete Wert der Teststatistik liegt im Verwerfungsbereich. Daher wird die Nullhypothese auf dem 5% Niveau verworfen.

Der P-Wert bei 2-seitiger Alternative $H_A : \mu \neq \mu_0$ kann wie folgt berechnet werden (der beobachtete Wert der Teststatistik ist $t = \frac{\sqrt{n}|\bar{x}_n - \mu_0|}{\hat{\sigma}_X}$):

$$\begin{aligned} P - \text{Wert} &= P(|T| > |t|) = P(T < -|t|) + P(T > |t|) = 2 \cdot P(T > |t|) = \\ &= 2 \cdot (1 - P(T \leq |t|)) = 2 \cdot (1 - F_{t_{n-1}}(|t|)) = 2 \cdot \left(1 - F_{t_{n-1}}\left(\frac{\sqrt{n}|\bar{x}_n - \mu_0|}{\hat{\sigma}_X}\right)\right), \end{aligned}$$

wobei $F_{t_{n-1}}$ die kumulative Verteilungsfunktion der t -Verteilung mit $n - 1$ Freiheitsgraden bezeichnet; also $F_{t_{n-1}}(t) = P(T \leq t)$ wobei $T \sim t_{n-1}$. Die Funktion $F_{t_{n-1}}$ für die t -Verteilung ist das Analogon zur Funktion $\Phi(z)$ bei der Standardnormalverteilung. Sowohl $F_{t_{n-1}}$ also auch $\Phi(z)$ sind am Ende des Skripts (Kap. 6) tabelliert. Die Tabelle für die t -Verteilung ist nicht sehr ausführlich. Sie müssen also schnell mal den Computer verwenden (die Funktionen `dt` für die Wa.dichte, `pt` für die kumulative Verteilungsfunktion und `qt` für das Quantil).

Beispiel (Forts.) Blutplättchen-Aggregation (siehe Abschnitt 4.2)

Wir betrachten die Differenzen $x_i = \text{Aggregation "nachher"} - \text{Aggregation "vorher"}$ ($i = 1, \dots, 11$) und wollen testen, ob der Unterschied grösser als null sein könnte. Da wir die Standardabweichung der Messmethode nicht kennen, machen wir einen t -Test:

1. **Modell:** X_i ist eine kontinuierliche Messgrösse;
 X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma_X^2)$, σ_X wird geschätzt
 $\hat{\sigma}_X = 7.98$

2. **Nullhypothese:** $H_0 : \mu = \mu_0 = 0$,
Alternative: $H_A : \mu > \mu_0$

3. **Teststatistik:**

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\hat{\sigma}_X}$$

Verteilung der Teststatistik unter H_0 : $T \sim t_{n-1}$

4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich für die Teststatistik:**

$$K = [t_{n-1;1-\alpha}, \infty) \text{ bei } H_A : \mu > \mu_0,$$

Aus der Tabelle entnehmen wir mit $\alpha = 0.05$ und $n = 11$: $t_{n-1;1-\alpha} = t_{10;0.95} = 1.81$. Der Verwerfungsbereich der Teststatistik ist also:

$$K = [1.81, \infty)$$

6. **Testentscheid:** Aus den $n = 11$ Daten haben wir $\bar{x} = 10.27$ und $\hat{\sigma}_X = 7.98$ errechnet. Damit ergibt sich für die Teststatistik der Wert $t = \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\hat{\sigma}_X} = \frac{\sqrt{11}(10.27-0)}{7.98} = 4.27$. Der beobachtete Wert der Teststatistik liegt im Verwerfungsbereich. Daher wird die Nullhypothese auf dem 5% Niveau verworfen.

Der P-Wert ist

$$P_{H_0}[T > 4.27] = 1 - F_{t_{10}}(4.27) = 0.00082.$$

Unsere Tabelle im Skript (Kap. 6) ist übrigens nicht genau genug, um $F_{t_{10}}(4.27)$ ablesen zu können. (Als Übung: Versuchen Sie mal $F_{t_{10}}(2.228)$ aus der Tabelle abzulesen. Sie sollten auf den Wert 0.975 kommen). Mit dem Computer geht es ganz leicht: `pt(q = 4.27, df = 10)` liefert $F_{t_{10}} \approx 0.99918$ und somit ist der p-Wert $1 - 0.99918 = 0.00082$. Dies bedeutet, dass der Einfluss von dem Rauchen einer Zigarette bezüglich der Blutplättchen-Aggregation hoch signifikant ist.

Wenn man anstelle der einseitigen die zweiseitige Alternative $H_A : \mu \neq \mu_0 = 0$ spezifiziert, so sind die Resultate wie folgt: das relevante Quantil für $\alpha = 0.05$ ist $t_{10;0.975} = 2.23$. Die Test-Entscheidung bleibt dieselbe: verwerfe H_0 auf dem 5% Signifikanz-Niveau. Der P-Wert ist doppelt so gross: $P_{H_0}[|T| > 4.27] = 2P_{H_0}[T > 4.27] = 0.0016$.

4.7.3 Vertrauensintervall für μ

Analog wie bei Zählraten in Kapitel 3.2.3 besteht das Vertrauensintervall aus denjenigen Werten μ , bei denen der entsprechende Test nicht verwirft. Das sind also alle Parameterwerte des Zufallmodells, bei denen die Daten recht wahrscheinlich oder plausibel sind.

Wir nehmen wiederum an, dass die Daten Realisierungen von (4.5) sind. Bei einem zweiseitigen t-Test hat der Verwerfungsbereich die Form $K = (-\infty, -t_{n-1;1-\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty)$. Der t-Test verwirft H_0 nicht, wenn der Wert der Teststatistik nicht im Verwerfungsbereich der Teststatistik ist. Wenn H_0 nicht verworfen wird, muss also gelten:

$$\begin{aligned} -t_{n-1;1-\frac{\alpha}{2}} &\leq \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\hat{\sigma}_X} \\ &\text{und} \\ t_{n-1;1-\frac{\alpha}{2}} &\geq \frac{\sqrt{n}(\bar{x}_n - \mu_0)}{\hat{\sigma}_X} \end{aligned}$$

Um das zweiseitige Vertrauensintervall von μ zu finden, müssen wir alle Werte von μ_0 finden, die obige Gleichungen erfüllen. Am einfachsten geht das, wenn wir beide Gleichungen nach μ_0 auflösen:

$$\begin{aligned} \mu_0 &\leq \bar{x}_n + \frac{\hat{\sigma}_X \cdot t_{n-1;1-\frac{\alpha}{2}}}{\sqrt{n}} \\ &\text{und} \\ \mu_0 &\geq \bar{x}_n - \frac{\hat{\sigma}_X \cdot t_{n-1;1-\frac{\alpha}{2}}}{\sqrt{n}} \end{aligned}$$

Dies führt dann auf die folgenden zweiseitigen Vertrauensintervalle (die dazugehörigen Tests sind zweiseitig mit Alternative $H_A : \mu \neq \mu_0$) zum Niveau $1 - \alpha$:

$$\left[\bar{x}_n - t_{n-1,1-\alpha/2} \frac{\hat{\sigma}_X}{\sqrt{n}}, \bar{x}_n + t_{n-1,1-\alpha/2} \frac{\hat{\sigma}_X}{\sqrt{n}} \right]$$

Analog kann man auch einseitige Vertrauensintervalle konstruieren. Sie enthalten alle Parameter, bei denen ein einseitiger Test nicht verwerfen würde. Beim t-Test sehen die einseitigen $(1-\alpha)$ -Vertrauensintervalle so aus:

$$\begin{aligned} \text{Falls } H_A : \mu < \mu_0: & (-\infty; \bar{x}_n + t_{n-1,1-\alpha} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}] \\ \text{Falls } H_A : \mu > \mu_0: & [\bar{x}_n - t_{n-1,1-\alpha} \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}; \infty) \end{aligned}$$

Beispiel (Forts.): Aggregation von Blutplättchen

Wir haben 10 Freiheitsgrade und $t_{10,0.975} = 2.23$. Das zweiseitige Konfidenzintervall für die Erhöhung der Blutplättchen-Aggregation nach dem Rauchen

einer Zigarette ist somit (%-ige Zunahme)

$$I = 10.27 \pm 2.23 \cdot 7.9761/\sqrt{11} = [4.91, 15.63].$$

Insbesondere ist die Null nicht im Intervall I : das heisst, der Wert $\mu = 0$ ist nicht mit den Daten kompatibel (was wir bereits vom t-Test (siehe oben) wissen).

4.7.4 Tests für μ bei nicht-normalverteilten Daten

Der z- und t-Test sind optimal falls die Daten Realisierungen von normalverteilten Zufallsvariablen sind wie in (4.5). Optimalität bedeutet hier, dass dies die Tests sind, welche die beste Macht haben.

Wir betrachten hier die allgemeinere Situation, in der die Daten Realisierungen sind von

$$X_1, \dots, X_n \text{ i.i.d. ,} \quad (4.6)$$

wobei X_i eine beliebige Verteilung hat. Wir bezeichnen mit μ einen Lageparameter der Verteilung (z.B. $\mu = \text{Median der Verteilung von } X_i$). Die Nullhypothese ist von der Form $H_0 : \mu = \mu_0$.

Der Vorzeichen-Test

Wir betrachten die Situation, wo die Daten Realisierungen von (4.6) sind, wobei die einzelnen X_i nicht normalverteilt sein müssen. Der Vorzeichentest testet Hypothesen über den Median der Verteilung von X_i , den wir hier mit μ bezeichnen; im Falle einer symmetrischen Verteilung ist $\mu = \mathcal{E}(X_i)$. Wenn μ der Median der Verteilung von X ist, dann ist die Wahrscheinlichkeit, dass eine Realisierung von X grösser als μ ist genauso gross wie die Wahrscheinlichkeit, dass eine Realisierung von X kleiner als μ ist. In anderen Worten: $P(X > \mu) = 0.5$. Der Vorzeichen-Test verwendet das folgendermassen:

1. **Modell:**

$$X_1, \dots, X_n \text{ i.i.d. ,} \quad (4.7)$$

wobei X_i eine beliebige Verteilung hat.

2. **Nullhypothese:** $H_0 : \mu = \mu_0$, (μ ist der Median)

Alternative: $H_A : \mu \neq \mu_0$ (oder einseitige Variante)

3. **Teststatistik:** V : Anzahl X_i s mit $(X_i > \mu_0)$

Verteilung der Teststatistik unter H_0 : $V \sim \text{Bin}(n, \pi_0)$ mit $\pi_0 = 0.5$

4. **Signifikanzniveau:** α

5. **Verwerfungsbereich für die Teststatistik:** $K = [0, c_u] \cup [c_o, n]$ falls $H_A : \mu \neq \mu_0$,

Die Grenzen c_u und c_o müssen mit der Binomialverteilung oder der Normalapproximation berechnet werden.

6. **Testentscheid:** Entscheide, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich der Teststatistik liegt.

Vielleicht ist es Ihnen schon aufgefallen: Der Vorzeichen-Test ist nichts anderes als ein Binomialtest. Wenn wir $\mu_0 = 0$ wählen, entspricht die Teststatistik gerade der Anzahl ”+” im Datensatz, daher der Name ”Vorzeichentest”. Wenn Sie den Binomialtest verstanden haben, müssen Sie für den Vorzeichen-Test also gar nichts neues lernen.

Beispiel (Forts.): Blutplättchen-Aggregation

Die Nullhypothese ist $H_0 : \mu = \mu_0 = 0$. Die realisierte Teststatistik ist dann $v = 10$ und der P-Wert bei einseitiger Alternative $H_A : \mu > \mu_0 = 0$ ist 0.005 (beim t-Test war der P-Wert = 0.00082).

Der Vorzeichentest stimmt immer, falls die Daten Realisierungen von (4.6) sind: das heisst, die Wahrscheinlichkeit für einen Fehler 1. Art ist kontrolliert durch α bei beliebiger Verteilung der X_i ’s.

Vom Standpunkt der Macht gibt es keine eindeutige Antwort, ob der Vorzeichen- oder der t-Test besser ist. Wenn die Verteilung der X_i langschwänzig ist, kann der Vorzeichentest grössere Macht haben. Weil der Vorzeichentest die Information nicht ausnützt, um wieviel die X_i von dem Wert μ_0 abweichen (siehe die Definition der Teststatistik V oben), kann die Macht aber auch wesentlich schlechter sein als beim t-Test.

Der Wilcoxon-Test

Der Wilcoxon-Test ist ein Kompromiss, der keine Normalverteilung voraussetzt wie der t-Test und die Information der Daten besser ausnützt als der Vorzeichen-Test.

Die Voraussetzung für den Wilcoxon-Test ist: Die Daten sind Realisierungen von (4.6) wobei die Verteilung der X_i ’s stetig und symmetrisch ist bezüglich $\mu = \mathcal{E}(X_i)$. Wir verzichten auf die Formel für die Teststatistik und die Berechnung der Verteilung der Teststatistik unter der Nullhypothese $\mu = \mu_0$, da der P-Wert mit statistischer Software berechnet werden kann.

Beispiel (Forts.): Blutplättchen-Aggregation

Die Nullhypothese ist $H_0 : \mu = \mu_0 = 0$. Der P-Wert bei einseitiger Alternative $H_A : \mu > \mu_0 = 0$ ist 0.002528.

Der Wilcoxon-Test ist in den allermeisten Fällen vorzuziehen: er hat in vielen Situationen oftmals wesentlich grössere Macht als der t- und als der Vorzeichen-Test, und selbst in den ungünstigsten Fällen ist er nie viel schlechter.

Wenn man trotzdem den t-Test verwendet, dann sollte man die Daten auch grafisch ansehen, damit wenigstens grobe Abweichungen von der Normalverteilung entdeckt werden. Insbesondere sollte der Normal-Plot (siehe Kap. 4.5.5) angeschaut werden.

4.8 Tests bei zwei Stichproben (Stahel, Kap. 8.8)

Wir besprechen hier Methoden, um einen Vergleich zweier Methoden (Gruppen, Versuchsbedingungen, Behandlungen) hinsichtlich der Lage der Verteilung machen.

4.8.1 Gepaarte Stichprobe

Struktur der Daten

Wenn möglich sollte man eine Versuchseinheit beiden Versuchsbedingungen unterwerfen: Es liegt eine **gepaarte Stichprobe** vor, wenn

- beide Versuchsbedingungen an derselben Versuchseinheit eingesetzt werden
- oder jeder Versuchseinheit aus der einen Gruppe genau eine Versuchseinheit aus der anderen Gruppe zugeordnet werden kann.

Die Daten sind dann von der folgenden Struktur:

$$\begin{aligned}x_1, \dots, x_n &\text{ unter Versuchsbedingung 1,} \\ y_1, \dots, y_n &\text{ unter Versuchsbedingung 2.}\end{aligned}$$

Notwendigerweise ist dann die Stichprobengrösse n für beide Versuchsbedingungen dieselbe. Zudem sind x_i und y_i abhängig, weil die Werte von der gleichen Versuchseinheit kommen.

Beispiel:

Wir testen den Muskelzuwachs durch ein Krafttraining. Dazu messen wir die Kraft von 10 Testpersonen zu Beginn des Trainings. Anschliessend durchlaufen alle Testpersonen ein 6-wöchiges Trainingsprogramm. Dann wird die Kraft erneut gemessen. Für jede Testperson gibt es also zwei Messungen: Vorher und nachher, die Zuordnung ist eindeutig. Somit handelt es sich um gepaarte Stichproben.

Beispiel:

Die Wirksamkeit von Augentropfen zur Reduktion des Augeninnendrucks soll untersucht werden. Wir haben 12 Patienten. Bei jedem Patienten wählen wir zufällig ein Auge aus. In dieses Auge kommen die Augentropfen mit dem Wirkstoff. In das andere Auge kommen Tropfen ohne Wirkstoff (Placebo). Für jede Testperson haben wir also zwei Messungen: Eine für das rechte und eine für das linke Auge; die Zuordnung ist eindeutig. Somit handelt es sich um gepaarte Stichproben.

Beispiel:

Wir haben eine Gruppe von 15 eineiigen Zwillingen, die sich für eine Studie für

ein Haarwuchsmittel gemeldet haben. Bei jedem Zwillingspaar wird eine Person zufällig ausgewählt und erhält das Medikament. Die andere Person des Zwillingspaares erhält ein Placebo. Nach drei Wochen misst man den Haarwuchs. Zu jeder Person aus der Gruppe mit Haarwuchsmittel kann man eindeutig eine Person aus der Gruppe ohne Haarwuchsmittel zuordnen. Somit handelt es sich um gepaarte Stichproben.

Beispiel:

Datensatz zu Blutplättchen-Aggregation, siehe Kapitel 4.2. Die Aggregation wurde für jede Person vor und nach dem Rauchen gemessen. Die Zuordnung aller Messungen in der Gruppe "vorher" zu den Messungen in der Gruppe "nachher" ist also eindeutig. Somit handelt es sich um gepaarte Stichproben.

Test

Bei der Analyse von gepaarten Vergleichen arbeitet man mit den Differenzen innerhalb der Paare,

$$u_i = x_i - y_i \quad (i = 1, \dots, n),$$

welche wir als Realisierungen von i.i.d. Zufallsvariablen U_1, \dots, U_n auffassen. Kein Unterschied zwischen den beiden Versuchsbedingungen heisst dann einfach $E[U_i] = 0$ (oder auch $\text{Median}(U_i) = 0$, je nach Test). Tests dafür sind in Kapitel 4.7 beschrieben: Falls die Daten normalverteilt sind, eignet sich ein t-Test. Sonst kommt ein Vorzeichentest oder ein Wilcoxon-Test in Frage. Dabei ist zu beachten, dass die vorausgesetzte Symmetrie für die Verteilung von U_i beim Wilcoxon-Test immer gilt unter der Nullhypothese, dass X_i und Y_i dieselbe Verteilung haben.

4.8.2 Ungepaarte Stichproben

Oft ist es nicht möglich, jeder Behandlungseinheit aus der einen Gruppe eine Behandlungseinheit aus der zweiten Gruppe eindeutig zuzuordnen. In diesem Fall ist eine Paarung nicht möglich und man spricht von einer ungepaarten Stichprobe. Auch hier muss die Zuordnung zur Behandlungsgruppe durch das Los erfolgen um systematische Fehler zu vermeiden. (vgl. Abschnitt 4.9 unten).

Struktur der Daten

Bei ungepaarten Stichproben hat man Daten x_1, \dots, x_n und y_1, \dots, y_m (siehe Kapitel 4.8.2), welche wir als Realisierungen der folgenden Zufallsvariablen auffassen:

$$\begin{aligned} X_1, \dots, X_n & \text{ i.i.d. ,} \\ Y_1, \dots, Y_m & \text{ i.i.d. ,} \end{aligned} \tag{4.8}$$

wobei auch alle X_i 's von allen Y_j 's unabhängig sind.

Bei einer solchen zufälligen Zuordnung von Versuchseinheiten zu einer von zwei verschiedenen Versuchsbedingungen spricht man von einer ungepaarten Stichprobe. Im Allgemeinen ist in einer ungepaarten Stichprobe $m \neq n$, aber nicht notwendigerweise. Entscheidend ist, dass x_i und y_i zu verschiedenen Versuchseinheiten gehören und als unabhängig angenommen werden können.

Beispiel:

Datensatz zu latenter Schmelzwärme von Eis in Kapitel 4.2. Wir haben die Schmelzwärme mit zwei verschiedenen Methoden hintereinander gemessen. Jede Messung ist entweder mit Methode A oder mit Methode B, aber nicht mit beiden gleichzeitig gemacht worden. Es gibt also keinen eindeutigen Zusammenhang zwischen den Messungen der Methode A und den Messungen der Methode B. Daher sind die beiden Stichproben ungepaart.

Beispiel:

Zufällige Zuordnung von 100 Testpatienten zu einer Gruppe der Grösse 50 mit Medikamenten-Behandlung und zu einer anderen Gruppe der Grösse 50 mit Placebo-Behandlung. Es gibt keine eindeutige Zuordnung von einem Patienten aus der Medikamenten-Gruppe zu einem Patienten in der Placebo-Gruppe. Daher handelt es sich um ungepaarte Stichproben, obwohl beide Gruppen gleich gross sind.

Test: Zwei-Stichproben t-Test bei gleichen Varianzen

Die beiden Stichproben können gleiche oder unterschiedliche Varianz haben. Wir behandeln nur den Fall mit gleicher Varianz im Detail und erwähnen den Fall mit ungleicher Varianz nur kurz.

Im Detail sieht der Zwei-Stichproben t-Test folgendermassen aus:

1. Modell:

$$\begin{aligned} X_1, \dots, X_n \text{ i.i.d.} &\sim \mathcal{N}(\mu_X, \sigma^2), \\ Y_1, \dots, Y_m \text{ i.i.d.} &\sim \mathcal{N}(\mu_Y, \sigma^2). \end{aligned} \quad (4.9)$$

2. Nullhypothese:

$$H_0 : \mu_X = \mu_Y.$$

Alternative:

$$\begin{aligned} H_A : \mu_X &\neq \mu_Y \text{ (zweiseitig)} \\ \text{oder } H_A : \mu_X &> \mu_Y \text{ (einseitig)} \\ \text{oder } H_A : \mu_X &< \mu_Y \text{ (einseitig)} \end{aligned}$$

3. Teststatistik:

$$T = \frac{\bar{X}_n - \bar{Y}_m}{S_{pool} \sqrt{1/n + 1/m}}$$

wobei

$$\begin{aligned} S_{pool}^2 &= \frac{1}{n+m-2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \right) = \\ &= \frac{1}{n+m-2} \left((n-1)\hat{\sigma}_x^2 + (m-1)\hat{\sigma}_y^2 \right). \end{aligned}$$

Verteilung der Teststatistik unter H_0 : $T \sim t_{n+m-2}$.

4. Signifikanzniveau: α

5. Verwerfungsbereich für die Teststatistik:

$$\begin{aligned} (-\infty, -t_{n+m-2, 1-\alpha/2}] \cup [t_{n+m-2, 1-\alpha/2}, \infty) & \text{ bei Alternative } H_A : \mu_X \neq \mu_Y, \\ [t_{n+m-2, 1-\alpha}, \infty) & \text{ bei Alternative } H_A : \mu_X > \mu_Y, \\ (-\infty, -t_{n+m-2, 1-\alpha}] & \text{ bei Alternative } H_A : \mu_X < \mu_Y. \end{aligned}$$

6. Testentscheid: Entscheide, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich der Teststatistik liegt.

Die Idee des Zwei-Stichproben t-Tests ist wie folgt. Man ersetzt die unbekannte Differenz $\mu_X - \mu_Y$ durch die Schätzung $\bar{X}_n - \bar{Y}_m$ und beurteilt, ob diese Schätzung "nahe bei" 0 liegt ("weit weg von" 0 würde Evidenz für H_A bedeuten). Dies wird so quantifiziert, dass man durch den geschätzten Standardfehler von $\bar{X}_n - \bar{Y}_m$ dividiert und dies als Teststatistik benutzt:

$$\begin{aligned} T &= \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\widehat{\text{Var}}(\bar{X}_n - \bar{Y}_m)}} \\ &= \frac{\bar{X}_n - \bar{Y}_m}{S_{pool} \sqrt{1/n + 1/m}}. \end{aligned}$$

Unter der Annahme (4.9) und der Null-Hypothese $\mu_X = \mu_Y$ gilt dann:

$$T \sim t_{n+m-2}.$$

Die Wahl des Nenners in der Teststatistik T ergibt sich aus

$$\text{Var}(\bar{X}_n - \bar{Y}_m) = \sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right). \quad (4.10)$$

Beweis von (4.10):

1. \bar{X}_n und \bar{Y}_m sind unabhängig, weil alle X_i 's von allen Y_j 's unabhängig sind.
2. Wegen der Unabhängigkeit von \bar{X}_n und \bar{Y}_m gilt:

$$\text{Var}(\bar{X}_n - \bar{Y}_m) = \text{Var}(\bar{X}_n) + \text{Var}(-\bar{Y}_m) = \text{Var}(\bar{X}_n) + \text{Var}(\bar{Y}_m).$$

$$3. \text{Var}(\bar{X}_n) = \sigma^2/n \text{ und } \text{Var}(\bar{Y}_m) = \sigma^2/m.$$

$$\text{Somit ist mit Schritt 2: } \text{Var}(\bar{X}_n - \bar{Y}_m) = \sigma^2(1/n + 1/m).$$

□

Beispiel: Schmelzwärme von Eis, siehe Kapitel 4.2.

1. **Modell:** X : Mit Methode A gemessene Schmelzwärme in cal/g.
 Y : Mit Methode B gemessene Schmelzwärme in cal/g.

$$X_1, \dots, X_n \text{ i.i.d. } \sim \mathcal{N}(\mu_X, \sigma^2), \quad n = 13$$

$$Y_1, \dots, Y_m \text{ i.i.d. } \sim \mathcal{N}(\mu_Y, \sigma^2), \quad m = 8.$$

2. **Nullhypothese:**

$$H_0 : \mu_X = \mu_Y.$$

Alternative:

$$H_A : \mu_X \neq \mu_Y \text{ (zweiseitig)}$$

3. **Teststatistik:**

$$T = \frac{\bar{X}_n - \bar{Y}_m}{S_{pool} \sqrt{1/n + 1/m}}$$

wobei

$$S_{pool}^2 = \frac{1}{n + m - 2} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \right).$$

Verteilung der Teststatistik unter H_0 : $T \sim t_{n+m-2}$.

4. **Signifikanzniveau:** $\alpha = 0.05$
5. **Verwerfungsbereich für die Teststatistik:**

$$(-\infty, -t_{n+m-2, 1-\alpha/2}] \cup [t_{n+m-2, 1-\alpha/2}, \infty) \quad \text{bei Alternative } H_A : \mu_X \neq \mu_Y,$$

6. **Testentscheid:** Zunächst berechnen wir den beobachteten Wert der Teststatistik. Für die Mittelwerte ergibt sich $\bar{x} = 80.021$, $\bar{y} = 79.979$. Für die Schätzung der Varianz ergibt sich: $s_{pool}^2 = 7.253 \cdot 10^{-4}$. Damit ist der beobachtete Wert der Teststatistik:

$$t = \frac{\bar{x} - \bar{y}}{s_{pool} \sqrt{1/n + 1/m}} = \frac{80.021 - 79.979}{\sqrt{7.253 \cdot 10^{-4}} \cdot \sqrt{1/13 + 1/8}} = 3.47.$$

Nun berechnen wir den konkreten Wert des Verwerfungsbereichs der Teststatistik. Aus der Tabelle entnehmen wir $t_{n+m-2, 1-\alpha/2} = t_{19, 0.975} = 2.093$. Daher ist der Verwerfungsbereich der Teststatistik:

$$(-\infty, -2.093] \cup [2.093, \infty)$$

Der beobachtete Wert der Teststatistik liegt also im Verwerfungsbereich der Teststatistik. Daher wird die Nullhypothese auf dem 5% Niveau verworfen.

4.8.3 Weitere Zwei-Stichproben-Tests bei ungepaarten Stichproben

Zwei-Stichproben t-Test bei ungleichen Varianzen

Anstelle der Annahme in (4.9) gelte:

$$\begin{aligned} X_1, \dots, X_n \text{ i.i.d.} &\sim \mathcal{N}(\mu_X, \sigma_X^2), \\ Y_1, \dots, Y_m \text{ i.i.d.} &\sim \mathcal{N}(\mu_Y, \sigma_Y^2). \end{aligned}$$

Die Verallgemeinerung des Zwei-Stichproben t-Tests für ungleiche Varianzen $\sigma_X^2 \neq \sigma_Y^2$ ist in der Literatur zu finden und in vielen statistischen Programmen implementiert. In den meisten Fällen erhält man ähnliche P-Werte wie unter der Annahme von gleichen Varianzen.

Zwei-Stichproben Wilcoxon-Test (Mann-Whitney Test)

Die Voraussetzungen für den Zwei-Stichproben Wilcoxon-Test, manchmal auch Mann-Whitney Test genannt, bezüglich (4.8) sind wie folgt:

$$\begin{aligned} X_1, \dots, X_n \text{ i.i.d.} &\sim F_X, \\ Y_1, \dots, Y_m \text{ i.i.d.} &\sim F_Y, \\ F_X &\text{ beliebige stetige Verteilungsfunktion, } F_Y(x) = F_X(x - \delta). \end{aligned}$$

Dies bedeutet, dass die Verteilung von Y_j die um δ verschobene Verteilung von X_i ist, denn: $P(Y_j \leq x + \delta) = F_Y(x + \delta) = F_X(x + \delta - \delta) = F_X(x) = P(X_i \leq x)$.

Die Berechnung des P-Werts eines Zwei-Stichproben Wilcoxon-Tests kann mittels Computer erfolgen. Aus den gleichen Gründen wie im Fall einer Stichprobe (siehe Kapitel 4.7.4) ist der Wilcoxon-Test im Allgemeinen dem t-Test vorzuziehen.

4.9 Versuchsplanung (Stahel, Kap. 14.1 - 14.2)

Genauso wichtig wie die Auswertung der Daten sind Überlegungen, wie man die Daten gewinnen soll. Bisher haben wir Vergleiche zwischen zwei "Behandlungen" besprochen (gepaart oder ungepaart). Allgemeiner geht es bei statistischen

Studien meist darum, wie sich eine oder mehrere Einflussgrößen auf eine Zielgrösse auswirken. Die statistischen Methoden dafür werden wir im nächsten Kapitel noch kurz behandeln.

Zunächst muss man unterscheiden zwischen **Beobachtungsstudien** und **Experimenten**. Bei Beobachtungsstudien werden die Grössen von Interesse bei den **Beobachtungseinheiten** (Patienten, Tiere, Standorte, etc.) passiv erfasst oder gemessen. Bei einem Experiment hingegen werden die erklärenden Variablen vom Experimentator für jede **Versuchseinheit** aktiv festgelegt. Dazu zählen auch Studien, in denen die Versuchseinheiten (Patienten) verschiedenen Behandlungen ausgesetzt werden, sofern der Forscher die Reihenfolge der Behandlungen frei wählen kann. Experimente sind prinzipiell vorzuziehen, da sie Schlüsse auf Ursache-Wirkungs-Beziehungen zulassen.

Will man den Effekt einer Behandlung untersuchen, so braucht es eine **Kontrollgruppe** in der gleichen Studie, die sich zu Beginn der Studie möglichst wenig von der Gruppe mit der neuen Behandlung unterscheidet. Man darf also nicht Versuchseinheiten aus früheren Studien nehmen und sie als die eine Gruppe in einem Zwei-Stichproben-Vergleich verwenden. Es gibt auch oft effizientere Methoden als der Vergleich von zwei unabhängigen Stichproben für den Nachweis eines Behandlungseffekts. So ist meistens eine Versuchsanordnung, bei der sowohl die Kontrollbehandlung und die neue Behandlung auf die gleiche Versuchseinheit angewendet wird (gepaarter Vergleich) effizienter. Weitere solche Möglichkeiten sind in der Literatur unter Versuchsplanung (design of experiments) beschreiben.

Wie soll man im Zwei-Gruppen-Vergleich die Zuordnung der Versuchseinheiten zu den beiden Gruppen vornehmen ?, bzw. im gepaarten Fall: In welcher Reihenfolge soll man die beiden Behandlungen durchführen ? Um eine bewusste oder unbewusste systematische Bevorzugung der einen Gruppe zu vermeiden, soll die Zuordnung zufällig erfolgen (sogenannte **Randomisierung**). Zufällig heisst dabei nicht willkürlich, sondern mit Hilfe von Zufallszahlen oder einem physikalischen Zufallsmechanismus.

Bei human-medizinischen Studien ist es ausserdem wichtig, dass das Experiment wenn möglich **doppelblind** durchgeführt wird. Das heisst, dass weder die Person, welche die Behandlung durchführt oder deren Erfolg beurteilt, noch die Versuchsperson die Gruppenzugehörigkeit kennen. Dies ist wichtig, um den Effekt von Voreingenommenheit bei der Beurteilung auszuschalten. Weiter soll die empfangene Behandlung von allen Patienten gleich wahrgenommen werden – man gibt in der Kontrollgruppe also z.B. eine Placebo-Pille oder injiziert eine Kochsalzlösung. Damit garantiert man, dass ein allfälliger Unterschied zwischen den Gruppen wirklich auf die spezifische Behandlungsart zurückzuführen ist, und nicht etwa auf die erhöhte Aufmerksamkeit, die der Patient erfährt.

Nicht immer ist ein randomisiertes, doppelblindes Experiment möglich (aus ethischen oder praktischen Gründen), und man muss auf Beobachtungsstudien zurückgreifen. Dies erschwert die Auswertung und Interpretation unter Umständen gewaltig, weil man Störeffekte praktisch nicht ausschliessen kann. Ein bekanntes Beispiel ist der Zusammenhang zwischen Rauchen und Lungenkrebs,

der lange umstritten war, weil die genetische Veranlagung und der Effekt des Lebensstils nicht auszuschliessen waren.

Mehr Wiederholungen reduzieren die Unsicherheit. Aus Kosten- und Zeitgründen will man jedoch möglichst wenige Wiederholungen machen. Die Statistik kann berechnen, wie viele Wiederholungen nötig sind, um einen Behandlungseffekt von vorgegebener Grösse mit einer vorgegebenen Wahrscheinlichkeit zu entdecken, sofern auch die Streuung bei gleicher Behandlung bekannt ist oder wenigstens abgeschätzt werden kann.

Eine Warnung vor sogenannten **Scheinwiederholungen** ist an dieser Stelle angebracht. Eine echte Wiederholung ist die kleinste Einheit, bei welcher eine Behandlung unabhängig von anderen Einheiten angewendet werden kann. Untersucht man z.B. den Effekt von einem Schadstoff im Wasser auf Fische, dann sind mehrere Fische im gleichen Aquarium Scheinwiederholungen, denn man kann bei Fischen im gleichen Aquarium die Exposition nicht individuell verändern. Ein allfälliger signifikanter Unterschied zwischen den Fischen in zwei verschieden behandelten Aquarien könnte auch durch irgendwelche unbeabsichtigte andere Unterschiede zwischen ihnen oder durch eine gegenseitige Beeinflussung der Fische im gleichen Aquarium verursacht sein. Man muss also pro Behandlung einige Aquarien haben, oder den Versuch auf geeignete Weise mehrmals wiederholen, damit man statistisch korrekt Unterschiede zwischen Behandlungen nachweisen kann. Scheinwiederholungen gibt es natürlich auch bei Beobachtungsstudien. Auch dort muss man sich fragen, was die Beobachtungseinheit ist, z.B. ein Individuum oder eine Gruppe.

4.10 Software

Praktisch alle Methoden, die in diesem Kapitel vorgestellt wurden, stehen in der Statistik-Software R (und auch in den meisten anderen üblichen Softwarepaketen) zur Verfügung. Sehen Sie sich mal die Hilfefiles der folgenden Funktionen an (wenn die Funktion z.B. `mean` heisst, dann können Sie mit `?mean` das zugehörige Hilfefile aufrufen).

4.10.1 Verschiedenes

Empirischer Mittelwert, empirische Varianz und empirische Standardabweichung lassen sich mit den Befehlen `mean`, `var` und `sd` berechnen. Quantile lassen sich mit der Funktion `quantile` berechnen. Für Histogramme, Boxplots, die empirische kumulative Verteilungsfunktion und Normal-Plots verwenden Sie die Funktionen `hist`, `boxplot`, `ecdf` und `qqnorm`.

Die Uniforme Verteilung, Exponentialverteilung, t-Verteilung und die Normalverteilung stehen unter den Kürzeln (siehe Kapitel 2.8) `unif`, `exp`, `t` und `norm` zur Verfügung.

4.10.2 Zwei-Stichproben t-Test für ungepaarte Stichproben

Sowohl der Ein-Stichproben t-Test als auch der Zwei-Stichproben t-Test (sowohl mit gleicher als auch mit ungleicher Varianz) sind unter `t.test` implementiert. Der Ein-Stichproben und Zwei-Stichproben Wilcoxon-Test ist unter `wilcox.test` verfügbar.

Beispiel Forts. Schmelzwärme: Wir berechnen den Zwei-Stichproben t-Test für ungepaarte Stichproben. Zunächst lesen wir die Daten ein:

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
        80.03, 80.02, 80.00, 80.02)
> y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)
```

Nun führen wir den t-Test mit durch. Wir nehmen zunächst an, dass die Streuung bei beiden Messmethoden gleich ist:

```
t.test(x, y, alternative = "two.sided", mu = 0, paired = FALSE,
       var.equal = TRUE, conf.level = 0.95)
```

Die ersten beiden Argumente enthalten die Daten der beiden Stichproben. Das Argument `alternative` gibt an, ob die Alternative einseitig (und wenn ja in welche Richtung mit `alternative = "greater"` und `alternative = "less"`) oder zweiseitig (mit `alternative = "two.sided"`) ist. Das Argument `mu` gibt an, welcher Unterschied in den Mittelwerten der beiden Gruppen in der Nullhypothese getestet werden soll. Wenn man testen will, ob die beiden Gruppenmittelwerte gleich sind, ist `mu=0` die richtige Wahl. `paired = FALSE` gibt an, dass es sich um zwei ungepaarte Stichproben handelt. `var.equal = TRUE` gibt an, dass die Streuungen in den beiden Stichproben gleich gross sind. Mit `conf.level = 0.95` wird ein 95%-Vertrauensintervall des Unterschieds zwischen den beiden Gruppenmittelwerten ausgegeben.

Dieses Beispiel haben wir in Kapitel 4.8.2 besprochen. Am besten überzeugen Sie sich selbst davon, dass die Software tatsächlich das gleiche Resultat wie die Rechnung von Hand liefert. Die Ausgabe des Computers sieht folgendermassen aus:

Two Sample t-test

```
data:  x and y
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

In der Zeile `t=...` steht zunächst der beobachtete Wert der Teststatistik: $t = 3.47$. Unter der Nullhypothese folgt die Teststatistik einer t-Verteilung mit $df = 19$ Freiheitsgraden. Das ergibt bei einer zweiseitigen Alternative (siehe Zeile `alternative hypothesis: ...`) einen P-Wert von 0.002551. Der Unterschied ist also auf dem auf dem 5% Signifikanzniveau signifikant, weil der P-Wert kleiner als 5% ist. Der Computer berechnet auch das 95%-Vertrauensintervall des Unterschieds in den Gruppenmittelwerten: Mit 95% Wahrscheinlichkeit ist der Gruppenmittelwert von `x` um eine Zahl im Bereich $[0.0167, 0.0673]$ grösser als der Gruppenmittelwert von `y` ⁵. In der letzten Zeile werden schliesslich noch die Mittelwerte der beiden Gruppen angegeben. Beachten Sie, dass kein Verwerfungsbereich ausgegeben wird.

4.10.3 Zwei-Stichproben t-Test für gepaarte Stichproben

Einen t-Test für gepaarte Stichproben kann man leicht durchführen, indem man das Argument `paired = TRUE` verwendet.

Beispiel Fortsetzung Blutplättchen Vergleichen Sie die Ergebnisse des Computers mit den Berechnungen, die wir in 4.7.2 von Hand durchgeführt haben. Sie sollten identisch sein. Zunächst lesen wir wieder die Daten ein:

```
> vorher <- c(25,25,27,44,30,67,53,53,52,60,28)
> nachher <- c(27,29,37,56,46,82,57,80,61,59,43)
```

Dann führen wir den t-Test für gepaarte Stichproben durch:

```
t.test(nachher, vorher, alternative = "two.sided", mu = 0, paired = TRUE,
       conf.level = 0.95)
```

Die Interpretation der Argumente ist wie im vorhergehenden Beispiel. Der Output ist (Interpretation ist ähnlich wie im vorhergehenden Beispiel):

Paired t-test

```
data:  nachher and vorher
t = 4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.91431 15.63114
sample estimates:
mean of the differences
      10.27273
```

Der Unterschied der Gruppenmittelwerte hat bei einer zweiseitigen Alternative (siehe Zeile `alternative hypothesis: ...`) einen P-Wert von 0.0016 und

⁵Die null ist nicht enthalten, also ist der Unterschied der Mittelwerte signifikant

ist somit auf dem 5% Signifikanzniveau signifikant. Der Wert der Teststatistik ist 4.27 und folgt unter der Nullhypothese einer t-Verteilung mit $df = 10$ Freiheitsgraden. Der Unterschied *nachher-vorher*⁶ ist 10.27. Ein 95%-Vertrauensintervall für diese Differenz ist: [4.91, 15.63].

4.10.4 t-Test für eine Stichprobe

Der t-Test für nur eine Stichprobe lässt sich leicht berechnen, indem man das zweite Argument im Funktionsaufruf einfach weglässt.

Beispiel (Forts. Schmelzwärme) Wir testen wie in Kapitel 4.7.2, ob die Beobachtungen in Gruppe A mit der Nullhypothese $H_0 : \mu = 80.00$ verträglich ist. Zunächst wieder die Dateneingabe:

```
> x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05,
        80.03, 80.02, 80.00, 80.02)
```

Und nun der t-Test für eine Stichprobe:

```
t.test(x, alternative = "two.sided", mu = 80.00, conf.level = 0.95)
```

Der Computer liefert:

One Sample t-test

```
data: x
t = 3.1246, df = 12, p-value = 0.008779
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 80.00629 80.03525
sample estimates:
mean of x
 80.02077
```

Der beobachtete Wert der Teststatistik ist 3.12 und folgt unter der Nullhypothese einer t-Verteilung mit $df = 12$ Freiheitsgraden. Der P-Wert mit einer zweiseitigen Alternative ist 0.008779 und ist somit auf dem 5% Signifikanzniveau signifikant. Der beobachtete Mittelwert der Daten ist 80.02. Ein 95%-Vertrauensintervall für den wahren Mittelwert der Messungen ist [80.006, 80.035].

⁶Allgemein: Erstes Argument minus zweites Argument im Funktionsaufruf

Kapitel 5

Regression

5.1 Lernziele

- Sie verstehen die Konzepte der einfachen und multiplen linearen Regression (wie ist die Form des Modells; wie kommen die zufälligen Fehler ins Modell; wie interpretiert man die Koeffizienten; wie schätzt man die Koeffizienten; wie testet man, ob ein Koeffizient signifikant ist)
- Sie können eine lineare Regression mit der Software R durchführen.
- Sie können prüfen, ob die Modellannahmen erfüllt sind.
- Sie verstehen jede Zeile im R-Output einer linearen Regression und können mit dem Output die konkrete Form des geschätzten Modells aufschreiben.

5.2 Einfache lineare Regression

Wir erklären das Modell der einfachen linearen Regression zunächst mit einem fiktiven Beispiel. Je dicker ein Roman (Hardcover) ist, desto teurer ist er in der Regel. Es gibt also einen Zusammenhang zwischen Seitenzahl x und Buchpreis y . Wir gehen in einen Buchladen und suchen zehn Romane verschiedener Dicke aus. Wir nehmen dabei je ein Buch mit der Seitenzahl 50, 100, 150, ..., 450, 500. Von jedem Buch notieren wir die Seitenzahl und den Buchpreis. Damit erhalten wir Tabelle 5.1. :

Aus der Tabelle sehen wir tatsächlich, dass dickere Bücher tendenziell mehr kosten. Abbildung 5.1(a) zeigt diesen Zusammenhang graphisch. Wenn wir einen formelmässigen Zusammenhang zwischen Buchpreis und Seitenzahl hätten, könnten wir Vorhersagen für Bücher mit Seitenzahlen, die wir nicht beobachtet haben, machen. Oder wir könnten herausfinden, wie teuer ein Buch mit "null" Seiten wäre (das wären die Grundkosten des Verlags, die unabhängig von der Seitenzahl sind: Einband, administrativer Aufwand für jedes Buch, etc.). Wie könnten wir diesen Zusammenhang mit einer Formel beschreiben? Auf den

	Seitenzahl	Buchpreis (SFr)
Buch 1	50	6.4
Buch 2	100	9.5
Buch 3	150	15.6
Buch 4	200	15.1
Buch 5	250	17.8
Buch 6	300	23.4
Buch 7	350	23.4
Buch 8	400	22.5
Buch 9	450	26.1
Buch 10	500	29.1

Tabelle 5.1: Zusammenhang zwischen Buchpreis und Seitenzahl (fiktiv).

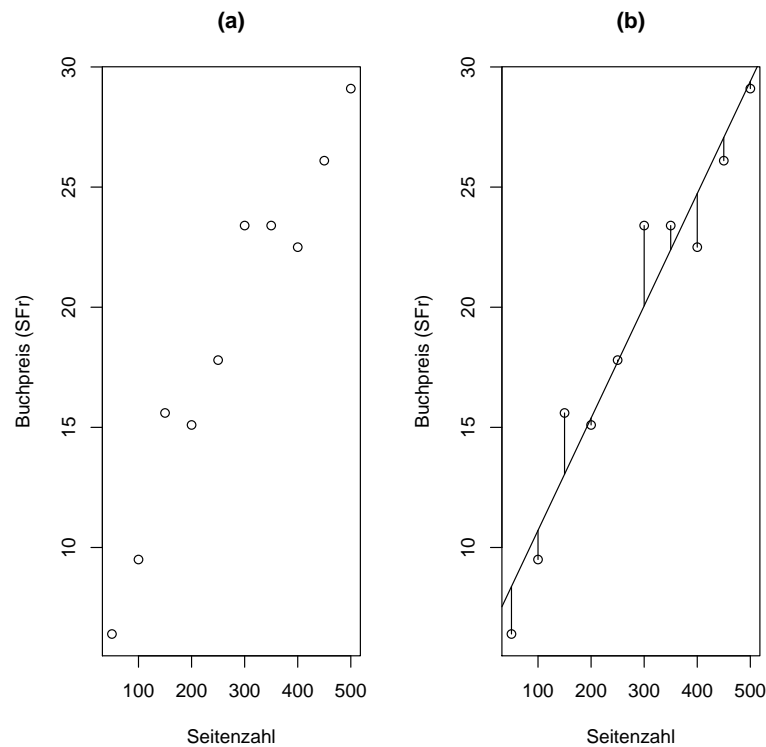


Abbildung 5.1: Zusammenhang zwischen Buchpreis und Seitenzahl (fiktiv).

ersten Blick scheint eine Gerade recht gut zu den Daten zu passen. Diese Gerade hätte die Form: $y = \beta_0 + \beta_1 x$, wobei y der Buchpreis und x die Seitenzahl sind. β_0 wären dann die Grundkosten des Verlags und β_1 wären die Kosten pro Seite. Versuchen Sie mit einem Lineal eine Gerade durch alle Punkte in Abb. 5.1(a) zu legen. Sie werden feststellen, dass das nicht möglich ist. Die Punkte folgen also nur ungefähr einer Geraden. Wie könnten wir eine Gerade finden, die möglichst gut zu allen Punkten passt? Hier gibt es verschiedene Möglichkei-

ten. Wir könnten die vertikalen Abstände zwischen Beobachtung und Gerade (siehe Abb. 5.1(b)) zusammenzählen und davon ausgehen, dass eine kleine Summe der Abstände eine gute Anpassung bedeutet. Diese Methode hat aber eine gravierende Schwäche: Wenn die Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegen, ist die Summe der Abstände etwa null. Dabei passt die Gerade gar nicht gut zu den Datenpunkten. Die positiven Abweichungen haben sich nur mit den negativen Abweichungen ausgelöscht. Wir müssen also das Vorzeichen der Abweichungen eliminieren, bevor wir zusammenzählen. Eine Möglichkeit besteht darin, den Absolutbetrag der Abweichungen aufzusummieren. Eine andere Möglichkeit besteht darin, die Quadrate der Abweichungen aufzusummieren. Letztere Methode hat sich durchgesetzt, weil man mit ihr viel leichter rechnen kann, als mit den Absolutbeträgen. Eine Gerade passt (nach unserem Gütekriterium) also dann am besten zu Punkten, wenn die Quadratsumme der vertikalen Abweichungen minimal ist. Dieses Vorgehen ist unter dem Namen **Methode der kleinsten Quadrate** bekannt. In unserem Fall errechnet der Computer die Werte $\beta_0 = 6.04$ und $\beta_1 = 0.047$. Die Grundkosten des Verlags sind also rund 6 SFr. Pro Seite verlangt der Verlag rund 5 Rappen.

Man könnte die Abweichungen von der Geraden auch auf eine andere Art erklären. Das Modell: $y = \beta_0 + \beta_1 x + E_i$, wobei $E \sim \mathcal{N}(0, \sigma^2)$ i.i.d. beschreibt auch eine Gerade zwischen x und y . Allerdings kann jeder Wert von y noch um den Wert E um die Gerade streuen. In diesem Modell nehmen wir an, dass die Streuung um die wahre Gerade für jedes Buch einer Normalverteilung mit Mittelwert null und Varianz σ^2 folgt. Nun könnten wir mit der Maximum-Likelihood Methode die optimalen Werte für β_0 , β_1 und σ finden. Verblüffenderweise stellt sich heraus, dass die Werte von β_0 und β_1 genau gleich sind wie die Werte, die wir mit der Methode der kleinsten Quadrate erhalten haben. Beide Methoden liefern also die gleichen Ergebnisse.

5.2.1 Das Modell der einfachen linearen Regression

In diesem Kapitel geben wir eine detaillierte Beschreibung vom Modell der einfachen linearen Regression:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad (i = 1, \dots, n),$$

$$E_1, \dots, E_n \text{ i.i.d. }, \mathcal{E}(E_i) = 0, \text{Var}(\mathcal{E}_i) = \sigma^2.$$

Die Y -Variable ist die **Zielvariable** (engl: response variable) und die x -Variable ist die **erklärende Variable** oder **Co-Variable** (engl: explanatory variable; predictor variable; covariate).

Wir nehmen an, dass für die erklärende Variable x gewisse Werte x_1, \dots, x_n vorgegeben sind. In der Regel ist die erklärende Variable nicht zufällig, sondern wird von den Experimentatoren bestimmt. Deshalb notieren wir sie mit einem Kleinbuchstaben. Anschliessend messen wir die Zielgrösse zu jedem Wert der erklärenden Variable. Diese hängt von x , aber auch von einem zufälligen Fehler E_i ab. Im Folgenden werden wir immer davon ausgehen, dass $E_i \sim \mathcal{N}(0, \sigma^2)$.

Die Zufallsvariablen E_i werden öfters als Fehler-Variablen oder Rausch-Terme bezeichnet. Sie besagen, dass der Zusammenhang zwischen der erklärenden und der Ziel-Variablen nicht exakt ist. Während die erklärenden Variablen x_i ($i = 1, \dots, n$) deterministisch sind, sind die Ziel-Variablen Y_i Zufallsvariablen, weil sie von der Zufallsvariablen E_i abhängen. Deshalb notieren wir Y_i mit einem Grossbuchstaben.

Das Modell ist in Abbildung 5.2 illustriert, wo für die Fehler-Variabel eine $\mathcal{N}(0, 0.1^2)$ -Verteilung spezifiziert wurde.

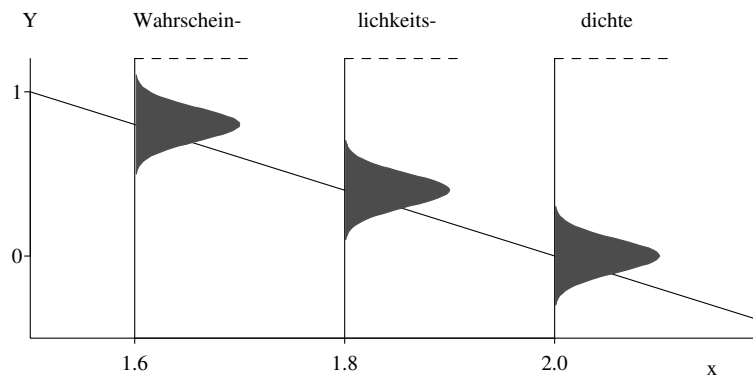


Abbildung 5.2: Veranschaulichung des Regressionsmodells $Y_i = 4 - 2x_i + E_i$ mit $E_i \sim \mathcal{N}(0, 0.1^2)$ für drei Beobachtungen.

Man nennt dieses Modell "einfach", weil nur eine x -Variable vorkommt. Später werden wir die "multiple" Regression kennenlernen, bei der mehrere x -Variablen vorkommen (z.B. Seitenzahl, Genre, Soft-/Hardcover, ...). Man nennt das Modell "linear", weil die Parameter β linear vorkommen. D.h., wenn man nach einem β ableitet, dann verschwindet es in der Ableitung. Die erklärenden Variablen können durchaus auf eine komplizierte, nichtlineare Art in dem Modell auftreten.

Beispiel: Bei den folgenden Modellen handelt es sich um lineare Modelle ($E_i \sim \mathcal{N}(0, \sigma^2)$):

- $y_i = \beta_0 + \beta_1 x_i + E_i$
- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$
- $\log(y_i) = \beta_0 + \beta_1 x_i + E_i$
- $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \beta_2 \sin(x_i^2) + E_i$

Beispiel: Bei den folgenden Modellen handelt es sich **nicht** um lineare Modelle. Prüfen Sie dies nach, indem Sie nach jedem Parameter ableiten. Wenn der Parameter noch in der Ableitung vorkommt, ist das Modell nichtlinear.

- $y_i = \beta_0 \exp(\beta_1 x_i) + E_i$ (β_1 verschwindet beim Ableiten nicht)

- $y_i = \log(\beta_0 + \beta_1 x_i + E_i)$ (weder β_0 noch β_1 verschwinden beim Ableiten)

Manchmal kann man ein nichtlineares Modell durch geschickte Transformation linearisieren.

Beispiel: Wie oben gesehen, ist das Modell $y_i = \log(\beta_0 + \beta_1 x_i + E_i)$ nichtlinear, weil weder β_0 noch β_1 beim Ableiten verschwinden. Allerdings können wir auf beide Seiten der Gleichung die Funktion \exp anwenden. Wir erhalten dann $\exp(y_i) = \beta_0 + \beta_1 x_i + E_i$. Nun ist das Modell linear. Manche Modelle sind allerdings nicht linearisierbar. $y_i = \beta_0 \exp(\beta_1 x_i) + E_i$ ist z.B. nicht linearisierbar.

Beachten Sie, dass alle besprochenen Modelle einen *additiven* Fehler ("+" E_i) hatten.

5.2.2 Parameterschätzungen

Die unbekannten Modell-Parameter in der einfachen linearen Regression sind β_0 , β_1 und auch die Fehlervarianz σ^2 . Die Methode der Kleinsten-Quadrate liefert die folgenden Schätzungen:

$$\hat{\beta}_0, \hat{\beta}_1 \text{ sind Minimierer von } \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

Die Lösung dieses Optimierungsproblem ergibt:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \\ \hat{\beta}_0 &= \bar{y}_n - \hat{\beta}_1 \bar{x}_n. \end{aligned}$$

Das Prinzip der Kleinsten-Quadrate liefert sogenannt erwartungstreue Schätzungen:

$$\mathcal{E}(\hat{\beta}_0) = \beta_0, \quad \mathcal{E}(\hat{\beta}_1) = \beta_1.$$

Das heisst, dass die Schätzungen keinen systematischen Fehler haben, sondern um die wahren Parameter herum streuen (wäre z.B. $\mathcal{E}(\hat{\beta}_1) > \beta_1$, dann wäre $\hat{\beta}_1$ systematisch zu gross). Man kann auch die Standardabweichungen der Schätzungen, die sogenannten Standardfehler berechnen. Von Interesse ist insbesondere das Resultat für $\hat{\beta}_1$:

$$s.e.(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}.$$

Für die Schätzung von σ^2 benützen wir das Konzept der Residuen. Falls wir Realisationen der Fehler-Terme E_i beobachten könnten, so könnten wir die empirische Varianzschätzung für σ^2 verwenden. Hier approximieren wir zuerst die unbeobachteten Fehler-Variablen E_i durch die **Residuen**:

$$R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad (i = 1, \dots, n).$$

Da $E_i = Y_i - (\beta_0 + \beta_1 x_i)$ scheint die Approximation $R_i \approx E_i$ vernünftig. Als Varianzschätzung benutzt man dann:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2. \quad (5.1)$$

Dabei ist zu beachten, dass (bei einfacher linearer Regression mit einem Achsenabschnitt β_0) gilt: $\sum_{i=1}^n R_i = 0$. Das heisst, die Varianzschätzung in (5.1) ist wie die empirische Varianz bei einer Stichprobe (siehe Kapitel 4.7.1), ausser dass wir den Faktor $1/(n-2)$ anstelle von $1/(n-1)$ nehmen. Dieser Faktor entspricht der folgenden Faustregel: $1/(n - \text{Anzahl Parameter})$, wobei die Anzahl Parameter ohne den zu schätzenden Varianz-Parameter zu zählen ist (in unserem Falle sind dies die Parameter β_0, β_1).

Bei einem Datensatz mit realisierten y_i ($i = 1, \dots, n$) werden die Schätzungen mit den Werten y_i anstelle von Y_i gerechnet. Zum Beispiel sind dann die realisierten Residuen von der Form $r_i = y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)$.

5.2.3 Tests und Konfidenzintervalle

Wir diskutieren hier die 2. und 3. Grundfragestellung (siehe Kapitel 3.1) im Kontext der einfachen linearen Regression. Dabei werden wir entscheidend mehr Schlussfolgerungen ziehen, als bloss eine best passende Regressionsgerade zu finden.

Der t-Test in der Regression

Wir betrachten hier als Beispiel den folgenden Datensatz. Es wurden $n = 111$ Messungen gemacht von mittlerer täglicher Temperatur (x -Variable) und mittlerem täglichem Ozongehalt (Y -Variable). Die Daten und die Regressionsgerade $\hat{\beta}_0 + \hat{\beta}_1 x$ sind in Abbildung 5.3 ersichtlich.

Die interessierende Frage in der Praxis lautet: Hat die Temperatur einen Einfluss auf den Ozongehalt? Für diese Fragestellung hat man einen statistischen Test entwickelt. Es stellt sich heraus, dass dieser Test mit dem t-Test in Kapitel 4.7.2 fast völlig identisch ist. Er wird daher auch t-Test genannt.

1. Modell:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad E_1, \dots, E_n \text{ i.i.d. } \mathcal{N}(0, \sigma^2). \quad (5.2)$$

2. Nullhypothese: $H_0 : \beta_1 = 0$

Alternative: $H_A : \beta_1 \neq 0$ (Es wird hier üblicherweise ein zwei-seitiger Test durchgeführt)

3. Teststatistik:

$$T = \frac{\text{beobachtet} - \text{erwartet}}{\text{geschätzter Standardfehler}} = \frac{\hat{\beta}_1 - 0}{\widehat{\text{s.e.}}(\hat{\beta}_1)}.$$

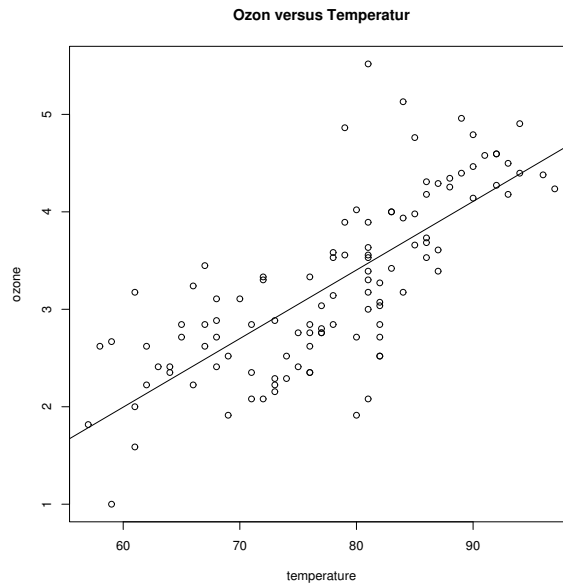


Abbildung 5.3: Streudiagramm und angepasste Regressionsgerade für den Ozon-Temperatur Datensatz.

Dabei ist der geschätzte Standardfehler

$$\widehat{\text{s.e.}}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}.$$

Verteilung der Teststatistik unter H_0 : $T \sim t_{n-2}$

4. **Signifikanzniveau:** α

5. **Verwerfungsbereich für die Teststatistik:**

$$K = (-\infty, -t_{n-2; 1-\frac{\alpha}{2}}] \cup [t_{n-2; 1-\frac{\alpha}{2}}, \infty)$$

6. **Testentscheid:** Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt.

Völlig analog erhält man auch einen Test für $H_0 : \beta_0 = 0$ bei zweiseitiger Alternative $H_A : \beta_0 \neq 0$.

P-Werte

Der P-Wert dieses zweiseitigen t-Test kann dann analog wie in Kapitel 4.7.2 berechnet werden (mit $n - 2$ anstelle von $n - 1$ Freiheitsgraden). Er wird auch von statistischer Software geliefert, sodass man ihn in der Praxis nicht von Hand ausrechnen muss.

Konfidenzintervalle

Basierend auf der Normalverteilungsannahme erhält man die folgenden zweiseitigen Konfidenzintervalle für β_i ($i = 0, 1$) zum Niveau $1 - \alpha$:

$$[\hat{\beta}_i - \widehat{\text{s.e.}}(\hat{\beta}_i)t_{n-2;1-\alpha/2}; \hat{\beta}_i + \widehat{\text{s.e.}}(\hat{\beta}_i)t_{n-2;1-\alpha/2}]$$

Für grosse Werte von n ist die Normalverteilung eine ausgezeichnete Approximation der t-Verteilung. Daher kann man $t_{n-2;1-\alpha/2}$ mit $\Phi^{-1}(1 - \alpha/2)$ approximieren. Sie sollten sich merken, dass $\Phi^{-1}(1 - 0.05/2) = \Phi^{-1}(0.975) \approx 2$. D.h., eine gute Faustregel für ein 95%-Vertrauensintervall ist (analog für β_1):

$$[\hat{\beta}_0 - 2 \cdot \widehat{\text{s.e.}}(\hat{\beta}_0); \hat{\beta}_0 + 2 \cdot \widehat{\text{s.e.}}(\hat{\beta}_0)]$$

Beispiel: Der Computer hat β_0 und β_1 geschätzt und liefert $\hat{\beta}_1 = 1.2$ und $\widehat{\text{s.e.}}(\hat{\beta}_1) = 0.3$. Diese Parameter wurden mit einem Datensatz berechnet, der $n = 10$ Daten hatte. Was ist ein 95%-Vertrauensintervall für den wahren Wert von β_1 ? Die Normalapproximation liefert das approximative Vertrauensintervall: $[1.2 - 2 \cdot 0.3; 1.2 + 2 \cdot 0.3] = [0.6; 1.8]$. Das genaue Vertrauensintervall erhält man unter Verwendung der t-Verteilung ($t_{n-2;0.975} = t_{8;0.975} = 2.31$): $[1.2 - 2.31 \cdot 0.3; 1.2 + 2.31 \cdot 0.3] = [0.51; 1.89]$. Wir sehen, dass das exakte Vertrauensintervall etwas grösser als das approximative Vertrauensintervall ist. Dieses Verhalten tritt in diesem Zusammenhang immer auf und ist umso deutlicher, je kleiner n ist.

5.2.4 Das Bestimmtheitsmass R^2

Die Güte eines Regressionsmodells kann mit dem sogenannten Bestimmtheitsmass R^2 quantifiziert werden. Dazu betrachten wir eine Beziehungen zwischen verschiedenen Variations-Quellen. Wenn wir mit $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ den Wert der angepassten Geraden beim Wert x_i bezeichnen, dann gilt

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SS_Y} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SS_E} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SS_R}. \quad (5.3)$$

(Das Besondere an dieser Gleichung ist, dass das Doppelprodukt $2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ gleich Null ist.) Dabei beschreibt SS_Y die totale Variation der Zielvariablen (ohne Einfluss der erklärenden Variablen x), SS_E die Variation des Fehlers (Residuen-Quadratsumme), und SS_R die Variation, welche durch die Regression erklärt wird (Einfluss der erklärenden Variablen x). Das Bestimmtheitsmass ist dann wie folgt definiert:

$$R^2 = \frac{SS_R}{SS_Y}, \quad (5.4)$$

und beschreibt den Anteil der totalen Variation, welche durch die Regression erklärt wird. Wegen (5.3) gilt, dass $0 \leq R^2 \leq 1$: falls R^2 nahe bei 1 ist, so

erklärt das Regressionsmodell viel der totalen Variation und ist somit gut; falls $R^2 \approx 0$ taugt das Regressionsmodell nicht besonders viel.

Das Bestimmtheitsmass entspricht der quadrierten Korrelation zwischen den beobachteten Werten der Zielgrösse und den geschätzten Werten der Zielgrösse.

$$R^2 = \hat{\rho}_{Y\hat{Y}}^2,$$

5.2.5 Allgemeines Vorgehen bei einfacher linearer Regression

Grob zusammengefasst kann bei einfacher linearer Regression folgendemassen vorgegangen werden.

1. Anpassen der Regressionsgeraden; d.h. Berechnung der Punktschätzer $\hat{\beta}_0, \hat{\beta}_1$.
2. Testen ob erklärende Variable x einen Einfluss auf die Zielvariable Y hat mittels t-Test für $H_0 : \beta_1 = 0$ und $H_a : \beta_1 \neq 0$. Falls dieser Test nicht-signifikantes Ergebnis liefert, so hat die erklärende Variable keinen signifikanten Einfluss auf die Zielvariable.
3. Testen ob Regression durch Nullpunkt geht mittels t-Test für $H_0 : \beta_0 = 0$ und $H_A : \beta_0 \neq 0$. Falls dieser Test nicht-signifikantes Ergebnis liefert, so benützt man das kleinere Modell mit Regression durch Nullpunkt.
4. Bei Interesse Angabe von Konfidenzintervallen für β_0 und β_1 .
5. Angabe des Bestimmtheitsmass R^2 . Dies ist in gewissem Sinne eine informellere (und zusätzliche) Quantifizierung als der statistische Test in Punkt 2.
6. Überprüfen der Modell-Voraussetzungen mittels Residuenanalyse. Dieser wichtige Schritt wird ausführlicher in Kapitel 5.2.6 beschrieben.

5.2.6 Residuenanalyse

Wir werden hier graphische Methoden beschreiben, basierend auf realisierten Residuen $r_i (i = 1, \dots, n)$, welche zur Überprüfung der Modell-Voraussetzungen für die einfache lineare Regression eingesetzt werden können. Die Modell-Voraussetzungen sind die folgenden; in Klammern stehen die Methoden, mit denen man die Annahmen prüfen kann:

1. $\mathcal{E}(E_i) = 0$.
Somit gilt $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$, das heisst: es gibt keinen systematischen Fehler im Modell.
Abweichungen von dieser Annahme könnten zum Beispiel durch einen komplizierteren Zusammenhang zwischen x und Y (z.B. $y_i = \beta_1 x_i^2 + E_i$) verursacht sein. (Prüfe im Tukey-Anscombe Plot)

2. E_1, \dots, E_n i.i.d.

i.i.d. steht für "independent and identically distributed".

Da die Fehler voneinander unabhängig sein müssen, dürfen sie z.B. keine "serielle Korrelationen" aufweisen; d.h., $\text{Corr}(E_i, E_j) = 0$ für $i \neq j$. Am einfachsten sieht man serielle Korrelationen, wenn man die Beobachtungen in ihrer Messreihenfolge aufzeichnet. (Prüfe in einem Plot, in dem die Beobachtungen in der Messreihenfolge aufgezeichnet werden.)

Da die Fehler gleich verteilt sein müssen, muss insbesondere auch die Varianz der Fehler gleich sein. (Prüfe im Tukey-Anscombe Plot)

3. E_1, \dots, E_n i.i.d. $\mathcal{N}(0, \sigma^2)$.

Die Fehler sind normalverteilt. Abweichungen könnte durch eine langschwänzige Fehlerverteilung verursacht sein. (Prüfe in einem Normalplot.)

Der Übergang zwischen Verletzen und Einhalten der Modellvoraussetzungen ist fließend. Wenn eine Modellannahme nur leicht verletzt ist, sind die Ergebnisse sicher noch brauchbar. Je deutlicher die Verletzung erkennbar ist, desto skeptischer sollte man allerdings gegenüber den geschätzten Werten (v.a. P-Werte und Vertrauensintervalle) sein.

Der Tukey-Anscombe Plot

Der wichtigste Plot in der Residuenanalyse ist der Plot der Residuen r_i gegen die angepassten Werte \hat{y}_i , der sogenannte Tukey-Anscombe Plot.

Im Idealfall: gleichmässige Streuung der Punkte um Null.

Abweichungen:

- kegelförmiges Anwachsen der Streuung mit \hat{y}_i

evtl. kann man die Zielvariable logarithmieren (falls Y_i 's positiv sind), d.h. man benutzt das neue Modell

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- Ausreisserpunkte

evtl. können robuste Regressions-Verfahren verwendet werden (siehe Literatur)

- unregelmässige Struktur

Indikation für nichtlinearen Zusammenhang

evtl. Ziel und/oder erklärende Variablen transformieren

Für den Ozon-Datensatz ist der Tukey-Anscombe Plot in Abbildung 5.6 gezeigt.

Nichtlineare Zusammenhänge können in der Praxis natürlich vorkommen: sie zeigen an, dass die Regressionsfunktion nicht korrekt ist. Abhilfe schaffen die Aufnahme zusätzlicher erklärender Variablen (z.B. quadratische Terme, siehe Kapitel 5.3.1) oder - wie bereits oben angedeutet - Transformationen der erklärenden und/oder der Ziel-Variablen. Ein einfaches Beispiel ist in Abbildung 5.4 gezeigt, bei dem es um den Zusammenhang zwischen Tiefe und Fließgeschwindigkeit von Bächen geht. Bei einfacher Regression zeigt der Tukey-Anscombe

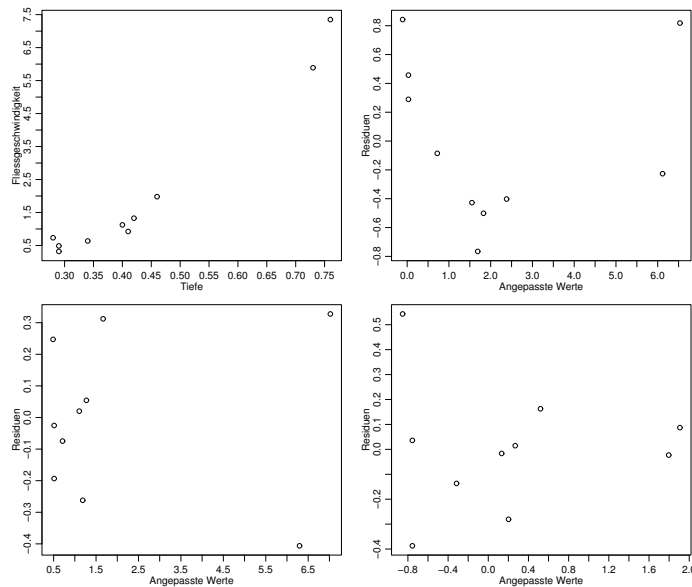


Abbildung 5.4: Streudiagramm von Tiefe und Fließgeschwindigkeit (oben links), Tukey-Anscombe Plots für einfache lineare Regression (oben rechts), für lineare Regression mit Fließgeschwindigkeit und Quadrat der Fließgeschwindigkeit als erklärende Variable (unten links) und für einfache lineare Regression mit logarithmierten Variablen $\log(Y)$ und $\log(x)$ (unten rechts).

Plot eine klare nichtlineare Struktur, die verschwindet, wenn man entweder einen quadratischen Term dazunimmt (siehe Kapitel 5.3.1) oder wenn man beide Variablen logarithmiert (d.h. einen Potenzzusammenhang anpasst mit dem Modell

$$\log(Y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i \quad (i = 1, \dots, n).$$

Mit so wenigen Daten kann man zwischen diesen beiden Modellen nicht unterscheiden. Die Nichtlinearität des Zusammenhangs ist natürlich auch im ursprünglichen Streudiagramm ersichtlich, wenn man genau hinschaut. Häufig sind aber Abweichungen von der Linearität im Tukey-Anscombe Plot besser zu sehen.

Damit Sie ein Gefühl für den Tukey-Anscombe Plot entwickeln können, sind in Abb. 5.5 neun Tukey-Anscombe Plots dargestellt. In der Bildunterschrift wird erklärt, wie sie zu interpretieren sind.

Plot bezüglich serieller Korrelation

Um die Unabhängigkeitsannahme der E_1, \dots, E_n zu überprüfen, kann der folgende Plot gemacht werden: plote r_i gegen die Beobachtungsnummer i .

Im Idealfall: gleichmässige Streuung der Punkte um Null.

Abweichungen:

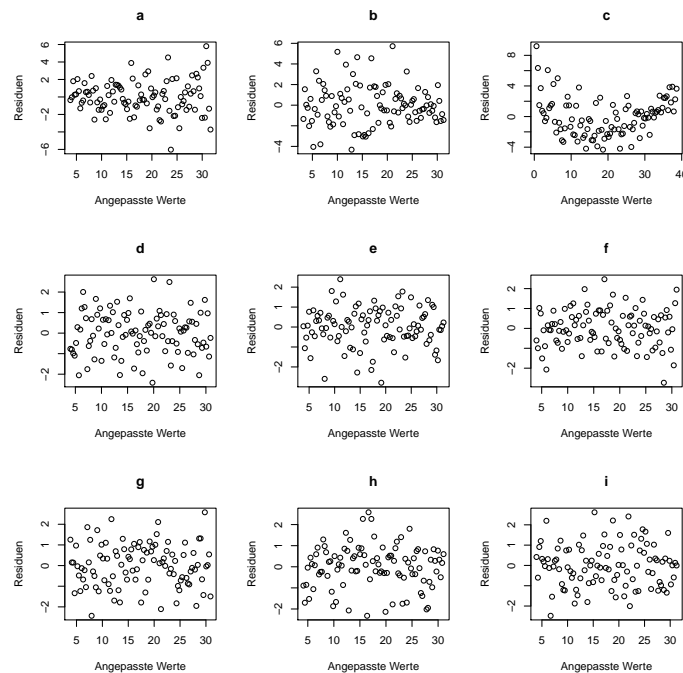


Abbildung 5.5: Tukey-Anscombe Plots. (a) Der Plot zeigt einen Trichter, der nach rechts grösser wird. D.h., die Fehlervarianz nimmt zu, wenn die beobachteten Werte der Zielgrösse gross sind. Damit ist die Annahme der gleichen Fehlervarianzen verletzt und die lineare Regression liefert keine zuverlässigen Werte. Evtl. kann man das Problem lösen, indem man die Zielgrösse transformiert. (b) Der Plot zeigt einen Trichter, der nach rechts kleiner wird. Wie in Fall (a) sind die Annahmen verletzt und die Regression liefert keine zuverlässigen Schätzwerte. (c) Die Punkte folgen einer U-förmigen Kurve. Das wahre Modell scheint komplizierter zu sein als das von uns angepasste. Evtl. kann man das Problem lösen, indem man einen quadratischen Term als erklärende Variable hinzufügt. (d-i) In diesen Plots ist kein Muster zu erkennen. Sie stellen einwandfreie Tukey-Anscombe Plots dar.

- langfristige Zonen mit durchwegs positiven oder negativen Residuen;
 die Punktschätzungen sind immer noch OK, aber die Tests und Konfidenzintervalle stimmen nicht mehr; evtl. Regression mit korrelierten Fehlern verwenden (siehe Literatur)

Für den Ozon-Datensatz ist der serielle Korrelations-Plot in Abbildung 5.6 gezeigt.

Der Normalplot

Mit dem Normalplot (siehe Kapitel 4.5.5) können wir die Normalverteilungsannahme in (5.2) überprüfen.

Im Idealfall: approximativ eine Gerade

Abweichungen:

- Abweichung von einer Geraden evtl. robuste Regression benutzen (siehe Literatur)

Für den Ozon-Datensatz ist der Normalplot in Abbildung 5.6 gezeigt.

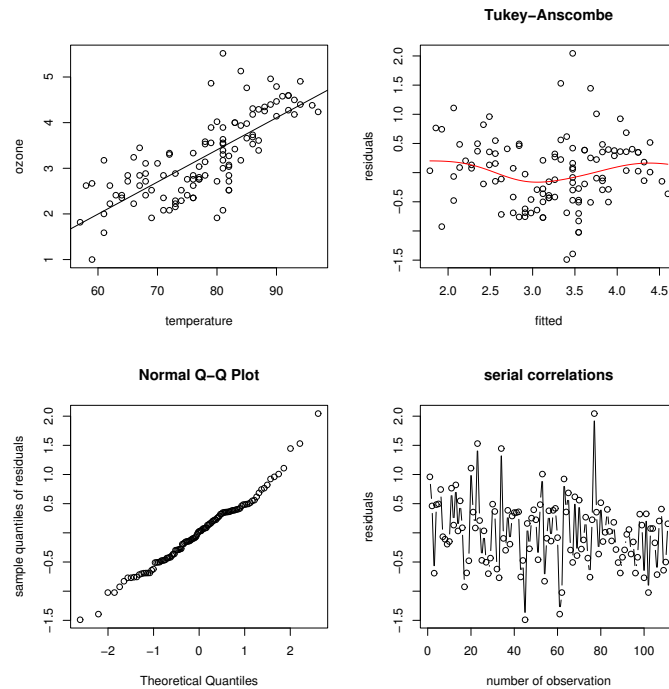


Abbildung 5.6: Ozon-Datensatz: Streudiagramm mit angepasster Regressionsgerade (oben links); Tukey-Anscombe Plot (oben rechts); Normalplot (unten links); serieller Korrelations-Plot (unten rechts).

Das Auffinden eines guten Modells

Oftmals werden mehrere Modelle in einer Art “workflow-feedback” Prozeß betrachtet und angepasst. Man beginnt mit einem ersten Modell; dann, aufgrund von Residuenanalyse wird das Modell modifiziert. Das modifizierte Modell (immer noch als linear angenommen in evtl. transformierten Variablen) wird wiederum mit linearer Regression angepasst, und mit Residuenanalyse wird das neue Modell beurteilt. Dieses Vorgehen wird iteriert bis man ein “zufriedenstellendes” Modell gefunden und angepasst hat.

5.3 Multiple lineare Regression

Oftmals hat man mehrere erklärende Variablen $x_{i,1}, \dots, x_{i,p-1}$ ($p > 2$).

5.3.1 Das Modell der multiplen linearen Regression

Das Modell ist wie folgt:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + E_i,$$

$$E_1, \dots, E_n \text{ i.i.d. , } \mathcal{E}(E_i) = 0, \text{ Var}(\mathcal{E}_i) = \sigma^2.$$

Wie bei der einfachen linearen Regression nehmen wir an, dass die erklärenden Variablen deterministisch sind. Es ist oftmals nützlich, das obige Modell in Matrix-Schreibweise darzustellen:

$$\begin{matrix} Y \\ n \times 1 \end{matrix} = \begin{matrix} X \\ n \times p \end{matrix} \times \begin{matrix} \beta \\ p \times 1 \end{matrix} + \begin{matrix} E \\ n \times 1 \end{matrix} \quad (5.5)$$

wobei X eine $(n \times p)$ -Matrix ist mit Spaltenvektoren $(1, 1, \dots, 1)^T, (x_{1,1}, x_{2,1}, \dots, x_{n,1})^T$ und letztendlich $(x_{1,p-1}, x_{2,p-1}, \dots, x_{n,p-1})^T$.

Beispiele von multipler linearer Regression sind unter anderen:

- **Einfache lineare Regression:** $Y_i = \beta_0 + \beta_1 x_i + E_i$ ($i = 1, \dots, n$).

$$p = 2 \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

- **Lineare Regression mit mehreren erklärenden Variablen:** $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + E_i$ ($i = 1, \dots, n$).

$$p = 3, \quad X = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

- **Lineare Regression mit quadratischen erklärenden Variablen:** $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$ ($i = 1, \dots, n$).

$$p = 3, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Beachte, dass das Modell linear in den Parametern ist. Daher wird es ein lineares Modell genannt.

- **Regression mit transformierten erklärenden Variablen:**

$$Y_i = \beta_0 + \beta_1 \log(x_{i2}) + \beta_2 \sin(\pi x_{i3}) + E_i \quad (i = 1, \dots, n).$$

$$p = 3, \quad X = \begin{pmatrix} 1 & \log(x_{12}) & \sin(\pi x_{13}) \\ 1 & \log(x_{22}) & \sin(\pi x_{23}) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_{n2}) & \sin(\pi x_{n3}) \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Wiederum ist das Modell linear in den Koeffizienten β_j und wird daher lineares Modell genannt (obwohl es in den x_{ij} 's nichtlinear ist).

Es gibt einen wichtigen Unterschied bei der Interpretation der Parameter bei der einfachen und der multiplen linearen Regression.

Einfache lineare Regression: β_1 ist die erwartete Zunahme der Zielgrösse, wenn x_1 um eine Einheit erhöht wird.

Multiple lineare Regression: β_i ist die erwartete Zunahme der Zielgrösse, wenn x_i um eine Einheit erhöht wird **und die übrigen Variablen fixiert werden.**

Es ist i.A. **nicht** möglich von den Koeffizienten einer multiplen linearen Regression auf die Koeffizienten einer einfachen linearen Regression mit nur einer der erklärenden Variablen zu schliessen. Die umgekehrte Richtung ist auch nicht möglich. Wenn man in einer multiplen Regression auch nur eine erklärende Variable hinzufügt oder weglässt, können sich alle Parameter völlig ändern. Man müsste die Parameter also erneut schätzen¹

5.3.2 Parameterschätzungen und t-Tests

Analog zur einfachen linearen Regression wird meist die Methode der Kleinsten Quadrate benutzt:

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1} \text{ sind Minimierer von } \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1}))^2.$$

Die eindeutige Lösung dieser Optimierung ist explizit darstellbar falls $p < n$ (d.h., es gibt weniger Variablen als Beobachtungen):

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

wobei $\hat{\beta}$ den $p \times 1$ Vektor $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})^T$ bezeichnet, und X, Y wie in (5.5).

¹Es gibt Spezialfälle, bei denen man von den Koeffizienten der multiplen linearen Regression auf die der einfachen linearen Regression und umgekehrt schliessen kann. In der Versuchsplanung nennt man eine solche Versuchsanordnung "orthogonales Design". In dieser Vorlesung werden wir das aber nicht behandeln.

Die Schätzung der Fehlervarianz ist

$$\hat{\sigma} = \frac{1}{n-p} \sum_{i=1}^n R_i^2, \quad R_i = Y_i - (\hat{\beta}_0 + \sum_{j=1}^{p-1} \hat{\beta}_j x_{i,j}).$$

Unter der Annahme, dass die Fehler normalverteilt sind wie in (5.2), können auch ähnlich zur einfachen Regression t-Tests für die folgenden Hypothesen gemacht werden:

$$H_{0,j} : \beta_j = 0; \quad H_{A,j} : \beta_j \neq 0 \quad (j = 0, \dots, p-1).$$

Der wesentliche Unterschied besteht aber in der Interpretation der Parameter:

β_j misst den linearen Effekt
der j -ten erklärenden Variable auf die Zielvariable Y
nach Elimination der linearen Effekte
aller anderen Variablen auf Y ($j = 1, \dots, p-1$)

Insbesondere impliziert dies, dass man die Koeffizienten β_j nicht einfach durch einzelne, individuelle simple lineare Regressionen von Y auf die j -te erklärende erhalten kann.

5.3.3 Der F-Test

Der (globale) F-Test quantifiziert die Frage, ob es mindestens eine erklärende Variable gibt, welche einen relevanten Effekt auf die Zielvariable (im Sinne der linear Regression) hat. Die folgende Nullhypothese wird beim (globalen) F-Test betrachtet:

$$H_0 : \beta_1 = \dots = \beta_{p-1} = 0$$

$$H_A : \text{mindestens ein } \beta_j \neq 0 \quad (j = 1, \dots, p-1).$$

Der P-Wert des (globalen) F-Tests wird in der Regel mit dem Computer berechnet.

Manchmal kann es sein, dass der F-Test einen sehr kleinen P-Wert (signifikant) liefert, die P-Werte der t-Tests aller Parameter aber sehr gross sind (nicht signifikant). Das kann auftreten, wenn zwei Variablen (oder mehr) stark korreliert sind. Am besten prüft man also die Korrelation zwischen allen erklärenden Variablen. Falls zwei Variablen stark korreliert sind, lässt man einfach eine von beiden weg (die andere bringt nämlich keine zusätzliche Information).

5.3.4 Das Bestimmtheitsmass R^2

Das Bestimmtheitsmass R^2 ist in der multiplen linearen Regression über die Formel (5.4) definiert (mit Hilfe der Zerlegung in (5.3)). Wie bei der einfachen

linearen Regression, entspricht das Bestimmtheitsmass der quadrierten Korrelation zwischen den beobachteten Werten der Zielgrösse und den geschätzten Werten der Zielgrösse.

$$R^2 = \hat{\rho}_{Y\hat{Y}}^2,$$

5.3.5 Residuenanalyse

Die Residuenanalyse geht völlig analog zu Kapitel 5.2.6. Das allgemeine Vorgehen bei multipler linearer Regression ist wie in Kapitel 5.2.5, unter Einbezug des F-Tests nach dem Schritt 1.

5.4 Software

Mit der Statistik-Software R (und auch mit den meisten anderen Statistik Programmen) kann man einfache und multiple Regressionen problemlos ausführen. Wir zeigen jeweils ein Beispiel und illustrieren den Computer-Output.

5.4.1 Einfache lineare Regression

Wir behandeln das fiktive Beispiel mit den Buchpreisen. Zunächst laden wir die Daten:

```
x <- c(50,100,150,200,250,300,350,400,450,500)
y <- c(9.9,10.7,13.3,15.2,16.4,23.6,23.5,21.1,28.9,29.1)
```

Anschliessend wird die Funktion `lm` verwendet um ein lineares Modell zu schätzen. Das erste Argument ist dabei eine Formel, die angibt, welche Variable die erklärende und welche Variable die Zielgrösse ist: $y \sim x$ gibt an, dass y die Zielgrösse und x die erklärende Variable ist. Wir speichern das Resultat unter der Variablen `fm` ab:

```
fm <- lm(y ~ x)
```

Mit der Funktion `summary` kann man das Ergebnis übersichtlich darstellen. `summary(fm)` liefert:

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.6958	-0.5944	-0.2203	0.9300	3.3048

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.793333   1.391060   4.884  0.00122 **
x            0.045006   0.004484  10.037 8.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.036 on 8 degrees of freedom
Multiple R-squared:  0.9264, Adjusted R-squared:  0.9172
F-statistic: 100.8 on 1 and 8 DF,  p-value: 8.254e-06

```

In den ersten beiden Zeilen wird der Funktionsaufruf reproduziert. In dem nächsten Abschnitt erhält man einen Überblick über die Residuen. Hier sollte man sich fragen, ob das Maximum oder Minimum aussergewöhnlich grosse Werte annimmt (das könnte z.B. auftreten, wenn man bei dem Eintippen der Daten den Dezimalpunkt bei einer Beobachtung um eine oder ein paar Positionen verschoben hat; also z.B. 103.0 statt 1.030). In dem Abschnitt **Coefficients** steht eine Tabelle mit allen wichtigen Informationen zu den geschätzten Parametern. In der ersten Spalte steht in der ersten Zeile **(Intercept)**. Diese Zeile behandelt den Parameter β_0 . In der zweiten Zeile steht **x**. Diese Zeile behandelt den Parameter β_1 . Falls es noch mehr erklärende Variablen geben würde, würden sie in den nächsten Zeilen folgen. Die zweite Spalte mit dem Titel **Estimate** beinhaltet die geschätzten Parameterwerte. In diesem Beispiel ist also $\beta_0 = 6.793$ und $\beta_1 = 0.045$. Die dritte Spalte mit dem Titel **Std. Error** zeigt den geschätzten Standardfehler des Parameters an. In diesem Beispiel ist also $\widehat{s.e.}(\hat{\beta}_0) = 1.391$ und $\widehat{s.e.}(\hat{\beta}_1) = 0.0044$. Wenn wir nun testen wollen, ob die Daten plausibel sind, wenn ein Parameter gleich null ist, müssen wir den t-Test aus Kapitel 5.2.3 machen. Die Teststatistik wurde von dem Computer schon berechnet. Sie ist gerade der Quotient aus geschätztem Wert und geschätzter Standardfehler. Für den Parameter β_0 ergibt sich: $t_0 = \frac{\beta_0}{\widehat{s.e.}(\hat{\beta}_0)} = \frac{6.793}{1.391} = 4.884$ und $t_1 = \frac{\beta_1}{\widehat{s.e.}(\hat{\beta}_1)} = \frac{0.045}{0.00484} = 10.037$. In der letzten Spalte werden die daraus berechneten P-Werte dargestellt. Der P-Wert für die Hypothese $H_0 : \beta_0 = 0$ (mit zweiseitiger Alternative) ist 0.00122. Der P-Wert für die Hypothese $H_0 : \beta_1 = 0$ (mit zweiseitiger Alternative) ist 8.25E-6 ($= 8.25 \cdot 10^{-6}$). Direkt nach dem P-Wert wird mit Sternchen dargestellt, auf welchem Signifikanzniveau der Parameter unterschiedlich von null ist (z.B. bedeutet '*', dass der Parameter auf dem 5% Signifikanzniveau von null verschieden ist; siehe die Zeile mit **Signif. codes**). Man sieht, dass beide Parameter hochsignifikant von null verschieden sind. Ein approximatives 95%-Vertrauensintervall für β_0 ist $[6.79 - 2 \cdot 1.39; 6.79 + 2 \cdot 1.39]$. Für das exakte 95% Vertrauensintervall muss man $t_{10-2;0.975} = 2.31$ in der Tabelle nachschlagen. Man erhält dann als exaktes 95%-Vertrauensintervall $[6.79 - 2.31 \cdot 1.39; 6.79 + 2.31 \cdot 1.39]$. Man kann auf drei Arten erkennen, dass β_0 auf dem 5%-Niveau (allgemein: α) von Null verschieden ist:

1. Der P-Wert ist kleiner als 5% (α)
2. Der t-Wert ist grösser als $t_{10-2;1-\alpha/2} \approx 2$

3. Das 95%-Vertrauensintervall ($1 - \alpha$ Vertrauensintervall) enthält die null nicht.

In dem letzten Abschnitt wird zunächst die Fehlervarianz σ^2 geschätzt. In diesem Beispiel ist $\hat{\sigma} = 2.036$. Nun haben wir genug Informationen, um das ganze Modell aufzuschreiben:

$$Y_i = 6.79 + 0.045x_i + E_i \quad (i = 1, \dots, n),$$

$$E_1, \dots, E_n \sim \mathcal{N}(0, 2.036^2)$$

Nach der Fehlervarianz steht noch die Information **on 8 degrees of freedom**. Sie sind wichtig für den t-Test. Die "degrees of freedom" entsprechen $n - p$, wobei n die Anzahl Beobachtungen und p die Anzahl β s ist, die man geschätzt hat. In unserem Fall ist $n = 10$ und $p = 2$. Daher sind die "degrees of freedom" gleich 8 (Sie könnten also aus dem Output alleine sehen, wie viele Daten man zur Verfügung hatte: degrees of freedom + Anzahl geschätzter β s).

Die Residuen des geschätzten Modells `fm` sind unter `fm$resid` gespeichert. Die vom Modell angepassten Werte sind unter `fm$fitted` gespeichert. Den Tukey-Anscombe Plot kann man also mit dem Befehl `plot(fm$fitted, fm$resid)` einfach darstellen. Für den Normalplot kann man den Befehl `qqnorm(fm$resid)` verwenden. Um die Residuen in ihrer Beobachtungsreihenfolge zu zeichnen², verwendet man den Befehl `plot(1:length(fm$resid), fm$resid)`.

Unter **Multiple R-squared** ist der Wert von R^2 angegeben. In dieser Vorlesung haben wir die Grösse **Adjusted R-squared** nicht besprochen. In der letzten Zeile kommt schliesslich noch der F-Test, den wir in Kapitel 5.3.3 behandelt haben. Die Nullhypothese $H_0 : \beta_0 = \beta_1 = 0$ kann also wegen des kleinen P-Wertes von $8.25 \cdot 10^{-6}$ verworfen werden.

5.4.2 Multiple Linear Regression

Der Computeroutput für die multiple lineare Regression ist praktisch identisch mit dem Output der einfachen linearen Regression. Der einzige Unterschied besteht darin, dass für jede Variable eine eigene Zeile in der Tabelle **Coefficients** erscheint.

Wir betrachten ein Beispiel wo die Asphalt-Qualität als Funktion von 6 erklärenden Variablen analysiert wird.

```
y = RUT : "rate of rutting" = change of rut depth in inches per million
          wheel passes
          ["rut" := 'Wagenspur', ausgefahrenes Geleise]
x1 = VISC : viscosity of asphalt
x2 = ASPH : percentage of asphalt in surface course
x3 = BASE : percentage of asphalt in base course
```

²Ich nehme an, dass der erste Eintrag in x und y zuerst gemacht wurde. Dann wurde der zweite Eintrag beobachtet usw.

```

x4 = RUN : '0/1' indicator for two sets of runs.
x5 = FINES: 10* percentage of fines in surface course
x6 = VOIDS: percentage of voids in surface course

```

Die Daten sind in Abbildung 5.7 dargestellt. Die Zusammenhänge werden linea-

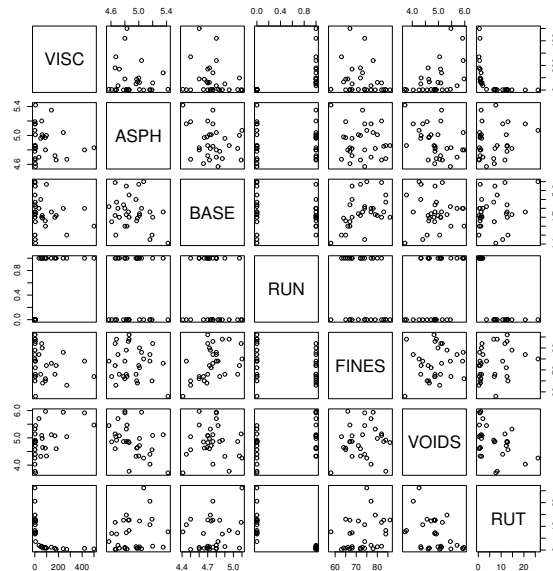


Abbildung 5.7: Paarweise Streudiagramme für den Asphalt-Datensatz. Die Zielvariable ist "RUT".

rer, wenn man die Zielvariable "RUT" logarithmiert und ebenfalls die erklärende Variable "VISC".

```

y = LOGRUT : log("rate of rutting") = log(change of rut depth in inches
              per million wheel passes)
              ["rut":= 'Wagenspur', ausgefahrenes Geleise]
x1 = LOGVISC : log(viscosity of asphalt)
x2 = ASPH : percentage of asphalt in surface course
x3 = BASE : percentage of asphalt in base course
x4 = RUN : '0/1' indicator for two sets of runs.
x5 = FINES: 10* percentage of fines in surface course
x6 = VOIDS: percentage of voids in surface course

```

Die transformierten Daten sind in Abbildung 5.8 dargestellt.

Mittels R wird ein multiples lineares Modell angepasst. Der Output sieht wie folgt aus:

```

Call:
lm(formula = LOGRUT ~ ., data = asphalt1)

```

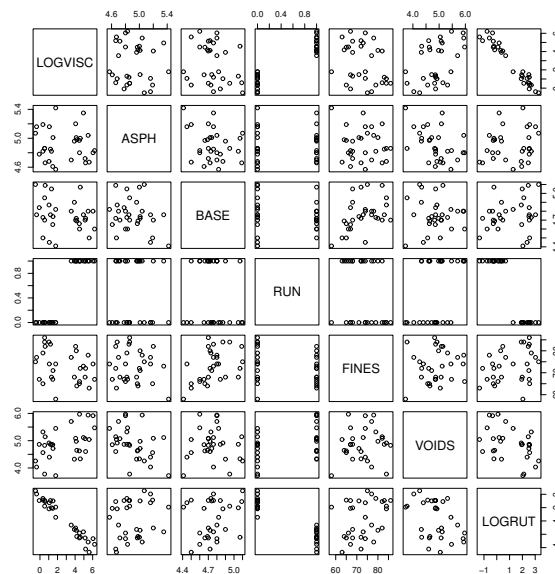


Abbildung 5.8: Paarweise Streudiagramme für den transformierten Asphalt-Datensatz. Die Zielvariable ist “LOGRUT”, die log-transformierte ursprüngliche Variable “RUT”. Die erklärende Variable “LOGVISC” ist ebenfalls die log-transformierte ursprüngliche Variable “VISC”.

Residuals:

	Min	1Q	Median	3Q	Max
	-0.48348	-0.14374	-0.01198	0.15523	0.39652

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-5.781239	2.459179	-2.351	0.027280	*
LOGVISC	-0.513325	0.073056	-7.027	2.90e-07	***
ASPH	1.146898	0.265572	4.319	0.000235	***
BASE	0.232809	0.326528	0.713	0.482731	
RUN	-0.618893	0.294384	-2.102	0.046199	*
FINES	0.004343	0.007881	0.551	0.586700	
VOIDS	0.316648	0.110329	2.870	0.008433	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2604 on 24 degrees of freedom

Multiple R-Squared: 0.9722, Adjusted R-squared: 0.9653

F-statistic: 140.1 on 6 and 24 DF, p-value: < 2.2e-16

Wir sehen, dass die Variablen “LOGVISC”, “ASPH” und “VOID” signifikant oder sogar hoch-signifikant sind; die Variable “RUN” ist bloss schwach signifikant. Der F-Test ist hoch-signifikant, das Bestimmtheitsmass R^2 sehr nahe bei

1. Die degrees of freedom sind hier $n - p = 24$ mit $p = 7$, d.h. $n = 31$. Die Residuenanalyse ist mittels Tukey-Anscombe und Normalplot in Abbildung 5.9 zusammengefasst: die Normalverteilungsannahme für die Fehler ist eine ver-

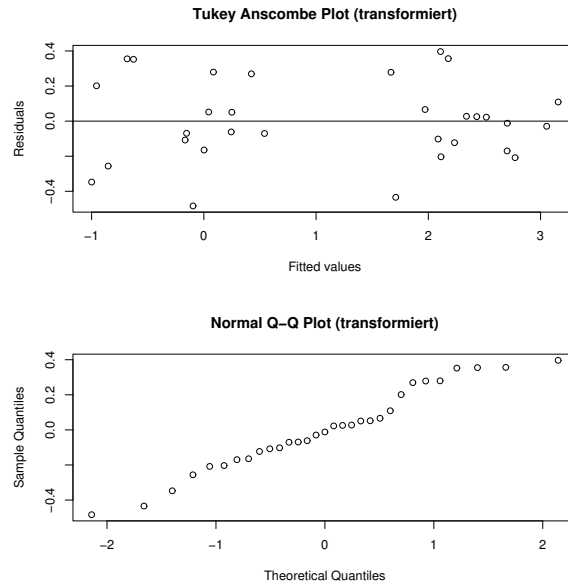


Abbildung 5.9: Tukey-Anscombe Plot (oben) und Normalplot (unten) beim Asphalt-Datensatz mit den transformierten Variablen “LOGRUT” und “LOG-VISC”.

nünftige Approximation. Der Tukey-Anscombe Plot zeigt keine systematische Variation ; das R^2 ist sehr nahe bei 1. Die multiple linear Regression erklärt also sehr viel der totalen Variation.

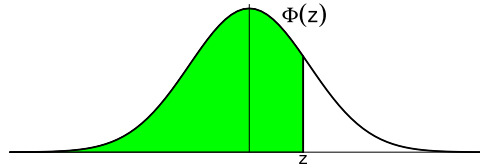
Ohne log-Transformationen, d.h. das untransformierte Modell wie in Abbildung 5.7, ist das Bestimmtheitsmass $R^2 = 0.7278$, also wesentlich schlechter als im transformierten Modell.

Kapitel 6

Tabellen

Tabelle der Kumulativen Normalverteilung

$$\Phi(z) = P[Z \leq z], \quad Z \sim \mathcal{N}(0, 1)$$

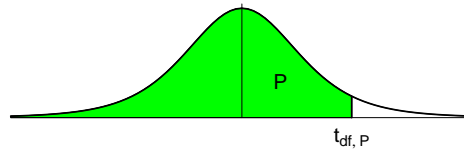


Bsp.: $P[Z \leq 1.96] = 0.975$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Achtung: In der Prüfung erhalten Sie evtl. eine andere/ausführlichere Tabelle

Perzentile der t-Verteilung



Bsp.: $t_{9; 0.975} = 2.262$

df	$t_{0.60}$	$t_{0.70}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
31	0.255	0.530	0.853	1.309	1.696	2.040	2.452	2.744
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738
33	0.255	0.530	0.853	1.308	1.693	2.035	2.445	2.733
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

Achtung: In der Prüfung erhalten Sie evtl. eine andere/ausführlichere Tabelle

Kapitel 7

Anhang

7.1 Normalapproximation des Binomialtests: Zwischenrechnung

Gemäss dem Zentralen Grenzwertsatz gilt: $\bar{X}_n \approx \mathcal{N}(\mu, \sigma_X^2/n)$ mit $\mu = \pi$ und $\sigma_X^2 = \pi(1 - \pi)$. Zu zeigen ist Gleichung (4.4):

$$P(|\bar{X}_n - \pi| > \varepsilon) \approx 2 - 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma_X}\right)$$

Im Folgenden verwenden wir folgende Notationen und Rechentricks:

1. Notation: $Z \sim \mathcal{N}(0, 1)$ und somit $P(Z \leq z) = \Phi(z)$
2. $P(|X| > c) = P((X < -c) \cup (X > c)) = P(X < -c) + P(X > c)$
3. $\frac{\bar{X}_n - \pi}{\sigma_X/\sqrt{n}} \sim \mathcal{N}(0, 1)$
4. $P(Z < z) = P(Z \leq z)$, weil $P(Z = z) = 0$
5. $P(Z < -z) = P(Z > z) = 1 - P(Z \leq z)$

Nun zeigen wir Gleichung (4.4). Zunächst werden wir standardisieren, sodass wir die Zufallsvariable Z erhalten. Anschliessend formen wir mit Rechentricks

um, bis wir das gewünschte Ergebnis erhalten:

$$\begin{aligned}
 P(|\bar{X}_n - \pi| > \varepsilon) &\stackrel{(2)}{=} P(\bar{X}_n - \pi < -\varepsilon) + P(\bar{X}_n - \pi > \varepsilon) = \\
 &= P\left(\frac{\bar{X}_n - \pi}{\sigma_X/\sqrt{n}} < -\frac{\varepsilon}{\sigma_X/\sqrt{n}}\right) + P\left(\frac{\bar{X}_n - \pi}{\sigma_X/\sqrt{n}} > \frac{\varepsilon}{\sigma_X/\sqrt{n}}\right) = \\
 &= P\left(\frac{\bar{X}_n - \pi}{\sigma_X/\sqrt{n}} < -\frac{\sqrt{n}\varepsilon}{\sigma_X}\right) + P\left(\frac{\bar{X}_n - \pi}{\sigma_X/\sqrt{n}} > \frac{\sqrt{n}\varepsilon}{\sigma_X}\right) = \\
 &\stackrel{(1,3)}{=} P(Z < -\frac{\sqrt{n}\varepsilon}{\sigma_X}) + P(Z > \frac{\sqrt{n}\varepsilon}{\sigma_X}) = \\
 &\stackrel{(5)}{=} 2 * P(Z > \frac{\sqrt{n}\varepsilon}{\sigma_X}) = \\
 &\stackrel{(5)}{=} 2 * (1 - P(Z \leq \frac{\sqrt{n}\varepsilon}{\sigma_X})) = \\
 &\stackrel{(1)}{=} 2 * (1 - \Phi(\frac{\sqrt{n}\varepsilon}{\sigma_X})) = \\
 &= 2 - 2\Phi(\frac{\sqrt{n}\varepsilon}{\sigma_X})
 \end{aligned}$$