ANOVA & Co

Kim Boström

23. September 2010

1 Konzepte und Bezeichnungen

Im folgenden werden Zufallsvariablen mit einen Hut gekennzeichnet, wie z.B. $\hat{x}$, und Datensätze (Stichproben, Samples) mit einer Tilde, wie z.B. $\tilde{x}$. Eine reelle Zufallsvariable $\hat{x}$ ist eindeutig festgelegt durch eine zugehörige *Wahrscheinlichkeitsdichte* $\rho(x)$ auf $\mathbb{R}$, so dass Wahrscheinlichkeiten von Ereignissen $X \subset \mathbb{R}$ durch das *Wahrscheinlichkeitsmaß*

$$P(X) := \int_X dx\, \rho(x) \tag{1}$$

bestimmt werden. Ereignisse werden durch Teilmengen von $\mathbb{R}$ repräsentiert. So wird z.B. das Ereignis "$\hat{x}$ ist kleiner gleich a" repäsentiert durch die Menge $X = \{x \in \mathbb{R} \mid x \leq a\}$, so dass in vereinfachender Schreibweise

$$P\{\hat{x} \leq a\} = \int_{-\infty}^{a} dx\, \rho(x) = p(a), \tag{2}$$

wobei p die Stammfunktion von ρ, die *kumulative Wahrscheinlichkeitsverteilung* ist. *Erwartungswerte* (Mittelwerte) von Funktionen f von $\hat{x}$ lassen sich berechnen durch

$$\langle f(\hat{x}) \rangle := \int dx\, \rho(x) f(x). \tag{3}$$

Die *Streuung* (Standardabweichung, Unsicherheit) von $\hat{x}$ ist definiert durch

$$\Delta\hat{x} := \sqrt{\langle \hat{x}^2 \rangle - \langle \hat{x} \rangle^2}. \tag{4}$$

Eine Zufallsvariable kann beliebig oft *realisiert* werden, etwa indem eine Messung mit unbestimmten Ausgang gemacht wird oder ein nicht kontrollierter Parameter in eine Messung einfließt. Endliche geordnete Mengen von R Realisationen einer Zufallvariablen $\hat{x}$ stellen einen unabhängigen *Datensatz* (Stichprobe, Sample) der Größe R dar,

$$\tilde{x} = (x_1, \ldots, x_R). \tag{5}$$

Wichtig ist, dass es sich bei den Elementen des Datensatzes $\tilde{x}$ um *unabhängige* Realisationen ein und derselben Zufallsvariablen $\hat{x}$ handelt, d.h. die Wahrscheinlichkeitsverteilung ist jedesmal dieselbe. Handelt es sich hingegen um *abhängige*

Werte, bei denen die Wahrscheinlichkeitsverteilung also *nicht* stets dieselbe ist, so hat man es stattdessen mit jeweils *einer* Realisation von *mehreren verschiedenen* Zufallsvariablen $\hat{x}(1), \ldots, \hat{x}(R)$ zu tun. Äquivalent dazu spricht man auch von *einer abhängigen* Zufallsvariablen $\hat{x}(i)$, wobei i den *Faktor* darstellt, von dem die Variable abhängt, und der R verschiedene *Ausprägungen* hat. Ein zugehöriger Datensatz hat dann die Form $\tilde{x}(i)$ und enthält für jede Ausprägung i mindestens eine Realisation. Es kann sein, dass pro Ausprägung mehrere Realisationen vorliegen, die dann wiederum als untereinander unabhängig verstanden werden. Wir wollen im folgenden mit R stets als die Anzahl unabhängiger Realisationen bezeichnen.
Reellwertige Funktionen eines Datensatzes $\tilde{x} = (x_1, \ldots, x_R)$ werden als *Statistiken* bezeichnet. Beliebte Statistiken sind der *Mittelwert*

$$\mu(\tilde{x}) = \frac{1}{R} \sum_{i=1}^{R} x_i \tag{6}$$

und die *Standardabweichung*

$$\sigma(\tilde{x}) = \sqrt{\frac{1}{R-1} \sum_{i=1}^{R} (x_i - \mu(\tilde{x}))^2} \tag{7}$$

eines Datensatzes $\tilde{x}$. Beide sind erwartungstreue unverzerrte Schätzer für den Erwartungswert bzw. für die Streuung der zugehörigen Zufallsvariablen $\hat{x}$.

2 t-Test

Der *Student'sche t-Test*, nach William Gosset aka "Student", untersucht, ob zwei Datensätze denselben Mittelwert haben. Es wird unterschieden zwischen einem *gepaarten (abhängigen)* und einem *ungepaarten (unabhängigen)* t-Test.

2.1 Ungepaarter t-Test

Der ungepaarte t-Test geht davon aus, dass zwei Datensätze den unabhängigen Realisationen zweier Zufallsvariablen entsprechen. Es handelt sich also um Datensätze $\tilde{x}$ und $\tilde{y}$, die von jeweils einer Zufallsvariable $\hat{x}$ bzw. $\hat{y}$ stammen. Die Nullhypothese ist

$$H: \quad \langle \hat{x} \rangle = \langle \hat{y} \rangle. \tag{8}$$

Dies lässt sich auch äquivalent durch *eine abhängige* Zufallsvariable $\hat{x}(i)$ ausdrücken, die von einem Faktor i mit zwei Ausprägungen kategorisiert wird, so dass die Nullhypothese nun die *Unabhängigkeit* des Erwartungswerts von diesem Faktor i behauptet,

$$H: \quad \langle \hat{x}(i) \rangle = \mu, \tag{9}$$

für $i = 1, 2$ und einen gemeinsamen Erwartungswert μ.

Hat die Zufallsgröße $\hat{x}(i)$ für jedes i die gleiche Anzahl von k Realisationen, dann hat der Datensatz $\tilde{x}$ die Form

$$\tilde{x} = x_r(i), \tag{10}$$

wobei $r = 1, \ldots, R$ die Realisationen indiziert.
In MATLAB muss der Datensatz $x_r(i)$ für den ungepaarten t-Test in zwei Datensätze `A` und `B` aufgeteilt werden, so dass

$$\mathtt{A} = \begin{bmatrix} x_1(1) \\ \vdots \\ x_R(1) \end{bmatrix}, \quad \mathtt{B} = \begin{bmatrix} x_1(2) \\ \vdots \\ x_R(2) \end{bmatrix}. \tag{11}$$

Der entsprechende MATLAB-Befehl für den Aufruf des ungepaarten t-Tests lautet dann

```
>> p = ttest2(A,B);
```

Der t-Test gibt dann für diese Datensätze die Wahrscheinlichkeit p aus, dass die vorliegenden Daten unter zutreffender Nullhypothese entstanden sind. In diesem Fall also ist p die Wahrscheinlichkeit dafür, dass die Datensätze $\tilde{x}(1)$ und $\tilde{x}(2)$ von Zufallvariablen $\hat{x}(1)$ bzw. $\hat{x}(2)$ stammen, die denselben Mittelwert haben. Mit wieder anderen Worten ist p die Wahrscheinlichkeit für einen *Fehler 1. Art* oder auch α-*Fehler*, der darin besteht, dass die Nullhypothese zurückgewiesen wird, obwohl sie zutrifft. Ist $p < \alpha$ für ein festgelegtes Signifikanzniveau α, dann kann die Nullhypothese auf diesem Signifikanzniveau zurückgewiesen werden, d.h. die Datensätze $\tilde{x}(1)$ und $\tilde{x}(2)$ haben einen α-signifikant verschiedenen Mittelwert. Zumeist setzt man $\alpha = 0.05$.

Hinweise Der ungepaarte t-Test setzt voraus:

i) Normalverteilung

ii) gleiche Varianz

Ist eine der Voraussetzungen verletzt, was durch unabhängige Tests entschieden werden muss, dann sollte ein nicht-parametrischer Test, etwa der *Wilcoxon-Rangsummentest* verwendet werden, der jedoch eine geringere Entscheidungsstärke besitzt.

2.2 Gepaarter t-Test

Der gepaarte t-Test geht davon aus, dass die beiden Datensätze *abhängigen* Realisationen einer abhängigen Zufallsvariable entsprechen. Es handelt sich also um Datensätze $\tilde{x}(1) = (x_1(1), \ldots, x_R(1))$ und $\tilde{x}(2) = (x_1(2), \ldots, x_R(2))$, die

den Realisationen einer abhängigen Zufallsvariable $\hat{x}(i,r)$ entsprechen, wobei $i = 1, 2$ und $r = 1, \ldots, R$. Die Nullhypothese ist

$$H: \quad \langle \hat{x}(i,r) \rangle = \mu. \tag{12}$$

Die Zufallsgröße $\hat{x}(i,r)$ liefert nun für jedes i, r genau *eine* Realisation, so dass der Datensatz $\tilde{x}$ die Form

$$\tilde{x} = x(i,r) \tag{13}$$

hat, wobei $i = 1, 2$ und $r = 1, \ldots, R$.
In MATLAB muss der Datensatz $x(i,r)$ für den gepaarten t-Test in zwei Datensätze `A` und `B` aufgeteilt werden, so dass

$$\texttt{A} = \begin{bmatrix} x(1,1) \\ \vdots \\ x(1,R) \end{bmatrix}, \quad \texttt{B} = \begin{bmatrix} x(2,1) \\ \vdots \\ x(2,R) \end{bmatrix}. \tag{14}$$

Der entsprechende MATLAB-Befehl für den Aufruf des gepaarten t-Tests lautet dann

```
>> p = ttest(A,B);
```

Äquivalent dazu ist der Aufruf

```
>> p = ttest(A-B);
```

Dies liegt daran, dass der gepaarte t-Test einem statistischen Test der *Differenzen* der miteinander gepaarten Einträge beider Datensätze hinsichtlich der Nullhypothese entspricht, dass diese Differenz um Null herum verteilt ist.
Der gepaarte t-Test gibt die Wahrscheinlichkeit p für den α-Fehler aus, d.h. für die irrtümliche Zurückweisung der Nullhypothese.

Hinweise Der gepaarte t-Test setzt voraus:

i) Normalverteilung

ii) gleiche Varianz

Ist eine der Voraussetzungen verletzt, was durch unabhängige Tests entschieden werden muss, dann sollte ein nicht-parametrischer Test, etwa der *Wilcoxon-Vorzeichen-Rang-Test* verwendet werden, der jedoch eine geringere Entscheidungsstärke besitzt.

3 ANOVA

Die *Analysis of Variance* (ANOVA) untersucht, ob ein gegebener Datensatz von einem oder mehreren Faktoren statistisch signifikant abhängt. Für mehr als zwei Faktoren steigt der Datenbedarf für eine sichere Schätzung der unterschiedlichen Abhängigkeiten stark an. Zumeist werden daher nur ein- oder zweifaktorielle ANOVAs durchgeführt.

3.1 Einfaktorielle ANOVA

Die *einfaktorielle* (im Englischen *one-way*) ANOVA geht davon aus, dass der Datensatz von genau einem Faktor abhängt. Die zugrunde liegende Zufallsgröße $\hat{x}$ hat also die Form

$$\hat{x} = \hat{x}(i), \tag{15}$$

wobei $i = 1, \ldots, I$ die *Ausprägungen* des Faktors indiziert. Die Nullhypothese H lautet

$$H: \quad \langle \hat{x}(i) \rangle = \mu \tag{16}$$

Mit anderen Worten: Gemäß H hängt der Mittelwert *nicht* vom Faktor i ab. Hat die Zufallsgröße $\hat{x}(i)$ für jedes i die gleiche Anzahl von R Realisationen, dann hat der Datensatz $\tilde{x}$ die Form

$$\tilde{x} = x_r(i), \tag{17}$$

wobei $r = 1, \ldots, R$ die Realisationen indiziert.
In MATLAB wird der Datensatz $x_r(i)$ durch eine Matrix `X` repräsentiert, bei welcher r die Zeilen und i die Spalten indiziert:

$$\texttt{X} = \begin{bmatrix} x_1(1) & \cdots & x_1(I) \\ \vdots & \ldots & \vdots \\ x_R(1) & \cdots & x_R(I) \end{bmatrix} \tag{18}$$

Der entsprechende MATLAB-Befehl für den Aufruf der einfaktoriellen ANOVA lautet dann

```
>> p = anova1(X);
```

Die ANOVA gibt dann für diesen Datensatz die Wahrscheinlichkeit p des α-Fehlers für die irrtümliche Zurückweisung der Nullhypothese H aus. Ist $p < \alpha$ für ein festgelegtes Signifikanzniveau α, dann kann die Nullhypothese auf diesem Signifikanzniveau zurückgewiesen werden, d.h. der Datensatz $\tilde{x}$ ist α-signifikant abhängig vom Faktor i.

Hinweise

1. Die ANOVA setzt voraus:

   i) Normalverteilung
   ii) gleiche Varianz

   Ist dies nicht in ausreichendem Maße der Fall (was durch separate Tests entschieden wird), so sollte man nicht-parametrische Verfahren anwenden, die allerdings eine geringere Teststärke besitzen.

2. Für $I = 2$ ist die einfaktorielle ANOVA identisch mit einem *ungepaarten t-Test*,

$$I = 2: \quad \text{Anova1}[x_r(i)] = \text{t-Test}[x_r(1), x_r(2)] \tag{19}$$

Beispiel Es werden $I \cdot S$ Versuchspersonen in I gleich große Gruppen zu je S Personen unterteilt. Die Gruppen unterscheiden sich hinsichtlich eines bestimmten Merkmals voneinander (z.B. Alter). Dieses Merkmal ist der *unabhängige Parameter* und stellt den *Faktor* dar, bezüglich welchem die ANOVA ausgewertet wird. Jede Versuchsperson wird M mal bezüglich einer bestimmten Größe vermessen (z.B. Reaktionszeit). Diese Größe ist der *abhängige Parameter*. Es handelt sich also für jedes i um $R = S \cdot M$ Realisierungen der Zufallvariablen $\hat{x}(i)$, wobei i der unabhängige Parameter (Faktor) ist. Die Versuchspersonen und die an ihnen durchgeführten Messungen innerhalb jeder Gruppe werden also "in einen Topf geschmissen" und auf Abhängigkeit vom Faktor i untersucht.

3.2 Zweifaktorielle ANOVA

Die *zweifaktorielle* (im Englischen *two-way*) ANOVA geht davon aus, dass der Datensatz von genau zwei Faktoren abhängt. Die zugrundeliegende Zufallsgröße $\hat{x}$ ist also gegeben durch

$$\hat{x} = \hat{x}(i, j), \tag{20}$$

mit $i = 1, \ldots, I$ und $j = 1, \ldots, J$, wobei I und J die jeweilige Anzahl der Ausprägungen der beiden Faktoren ist. Es gibt zwei Nullhypothesen H_A und H_B, die jeweils den ersten bzw. den zweiten Faktor betreffen:

$$H_A: \quad \langle \hat{x}(i, j) \rangle = \mu_1(j) \tag{21}$$

$$H_B: \quad \langle \hat{x}(i, j) \rangle = \mu_2(i) \tag{22}$$

Mit anderen Worten: Gemäß H_A hängt der Mittelwert nicht vom ersten Faktor i ab; gemäß H_B hängt der Mittelwert nicht vom zweiten Faktor j ab. Treffen beide Nullhypothesen zu, dann ist der Mittelwert unabhängig von *beiden* Faktoren, also konstant:

$$H_A \wedge H_B: \quad \langle \hat{x}(i, j) \rangle = \mu, \tag{23}$$

denn es ist ja dann $\mu_1(j) = \mu_2(i)$, was nur durch eine konstante Funktion erfüllt wird, so dass also $\mu_1(j) = \mu_2(i) = \mu$. Wird eine der beiden Nullhypothesen zurückgewisen, dann bedeutet dies, dass der Mittelwert von dem jeweiligen Faktor signifikant abhängt. Werden beide Nullhypothesen zurückgewiesen, dann hängt der Mittelwert signifikant von beiden Faktoren ab. In diesem Fall kann es eine *Interaktion* zwischen den Faktoren geben. Dies wird durch eine dritte Nullhypothese H_{AB} zum Ausdruck gebracht,

$$H_{AB}: \quad \langle \hat{x}(i, j) \rangle = \mu_1(j) + \mu_2(i). \tag{24}$$

Mit anderen Worten: Gemäß H_{AB} hängt der Mittelwert *rein additiv* von beiden Faktoren ab. Wird H_{AB} verworfen, dann bedeutet dies, dass der Mittelwert *nicht rein additiv* von beiden Faktoren abhängt, d.h. dass es eine Interaktion zwischen beiden Faktoren gibt.

H_{AB} ist nicht unabhängig von H_A und H_B. Nur wenn H_A und H_B beide zurückgewiesen werden, kann H_{AB} ebenfalls verworfen werden, weil eine Interaktion die Abhängigkeit von *beiden* Faktoren impliziert, d.h.

$$\neg H_{AB} \quad \rightarrow \quad \neg H_A \wedge \neg H_B. \tag{25}$$

Dies ist logisch äquivalent zu

$$H_A \vee H_B \quad \rightarrow \quad H_{AB}, \tag{26}$$

was sich folgendermaßen lesen lässt: Wenn einer von beiden Faktoren keine Rolle spielt, kann es keine Interaktion geben. Es ist möglich, dass H_A und H_B beide verworfen werden, H_{AB} hingegen nicht. Dies bedeutet dann, dass es eine rein additive Abhängigkeit von beiden Faktoren gibt und damit also keine Interaktion. Hat die Zufallsgröße $\hat{x}(i,j)$ für jedes i,j die gleiche Anzahl von R Realisationen, dann hat der Datensatz die Form

$$\tilde{x} = x_r(i,j) \tag{27}$$

In MATLAB kann man dies in eine Matrix `X` schreiben, wobei man beachten muss, dass der erste Faktor i in den *Spalten* steht! Der zweite Faktor j steht in den Zeilen, allerdings so, dass für jede Ausprägung von j jeweils R Datenzeilen für die einzelnen Realisationen untereinander stehen. Es ergibt sich also:

$$\mathtt{X} = \underbrace{\begin{bmatrix} x_1(1,1) & \cdots & x_1(I,1) \\ \vdots & \ldots & \vdots \\ x_R(1,1) & \cdots & x_R(I,1) \\ \vdots & \ldots & \vdots \\ x_1(1,J) & \cdots & x_1(I,J) \\ \vdots & \ldots & \vdots \\ x_R(1,J) & \cdots & x_R(I,J) \end{bmatrix}}_{i=1 \qquad\qquad i=I} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} j=1 \quad (R \text{ Realisationen}) \\ \vdots \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} j=J \quad (R \text{ Realisationen}) \end{matrix} \tag{28}$$

Der entsprechende MATLAB-Befehl für die Matrix ist

```
for i=1:I
   for j=1:J
      for r=1:R
         X(r+(j-1)*R,i) = data(i,j,r);
      end
   end
end
```

wobei `data(i,j,r)` die r-te Realisation der Zufallsvariablen, also z.B. der r-te Messwert eines bestimmten Merkmals, für die Kombination (i,j) der beiden Faktoren ist. Der MATLAB-Befehl für die zweifaktorielle ANOVA lautet dann

```
>> [p_A, p_B, p_AB] = anova2(X,R);
```

Die ANOVA gibt dann für diesen Datensatz die Wahrscheinlichkeiten p_A, p_B, p_{AB} der entsprechenden Nullhypothesen H_A, H_B, H_{AB} aus.

Hinweise

1. Für $J = 1$ ist der p-Wert p_A des ersten Faktors identisch mit dem p-Wert einer einfaktoriellen ANOVA. Das liegt daran, dass der zweite Faktor j nur eine einzige Ausprägung j_0 hat, effektiv also nur noch der erste Faktor i übrig bleibt. Die zugrunde liegende Zufallsvariable hat die Form $\hat{x}'(i) := \hat{x}(i, j_0)$ und die Realisierungen $\tilde{x}'_r(i) := x_r(i, j_0)$ mit $i = 1, \ldots, I$ und $r = 1, \ldots, R$, und es gilt

$$\mathrm{Anova2}[x_r(i, j_0)] = \mathrm{Anova1}[x'_r(i)] \tag{29}$$

2. Für $I = 2$ und $R = 1$ ist der p-Wert p_A des ersten Faktors i identisch mit dem p-Wert eines *gepaarten t-Tests*, welcher den zweiten Faktor j als Index für die Realisationen einer gepaarten Stichprobe interpretiert. Mit $x'_j(i) := x_1(i, j)$ gilt also

$$I = 2: \quad \mathrm{Anova2}[x_1(i, j)] = \text{t-Test}[x'_j(1) - x'_j(2)] \tag{30}$$

   Dies liegt daran, dass ein gepaarter t-Test davon ausgeht, dass es sich bei den Realisationen $x'_j(i)$ der Zufallsvariablen $\hat{x}'(i)$ um *Paare* $(x'_j(1), x'_j(2))$ handelt, die jeweils miteinander verglichen werden. Der Index j, der die Realisation abzählt, ist also gewissermaßen ein eigener *Faktor*, nämlich die *Identität der Versuchsperson*.

Beispiele

1. Es werden $J \cdot S$ Versuchspersonen in J gleich große Gruppen zu je S Personen unterteilt. Die Gruppen unterscheiden sich hinsichtlich zweier Merkmale i und j voneinander (z.B. Alter und Geschlecht). Diese Merkmale sind die unabhängigen Parameter und stellen die beiden Faktoren dar, bezüglich welcher die ANOVA ausgewertet wird. Jede Versuchsperson wird M mal bezüglich einer bestimmten Größe, d.h. einem abhängigen Parameter, vermessen (z.B. Reaktionszeit). Es handelt sich also für jede Kombination i, j der beiden Faktoren um $R = S \cdot M$ Realisierungen der Zufallvariablen $\hat{x}(i, j)$, der Datensatz hat also die Form $x_r(i, j)$ mit $r = 1, \ldots, S \cdot M$. Die Versuchspersonen und die an ihnen durchgeführten Messungen innerhalb jeder Gruppe werden also "in einen Topf geschmissen" und auf Abhängigkeit von den beiden Faktoren i und j untersucht.

2. Wie oben, nur dass diesmal die Versuchspersonen und die Messungen an ihnen *nicht* in einen Topf geschmissen werden und nur *ein* Faktor i untersucht wird. Die Identität jeder Versuchsperson stellt nun formal einen zusätzlichen Faktor j dar, so dass die zweifaktorielle ANOVA neben der Abhängigkeit vom kontrollierten Parameter i auch die inter-subjektive Abhängigkeit j bestimmt. Die zugrunde liegende Zufallsvariable lautet also $\hat{x}(i, j)$, wobei i den unabhängigen Parameter und j die Versuchsperson indiziert. Werden pro Versuchsperson M unabhängige Messungen durchgeführt, handelt es sich um $R = M$ Realisationen von $\hat{x}(i, j)$ und der Datensatz hat daher die Form $x_r(i, j)$ mit $r = 1, \ldots, M$ und $j = 1, \ldots, S$. Zumeist macht man in dieser Konstellation aber nur genau eine Messung pro Versuchsperson j und Versuchsbedingung i: Man misst die Versuchsperson z.B. einmal mit und einmal ohne Medikation. Oder man misst die Reaktionszeit der Versuchsperson unter Einfluss der Drogen a,b,c. Wenn man für solche *abhängigen Stichproben* nur zwei Versuchsbedingungen hat, also $I = 2$, dann ist die zweifaktorielle ANOVA äquivalent zu einem gepaarten t-test. Man könnte daher im Falle beliebiger Anzahl I von Versuchsbedingungen allgemein von einer "gepaarten ANOVA" sprechen (auch wenn das Wort "Paar" hier nichts mit einer Zwei-heit zu tun hätte), um damit zum Ausdruck zu bringen, dass es sich um die Analyse *abhängiger Stichproben* handelt.