

Kurz & gut*

Statistik für Experimente

von Joachim Selke[◊]

In diesem Artikel werden die Grundbegriffe der mathematischen Statistik vorgestellt. Der Text zielt auf Studierende ab, die bereits über Vorkenntnisse aus dem Bereich der mathematischen Stochastik verfügen.

Wozu Statistik?

Statistische Verfahren werden in Experimenten hauptsächlich angewendet, um den Wert bisher unbekannter quantitativer Größen zu schätzen oder um eine Aussage auf ihre Gültigkeit hin zu überprüfen. Beispiele dafür sind die Ermittlung von Einschaltquoten und die Frage, ob ein Medikament eine größere Heilwirkung erzielt als ein anderes.

Um derartige Informationen exakt zu ermitteln, ist es in den meisten Fällen nötig, eine sehr große – wenn nicht sogar unendliche – Anzahl von Datensätzen zu erheben. Für die exakte Ermittlung von Einschaltquoten etwa müßten alle Menschen im Sendegebiet befragt werden.

Da in der Praxis jedoch zumeist nur wenige Daten erhoben werden können und Näherungswerte für die gesuchten Informationen akzeptabel sind, wird mit Methoden der Statistik aus der Gestalt der vorliegenden Datensätze die Gestalt aller Datensätze geschätzt. Es ist beispielsweise eine vernünftige Annahme, daß die Einschaltquoten in einer zufällig ausgewählten (nicht zu kleinen) Menge von Menschen im Sendegebiet in etwa den Einschaltquoten in der Menge aller Menschen im Sendegebiet entsprechen.

Von großer Bedeutung für die Interpretation der auf diese Weise geschätzten Ergebnisse ist es dabei, daß Aussagen zur Qualität derselben gemacht werden können. Schätzungen, die in vielen Fällen stark von den exakten Werten abweichen, sind nicht zu gebrauchen. Umgekehrt stellt sich bereits bei der Planung von Experimenten die Frage, wieviele Datensätze benötigt werden, um bei der Auswertung verlässliche Ergebnisse zu erhalten.

Modellbildung

Um Situationen wie die oben geschilderten mathematisch untersuchen zu können, ist ein allem zugrundeliegendes Modell erforderlich.

*Dieser Artikel weicht von den für Publikationen aus der Reihe „Kurz & gut“ üblichen Vorgaben zu Gestaltung und Umfang ab. Die Gründe dafür sind der Platzbedarf für die verwendeten mathematischen Formeln sowie die Komplexität des dargestellten Themas.

◊E-Mail: mail@joachim-selke.de

Experimente sollen eine Aussage über eine Menge von Objekten ermöglichen. Diese Menge, also die Menge der potentiellen Untersuchungsobjekte, wird Grundgesamtheit genannt.

In jedem Experiment wird eine Teilmenge der Grundgesamtheit ausgewählt, dies ist die sogenannte Stichprobe, also die Menge der tatsächlichen Untersuchungsobjekte.

Jedes Untersuchungsobjekt wird nun einer Messung unterzogen, die die Ausprägung eines vorher festgelegten quantifizierbaren Merkmals des jeweiligen Objektes liefert. Ist Ω die Grundgesamtheit, so kann das Ergebnis der Messung eines Objektes $\omega \in \Omega$ als Wert einer (Meß-)Funktion $f : \Omega \rightarrow \mathbb{R}$ an der Stelle ω betrachtet werden.¹

Ist $\{\omega_1, \omega_2, \dots, \omega_n\}$ die Stichprobe (mit Umfang $n \in \mathbb{N}$) und f eine passende Meßfunktion, so werden die zugehörigen sogenannten Stichprobenwerte $f(\omega_1), f(\omega_2), \dots, f(\omega_n)$ häufig mit x_1, x_2, \dots, x_n bezeichnet. In der Regel wird auch das Tupel (x_1, x_2, \dots, x_n) zur Vereinfachung als Stichprobe bezeichnet.

Deskriptive Statistik

Aufgabe der sogenannten deskriptiven² Statistik ist die Charakterisierung von Stichproben durch Kenngrößen. Es werden dabei jedoch ausschließlich Aussagen über die Stichprobe getroffen, eine Verallgemeinerung dieser Aussagen auf die Grundgesamtheit ist unzulässig.

Wichtige Kenngrößen sind der Mittelwert \bar{x} , der Median \tilde{x} und die Standardabweichung σ , die wie folgt definiert sind:

$$\bar{x} := \frac{1}{n} \cdot \sum_{i=1}^n x_i \qquad \hat{\sigma}^2 := \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\tilde{x} := \begin{cases} x_{(\frac{n+1}{2})} & \text{für } n \text{ ungerade,} \\ \frac{1}{2} \cdot \left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) & \text{für } n \text{ gerade,} \end{cases}$$

wobei $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die der Größe nach geordneten Stichprobenwerte sind. Die Größe $\hat{\sigma}^2$ wird Stichprobenvarianz genannt.

Der Median wird im Gegensatz zum Mittelwert als robust bezeichnet, da Ausreißer den Median kaum beeinflussen, jedoch große Änderungen des Mittelwerts bewirken können. Zieht beispielsweise ein Multimillionär in ein kleines Dorf, so treibt dies das mittlere Vermögen der Einwohner dieses Dorfes in der Regel nach oben, obwohl sich dadurch das Vermögen eines „typischen“ Einwohners nicht nennenswert geändert hat.

Standardabweichung und Stichprobenvarianz dienen als Maß dafür, wie stark die Stichprobenwerte um den Mittelwert herum streuen.

Ist die Grundgesamtheit endlich, so kann durch eine sogenannte Vollerhebung jedes potentielle Untersuchungsobjekt auch tatsächlich einer Messung unterzogen werden; in der

¹Anstatt \mathbb{R} können auch andere Mengen betrachtet werden. Um die Übersichtlichkeit zu erhöhen, wird in dieser Arbeit auf eine Verallgemeinerung der Definition verzichtet.

²beschreibenden

Praxis ist dies selbstverständlich meist nur für kleine Grundgesamtheiten möglich. Nur unter dieser Bedingung kann die deskriptive Statistik Aussagen über die Grundgesamtheit liefern. Dafür handelt es sich dabei dann auch um exakte Aussagen.

Induktive Statistik

Aufgabe des sogenannten induktiven³ Statistik ist es, von Eigenschaften der Stichprobe auf Eigenschaften der Grundgesamtheit zu schließen.

Ansatzpunkt ist dabei die Annahme, daß eine (große) zufällig ausgewählte Teilmenge der Grundgesamtheit mit hoher Wahrscheinlichkeit der Grundgesamtheit ähnelt. Es ist daher (mit geringer Irrtumswahrscheinlichkeit) möglich, von den Eigenschaften einer solchen Teilmenge auf die Eigenschaften der Grundgesamtheit zu schließen.

Das mathematische Modell ergibt sich daher wie folgt. Sei Ω die Grundgesamtheit und Y eine Zufallsgröße, die auf Ω gleichverteilt ist.⁴ Durch Anwendung einer Funktion $f : \Omega \rightarrow \mathbb{R}$ erhält man eine Zufallsvariable $X := f(Y)$ mit einer von f abhängigen Verteilung. Der Wertebereich von X ist damit \mathbb{R} .

Liegt nun eine n -elementige Folge von Realisierungen (x_1, x_2, \dots, x_n) der Zufallsvariablen X vor, so stellt die induktive Statistik Methoden zur Verfügung, mit denen die Verteilung von X aus dieser Folge geschätzt werden kann. Eine solche Folge (x_1, x_2, \dots, x_n) läßt sich als Stichprobe auffassen.

Angemerkt sei an dieser Stelle, daß die im folgenden vorgestellten Methoden der induktiven Statistik in einem Experiment nur dann anwendbar sind, wenn die Auswahl der tatsächlichen Untersuchungsobjekte wie im Modell auf einer gleichverteilten Ziehung aus der Grundgesamtheit beruht. Unter speziellen Bedingungen ist jedoch eine gleichverteilte Ziehung aus einer Teilmenge Δ der Grundgesamtheit Ω ausreichend. Diese liegen vor, wenn $f(Y_\Omega)$ und $f(Y_\Delta)$ eine ähnliche Verteilung besitzen, wobei $f : \Omega \rightarrow \mathbb{R}$ eine Funktion ist, Y_Ω eine auf Ω und Y_Δ eine auf Δ gleichverteilte Zufallsvariable.

Ist beispielsweise die Grundgesamtheit die Menge aller Studierenden an deutschen Hochschulen, so darf die Stichprobe in einem Experiment, das mit Methoden der induktiven Statistik ausgewertet werden soll, nur dann aus der Menge der Studierenden an der Universität Hannover erhoben werden, wenn sich beide Mengen bezüglich der Ausprägungen des zu messenden Merkmals ähneln.

Hierbei handelt es sich um eine Grundvoraussetzung, die für jedes Experiment überprüft werden muß.

Im folgenden bezeichne Ω stets die Grundgesamtheit, Y eine auf Ω gleichverteilte Zufallsvariable und $f : \Omega \rightarrow \mathbb{R}$ die Meßfunktion. Zudem seien $X := f(Y)$ und (x_1, x_2, \dots, x_n) mit $n \in \mathbb{N}$ eine Folge von Realisierungen von X .

³In der mathematischen Logik wird das Folgern vom Speziellen auf das Allgemeine als Induktionsschluß bezeichnet.

⁴Nicht auf jeder Menge läßt sich mathematisch korrekt eine gleichverteilte Zufallsgröße konstruieren. Um das vorgestellte Modell möglichst übersichtlich zu halten, wird diese Tatsache hier jedoch vernachlässigt. Anschaulich sollte klar sein, was unter einer Gleichverteilung zu verstehen ist.

Parametrische und nicht-parametrische Statistik

Neben der Stichprobe sind in Experimenten häufig zusätzliche Informationen über die Verteilung von X vorhanden. So kann beispielsweise der Wertebereich von X oder sogar der Typ der Verteilung (wie etwa Binomial- oder Normalverteilung) bekannt sein.

Für den Spezialfall, daß nur eine bekannte Menge von Verteilungen für X in Frage kommt, deren Elemente sich durch einen (häufig reellwertigen) Parameter vollständig beschreiben lassen, liefert die sogenannte parametrische Statistik Verfahren zur Analyse der Stichprobe.

Eine solche Menge von Verteilungen (genauer: Wahrscheinlichkeitsdichten) ist zum Beispiel

$$\left\{ x \mapsto \lambda e^{-\lambda x} \mid \lambda \in (0, \infty) \right\},$$

die Klasse aller Wahrscheinlichkeitsdichten zu exponentialverteilten Zufallsvariablen.

Aufgabe der parametrischen Statistik ist es demnach, den („wahren“) Parameter der Verteilung von X zu ermitteln. Dieser Parameter wird häufig (und im folgenden auch hier) mit ϑ bezeichnet. Die (bekannte) Menge, aus der dieser Parameter stammt, trägt oft die Bezeichnung Θ , beliebige Elemente aus Θ werden mit ϑ bezeichnet.

Da in Experimenten häufig durch den Aufbau des Experimentes und die gemessenen Größen eine vernünftige Annahme über die Verteilung von X bis auf einen Parameter gemacht werden kann⁵, findet hier zumeist die parametrische Statistik Anwendung. Auf diese wird im folgenden näher eingegangen.

Liegen keine derartigen Informationen über die Verteilung von X vor, so finden Methoden der sogenannten nichtparametrischen Statistik Anwendung. Hierbei geht es beispielsweise um die Erkennung von Mustern in der Stichprobe oder um die Ermittlung des Typs der Verteilung von X . Es findet keine Annahme über die Verteilung von X statt.

Schätzer

Die Schätztheorie beschäftigt sich als Teilgebiet der parametrischen Statistik mit der Schätzung eines quantifizierbaren Merkmals der Verteilung von X aus den Stichprobenwerten. Auf das Modell bezogen soll für eine bekannte Funktion $g : \Theta \rightarrow \mathbb{R}$ der Wert $g(\vartheta)$ geschätzt werden. Als Schätzer für diesen Wert (bei einer Stichprobe vom Umfang n) wird jede Funktion $\hat{g}_n : \mathbb{R}^n \rightarrow \mathbb{R}$ bezeichnet.

Es lassen sich viele Kriterien angeben, die „gute“ Schätzer erfüllen sollten (insbesondere, daß die Qualität der Schätzung mit wachsendem n ebenfalls steigt). Für diese sei jedoch auf die Literatur verwiesen. Im folgenden wird ein naheliegender und in den meisten Fällen anwendbarer Ansatz vorgestellt, mit dem sich eine bestimmte Art von Schätzer konstruieren läßt, der sogenannte Maximum-Likelihood-Schätzer.

Für jeden Parameter $\vartheta \in \Theta$ läßt sich (unter der Annahme, daß diese Realisierung voneinander unabhängig sind) die Wahrscheinlichkeit berechnen, mit der die Stichprobe

⁵Die Lebensdauer elektronischer Bauteile ist beispielsweise zumeist exponentialverteilt, die Größe von Webressourcen hingegen Pareto-verteilt.

von einer Zufallsvariablen realisiert wird, die die durch ϑ definierte Verteilung besitzt.⁶ Sei $\tilde{\vartheta} \in \Theta$ so gewählt, daß diese Wahrscheinlichkeit für $\tilde{\vartheta}$ maximal ist. Ein Maximum-Likelihood-Schätzer wählt als Schätzung für $g(\hat{\vartheta})$ den Wert $g(\tilde{\vartheta})$.

Maximum-Likelihood-Schätzer legen der Schätzung also die Verteilung zugrunde, unter der das Auftreten der Stichprobe am wahrscheinlichsten ist (bezogen auf alle für möglich gehaltenen Verteilungen). Diese Vorgehensweise ist nicht zwingend, besitzt aber eine durchaus nachvollziehbare Idee.

Konfidenzintervalle

Neben der reinen Schätzung eines Wertes durch einen Schätzer wie oben ist selbstverständlich auch die Qualität dieser Schätzung von großer Bedeutung.

Liefert der Schätzer beispielsweise einen Wert, der als Schätzung genauso oder ähnlich plausibel ist wie ein anderer Wert, so sollte diese Information nicht verloren gehen. Daher gibt man neben dem Schätzwert häufig Zahlenbereiche an, die den wahren zu schätzenden Wert „mit großer Wahrscheinlichkeit“ enthalten.

Formal wird nach Funktionen $L, U : \mathbb{R}^n \rightarrow \mathbb{R}$ gesucht, so daß das (zufällige!) Intervall

$$[L(X_1, X_2, \dots, X_n), U(X_1, X_2, \dots, X_n)]$$

(für voneinander unabhängige und identisch zu X verteilte Zufallsvariablen X_1, X_2, \dots, X_n) den wahren zu schätzenden Wert $g(\hat{\vartheta})$ mindestens mit einer vorher festgelegten Wahrscheinlichkeit $\alpha \in [0, 1]$ umfaßt. Diese Bedingung soll für alle $\hat{\vartheta} \in \Theta$ gelten. Ein solches Intervall wird Konfidenzintervall zum Niveau α genannt.

Bei „gut“ gewählten Konfidenzintervallen verkleinert sich die Breite des Intervalls mit wachsendem n , vergrößert sich jedoch mit wachsendem α .

Für eine (zufällige) Realisierung (x_1, x_2, \dots, x_n) umfaßt das Konfidenzintervall also mit einer Wahrscheinlichkeit von $100\alpha\%$ den Wert $g(\hat{\vartheta})$. Geht man nun davon aus, daß es sich bei der zu untersuchenden Stichprobe um eine „typische“ Realisierung handelt und α groß ist, so kann man sich ziemlich sicher sein, daß auch in diesem Fall das Intervall

$$[L(x_1, x_2, \dots, x_n), U(x_1, x_2, \dots, x_n)]$$

den Wert $g(\hat{\vartheta})$ umfaßt. Die Zahl $1 - \alpha$ läßt sich dann als „Irrtumswahrscheinlichkeit“ deuten, obwohl hier selbstverständlich nicht mehr von Wahrscheinlichkeiten die Rede sein kann: der Wert $g(\hat{\vartheta})$ ist entweder in diesem Intervall enthalten oder nicht.

Tests

Häufig stellt sich die Frage, ob die Grundgesamtheit eine bestimmte Eigenschaft E besitzt. Diese Frage läßt sich als eine Ja-Nein-Frage formulieren, deren Antwort aus der Stichprobe

⁶Für Zufallsvariablen mit stetiger Verteilung ist diese Wahrscheinlichkeit stets 0. In diesem Fall betrachtet man statt dem Produkt der Einzelwahrscheinlichkeiten das Produkt der Werte der Dichtefunktion an den Stichprobenwerten.

durch eine sogenannte Testfunktion

$$\varphi : \mathbb{R}^n \rightarrow \{0, 1\}$$

ermittelt werden soll. Dabei steht der Wert 0 für die Antwort „Nein“, der Wert 1 für die Antwort „Ja“.

Mit diesem Vorgehen lassen sich zwei verschiedene Typen von Fehlern machen. Besitzt die Grundgesamtheit nicht die Eigenschaft E , liefert der Test jedoch die Antwort 1, so spricht man von einem Fehler erster Art. Ein Fehler zweiter Art liegt vor, wenn die Grundgesamtheit die Eigenschaft E besitzt, der Test hingegen 0 ergibt.

Ein Test sollte daher so konstruiert werden, daß diese Fehler nur mit geringer „Wahrscheinlichkeit“ auftreten. Wie bei den Konfidenzintervallen kann auch hier erst von einer Wahrscheinlichkeit gesprochen werden, wenn wir das Testergebnis als Zufallsvariable $\varphi(X_1, X_2, \dots, X_n)$ auffassen (für voneinander unabhängige und identisch zu X verteilte Zufallsvariablen X_1, X_2, \dots, X_n).

Leider lassen sich nicht beide Fehlerwahrscheinlichkeiten gleichzeitig beliebig klein machen. Soll die Wahrscheinlichkeit für einen Fehler erster Art minimiert werden, so sollte der Test immer die Antwort 0 liefern. Dies erhöht jedoch die Wahrscheinlichkeit für einen Fehler zweiter Art. Umgekehrt führt eine Minimierung der Wahrscheinlichkeit für einen Fehler zweiter Art zu einer Erhöhung der Wahrscheinlichkeit für einen Fehler erster Art.

Aus diesem Grund wird eine der beiden Fehlerwahrscheinlichkeiten durch einen vorher gewählten Höchstwert $1 - \alpha$ für ein fest gewähltes $\alpha \in [0, 1]$ begrenzt. Der Test wird dann so konstruiert, daß unter dieser Vorgabe die andere der beiden Fehlerwahrscheinlichkeiten minimal wird.

In der Praxis wird zumeist die Wahrscheinlichkeit für einen Fehler erster Art nach oben beschränkt. Dies hat den Zweck, daß das Testergebnis 1 bei großem α auch mit großer Wahrscheinlichkeit korrekt ist. Wenn also die Eigenschaft E durch den Test festgestellt wird, so kann dieses Resultat als verläßlich angenommen werden.

Der Wert α wird dabei das Niveau des verwendeten Tests genannt. Die Eigenschaft E wird als Alternative bezeichnet, da ihr Gegenteil zumeist die bislang übliche Meinung (auch Standardmeinung, Hypothese oder Nullhypothese genannt) oder die Annahme darstellt, daß keine systematische Abweichung vom Zufalls vorliegt.

Soll beispielsweise getestet werden, ob ein neues Medikament wirksamer ist als ein bisher übliches, so ist die Hypothese „das neue Medikament ist nicht wirksamer als das bisherige“ und die Alternative „das neue Medikament ist wirksamer als das bisherige“. Diese Wahl stellt sicher, daß die Hypothese (bei großem α) nur mit geringer Wahrscheinlichkeit zu Unrecht durch den Test abgelehnt wird.

Für die Anwendung von statistischen Tests in der Auswertungsphase von Experimenten ist es von großer Bedeutung, daß das Niveau der zu verwendenden Tests bereits vor Datenerhebung festgelegt wird. Üblich sind hier 95 % oder sogar 99 %. Bei der Wahl des Niveaus sollte berücksichtigt werden, wieviele Datensätze erhoben werden müssen, um mit den Tests überhaupt die Alternative belegen zu können.

Um diese Anforderung etwas zu lockern, wurden die sogenannten P-Werte eingeführt. Zu jedem Test und jeder Stichprobe läßt sich das kleinste Niveau ermitteln, bei dem ein

dazugehöriger (und nach einer festgelegten Methode konstruierter) Test die Hypothese abgelehnt. Dieses Niveau wird dabei als P-Wert bezeichnet. Dies stellt eine Vereinfachung in der Anwendung von Tests dar, entbindet jedoch nicht von der Notwendigkeit, das Niveau des Tests bereits vor der Erhebung der Daten festzusetzen.

Tests sind somit ein wichtiges Hilfsmittel, um das „Vertrauen“ in die Ergebnisse eines Experimentes zu beziffern. Sie stellen jedoch keinen absoluten Beweis für die Gültigkeit einer Hypothese oder Alternative dar, denn dies ist wegen der oben erwähnten Fehlerwahrscheinlichkeiten nicht möglich.

Studentisches Beispiel

In einer mündlichen Prüfung soll festgestellt werden, welche Note dem Wissensstand der Prüflings angemessen ist. Dazu wird das folgende mathematische Modell entwickelt:

Die Grundgesamtheit ist die Menge aller Fragen, die dem Prüfling gestellt werden können. Dabei wird angenommen, daß alle Fragen voneinander unabhängig sind und den gleichen Schwierigkeitsgrad besitzen.⁷

Der Prüfer kann auf jeder Frage eine „Messung“ ausführen, indem er diese Frage dem Prüfling stellt. Das Ergebnis der Messung ist die Qualität der Antwort des Prüflings, die mittels der üblichen Notenskala

$$S := \{1,0, 1,3, 1,7, 2,0, 2,3, 2,7, 3,0, 3,3, 3,7, 4,0, 4,3, 4,7, 5,0\}$$

angegeben wird (hier ergänzt um die Hilfwerte 4,3 und 4,7, um die „Lücken“ in der Skala zu besetzen). Dabei wird angenommen, daß der Prüfer jeder Antwort objektiv und korrekt auf diese Weise beurteilen kann.

Es wird zudem davon ausgegangen, daß der Prüfling eine „wahre“ Note $\hat{\vartheta}$ besitzt, um deren Wert die Qualität seiner Antworten näherungsweise (da auf obige Skala diskretisiert) normalverteilt ist. Die Varianz dieser Normalverteilung wird als bekannt angenommen. Für alle $\vartheta \in S$ sei X_{ϑ} eine Zufallsvariable mit der zum Parameter ϑ gehörigen Verteilung.

Es ergibt sich, daß die Qualität der Antworten des Prüflings $X_{\hat{\vartheta}}$ eine der in Tabelle 1 aufgeführten Verteilungen besitzt.⁸

Als Beispiel für die Verwendung von Schätzern, Konfidenzintervallen und Tests wird nun die 5-elementige Stichprobe

$$(1,7, 2,7, 3,0, 3,0, 3,0)$$

untersucht.

⁷Diese Annahmen sind selbstverständlich zumindest zweifelhaft, sorgen aber für eine deutliche Vereinfachung der folgenden Berechnungen. Auf die Gültigkeit dieser Annahmen in der Realität wird hier nicht eingegangen.

⁸Die Konstruktion dieser Werte erfolgte dadurch, daß die Wahrscheinlichkeitsdichten von Normalverteilungen jeweils mit Standardabweichung 1,5 für die Erwartungswerte 1, 2, ..., 12 und 13 auf Intervalle der Form $[i - 0,5, i + 0,5]$ mit $i \in \{1, 2, \dots, 13\}$ diskretisiert wurden. Diese Intervalle entsprechen dann den Werten auf der obigen Notenskala.

ϑ	P($X_\vartheta = \dots$)												
	1,0	1,3	1,7	2,0	2,3	2,7	3,0	3,3	3,7	4,0	4,3	4,7	5,0
1,0	0,63	0,21	0,11	0,04	0,01								
1,3	0,37	0,26	0,21	0,11	0,04	0,01							
1,7	0,16	0,21	0,26	0,21	0,11	0,04	0,01						
2,0	0,05	0,11	0,21	0,26	0,21	0,11	0,04	0,01					
2,3	0,01	0,04	0,11	0,21	0,26	0,21	0,11	0,04	0,01				
2,7		0,01	0,04	0,11	0,21	0,26	0,21	0,11	0,04	0,01			
3,0			0,01	0,04	0,11	0,21	0,26	0,21	0,11	0,04	0,01		
3,3				0,01	0,04	0,11	0,21	0,26	0,21	0,11	0,04	0,01	
3,7					0,01	0,04	0,11	0,21	0,26	0,21	0,11	0,04	0,01
4,0						0,01	0,04	0,11	0,21	0,26	0,21	0,11	0,05
4,3							0,01	0,04	0,11	0,21	0,26	0,21	0,16
4,7								0,01	0,04	0,11	0,21	0,26	0,37
5,0									0,01	0,04	0,11	0,21	0,63

Tabelle 1: Die für $X_{\hat{\vartheta}}$ in Frage kommenden Verteilungen (Wahrscheinlichkeiten gerundet).

Konstruiert man einen Maximum-Likelihood-Schätzer für $\hat{\vartheta}$, so liefert dieser (für voneinander unabhängige und identisch zu X_ϑ verteilte Zufallsvariablen $X_{\vartheta,1}, X_{\vartheta,2}, \dots, X_{\vartheta,5}$) die in Tabelle 2 dargestellten Resultate.⁹

Damit liefert der Maximum-Likelihood-Schätzer den Wert 2,7 als Schätzung für $\hat{\vartheta}$.

Bestimmt man nun die Mindestwahrscheinlichkeit (bezogen auf $\vartheta \in S$) dafür, daß der Maximum-Likelihood-Schätzer für $(X_{\vartheta,1}, X_{\vartheta,2}, X_{\vartheta,3}, X_{\vartheta,4}, X_{\vartheta,5})$ den korrekten Parameter ermittelt, so ergibt als Wert etwa 0,51. Mit einer „Sicherheit“ von rund 51 % ist die obige Schätzung also korrekt.¹⁰ Da dies kein besonders großer Wert ist, sind zugehörige Konfidenzintervalle von Interesse.

Ein Konfidenzintervall zum Niveau 50 % enthält gemäß der obigen Feststellung also nur den Wert 2,7. Für das Niveau 95 % läßt sich ein Konfidenzintervall angeben, das durch die Werte 2,3, 2,7 und 3,0 gebildet wird. Benötigt man gar ein Konfidenzintervall zum Niveau 99 %, so ist die Menge $\{2,0, 2,3, 2,7, 3,0, 3,3\}$ ein solches. Hier wird deutlich, daß die Breite guter Konfidenzintervalle mit zunehmender gewünschter „Sicherheit“ steigt. Die Breite eines Konfidenzintervalls zu einem vorgegebenen Niveau läßt sich in der Regel bereits vor der Erhebung der Daten bestimmen. Bei einer Stichprobe vom Umfang 100 beispielsweise hat hier ein gutes Konfidenzintervall zum Niveau 99 % die Breite 1.

Zum Abschluß des Beispiels soll noch zum Niveau 99 % getestet werden, ob $\hat{\vartheta} \geq 4,0$ gilt. Ein Test $\varphi : S^5 \rightarrow \{0, 1\}$ kann beispielsweise auf dem Maximum-Likelihood-Schätzer

⁹Auf eine Herleitung dieser und auch der in der Folge ermittelten Werte wird hier verzichtet. Es handelt sich um numerisch bestimmte Größen.

¹⁰An dieser Stelle sei noch einmal angemerkt, daß hier nicht von einer Wahrscheinlichkeit gesprochen werden kann. Der Parameter $\hat{\vartheta}$ ist entweder gleich 2,7 oder er ist es nicht.

ϑ	$P\left(\left(X_{\vartheta,1}, X_{\vartheta,2}, X_{\vartheta,3}, X_{\vartheta,4}, X_{\vartheta,5}\right) = (1,7, 2,7, 3,0, 3,0, 3,0)\right)$
1,0	$2,1 \cdot 10^{-16}$
1,3	$3,3 \cdot 10^{-12}$
1,7	$6,0 \cdot 10^{-9}$
2,0	$1,3 \cdot 10^{-6}$
2,3	$3,2 \cdot 10^{-5}$
2,7	$9,3 \cdot 10^{-5}$
3,0	$3,2 \cdot 10^{-5}$
3,3	$1,3 \cdot 10^{-6}$
3,7	$6,0 \cdot 10^{-9}$
4,0	$3,3 \cdot 10^{-12}$
4,3	$2,1 \cdot 10^{-16}$
4,7	$1,5 \cdot 10^{-21}$
5,0	$1,3 \cdot 10^{-27}$

Tabelle 2: Gerundete Resultate des Maximum-Likelihood-Schätzers für die Stichprobe aus dem Beispiel.

basieren. Berechnungen ergeben, daß das gewünschte Niveau eingehalten wird, wenn dieser Test als

$$\varphi(x_1, x_2, x_3, x_4, x_5) := \begin{cases} 1 & \text{falls } M_5(x_1, x_2, x_3, x_4, x_5) \geq 3,3, \\ 0 & \text{sonst} \end{cases}$$

definiert ist, wobei M_5 der Maximum-Likelihood-Schätzer für Stichproben vom Umfang 5 ist.¹¹

Hier gilt $\varphi(1,7, 2,7, 3,0, 3,0, 3,0) = 1$. Somit kann $\hat{\vartheta} \geq 4,0$ mit einer „Sicherheit“ von mindestens 99 % angenommen werden.

Und in der Tat wurde die Beispielstichprobe von einem Zufallsgenerator erzeugt, der Realisierungen der Zufallsvariablen $X_{2,7}$ simuliert.

¹¹Die Wahrscheinlichkeit dafür, daß $M_5(X_{4,3,1}, X_{4,3,2}, X_{4,3,3}, X_{4,3,4}, X_{4,3,5})$ einen Wert größer als oder gleich 3,3 annimmt, beträgt nämlich weniger als 1 %. Dies ist gleichzeitig die Wahrscheinlichkeit für einen Fehler erster Art, also das Niveau des Tests.