Statistische Verfahren

Schumacher

Semester: WS 2008/09

Vorwort

Dieses Dokument wurde als Skript für die auf der Titelseite genannte Vorlesung erstellt und wird jetzt im Rahmen des Projekts „Vorlesungsskripte der Fakultät für Mathematik und Informatik“ weiter betreut. Das Dokument wurde nach bestem Wissen und Gewissen angefertigt. Dennoch garantiert weder der auf der Titelseite genannte Dozent, die Personen, die an dem Dokument mitgewirkt haben, noch die Mitglieder des Projekts für dessen Fehlerfreiheit. Für etwaige Fehler und dessen Folgen wird von keiner der genannten Personen eine Haftung übernommen. Es steht jeder Person frei, dieses Dokument zu lesen, zu verändern oder auf anderen Medien verfügbar zu machen, solange ein Verweis auf die Internetadresse des Projekts `http://uni-skripte.lug-jena.de/` *enthalten ist.*

Diese Ausgabe trägt die Versionsnummer 2589 und ist vom 4. Dezember 2009. Eine neue Ausgabe könnte auf der Webseite des Projekts verfügbar sein.

Jeder ist dazu aufgerufen, Verbesserungen, Erweiterungen und Fehlerkorrekturen für das Skript einzureichen bzw. zu melden oder diese selbst einzupflegen – einfach eine E-Mail an die Mailingliste `<uni-skripte@lug-jena.de>` *senden. Weitere Informationen sind unter der oben genannten Internetadresse verfügbar.*

Hiermit möchten wir allen Personen, die an diesem Skript mitgewirkt haben, vielmals danken:

- *Jens Kubieziel* `<jens@kubieziel.de>` *(2008)*

Inhaltsverzeichnis

Auflistung der Theoreme

Sätze

Definitionen und Festlegungen

1. Einführung

Der Studierende soll ein Gefühl für die statistische Modellierung bekommen. Es sollen viele verschiedene Anwendungsgebiete gezeigt werden.

Die Webseite zur Vorlesung ist http://www.stochastik.uni-jena.de/ unter dem Stichpunkt „Lehre“. Derzeit ist dort eine Literaturliste zu finden. Wir werden in mit dem Statistikpaket GNU R beschäftigen.

Beispiel 1.1 (Schätzung der Größe einer biologischen Population)
Tiere kann man im Gegensatz zum Menschen nicht einfach abzählen. Daher werden hier statistische Verfahren benutzt. Beispielsweise wird hierfür das Fang-Wiederfang-Verfahren genutzt.

Wir haben eine Population der Größe N. In der ersten Fangperiode werden Individuen markiert. Die Anzahl der markierten Individuen werde mit m bezeichnet. In der zweiten Fangperiode werden Individuen gefangen. Die Anzahl der gefangenen Individuen werde mit c bezeichnet. Davon ist eine Anzahl r markiert.

Etwa-Zeichen prüfen

Daraus kann man sich einen heuristischen Schätzer für die Populationsgröße überlegen. Es ist $m/N \sim r/c$. Die Formel wird nach N umgestellt und es ergibt sich $\hat{N} = \frac{m \cdot c}{r}$. Der Schätzer heißt **Peterson-Lincoln-Schätzer**.

1.1. Statistische Modellierung

Wir nutzen ein Urnenmodell mit N Kugeln. Davon sind m rot und es werden c Kugeln entnommen. Dabei ist X die Anzahl der markierten Kugeln unter den c entnommenen.

Folgende Voraussetzungen müssen gegeben sein:

- gute Durchmischung der Population
- Ziehen mit Zurücklegen, d. h. gefangene Tiere werden wieder zurück in die Population gegeben
- Population ist im Untersuchungszeitraum geschlossen (keine Änderung der Größe)
- kein Verlust der Markierung

Es ist X mit den Parametern c und $p = {}^{m}/_{N}$ binomialverteilt. Dann ist $P(X = r) = \binom{c}{r}({}^{m}/_{N})^{r}(1 - {}^{m}/_{N})^{c-r}$. Um das N zu erhalten, verwenden wir die Maximum-Likelihood-Methode.

1.2. Parameterschätzung

Wir fassen $P(X = r)$ als Funktion des unbekannten Parameters N auf. Dies nennt sich **Likelihood-Funktion**: $\mathcal{L}(N|r) = \binom{c}{r}({}^{m}/_{N})^{r}(1 - {}^{m}/_{N})^{c-r}$. Das Maximum der Likelihood-Funktion bestimmen wir durch Ableitung nach N. Jedoch ist die Ableitung nicht unbedingt einfach. Ein leichterer Weg ist durch Übergang zur Log-Likelihood-Funktion: $\ell(N|r) = \log\binom{c}{r} + r\log(m) - r\log N + (c-r)\log(N-m) - (c-r)\log N$. Die Ableitung ergibt: $\frac{\partial}{\partial N}\ell(N|r) = -{}^{r}/_{N} + \frac{(c-r)}{N-m} - \frac{(c-r)}{N} = \frac{(c-r)}{N-m} - \frac{c}{N} = 0$. Durch Umstellen ergibt sich $\hat{N} = \frac{m \cdot c}{r}$.

Neben dem Schätzer haben wir hier auch eine Aussage über die Verteilung erhalten: $\hat{N} = \frac{m \cdot c}{X}$. Vom X wissen wir, dass es binomialverteilt ist. Somit ist $\hat{N}$ eine Zufallsvariable, deren Verteilung eine Transformation der Binomialverteilung ist. Darin liegt der Vorteil des statistischen Modells gegenüber der obigen heuristischen Variante.

Dadurch können wir prinzipiell statistische Eigenschaften des Schätzers untersuchen, wie beispielsweise Erwartungstreue, Konfidenzintervalle.

1.3. Aussagen über die Verteilung von Parameterschätzern

Die Aussagen sind nicht immer exakt, sondern in vielen Fällen asymptotisch. Wir betrachen allgemeine Aussagen zu Eigenschaften der Maximum-Likelihood-Schätzer.

Das obige Beispiel kann erweitert werden. So kann es mehr als zwei Fangperioden geben. Wie lassen sich nun drei Fangperioden statistisch modellieren? Dazu verwendet man „Fanggeschichten“. Dies sind Folgen von Nullen und Einsen, die jedes Individuum zugeordnet bekommt. Eine Geschichte von 111 bedeutet, dass das Individuum in allen drei Perioden gefangen wurde. Ein Individuum mit der Markierung 001 wurde nur in der letzten Periode gefangen. Insgesamt existieren acht solcher Fanggeschichten.

$$
\begin{aligned}
111 &= X_1 \\
110 &= X_2 \\
101 &= X_3 \\
011 &= X_4 \\
100 &= X_5 \\
010 &= X_6 \\
001 &= X_7 \\
000 &\quad \text{unbekannt}
\end{aligned}
$$

Wir kennen die Variablen $X_1, \ldots, X_7$, für $X_8 = N - (X_1 + \cdots + X_7)$. Es gilt, $P((X_1 = n_1, \ldots, X_7 = n_7, X_8 = n_8)) = \frac{N!}{n_1! \cdot \ldots \cdot n_8!} q_1^{n_1} q_2^{n_2} \ldots q_8^{n_8}$. Dies bezeichnet man als **Multinomialverteilung**. Dabei ist q_i die Wahrscheinlichkeit, dass ein zufällig gewähltes Individuumm die Fanggeschichte i hat. Die Likelihood-Funktion dazu ist: $\mathcal{L}(N, q_1, \ldots, q_8 | n_1, \ldots, n_7) = \frac{N!}{n_1! \cdot \ldots \cdot n_7! (N - n_1 - \cdots - n_7)!} q_1^{n_1} q_2^{n_2} \ldots q_7^{n_7} q_8^{N - n_1 - \cdots - n_7}$. Hier haben wir nun mehr Parameter als Beobachtugen. Daher muss das Modell ein wenig reduziert werden.

Es ist q_1 die Wahrscheinlichkeit, dass das Individuumm die Fanggeschichte 111 hat. Wir nehmen an, dass p die konstante Fangwahrscheinlichkeit für alle Individuen in allen Fangperioden ist. Dann ist $q_1 = p \cdot p \cdot p = p^3, q_2 = pp(1-p) = p^2(1-p), \ldots$ und die Likelihood-Funktion ist $\mathcal{L}(N, p | n_1, \ldots, n_7) = \frac{N!}{n_1!, \ldots, n_7! (N - n_1 - \cdots - n_7)!} p^{3n_1 + 2n_2 + 2n_3 + 2n_4 + n_5 + n_6 + n_7} (1 - p)^{n_2 + n_3 + n_4 + 2n_5 + 2n_6 + 2n_7 + 3(N - n_1 - \cdots - n_7)}$. Hierfür gibt es jedoch *keine* analytische Lösung. Das Modell bezeichnet man als M_0.

Zur Verschärfung kann man die Annahme treffen, dass es drei unterschiedliche Fangwahrscheinlichkeiten, je nach Fangperiode, gibt. Dann ist $q_1 = p_1 p_2 p_3, q_2 = p_1 p_2 (1 - p_3), q_3 = p_1 (1 - p_2) p_3, \ldots$ und die Likelihood-Funktion ist $\mathcal{L}(N, p_1, p_2, p_3 | n_1, \ldots, n_7) = \frac{N!}{n_1!, \ldots, n_7! (N - n_1 - \cdots - n_7)!} p_1^{n_1 + n_2 + n_3 + n_5} (1 - p_1)^{n_4 + n_6 + n_7 + (N - n_1 - \cdots - n_7)} p_2^{n_1 + n_2 + n_4 + n_6} (1 - p_2)^{n_3 + n_5 + n_7 + (N - n_1 - \cdots - n_7)} p_3)^{n_2 + n_5 + n_6 + (N - n_1 - \cdots - n_7)}$. Das Modell bezeichnet man als M_t.

Die Wahl des richtigen Modells steht an zentraler Stelle der statistischen Modellierung.

1.4. Test der Hypothesen

Zum Abschluss steht der Test der Hypothesen.

Also ergeben sich die folgenden sechs Punkte:

1. Statistische Modellbildung
2. Parameterschätzung

3. Verteilung von Parameterschätzern
4. Modellanpassung
5. Test der Hypothesen
6. Konfidenzintervalle/Prognoseintervalle

1.5. Beispiele für statistische Modelle

Beispiel 1.2 (Körpergewicht von Säuglingen)
Man betrachtet das Körpergewicht von Säuglingen im ersten Lebensjahr. Also Ergebnis von Messungen ergeben sich Punktepaare (x_i, y_i) für $i = 1, \dots, n$. Dabei steht das x_i für das Alter in Tagen und das y_i für das Körpergewicht. Wir fassen das y_i als Realisierung einer zufälligen Größe Y_i auf. Es ist $y_i = b_1 + \beta_2 x_i + \varepsilon_i$ eine lineare Funktion. Das ε_i wird als **Störterm** bezeichnet. Typischerweise nimmt man an, dass die ε_i unabhängig und identisch verteilt sind. Die Standardannahme für die Verteilung ist $\varepsilon_i \sim N(O, \sigma^2)$. Das Modell wird als **einfache lineare Regression** bezeichnet.

Die Zufallsgröße Y_i hat einen Erwartungswert $\mathbb{E}Y_i = \mu_i = \beta_1 + \beta_2 x_i$ (deterministischer Teil). Im statistischen Teil ergibt sich $Y_i \sim N(\mu_i, \sigma^2)$.

Beispiel 1.3 (SO_2-Belastung der Luft)
Wir haben y_i als den SO_2-Gehalt der Luft. Weiter ist x_{i2} die mittlere Jahrestemperatur, x_{i3} Anzahl der Betriebe mit mehr als 20 Beschäftigten, x_{i4} Einwohnerzahl, x_{i5} mittlerer Jahresniederschlag, x_{i6} mittlere Windgeschwindigkeit und x_{i7} die mittlere Anzahl von Regentagen. Als Daten haben wir nun Angaben in Form von Vektoren $(y_i, x_{i2}, \dots, x_{i7})$. Die Zufallsvariable ist im wesentlichen $Y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_7 x_{i7} + \varepsilon_i$. Die ε_i sind wieder unabhängig und identisch verteilt und es gilt $\varepsilon_i \sim N(O, \sigma^2)$.

Beispiel 1.4 (Geburtsgewicht von Säuglingen)
Diese werden getrennt nach ♂und ♀. Es ist

$$\mu_i = \begin{cases} \beta_{11} + \beta_{21} x_i & \text{für } ♂ \\ \beta_{12} + \beta_{22} x_i & \text{für } ♀ \end{cases}$$

Eine typische Hypothese wäre $H_0 \colon \beta_{21} = \beta_{22}$ oder $H_0 \colon \beta_{11} = \beta_{12}$. Es wird eine so genannte Dummyvariable eingeführt. Das ist eine Indikatorvariable mit $\mathbb{1}_i = \begin{cases} 1 & \text{falls } ♀ \\ 0 & \text{sonst} \end{cases}$.
Damit folgt $\mu_i = \beta_1 + \beta_2 \mathbb{1}_i + \beta_3 x_i + \beta_4 \mathbb{1}_i x_i$, d. h.

$$\mu_i = \begin{cases} \beta_1 + \beta_3 & \text{für } ♂ \\ (\beta_1 + \beta_2) + (\beta_3 + \beta_4) x_i & \text{sonst} \end{cases}$$

Damit können wir die Hypothesen einfach übersetzen: $H_0 \colon \beta_4 = 0$ bzw. $H_0 \colon \beta_2 = 0$. Das wird klassisch als **Kovarianzanalyse** bezeichnet.

Beispiel 1.5 (Reaktionszeit von Busfahrern)
Es wird die Reaktionszeit von Busfahrern auf unterschiedlichen Linien betrachtet. Das $y_i = \alpha_k + \varepsilon_{ik}$ mit $k = 1, \dots, K$ der Nummer der Linie und $i = 1, \dots, n$ Numerierung der Busfahrer. Es ist $\mathbb{E}Y_i = \alpha_k = \begin{cases} \alpha_1 & k = 1 \\ \vdots & \vdots \\ \alpha_k & k = K \end{cases}$. Die Prüfung wird wiederum mit einer Dummyvariable vorgenommen. Es ist $\mathbb{1}_{ik} = \begin{cases} 1 & \text{Fahrer } i \text{ fährt Linie } k \\ 0 & \text{sonst} \end{cases}$. Dann ist $Y_i = \sum_{k=1}^{K} \alpha_k \mathbb{1}_{ik} + \varepsilon_i$. Die Annahme für die Verteilung von ε_i ist $\varepsilon_i \sim N(O, \sigma^2)$ unabhängig und identisch verteilt. Es ist $\mu_i = \sum_{k=1}^{K} \alpha_k \mathbb{1}_{ik}$. Es interessiert die Hypothese $H_0 \colon \alpha_1 = \dots = \alpha_k$. Dieser Ansatz heißt **Varianzanalyse**.

Beispiel 1.6 (Vorkommen der blauflügligen Ödlandschrecke (Oedipoda caerulescens))
Die Zielgröße y_i ist das Vorkommen. Das kann nur die Werte 0 oder 1 annehmen. Weiter bezeichnet x_i den Offenbodenanteil. Das bisherige Herangehen muss hier geändert werden. Denn das y kann nur die Werte 0 oder 1 annehmen. Der obige Ansatz kann aber immer beliebige Werte annehmen.

Jetzt beginnen wir mit dem stochastischen Teil. Wir nutzen die Binomialverteilung $y_i \sim \text{Bin}(1, \mu_i)$ mit $\mu_i \in (0, 1)$. Im deterministischen Teil ergibt sich $\mu_i = h(\underbrace{\beta_1 + \beta_2 x_i}_{\textbf{linearer Prädiktor}})$.
Das h heißt **Responsefunktion**. Was wäre eine vernünftige Responsefunktion? Sie sollte monoton, stetig sein und Wert in $(0, 1)$ haben. Vernünftige Kandidaten sind stetige Verteilungsfunktionen. Typisch wäre $h = \Phi(\cdot)$, also die Verteilungsfunktion der Standard-Normalverteilung oder $h(x) = \frac{1}{1+e^{-x}}$ Verteilungsfunktion der logistischen Verteilung. Diese Klasse von Modellen werden als verallgemeinerte lineare Modelle bezeichnet.

Beispiel 1.7 (Entwicklung des Körpergewichts von mehreren Säuglingen im ersten Lebe
Wir setzen y_{ij} das Körpergewicht von Kind i zum Zeitpunkt x_j.

Graf mit verschiedenen Gewichtspunkten, x_j aufd x-Achse und y_{ij} auf y-Achse

Es ist $y_{ij} = \beta_1 + \beta_2 x_j + \varepsilon_{ij}$. Dann ist nicht Anname, dass die ε_{ij} unabhängig und identisch verteilt sind, *nicht vernünftig*. Wir haben eine Gruppierung der Daten. Das führt zu stochastischer Abhängigkeit. Die Einführung von Dummyvariablen birgt ein Problem. Es ist zwar für jedes Kind möglich. Aber wir wollen keine Aussage für *jedes* Kind, sondern eine allgemeine Aussage wie sich die Kinder im ersten Lebensjahr entwickeln.

Hierfür nutzt man gemischte Modelle. Dabei wird mehr Zufall eingeführt, indem man die individuellen Unterschiede als zufällig modellieren. Wir haben $y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})x_{ij} + \varepsilon_{ij}$. In diesem Modell können wir nun sagen, $\varepsilon_{ij} \sim N(O, \sigma^2), b_{1i} \sim N(O, \sigma_1^2), b_{2i} \sim N(O, \sigma_2^2)$. Zwischen den Größen b_{1i} und b_{2i} gibt es eine Korrelation $\mathrm{cov}(b_{1i}, b_{2i}) = \rho\sigma_1\sigma_2$. Aber die b_{1i}, b_{2i} sind unabhängig von den ε_{ij}.

2. Exponentialfamilien

Definition 2.1
Ein Zufallsvektor $Y = (Y_1, \ldots, Y_n)$ besitz eine Verteilung aus einer **Exponentialfamilie**, falls er eine Dichte bzw. Wahrscheinlichkeitsfunktion besitzt, die sich in der Form

$$f(y|\Theta) = f((y_1, \ldots, y_n)|(\Theta_1, \ldots, \Theta_n)) = h(y) \exp\{\sum_{i=1}^{k} \eta_i(\Theta) T_i(y) - c(\Theta)\}$$

darstellen lässt. Die Familie heißt ***k*-parametrige Exponentialfamilie**, falls k die kleinste Zahl, für die so eine Darstellung möglich ist, ist.

A. Übungsaufgaben

1. Überlegen Sie sich, welche Arten der Verteilung Sie kennen.
   - Normalverteilung
   - Exponentialverteilung
   - Binomialverteilung
   - Poissonverteilung
   - Gleichverteilung
   - Gammaverteilung
   - hypergeometrische Verteilung
   - Bernoulliverteilung
   - χ^2-Verteilung
   - logistische Verteilung
   - Fischersche F-Verteilung
   - Studentsche t-Verteilung
   - Lorentzverteilung
   - Erlangverteilung
   - Weibullverteilung
   - Boltzmannverteilung
   - Multinomialverteilung
   - geometrische Verteilung

Literaturverzeichnis

[1] Math mode von Herbert Voss,
http://www.dante.de/CTAN//info/math/voss/mathmode/Mathmode.pdf

[2] Short Math Guide for LaTeX,
ftp://ftp.ams.org/pub/tex/doc/amsmath/short-math-guide.pdf

Index