

1 Einführung

Übersicht

| | |
|---|----|
| 1.1 Statistische Inferenz | 1 |
| 1.2 Beispiele | 2 |
| 1.3 Statistische Modelle | 9 |
| 1.4 Inhalt und Notation der Kapitel | 12 |

1.1 Statistische Inferenz

In diesem Buch werden statistische Methoden beschrieben, die es ermöglichen, aus Daten zu lernen. Ein Schwerpunkt liegt auf konstruktiven Konzepten, die also möglichst universell anwendbar sind. Dazu zählen insbesondere Methoden, die auf dem zentralen Konzept – der Likelihood – basieren: einerseits die (reine) Likelihood-Inferenz, andererseits die Bayes-Inferenz, bei der die Likelihoodfunktion mit Vorwissen kombiniert wird. In beiden Fällen muss zunächst ein statistisches Modell formuliert werden.

Man kann grob drei statistische Problemstellungen unterscheiden. Zum Ersten will man, bei gegebenen Daten und unter Annahme eines bestimmten Modells, häufig Aussagen über unbekannte Parameter des Modells treffen, die Parameter also in gewissem Sinne „schätzen“. Zweitens will man möglicherweise untersuchen, welches von verschiedenen Modellen die vorliegenden Daten besser beschreibt. Drittens kann es von Interesse sein, aus den gegebenen Daten zu lernen, um zukünftige Beobachtungen vorherzusagen. Diese drei Themenkreise bezeichne ich als das *Schätzproblem*, das *Modellwahlproblem* und das *Prognoseproblem*.

Moderne statistische Inferenz ist ohne den Einsatz des Computers heutzutage undenkbar. Verschiedene numerische Methoden, insbesondere zur Optimierung und Integration von Funktionen, werden unentwegt zur Lösung von statistischen Problemen verwendet. Dieses Lehrbuch legt Wert auf die praktische Umsetzung der diskutierten Methoden, daher werden an vielen Stellen Beispielprogramme in der statistischen Programmiersprache R geliefert. In einem eigenen Anhang werden die verwendeten numerischen Methoden skizziert. Ein weiterer Anhang gibt die verschiedenen Wahrscheinlichkeitsverteilungen wieder, die im Buch verwendet werden.

Einige Beispiele aus der realen Welt werden sich wie ein roter Faden durch dieses Buch ziehen und sollen im Folgenden kurz vorgestellt werden.

1.2 Beispiele

1.2.1 Anteilsschätzung

Eine der ältesten statischen Problemstellungen betrifft die Schätzung von Wahrscheinlichkeiten, die eng mit der Anteilsschätzung verknüpft ist. Man betrachtet üblicherweise ein Zufallsexperiment, bei dem ein bestimmtes Ereignis mit einer unbekannten Wahrscheinlichkeit π eintritt. Beispielsweise könnte das betrachtete Ereignis das Auftreten eines bestimmten genetischen Defekts bei einer zufällig ausgewählten Person aus einer definierten *Grundgesamtheit*, beispielsweise das Auftreten des Klinefelter-Syndroms bei allen männlichen Neugeborenen der Schweiz, sein. Das Experiment wird nun üblicherweise mehrmals, sagen wir n -mal, wiederholt. In obigem Beispiel würden also n männliche Neugeborene genetisch untersucht werden. Bei x Neugeborenen, $x \in \{0, 1, \dots, n\}$, konnte der genetische Defekt festgestellt werden, somit hatten $n - x$ Neugeborene den Defekt nicht. Man möchte nun aus dieser *Stichprobe* Rückschlüsse auf den Anteil π in der zugrundeliegenden Grundgesamtheit treffen.

Zu bemerken ist, dass Wahrscheinlichkeiten π häufig in *Chancen* („Odds“) angegeben werden, wobei die *Chance* sich durch $\omega = \pi/(1 - \pi)$ ergibt, also gleich dem Verhältnis der Wahrscheinlichkeiten des betrachteten Ereignisses und des zugehörigen Gegenereignisses ist. Beispielsweise entspricht eine Wahrscheinlichkeit von $\pi = 0.5$ einer Chance von $\omega = 1$ (in Worten 1:1), einer Chance von $\omega = 9$ (9:1) entspricht beispielsweise eine Wahrscheinlichkeit von $\pi = 0.9$. Allgemein kann man bei gegebener Chance ω die zugehörige Wahrscheinlichkeit π durch $\pi = \omega/(1 + \omega)$ berechnen.

1.2.2 Vergleich von Anteilen

Eng mit der Anteilsschätzung verknüpft ist der Vergleich von Anteilen. Beispielsweise könnte es von Interesse sein, in einer klinischen Studie die Erkrankungs-Häufigkeiten in der Behandlungs- und Kontrollgruppe zu vergleichen. Somit liegen zwei unbekannte Erkrankungs-Wahrscheinlichkeiten π_1 und π_2 vor, mit zugehörigen Beobachtungen x_1 und x_2 bei Stichprobenumfängen n_1 und n_2 . Zum Vergleich der zwei Wahrscheinlichkeiten werden verschiedene Größen verwendet. Es ist naheliegend, die *Risikodifferenz* $\pi_1 - \pi_2$ und das *relative Risiko* π_1/π_2 zu betrachten. Beliebiger ist auch die Verwendung der *relativen Chance* („Odds Ratio“) $(\pi_1/(1 - \pi_1))/(\pi_2/(1 - \pi_2)) = \omega_1/\omega_2$, dem Quotienten der zugehörigen Chancen ω_1 und ω_2 . Dem Fall, dass die zwei Wahrscheinlichkeiten

Tab. 1.1: Erkrankungshäufigkeiten (absolute Zahlen) von Präeklampsie in neun randomisierten, kontrollierten klinischen Studien in den Behandlungsgruppen mit Diuretika und den Kontrollgruppen mit Scheinpräparat. Odds Ratios (OR) von Behandlungs- gegen Kontrollgruppen sind ebenfalls angegeben. Die Studien sind mit dem Hauptautor bezeichnet.

| Studie | Behandlung | | Kontrollen | | OR |
|-----------|------------|------------|------------|------------|------|
| Weseley | 11% | (14/131) | 10% | (14/136) | 1.04 |
| Flowers | 5% | (21/385) | 13% | (17/134) | 0.40 |
| Menzies | 25% | (14/57) | 50% | (24/48) | 0.33 |
| Fallis | 16% | (6/38) | 45% | (18/40) | 0.23 |
| Cuadros | 1% | (12/1011) | 5% | (35/760) | 0.25 |
| Landesman | 10% | (138/1370) | 13% | (175/1336) | 0.74 |
| Krans | 3% | (15/506) | 4% | (20/524) | 0.77 |
| Tervila | 6% | (6/108) | 2% | (2/103) | 2.97 |
| Campbell | 42% | (65/153) | 39% | (40/102) | 1.14 |

lichkeiten gleich sind, d. h. $\pi_1 = \pi_2$, entspricht somit eine Risikodifferenz von Null, sowie ein relatives Risiko als auch eine relative Chance von Eins.

Häufig will man nun durch statistische Verfahren untersuchen, ob man das einfachere Modell mit einem Parameter $\pi = \pi_1 = \pi_2$ dem Komplexeren mit zwei Parametern π_1 und π_2 vorziehen kann. Beispielsweise wurde in Studien zur Vorbeugung von Präeklampsie (einer Erkrankung in der Schwangerschaft, die mit Bluthochdruck und Wassereinlagerungen einhergeht) die Wirkung von Diuretika-Medikamenten, die Wasser ausschwemmen, untersucht. Um die Wirkung von Placebo-Effekten zu trennen, wurden parallel Frauen zufällig als Kontrollen ausgewählt, die stattdessen ein Scheinpräparat ("Plazebo") einnahmen. Die Erkrankungszahlen aus neun von Collins et al. (1985) als seriös bewerteten Studien sind in Tabelle 1.1 aufgeführt. Die Frage ist also, ob sich die Inzidenzen in den Behandlungsgruppen von denen in den entsprechenden Kontrollgruppen unterscheiden, d. h. ob die Diuretika einen positiven (oder negativen) Effekt haben.

Um alle bisherigen Studienergebnisse auf einmal zu verwerten, können Techniken der *Meta-Analyse* eingesetzt werden. Diese erlauben es, die Erkenntnisse aus den einzelnen Studien statistisch zusammenzufassen. Dabei wird zu differenzieren sein zwischen der Annahme eines gleichen Behandlungseffekts in allen Studien und der Annahme von Heterogenität in den Studien, d. h. der Annahme von unterschiedlichen wahren Behandlungseffekten in den Einzelstudien.

1.2.3 Schätzung der Größe einer Population

Die *Rückfangmethode* („Capture-recapture method“) ist ein Verfahren zur Abschätzung der Größe einer Population von Individuen und soll hier am Beispiel von Fischen in einem See erläutert werden. Um herauszufinden, wieviele Fische N im See leben, fängt man zunächst M Fische, markiert diese und lässt sie wieder frei. Einige Zeit später fängt man n Fische und zählt die x markierten Fische in dieser Stichprobe. Durch intuitive Überlegungen kommt man zu dem Schluss, dass der Anteil der markierten Fische in der zweiten Stichprobe ungefähr gleich dem Anteil aller markierten Fische in der Gesamtpopulation sein sollte, d. h.

$$\frac{x}{n} \approx \frac{M}{N}.$$

Somit ergibt sich der heuristische Schätzwert $\hat{N} \approx M \cdot n/x$ für die Anzahl der Fische N im See. Dieser wird sich in Beispiel 2.2 auch statistisch begründen lassen. Man beachte aber, dass \hat{N} für $x = 0$ nicht definiert ist. Des Weiteren werden wir auch Intervalle angeben können, in denen mit hoher Sicherheit (was das genau bedeutet wird noch zu präzisieren sein) plausible Werte für die unbekannte Anzahl an Fischen liegen.

1.2.4 Überprüfung eines genetischen Gleichgewichtszustandes

Das *Hardy-Weinberg-Gleichgewicht* (nach GODFREY H. HARDY (1879–1944) und WILHELM WEINBERG (1862–1937)) ist eine zentrale Modellannahme der Populationsgenetik. Hierzu wird eine Population von diploiden Organismen betrachtet, d. h. Organismen mit doppeltem Chromosomensatz wie z. B. Menschen. Wir betrachten nun einen bestimmten Ort (*Locus*) auf einem Chromosom mit zwei Allelen A und a , ein Allel bezeichnet hierbei die möglichen Ausprägungen eines Gens. Ein Individuum hat somit die drei möglichen Genotypen AA , Aa und aa ; dabei wird nicht berücksichtigt, welches Allel von Vater oder Mutter vererbt wurde, sodass aA nicht von Aa unterscheidbar ist. Seien nun π_1, π_2 und π_3 (mit $\pi_i \geq 0$ und $\pi_1 + \pi_2 + \pi_3 = 1$) die Häufigkeiten der drei Genotypen in der Population. Das Hardy-Weinberg-Gleichgewicht, das unter bestimmten Annahmen an den Vererbungsprozess in der Population hergeleitet werden kann, fordert nun eine spezielle Form für die Wahrscheinlichkeiten π_1, π_2 und π_3 :

$$\pi_1 = q^2, \quad \pi_2 = 2q(1-q) \quad \text{und} \quad \pi_3 = (1-q)^2 \quad (1.1)$$

mit $q \in (0, 1)$. Im Hardy-Weinberg-Gleichgewicht gibt es somit nur einen unbekannten Parameter, q , der als Häufigkeit des Allels A interpretiert werden kann, während im Allgemeinen zwei Parameter unbekannt sind (der dritte ergibt sich über $\pi_1 + \pi_2 + \pi_3 = 1$). Eine Veranschaulichung des Hardy-Weinberg-Gleichgewichts ist im *de Finetti-Diagramm* möglich, vgl. Abbildung 1.1.

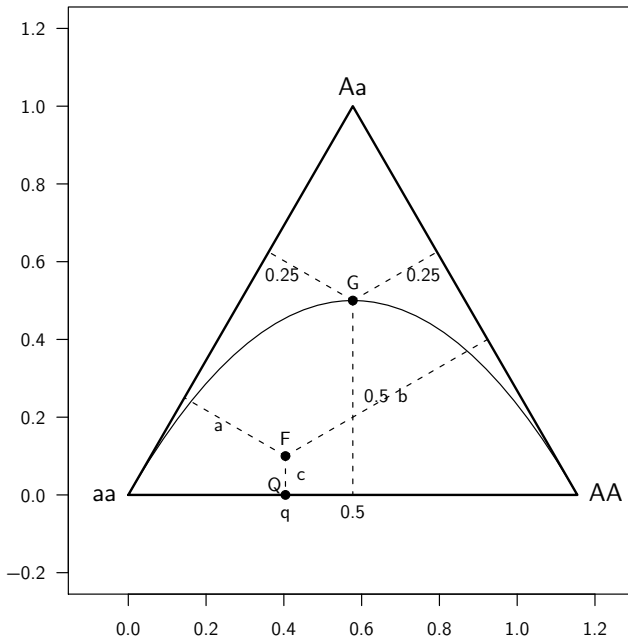


Abb. 1.1: Das nach dem italienischen Statistiker Bruno de Finetti (1906–1985) benannte de Finetti-Diagramm stellt die relativen Genotyphäufigkeiten $P(AA) = \pi_1$, $P(aa) = \pi_3$ und $P(Aa) = \pi_2$ in einer biallelischen, diploiden Population als Längen der Lote a , b und c von einem inneren Punkt F eines gleichseitigen Dreiecks auf die Seiten dar. Dies ist sinnvoll, da sich diese Längen von jedem inneren Punkt aus zur Höhe des Dreiecks, die auf 1 gesetzt wird, summieren. Das Verhältnis der Streckenlänge $\overline{aa, Q}$ zur Seitenlänge $\overline{aa, AA}$ gibt die relative Allelhäufigkeit q von A an. Die Populationen im Hardy-Weinberg-Gleichgewicht liegen auf dem Unterraum der Parabel $2q(1 - q)$; Punkt G stellt beispielsweise eine solche Population mit $q = 0.5$ dar, während in der Population F deutlich weniger Heterozygote Aa als im Hardy-Weinberg-Gleichgewicht erwartet vorhanden sind.

Häufig stellt sich nun die Frage, ob sich eine bestimmte Population bzgl. eines bestimmten Genotypes im Hardy-Weinberg-Gleichgewicht befindet. Im Rahmen einer Studie in Island wurden beispielsweise $n = 747$ Personen zufällig ausgewählt und auf eine bestimmte Blutgruppenkategorie hin untersucht. Die Häufigkeiten der drei Genotypen AA , Aa und aa ergaben sich zu $x_1 = 233$, $x_2 = 385$ und $x_3 = 129$. Einerseits interessiert man sich nun für die statistische Schätzung des unbekannten Parameters q unter Annahme des Hardy-Weinberg-Gleichgewichts in der zugrundeliegenden Population. Wir werden verschiedene Verfahren kennenlernen, wie man diesen Parameter und die daraus abgeleiteten Parameter π_1 , π_2 und π_3 schätzen kann. Andererseits möchte man

Tab. 1.2: Verteilung der positiven Testergebnisse beim Screening von 196 Darmkrebspatienten

| | | | | | | | |
|--|---|----|----|----|----|----|----|
| Anzahl k der positiven Tests | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Häufigkeit Z_k | ? | 37 | 22 | 25 | 29 | 34 | 49 |

auf Grundlage der vorliegenden Daten untersuchen, ob sich die Population tatsächlich im Hardy-Weinberg-Gleichgewicht befindet. Auch hier werden wir verschiedene statistische Verfahren kennenlernen, die diese Frage beantworten.

1.2.5 Schätzung von diagnostischen Kenngrößen

In der Medizin versteht man unter *Screening* eine Untersuchung, die als Reihenuntersuchung bei möglichst vielen Menschen eine möglichst frühe Angabe zur Wahrscheinlichkeit des Vorliegens von bestimmten Krankheiten (oder Risikofaktoren) ermöglichen soll und somit meist als „Vorsorgeuntersuchung“ bezeichnet wird, obgleich beim Vorliegen auffälliger Werte erst durch nachfolgende diagnostische Untersuchungen die Früherkennung von Krankheiten möglich ist.

Das verwendete Testverfahren soll hohe

$$\text{Sensitivität} = P(\text{Test positiv} \mid \text{Proband krank})$$

$$\text{und Spezifität} = P(\text{Test negativ} \mid \text{Proband gesund})$$

aufweisen, d. h. der Test soll die gesuchte Erkrankung (bzw. die bestehenden Risikofaktoren) mit möglichst großer Sicherheit nachweisen oder ausschließen können. Die Untersuchung muss außerdem zeit- und kostengünstig sein und sollte die Probanden möglichst wenig belasten, um in der Bevölkerung akzeptiert zu werden.

Besonders sinnvoll sind Screening-Untersuchungen zur Erkennung von Krankheiten, die rechtzeitig erkannt deutlich besser zu behandeln sind als im Spätstadium, wie z. B. Darmkrebs. In Australien wurden im Rahmen einer Studie 38 000 Patienten an sechs aufeinanderfolgenden Tagen mit einem einfachen diagnostischen Test (einer Stuhluntersuchung) auf dieses Karzinom untersucht. Bei den 3000 Patienten, die mindestens ein positives Testergebnis hatten, wurde mit einer Darmspiegelung das Testergebnis überprüft. Man erhielt eine Häufigkeitsverteilung für die Anzahl der positiven Testergebnisse Z_k bei 196 Krebskranken, vgl. Tabelle 1.2.

Man beachte, dass die Häufigkeit der kranken Patienten mit keinem positiven Testergebnis aufgrund des gewählten Studiendesigns nicht erfasst werden konnte. Die damit eng verbundene

$$\text{Falsch-Negativ-Rate} = P(\text{Test negativ} \mid \text{Proband krank})$$

ist aber von zentralem Interesse und kann mit statistischen Methoden geschätzt werden.

1.2.6 Schätzung von relativem Risiko aus Krebsregisterdaten

Krebsregister sammeln Informationen zur Inzidenz oder Mortalität von bestimmten Krebserkrankungen. Beispielsweise beschreibt der vorliegende Datensatz die Inzidenz von Lippenkrebs bei Männern in 56 administrativen Regionen von Schottland im Zeitraum 1975 bis 1980. Die absolute Anzahl der Fälle variiert häufig sehr stark, einerseits durch Unterschiede in der zugrundeliegenden Bevölkerungszahl, andererseits aber möglicherweise auch durch eine unterschiedliche Altersverteilung in den einzelnen Regionen. Beispielsweise wird man in einer Region mit eher junger Bevölkerung weniger Krebsfälle erwarten als in einer Region mit einer eher älteren Bevölkerung, selbst wenn die Bevölkerungszahl in den beiden Regionen gleich groß ist. Daher berechnet man häufig durch Altersstandardisierung die Anzahl der erwarteten Fälle in den einzelnen Regionen, wobei hierbei die unterschiedliche Bevölkerungszahl und Altersverteilung berücksichtigt wird. Der Quotient der beobachteten zur erwarteten Anzahl, die sogenannte *standardisierte Inzidenzrate* (SIR), wird dann häufig zum Vergleich des zugrundeliegenden relativen Risikos verwendet. Abbildung 1.2 zeigt für unser Beispiel die geographische Variation der SIRs in Schottland. Die einzelnen SIRs sind aber bei seltenen Krankheiten häufig recht unzuverlässig, wenn die zugrundeliegende Anzahl der Fälle sehr klein ist. Im Extremfall wurde möglicherweise gar kein Fall beobachtet, dann ist SIR gleich 0, dies ist bei dem vorliegenden Datensatz in zwei Regionen der Fall. Generell werden extreme Werte der SIRs eher in bevölkerungsarmen als in bevölkerungsreichen Regionen zu erwarten sein. Dies ist auch aus Abbildung 1.3 ersichtlich, in der die SIRs gegen die Anzahl der erwarteten Fälle aufgetragen sind, hier variieren die SIRs bei einer kleineren Anzahl an erwarteten Fällen mehr. Daher ist es wichtig, mit statistischen Methoden zusätzlich Intervalle für das zugrundeliegende relative Risiko in den einzelnen Regionen zu berechnen, die plausible Werte für das wahre relative Risiko angeben. Weiterhin können statistische Methoden auch die Frage beantworten, ob das zugrundeliegende relative Risiko in allen Regionen gleich ist oder ob es tatsächlich Heterogenität in dem Auftreten der Krankheit gibt.

1.2.7 Statistische Analyse einer Therapiestudie

In einer randomisierten kontrollierten Studie („randomized controlled trial“, RCT) sollte die Frage untersucht werden, ob der Zusatz von drei Fischmahlzeiten pro Woche zu einer cholesterinarmen Kost zu einer stärkeren Senkung des Cholesterinspiegels führt als die unbelassene cholesterinarme Kost. Als Zielgröße wurde die Änderung des Gesamt-Cholesterinspiegels 28 Tage nach Beginn der Diät verwendet. Die Patienten wurden zufällig auf die beiden Behandlungsgruppen verteilt. Da ein Einfluss der Gesamtkalorienzufuhr auf den Effekt des Fischzusatzes nicht ausgeschlossen werden konnte, wurden die Patienten in zwei weitere Gruppen eingeteilt: normalgewichtige und über-

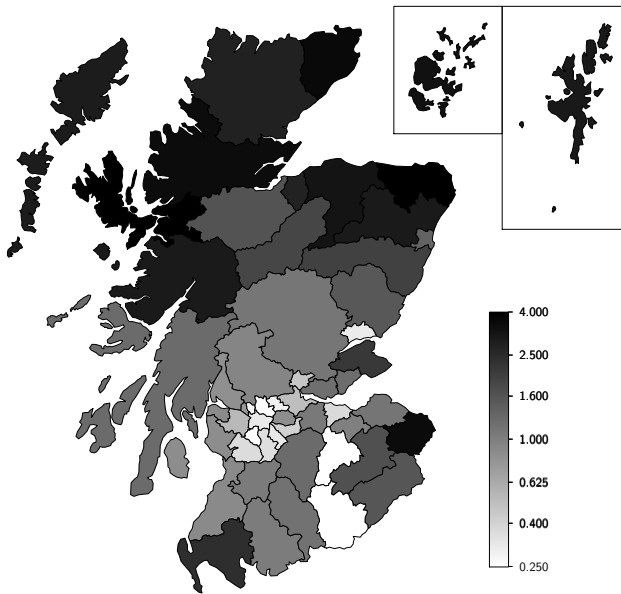


Abb. 1.2: Die geographische Verteilung der standardisierten Inzidenzraten (SIR) von Lippenkrebs in Schottland. Man beachte, dass einzelne SIRs sogar außerhalb des Skalenbereiches $[0.25, 4]$ liegen. Diese sind mit weiß bzw. schwarz markiert.

gewichtige Patienten, wobei für letztere eine kalorienreduzierte Kost vorgesehen war. Die Daten sind in Tabelle 1.3 zusammengefasst. Es stellt sich die Frage, ob die empiri-

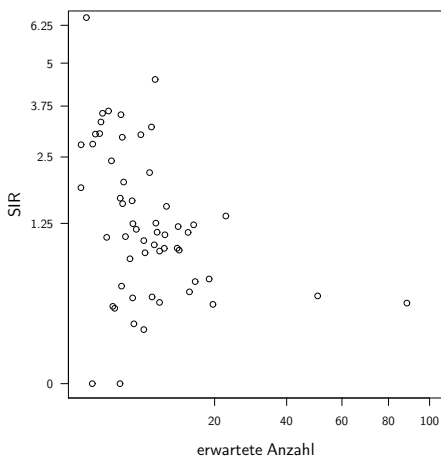


Abb. 1.3: Plot der standardisierten Inzidenzraten (SIR) gegen die erwartete Anzahl an Fällen von Lippenkrebs. Beide Variablen sind zur besseren Verdeutlichung auf einer wurzeltransformierten Skala aufgetragen.

Tab. 1.3: Senkung des Cholesterinspiegels 28 Tage nach Beginn der Diät (in mg pro 100 ml) in der Fischstudie

| Gewicht | Behandlung | Anzahl Patienten | Cholesterinsenkung | |
|-----------------|------------|------------------|--------------------|--------------------|
| | | | Mittelwert | Standardabweichung |
| Normalgewichtig | Ohne Fisch | 97 | 23.2 | 32.05 |
| Normalgewichtig | Mit Fisch | 98 | 21.0 | 30.43 |
| Übergewichtig | Ohne Fisch | 50 | 40.8 | 34.23 |
| Übergewichtig | Mit Fisch | 63 | 34.6 | 37.04 |

schen Unterschiede in der mittleren Cholesterinsenkung in den einzelnen Gruppen nur zufällig zustande kam, oder ob es tatsächlich Unterschiede in den zugrundeliegenden Behandlungen gibt.

1.2.8 Statistische Analyse von Überlebenszeiten

Die vorliegenden Daten beschreiben die Überlebenszeit (in Tagen) von 184 Patienten, die an Primärer biliärer Zirrhose (PBC) litten. Die Primär biliäre Zirrhose ist eine relativ seltene Autoimmunerkrankung der Leber, die in ca. 90% der Fälle Frauen betrifft. Unbehandelt kann eine PBC zum Tode führen, allerdings ist der Verlauf von Patient zu Patient sehr unterschiedlich. In der zugrunde liegenden randomisierten Studie wurde ein neues Medikament mit einer Placebo-Behandlung verglichen. Wir werden nur die Placebo-Gruppe betrachten, in der 94 Patienten eingeschlossen waren, vgl. Tabelle 1.4. Von diesen waren 47 (50%) *zensierte Beobachtungen*, d. h. von diesen 47 Patienten war nur bekannt, dass sie bis zu einem gewissen Zeitpunkt (der angegebenen Überlebenszeit) überlebt haben. Zu diesem Zeitpunkt sind sie jedoch nicht an der betrachteten Krankheit gestorben. In Abbildung 1.4 ist die Datensituation veranschaulicht.

1.3 Statistische Modelle

Einen zentralen Platz in der statistischen Analyse von Daten bildet die Formulierung eines geeigneten stochastischen Modells, dem die beobachteten Daten zugrundeliegen. Man spricht auch von einem *statistischen Modell*. Meist beinhaltet das statistische Modell unbekannte Parameter, die aus den Daten geschätzt werden. In diesem Buch werden wir uns ausschließlich mit *parametrischen Modellen* beschäftigen, bei der die Anzahl der unbekannten Parameter fest ist. Bei *nichtparametrischen Modellen* hängt

Tab. 1.4: Überlebenszeiten der 94 Patienten unter Placebo-Behandlung in Tagen. Zensierte Werte sind mit einem Plus gekennzeichnet – hier ist unbekannt, wie lange der Patient darüber hinaus gelebt hat.

| | | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 8+ | 9 | 38 | 96 | 144 | 167 | 177 | 191+ | 193 | 201 |
| 207 | 251 | 287+ | 335+ | 379+ | 421 | 425 | 464 | 498+ | 500 |
| 574+ | 582+ | 586 | 616 | 630 | 636 | 647 | 651+ | 688 | 743 |
| 754 | 769+ | 797 | 799+ | 804 | 828+ | 904+ | 932+ | 947 | 962+ |
| 974 | 1113+ | 1219 | 1247 | 1260 | 1268 | 1292+ | 1408 | 1436+ | 1499 |
| 1500 | 1522 | 1552 | 1554 | 1555+ | 1626+ | 1649+ | 1942 | 1975 | 1982+ |
| 1998+ | 2024+ | 2058+ | 2063+ | 2101+ | 2114+ | 2148 | 2209 | 2254+ | 2338+ |
| 2384+ | 2387+ | 2415+ | 2426 | 2436+ | 2470 | 2495+ | 2500 | 2522 | 2529+ |
| 2744+ | 2857 | 2929 | 3024 | 3056+ | 3247+ | 3299+ | 3414+ | 3456+ | 3703+ |
| 3906+ | 3912+ | 4108+ | 4253+ | | | | | | |

die Anzahl der unbekannten Parameter von der Anzahl der Beobachtungen ab, oder ist sogar unendlich groß.

Die Formulierung eines statistischen Modells geschieht über inhaltliche Überlegungen. Häufig nimmt man dabei Approximationen in Kauf, die eine gewisse Vereinfachung der Problemstellung ermöglichen. Ein typisches Beispiel ist die Anteilsschätzung, bei der aus einer Grundgesamtheit vom Umfang N eine Stichprobe vom Umfang n ohne Zurücklegen gezogen wird. Korrekt wäre hier die Annahme einer hypergeometrischen Verteilung für die Anzahl der Einheiten in der Stichprobe mit einer gewissen Eigenschaft, beispielsweise blauen Kugeln aus einer Urne. Nimmt man an, dass N im Vergleich zu n sehr groß ist, so kann man die hypergeometrische durch eine Binomial-Verteilung approximieren, die einfacher zu handhaben ist, aber annimmt, dass die

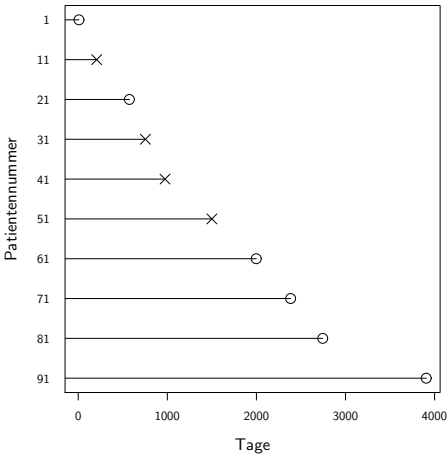


Abb. 1.4: Veranschaulichung der Datensituation bei teilweise zensierten Überlebenszeiten anhand der 10 Beobachtungen aus der ersten Spalte von Tabelle 1.4. Nur bei mit einem Kreuz markierten Endpunkten ist der Patient an PBC gestorben, alle anderen sind aus anderen Gründen aus der Studie ausgeschieden. Dies ist mit einem Kreis gekennzeichnet; bekannt ist bei diesen zensierten Zeiten nur, dass die wahre Überlebenszeit länger war.

Stichprobe *mit Zurücklegen* erhoben wurde. Der Unterschied zwischen beiden statistischen Modellen ist aber unter der oben getroffenen Annahme vernachlässigbar. Eine solches binomiales statistisches Modell werden wir zur Anteilsschätzung wie auch zum Vergleich von Anteilen annehmen.

Bei der Rückfangmethode ist N unbekannt, ja ist von zentralem Interesse. Daher ist hier die Annahme des hypergeometrischen Modells naheliegend, da einerseits eine Approximation, deren Gültigkeit unklar ist, vermieden wird, andererseits die Größe, die wir aus den Daten schätzen wollen, durch einen Parameter im hypergeometrischen Modell repräsentiert wird.

Ein statistisches Modell kann auch mit statistischen Methoden hinterfragt werden. So werden wir Verfahren kennenlernen, die es erlauben zu untersuchen, ob eine Population, aus der wir eine Stichprobe gezogen haben, sich im Hardy-Weinberg-Gleichgewicht befindet oder nicht. Bei der Analyse der Daten aus der Fischstudie kann man sich ebenfalls mindestens zwei Modelle vorstellen: in einem hat der zusätzliche Konsum von Fisch keinen Einfluss auf das Ausmaß der Cholesterinsenkung, so dass sich die Behandlungsgruppen mit und ohne Fischmahlzeiten nicht unterscheiden. In dem anderen gibt es einen Effekt des Fischkonsums, hier werden sich die zwei Behandlungsgruppen zumindest durch unterschiedliche Mittelwerte unterscheiden. Zufällige Variation zwischen den einzelnen Probanden muss ebenfalls berücksichtigt werden, wobei man üblicherweise von normalverteilten Beobachtungen mit unbekannter Varianz ausgeht.

Bei der Analyse von (stetigen) Überlebenszeiten ist eine Normalverteilungsannahme fragwürdig, sind doch Überlebenszeiten immer positive Zahlen. Stetige Verteilungen mit positivem Träger, wie etwa die Exponential- oder die Gammaverteilung erscheinen hier auf den ersten Blick adäquat.

Auch bei der Analyse von Zähldaten, wie etwa die Anzahl der Lippenkrebsfälle in Beispiel 1.2.6, muss man eine passende Verteilung wählen. Im einfachsten Fall wählt man eine Poisson-Verteilung, die aber annimmt, dass Erwartungswert und Varianz übereinstimmen. In vielen Fällen ist aber die Varianz größer als der Erwartungswert, man spricht von *Überdispersion*. Eine Verteilungsklasse für Zähldaten, die dies zulässt, ist die negative Binomialverteilung.

Statistische Modelle können aber noch deutlich komplexer werden, wenn die zu analysierenden Daten dies erfordern. So muss man bei der Analyse von Überlebenszeiten auch zensierte Beobachtungen berücksichtigen. Unter geeigneten Modellannahmen ist dies durchaus möglich. Ähnlich verhält es sich bei Beispiel 1.2.5, bei dem durch das gewählte Studiendesign Darmkrebspatienten mit keinem positivem Testergebnis nicht vorkommen können. Die Schätzung des Anteils dieser Darmkrebspatienten an der gesamten Anzahl an Darmkrebspatienten, die Falsch-Negativ-Rate des Tests, ist hier sogar von zentralen Interesse.

1.4 Inhalt und Notation der Kapitel

In Kapitel 2 wird zunächst der zentrale Begriff der Likelihoodfunktion eingeführt. Die klassische Likelihood-Inferenz wird in den Kapiteln 3 und 4 besprochen, diese ermöglicht das Studium der frequentistischen Eigenschaften des Maximum-Likelihood-Schätzers und zugehöriger Konfidenzintervalle. Komplementär dazu agiert der Bayes-Ansatz, der es durch Annahme einer Priori-Verteilung erreicht, die Unsicherheit in der Schätzung der Daten zu quantifizieren. Kapitel 5 beschreibt hierzu das grundsätzliche Vorgehen, Kapitel 6 diskutiert den Einsatz moderner numerischer Methoden in der Bayes-Inferenz. Die abschließenden Kapitel 7 und 8 geben Einführungen in das Modellwahl- bzw. das Prognoseproblem, wobei sowohl klassische als auch Bayesianische Ansätze besprochen werden. Jedes Kapitel wird mit Aufgaben und Hinweisen auf weiterführende Literatur abgeschlossen.

An dieser Stelle sollen nun auch einige notationelle Erläuterungen folgen. Mathematische Sätze werden immer durch kursive Schrift hervorgehoben. Das Ende eines Beweises wird mit einem offenen Quadrat (\square) markiert, während ein geschlossenes Quadrat (\blacksquare) den letzten Absatz eines Beispiels beendet. Definitionen werden mit einer Raute (\blacklozenge) abgeschlossen.

Durch Fettdruck werden vektorielle Parameter $\boldsymbol{\theta}$ von skalaren Parametern θ unterschieden. Analog werden univariate Zufallsvariablen X_i zu einer Stichprobe $X = (X_1, \dots, X_n)$ zusammengefasst, während n multivariate Zufallsvariablen $\mathbf{X}_i = (X_{i1}, \dots, X_{ik})^T$ (auch Zufallsvektoren genannt) als $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ abgekürzt werden.

Abkürzungen für gängige Verteilungen finden sich im Anhang A.3. So bezeichnet beispielsweise $X \sim \text{Bin}(n, \pi)$ eine binomialverteilte Zufallsvariable mit Parametern n und π , siehe Tabelle A.1.

Literatur

Die Schätzung und der Vergleich von Anteilen ist ausführlich in Connor und Imrey (2005) besprochen. Für den Einsatz von Rückfangmethoden siehe etwa Seber (1982). Das Hardy-Weinberg-Gleichgewicht ist beispielsweise in Lange (2002) diskutiert, die Daten stammen aus Falconer und Mackay (1996). Der Datensatz zum Lippenkrebs in Schottland ist Clayton und Bernardinelli (1992) entnommen, der Datensatz zum

Darmkrebs-Screening ist aus Lloyd und Frommer (2004). Die Daten zur Fischstudie stammen aus Schumacher und Schulgen (2002). Kirkwood und Sterne (2003) beschreiben die randomisierte Studie zur Behandlung der primären biliären Zirrhose.

Methoden der statistischen Inferenz

Likelihood und Bayes

Held, L.

2008, XII, 304 S. 60 Abb., Softcover

ISBN: 978-3-8274-1939-2