Varianzanalyse

von Peter Pfaffelhuber

Version: 24. November 2015

Bei der (ein-faktoriellen) Varianzanalyse will man Unterschiede zwischen metrischen Merkmalen X verschiedener Gruppen herausfinden. (Man stelle sich etwa die Wirkung p verschiedener Behandlungsmethoden auf ein biometrisches Merkmal bei einer Krankheit vor.) Man betrachtet also p Populationen und Stichprobengrößen $n_1, ..., n_p$. Die zu messende metrische Größe X wird auch *Faktor* genannt, die einzelnen Gruppen als *Levels* oder *Faktorstufen*.

⟨ex:ins1⟩ **Beispiel 1 (Insektensprays).** Wir verwenden den in R verfügbaren Datensatz `InsectSprays` mittels

```
> attach(InsectSprays)
> a<-data(InsectSprays)
```

Der Datensatz enthält eine Untersuchung von sechs verschiedenen Insektensprays und deren Auswirkungen auf die gefundene Zahl der Insekten auf einem damit behandelten Gebiet. Mit den obigen Befehlen stehen nun die Variablen `spray` und `count` zur Verfügung.

```
> spray
 [1] A A A A A A A A A A A A B B B B B B B B B B B C C C C C C C C C C C C D D
[39] D D D D D D D D D D E E E E E E E E E E E E F F F F F F F F F F F F
Levels: A B C D E F
> count
 [1] 10  7 20 14 14 12 10 23 17 20 14 13 11 17 21 11 16 14 17 17 19 21  7 13  0
[26]  1  7  2  3  1  2  1  3  0  1  4  3  5 12  6  4  3  5  5  5  5  2  4  3  5
[51]  3  5  3  6  1  1  3  2  6  4 11  9 15 22 15 16 13 10 26 26 24 13
```

Die sechs verschiedenen Sprays sind mit `A` bis `F` gekennzeichnet. Da wir uns damit befassen wollen, ob die verschiedenen Sprays gleiche oder unterschiedliche Effekte auf die Insektenzahlen haben, verschaffen wir uns zunächst einen Überblick über die Daten.[1]

```
> tapply(count, spray, mean)
        A         B         C         D         E         F
14.500000 15.333333  2.083333  4.916667  3.500000 16.666667
```

[1] Der Aufruf von `tapply(count, spray, mean)` wendet die Funktion `mean` auf die Zielvariable `count` an, wobei sie die Faktoren `spray` unterscheidet. Die Vektoren `count` und `spray` müssen gleich lang sein. Etwa liefert

```
> tapply(count, spray, length)
A B C D E F
12 12 12 12 12 12
```

da alle Faktoren 12-mal in `spray` vorkommen.

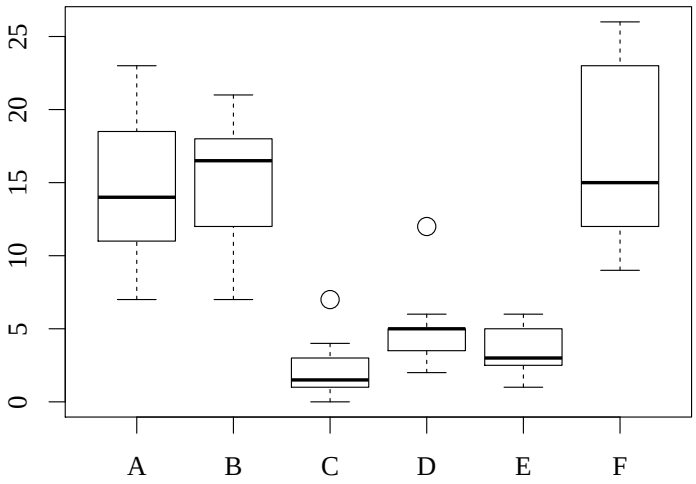

Abbildung 1: Der Box-Plot der Auswirkungen sechs verschiedener Insektensprays.
⟨fig4⟩

Beispielsweise sehen wir so, dass bei Gruppen `E` und `F` die Mittelwerte stark voneinander abweichen.

Eine weitere sinnvolle Methode, sich einen Überblick über die Daten zu verschaffen, ist es, eine Grafik zu erstellen. In unserem Fall bietet es sich an, einen *Box(-Whisker)-Plot* zu verwenden. Dieser wird von

```
> boxplot(count ~ spray)
```

erzeugt; siehe Abbildung 1. Hier wird für alle Levels eine Box angelegt. Der horizontale Strich innerhalb der *Box* stellt den Median dar, die Begrenzungen der *Box* das erste und dritte Quartil (und die *Box* damit den Interquartilbereich). Die *Whiskers*[2] reichen maximal bis zum kleinsten bzw. größten Wert der Daten und sind maximal die anderthalbfache Breite des Interquartilbereiches lang. Daten außerhalb dieses Bereichs werden als einzelne Punkte dargestellt.

Das Modell

Für die Varianzanalyse habe im Modell der Faktor innerhalb der Population k einen Mittelwert von β_k, $k = 1, ..., p$. Weiterhin werden wir wie auch bei der Regression eine gemeinsame Varianz von σ^2 annehmen. Die Modellannahmen lauten also

$$Y_{ki} = \beta_k + \epsilon_{ki}, \qquad k = 1, ..., p, i = 1, ..., n_k,$$

mit $\epsilon \sim N(0, \sigma^2 I)$ und $n = n_1 + \cdots + n_p$. Ziel der Varianzanalyse ist es, die Nullhypothese

$$H_0 := \beta_1 = \cdots = \beta_p$$

gegen $H_1 : \beta_k \neq \beta_\ell$ *für ein Paar* k, ℓ zu testen. Hierfür berechnen wir zunächst die Stichprobenmittel innerhalb der *Faktorstufen* sowie das Gesamtmittel

$$\overline{Y}_{k\bullet} := \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}, \qquad \overline{Y} := \frac{1}{n} \sum_{k=1}^{p} \sum_{i=1}^{n_k} Y_{ki}.$$

[2] Dies bezeichnet auch die Schnurrhaare einer Katze.

Beispiel 2 (Insektensprays, Modellannahmen). Da wir annehmen, dass $Y_{ki} \sim N(\beta_i, \epsilon^2)$, ist

$$\overline{Y}_{k\bullet} \sim N(\beta_k, \sigma^2/n_k),$$

insbesondere sollten $\sqrt{n_k}\overline{Y}_{k\bullet}$ dieselben Varianzen haben. Diese können wir erwartungstreu schätzen, indem wir die empirischen Varianzen innerhalb des k-ten Levels der Stichprobe betrachten.

```
> tapply(count, spray, sd)
       A        B        C        D        E        F
4.719399 4.271115 1.975225 2.503028 1.732051 6.213378
```

Für Level `C` ergibt sich eine kleine Varianz, so dass sich eine Abweichung der Modellannahmen entstehen könnte. Im Moment gehen wir noch nicht darauf ein, die Hypothese der gleichen Varianzen zu testen.

Eine weitere Möglichkeit, die (Un-)Gleichheit der Varianzen zu sehen, ist die grafische Darstellung der Residuen $Y_{ki} - \overline{Y}_{k\bullet}$. Bereits in Abbildung 1 sieht man jedoch, dass kleineres $\overline{Y}_{k\bullet}$ mit einer eher kleineren Streuung einhergeht. □

Zurück zum Test von H_0 gegen H_1. Grundgedanke der Varianzanalyse ist die Varianzzerlegung, also die Zerlegung der Stichprobenvarianz (Sum of sQuares Total)

$$SQT := \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y})^2 \qquad \text{(Sum of sQuares Total)}.$$

Diese Zerlegung ist folgendermaßen gegeben.

Proposition 3 (Varianzzerlegung). *Es gilt*

$$SQT = SQE + SQR$$

für SQT wie oben und

$$SQE := \sum_{k=1}^{p} n_k(\overline{Y}_{k\bullet} - \overline{Y})^2 \qquad \textit{(Sum of sQuares Explained)},$$

$$SQR := \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k\bullet})^2. \qquad \textit{(Sum of sQuares Residual)}$$

Beweis. Wir schreiben

$$\begin{aligned}\sum_{k=1}^{p} n_k(\overline{Y}_{k\bullet} - \overline{Y})^2 + \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k\bullet})^2 &= \sum_{k=1}^{p} n_k(Y_{k\bullet}^2 - \overline{Y}^2) + \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki}^2 - \overline{Y}_{k\bullet}^2) \\ &= \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki}^2 - \overline{Y}^2) = \sum_{k=1}^{p}\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y})^2.\end{aligned}$$

□

Mit diesem Resultat geben wir nun eine Teststatistik an, mit der man H_0 testen kann.

⟨T:testAnova⟩ **Theorem 4 (Varianzanalyse).** *Ist $\beta_1 = \cdots = \beta_p = 0$, so ist*

$$SQT/\sigma^2 \sim \chi^2_{n-1}, \qquad SQE/\sigma^2 \sim \chi^2_{p-1}, \qquad SQR/\sigma^2 \sim \chi^2_{n-p},$$

und

$$\frac{SQE/(p-1)}{SQR/(n-p)} \sim F_{p-1,n-p}.$$

Beweis. OBdA sei $\mu = 0, \sigma^2 = 1$, da der allgemeine Fall durch lineare Transformation in diesen überführt werden kann. Für die erste Aussage ändern wir die Nummerierung der Y's zu $Y_1, ..., Y_n$. Sei $O \in \mathbb{R}^{n\times n}$ eine orthogonale Matrix mit $O_{11} = \cdots = O_{1n} = 1/\sqrt{n}$. Dann ist $Z := OY \sim N(0,1)$ und $Z_1 = (OY)_1 = \sqrt{n}\overline{Y}$ sowie

$$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^n Y_i^2 - Z_1^2 = \sum_{i=2}^n Z_i^2 \sim \chi^2_{n-1}.$$

Die Aussage über SQR ergibt sich analog, da $\sum_{i=1}^{n_k}(Y_{ki} - \overline{Y}_{k\bullet})^2 \sim \chi^2_{n_k-1}$, $k = 1, ..., p$ und diese Zufallsvariablen für verschiedene k unabhängig sind. Bei der Aussage über SQE setzen wir $W_k := \sqrt{n_k}\overline{Y}_{k\bullet}$. Dann ist $W_1, ..., W_p$ unabhängig mit $W_k \sim N(0,1)$ und genau wie oben folgt $SQE \sim \chi^2_{p-1}$. Es bleibt noch, die Unabhängigkeit von SQR und SQE zu zeigen. Hierzu bemerken wir, dass SQE eine Funktion von $(\overline{Y}_{k\bullet} - \overline{Y})_{k=1,...,p}$ ist, und SQR eine Funktion von $(Y_{ki} - \overline{Y}_{k\bullet})_{k=1,...,p,i=1,...,n_k}$. Diese beiden Vektoren sind unabhängig, da

$$\mathbb{COV}[n_k(\overline{Y}_{k\bullet} - \overline{Y}), Y_{\ell i} - \overline{Y}_{\ell\bullet}] = \delta_{k\ell} - \delta_{k\ell} - \frac{n_k}{n} + \frac{n_k}{n} = 0.$$

□

?⟨ex:ins2⟩? **Beispiel 5 (Insektensprays).** Wir führen nun eine Varianzanalyse für das Beispiel 1 durch. Dies funktioniert mit der Funktion `aov` (was für *Analysis Of Variance* steht)

```
> aov.out = aov(count ~ spray, data=InsectSprays)
> summary(aov.out)
```

Dies liefert

```
            Df Sum Sq Mean Sq F value Pr(>F)
spray        5   2669   533.8    34.7 <2e-16 ***
Residuals   66   1015    15.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Da es $p = 6$ Faktoren gibt, wird die Anzahl von $p - 1 = 5$ Freiheitsgraden für SQE berichtet. Dabei ist $SQE = 2669$ und $SQR = 1015$ mit $n - p = 66$ Freiheitsgraden. Zähler und Nenner der Teststatistik werden in der nächsten Spalte berichtet, also $SQE/(p-1)$ und $SQR/(n-p)$. Der entsprechende p-Wert ist

```
> 1 - pF(34.7,5,66)
[1] 0
```

was R kleiner als $2 \cdot 10^{-16}$ berichtet, der internen Rechengenauigkeit.

Die Varianzanalyse verläuft nach dem letzten Theorem folgendermaßen ab:

Modell der einfaktoriellen Varianzanalyse

Annahme	$X_{ki} = \beta_k \epsilon_{ki} \quad (k = 1, \ldots, p, i = 1, \ldots, n_k)$
Dabei sind	
$X_{11}, \ldots, X_{pn_p}$	gegebene Merkmalsausprägungen eines Merkmals gemessen in Levels $1, \ldots, p$
β_k	erwarteter Effekt der k-ten Faktorstufe auf die Ausprägung des Merkmals
$\epsilon_{11}, \ldots, \epsilon_{p,n_p}$	Zufallsvariablen, die die Abweichung der Messdaten des k-ten Levels messen. Diese sind unabhängig, identisch verteilt mit $\epsilon_{ki} \sim N(0, \sigma^2)$.
Hypothesen	$H_0 : \beta_1 = \ldots = \beta_p = 0 \quad$ gegen $\quad H_1 : \beta_k \neq \beta_\ell$ für ein Paar k, ℓ
Teststatistik	$F = \dfrac{SQE/(p-1)}{SQR/(n-p)} \sim F(p-1, n-p)$
Ablehnungsbereich	$F >$ $(1-\alpha)$-Quantil von $F(p-1, n-p)$
p-Wert	$1 - P_{F(p-1,n-p)}(F)$

Verbindung zu Regression

Die Varianzanalyse lässt sich mit der Regression vergleichen. Wir können die Modellannahmen auch schreiben als $Y = x\beta + \epsilon$ mit

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & & \cdots & 0 \\ \vdots & \vdots & & & \vdots \\ 1 & 0 & & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & & & \cdots & 1 \\ \vdots & & & & \vdots \\ 0 & & & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \vdots \\ \epsilon_{p1} \\ \vdots \\ \epsilon_{pn_p} \end{pmatrix}.$$

In diesem Fall ist

$$x^\top x = \begin{pmatrix} n_1 & 0 & \cdots & \cdots & 0 \\ 0 & n_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & \cdots & n_p \end{pmatrix}, \qquad (x^\top x)^{-1} = \begin{pmatrix} n_1^{-1} & 0 & \cdots & \cdots & 0 \\ 0 & n_2^{-1} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & \cdots & n_p^{-1} \end{pmatrix},$$

und damit ist

$$\hat\beta = (x^\top x)^{-1} x^\top Y = \Big(\frac{1}{n_1}(Y_{11} + \cdots + Y_{1n_1}), \cdots, \frac{1}{n_p}(Y_{p1} + \cdots + Y_{pn_p})\Big)^\top =: \big(\overline{Y}_{1\bullet}, \cdots, \overline{Y}_{p\bullet}\big)^\top$$

der kleinste-Quadrate-Schätzer von β. Das ist auch nicht erstaunlich, ist doch $\overline{Y}_{k\bullet}$ der Mittelwert der Beobachtungen in Klasse k. Weiter schreiben wir

$$\hat Y = x\hat\beta = (\underbrace{\overline{Y}_{1\bullet}, \cdots, \overline{Y}_{1\bullet}}_{n_1-\text{mal}}, \cdots, \underbrace{\overline{Y}_{p\bullet}, \cdots, \overline{Y}_{p\bullet}}_{n_p-\text{mal}}, \cdots,)^\top,$$

$$RSS = (Y - \hat Y)^\top (Y - \hat Y) = \sum_{k=1}^{p} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_{k\bullet})^2 = SQR.$$

Aus dem Beweis von Proposition 4.2 aus dem Skript *Regression* folgt damit die Varianzzerlegung

$$\sum_{k=1}^{p} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y})^2 = \sum_{k=1}^{p} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_{k\bullet})^2 + \sum_{k=1}^{p} n_k (\overline{Y}_{k\bullet} - \overline{Y})^2$$

Den Test $\beta_1 = \cdots = \beta_p$ werden wir nun für den einfacheren Fall $p = 3$ und $n_1 = n_2 = n_3 =: q$ anhand von Korollar 6.5 aus dem Skript *Regression* erklären. Hierzu setzen wir $A\beta - \gamma = 0$ für $\gamma = 0$ und

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}$$

(und also $m = 2$). Nun ist für den Zähler der Statistik aus Korollar 6.5 des Skripts zur *Regression*

$$A(x^\top x)^{-1} A^\top = \frac{1}{q} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \qquad (A(x^\top x)^{-1} A^\top)^{-1} = \frac{q}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

und damit

$$\begin{aligned} A\hat\beta &= (\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}, \overline{Y}_{2\bullet} - \overline{Y}_{3\bullet})^\top, \\ (A\hat\beta)^\top (A(x^\top x)^{-1} A^\top)^{-1} A\hat\beta &= \frac{q}{3} (A\hat\beta)^\top \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} A\hat\beta \\ &= \frac{2q}{3}((\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})^2 + (\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet})(\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}) + (\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet})^2) \\ &= \frac{2q}{3}(\overline{Y}_{1\bullet}^2 + \overline{Y}_{2\bullet}^2 + \overline{Y}_{3\bullet}^2 - \overline{Y}_{1\bullet}\overline{Y}_{2\bullet} - \overline{Y}_{1\bullet}\overline{Y}_{3\bullet} - \overline{Y}_{2\bullet}\overline{Y}_{3\bullet}) \\ &= \frac{q}{3}(3(\overline{Y}_{1\bullet}^2 + \overline{Y}_{2\bullet}^2 + \overline{Y}_{3\bullet}^2) - (\overline{Y}_{1\bullet} + \overline{Y}_{2\bullet} + \overline{Y}_{3\bullet})^2) \\ &= q(\overline{Y}_{1\bullet}^2 + \overline{Y}_{2\bullet}^2 + \overline{Y}_{3\bullet}^2 - 3\overline{Y}^2) = SQE \end{aligned}$$

und

$$\widehat{\sigma^2} = \frac{RSS}{n-p} = \frac{SQR}{n-p}.$$

Damit folgt, dass die Statistik aus Korollar 6.5 identisch mit der aus Theorem 4 ist.

Erweiterungen

Bemerkung 6 (Untersuchung bei signifikantem Ergebnis, Tukey's Test). Kann man nun H_0 ablehnen, stellt man sich sofort die Frage, zwischen welchen Levels der Unterschied der Mittelwerte denn für dieses Ergebnis entscheidend war. Hierfür kann man einen *Post-Hoc*-Test an die Varianzanalyse anschließen. Einer dieser Tests ist *Tukey's Test*. Er basiert auf der t-Range-Statistik. Im Modell der Varianzanalyse ist für zwei Gruppen k, ℓ (falls $n_1 = \cdots = n_p$) ist

$$Q := \frac{\max_k \hat{\beta}_k - \min_k \hat{\beta}_k}{\sqrt{\widehat{\sigma^2}/n_k}}$$

Um etwas über diese Verteilung zu erfahren, stehen die R-Befehle `ptukey` und `qtukey` zur Verfügung. Ausführung des Tukey-Tests für den `InsectSprays`-Datensatz liefert

```
> TukeyHSD(aov.out)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = count ~ spray)

$spray
           diff        lwr       upr     p adj
B-A   0.8333333  -3.866075  5.532742 0.9951810
C-A -12.4166667 -17.116075 -7.717258 0.0000000
D-A  -9.5833333 -14.282742 -4.883925 0.0000014
E-A -11.0000000 -15.699409 -6.300591 0.0000000
F-A   2.1666667  -2.532742  6.866075 0.7542147
C-B -13.2500000 -17.949409 -8.550591 0.0000000
D-B -10.4166667 -15.116075 -5.717258 0.0000002
E-B -11.8333333 -16.532742 -7.133925 0.0000000
F-B   1.3333333  -3.366075  6.032742 0.9603075
D-C   2.8333333  -1.866075  7.532742 0.4920707
E-C   1.4166667  -3.282742  6.116075 0.9488669
F-C  14.5833333   9.883925 19.282742 0.0000000
E-D  -1.4166667  -6.116075  3.282742 0.9488669
F-D  11.7500000   7.050591 16.449409 0.0000000
F-E  13.1666667   8.467258 17.866075 0.0000000
```

Hier sieht man, welche paarweisen Vergleiche signifikant sind, wenn man die letzte Spalte betrachtet. Zu beachten ist hier, dass *gleichzeitig* insgesamt 15 Tests zum Signifikanzniveau 5% durchgeführt werden würden, so dass mit mindestens einem signifikanten Ergebnis zu

rechnen ist, auch wenn H_0 zutrifft. R reagiert darauf, indem das Signifikanzniveau angepasst ist. Auf dieses Thema werden wir später zurückkommen.

Bemerkung 7 (Ungleiche Varianzen). Neben der Varianzanalyse von oben gibt es noch die Möglichkeit, in R eine Varianzanalyse ohne die Annahme der gleichen Varianzen durchzuführen.

```
> oneway.test(count~spray)

One-way analysis of means (not assuming equal variances)

data:  count and spray
F = 36.0654, num df = 5.000, denom df = 30.043, p-value = 7.999e-12
```

Hier wird die Anzahl der Freiheitsgrade von SQR an die unterschiedlichen Varianzen angepasst.

Bemerkung 8 (Zwei-faktorielle Varianzanalyse). Gibt es nicht nur einen Faktor, sondern zwei, hilft die Zwei-faktorielle Varianzanalyse weiter. Hinter dieser steckt das Modell

$$Y_{k\ell i} = \beta_{k\bullet} + \beta_{\bullet\ell} + \epsilon_{k\ell i},$$

wobei $\beta_{k\bullet}$ den Effekt des Levels k für den ersten Faktor und $\beta_{\bullet\ell}$ den Effekt des Levels ℓ für den zweiten Faktor beschreibt. Ähnliche Tests wie oben können auch für eine zwei-faktorielle Varianzanalyse durchgeführt werden.