ANOVA zu Fuß

Kim Boström

24. September 2010

Wenn man keine Rohdaten hat, sondern nur Mittelwerte und Standardabweichungen, oder wenn man eine variable Anzahl von Messwiederholungen hat, dann kann man MATLABs ANOVA-Funktionen nicht benutzen. Aber es gibt eine Rettung: die ANOVA zu Fuß ausrechnen! Das wollen wir im folgenden tun und zwar für die einfaktorielle und zweifaktorielle ANOVA.

1 Die Idee dahinter

Erhobene Daten hängen von bestimmten Einflüssen, den *Faktoren* ab. Alle Faktoren, die einem unbekannt sind oder die sich der Kontrolle entziehen, fasst man zusammen zum Faktor *Zufall*. Die Streuung der Daten, die sich aus dem Einfluss dieses unbekannten Faktors ergibt, bezeichnet man als *zufällige* oder *unerklärte Streuung*. Jeder weitere Faktor, den man kennt oder kontrollieren kann, verursacht eine nicht-zufällige oder *erklärte Streuung*. Ebenso verursacht jede Kombination von bekannten oder kontrollierten Faktoren eine nicht-zufällige Streuung, die der *Wechselwirkung* der Faktoren geschuldet ist. Die Gesamtstreuung ergibt sich aus der Summe von zufälliger Streuung und allen erklärten Streuungen. Hat nun die durch einen bestimmten Faktor oder eine bestimmte Kombination aus Faktoren erklärte Streuung einen signifikanten Anteil an der Gesamtstreuung, dann hat folglich dieser Faktor bzw. diese Kombination von Faktoren einen *signifikanten Einfluss*. Die ANOVA ist ein mathematisches Werkzeug zum Aufspüren dieser signifikanten Einflüsse.
Der durch die ANOVA ermittelte p-Wert eines bestimmten Faktors oder einer bestimmten Faktorkombination entspricht der Wahrscheinlichkeit, dass der *empirisch ermittelte* Anteil der durch diesen Faktor bzw. dieser Faktorkombination erklärten Streuung an der Gesamtstreuung *in Wirklichkeit durch reinen Zufall* zustande gekommen ist. Liegt der p-Wert unterhalb eines bestimmten *Signifikanzniveaus* α, dann hat der untersuchte Faktor bzw. die Faktorkombination einen α-signifikanten Einfluss auf die erhobenen Daten.

1.1 Empirische Streuung und Varianz

Die *empirische Streuung* eines Datensatzes $x = x_1, \ldots, x_N$ um eine Funktion $f = f_1, \ldots, f_N$ ist ein Maß für die Abweichung des Datensatzes von seinen

durch f vorhergesagten Werten. Sie ist definiert durch die Quadratsumme der Differenzen zwischen Datenpunkten und Funktionswerten,

$$S_f^2(x) := \sum_{n=1}^{N} (x_n - f_n)^2. \tag{1}$$

Eine besonders häufige Vorhersage ist der *empirische Mittelwert*

$$\overline{x} := \frac{1}{N} \sum_{n=1}^{N} x_n. \tag{2}$$

Die Streuung des Datensatzes um seinen Mittelwert $\overline{x}$ ist also gegeben durch

$$S^2(x) := \sum_{n=1}^{N} (x_n - \overline{x})^2. \tag{3}$$

Die *empirische Varianz* s_f^2 eines Datensatzes x bezüglich einer Funktion f entspricht der empirischen Streuung pro Freiheitsgrad. Sei D_S die Anzahl der Freiheitsgrade der Streuung S, dann ist die empirische Varianz definiert durch

$$s_f^2(x) := \frac{S_f^2(x)}{D_S}. \tag{4}$$

Jeder Datenpunkt liefert einen Freiheitsgrad der Streuung um f. Für den Fall $f = \overline{x}$ wird der Mittelwert aus den Datenpunkten berechnet, also hat S einen Freiheitsgrad weniger und damit insgesamt $D_S = N - 1$ Freiheitsgrade[1]. Somit bekommt die empirische Varianz ihre vertraute Form

$$s^2(x) = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \overline{x})^2, \tag{5}$$

von welcher sich wiederum die *empirische Standardabweichung*

$$s(x) = \sqrt{S^2(x)} \tag{6}$$

ableitet. Das Wort "empirisch" bezeichnet die Abhängigkeit von empirisch erhobenen Datensätzen. Skalare Funktionen von Datensätzen heissen *Statistiken*. Die empirische Streuung und Varianz sind also Beispiele für Statistiken. In Abgrenzung zu Statistiken gibt es Funktionen von *Zufallsvariablen*, welche theoretische Konstrukte sind, die durch *Wahrscheinlichkeitsverteilungen* definiert sind. Man konzipiert die empirisch erhobenen Datensätze als Realisationen von zugrunde liegenden theoretischen Zufallsvariablen.

[1] Zur Veranschaulichung stelle man sich den Extremfall eines einzelnen Datenpunktes vor. Dann ist $S(x)$ immer Null, egal welchen Wert der Datenpunkt hat. Somit hat für $N = 1$ die Streuung offenbar $N - 1 = 0$ Freiheitsgrade.

1.2 Einfaktorielle ANOVA

Die einfaktorielle ANOVA basiert auf der Annahme, dass die Streuung der Daten zum Teil durch den Einfluss eines einzelnen Faktors a erklärt werden kann.
Faktoren haben der Einfachheit halber nur ganzzahlige Werte $a = 1, \ldots, A$, wobei A die Anzahl der *Faktorstufen* bzw. *Faktorausprägungen* ist. Nehmen wir an, zu jedem Wert von a wurden $N(a)$ unabhängige Messungen (Messwiederholungen) gemacht, dann liegen insgesamt

$$N = \sum_{n=1}^{A} N(a) \tag{7}$$

Datenpunkte vor, die sich in der Form $x = \{x_n(a)\}$ darstellen lassen. Häufig liegen zu jedem a gleich viele Messwiederholungen $N(a) = R$ vor, so dass dann $N = A \cdot R$.
Zuerst berechnen wir die zu den einzelnen Faktorwerten a gehörigen Mittelwerte,

$$\overline{x}(a) = \frac{1}{N(a)} \sum_{n=1}^{N(a)} x_n(a), \tag{8}$$

und daraus den Gesamtmittelwert,

$$\overline{x} = \frac{1}{A} \sum_{a=1}^{A} \overline{x}(a). \tag{9}$$

Aus beiden berechnen wir die durch den Faktor a erklärte Streuung,

$$S_a^2 = \sum_{a=1}^{A} N(a)(\overline{x}(a) - \overline{x})^2. \tag{10}$$

sowie die zufällige *Reststreuung*,

$$S_r^2 = \sum_{a=1}^{A} \sum_{n=1}^{N(a)} (x_n(a) - \overline{x}(a))^2, \tag{11}$$

so dass wir die Gesamtstreuung S_t^2 durch Addition der letzten beiden Ausdrücke erhalten,

$$S_t^2 = S_a^2 + S_r^2, \tag{12}$$

was äquivalent zu

$$S_t^2 = \sum_{a=1}^{A} \sum_{n=1}^{N(a)} (x_n(a) - \overline{x})^2 \tag{13}$$

ist. Als nächstes überlegen wir uns die Anzahl der Freiheitsgrade. Die durch a erklärte Streuung S_a^2 hat $D_a = A - 1$ Freiheitsgrade. Die Reststreuung S_r^2 hat

$$D_r = \sum_{a=1}^{A} \big(N(a) - 1\big) \tag{14}$$

und die Gesamtstreuung

$$D_t = \left(\sum_{a=1}^{A} N(a)\right) - 1 \tag{15}$$

Freiheitsgrade. Daraus ergeben sich die empirischen Varianzen

$$s_a^2 = \frac{S_a^2}{D_a}, \quad s_r^2 = \frac{S_r^2}{D_r}, \quad s_t^2 = \frac{S_t^2}{D_t}. \tag{16}$$

Den zum Faktor a gehörigen empirischen F-Wert bestimmen wir durch

$$F_a = \frac{\text{erklärte Varianz}}{\text{Restvarianz}} = \frac{s_a^2}{s_r^2}. \tag{17}$$

Schliesslich berechnen wir den p-Wert zu

$$p_a = 1 - P_F(F; D_a, D_r), \tag{18}$$

wobei P_F die *Fisher-Verteilung* für gegebene Freiheitsgrade D_a und D_r ist. Liegt der p-Wert unterhalb dem vorgegebenen Signifikanz-Niveau α, dann hat der Faktor a einen auf diesem Niveau signifikanten Einfluss auf die Daten.
Falls MATLAB zur Verfügung steht, findet man die Fisher-Verteilung über den Befehl `fcdf`. Falls MATLAB (oder ähnliche Programme, die die Fisher-Verteilung ausrechnen können) nicht zur Verfügung steht, dann muss man sich eine Tabelle mit F-Werten besorgen. Da die F-Verteilung zwei Parameter hat (die beiden Freiheitsgrade), handelt es sich zumeist um recht große, unhandliche Tabellenwerke. Hier schlägt man bei dem gewünschten α und den berechneteten Freiheitsgraden D_a und D_r den *theoretischen F-Wert* nach. Liegt dieser theoretische F-Wert *unter* dem soeben berechneten empirischen F-Wert, dann kann (und muss) man auf Signifikanz schliessen.

1.3 Falls keine Rohdaten vorliegen...

...müssen zumindest folgende Informationen zur Verfügung stehen: Die Liste der Mittelwerte $\overline{x}(a)$, der Standardabweichungen $s(a)$, sowie der Anzahl der Messwiederholungen $N(a)$, und zwar für jede Faktorstufe $a = 1, \ldots, A$. Dann kann man die Berechnung (8) einfach überspringen. Da ferner die zu jeder Faktorstufe a gehörige empirische Varianz definiert ist durch

$$s^2(a) = \frac{1}{N(a) - 1} \sum_{n=1}^{N(a)} (x_n(a) - \overline{x}(a))^2, \tag{19}$$

lässt sich Gleichung (11) ersetzen durch den Ausdruck

$$S_r^2 = \sum_{a=1}^{A} \big(N(a) - 1\big) s^2(a). \tag{20}$$

Alles in allem lautet der Ausdruck für den empirischen F-Wert dann

$$F = \frac{D_r}{D_a} \cdot \frac{\sum_{a=1}^{A} N(a)(\overline{x}(a) - \overline{x})^2}{\sum_{a=1}^{A} (N(a) - 1) s^2(a)}. \tag{21}$$

Es lässt sich also die einfaktorielle ANOVA auch ohne vorliegende Rohdaten durchführen; alles, was man braucht sind die Mittelwerte, Standardabweichungen und Anzahl der Messwiederholungen.

1.4 Zusammenhang mit dem Bestimmtheitsmaß

Schauen wir uns nochmal die Gleichung (12) an,

$$S_t^2 = S_a^2 + S_r^2, \tag{22}$$

in Worten:

$$\text{Gesamtstreuung} = \text{erklärte Streuung} + \text{Reststreuung}. \tag{23}$$

Durch Umstellen erhält man

$$\frac{S_a^2}{S_t^2} = 1 - \frac{S_r^2}{S_t^2} \tag{24}$$

und dies entspricht genau der Definition des Bestimmtheitsmaßes,

$$R^2 = 1 - \frac{S_r^2}{S_t^2}, \tag{25}$$

oder in Worten

$$\frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}} = 1 - \frac{\text{Reststreuung}}{\text{Gesamtstreuung}}. \tag{26}$$

Das Bestimmtheitsmaß R^2 entspricht also dem *Anteil von erklärter Streuung an der Gesamtstreuung*,

$$R^2 = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}}. \tag{27}$$

Wir können also neben dem Signifikanztest, der rein qualitativ sagt, ob der Einfluß eines Faktors höchstwahrscheinlich vorliegt oder nicht, noch eine quantitative Aussage treffen, *wie stark* der Einfluss des Faktors ist. Die der ANOVA zugrunde liegende Nullhypothese besagt, dass der Einfluss exakt Null ist. Der

berechnete p-Wert gibt an, wie wahrscheinlich diese Nullhypothese angesichts der empirisch gewonnenen Daten ist. Das Bestimmtheitsmaß R^2 gibt an, *wie groß* der Einfluss des Faktors auf die gewonnenen Daten *scheinbar* ist.
Unter Verwendung der Definition

$$F = \frac{D_r}{D_a} \cdot \frac{S_a^2}{S_r^2} \tag{28}$$

und Gleichung (25) kann man folgende Relationen herleiten:

$$F = \frac{D_r}{D_a} \cdot \frac{R^2}{1-R^2}, \quad R^2 = \frac{1}{1 + \frac{D_r}{D_a} \cdot \frac{1}{F}} \tag{29}$$

Man kann also Bestimmtheitsmaß und F-Wert ineinander umrechnen. Dabei gilt: Je größer F, umso näher liegt R^2 an 1 und umgekehrt,

$$R^2 \to 1 \quad \Leftrightarrow \quad F \to \infty. \tag{30}$$

1.5 Zusammenhang mit Modellanpassung

Die Größe R^2 ist bekannt insbesondere im Zusammenhang mit Modellanpassungen. Hier gibt sie an, wie gut ein bestimmtes Modell zu den Daten passt. Während der Modellanpassung werden die Parameter so gewählt, dass R^2 möglichst nahe 1 liegt (*model fit*). Anschliessend wird verglichen, wie gut das so angepasste Modell zu neuen, bei der Anpassung nicht verwendeten Daten passt (*model prediction*). Hier gibt R^2 dann an, wie gut die Vorhersage des Modell ist.
Sei also $x_n(t)$ ein gemessener zeitabhängiger Datensatz bestehend aus $n = 1, \dots N$ Zeitreihen, und $f(t)$ die vom Modell gelieferte Funktion. Dann ist R^2 definiert durch

$$R_f^2 = 1 - \frac{\sum_{t=1}^{T} \sum_{n=1}^{N} (x_n(t) - f(t))^2}{\sum_{t=1}^{T} \sum_{n=1}^{N} (x_n(t) - \overline{x})^2}, \tag{31}$$

wobei

$$\overline{x} = \sum_{t=1}^{T} \sum_{n=1}^{N} x_n(t) \tag{32}$$

und wobei wir der Einfachheit halber diskrete Zeitpunkte $t = 1, \dots, T$ angenommen haben. Vergleichen wir dies mit dem Bestimmtheitsmaß, das wir oben im Zusammenhang mit der einfaktoriellen ANOVA gewonnen haben,

$$R_a^2 = 1 - \frac{\sum_{a=1}^{A} \sum_{n=1}^{N} (x_n(a) - \overline{x}(a))^2}{\sum_{a=1}^{A} \sum_{n=1}^{N} (x_n(a) - \overline{x})^2}, \tag{33}$$

dann bemerken wir, dass unter folgenden Transformationen die Definitionen ineinander übergehen, wenn wir den Faktor a mit der Zeit t und den Mittelwert $\overline{x}(a)$ mit der Modellfunktion $f(t)$ ersetzen. Man kann also sagen, dass

in der ANOVA ein bestimmtes Modell überprüft wird, welches versucht, die empirischen Daten aus einem von einem Faktor a abhängigen Mittelwert vorherzusagen. Umgekehrt kann man sagen, dass in der Modellanpassung die Zeit als Faktor angenommen wird und davon ausgehend eine zeitabhängige Funktion zur Anpassung an die Daten gesucht wird. Liegen N Zeitreihen vor, so ist sicherlich der Mittelwert

$$\overline{x}(t) = \frac{1}{N} \sum_{n=1}^{N} x_n(t) \tag{34}$$

eine gute Anpassungsfunktion, die das entsprechende R^2 minimiert. Sie hängt allerdings von einem maximalen Satz von $N \cdot T$ Parametern ab und ist daher als Modellfunktion völlig unbrauchbar. Je mehr man vom zugrunde liegenden System versteht, umso weniger Parameter sollte die Anpassungsfunktion benötigen.