SS 2004

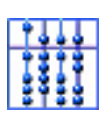
Diskrete Strukturen II

Ernst W. Mayr

Fakultät für Informatik

TU München

http://www14.in.tum.de/lehre/2004SS/ds/index.html.de





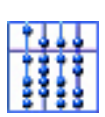

Verschiedene Approximationen der Binomialverteilung

Sei $H_n \sim \mathrm{Bin}(n,p)$ eine binomialverteilte Zufallsvariable mit der Verteilungsfunktion F_n. Für $n \longrightarrow \infty$ gilt

$$F_n(t) = \Pr[H_n/n \leq t/n]$$

$$\rightarrow \quad \Phi\left(\frac{t/n - p}{\sqrt{p(1-p)/n}}\right) = \Phi\left(\frac{t - np}{\sqrt{p(1-p)n}}\right) .$$

Wir können F_n somit für große n durch Φ approximieren. Diese Approximation ist in der Praxis deshalb von Bedeutung, da die Auswertung der Verteilungsfunktion der Binomialverteilung für große n sehr aufwendig ist, während für die Berechnung der Normalverteilung effiziente numerische Methoden vorliegen.

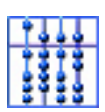


Beispiel: Wenn man die Wahrscheinlichkeit berechnen möchte, mit der bei 10^6 Würfen mit einem idealen Würfel mehr als 500500-mal eine gerade Augenzahl fällt, so muss man eigentlich folgenden Term auswerten:

$$T := \sum_{i=5{,}005\cdot 10^5}^{10^6} \binom{10^6}{i} \left(\frac{1}{2}\right)^{10^6}.$$

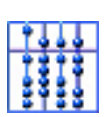
Dies ist numerisch kaum effizient möglich.

Die numerische Integration der Dichte φ der Normalverteilung ist hingegen relativ einfach. Auch andere Approximationen der Verteilung Φ, beispielsweise durch Polynome, sind bekannt. Entsprechende Funktionen werden in zahlreichen Softwarebibliotheken als „black box“ angeboten.



Mit der Approximation durch die Normalverteilung erhalten wir

$$
\begin{aligned}
T &\approx 1 - \Phi\left(\frac{5{,}005 \cdot 10^5 - 5 \cdot 10^5}{\sqrt{2{,}5 \cdot 10^5}}\right) \\
&= 1 - \Phi\left(\frac{5 \cdot 10^2}{5 \cdot 10^2}\right) \\
&= 1 - \Phi(1) \approx 0{,}1573 \; .
\end{aligned}
$$



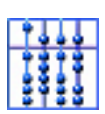
Bei der Approximation der Binomialverteilung mit Hilfe von Korollar 46 führt man oft noch eine so genannte Stetigkeitskorrektur durch. Zur Berechnung von $\Pr[X \leq x]$ für $X \sim \mathrm{Bin}(n,p)$ setzt man

$$\Pr[X \leq x] \approx \Phi\left(\frac{x+0{,}5-np}{\sqrt{np(1-p)}}\right)$$

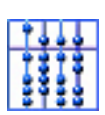
statt

$$\Pr[X \leq x] \approx \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$$

an.

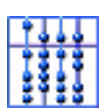


Der Korrekturterm läßt sich in der Histogramm-Darstellung der Binomialverteilung veranschaulichen. Die Binomialverteilung wird dort durch Balken angegeben, deren Fläche in etwa der Fläche unterhalb der Dichte φ von $\mathcal{N}(0,1)$ entspricht. Wenn man die Fläche der Balken mit „$X \leq x$“ durch das Integral von φ approximieren möchte, so sollte man bis zum Ende des Balkens für „$X = x$“ integrieren und nicht nur bis zur Mitte. Dafür sorgt der Korrekturterm $0{,}5$.



Approximationen für die Binomialverteilung

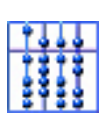
- Approximation durch die Poisson-Verteilung: $\mathrm{Bin}(n,p)$ wird approximiert durch $\mathrm{Po}(np)$. Diese Approximation funktioniert sehr gut für seltene Ereignisse, d. h. wenn np sehr klein gegenüber n ist. Als Faustregel fordert man $n \geq 30$ und $p \leq 0{,}05$.
- Approximation durch die Chernoff-Schranken: Bei der Berechnung der tails der Binomialverteilung liefern diese Ungleichungen meist sehr gute Ergebnisse. Ihre Stärke liegt darin, dass es sich bei den Schranken nicht um Approximationen, sondern um echte Abschätzungen handelt. Dies ist vor allem dann wichtig, wenn man nicht nur numerische Näherungen erhalten möchte, sondern allgemeine Aussagen über die Wahrscheinlichkeit von Ereignissen beweisen möchte.



- Approximation durch die Normalverteilung: Als Faustregel sagt man, dass die Verteilungsfunktion $F_n(t)$ von $\mathrm{Bin}(n,p)$ durch

$$F_n(t) \approx \Phi((t-np)/\sqrt{p(1-p)n})$$

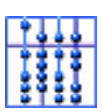
approximiert werden kann, wenn $np \geq 5$ und $n(1-p) \geq 5$ gilt.



3 Induktive Statistik

3.1 Einführung

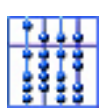
Das Ziel der induktiven Statistik besteht darin, aus gemessenen Zufallsgrößen auf die zugrunde liegenden Gesetzmäßigkeiten zu schließen. Im Gegensatz dazu spricht man von deskriptiver Statistik, wenn man sich damit beschäftigt, große Datenmengen verständlich aufzubereiten, beispielsweise durch Berechnung des Mittelwertes oder anderer abgeleiteter Größen.



3.2 Schätzvariablen

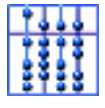
Wir betrachten die Anzahl X von Lesezugriffen auf eine Festplatte bis zum ersten Lesefehler und nehmen an, dass $\Pr[X = i] = (1 - p)^{i-1}p$, setzen also für X eine geometrische Verteilung an. Dahinter verbirgt sich die Annahme, dass bei jedem Zugriff unabhängig und mit jeweils derselben Wahrscheinlichkeit p ein Lesefehler auftreten kann.

Unter diesen Annahmen ist die Verteilung der Zufallsvariablen X eindeutig festgelegt. Allerdings entzieht sich der numerische Wert des Parameters p noch unserer Kenntnis. Dieser soll daher nun empirisch geschätzt werden. Statt p können wir ebensogut $\mathbb{E}[X]$ bestimmen, da wir daraus nach den Eigenschaften der geometrischen Verteilung p mittels $p = \frac{1}{\mathbb{E}[X]}$ berechnen können.

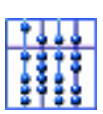


Dazu betrachten wir n baugleiche Platten und die zugehörigen Zufallsvariablen X_i (für $1 \leq i \leq n$), d. h. wir zählen für jede Platte die Anzahl von Zugriffen bis zum ersten Lesefehler. Die Zufallsvariablen X_i sind dann unabhängig und besitzen jeweils dieselbe Verteilung wie X. Wir führen also viele Kopien eines bestimmten Zufallsexperiments aus, um Schlüsse auf die Gesetzmäßigkeiten des einzelnen Experiments ziehen zu können.

Dies ist das Grundprinzip der induktiven Statistik. Die n Messungen heißen Stichproben, und die Variablen X_i nennt man Stichprobenvariablen.



Grundprinzip statistischer Verfahren. Wir erinnern an das Gesetz der großen Zahlen (Satz 23) bzw. den Zentralen Grenzwertsatz (Satz 45). Wenn man ein Experiment genügend oft wiederholt, so nähert sich der Durchschnitt der Versuchsergebnisse immer mehr dem Verhalten an, das man „im Mittel“ erwarten würde. Je mehr Experimente wir also durchführen, umso genauere und zuverlässigere Aussagen können wir über den zugrunde liegenden Wahrscheinlichkeitsraum ableiten. Auf diesem Grundprinzip beruhen alle statistischen Verfahren.



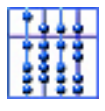
Um $\mathbb{E}[X]$ empirisch zu ermitteln, bietet es sich an, aus den Zufallsvariablen X_i das arithmetische Mittel $\overline{X}$ zu bilden, das definiert ist durch

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Es gilt

$$\mathbb{E}[\overline{X}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X] = \mathbb{E}[X].$$

$\overline{X}$ liefert uns also im Mittel den gesuchten Wert $\mathbb{E}[X]$. Da wir $\overline{X}$ zur Bestimmung von $\mathbb{E}[X]$ verwenden, nennen wir $\overline{X}$ einen Schätzer für den Erwartungswert $\mathbb{E}[X]$. Wegen der obigen Eigenschaft ist $\overline{X}$ sogar ein so genannter erwartungstreuer Schätzer.

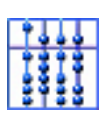


Definition: Gegeben sei eine Zufallsvariable X mit der Dichte $f(x;\theta)$. Eine Schätzvariable oder kurz Schätzer für den Parameter θ der Dichte von X ist eine Zufallsvariable, die aus mehreren (meist unabhängigen und identisch verteilten) Stichprobenvariablen zusammengesetzt ist. Ein Schätzer U heißt erwartungstreu, wenn gilt

$$\mathbb{E}[U] = \theta.$$

Bemerkung:

Die Größe $\mathbb{E}[U - \theta]$ nennt man Bias der Schätzvariablen U. Bei erwartungstreuen Schätzvariablen ist der Bias gleich Null.



Der Schätzer $\overline{X}$ ist also ein erwartungstreuer Schätzer für den Erwartungswert von X. Ein wichtiges Maß für die Güte eines Schätzers ist die mittlere quadratische Abweichung, kurz MSE für mean squared error genannt. Diese berechnet sich durch $MSE := \mathbb{E}[(U - \theta)^2]$. Wenn U erwartungstreu ist, so folgt $MSE = \mathbb{E}[(U - \mathbb{E}[U])^2] = \mathrm{Var}[U]$.

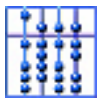
Definition: Wenn die Schätzvariable A eine kleinere mittlere quadratische Abweichung besitzt als die Schätzvariable B, so sagt man, dass A effizienter ist als B.

Eine Schätzvariable heißt konsistent im quadratischen Mittel, wenn MSE $\longrightarrow 0$ für $n \longrightarrow \infty$ gilt. Hierbei bezeichne n den Umfang der Stichprobe.

Für $\overline{X}$ erhalten wir wegen der Unabhängigkeit von $X_1, \ldots, X_n$

$$MSE = \mathrm{Var}[\overline{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{1}{n}\mathrm{Var}[X].$$

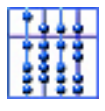


Bei jeder Verteilung mit endlicher Varianz folgt $MSE = \mathcal{O}(1/n)$ und somit $MSE \to 0$ für $n \to \infty$. Der Schätzer $\overline{X}$ ist also konsistent.

Aus der Konsistenz von $\overline{X}$ im quadratischen Mittel können wir mit Hilfe des Satzes von Chebyshev (siehe Satz 22) folgende Konsequenz ableiten. Sei $\varepsilon > 0$ beliebig, aber fest. Dann gilt

$$\Pr[|\overline{X} - \theta| \geq \varepsilon] = \Pr[|\overline{X} - \mathbb{E}[X]| \geq \varepsilon] \leq \frac{\mathrm{Var}[\overline{X}]}{\varepsilon^2} \to 0$$

für $n \to \infty$. Für genügend große n liegen also die Werte von $\overline{X}$ beliebig nahe am gesuchten Wert $\theta = \mathbb{E}[X]$. Diese Eigenschaft nennt man auch schwache Konsistenz, da sie aus der Konsistenz im quadratischen Mittel folgt.

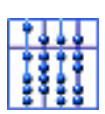


Als nächstes betrachten wir eine weitere von $\overline{X}$ abgeleitete Schätzvariable:

$$S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Wir zeigen, dass S^2 ein erwartungstreuer Schätzer für die Varianz von X ist. Sei $\mu := \mathbb{E}[X] = \mathbb{E}[X_i] = \mathbb{E}[\overline{X}]$.

$$\begin{aligned}
(X_i - \overline{X})^2 &= (X_i - \mu + \mu - \overline{X})^2 \\
&= (X_i - \mu)^2 + (\mu - \overline{X})^2 + 2(X_i - \mu)(\mu - \overline{X}) \\
&= (X_i - \mu)^2 + (\mu - \overline{X})^2 - \frac{2}{n}\sum_{j=1}^{n}(X_i - \mu)(X_j - \mu) \\
&= \frac{n-2}{n}(X_i - \mu)^2 + (\mu - \overline{X})^2 - \frac{2}{n}\sum_{j\neq i}(X_i - \mu)(X_j - \mu).
\end{aligned}$$



Für je zwei unabhängige Zufallsvariablen X_i, X_j mit $i \neq j$ gilt

$$\begin{aligned}\mathbb{E}[(X_i - \mu)(X_j - \mu)] &= \mathbb{E}[X_i - \mu] \cdot \mathbb{E}[X_j - \mu] \\ &= (\mathbb{E}[X_i] - \mu) \cdot (\mathbb{E}[X_j] - \mu) = 0 \cdot 0 = 0.\end{aligned}$$

Daraus folgt

$$\begin{aligned}\mathbb{E}[(X_i - \overline{X})^2] &= \frac{n-2}{n} \cdot \mathbb{E}[(X_i - \mu)^2] + \mathbb{E}[(\mu - \overline{X})^2] \\ &= \frac{n-2}{n} \cdot \mathrm{Var}[X_i] + \mathrm{Var}[\overline{X}].\end{aligned}$$

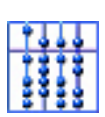
Wegen $\mathrm{Var}[X_i] = \mathrm{Var}[X]$ und $\mathrm{Var}[\overline{X}] = \frac{1}{n}\mathrm{Var}[X]$ folgt nun

$$\mathbb{E}[(X_i - \overline{X})^2] = \frac{n-1}{n} \cdot \mathrm{Var}[X],$$

und somit gilt für S^2

$$\begin{aligned} \mathbb{E}[S^2] &= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[(X_i - \overline{X})^2] \\ &= \frac{1}{n-1} \cdot n \cdot \frac{n-1}{n} \cdot \mathrm{Var}[X] = \mathrm{Var}[X]. \end{aligned}$$

S^2 ist also eine erwartungstreue Schätzvariable für die Varianz von X.



Die vorangegangene Rechnung erklärt, warum man als Schätzer nicht

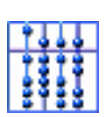
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \overset{!}{=} S^2$$

verwendet, wie man vielleicht intuitiv erwarten würde.

Definition: Die Zufallsvariablen

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ und } S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

heißen Stichprobenmittel bzw. Stichprobenvarianz der Stichprobe $X_1, \ldots, X_n$. $\overline{X}$ und S^2 sind erwartungstreue Schätzer für den Erwartungswert bzw. die Varianz.



Maximum-Likelihood-Prinzip zur Konstruktion von Schätzvariablen

Wir betrachten nun ein Verfahren zur Konstruktion von Schätzvariablen für Parameter von Verteilungen. Sei

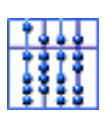
$$\vec{X} = (X_1, \ldots, X_n).$$

Bei $X_1, \ldots, X_n$ handelt es sich um unabhängige Kopien der Zufallsvariablen X mit der Dichte $f(x;\theta)$. Hierbei sei θ der gesuchte Parameter der Verteilung. Wir setzen

$$f(x;\theta) = \Pr[X = x],$$

wobei θ ein Parameter der Verteilung ist.

Wenn wir den Parameter explizit angeben wollen, so schreiben wir dafür auch $f(x;\theta) = \Pr_\theta[X = x]$. Eine Stichprobe liefert für jede Variable X_i einen Wert x_i. Diese Werte fassen wir ebenfalls zu einem Vektor $\vec{x} = (x_1, \ldots, x_n)$ zusammen.



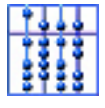
Der Ausdruck

$$L(\vec{x};\theta) \quad := \quad \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \Pr_\theta[X_i = x_i]$$

$$\overset{\text{unabh.}}{=} \quad \Pr_\theta[X_1 = x_1, \ldots, X_n = x_n]$$

entspricht der Wahrscheinlichkeit, dass wir die Stichprobe $\vec{x}$ erhalten, wenn wir den Parameter mit dem Wert θ belegen.

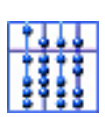
Wir betrachten nun eine feste Stichprobe $\vec{x}$ und fassen $L(\vec{x};\theta)$ somit als Funktion von θ auf. In diesem Fall nennen wir L die Likelihood-Funktion der Stichprobe.



Es erscheint sinnvoll, zu einer gegebenen Stichprobe $\vec{x}$ den Parameter θ so zu wählen, dass $L(x;\theta)$ maximal wird.

Definition: Ein Schätzwert $\widehat{\theta}$ für den Parameter einer Verteilung $f(x;\theta)$ heißt Maximum-Likelihood-Schätzwert (ML-Schätzwert) für eine Stichprobe $\vec{x}$, wenn gilt

$$L(\vec{x};\theta) \leq L(\vec{x};\widehat{\theta}) \text{ für alle } \theta.$$



Beispiel: Wir konstruieren mit der ML-Methode einen Schätzer für den Parameter p der Bernoulli-Verteilung. Es gilt $\Pr_p[X_i = 1] = p$ und $\Pr_p[X_i = 0] = 1 - p$. Daraus schließen wir, dass $\Pr_p[X_i = x_i] = p^{x_i}(1-p)^{1-x_i}$, und stellen die Likelihood-Funktion

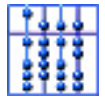
$$L(\vec{x}; p) = \prod_{i=1}^{n} p^{x_i} \cdot (1-p)^{1-x_i}$$

auf.

Wir suchen als Schätzer für p den Wert, an dem die Funktion L maximal wird. Wir erhalten

$$\begin{aligned} \ln L(\vec{x}; p) &= \sum_{i=1}^{n} (x_i \cdot \ln p + (1-x_i) \cdot \ln(1-p)) \\ &= n\bar{x} \cdot \ln p + (n - n\bar{x}) \cdot \ln(1-p). \end{aligned}$$

Hierbei bezeichnet $\bar{x}$ das arithmetische Mittel $\frac{1}{n}\sum_{i=1}^{n} x_i$.



Wir finden das Maximum durch Nullsetzen der Ableitung:

$$\frac{\mathrm{d}\ln L(\vec{x};p)}{\mathrm{d}p} = \frac{n\bar{x}}{p} - \frac{n-n\bar{x}}{1-p} = 0.$$

Diese Gleichung hat die Lösung $p = \bar{x}$.