Ausreißer bei ballistischen Versuchen und die Normalverteilung

Prof. Dr. Andreas Rudolph
Universität der Bundeswehr
München
WE 2 Mathematik und Informatik
FB BW
Werner-Heisenberg-Weg 39
85577 Neubiberg
Email: Andreas.Rudolph@unibw.de

17. Juni 2011

Inhaltsverzeichnis

1 Einleitung

In diesem Dokument wollen wir folgende Ausgangssituation untersuchen:

Es soll auf eine Scheibe geschossen werden, bei der ein Ursprung mit den Koordinaten $\boldsymbol{\mu}_0 = (\mu_1; \mu_2)^T$ festgelegt ist.

Diese Scheibe stehe in einer festgelegten Entfernung d_0.

Eine Waffe soll aufgrund ihrer Visiereinstellung so auf d_0 eingeschossen sein, dass sie prinzipiell in den Ursprung $\boldsymbol{\mu}_0$ trifft, wenn man aus der Entfernung d_0 auf die Scheibe schießt.

Da die Waffe aufgrund von Zufallseinflüssen nicht bei jedem Schuss exakt in den Ursprung treffen kann, somit die Einschläge $\mathbf{X}_i = (X_{i1}; X_{i2})^T$ für $i = 1, \ldots, n$ um den Ursprung streuen[1], unterstellt man für die Verteilung der Zufallsvektoren (siehe z. B. Hauck (1982), S. 117) eine zweidimensionale Normalverteilung.

[1] Wir fassen somit die Einschläge als zweidimensionale Zufallsvektoren auf

Hiermit ergeben sich folgende Fragen:

1.) Sind die Einschläge kompatibel mit der Annahme einer zweidimensionalen Normalverteilung bzw. wie kann man dies gegebenenfalls überprüfen oder weichen sie doch erheblich davon ab, und insbesondere

2.) wenn es Ausreißer gibt, wie kann man diese identifizieren?

Diese beiden Fragestellungen wollen wir in den nachfolgenden Abschnitten untersuchen.

2 Die zweidimensionale Normalverteilung

Um die obige Fragestellungen beantworten zu können, muss zuerst eine sogenannte Verteilungsannahme getroffen werden.

Hier unterstellt man für die Koordinaten $\mathbf{X} = (X_1, X_2)^T$ der Einschläge auf der Scheibe eine sogenannte zweidimensionalen Normalverteilung.

Eine derartige zweidimensionale Normalverteilung ist durch ihre Dichte festgelegt. Sie lautet

$$f(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \det(\boldsymbol{\Sigma})}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})) \quad \text{mit } \mathbf{x} = (x_1; x_2)^T$$

siehe Morrison (1976), S. 86 ff.

Hierbei ist $\boldsymbol{\Sigma}$ die sogenannte Kovarianzmatrix, sie gibt die Abhängigkeiten innerhalb des Zufallsvektors $\mathbf{X}$ an (d. h. zwischen den beiden Komponenten des Zufallsvektors) und ist eine symmetrische 2×2-Matrix, und $\boldsymbol{\mu}$ der Erwartungswertvektor, dieser gibt den Schwerpunkt der Verteilung an, somit den Punkt auf der Scheibe, an dem sich die Einschläge konzentrieren, und ist damit ein zweidimensionaler Vektor.

Allgemein besteht der Zufallsvektor einer multivariaten Normalverteilung aus p Komponenten, damit ist die Kovarianzmatrix $\boldsymbol{\Sigma}$ eine $p \times p$-Matrix und der Erwartungswertvektor $\boldsymbol{\mu}$ ein p-dimensionaler Vektor. An den statistischen Bedeutungen ändert sich allerdings nichts.

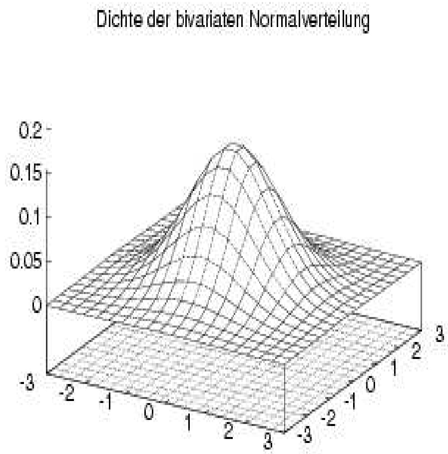
Ihre Dichte sieht dann wie folgt aus

$$f(x_1, \ldots, x_p) = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})) \quad \text{mit } \mathbf{x} = (x_1, \ldots, x_p)^T$$

Die Dichte dient dazu, gegebenenfalls die Wahrscheinlichkeit auszurechnen, dass sich innerhalb eines Rechtecks $[a_1; b_1] \times [a_2; b_2]$ Einschläge befinden (dazu müssen allerdings die Kovarianzmatrix $\boldsymbol{\Sigma}$ und der Erwartungswertvektor $\boldsymbol{\mu}$ bekannt sein oder zumindest vorab geschätzt worden sein). Hierzu wird das folgende Doppelintegral ausgewertet:

$$Prob((X_1, X_2) \in [a_1; b_1] \times [a_2; b_2]) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_2 dx_1$$

Die Dichte der zweidimensionalen Normalverteilung hat in etwa folgendes Aussehen:



3 Der Q-Q-Plot

Um nun die Verteilungsannahme der zweidimensionalen Normalverteilung zu testen, gehen wir davon aus, dass n unabhängig und identisch verteilte zweidimensionale Zufallsvektoren $\mathbf{x}_1, \ldots,$ $\mathbf{x}_n$ vorliegen.

Hieraus berechnen wir zunächst den Schwerpunkt

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

und die empirische Kovarianzmatrix $\mathbf{S}$ als Schätzer für die Kovarianzmatrix $\mathbf{\Sigma}$:

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

siehe hierzu Hartung and Elpelt (1984), S. 602 ff.

Hiermit berechnen wir die sogenannten Mahalanobis Abstände

$$d_i := (\mathbf{x}_i - \overline{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \quad \text{für } i = 1, \ldots, n$$

Sind nun die Daten multivariat normalverteilt, dann genügen die Abstände d_i näherungsweise einer χ^2-Verteilung mit (allgemein: p in unserem zweidimensionalen Fall) 2 Freiheitsgraden.

Nun ordnet man die Abstände der Größe nach aufsteigend an

$$d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(n)}$$

und bestimmt die $\frac{i}{n+1}$-Fraktile

$$\chi^2_{p;\frac{i}{n+1}}$$

der χ^2-Verteilung mit (allgemein: p in unserem zweidimensionalen Fall) 2 Freiheitsgraden.

Unter der Voraussetzung einer multivariaten Normalverteilung sollten die Punkte

$$(\chi^2_{p;\frac{i}{n+1}}; d_{(i)}), \quad i = 1, \ldots, n$$

in etwa auf einer Gerade mit Steigung 1 durch den Ursprung liegen (siehe Hartung and Elpelt (1984), S. 603).

Wesentliche Abweichungen von der Normalverteilung schlagen sich dann in Abweichungen von der Geraden nieder.

Sollten in den Daten Ausreißer auftreten, dann wird dies im dem Q-Q-Plot dadurch deutlich, dass die zugehörigen Punkte “sehr weit“ nach oben von der Geraden liegen.

4 Eine Anwendung

Hier ein Datensatz, bei dem 10 Schuss auf die Scheibe abgegeben wurden (Entfernung 100 Meter, Kaliber .308 TIG):

x-Koordinate in mm	y-Koordinate in mm
3	6
9	9
13	24
18	−14
−9	5
−22	−2
−19	−11
−14	−22
−45	−49
0	0,1

Damit ist der Schwerpunkt gleich $\overline{\mathbf{x}} = (-6{,}6; -5{,}39)^T$.

Hieraus ergibt sich für die Schätzung der Kovarianzmatrix $\mathbf{S}$ und damit für deren Inverse:

$$\mathbf{S} = \begin{pmatrix} 366{,}04 & 280{,}47 \\ 280{,}47 & 403{,}72 \end{pmatrix} \Rightarrow \mathbf{S}^{-1} = \frac{1}{69114{,}24} \begin{pmatrix} 403{,}72 & -280{,}47 \\ -280{,}47 & 366{,}04 \end{pmatrix}$$

Hieraus ergibt sich für die Abstände $d_{(1)}, \ldots, d_{(10)}$:

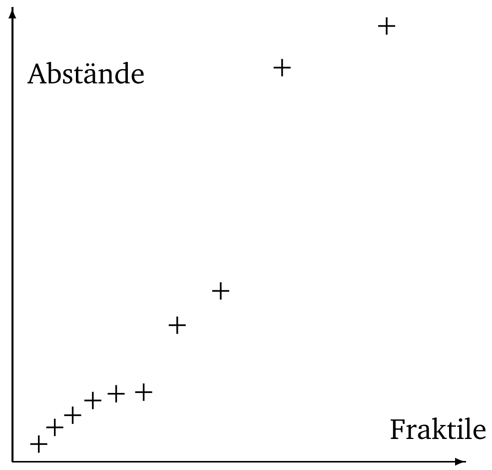
$$d_{(1)} = 0,119 \leq d_{(2)} = 0,339 \leq d_{(3)} = 0,499 \leq d_{(4)} = 0,695 \leq d_{(5)} = 0,784 \leq d_{(6)} = 0,807$$

$$\leq d_{(7)} = 1,689 \leq d_{(8)} = 2,144 \leq d_{(9)} = 5,093 \leq d_{(10)} = 5,642$$

Wir ordnen nun den Fraktilen $\chi^2_{2;\frac{i}{n+1}}$ die Mahalanobis-Abstände zu und erhalten den nachstehenden Datensatz (die erste Spalte beinhaltet die Fraktile, die zweite Spalte die Abstände):

Fraktile	Abstände
0,1906	0,119
0,4013	0,339
0,6369	0,499
0,9040	0,695
1,2123	0,784
1,5769	0,807
2,0232	1,689
2,5986	2,144
3,4095	5,093
4,7958	5,642

Plottet man nun Fraktile gegen Abstände, so erhält man folgende Darstellung:



Die letzten beiden Punkte lassen zwei Ausreißer vermuten, die anderen Punkte bestätigen in etwa die Annahme einer zweidimensionalen Normalverteilung.

Literatur

J. Hartung and B. Elpelt. *Multivariate Statistik*. Oldenbourg, 1984.

G. Hauck. *Der Flug ungelenkter Geschosse und Raketen*. Militärverlag der Deutschen Demokratischen Republik, 1982.

D. F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill Book Company, 1976.