# LERNZIELE

Die statistische Analyse und das statistische Verständnis bilden die Grundlage für datengestützte Methoden und

Ansätze des maschinellen Lernens. Statistik: Inferenzstatistik gibt eine gründliche Einführung

zu Punktschätzern und erörtert verschiedene Techniken zur Schätzung und Optimierung von Parametern. Besonderes Augenmerk wird auf eine detaillierte Diskussion von statistischen und systematischen Unsicherheiten sowie der Ausbreitung von Unsicherheiten gelegt. Die Bayessche Statistik ist für datengestützte Ansätze von grundlegender Bedeutung, und dieser Kurs befasst sich eingehend mit Bayesschen Techniken wie der Bayesschen Parameterschätzung und vorherigen Wahrscheinlichkeitsfunktionen. Darüber hinaus gibt dieser Kurs einen detaillierten Überblick über statistische Tests und Entscheidungstheorie mit folgenden Schwerpunkten

Aspekte wie A/B-Tests, Hypothesentests, p-Werte und Mehrfachtests, die für statistische Analyseansätze in einem breiten Spektrum praktischer Anwendungen grundlegend sind.

Der Inhalt dieses Lehrbuchs vermittelt Ihnen ein Verständnis für Punktschätzungsmethoden,

Anwendung der Maximum-Likelihood-Methode und der Methode der gewöhnlichen kleinsten Quadrate zur Schätzung der Parameter,

das Konzept der statistischen und systematischen Fehler zu verstehen, Methoden der Fehlerfortpflanzung anzuwenden, Bayes'sche Inferenz und nicht-parametrische Techniken zu nutzen, statistische Tests zu bewerten und die Grundlagen der statistischen Entscheidungstheorie zu verstehen.

T 1

# Einheit 1

# PUNKTSCHÄTZUNG

STUDIENZIELE

Nach Abschluss dieser Einheit werden Sie gelernt haben...

- wie man Parameter mit Hilfe der Methode der Momente, der maximalen Wahrscheinlichkeit, der gewöhnlichen kleinsten Quadrate und der Wiederholungsstichprobenverfahren schätzt.

- wie man feststellt, ob eine Statistik für die Schätzung eines Parameters ausreichend ist.

- die Definitionen und Interpretationen der Likelihood-, Log-Likelihood- und negativen Log-Likelihood-Funktionen.

- die Annahmen, die zur Bestimmung der gewöhnlichen Kleinstquadrat-Schätzungen für Modell-/Funktionsparameter erforderlich sind.

- wie man die Bootstrap- und Jackknife-Techniken für Punktschätzungen verwendet.

- wie man die Unsicherheiten mit Hilfe von Bootstrap- und Jackknife-Techniken schätzt.

# 1. PUNKTESCHÄTZUNG

### Einführung

Angenommen, wir haben eine Stichprobe von 100 Zahlen, die aus einer Exponentialverteilung stammen.

Erinnern Sie sich daran, dass die Exponentialverteilung durch ihren Ratenparameter λ charakterisiert ist. Mit anderen Worten: Sobald wir diesen Parameter kennen, wissen wir alles, was es über die Verteilung zu wissen gibt. Wie können wir die Stichprobe verwenden, um einen Schätzwert für λ zu finden? Wie können wir nach der Ermittlung des Schätzwerts die Qualität der von uns verwendeten Methode (Schätzer) bewerten? Ist mit dieser Methode eine große Unsicherheit verbunden? Können wir die Stichprobe wegwerfen, wenn wir unsere Schätzung haben, oder werden die einzelnen Datenpunkte noch weitere Informationen liefern? Diese Einheit wird uns helfen, solche Fragen zu beantworten.

Bei statistischen Schlussfolgerungen geht es darum, aus den Informationen einer beobachteten Stichprobe Rückschlüsse auf die Grundgesamtheit zu ziehen, aus der die Stichprobe entnommen wurde. Populationen werden durch numerische Maße, die sogenannten Parameter, charakterisiert. Das Ziel der Punktschätzung besteht darin, die relevanten Parameter zu schätzen. Viele Ergebnisse aus der Wahrscheinlichkeitsrechnung spielen eine wichtige Rolle bei den Werkzeugen, die wir für statistische Schlussfolgerungen entwickeln und verwenden. Daher werden wir die relevanten Ergebnisse aus der Wahrscheinlichkeitsrechnung überprüfen und in Erinnerung rufen, wo es angebracht ist.

Im ersten Abschnitt lernen wir, wie man die Methode der Momente verwendet, um Punktschätzungen der interessierenden Parameter zu finden. Bei dieser Methode wird der interessierende Parameter mit den Momenten der zugrundeliegenden Verteilung in Beziehung gesetzt, und dann werden die zugehörigen Stichprobenmomente verwendet, um den Schätzer zu erstellen. Im nächsten Abschnitt lernen wir, wie man herausfindet, ob eine aggregierte Größe (eine Statistik) auf der Grundlage der Daten alle relevanten Informationen über den interessierenden Parameter erfasst, den wir schätzen wollen. Eine solche Größe wird als hinreichende Statistik bezeichnet.

In Abschnitt 1.3 entwickeln wir eine alternative Methode zur Schätzung von Parametern von Interesse: die Methode der maximalen Wahrscheinlichkeit (maximum likelihood). Auf einer hohen Ebene zielt diese Methode darauf ab, eine Schätzung des interessierenden Parameters zu finden, indem die Wahrscheinlichkeit für die Beobachtung der gegebenen Daten maximiert wird. Wir untersuchen eine der am häufigsten verwendeten Funktionen in der Statistik, die Likelihood-Funktion, und wie diese die Strategie zur Ermittlung der Punktschätzung quantifiziert.

Im nächsten Abschnitt werden wir die allgemeine Idee der gewöhnlichen kleinsten Quadrate vorstellen. Sie haben diese Methode vielleicht schon bei einer einfachen Regression gesehen. Der Vorteil dieser Methode gegenüber der Maximum-Likelihood-Methode ist, dass wir nichts über die Verteilung wissen müssen, die die gegebenen Daten erzeugt hat. Stattdessen müssen wir nur die funktionale oder Modellabhängigkeit kennen.

In Abschnitt 1.5 schließlich werden zwei gängige Wiederholungsstichprobenverfahren untersucht: der Bootstrap und das Jackknife. Obwohl ihre Anwendung sehr vielfältig ist, konzentrieren wir uns darauf, wie man die gegebenen Daten auf verschiedene Weise nutzen kann, um Schätzungen für die interessierenden Parameter zu erhalten, ohne die zugrunde liegende Verteilung zu kennen. Die beiden Vorteile dieser Techniken bestehen darin, dass sie (i) bei kleinen Stichprobengrößen, bei denen der zentrale Grenzwertsatz möglicherweise nicht gilt, recht gut funktionieren und (ii) Schätzungen der mit den Schätzungen verbundenen Unsicherheiten liefern. Zusammenfassend lässt sich sagen, dass diese Einheit eine Vielzahl von Möglichkeiten aufzeigt, wie Sie Stichprobendaten verwenden können, um Punktschätzungen (einzelne Zahlen) von unbekannten Größen, die die Daten beschreiben, zu berechnen.

## 1.1 Methode der Momente

Die Methode der Momente ist eine der einfachsten Techniken zur Ableitung von Punktschätzern. Wie der Name schon sagt, bezieht sie sich auf die Momente einer Zufallsvariablen. X sei eine Zufallsvariable; ihr erstes Moment ist einfach ihr Erwartungswert μ = μ 1 = E X . Ihr zweites Moment ist der Erwartungswert ihres Quadrats μ 2 = E X2 . Im Allgemeinen ist das k-te Moment gegeben durch

XXX

Beginnen wir mit dem ersten Moment und ersetzen den Erwartungswert E durch den Durchschnitt wie folgt.

Man nehme n Kopien der Zufallsvariablen X und bezeichne diese Kopien mit X1,X2, ...,Xn.

Das (erste) Stichprobenmoment ist dann

XXX

Diese Gleichung ist ein Schätzer des ersten Moments. Ein Schätzer ist eine Zufallsvariable, die darauf abzielt, einen unbekannten, aber nicht zufälligen Parameter zu schätzen. Wenn wir n Realisierungen von X, einer Zufallsstichprobe, beobachtet haben, die durch x1, x2, ..., xn gegeben sind, dann können wir die Schätzung des ersten Moments auf der Grundlage dieser Daten berechnen, indem wir jedes Xi durch xi ersetzen, um die folgende Gleichung zu erhalten:

XXX

Es ist zu beachten, dass der Schätzer m 1 eine Zufallsvariable ist, während die Schätzung m 1 nicht zufällig ist.

Wir können analoge Gleichungen für die Schätzer und die Schätzungen der höheren Momente erhalten. Der Schätzer für das k-te Moment der Stichprobe lautet

XXX

und die entsprechende Schätzung auf der Grundlage der beobachteten Daten lautet

XXX

#### Beispiel 1.1.1

Berechnen Sie die ersten und zweiten Stichprobenmomente der beobachteten Daten, die in der nachstehenden Tabelle aufgeführt sind.

Tabelle 1: Beispieldaten für Beispiel 1.1.1

XXX

#### Lösung

Mit diesen Daten werden wir Gleichungen für m 1 und m k verwenden. Das erste Stichprobenmoment ist

xxx

Das zweite Stichprobenmoment ist

xxx

Eine der Hauptaufgaben in diesem Abschnitt besteht darin, einen oder mehrere Parameter einer Verteilung zu schätzen.

Wir müssen einen Weg finden, um die unbekannten Parameter mit einem oder mehreren Momenten in Beziehung zu setzen.

Danach können wir die Stichprobenmomente verwenden, um den oder die unbekannten Parameter zu schätzen. Beginnen wir zu diesem Zweck mit einem einfachen Beispiel.

#### Beispiel 1.1.2

X1, X2, ..., Xn seien eine Zufallsstichprobe (unabhängige Variablen) aus der Gleichverteilung U 0, θ , wobei θ unbekannt ist. Verwenden Sie die Methode der Momente, um einen Schätzer für θ zu finden. Verwenden Sie anschließend die unten angegebenen Daten, um θ mit Hilfe des gefundenen Schätzers zu schätzen.

Tabelle 2: Beispieldaten für Beispiel 1.1.2

xxx

#### Lösung

Es sei daran erinnert, dass für X U a, b das erste Moment μ 1 = E X = a + b 2 ist. In diesem Fall gilt

μ 1 = θ2 und das Ersetzen von μ 1 m 1 ergibt θ = 2m. Der Schätzer, den wir erhalten, ist daher

xxx

Anhand der beobachteten Daten können wir die Schätzung wie folgt berechnen

xxx

Wir wollen die Leistungsfähigkeit des Momentenschätzers für verschiedene Stichprobengrößen untersuchen:

10, 100, 1, 000, 10, 000, 100, 000

Wir simulieren 100 Stichproben aus U 0, 5 entsprechen jeder dieser Stichproben und berechnen die Momentenmethode-Schätzung für θ aus Beispiel 1.1.2. Betrachten Sie die folgende Abbildung und beachten Sie die Variation der Schätzungen für verschiedene Stichprobengrößen. In der ersten Darstellung ist jeder Punkt eine Momentenschätzung für θ, wobei der Wert dieser Schätzung auf der vertikalen Achse und der Umfang der zur Schätzung verwendeten Stichprobe auf der horizontalen Achse steht. Die verschiedenen Stichprobengrößen sind farblich kodiert. Man beachte, wie die Punkte für N = 10 verstreut sind, für N = 100 weniger verstreut, und schließlich, wenn N erhöht wird, gruppieren sich die 100 Schätzungen sehr eng um den wahren Wert von 5.

Im mittleren Diagramm sind die Histogramme dieser Punkte dargestellt. Der Wert der Schätzung befindet sich auf der horizontalen Achse und die Häufigkeit der Bins auf der vertikalen Achse. Beachten Sie, dass Werte, die weit von der Mitte entfernt sind, bei größeren Werten von N weniger wahrscheinlich sind, und die Form der Verteilung ist symmetrisch und ähnelt der Gauß-Verteilung.

In der abschließenden Darstellung berechnen wir die Stichprobenvarianz jeder der 100 Schätzungen aus jeder der Stichprobengrößen. Mit anderen Worten, jeder Punkt stellt die Stichprobenvarianz von 100 Schätzungen von θ dar, wobei die vertikale Achse die Werte der Stichprobenvarianz und die horizontale Achse die Anzahl der Punkte angibt, die zur Erzeugung jeder der 100 Schätzungen verwendet wurden. Wie Sie sehen können, ist die Stichprobenvarianz klein, wenn N groß ist, was das Clusterverhalten bestätigt, das wir in den beiden obigen Diagrammen beobachtet haben.

Abbildung 1: Momentenmethode-Schätzungen für θ aus U 0, θ . Wahres θ = 5 (Streudiagramm von

Einschätzungen)

Abbildung 2: Momentenmethode-Schätzungen für θ aus U 0, θ . Wahres θ = 5 (Histogramme von

Einschätzungen)

Abbildung 3: Momentenmethode-Schätzungen für θ aus U 0, θ . Wahr θ = 5 (Varianz von

Einschätzungen)

Das nächste Beispiel ist die geometrische Verteilung. Erinnern Sie sich, dass die geometrische Verteilung die Anzahl der Fehlschläge modelliert, bevor der erste Erfolg eintritt. Wenn X geometrisch p ist, dann ist ihr Erwartungswert E X = 1 - p p .

#### Beispiel 1.1.3

X1,X2, ...,Xn seien iid von Geometric p . Finden Sie den Momentenschätzer für p.

Verwenden Sie die folgenden Daten, um den Schätzwert für p zu ermitteln. Erinnern Sie sich, dass iid für "unabhängig und identisch verteilt" steht.

Tabelle 3: Beispieldaten für Beispiel 1.1.3

xxx

#### Lösung

Wie bereits erwähnt, ist E X = 1 - p p für X Geometrisch p . Somit ergibt der Stichprobenmomentschätzer

xxx

Unter Verwendung der gegebenen Daten ist die Schätzung für p gegeben durch

xxx

Ähnlich wie bei der Simulation für Beispiel 1.1.2 simulieren wir Sätze von 100 Stichproben mit unterschiedlichem Stichprobenumfang und berechnen die Methode der Momentenschätzung aus Beispiel 1.1.3. Untersuchen Sie die nachstehende Abbildung und beachten Sie die Merkmale der Methode der Augenblicksschätzungen für verschiedene Stichprobengrößen.

Abbildung 4: Momentenmethode-Schätzungen für θ aus geometrischem p . Wahres p = 0,4(Streuung

Plot der Schätzungen)

Abbildung 5: Momentenmethode-Schätzungen für θ aus geometrischem p . Wahres p = 0,4

(Histogramme von Schätzungen)

Abbildung 6: Momentenmethode-Schätzungen für θ aus geometrischem p . Wahres p = 0,4 (Varianz

von Schätzungen)

In allen bisher besprochenen Beispielen war das erste Moment ausreichend, um einen Schätzer für den unbekannten Parameter zu erhalten. Wir erörtern nun ein Beispiel, bei dem wir das zweite Moment verwenden müssen. Wir erinnern uns, dass die Varianz einer Zufallsvariablen V X = E X - E X = E X2 - E X 2 ist. Wenn X einer Gaußverteilung folgt, X N μ, σ , dann ist E X = μ und V X = σ2.

#### Beispiel 1.1.4

Sei X1, . . . ,Xn seien iid aus N 0, σ . Finden Sie den Momentenschätzer für σ2.

#### Lösung

Da μ = 0 ist, haben wir σ2 = E X2 ; daher ist der Momentenschätzer für σ

Xxx

#### Beispiel 1.1.5

Seien X1, ...,Xn iid aus N μ, σ mit unbekanntem μ und unbekanntem σ. Finden Sie den Momentenschätzer für μ und σ.

#### Lösung

Wir wissen, dass für X N μ, σ , E X = μ, und V X = σ2. Daher sind die Momentenschätzer der Methode gegeben durch

**Unverzerrter Schätzer**

Ein Schätzer ist dann und nur dann unvoreingenommen, wenn seine

der erwartete Wert mit dem Zielparameter übereinstimmt.

xxx

#### Unvoreingenommenheit

Punktschätzungen bieten nicht nur eine Möglichkeit zur Schätzung unbekannter Parameter, sondern haben auch bestimmte Eigenschaften, anhand derer wir ihre Qualität bewerten. Eine der Dimensionen, anhand derer wir einen Punktschätzer bewerten, ist, ob er **unvoreingenommen ist**. Der Schätzer für θ in U 0, θ aus Beispiel 1.1.2 war gegeben durch

xxx

Der Erwartungswert dieses Schätzers ist

xxx

Daher ist dieser Schätzer unverzerrt.

#### Beispiel 1.1.6

Zeigen Sie, dass der Schätzer für μ aus Beispiel 1.1.5 unverzerrt ist.

#### Lösung

Auch hier brauchen wir nur den Erwartungswert des Schätzers zu berechnen

xxx

Daher ist der Schätzer unverzerrt.

Die Statistik 1nΣi = 1nXi kommt häufig vor. Diese Statistik wird als **Schätzer des Stichprobenmittelwerts** bezeichnet und wird mit X oder Xn bezeichnet. Die letztere Schreibweise wird verwendet, um den Umfang der Stichprobe explizit anzugeben. Kurz gesagt, wir haben definiert

**Schätzer für den Stichprobenmittelwert**

Diese Zufallsvariable schätzt den Mittelwert der Verteilung, aus der eine Stichprobe gezogen wird, indem der Quotient zwischen der Summe der Variablen und dem Stichprobenumfang berechnet wird.

xxx

Zusammenfassend haben wir gesehen, dass X ein unverzerrter Schätzer für μ für die Gaußsche Verteilung ist. Einige Schätzer, die aus der Methode der Momente stammen, sind unverzerrt, andere jedoch nicht.

Bevor wir ein Beispiel für einen verzerrten Schätzer geben, müssen wir einige vorläufige Ergebnisse aus der Wahrscheinlichkeitsrechnung diskutieren.

Die Varianz einer Summe von unabhängigen Zufallsvariablen ist die Summe der Varianzen V X1 + ⋯Xn = V X1 + ⋯ + V Xn . Auch die Varianz von nicht zufälligen Vielfachen einer Zufallsvariablen ist das Quadrat des Vielfachen mal der Varianz der Zufallsvariablen V c . X = c2V X . Mit Hilfe dieser beiden Fakten können wir die Varianz von Xn berechnen.

Xxx

Erinnern Sie sich nun daran, dass V[X] = E[X2] - E[X]2. Daher ist E[X2] = V[X] + E[X]2. Anhand dieser Tatsache lässt sich das zweite Moment von X wie folgt berechnen

xxx

#### Beispiel 1.1.7

Zeigen Sie, dass der Schätzer für σ2 aus Beispiel 1.1.5 verzerrt ist.

**Varianz der Stichprobe**

Diese Zufallsvariable dient als Schätzer für die Varianz einer Verteilung, aus der eine Stichprobe gezogen wird.

#### Lösung

Wir werden den Erwartungswert von σ2 berechnen:

Xxx

Wir haben festgestellt, dass der Schätzer verzerrt ist.

Der Faktor n - 1n macht den Schätzer verzerrt. Wenn wir also den Schätzer mit nn - 1 multiplizieren, wird der resultierende Schätzer unverzerrt sein:

xxx

Dieser unverzerrte Schätzer, Sn2 für σ2, wird als **Stichprobenvarianz** bezeichnet. In der Tat sind die Schätzer Xn und Sn2 unverzerrte Schätzer des Mittelwerts bzw. der Varianz der Grundgesamtheit, unabhängig davon, ob die Verteilung gaußförmig ist oder nicht.

## 1.2 Suffiziente Statistiken

Im vorangegangenen Abschnitt haben wir zwei wichtige Schätzer definiert: den Stichprobenmittelwert (X) und die Stichprobenvarianz (Sn2). Diese beiden Größen fassen alle Informationen über die jeweiligen Parameter (Mittelwert und Varianz der Grundgesamtheit) zusammen, die die Stichprobe enthält. Mit anderen Worten: Wenn wir diese Größen haben, spielen die einzelnen Werte aus den Stichprobendaten keine Rolle mehr, wenn es darum geht, zusätzliche Informationen über die interessierenden Parameter zu erhalten. In diesem Sinne werden die Statistiken X und Sn2 als suffizient (erschöpfend) bezeichnet. Bevor wir definieren, was eine suffiziente Statistik ist, sollten wir uns die Definition einer Statistik in Erinnerung rufen.

**Statistik**

Diese Zufallsvariable ist definiert durch eine Funktion einer Zufallsstichprobe, in der Regel zur Schätzung eines unbekannten Parameters von Interesse.

X1, ...,Xn sei eine Folge von Zufallsvariablen. Eine **Statistik** dieser Folge ist eine Funktion g:ℝn ℝ von X1, ...,Xn. Mit anderen Worten: Wenn U eine Statistik dieser Folge ist, dann ist U = g X1, ...,Xn .

**Suffiziente Statistik**

Diese Statistik enthält alle die Informationen, die eine Zufallsstichprobe in Bezug auf die Schätzung eines unbekannten Parameters von Interesse liefert.

Bei einer Stichprobe X1, . . . , Xn, sind der Stichprobenmittelwert Xn und die Stichprobenvarianz Sn 2 zwei Beispiele für Statistiken dieser Stichprobe. Wir können auch andere Statistiken definieren. Ein Beispiel ist das Maximum Xmax, ein anderes das Minimum Xmin und ein weiteres der Median Xmid.

Alle diese Größen sind Funktionen der Zufallsstichprobe und damit Statistiken. Je nach dem interessierenden Parameter sind einige dieser Größen suffizient und andere nicht. Wir sind nun bereit, suffiziente Statistiken zu definieren (Hogg et al., 2019).

X1, ...,Xn seien iid aus einer Wahrscheinlichkeitsverteilung, die durch einen unbekannten Parameter θ gekennzeichnet ist. Eine Statistik U gilt als **suffizient** für θ, wenn die bedingte Verteilung von X1, ...,Xn bei U unabhängig von θ ist.

#### Beispiel 1.2.1

Angenommen, wir haben eine kleine Stichprobe von nur zwei: X1,X2 iid aus Bernoulli p mit unbekanntem p. Zeigen Sie, dass der Stichprobenmittelwert U = X1 + X2 eine suffiziente Statistik für p ist.

#### Lösung

Da die Variablen unabhängig sind, ist die gemeinsame PMF von X1,X2 einfach das Produkt der einzelnen Dichten:

xxx

Für die PMF von U gilt nun, dass U Binomial 2, p :

xxx

Die gemeinsame bedingte Dichte von X1,X2 U = u ist gegeben durch

Xxx

Da diese gemeinsame bedingte Dichte nicht von p abhängt, ist die Statistik U = X1 + X2 eine suffiziente Statistik für diesen Parameter (p).

Es kann recht schwierig sein, die Verteilung einer Statistik U zu finden. Das macht es schwierig zu wissen, ob U eine suffiziente Statistik für die Schätzung eines unbekannten Parameters θ ist. Um diese Herausforderung zu umgehen, werden wir ein Ergebnis einführen, das uns nicht nur helfen wird zu bestimmen, ob eine gegebene Statistik für die Schätzung eines Parameters suffiziente ist, sondern uns auch helfen kann, eine suffiziente Statistik zu ermitteln. Bevor wir dieses Ergebnis vorstellen, benötigen wir ein weiteres Konzept.

**Likelihood** Dies ist ein quantifizierter Wert,wie wahrscheinlich die Beobachtung einer gegebenen Stichprobe ist.

Die gemeinsame Verteilung einer Folge von Zufallsvariablen spielt in vielen Anwendungen der statistischen Inferenz eine Schlüsselrolle. Wenn diese gemeinsame Verteilung von einem oder mehreren unbekannten Parametern abhängt, kann sie als eine Funktion des Parameters / der Parameter betrachtet werden. Diese Funktion ist die Likelihood der Beobachtung der Daten in Abhängigkeit von dem/den Parameter(n). Hier ist die formale Definition:

**Likelihood-Funktion**

Diese Funktion errechnet, wie wahrscheinlich die Beobachtung einet gegebenen Stichprobe ist, basierend auf einem Wert des/der interessierenden Parameter(s).

x1, ..., xn seien eine Stichprobe von Beobachtungen aus der Folge von Zufallsvariablen X1, ...,Xn. Nehmen wir an, dass die Verteilung jedes Xi (für i = 1, ..., n) von einem oder mehreren unbekannten Parametern θ abhängt. Die **Likelihood** der Stichprobendaten ist die gemeinsame Dichte (PMF) von X1, ...,Xn, die bei x1, ...,Xn bewertet wird. Die **Likelihood-Funktion** ist gegeben durch

Xxx

wobei f die gemeinsame Dichte bezeichnet.

Angenommen, die Zufallsvariablen X1, ... ,Xn sind unabhängig und ihre jeweilige Verteilung hängt von unbekannten Parametern θi , i = 1, ..., n ab. Die Dichte von Xi wird mit fi x θi bezeichnet. Die Likelihood-Funktion der Stichprobendaten x1, ..., xn, die anhand der Xi beobachtet werden, ist dann wiederum die gemeinsame Dichte. Da die Xi unabhängig sind, ist die gemeinsame Dichte einfach das Produkt der Randverteilungen. Die Likelihood-Funktion ist daher gegeben durch

xxx

In der Praxis ist es oft der Fall, dass die beobachteten Daten unabhängig sind und aus identisch verteilten Zufallsvariablen stammen. Mit anderen Worten, wir haben eine Zufallsstichprobe x1, ..., xn, die aus X1, ...,Xn, iid beobachtet wird. Die Verteilung jedes Xi, i = 1, ..., n ist gleich, und die Dichte (PMF) hängt von einem oder mehreren unbekannten Parametern θ ab. Um die Likelihood-Funktion in diesem einfachen Fall zu schreiben, können wir einfach das Ergebnis der obigen Gleichung verwenden, ohne dass wir Indizes benötigen. Dies liegt daran, dass fi ≡ fj für alle i, j. Zusammenfassend lässt sich sagen, dass die Likelihood-Funktion für Stichprobendaten, die aus einer unabhängigen und identisch verteilten Zufallsvariablen beobachtet werden, gegeben ist durch

xxx

#### Beispiel 1.2.2

Schreiben Sie die Likelihood-Funktion der iid-Stichprobe 2, 3, 2, 1 aus einer Poisson-Verteilung mit unbekanntem Parameter λ auf. Es sei daran erinnert, dass die PMF von X Poisson λ gegeben ist durch

xxx

#### Lösung

Da die Stichprobe iid ist, können wir die obige Gleichung verwenden, um die Likelihood-Funktion zu schreiben:

Xxx

Betrachten wir nun die Statistik U = X1 + X2 + X3 + X4 für das vorherige Beispiel. Ist diese Statistik suffizient für die Schätzung von λ? Wie bereits erwähnt, erfordert die Anwendung der Definition der suffizienten Statistik die Ermittlung der Verteilung von U, was schwierig sein kann. Da wir nun mit einigen Kenntnissen über die Likelihood-Funktion ausgestattet sind, können wir ein wichtiges Ergebnis nutzen, das auf diesem Instrument beruht. Dieses Ergebnis stellt die Verbindung zwischen der Likelihood-Funktion und der Frage her, ob eine Statistik zur Schätzung eines unbekannten Parameters suffizient oder nicht. Das Folgende ist in Hogg et al. (2019) zu finden.

#### Theorem 1.2.1

Angenommen, X1, ..., Xn ist eine Zufallsstichprobe, x1, ..., xn, der entsprechenden beobachteten Daten und U eine Statistik. l θ = l θ x1, ..., xn sei die Likelihood-Funktion. U ist eine suffiziente Statistik für die Schätzung von θ, wenn und nur wenn l . in ein Produkt zweier nicht-negativer Funktionen zerlegt werden kann als

xxx

wobei g u, θ nicht von den beobachteten Daten und h nicht von θ abhängt.

Kehren wir nun zu unserer ursprünglichen Frage zurück: Ist die Statistik U = X1 + X2 + X3 + X + 4 suffizient für die Schätzung von λ in Beispiel 1.2.1? Auf der Grundlage des vorherigen Ergebnisses müssen wir prüfen, ob die Likelihood-Funktion eine bestimmte Form der Faktorisierung genießt:

xxx

In der Tat hängt g nicht von den Daten ab, außer über u, und h hängt nicht von λ ab. Diesem Theorem zufolge ist U eine suffiziente Statistik für die Schätzung von λ. Tatsächlich können wir dieses Ergebnis für eine beliebige Stichprobengröße erweitern. X1, ...,Xn sei eine Zufallsstichprobe aus einer Poisson-Verteilung mit dem Parameter λ. U = Σi = 1nXi ist eine suffiziente Statistik für λ.

#### Beispiel 1.2.3

X1, ...,Xn sei eine Zufallsstichprobe aus einer Gauß-Verteilung mit unbekanntem Mittelwert μ und bekannter Varianz σ2. Zeigen Sie, dass der Stichprobenmittelwertschätzer U = X eine suffiziente Statistik für die Schätzung von μ ist. Es sei daran erinnert, dass die Dichte eines Gaußschen X N μ, σ gegeben ist durch

xxx

#### Lösung

Zunächst schreiben wir die Likelihood-Funktion für eine beobachtete Stichprobe x1, ..., xn, die X1, ...,Xn entspricht:

xxx

Wir setzen nun u = X = 1nΣi = 1n Xi, so dass nu = Σi = 1n Xi. Die Likelihood-Funktion kann wie folgt geschrieben werden

Xxx

Nach dem Faktorisierungskriterium wissen wir also, dass U = X eine suffiziente Statistik zur Schätzung von μ ist, wenn σ2 bekannt ist.

Mit einer ähnlichen Berechnung kann gezeigt werden, dass U = 1 n - 1Σ i = 1 n Xi - μ 2 eine suffiziente Statistik für die Schätzung von σ2 ist, wenn μ bekannt ist. Schließlich sind U1 = X = 1n Σ i = 1 n Xi und U2 = Sn 2 = 1 n - 1Σ i = 1 n Xi - X 2 gemeinsam eine suffiziente Statistik für die Schätzung von μ und σ2. Um mit Hilfe des Faktorisierungskriteriums gemeinsame suffiziente Statistiken U1 und U2 für die Schätzung der Parameter θ1 und θ2 zu ermitteln, ersetzen wir einfach g u, θ durch g u1, u2, θ1, θ2 .

## 1.3 Maximum Likelihood

**Maximum-Likelihood-Schätzer**

Eine Funktion einer beobachteten Stichprobe, die die Likelihood-Funktion maximiert.

Beachten Sie, dass es sich um eine nicht zufällige

Menge.

Die Likelihood-Funktion misst die Likelihood der Beobachtung einer gegebenen Stichprobe in Abhängigkeit von einem (möglicherweise unbekannten) Parameter. Angenommen, wir erhalten beobachtete Daten x1, ..., xn, die X1, ..., Xn entsprechen, deren Verteilung von einem unbekannten Parameter θ abhängt. Mit Hilfe der Likelihood-Funktion können wir bestimmen, welcher Wert von θ aus einer Menge möglicher Werte am besten zu unserer beobachteten Stichprobe passt. Mit anderen Worten: Wenn wir zwei Schätzungen für θ haben, θ1 und θ2, so dass l θ1 > l θ2 , dann haben die gegebenen Daten eine höhere Likelihood, aus einer Verteilung mit dem Parameter θ = θ1 beobachtet zu werden als aus einer Verteilung mit dem Parameter θ = θ2. Wären dies unsere einzigen Möglichkeiten für θ, würden wir daher θ1 als Punktschätzung für θ wählen. Auf diese Weise hätten wir die Maximum-Likelihood-Methode zur Bestimmung der Punktschätzung für θ verwendet. Der Schätzer θ MLE = θ MLE X1, ...,Xn wird als **Maximum-Likelihood-Schätzer** für θ bezeichnet. Analog dazu wird θ MLE = θ MLE x1, ..., xn als Maximum-Likelihood-Schätzvariable bezeichnet.

**Maximum-Likelihood-**

**Schätzung**

Diese Funktion einer beobachteten Stichprobe maximiert die Likelihood-Funktion.

Es ist zu beachten, dass es sich um eine nicht zufällige Größe handelt.

#### Beispiel 1.3.1

Angenommen, die unten angegebene Stichprobe ist iid aus einer Poisson-Verteilung mit unbekanntem Parameter λ. Geben Sie die Likelihood-Funktion für diese Stichprobe an. Bestimmen Sie, welcher der beiden Werte 3, 4 die beobachteten Daten plausibler macht.

Tabelle 4: Stichprobendaten für Beispiel 1.3.1

#### Lösung

Die Likelihood-Funktion lautet

Xxx

Wir berechnen die Likelihood bei diesen Werten: l 3 = 1 . 09 Å~ 10-9 und l 4 = 2 . 07 Å~ 10-9.

Da l 4 > l 3 ist, ist λ = 4 eine bessere Likelihood-Schätzung für λ als λ = 3.

Ausgehend von der Lösung des Beispiels 1.3.1 gibt es in der Praxis nicht nur zwei Möglichkeiten für den Parameter, sondern einen kontinuierlichen Bereich von Möglichkeiten. Wir können definitiv nicht alle möglichen Werte ausprobieren! Stattdessen verwenden wir die Infinitesimalrechnung, um das Maximum der Likelihood-Funktion zu ermitteln. Dieser Wert wird als Maximum-Likelihood-Schätzung für λ bezeichnet. Die folgende Abbildung zeigt den Graphen der Likelihood-Funktion l λ aus Beispiel 1.3.1. Sie zeigt die Werte des Parameters λ auf der horizontalen Achse und die Likelihood-Werte auf der vertikalen Achse. Unser Ziel ist es, die Spitze (das Maximum) dieser Funktion zu finden und dann abzulesen, wo diese Spitze auf der horizontalen Achse auftritt. In unserem Fall liegt dieser Punkt bei 3,7. Aus diesem Grund können wir dies als unsere Maximum-Likelihood-Schätzung bezeichnen: λ MLE =.

Abbildung 7: Likelihood-Funktion für die Poisson-Stichprobe aus Beispiel 1.3.1

Es sei daran erinnert, dass die Abbildung z logz für z > 0 nicht abnehmend ist. Daher ist der Maximierer der Log-Likelihood, ll λ = logl λ derselbe wie der der Likelihood l λ . Außerdem ist die **Log-Likelihood** einfacher zu handhaben, weil sie Produkte in Summen von Logarithmen umwandelt; Summen abzuleiten ist viel einfacher, als Produkte abzuleiten. In der Praxis verwenden wir Computeralgorithmen, um den Maximierer von Funktionen zu finden. Die Likelihood-Funktion arbeitet mit sehr kleinen Zahlen (in der Größenordnung), und wir könnten auf Unterlaufprobleme stoßen, wenn die Werte, mit denen wir arbeiten, kleiner sind als die kleinste Zahl, die ein Computer darstellen kann. Die Log-Likelihood-Werte hingegen vermeiden kleine Zahlen und sind besser für numerische Verfahren geeignet.

**Log-Likelihood**

Der (natürliche) Logarithmus der Likelihood- Funktion lässt denselben Maximierer zu wie die Likelihood-

Funktion, ist aber einfacher zu handhaben.

Die folgende Abbildung zeigt die Log-Likelihood-Funktion für dieses Poisson-Beispiel. Ähnlich wie bei dem Graphen der Likelihood-Funktion befinden sich die Werte des Parameters (λ) auf der horizontalen Achse. Die Werte auf der vertikalen Achse sind die Werte der Log-Likelihood-Funktion ll λ . Wie bei der Likelihood-Funktion sind wir daran interessiert, den Wert auf der x-Achse zu finden, bei dem das Maximum auftritt. Dies ist derselbe Wert, den wir aus dem Likelihood-Diagramm erhalten.

Abbildung 8: Log-Likelihood-Funktion für die Poisson-Stichprobe aus Beispiel 1.3.1

Die Log-Likelihood-Funktion, die der Likelihood aus der Lösung von Beispiel 1.3.1 entspricht, ist gegeben durch

xxx

Ihre erste und zweite Ableitung lauten xxx bzw. xxx. Da ll′′ λ < 0 für jedes λ ist, wissen wir, dass jede Nullstelle von l′ ein lokaler Maximierer ist. In diesem Fall lässt ll′ genau eine Null zu:

Xxx

Daher ist λ MLE = 3 . 7 der (globale) Maximierer der Log-Likelihood und damit der Likelihood. Aus diesem Grund bezeichnen wir sie als die Maximum-Likelihood-Schätzung für λ. Jeder Maximierer einer Funktion ϕ ist der Minimierer von -ϕ. Da die meisten Optimierungsalgorithmen für die Minimierung von Funktionen geschrieben werden, arbeiten wir in der Praxis oft mit der **negativen** Log-Likelihood-Funktion, die lautet

**Negative Log-Likelihood**

Mit dem Negativ der Log-Likelihood-Funktion werden numerische Verfahren

geschrieben, um die objektive Funktion zu minimieren, daher wird in der Praxis die negative Log-Likelihood verwendet.

Der Minimierer dieser Funktion ist der gleiche wie der Maximierer der Log-Likelihood-

(und der Liklihood-)Funktion.

xxx

Die folgende Abbildung zeigt den Graphen der negativen Log-Likelihood-Funktion für das Poisson-Beispiel. Die Form dieses Graphen ist die umgekehrte Form der Log-Likelihood-Funktion. Mit Hilfe dieser Funktion können wir also die MLE bestimmen, indem wir das Minimum finden. Es ist zu beachten, dass 3.7 das Minimum dieser Funktion ist. Im weiteren Verlauf dieses Abschnitts werden wir den/die MLE-Punktschätzer finden, indem wir mit der negativen Log-Likelihood arbeiten.

Abbildung 9: Negative Log-Likelihood-Funktion für die Poisson-Stichprobe aus Beispiel 1.3.1

#### Beispiel 1.3.2

Betrachten Sie die Zufallsstichprobe X1, ...,Xn iid aus der Exponentialverteilung mit unbekannter Rate λ. Finden Sie den MLE-Schätzer durch Minimierung der negativen Log-Likelihood-Funktion. Es sei daran erinnert, dass die Dichte von X Exp λ gegeben ist durch

xxx

#### Lösung

Die Likelihood-Funktion, die einer beobachteten Stichprobe von D = x1, ..., xn entspricht, ist gegeben durch

Xxx

Die negative Log-Likelihood-Funktion lautet

Xxx

Wie zuvor finden wir die Nullstelle(n) der ersten Ableitung:

Xxx

Die zweite Ableitung bestätigt, dass es sich um den Minimierer handelt und die MLE-Schätzung daher tatsächlich dem obigen Ausdruck entspricht. Der entsprechende MLE-Schätzer für die Rate einer exponentiellen Stichprobe lautet

xxx

#### Beispiel 1.3.3

Angenommen, wir wollen die Wahrscheinlichkeit für einen Kopfwurf, p, einer Münze schätzen. Wir werfen eine Münze n-mal und bezeichnen Xi = 1, wenn die Münze Kopf zeigt, und Xi = 0, wenn die Münze Zahl zeigt. Wir erhalten somit eine Folge X1, ...,Xn. x1, ..., xn sollen die beobachteten Werte bezeichnen. Bestimmen Sie den MLE-Schätzer und den Schätzer für p.

#### Lösung

Unter der Annahme, dass die Münzwürfe unabhängig sind, und da wir dieselbe Münze verwenden, sind X1, ...,Xn iid aus Bernoulli p . Erinnern wir uns, dass die PMF für X Bernoulli p gegeben ist durch

Xxx

Die Likelihood der Stichprobe lautet

xxx

Die negative Log-Wahrscheinlichkeit lautet

Xxx

wobei wir x = 1n gesetzt haben

Σ

i = 1

n

xi. Als Nächstes finden wir die Nullstelle(en) der negativen Log-Likelihood-Funktion:

xxx

Die Berechnung der zweiten Ableitung von nll p bestätigt, dass dies tatsächlich der gesuchte Minimierer und damit die Maximum-Likelihood-Schätzung ist. Der entsprechende Maximum-Likelihood-Schätzer lautet

xxx

Unsicherheiten des MLE

Um die Qualität eines Schätzers θ MLE zu bewerten, müssen wir uns die damit verbundene Unsicherheit ansehen: V θ MLE . Es sei daran erinnert, dass die Varianz von X Bernoulli p durch V X = p 1 - p gegeben ist.

Da die Xi' aus dem Münzwurf-Beispiel unabhängig sind, erhalten wir

Xxx

Da wir p nicht wirklich kennen, können wir die Varianz schätzen mit

xxx

Wie sich herausstellt, brauchen wir die Unsicherheit nicht direkt auf diese Weise zu berechnen. Die mit einer MLE verbundene Unsicherheit ist der Kehrwert des Erwartungswerts der zweiten Ableitung der negativen Log-Likelihood, die mit der MLE bewertet wird:

xxx

Außerdem wird diese Approximation für größere Werte von n immer besser. Wenden wir dies auf unser Münzwurf-Beispiel an. Die zweite Ableitung der negativen Log-Likelihood für das Beispiel des Münzwurfs lautet

xxx

Mit den Erwartungswerten auf beiden Seiten (es sei daran erinnert, dass E X = p), ergibt sich

Xxx

Unter Verwendung der Gleichung zur Berechnung der Varianz der MLE ergibt sich schließlich

xxx

Die entsprechende Standardabweichung kann geschätzt werden als σ =

p MLE 1 - p MLE

n .

In der nachstehenden Abbildung wird die negative Log-Likelihood-Funktion für eine Zufallsstichprobe aus Bernoulli(p) dargestellt und die Schätzung p anhand des Ergebnisses von Beispiel 1.3.2 sowie die Schätzungen p Å} σ hervorgehoben. Auch hier ist die MLE-Schätzung von p der Minimierer dieser Funktion.

Abbildung 10: Negative Log-Likelihood für eine Bernoulli-Stichprobe, die p Å} σ zeigt

#### Beispiel 1.3.4

Verwenden Sie die Formel für die Varianz von MLE, um die mit dem MLE-Schätzwert für λ aus Beispiel 1.3.1 verbundene Unsicherheit zu schätzen.

#### Lösung

Wir beginnen damit, die negative Log-Likelihood-Funktion unter Vewendung des Ergebnisses der Gleichung aus Beispiel 1.3.1 aufzuschreiben:

xxx

Die Ableitungen lauten

xxx

und

xxx

Wendet man nun die Gleichung für die Varianz der MLE an, so ergibt sich

Xxx

Daher ist σ = 3 . 7 ≈ 0 . 608

Die nachstehende Abbildung zeigt die Auftragung der negativen Log-Likelihood. Die MLE-Schätzung ist der Minimierer dieser Funktion. Die Abbildung zeigt auch einen Bereich von Schätzungen innerhalb einer Standardabweichung der MLE.

Abbildung 11: Negative Log-Likelihood für eine Poisson-Stichprobe, die λ Å} σ zeigt

## 1.4 Gewöhnliche Methode der kleinsten Quadrate (Ordinary Least Squares)

Die Maximum-Likelihood-Methode erfordert Kenntnisse über die zugrundeliegende Verteilung, die die Stichprobendaten erzeugt hat. Diese Verteilung wird verwendet, um die Maximum-Likelihood-Schätzung für den unbekannten Parameter von Interesse zu ermitteln. Im Gegensatz dazu erfordert die gewöhnliche Methode der kleinsten Quadrate (Ordinary Least Squares) keine Kenntnis der zugrundeliegenden Verteilung; stattdessen muss das Modell bekannt sein, das die Daten erzeugt hat.

Eine Produktionsstätte stellt Aluminiumdosen her. Ein neues kamerabasiertes Messsystem soll eingesetzt werden, um zur Qualitätssicherung den Durchmesser der produzierten Dosen zu messen. Vor dem Einsatz dieses Systems in der Produktion wird ein Test durchgeführt, um die Genauigkeit des kamera- (bild-)basierten Messsystems zu messen. Für diesen Test wird eine einzelne Aluminiumdose mit einem Durchmesser von fünf Zentimetern verwendet. Das Messsystem wird verwendet, um den Durchmesser dieser Dose unter verschiedenen Bedingungen zu messen.

Zehn Messungen des Kamerasystems werden aufgezeichnety1, ..., y10 = 5 . 13, 5 . 07, 4 . 85, 5 . 00, 5 . 06, 4 . 93, 5 . 03, 5 . 01, 5 . 00, 4 . 98 . Um die Güte dieses Systems zu bewerten, berechnen wir die Differenz (Residuen) zwischen diesen Messungen und dem bekannten wahren Wert von y = 5. Um größere Residuen stärker zu bestrafen als kleinere Residuen, berechnen wir die Quadrate der Residuen: y - yi

2. Schließlich addieren wir diese quadrierten Residuen, um ein Maß für die Güte des kamerabasierten Messsystems zu erhalten. Die Berechnungen sind in der nachstehenden Tabelle aufgeführt.

*Tabelle 5: Beobachtete Messungen, Residuen und Quadrat-Residuen*

Die Summe der quadrierten Residuen beträgt Σ

i = 1

10

y - yi

2 = 0 . 0542. Bei der gewöhnlichen Methode der kleinsten Quadrate besteht unser Ziel darin, das Modell immer wieder zu optimieren, bis yi die niedrigste Summe der Quadrate die angibt. In der Praxis basiert das Modell auf Parametern, die wir schätzen müssen. Wir optimieren unser Modell, indem wir die Parameter so anpassen, dass die Quadratsumme der Residuen reduziert wird.

Nehmen wir an, dass wir für unser Kamera-Beispiel einige dieser Parameter optimieren und neue Messwerte y1

(neu), ..., yn

(neu) = 5 . 095 . 054 . 895 . 5 . 044 . 955 . 025 . 015 . 4 . 99 erhalten. Wenn wir die Berechnungen mit dieser neuen Menge von Messwerten wiederholen, erhalten wir

Σ

i = 1

10

y - yi

(neu) 2 = 0 . 0274. Da die neue Quadratsumme niedriger ist als die ursprüngliche, haben wir unser Modell verbessert. Wenn dies das Beste ist, was dieses Modell erreichen kann, d. h. wenn jede Änderung der Parameter des Modells immer zu einer höheren Quadratsumme führt, dann wird dieses Modell als die Lösung der kleinsten Quadrate bezeichnet.

Sehen Sie sich die beiden folgenden Abbildungen an, die die Messwerte bzw. die quadrierten Residuen zeigen. In der ersten Abbildung stellt jeder Punkt eine Vorhersage (Schätzung) für den Durchmesser der erkannten Dose dar. Die Werte der Schätzungen befinden sich auf der vertikalen Achse. Die Schätzungen sind in zwei verschiedenen Farben dargestellt, wobei jede Farbe für ein anderes Modell steht. Es ist zu beachten, dass die Schätzungen des "neuen" Modells im Vergleich zu dem anderen Modell weniger stark um den Zielwert von 5,0 streuen. In der letztgenannten Abbildung stellt jeder Punkt in den Diagrammen eine Vorhersage dar, wobei die vertikale Achse das quadrierte Residuum und die horizontale Achse den vorhergesagten Wert misst. Das "neue" Modell hat im Vergleich zu dem anderen Modell kleinere quadrierte Residuen, die eng um den Wert Null herum angeordnet sind.

Abbildung 12: Messwerte aus zwei Modellen

Abbildung 13: Quadrierte Residuen aus zwei Modellen

Nehmen wir nun an, dass wir x1, y1 , ..., xn, yn beobachtet haben. Nehmen wir an, dass eine funktionale Abhängigkeit zwischen xi und yi besteht, so dass yi = f xi θ , wobei θ der/die Parameter ist/sind, der/die diese Funktion vollständig bestimmt. Die Kleinste-Quadrate-Schätzung θ(OLS) minimiert die Quadratsumme der Residuen:

xxx

Mit anderen Worten: θ(OLS) = minθC θ . Die OLS-Methode liefert ein gutes Ergebnis, wenn die funktionale Abhängigkeit oder das verwendete Modell für die Daten geeignet ist. Wenn dies der Fall ist, dann ist die OLS-Schätzung des Parameters, der die beste Funktion oder das beste Modell bestimmt, für die Zwecke gut geeignet. Wenn jedoch die funktionale Abhängigkeit oder das Modell von vornherein falsch ist, dann würde auch die OLS-Schätzung der Parameter zu einem unbrauchbaren Ergebnis führen. Die Abbildung mit dem Titel "Beobachtete bivariate Daten" zeigt einen beobachteten Datensatz von 100 Punkten. Die beobachteten Daten haben zwei Variablen, x und y, die im Diagramm als horizontale bzw. vertikale Werte dargestellt sind. Ziel ist die Anpassung einer Kurve an diese Daten. In der Abbildung mit dem Titel "Kostenfunktionen für zwei Modelle" sind zwei Graphen von zwei Kostenfunktionen dargestellt, von denen eine von einem guten Modell und die andere von einem schlechten Modell ausgeht. Es ist zu beachten, dass die Kostenfunktion des guten Modells niedrigere Werte aufweist als der niedrigste Kostenwert des schlechten Modells. Die OLS-Anpassung für jedes dieser Modelle ergibt sich aus der Suche nach dem Wert des Parameters θ (auf der horizontalen Achse), der die Kosten C θ (auf der vertikalen Achse) minimiert. Schließlich zeigen wir in der Abbildung "Angepasste Kurven für bivariate Daten" drei Kurven, die an die Daten angepasst wurden. Die am besten angepasste Kurve ergibt sich aus einer guten Modellannahme und der OLS-Schätzung von θ. Das zweitbeste Modell ergibt sich aus dem guten Modell und einer anderen Schätzung von θ (nicht der OLS). Die am schlechtesten angepasste Kurve ergibt sich aus einer schlechten Modellannahme, obwohl wir die OLS-Schätzung für θ gewählt haben. Daher ist die Modellannahme sehr wichtig, wenn man sich für die Verwendung von OLS zur Ermittlung der Punktschätzungen unbekannter Parameter entscheidet.

Abbildung 14: Beobachtete bivariate Daten

Abbildung 15: Kostenfunktionen für zwei Modelle

Abbildung 16: Angepasste Kurven für bivariate Daten

## 1.5 Techniken der Stichprobenwiederholung (Resampling)

In allen vorangegangenen Abschnitten haben wir verschiedene Möglichkeiten kennengelernt, einen unbekannten Parameter der Grundgesamtheit anhand einer beobachteten Stichprobe zu schätzen. Wenn der Stichprobenumfang groß ist, bietet der zentrale Grenzwertsatz eine Möglichkeit, die mit einer Parameterschätzung verbundene Unsicherheit zu messen. Wenn der Stichprobenumfang nicht groß ist, aber die Verteilung der Grundgesamtheit, aus der die Stichprobe gezogen wurde, bekannt ist (mit Ausnahme des unbekannten Parameters), dann können wir die Eigenschaften dieser Verteilung nutzen, um die mit unserer Punktschätzung verbundene Unsicherheit zu messen. Was aber, wenn der Stichprobenumfang zu gering ist, um den zentralen Grenzwertsatz anzuwenden, und wir keine Ahnung von der Verteilungsfamilie der Grundgesamtheit haben? Für dieses Szenario wurden Resampling-Verfahren eingeführt.

Resampling-Techniken sind Parameterschätzungsverfahren, die Teilstichproben einer beobachteten Stichprobe verwenden. In diesem Abschnitt werden wir die Bootstrap- und die Jackknife-Methode im Detail besprechen.

### Das Bootstrap-Verfahren

Bezeichnen wir eine beobachtete Stichprobe mit x = x1, ..., xn . Bei der Bootstrap- Stichprobenwiederholung werden n Zahlen für die beobachtete Stichprobe nacheinander und mit Ersatz ausgewählt. Wenn die beobachtete Stichprobe beispielsweise 1, 2, 3, 4, 5 lautet, dann könnte eine Bootstrap-Stichprobe 5, 1, 1, 3, 2 lauten. Dieses Verfahren wird eine feste Anzahl von Malen wiederholt, die im Voraus festgelegt wird. Die folgende Tabelle zeigt eine beobachtete Stichprobe zusammen mit zehn Bootstrap-Stichproben.

Tabelle 6: Eine Zufallsstichprobe von fünf Zahlen und zehn Bootstrap-Stichproben

Angenommen, wir wollen den Mittelwert der Grundgesamtheit berechnen, aus der die Daten beobachtet wurden. Eine Punktschätzung wäre der Stichprobenmittelwert unter Verwendung der beobachteten Stichprobe

x = 7.0+2.9 + 2.3 + 5.5 + 7.2

5 = 4.98

Für eine Schätzung mit der Bootstrap-Methode würden wir den Mittelwert jeder der Bootstrap-Stichproben berechnen (die Ergebnisse sind in der Tabelle unten aufgeführt) und dann den Mittelwert dieser Mittelwerte berechnen.

Xxx

Tabelle 7: Stichprobenmittelwerte der Bootstrap-Stichproben aus der vorherigen Tabelle

Der Vorteil dieser (zehn) Schätzungen besteht darin, dass wir sie verwenden können, um die mit der endgültigen Schätzung x(boot) verbundene Unsicherheit zu ermitteln. In unserem Fall berechnen wir den quadrierten Standardfehler der Bootstrap-Schätzung wie folgt:

xxx

Äquivalent dazu haben wir SE x boot = 0,925 = 0,9618.

Es folgt eine Zusammenfassung des allgemeinen Ansatzes unter Verwendung einer beobachteten Stichprobe, die durch x = x1, ..., xn gegeben ist, unter Verwendung von B Bootstrap-Stichproben. Zunächst wird die Punktschätzung des Zielparameters θ berechnet und mit θ bezeichnet. Anschließend werden für jedes b von 1 bis B die Bootstrap-Punktschätzungen θ b aus x b = xi1, xi2, ..., xin berechnet, wobei ij mit gleicher Wahrscheinlichkeit und mit Ersetzung aus 1, ..., n gezogen werden. Anschließend wird die Unsicherheit der ursprünglichen Punktschätzung unter Verwendung der Bootstrap-Schätzungen als Standardfehler nach folgender Formel berechnet

xxx

Für eine Zufallsstichprobe von X1, ...,X10 ist die **geordnete Statistik** durch X 1 , ...,X 10 gegeben, wobei X 1 ≤ X 2 ≤ ⋯ ≤ X 10 . Das Maximum ist also Xmax = X 10 , und das Minimum ist Xmin = X 1 .

#### Beispiel 1.5.1

Verwenden Sie die beobachtete Stichprobe zusammen mit den zehn Bootstrap-Stichproben, um die Bootstrap-Schätzung des Maximums, Xmax = X 5 , zu ermitteln.

#### Lösung

Zunächst berechnen wir die Punktschätzung aus der beobachteten Stichprobe: xmax = 7 . 2. Als Nächstes berechnen wir die Punktschätzungen aus den zehn Bootstrap-Schätzungen. Dabei handelt es sich lediglich um die Maxima aus jeder der Stichproben (siehe Tabelle unten).

Tabelle 8: Maximalwerte der Bootstrap-Stichproben

Der Standardfehler mittels Bootstrap ergibt sich aus der nachstehenden Gleichung:

Xxx

In der Praxis sind B = 10 Bootstrap-Stichproben möglicherweise nicht ausreichend. Andererseits würde eine Stichprobe mit einem Umfang von 12 eine Gesamtzahl von etwa 1,3 Millionen Bootstrap-Teilstichproben aufweisen, was die Verwendung aller möglichen Bootstrap-Stichproben rechnerisch undurchführbar macht. In der Regel reichen B = 30 bis B = 100 für praktische Anwendungen aus.

### Das Jackknife-Verfahren

**delete-1-Jackknife-Replikation**

Diese Teilstichprobe erhält man durch Entfernen eines Wertes aus einer gegebenen Stichprobe.

Bei einer beobachteten Stichprobe x = x1, x2, ..., xn ist die **delete-1-Jackknife-Replikation** eine Teilstichprobe der beobachteten Daten, bei der ein Datenpunkt gelöscht wurde. Zum Beispiel ist x - 1 = x2, ..., xn eine dieser Jackknife-Replikationen und x - 3 = x1, x2, x4, ..., xn eine weitere. Insgesamt gibt es n Jackknife-Replikationen:

xxx

Angenommen, g:ℝn ℝ ist eine Funktion, die eine Statistik θ = g x = g x1, ..., xn berechnet, um einen unbekannten Parameter θ zu schätzen. Die entsprechende Funktion g:ℝn - 1 ℝ berechnet eine Schätzung θ - i durch Anwendung von g auf eine Jackknife-Replikation x - i ;

θ - i = g x - i = g x1, ..., xi - 1, xi + 1, ..., xn .

Angenommen, g . berechnet den Stichprobenmittelwert, dann ist θ = 1n x1 + ..., xn . Die entsprechende Funktion g . berechnet den Stichprobenmittelwert für die n - 1 Punkte aus einer Jackknife-Replikation:

θ - 1 = g x - 1 = 1

n - 1 x2 + ..., xn . Die Jackknife-Schätzung von θ ist der Stichprobenmittelwert aller Jackknife-Schätzungen:

xxx

#### Beispiel 1.5.2

Schätzen Sie angesichts der beobachteten Daten 3, 7, 1, 0, 4 aus X1, ...,X5 iid-Kopien von X θ = E X2 mit der Jackknife-Methode.

#### Lösung

Legen wir x = 3, 7, 1, 0, 4 fest. Das Stichprobenmoment über diese Stichprobe ergibt eine Schätzung von θ:

Xxx

Die fünf Jackknife-Replikationen und ihre entsprechenden Schätzungen lauten

Xxx

Die Jackknife-Schätzung schließlich ist gegeben durch

Xxx

Der Vorteil der Jackknife-Resampling-Methode besteht darin, dass wir eine Schätzung der Unsicherheit der Jackknife-Schätzung erhalten können, was ohne weitere Kenntnisse über die Verteilung, aus der die Daten entnommen wurden, nicht möglich ist. Wir messen die mit der Jackknife-Schätzung verbundene Unsicherheit, θ cot, unter Verwendung des Jackknife-Standardfehlers, der gegeben ist durch

xxx

#### Beispiel 1.5.3

Verwenden Sie die obige Formel für den Jackknife-Standardfehler, um den Standardfehler der Jackknife-Schätzung aus Beispiel 1.5.2 zu berechnen.

#### Lösung

Hier haben wir n = 5,

θ - 1 = 13,2, θ - 2 = 5,2, θ - 3 = 14,8, θ - 4 = 15, θ - 5 = 11,8 und θ . = 12:

xxx

ergibt SE θ jack = 7,187.

### Zusammenfassung

In dieser Lektion haben wir verschiedene Methoden zur Schätzung von Parametern vorgestellt: die Momentenmethode, die Maximum-Likelihood-Methode, die gewöhnliche Methode der kleinsten Quadrate (OLS) und Resampling-Methoden. Die Momentenmethode setzt die Momente einer Verteilung mit dem interessierenden Parameter in Beziehung und verwendet dann die Stichprobenmomente, um einen Schätzer zu ermitteln. Es sei daran erinnert, dass das k-te Moment einer Zufallsvariablen X μ k = E Xk ist, und wenn wir eine Folge von Kopien X1, ...,Xn haben, lautet der Stichprobenmomentschätzer m k = 1n

Σ

i = 1

n

Xi

k. Die entsprechende Schätzung mit beobachteten Daten

x1, ..., xn lautet m k = 1n

Σ

i = 1

n

xi

k. Angenommen, wir wollen einen unbekannten Parameter θ mit der Momentenmethode schätzen. Wir ermitteln eine Funktion f, die die Momente μ k mit θ in Beziehung setzt, θ = f μ 1 , μ 2 , ..., μ K und dann ist der Schätzer gegeben durch θ = f m 1 ,m 2 , ...,m K . Schließlich ist die Schätzung auf der Grundlage der beobachteten Daten gegeben durch θ = f m 1 ,m 2 , ...,m K .

Eine wichtige Eigenschaft eines jeden Punktschätzers ist, ob er unverzerrt ist oder nicht: E θ = θ. Wir haben gezeigt, dass einige Schätzungen nach der Momentenmenthode unverzerrt sind, andere dagegen nicht. Zum Beispiel ist der Schätzer des Stichprobenmittelwerts X ein unverzerrter Schätzer für den Mittelwert (μ) einer Gauß-Verteilung. Der Momentenschätzer für die Varianz σ2 einer Gauß-Verteilung ist jedoch nicht unverzerrt, wenn der Mittelwert unbekannt ist.

Eine Statistik ist eine Zufallsvariable und eine Funktion einer Zufallsstichprobe X1, ...,Xn. In unserem Zusammenhang ist eine Statistik ein Schätzer für einen unbekannten Parameter. Umfasst diese Statistik alle Informationen, die die Stichprobe für die Schätzung des unbekannten Parameters zu bieten hat, wird sie als suffiziente Statistik bezeichnet. Wir haben uns die Definition einer suffizienten Statistik auf der Grundlage der gemeinsamen Verteilung der Stichprobe unter der Bedingung angesehen, dass die Statistik frei von dem Parameter ist. Eine praktischere Methode zur Bestimmung der Suffizienz ist die über das Kriterium der Likelihood-Faktorisierung.

Die Maximum-Likelihood-Methode verwendet die Verteilungsfamilie einer Stichprobe, um einen Punktschätzer als Maximierer der Likelihood-Funktion abzuleiten. Wenn die Daten unabhängig sind, ist die Likelihood-Funktion das Produkt der Dichten (oder PMS) der Zufallsvariablen in der Stichprobe. Dies ist in der Praxis häufig der Fall. Die Log-Likelihood ist die Funktion, die verwendet wird, weil sie algebraische und rechnerische Effizienz bietet. Die negative Log-Likelihood schließlich wird verwendet, wenn numerische (Computer-)Algorithmen zur Ermittlung der MLE-Schätzungen eingesetzt werden. Wie bei der Momentenmethode sind einige MLE-Schätzer unverzerrt, während andere es nicht sind. Eine der schönen Eigenschaften von MLE-Schätzern (im Gegensatz zu Momentenschätzern) liegt darin, dass sie immer suffiziente Statistiken oder Funktionen aller suffizienten Statistiken sind. Mit anderen Worten: Wenn wir eine MLE erhalten, wissen wir, dass wir alle Informationen, die die Stichprobe in Bezug auf den unbekannten Parameter von Interesse zu bieten hat, zusammengefasst haben.

Im Gegensatz zur Maximum-Likelihood-Methode erfordert die gewöhnliche Methode der kleinsten Quadrate (OLS) zur Schätzung von Parametern keine Kenntnisse über die zugrundeliegende Verteilung, die die Zufallsstichprobe erzeugt hat. Die einzige Voraussetzung für OLS ist die funktionale/Modellabhängigkeit innerhalb der Zufallsstichprobe. Die OLS-Schätzungen der interessierenden Parameter werden durch Minimierung der Summe der quadrierten Residuen ermittelt.

Im letzten Abschnitt haben wir zwei Methoden der Stichprobenwiederholung (Resampling) erörtert: das Bootstrap- und das Jackknife-Verfahren. Beide Methoden bieten eine Möglichkeit zur Berechnung der Unsicherheit einer Punktschätzung, wenn der Stichprobenumfang nicht ausreicht, um die Theorie der großen Stichprobe unter Verwendung des zentralen Grenzwertsatzes anzuwenden, oder wenn die Verteilung der zugrundeliegenden (Grundgesamtheit) unbekannt ist.