§11 Varianzanalyse pfad

▼

```
SetDirectory[$UserDocumentsDirectory];


<< StatisticalPlots`;
<< ANOVA`;

BartlettTest[daten_] := Module[{n, s, dat, nl, vl, am, gm},
    n = Length[daten];
    s = Length[Union[Part[daten, All, 1]]];
    dat = Table[Part[Select[daten, #[[1]] == i &], All, 2], {i, 1, s}];
    nl = Table[Length[dat[[i]]], {i, 1, s}];
    vl = Table[CentralMoment[dat[[i]], 2], {i, 1, s}];
    am = nl.vl / n;
    gm = Exp[nl.Log[vl] / n];
    Print["PValue→", 1 - CDF[ChiSquareDistribution[s - 1], n Log[am / gm]] // N]]
```

Bei wissenschaftlichen Untersuchungen geht man oft davon aus, dass ein oder mehrere Faktoren eine Messung beeinflussen können. Ein typisches Beispiel dafür ist der Hektarertrag einer bestimmten Weizensorte, der von den Faktoren Düngung, Bodenbeschaffenheit, Klimabedingung usw abhängt.

Will man den Einfluss eines oder mehrerer Faktoren auf das Messergebnis analysieren, so müssen alle anderen Einflussgrößen entweder konstant gehalten oder ausgeschaltet werden. Im allgemeinen wird dies aber nicht möglich sein. In der Regel treten nämlich neben den beobachtbaren Einflussgrößen auch noch nicht beobachtbare sowie zufällige Einflussgrößen auf. Damit setzt sich das Ergebnis eines einzelnen Versuchs stets aus zwei Teilen zusammen. Der erste Teil hängt dabei nur von den entsprechenden Kombinationen der Stufen ab, in denen die betrachteten Einflussgrößen wirken. Der zweite Teil ist eine zufällige Größe die alle übrigen nicht beobachtbaren und zufälligen Einflussgrößen berücksichtigt.

11.1 Einfache Varianzanalyse

Bei der einfachen Varianzanalyse hat man es mit einem Faktor S zu tun, der in s Stufen auf den Mittelwert von normalverteilten Messwerten einwirken kann. Wir treffen dazu die folgenden Annahmen:

■ Falls der Faktor S in der i-ten Stufe wirkt, so lässt sich das Messergebnis durch eine $\mathcal{N}[\mu_i, \sigma]$-verteilte Zufallsvariable X_i beschreiben. Der Faktor S wirkt also nur auf den Mittelwert ein. Die nicht beobachtbaren bzw die zufälligen Einflussgrößen verursachen eine $\mathcal{N}[0, \sigma]$-verteilte Abweichung vom jeweiligen Mittelwert μ_i.

■ Für den Fall, dass der Faktor S in der i-ten Stufe wirkt, liegen $n_i \geq 2$ Messungen vor, die sich durch die vollständig unabhängigen Zufallsvariablen $X_{i\,1}, X_{i\,2}, \ldots, X_{i\,n_i}$ beschreiben lassen.

Soll geprüft werden, ob der Faktor S einen Einfluss auf die Messwerte hat, so hat man es mit dem a-priori Modell

$$\mathfrak{P}_{X_1} \times \mathfrak{P}_{X_2} \times \ldots \times \mathfrak{P}_{X_s} = \{\{\mathcal{N}[\mu_1, \sigma], \ldots, \mathcal{N}[\mu_s, \sigma]\} \mid \mu_1, \ldots, \mu_s \in \mathbb{R},\ \sigma > 0\}$$

zu tun, für das die Hypothese $\mathcal{H}_0$ …"alle μ_i sind gleich" gegen die Alternative $\mathcal{H}_1$ …"nicht alle μ_i sind gleich" getestet werden muss. Natürlich könnte man für dieses Testproblem den entsprechenden Maximum-Likelihood-Quotiententest entwickeln. Da dieser Test jedoch einen großen Stichprobenumfang benötigt und auch dann nur annähernd das vorgegebene Signifikanzniveau besitzt, werden wir auf andere Weise einen Test für diese Fragestellung erarbeiten, der auch bei kleinem Stichprobenumfang exakt das vorgegebene Signifikanzniveau besitzt.

Unter Verwendung der Reparametrisierung $\mu_i = \mu + \xi_i$ mit

$$\mu = \frac{1}{n_\bullet} \sum_{i=1}^{s} n_i \mu_i \qquad \text{und} \qquad \xi_i = \mu_i - \mu$$

wobei $n_\bullet = n_1 + n_2 + \ldots + n_s$ bezeichnet und offenbar $\sum_{i=1}^{s} n_i \xi_i = 0$ gilt, lässt sich dieses Testproblem in der folgenden Weise formulieren: Gesucht ist ein Test für die Hypothese $\mathcal{H}_0$ …"alle ξ_i sind 0" gegen die Alternative $\mathcal{H}_1$ …"nicht alle ξ_i sind 0". Mit den beiden Statistiken

$$\overline{X}_{i\bullet} = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik} \qquad \text{und} \qquad \overline{X}_{\bullet\bullet} = \frac{1}{n_\bullet} \sum_{i=1}^{s} \sum_{k=1}^{n_i} X_{ik}$$

(man beachte, dass es sich dabei um erwartungstreue Schätzer für die Parameter μ_i bzw μ handelt) führen wir nun drei für die einfache Varianzanalyse fundamentale Bezeichnungen ein:

11.1.1 Definition:

a) Die Statistik SSS (**S**um of **S**quares of factor **S**) mit

$$\text{SSS} = \sum_{i=1}^{s} n_i \, (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2$$

ist ein sinnvoller Schätzer für den Wert $\sum_{i=1}^{s} n_i \, \xi_i^2$. Die Statistik SSS ist somit genau dann klein (nur wenig größer als Null), wenn die Hypothese $\mathcal{H}_0$ zutrifft, der Faktor S also keinen Einfluss hat. Dabei hängt dieses "klein sein" aber auch von der (unbekannten) Streuung σ ab und ist daher noch geeignet zu relativieren.

b) Die Statistik SSE (**S**um of **S**quares of **E**rror) mit

$$\text{SSE} = \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\bullet})^2$$

ist ein sinnvoller Schätzer für den Wert $n_\bullet \, \sigma^2$. Mit Hilfe der Statistik SSE lässt sich dieses "klein sein" der Statistik SSS in der üblichen Weise relativieren.

c) Die Statistik SStotal (**S**um of **S**quares)

$$\text{SStotal} = \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{\bullet\bullet})^2$$

entspricht der Summe der quadratischen Abweichungen der einzelnen Messwerte vom Gesamtmittelwert. Dabei gilt die als Zerlegung der Varianz bekannte Formel

$$\text{SStotal} = \text{SSS} + \text{SSE}$$

▼

Wir zeigen die Zerlegung der Varianz: Beachtet man, dass für alle $1 \le i \le s$ offenbar

$$\sum_{k=1}^{n_i} \left(X_{ik} - \overline{X}_{i\bullet}\right) = n_i \, \overline{X}_{i\bullet} - n_i \, \overline{X}_{i\bullet} = 0$$

ist, so gilt

$$\text{SStotal} = \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\bullet} + \overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 =$$

$$= \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\bullet})^2 + 2 \sum_{i=1}^{s} \sum_{k=1}^{n_i} \left(X_{ik} - \overline{X}_{i\bullet}\right)\left(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet}\right) + \sum_{i=1}^{s} \sum_{k=1}^{n_i} (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 =$$

$$= \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{ik} - \overline{X}_{i\bullet})^2 + \sum_{i=1}^{s} n_i \, (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 = \text{SSE} + \text{SSS}$$

Von zentraler Bedeutung für die einfache Varianzanalyse ist der folgende

11.1.2 Satz: Falls die Hypothese $\mathcal{H}_0$ zutrifft, also alle ξ_i gleich 0 sind, so genügt die Testgröße

$$\frac{n_\bullet - s}{s-1}\,\frac{\text{SSS}}{\text{SSE}} = \frac{n_\bullet - s}{s-1}\,\frac{\sum_{i=1}^{s} n_i\,(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2}{\sum_{i=1}^{s}\sum_{k=1}^{n_i}(X_{ik} - \overline{X}_{i\bullet})^2}$$

einer Fisher F Verteilung mit den Parametern $s-1$ und $n_\bullet - s$.

▼

Beweis: Für jedes $1 \le i \le s$ bezeichne $\mathbf{\Delta}_i$ die Matrix

$$\mathbf{\Delta}_i = \frac{1}{n_i}\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n_i}_{n_i}$$

Außerdem bezeichne

$$\mathbf{\Gamma}_1 = \frac{1}{n_\bullet}\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{n_\bullet}_{n_\bullet} \qquad \text{und} \qquad \mathbf{\Gamma}_2 = \begin{pmatrix} \mathbf{\Delta}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_2 & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Delta}_s \end{pmatrix} \in \mathbb{R}^{n_\bullet}_{n_\bullet}$$

a) Offenbar gilt (mit $\mathbf{E}_k$ bezeichnen wir die $k \times k$ Einheitsmatrix)

$$(\mathbf{E}_{n_\bullet} - \mathbf{\Gamma}_2) + (\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1) + \mathbf{\Gamma}_1 = \boldsymbol{E}_{n_\bullet}$$

b) Weiters gilt

$$\text{Rg}[\boldsymbol{E}_{n_\bullet} - \mathbf{\Gamma}_2] + \text{Rg}[\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1] + \text{Rg}[\mathbf{\Gamma}_1] = n_\bullet$$

Diese Tatsache wird durch die folgenden Überlegungen klar:

i) Die Matrix $\boldsymbol{E}_{n_\bullet} - \mathbf{\Gamma}_2$ besitzt den Rang $n_\bullet - s$:

Die Matrix $\boldsymbol{E}_{n_\bullet} - \mathbf{\Gamma}_2$ ist eine Block-Diagonalmatrix, bei der die einzelnen Blöcke $\boldsymbol{E}_{n_i} - \mathbf{\Delta}_i$ den Rang $n_i - 1$ besitzen. Addiert man nämlich alle Zeilen der Matrix $\boldsymbol{E}_{n_i} - \mathbf{\Delta}_i$, so erhält man eine Zeile, welche aus lauter Nullen besteht. Also ist der Rang der Matrix $\boldsymbol{E}_{n_i} - \mathbf{\Delta}_i$ kleiner als n_i. Subtrahiert man die jeweils folgende Spalte der Matrix $\boldsymbol{E}_{n_i} - \mathbf{\Delta}_i$ von der vorhergehenden, so erhält man eine Dreiecks-Matrix mit lauter Einsern in der Hauptdiagonale. Also ist der Rang der Matrix $\boldsymbol{E}_{n_i} - \mathbf{\Delta}_i$ mindestens $n_i - 1$.

ii) Die Matrix $\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1$ besitzt den Rang $s-1$:

Streicht man bei der Matrix $\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1$ gleiche Zeilen, so bleibt eine Matrix $\mathbf{M}$ mit s Zeilen übrig. Summiert man das n_i-fache der i-ten Zeilen dieser Matrix $\mathbf{M}$, so erhält man eine Zeile, welche aus lauter Nullen besteht. Also ist der Rang der Matrix $\mathbf{M}$ und damit auch der Rang der Matrix $\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1$ kleiner als s. Subtrahiert man die jeweils folgende Spalte der Matrix $\mathbf{M}$ von der vorhergehenden, so erhält man eine Dreiecks-Matrix, bei der in der Hauptdiagonalen die Werte $1/n_1$, $1/n_2$, …, $1/n_{s-1}$ stehen. Also ist der Rang der Matrix $\mathbf{M}$ und damit auch der Rang der Matrix $\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1$ mindestens $s-1$.

iii) Die Matrix $\mathbf{\Gamma}_1$ besitzt lauter gleiche Zeilen und hat damit den Rang 1.

c) Mit den vollständig unabhängigen und identisch $\mathcal{N}[0, 1]$-verteilten Zufallsvariablen $Z_{ik} = \left(X_{ik} - \mu\right)/\sigma$ und der Abkürzung $\vec{Z} = \{Z_{11}, Z_{12}, \dots, Z_{1\,n_1}; Z_{21}, Z_{22}, \dots, Z_{2\,n_2}; \dots; Z_{s\,1}, Z_{s\,2}, \dots, Z_{2\,n_s}\}$ gilt

$$\frac{1}{\sigma^2}\,\text{SSS} = \frac{1}{\sigma^2}\sum_{i=1}^{s} n_i\,(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{s} n_i\,(\overline{Z}_{i\bullet} - \overline{Z}_{\bullet\bullet})^2 =$$

$$= \{\overline{Z}_{1\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \overline{Z}_{1\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}; \ldots; \overline{Z}_{s\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \overline{Z}_{s\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}\} \cdot$$

$$\cdot \{\overline{Z}_{1\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \overline{Z}_{1\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}; \ldots; \overline{Z}_{s\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \overline{Z}_{s\,\blacksquare} - \overline{Z}_{\blacksquare\,\blacksquare}\}^t =$$

$$= \{\overline{Z}_{1\,\blacksquare}, \ldots, \overline{Z}_{1\,\blacksquare}; \ldots; \overline{Z}_{s\,\blacksquare}, \ldots, \overline{Z}_{s\,\blacksquare}\} \cdot \{\overline{Z}_{1\,\blacksquare}, \ldots, \overline{Z}_{1\,\blacksquare}; \ldots; \overline{Z}_{s\,\blacksquare}, \ldots, \overline{Z}_{s\,\blacksquare}\}^t -$$

$$- \{\overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \overline{Z}_{\blacksquare\,\blacksquare}; \ldots; \overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \blacksquare\blacksquare\} \cdot \{\overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \overline{Z}_{\blacksquare\,\blacksquare}; \ldots; \overline{Z}_{\blacksquare\,\blacksquare}, \ldots, \blacksquare\blacksquare\}^t =$$

$$= \vec{Z} \cdot \mathbf{\Gamma}_2 \cdot \vec{Z}^t - \vec{Z} \cdot \mathbf{\Gamma}_1 \cdot \vec{Z}^t = \vec{Z} \cdot (\mathbf{\Gamma}_2 - \mathbf{\Gamma}_1) \cdot \vec{Z}^t$$

sowie

$$\frac{1}{\sigma^2}\,\mathrm{SSE} = \frac{1}{\sigma^2} \sum_{i=1}^{s} \sum_{k=1}^{n_i} (X_{i\,k} - \overline{X}_{i\,\blacksquare})^2 = \sum_{i=1}^{s} \sum_{k=1}^{n_i} (Z_{i\,k} - \overline{Z}_{i\,\blacksquare})^2 =$$

$$= \{Z_{1\times 1} - \overline{Z}_{1\,\blacksquare}, \ldots, Z_{1\,n_1} - \overline{Z}_{1\,\blacksquare}; \ldots; Z_{s\,1} - \overline{Z}_{s\,\blacksquare}, \ldots, Z_{s\,n_s} - \overline{Z}_{s\,\blacksquare}\} \cdot$$

$$\cdot \{Z_{1\times 1} - \overline{Z}_{1\,\blacksquare}, \ldots, Z_{1\,n_1} - \overline{Z}_{1\,\blacksquare}; \ldots; Z_{s\,1} - \overline{Z}_{s\,\blacksquare}, \ldots, Z_{s\,n_s} - \overline{Z}_{s\,\blacksquare}\}^t =$$

$$= \{Z_{1\times 1}, \ldots, Z_{1\,n_1}; \ldots; Z_{s\,1}, \ldots, Z_{s\,n_s}\} \cdot \{Z_{1\times 1}, \ldots, Z_{1\,n_1}; \ldots; Z_{s\,1}, \ldots, Z_{s\,n_s}\}^t -$$

$$- \{\overline{Z}_{1\,\blacksquare}, \ldots, \overline{Z}_{1\,\blacksquare}; \ldots; \overline{Z}_{s\,\blacksquare}, \ldots, \overline{Z}_{s\,\blacksquare}\} \cdot \{\overline{Z}_{1\,\blacksquare}, \ldots, \overline{Z}_{1\,\blacksquare}; \ldots; \overline{Z}_{s\,\blacksquare}, \ldots, \overline{Z}_{s\,\blacksquare}\}^t =$$

$$= \vec{Z} \cdot \mathbf{E} \cdot \vec{Z}^t - \vec{Z} \cdot \mathbf{\Gamma}_2 \cdot \vec{Z}^t = \vec{Z} \cdot (\mathbf{E} - \mathbf{\Gamma}_2) \cdot \vec{Z}^t$$

Aus dem Satz von Cochran zusammen mit Satz 23.5.1 folgt die Aussage des Satzes unmittelbar.

Damit ist offensichtlich, wie sich die Frage, ob der Faktor *S* tatsächlich einen Einfluss auf den Mittelwert der normalverteilten Messwerte besitzt, überprüfen lässt:

11.1.3 Die einfache Varianzanalyse wird verwendet, wenn geprüft werden soll, ob der Faktor *S*, welcher in *s* Stufen wirken kann, tatsächlich einen Einfluss auf den Mittelwert von normalverteilten Messwerten besitzt:

$\mathcal{P}_{X_1} \times \mathcal{P}_{X_2} \times \ldots \times \mathcal{P}_{X_S}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\{\mathcal{N}[\mu_1, \sigma], \ldots, \mathcal{N}[\mu_S, \sigma]\} \mid \mu_1, \ldots, \mu_S \in \mathbb{R},\ \sigma > 0\}$	alle μ_i sind gleich $\sigma > 0$	nicht alle μ_i sind gleich $\sigma > 0$	$\frac{n_\blacksquare - s}{s-1} \frac{\mathrm{SSS}}{\mathrm{SSE}} > f_{s-1,n_\blacksquare - s;1-\alpha}$

Dabei bezeichnet $f_{n,m;q}$ das q-Quantil der $\mathcal{F}[n, m]$-Verteilung.

▼

Im Fall $s = 2$ stimmt die Fragestellung der einfachen Varianzanalyse mit der Fragestellung des *t*-Tests für zwei Grundgesamtheiten überein. Damit kann man die einfache Varianzanalyse in gewisser Weise als Verallgemeinerung des *t*-Tests für zwei Grundgesamtheiten ansehen.

Die mit der einfachen Varianzanalyse verbundenen Berechnungen sind rechenintensiv. Daher ist die einfache Varianzanalyse auch in *Mathematica* implementiert. Man lade dazu zuerst das Paket ANOVA` und verwende den Befehl ANOVA in der folgenden Form:

```
<< ANOVA`
```

ladet das Paket ANOVA`.

■ ANOVA[*daten*]

führt für das Datenmaterial *daten* eine einfache Varianzanalyse durch. Das Datenmaterial *daten* muss dabei die Form $\{\{s_1, x_1\}, \{s_2, x_2\}, \ldots\}$ besitzen, wobei der Eintrag s_i den zum i-ten Messwert x_i gehörenden Wert des Faktors *S* bezeichnet.

Der von *Mathematica* gelieferte Output hat dabei eine leicht verständliche Gestalt: Neben den Freiheitsgraden (DF) und den Schätzwerten der Statistiken SSS, SSE, SStotal (SumOfSq) sowie $SSS/(s-1)$ und $SSE/(n_\bullet - s)$ (MeanSq) werden die Testgröße (FRatio) und der zugehörige p-Wert (PValue) ausgegeben.

Außerdem werden die Schätzwerte $x_{\bullet\bullet}$ und $x_{i\bullet}$ von μ und μ_i angeführt. Sollen diese Schätzwerte nicht ausgegeben werden, so verwende man die Option CellMeans → False.

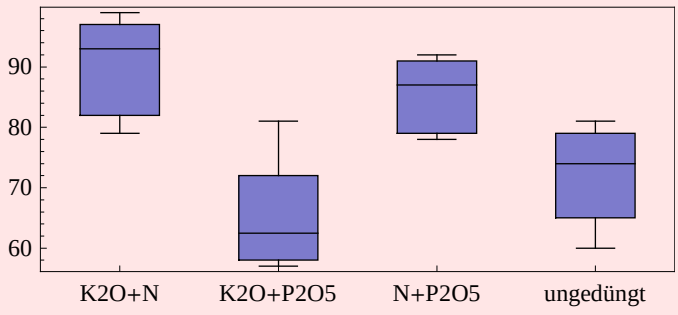
Wir demonstrieren die Verwendung der einfachen Varianzanalyse an einigen konkreten Beispielen:

11.1.4 Beispiel: Es soll untersucht werden, ob die Art des Düngemittels (es stehen die drei Düngemittel K2O + N, K2O + P2O5, N + P2O5 zur Verfügung; zum Vergleich wurden einige Parzellen nicht gedüngt) einen Einfluss auf den mittleren Ernteertrag von Weizen hat. Dazu wurde ein Feld in einige gleich große, gleichwertige Parzellen eingeteilt und diese Parzellen mit unterschiedlichen Düngemitteln behandelt. Der dabei erzielte Ernteertrag findet sich im Datenmaterial düngemittel.

▼

Lösung: Wir veranschaulichen dieses Datenmaterial zuerst graphisch durch ein Box-Plot:

```
düngemittel = Rest[<< "düngemittelfile"];
e1 = Select[düngemittel, #1[[2]] == 1 &][[All, 3]];
e2 = Select[düngemittel, #1[[2]] == 2 &][[All, 3]];
e3 = Select[düngemittel, #1[[2]] == 3 &][[All, 3]];
e4 = Select[düngemittel, #1[[2]] == 4 &][[All, 3]];
BoxWhiskerPlot[e1, e2, e3, e4, BoxLabels → {"K2O+N", "K2O+P2O5", "N+P2O5", "ungedüngt"}, ImageSize → {
Clear[e1, e2, e3, e4]
```



An Hand dieser Graphik erkennen wir, dass der Faktor Düngemittel sehr wohl einen Einfluss auf den mittleren Ernteertrag von Weizen haben dürfte. Um diese Vermutung zu bestätigen, wollen wir dieses Datenmaterial einer Varianzanalyse unterziehen. Dazu ist erforderlich, zuerst mit Hilfe des Bartlett-Tests zu überprüfen, ob die Streuungen der Ernteerträge, welche mit den einzelnen Düngemitteln erzielt werden, übereinstimmen (eine Überprüfung, ob die einzelnen Grundgesamtheiten normalverteilt sind, ist wegen des geringen Stichprobenumfangs nicht sinnvoll):

```
BartlettTest[düngemittel[[All, {2, 3}]]]
```

```
PValue→0.876343
```

Da der p-Wert des Bartlett-Tests groß ist, können wir annehmen, dass die Streuungen der Ernteerträge, welche mit den einzelnen Düngemitteln erzielt werden, annähernd gleich sind (diese Tatsache erkennt man natürlich auch am Boxplot). Wir führen nun für unser Datenmaterial eine Varianzanalyse durch:

```
ANOVA[düngemittel[[All, {2, 3}]]]
Clear[düngemittel]
```

```
{ANOVA → |        | DF | SumOfSq  | MeanSq   | FRatio | PValue       |
          | Model  | 3  | 2585.    | 861.667  | 13.1693| 0.0000388667 |
          | Error  | 22 | 1439.46  | 65.4301  |                       |
          | Total  | 25 | 4024.46  |                                  |

 CellMeans → {All      77.5385
              Model[1] 89.7143
              Model[2] 66.
              Model[3] 85.4
              Model[4] 72.1667}
```

Da der p-Wert sehr klein ist, wird die Hypothese $\mathcal{H}_0$ …"alle μ_i sind gleich" deutlich abgelehnt. Der Faktor Düngemittel hat somit einen sehr signifikanten Einfluss auf den mittleren Ernteertrag von Weizen. An den einzelnen Zellenmittelwerten erkennt man außerdem, dass der mittlere Ernteertrag bei Verwendung des Düngemittels K2O+N am größten ist.

11.1.5 Beispiel: In fünf Laboratorien wurde mit Hilfe des Michelson-Versuches jeweils zwanzig mal die Lichtgeschwindigkeit bestimmt und jener die Geschwindigkeit von 299 000 km/sek übersteigende Wert tabelliert (vergleiche dazu das Datenmaterial michelson). Man prüfe, ob der Faktor "Laboratorium" einen Einfluss auf den Mittelwert der Messwerte hat.

▼

Lösung: Wir haben uns mit diesem Datenmaterial bereits in Beispiel 10.2.4 beschäftigt und gesehen, dass die in den einzelnen Laboratorien ermittelten Messwerte in signifikant unterschiedlicher Weise streuen. Damit dürfte die einfache Varianzanalyse für diese Fragestellung eigentlich nicht herangezogen werden. Da aber die einfache Varianzanalyse recht robust ist (sowohl hinsichtlich der Tatsache, dass die einzelnen Grundgesamtheiten möglicherweise nicht normalverteilt sind, als auch hinsichtlich einer Verletzung der Gleichheit der Streuungen dieser Grundgesamtheiten), dürfen wir die Varianzanalyse auch in dieser Situation verwenden:

```
geschwindigkeit = Rest[<< michelsonfile];
ANOVA[geschwindigkeit, CellMeans → False]
Clear[geschwindigkeit]
```

```
ANOVA → |        | DF | SumOfSq   | MeanSq   | FRatio | PValue     |
        | Model  | 4  | 94 514.   | 23 628.5 | 4.2878 | 0.00311445 |
        | Error  | 95 | 523 510.  | 5510.63  |                      |
        | Total  | 99 | 618 024.  |                                    |
```

Da der p-Wert klein ist, wird die Hypothese $\mathcal{H}_0$ …"alle μ_i sind gleich" abgelehnt. Der Faktor "Labor" hat somit einen signifikanten Einfluss auf die Mittelwerte der in den einzelnen Laboratorien gemessenen Lichtgeschwindigkeit.

11.1.6 Beispiel: In Beispiel 2.2.4 haben wir uns mit dem Datenmaterial *stahl* befasst und die Abhängigkeit des quantitativen Merkmals Zugfestigkeit vom qualitativen Merkmal Sorte graphisch veranschaulicht. Wir sind dabei zu der Auffassung gelangt, dass die Zugfestigkeit von Stahlblechen der Sorte *B* insgesamt etwas höher sein dürfte, als die Zugfestigkeit von Stahlblechen der beiden anderen Sorten. Man prüfe nun mit Hilfe der Varianzanalyse, ob der Faktor "Sorte" tatsächlich einen Einfluss auf den Mittelwert der Zugfestigkeit von Stahlblechen hat.

▼

Lösung: Wir unterziehen die erste Spalte (Sorte) und dritte Spalte (Zugfestigkeit) des Datenmaterials *stahl* einer Varianzanalyse:

```
zugfestigkeit = Rest[<< stahlfile];
ANOVA[Part[zugfestigkeit, All, {1, 3}], CellMeans → False]
Clear[zugfestigkeit]
```

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	8545.11	4272.55	2.13038	0.125243
Error	83	166 460.	2005.54		
Total	85	175 005.			

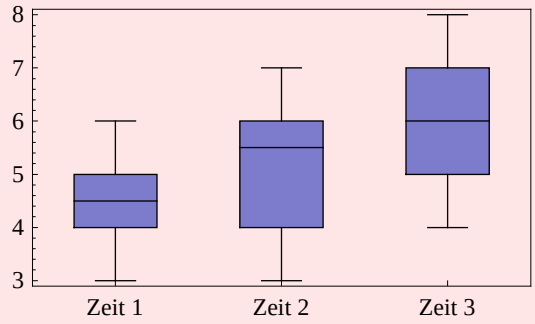
Da der *p*-Wert deutlich größer als 0.05 ist, kann nicht behauptet werden, dass der Faktor "Sorte" einen signifikanten Einfluss auf die mittlere Zugfestigkeit besitzt. Die seinerzeit durch das Boxplot nahe gelegte Vermutung, Stahlbleche der Sorte *B* hätten im Durchschnitt eine signifikant höhere Zugfestigkeit als Stahlbleche der beiden anderen Sorten, ist somit nicht signifikant belegbar.

11.1.7 Beispiel: Dem Produktionsprozess eines gewissen Erzeugnisses wurden in größeren Zeitabständen drei Stichproben von je acht Einheiten entnommen und davon eine bestimmte Messgröße ermittelt (man vergleiche dazu das Datenmaterial messgröße). Es soll geprüft werden, ob die im Laufe der Zeit eintretende Veränderung der Einstellung der Maschinen die Messwerte signifikant beeinflusst.

▼

Lösung: Wir veranschaulichen den Einfluss des Faktors "Zeit" zuerst graphisch durch ein Box-Plot

```
messgröße = Rest[<< "messgrößefile"];
z1 = Select[messgröße, #1[[1]] == 1 &][[All, 2]];
z2 = Select[messgröße, #1[[1]] == 2 &][[All, 2]];
z3 = Select[messgröße, #1[[1]] == 3 &][[All, 2]];
BoxWhiskerPlot[z1, z2, z3, BoxLabels → {"Zeit 1", "Zeit 2", "Zeit 3"}, ImageSize → {250, 120}]
Clear[z1, z2, z3]
```



und erkennen, dass der Faktor "Zeit" einen deutlichen Einfluss auf die Mittelwerte der Messwerte haben dürfte. Um diesen Einfluss auch quantitativ fassen zu können, unterziehen wir unser Datenmaterial einer einfachen Varianzanalyse:

```
ANOVA[messgröße, CellMeans → False]
Clear[messgröße]
```

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	12.25	6.125	3.85393	0.0375226
Error	21	33.375	1.58929		
Total	23	45.625			

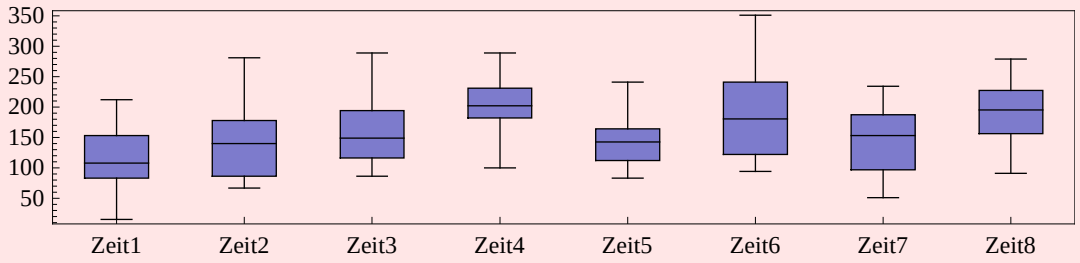
Da der p-Wert kleiner als 0.05 ist, können wir behaupten, dass der Faktor "Zeit" die Mittelwerte der Messwerte signifikant beeinflusst.

11.1.8 Beispiel: Seit dem Jahr 1850 gibt es Aufzeichnungen über die Niederschlagsmengen der Wintermonate in Linz (man vergleiche dazu das Datenmaterial wetter). Es soll geprüft werden, ob der Faktor "Zeit" einen signifikanten Einfluss auf diese Niederschlagsmengen hat.

▼

Lösung: Wir unterteilen den zur Verfügung stehenden Zeitbereich von 144 Jahren mit Hilfe von Partition in 8 Zeitbereiche von jeweils 18 aufeinanderfolgende Jahre und veranschaulichen den Einfluss des Faktors "Zeit" zuerst graphisch durch ein Box-Plot:

```
wetter = Partition[Part[Rest[<< "wetterfile"], All, 2], 18];
BoxWhiskerPlot[Transpose[wetter],
  BoxLabels → {"Zeit1", "Zeit2", "Zeit3", "Zeit4", "Zeit5", "Zeit6", "Zeit7", "Zeit8"},
  ImageSize → {400, 100}]
```



An dieser Graphik erkennt man, dass der Faktor "Zeit" einen Einfluss auf den Mittelwert der gemessenen Niederschlagsmengen haben dürfte. Um diesen Einfluss auch quantitativ fassen zu können, unterziehen wir unser Datenmaterial einer einfachen Varianzanalyse (wobei wir zuerst das Datenmaterial mit Hilfe von Table und Flatten in jene Form bringen, welche für die Anwendung des Befehls ANOVA erforderlich ist):

```
daten = Flatten[Table[{i, wetter[[i, k]]}, {k, 1, 18}, {i, 1, 8}], 1];
ANOVA[daten, CellMeans → False]
Clear[wetter, daten]
```

```
                DF     SumOfSq     MeanSq      FRatio      PValue
ANOVA → Model   7      105 303.    15 043.3    4.87156     0.0000628614
        Error   136    419 968.    3088.
        Total   143    525 271.
```

Der Einfluss des Faktors "Zeit" ist demnach sehr signifikant.

Führt die Varianzanalyse zur Ablehnung der Hypothese $\mathcal{H}_0$ …" alle μ_i sind gleich", so wird man als nächstes fragen, welche der μ_i's nun tatsächlich voneinander verschieden sind. Es wäre aber falsch, alle möglichen Paare von μ_i's mit Hilfe des t-Tests für zwei Grundgesamtheiten mit Signifikanz α auf Gleichheit zu überprüfen. Da man dabei nämlich $s\,(s-1)/2$ Tests durchführen müsste und bei jedem dieser Tests mit der Wahrscheinlichkeit α eine Fehlentscheidung 1. Art auftreten kann, hätte man insgesamt mit einer viel größeren Fehlerwahrscheinlichkeit zu rechnen. Eine einfache Möglichkeit würde darin bestehen, das zulässige Signifikanzniveau α auf diese Tests gleichmäßig aufzuteilen. Diese Methode ist aber problematisch, denn

- es wird der eventuell unterschiedliche Stichprobenumfang nicht berücksichtigt;
- es wird nicht berücksichtigt, dass sich diese Tests auf zum Teil gleiche Grundgesamtheiten beziehen;
- es wird nicht berücksichtigt, dass sich manche Paare von μ_i's mehr unterscheiden als andere.

Im Laufe der Zeit wurden zahlreiche effizientere Methoden entwickelt, mit denen man simultan prüfen kann, für welche Paare von Grundgesamtheiten sich die zugehörigen Mittelwerte μ_i signifikant unterscheiden, wobei die Wahrscheinlichkeit von Fehlentscheidungen erster Art insgesamt durch das Signifikanzniveau α beschränkt ist - man spricht in diesem Zusammenhang vom Familien-Signifikanzniveau. Wir wollen diese Methoden nicht im Einzelnen besprechen, sondern nur zeigen, wie sich die beiden von C. E. BONFERRONI bzw J. W. TUKEY entwickelten

Methoden mit Hilfe von *Mathematica* aufrufen lassen:

Verwendet man beim Befehl ANOVA die Option PostTests → {Bonferroni, Tukey}, so gibt *Mathematica* Listen jener Paare von Grundgesamtheiten aus, deren Mittelwerte sich bei Verwendung der Methode von Bonferroni bzw Tukey signifikant voneinander unterscheiden. Standardmäßig wird dabei als Familien-Signifikanzniveau der Wert $\alpha = 0.05$ verwendet. Mit der Option SignificanceLevel lässt sich dieses Familien-Signifikanzniveau in beliebiger Weise abändern.

11.1.9 Beispiel: Wir behandeln nochmals die Fragestellung von Beispiel 11.1.4, fragen nun aber auch, bei welchen Paaren von Düngemitteln sich die mittleren Ernteerträge deutlich voneinander unterscheiden.

▼

Lösung: Auf Grund der in Beispiel 11.1.4 angeführten Box-Plots vermuten wir, dass sich die Mittelwerte der Grundgesamtheiten {1, 2}, {2, 3} und {1, 4} deutlich voneinander unterscheiden. Weniger deutlich scheinen sich hingegen die Mittelwerte der Grundgesamtheiten {3, 4} zu unterscheiden. Kein deutlicher Unterschied scheint schließlich zwischen den Mittelwerten der Grundgesamtheiten {1, 3} und {2, 4} zu bestehen. Wir wollen diese Vermutungen nun durch eine Varianzanalyse bestätigen und verwenden dazu den Befehl ANOVA mit der Option PostTests → {Bonferroni, Tukey}:

```
düngemittel = Rest[<< "düngemittelfile"];
ANOVA[düngemittel[[All, {2, 3}]], PostTests → {Bonferroni, Tukey}]
```

```
{ANOVA → 
              DF    SumOfSq    MeanSq     FRatio     PValue
  Model       3     2585.      861.667    13.1693    0.0000388667
  Error       22    1439.46    65.4301
  Total       25    4024.46                                      ,

 CellMeans → {All       77.5385
              Model[1]  89.7143
              Model[2]  66.      ,
              Model[3]  85.4
              Model[4]  72.1667

 PostTests → {Model → {Bonferroni  {{1, 2}, {2, 3}, {1, 4}}
                       Tukey       {{1, 2}, {2, 3}, {1, 4}}}}}
```

Unsere Vermutung wird somit sowohl mit der von Bonferroni als auch mit der von Tukey entwickelten Methode bestätigt. Würde man jedoch als Familien-Signifikanzniveau nicht den Standardwert $\alpha = 0.05$ sondern den Wert $\alpha = 0.07$ verwenden

```
ANOVA[düngemittel[[All, {2, 3}]], PostTests → {Bonferroni, Tukey}, SignificanceLevel → 0.07]
Clear[ertrag, düngemittel]
```

```
{ANOVA → 
              DF    SumOfSq    MeanSq     FRatio     PValue
  Model       3     2585.      861.667    13.1693    0.0000388667
  Error       22    1439.46    65.4301
  Total       25    4024.46                                      ,

 CellMeans → {All       77.5385
              Model[1]  89.7143
              Model[2]  66.      ,
              Model[3]  85.4
              Model[4]  72.1667

 PostTests → {Model → {Bonferroni  {{1, 2}, {2, 3}, {1, 4}}
                       Tukey       {{1, 2}, {2, 3}, {1, 4}, {3, 4}}}}}
```

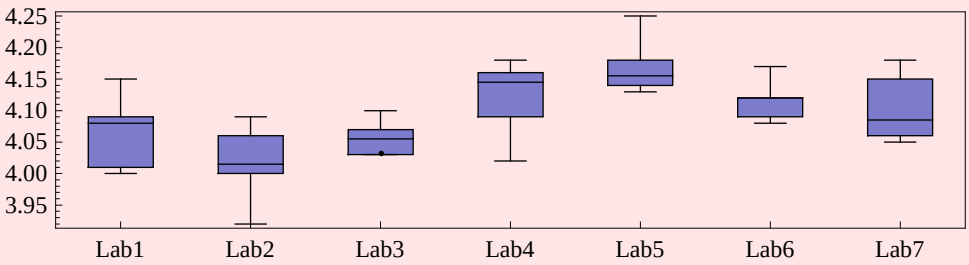
so liefert die von Tukey entwickelte Methode, dass sich auch der mittlere Ernteertrag jener Parzellen, welche mit N + P2O5 gedüngt wurden, deutlich von dem der ungedüngten Zellen unterscheiden.

11.1.10 Beispiel: Aus den Tabletten eines bestimmten Herstellers, welche eine Nominaldosis von 4 mg Wirkstoff enthalten, wurde durch Zermahlen und Vermengen eine homogene Mischung hergestellt. Sieben Laboratorien bekamen jeweils 10 Portionen dieser Mischung, deren Gewicht gleich dem Normgewicht einer Tablette entsprach, und hatten die Aufgabe, die Dosis dieses Wirkstoffes zu ermitteln (vergleiche dazu das Datematerial tabletten). Gibt es einen signifikanten Unterschied zwischen den Mittelwerten der von den einzelnen Laboratorien ermittelten Messwerten? Wenn ja, welche Paare von Laboratorien arbeiten signifikant unterschiedlich?

▼

Lösung: Wir veranschaulichen das Datenmaterial tabletten graphisch durch ein Box-Plot:

```
tabletten = Rest[<< "tablettenfile"];
dosis = Table[Select[tabletten, #1[[1]] == i &][[All, 2]], {i, 1, 7}];
BoxWhiskerPlot[Transpose[dosis],
  BoxLabels → {"Lab1", "Lab2", "Lab3", "Lab4", "Lab5", "Lab6", "Lab7"}]
```



Auf Grund dieses Plots vermuten wir, dass der Faktor Laboratorium einen deutlichen Einfluss auf den Mittelwert der gemessenen Dosis haben dürfte. Wir unterziehen dieses Datenmaterial daher einer Varianzanalyse

```
ANOVA[tabletten, CellMeans → False, PostTests → {Bonferroni, Tukey}]
Clear[dosis, tabletten]
```

```
                    DF    SumOfSq     MeanSq        FRatio    PValue
{ANOVA → Model      6     0.153434    0.0255724     11.902    6.9746×10^-9 , PostTests →
         Error      63    0.13536     0.00214857
         Total      69    0.288794

  {Model → Bonferroni {{2, 4}, {1, 5}, {2, 5}, {3, 5}, {2, 6}, {2, 7}, {5, 7}}
           Tukey      {{2, 4}, {3, 4}, {1, 5}, {2, 5}, {3, 5}, {2, 6}, {2, 7}, {5, 7}}}}
```

und erkennen: Da der *p*-Wert extrem klein ist, hat der Faktor Laboratorium einen sehr deutlichen Einfluss auf den Mittelwert der gemessenen Dosis. Außerdem zeigt sich, dass die Laboratorien {1, 5}, {2, 4}, {2, 5}, {2, 6}, {2, 7}, {3, 5} und {5, 7} jedenfalls signifikant unterschiedlich arbeiten. Nach Tukey arbeiten zusätzlich auch die beiden Laboratorien {3, 4} in deutlich unterschiedlicher Weise.

11.2 Zweifache Varianzanalyse (balanzierte Versuchspläne)

Bei der zweifachen Varianzanalyse hat man es mit zwei Faktoren *S* und *T* zu tun, welche in *s* bzw *t* Stufen auf den Mittelwert von normalverteilten Messwerten einwirken können. Wir treffen dazu die folgenden Annahmen:

■ Falls der Faktor *S* in der *i*-ten Stufe und der Faktor *T* in der *j*-ten Stufe wirkt, so lässt sich das Messergebnis durch eine $\mathcal{N}[\mu_{ij}, \sigma]$-verteilte Zufallsvariable X_{ij} beschreiben. Die beiden Faktoren *S* und *T* wirken also nur auf die Mit-

telwert ein. Die nicht beobachtbaren bzw die zufälligen Einflussgrößen verursachen eine $\mathcal{N}[0, \sigma]$-verteilte Abweichung vom jeweiligen Mittelwert μ_{ij}.

■ Für den Fall, dass der Faktor S in der i-ten und der Faktor T in der j-ten Stufe wirkt, liegen $n_{ij} \geq 2$ Messungen vor, die sich durch die vollständig unabhängigen Zufallsvariablen $X_{ij1}, \ldots, X_{ijn_{ij}}$ beschreiben lassen.

■ Wir beschränken uns in diesem Abschnitt auf den Fall, dass alle n_{ij} gleich n sind, in jeder Zelle also gleich viele Messwerte vorliegen. Man spricht in diesem Fall von einem balanzierten Versuchsplan.

Wir haben es also mit dem a-priori Modell

$$\mathcal{P}_{X_{11}} \times \ldots \times \mathcal{P}_{X_{st}} = \{\{\mathcal{N}[\mu_{11}, \sigma], \ldots, \mathcal{N}[\mu_{st}, \sigma]\} \mid \mu_{ij} \in \mathbb{R},\ \sigma > 0\}$$

zu tun (der erste Index bezieht sich dabei auf jene Stufe, in welcher der Faktor S wirkt, der zweite Index bezieht sich auf jene Stufe, in welcher der Faktor T wirkt), für das geprüft werden soll, ob der Faktor S bzw der Faktor T tatsächlich einen Einfluss auf den Mittelwert der Messungen hat und ob zwischen diesen beiden Faktoren S und T eine Wechselwirkung besteht.

Verwendet man die Reparametrisierung $\mu_{ij} = \mu + \xi_i + \eta_j + \zeta_{ij}$ mit

$$\mu = \frac{1}{st}\sum_{i=1}^{s}\sum_{j=1}^{t}\mu_{ij}; \qquad \xi_i = \frac{1}{t}\sum_{j=1}^{t}\mu_{ij} - \mu; \qquad \eta_j = \frac{1}{s}\sum_{i=1}^{s}\mu_{ij} - \mu; \qquad \zeta_{ij} = \mu_{ij} - \mu - \xi_i - \eta_j$$

wobei offenbar sowohl $\sum_{i=1}^{s}\xi_i = 0$ also auch $\sum_{j=1}^{t}\eta_j = 0$ ist und außerdem für alle $1 \leq i \leq s$ und alle $1 \leq j \leq t$ einerseits $\sum_{j=1}^{t}\zeta_{ij} = 0$ und andererseits $\sum_{i=1}^{s}\zeta_{ij} = 0$ gilt, so wird klar, dass die Größen ξ_i den Einfluss des Faktors S, die Größen η_j den Einfluss des Faktors T und die Größen ζ_{ij} die Wechselwirkung zwischen den beiden Faktoren S und T beschreiben. Damit entsprechen die oben erwähnten Fragestellungen den drei Testproblemen

$\mathcal{H}_0^S$	… alle ξ_i sind 0	gegen	$\mathcal{H}_1^S$ … nicht alle ξ_i sind 0
$\mathcal{H}_0^T$	… alle η_j sind 0	gegen	$\mathcal{H}_1^T$ … nicht alle η_j sind 0
$\mathcal{H}_0^{ST}$	… alle ζ_{ij} sind 0	gegen	$\mathcal{H}_1^{ST}$ … nicht alle ζ_{ij} sind 0

Mit den Statistiken

$$\overline{X}_{ij\blacksquare} = \frac{1}{n}\sum_{k=1}^{n}X_{ijk} \qquad \overline{X}_{i\blacksquare\blacksquare} = \frac{1}{nt}\sum_{j=1}^{t}\sum_{k=1}^{n}X_{ijk}$$

$$\overline{X}_{\blacksquare j\blacksquare} = \frac{1}{ns}\sum_{i=1}^{s}\sum_{k=1}^{n}X_{ijk} \qquad \overline{X}_{\blacksquare\blacksquare\blacksquare} = \frac{1}{nts}\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{n}X_{ijk}$$

(es handelt sich dabei offenbar um erwartungstreue Schätzer für die Parameter μ_{ij}, $\mu + \xi_i$, $\mu + \eta_j$ und μ) führen wir einige für die zweifache Varianzanalyse fundamentale Bezeichnungen ein:

11.2.1 Definition:

a) Die Statistik SSS (**S**um of **S**quares of factor **S**) mit

$$\text{SSS} = t\,n \sum_{i=1}^{s} (\overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,\bullet\,\bullet})^2$$

ist ein sinnvoller Schätzer für den Wert $t\,n \sum_{i=1}^{s} \xi_i^2$. Die Statistik SSS ist somit genau dann klein, wenn der Faktor S keinen Einfluss hat.

b) Die Statistik SST (**S**um of **S**quares of factor **T**) mit

$$\text{SST} = s\,n \sum_{j=1}^{t} (\overline{X}_{\bullet\,j\,\bullet} - \overline{X}_{\bullet\,\bullet\,\bullet})^2$$

ist ein sinnvoller Schätzer für den Wert $s\,n \sum_{j=1}^{t} \eta_j^2$. Die Statistik SST ist somit genau dann klein, wenn der Faktor T keinen Einfluss hat.

c) Die Statistik SS(ST) (**S**um of **S**quares of interaction between factor **S** and **T**) mit

$$\text{SS(ST)} = n \sum_{i=1}^{s} \sum_{j=1}^{t} (\overline{X}_{i\,j\,\bullet} - \overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,j\,\bullet} + \overline{X}_{\bullet\,\bullet\,\bullet})^2$$

ist ein sinnvoller Schätzer für den Wert $n \sum_{i=1}^{s} \sum_{j=1}^{t} \zeta_{i\,j}^2$. Die Statistik SS(ST) ist somit genau dann klein, wenn zwischen den beiden Faktoren S und T keine Wechselwirkung besteht.

d) Die Statistiken SSE bzw SSE' (**S**um of **S**quares of **E**rror) mit

$$\text{SSE} = \sum_{i=1}^{s} \sum_{j=1}^{t} \sum_{k=1}^{n} (X_{i\,j\,k} - \overline{X}_{i\,j\,\bullet})^2$$

bzw

$$\text{SSE}' = \sum_{i=1}^{s} \sum_{j=1}^{t} \sum_{k=1}^{n} (X_{i\,j\,k} - \overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,j\,\bullet} + \overline{X}_{\bullet\,\bullet\,\bullet})^2 = \text{SSE} + \text{SS(ST)}$$

sind sinnvolle Schätzer für den Wert $s\,t\,n\,\sigma^2$. Mit der Statistik SSE lässt sich dieses "klein sein" der Statistiken SSS, SST und SS(ST) in der üblichen Weise relativieren. Im Fall, dass eine Wechselwirkung zwischen den beiden Faktoren S und T prinzipiell unmöglich ist (also alle $\zeta_{i\,j}$ gleich Null sind), lässt sich dieses "klein sein" der Statistiken SSS und SST mit Hilfe der Statistik SSE' relativieren.

e) Die Statistik SStotal (total **S**um of **S**quares)

$$\text{SStotal} = \sum_{i=1}^{s} \sum_{j=1}^{t} \sum_{k=1}^{n} (X_{i\,j\,k} - \overline{X}_{\bullet\,\bullet\,\bullet})^2$$

entspricht der Summe der quadratischen Abweichungen der einzelnen Messwerte vom Gesamtmittelwert. Dabei gilt wieder die als Zerlegung der Varianz bekannte Formel

$$\text{SStotal} = \text{SSS} + \text{SST} + \text{SS(ST)} + \text{SSE} = \text{SSS} + \text{SST} + \text{SSE}'$$

Analog zu Satz 11.1.2 lassen sich die folgenden, für die zweifache Varianzanalyse zentralen Sätze beweisen (man beachte, dass diese Sätze nur für balanzierte Versuchspläne gelten):

11.2.2 Satz: Für die Verteilung der Testgrößen der zweifachen Varianzanalyse **mit** Wechselwirkung (in diesem Fall lassen sich die μ_{ij} in der Form $\mu_{ij} = \mu + \xi_i + \eta_j + \zeta_{ij}$ reparametrisieren) gilt:

a) Falls die Hypothese $\mathcal{H}_0^S$ zutrifft, also alle ξ_i gleich 0 sind, so genügt die Testgröße

$$\frac{s\,t\,(n-1)}{s-1}\,\frac{\mathrm{SSS}}{\mathrm{SSE}} = \frac{s\,t\,(n-1)}{s-1}\,\frac{t\,n\sum_{i=1}^{s}(\overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,\bullet\,\bullet})^2}{\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{n}(X_{i\,j\,k} - \overline{X}_{i\,j\,\bullet})^2}$$

einer Fisher F Verteilung mit den Parametern $s-1$ und $s\,t\,(n-1)$.

b) Falls die Hypothese $\mathcal{H}_0^T$ zutrifft, also alle η_j gleich 0 sind, so genügt die Testgröße

$$\frac{s\,t\,(n-1)}{t-1}\,\frac{\mathrm{SST}}{\mathrm{SSE}} = \frac{s\,t\,(n-1)}{t-1}\,\frac{s\,n\sum_{j=1}^{t}(\overline{X}_{\bullet\,j\,\bullet} - \overline{X}_{\bullet\,\bullet\,\bullet})^2}{\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{n}(X_{i\,j\,k} - \overline{X}_{i\,j\,\bullet})^2}$$

einer Fischer F Verteilung mit den Parametern $t-1$ und $s\,t\,(n-1)$.

c) Falls die Hypothese $\mathcal{H}_0^{ST}$ zutrifft, also alle ζ_{ij} gleich 0 sind, so genügt die Testgröße

$$\frac{s\,t\,(n-1)}{(s-1)\,(t-1)}\,\frac{\mathrm{SS(ST)}}{\mathrm{SSE}} = \frac{s\,t\,(n-1)}{(s-1)\,(t-1)}\,\frac{n\sum_{i=1}^{s}\sum_{j=1}^{t}(\overline{X}_{i\,j\,\bullet} - \overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,j\,\bullet} + \overline{X}_{\bullet\,\bullet\,\bullet})^2}{\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{n}(X_{i\,j\,k} - \overline{X}_{i\,j\,\bullet})^2}$$

einer Fischer F Verteilung mit den Parametern $(s-1)\,(t-1)$ und $s\,t\,(n-1)$.

11.2.3 Satz: Für die Verteilung der Testgrößen der zweifachen Varianzanalyse **ohne** Wechselwirkung (in diesem Fall lassen sich die μ_{ij} in der Form $\mu_{ij} = \mu + \xi_i + \eta_j$ reparametrisieren) gilt:

a) Falls die Hypothese $\mathcal{H}_0^S$ zutrifft, also alle ξ_i gleich 0 sind, so genügt die Testgröße

$$\frac{s\,t\,n-s-t+1}{s-1}\,\frac{\mathrm{SSS}}{\mathrm{SSE'}} = \frac{s\,t\,n-s-t+1}{s-1}\,\frac{t\,n\sum_{i=1}^{s}(\overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,\bullet\,\bullet})^2}{\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{n}(\overline{X}_{i\,j\,k} - \overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,j\,\bullet} + \overline{X}_{\bullet\,\bullet\,\bullet})^2}$$

einer Fisher F Verteilung mit den Parametern $s-1$ und $s\,t\,n-s-t+1$.

b) Falls die Hypothese $\mathcal{H}_0^T$ zutrifft, also alle η_j gleich 0 sind, so genügt die Testgröße

$$\frac{s\,t\,n-s-t+1}{t-1}\,\frac{\mathrm{SST}}{\mathrm{SSE'}} = \frac{s\,t\,n-s-t+1}{t-1}\,\frac{s\,n\sum_{j=1}^{t}(\overline{X}_{\bullet\,j\,\bullet} - \overline{X}_{\bullet\,\bullet\,\bullet})^2}{\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{n}(\overline{X}_{i\,j\,k} - \overline{X}_{i\,\bullet\,\bullet} - \overline{X}_{\bullet\,j\,\bullet} + \overline{X}_{\bullet\,\bullet\,\bullet})^2}$$

einer Fischer F Verteilung mit den Parametern $s-1$ und $s\,t\,n-s-t+1$.

Damit ist offensichtlich, wie sich prüfen lässt, ob die Faktoren *S* bzw *T* tatsächlich einen Einfluss auf den Mittelwert der normalverteilten Messwerte besitzen bzw ob zwischen diesen beiden Faktoren eine Wechselwirkung besteht:

11.2.4 Die zweifache Varianzanalyse mit Wechselwirkung wird verwendet, wenn geprüft werden soll, ob die Faktoren S bzw T, welche in s bzw t Stufen wirken können, tatsächlich einen Einfluss auf den Mittelwert von normalverteilten Messwerten besitzen und ob zwischen diesen beiden Faktoren S und T eine Wechselwirkung besteht (dabei geht man vom Modell $\mu_{ij} = \mu + \xi_i + \eta_j + \zeta_{ij}$ aus):

$\mathcal{P}_{X_{11}} \times \ldots \times \mathcal{P}_{X_{st}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\{\mathcal{N}[\mu_{ij}, \sigma]\}\,\vert\, \mu_{ij} \in \mathbb{R}, \sigma > 0\}$	alle ξ_i sind 0, $\sigma > 0$	nicht alle ξ_i sind 0, $\sigma > 0$	$\frac{st(n-1)}{s-1}\,\frac{\mathrm{SSS}}{\mathrm{SSE}} > f_{s-1,st(n-1);1-\alpha}$
	alle η_j sind 0, $\sigma > 0$	nicht alle η_j sind 0, $\sigma > 0$	$\frac{st(n-1)}{t-1}\,\frac{\mathrm{SST}}{\mathrm{SSE}} > f_{t-1,st(n-1);1-\alpha}$
	alle ζ_{ij} sind 0, $\sigma > 0$	nicht alle ζ_{ij} sind 0, $\sigma > 0$	$\frac{st(n-1)}{(s-1)(t-1)}\,\frac{\mathrm{SS(ST)}}{\mathrm{SSE}} > f_{(s-1)(t-1),st(n-1);1-\alpha}$

Dabei bezeichnet $f_{n,m;q}$ das q-Quantil der $\mathcal{F}[n, m]$-Verteilung.

11.2.5 Die zweifache Varianzanalyse ohne Wechselwirkung wird verwendet, wenn geprüft werden soll, ob die Faktoren S bzw T, welche in s bzw t Stufen wirken können, tatsächlich einen Einfluss auf den Mittelwert von normalverteilten Messwerten besitzen, wobei vorausgesetzt wird, dass zwischen diesen beiden Faktoren S und T prinzipiell keine Wechselwirkung besteht (man geht also vom Modell $\mu_{ij} = \mu + \xi_i + \eta_j$ aus):

$\mathcal{P}_{X_{11}} \times \ldots \times \mathcal{P}_{X_{st}}$	$\mathcal{H}_0$	$\mathcal{H}_1$	Ablehnungsbereich
$\{\{\mathcal{N}[\mu_{ij}, \sigma]\}\,\vert\, \mu_{ij} \in \mathbb{R},\ \sigma > 0\}$	alle ξ_i sind 0, $\sigma > 0$	nicht alle ξ_i sind 0, $\sigma > 0$	$\frac{stn-s-t+1}{s-1}\,\frac{\mathrm{SSS}}{\mathrm{SSE'}} > f_{s-1,stn-s-t+1;1-\alpha}$
	alle η_j sind 0, $\sigma > 0$	nicht alle η_j sind 0, $\sigma > 0$	$\frac{stn-s-t+1}{t-1}\,\frac{\mathrm{SST}}{\mathrm{SSE'}} > f_{t-1,stn-s-t+1;1-\alpha}$

Dabei bezeichnet $f_{n,m;q}$ das q-Quantil der $\mathcal{F}[n, m]$-Verteilung.

Wesentlich für ein richtiges Verständnis der zweifachen Varianzanalyse ist folgende Bemerkung:

11.2.6 Bemerkung: Die Frage, ob der Faktor S tatsächlich einen Einfluss auf den Mittelwert von normalverteilten Messwerten hat, lässt sich überprüfen

a) mit einer **einfachen** Varianzanalyse (man berücksichtigt den Faktor T überhaupt nicht);

b) mit einer **zweifachen** Varianzanalyse **ohne** Wechselwirkung (man berücksichtigt zwar den Faktor T, geht aber davon aus, dass eine Wechselwirkung zwischen den beiden Faktoren S und T prinzipiell unmöglich ist);

c) mit einer **zweifachen** Varianzanalyse **mit** Wechselwirkung (man berücksichtigt sowohl den Faktor T als auch eine mögliche Wechselwirkung zwischen den beiden Faktoren S und T).

In allen drei Fällen hat die Statistik SSS den gleichen Wert. Der Unterschied dieser drei Modelle besteht in der Verwendung von unterschiedlichen Schätzern für den Wert $stn\sigma^2$. Während man dafür im ersten Fall die Statistik SSE + SS(ST) + SST verwendet (diese Statistik entspricht der Statistik SSE der einfachen Varianzanalyse), wird im zweiten Fall die Statistik SSE' = SSE + SS(ST) und im dritten Fall die Statistik SSE verwendet. Das Ergebnis der Varianzanalyse hängt damit vom gewählten Modell wesentlich ab und kann möglicherweise zu verschiedenen Entscheidungen führen!

Die zweifache (mehrfache) Varianzanalyse lässt sich ebenfalls mit dem Befehl ANOVA behandeln. Man hat dabei zusätzlich zu dem zu analysierenden Datenmaterial *daten* auch das verwendete Modell *model* sowie die zu untersuchenden Faktoren *factors* einzugeben (man beachte, dass dieser Befehl auch auf nicht-balanzierte Stichprobenpläne angewendet werden kann; wie dabei die Ergebnisse zu interpretieren sind, werden wir später behandeln):

```
<< Statistics`ANOVA`
```

ladet das Paket ANOVA`.

■ ANOVA[*daten, model, factors*]

führt für das Datenmaterial *daten* eine zweifache Varianzanalyse durch. Das Datenmaterial *daten* muss dabei die Form $\{\{s_1, t_1, x_1\}, \{s_2, t_2, x_2\}, \ldots\}$ besitzen, wobei die Einträge s_i bzw t_i die zum i-ten Messwert x_i gehörenden Werte der Faktoren S bzw T bezeichnen.

Falls zwischen den beiden Faktoren S und T eine mögliche Wechselwirkung berücksichtigt werden soll, so hat die Liste *model* die Form $\{S, T, S\,T\}$; besteht zwischen diesen beiden Faktoren prinzipiell keine Wechselwirkung, so hat die Liste *model* die Form $\{S, T\}$; möchte man den Einfluss des Faktors S mit einer einfachen Varianzanalyse prüfen, so hat die Liste model die Form $\{S\}$.

In der Liste *factors* werden die zu untersuchenden Faktoren S und T aufgelistet; die Liste *factors* hat bei der zweifachen Varianzanalyse somit die Form $\{S, T\}$.

Der von *Mathematica* gelieferte Output hat wieder eine leicht verständliche Gestalt: Neben den Freiheitsgraden (DF) und den Schätzwerten der Statistiken SSS, SST, … (SumOfSq) sowie $\mathrm{SSS}/(s-1)$, $\mathrm{SST}/(t-1)$, … (MeanSq) werden die entsprechenden Testgrößen (FRatio) mit den zugehörigen p-Werte (PValue) ausgegeben.

Außerdem werden die Schätzwerte $x_{\blacksquare\blacksquare\blacksquare}$, $x_{i\blacksquare\blacksquare}$, $x_{\blacksquare j\blacksquare}$, $x_{i j\blacksquare}$ von μ, $\mu+\xi_i$, $\mu+\eta_j$, $\mu_{i j}$ angeführt. Sollen diese Schätzwerte nicht ausgegeben werden, so verwende man die Option CellMeans → False.

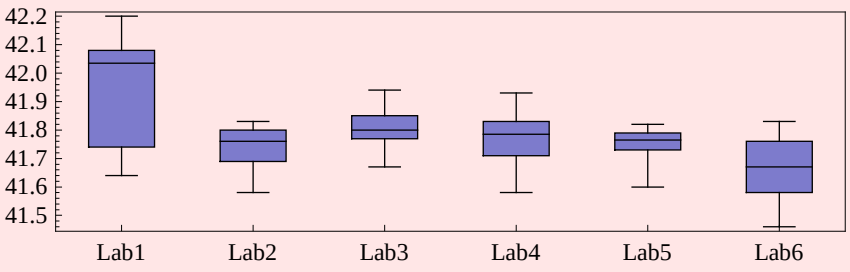
Wir demonstrieren die zweifache Varianzanalyse an zwei konkreten Beispielen:

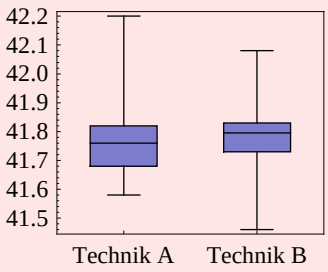
11.2.7 Beispiel: In sechs Laboratorien wurde der Fettgehalt von getrockneten Eiern unter Verwendung von zwei verschiedenen Techniken jeweils vier mal ermittelt (bei unserem Datenmaterial handelt es sich also um einen balanzierten Versuchsplan). Die dabei erzielten Messergebnisse finden sich im Datenmaterial fettgehalt. Man prüfe, ob der Faktor "Labor" bzw der Faktor "Technik" den Mittelwert des gemessenen Fettgehalts signifikant beeinflusst und ob zwischen diesen beiden Faktoren eine Wechselwirkung besteht.

▼

Lösung: Wir veranschaulichen den Einfluss der Faktoren "Labor" bzw "Technik" graphisch durch Box-Plots:

```
fettgehalt = Rest[<< "fettgehaltfile"];
labor = Table[Select[fettgehalt, #1[[1]] == i &][[All, 3]], {i, 1, 6}];
technik = Table[Select[fettgehalt, #1[[2]] == i &][[All, 3]], {i, 1, 2}];
BoxWhiskerPlot[Transpose[labor], BoxLabels → {"Lab1", "Lab2", "Lab3", "Lab4", "Lab5", "Lab6"}, ImageSize
BoxWhiskerPlot[Transpose[technik], BoxLabels → {"Technik A", "Technik B"}, ImageSize → {120, 100}]
Clear[labor, technik]
```





An Hand dieser Graphiken erkennt man, dass der Faktor "Labor" offensichtlich einen deutlichen Einfluss auf den ermittelten Fettgehalt haben dürfte, während der Einfluss des Faktors "Technik" aber eher ist. Diese Vermutung

soll nun durch eine Varianzanalyse überprüft werden. (Man beachte, dass die Meßwerte des Fettgehalts in den einzelnen Laboratorien offenbar in stark unterschiedlicher Weise streuen und damit die Voraussetzungen der Varianzanalyse verletzt sind. Trotzdem darf in dieser Situation die Varianzanalyse verwendet werden, da es sich bei der Varianzanalyse um ein sehr robustes Verfahren handelt.)

a) Wir prüfen zuerst mit Hilfe einer **einfachen** Varianzanalyse, ob der Faktor "Labor" bzw der Faktor "Technik" einen Einfluss auf den ermittelten Fettgehalt hat:

```
ANOVA[fettgehalt, {Labor}, {Labor, Technik}, CellMeans → False]
```

```
ANOVA →
          DF   SumOfSq     MeanSq       FRatio     PValue
Labor     5    0.443025    0.088605     6.41539    0.000164084
Error     42   0.580075    0.0138113
Total     47   1.0231
```

```
ANOVA[fettgehalt, {Technik}, {Labor, Technik}, CellMeans → False]
```

```
ANOVA →
          DF   SumOfSq        MeanSq         FRatio       PValue
Technik   1    0.000408333    0.000408333    0.0183666    0.892789
Error     46   1.02269        0.0222324
Total     47   1.0231
```

Im ersten Fall ist der *p*-Wert sehr klein; der Faktor "Labor" hat somit einen sehr signifikanten Einfluss auf den ermittelten Fettgehalt. Im zweiten Fall ist der *p*-Wert jedoch groß; der Faktor "Technik" hat somit keinen signifikanten Einfluss auf den ermittelten Fettgehalt.

b) Nun prüfen wir mit Hilfe einer **zweifachen** Varianzanalyse **ohne** Wechselwirkung, ob der Faktor "Labor" bzw der Faktor "Technik" einen Einfluss auf den ermittelten Fettgehalt hat:

```
ANOVA[fettgehalt, {Labor, Technik}, {Labor, Technik}, CellMeans → False]
```

```
ANOVA →
          DF   SumOfSq        MeanSq         FRatio       PValue
Labor     5    0.443025       0.088605       6.26706      0.000209827
Technik   1    0.000408333    0.000408333    0.0288815    0.865889
Error     41   0.579667       0.0141382
Total     47   1.0231
```

Während der *p*-Wert für den Faktor "Labor" wieder sehr klein ist, der Faktor "Labor" also auf den ermittelten Fettgehalt einen sehr signifikanten Einfluss hat, ist der *p*-Wert für den Faktor "Technik" wieder groß, also besitzt der Faktor "Technik" keinen signifikanten Einfluss auf den ermittelten Fettgehalt.

c) Schließlich prüfen wir mit Hilfe einer **zweifachen** Varianzanalyse **mit** Wechselwirkung, ob der Faktor "Labor" bzw der Faktor "Technik" einen Einfluss auf den ermittelten Fettgehalt hat und ob zwischen diesen beiden Faktoren eine Wechselwirkung besteht:

```
ANOVA[fettgehalt, {Labor, Technik, Labor Technik}, {Labor, Technik}, CellMeans → False]
Clear[fettgehalt]
```

```
ANOVA →
                 DF   SumOfSq        MeanSq         FRatio      PValue
Labor            5    0.443025       0.088605       5.81016     0.000494735
Technik          1    0.000408333    0.000408333    0.026776    0.870935
Labor Technik    5    0.0306667      0.00613333     0.402186    0.844049
Error            36   0.549          0.01525
Total            47   1.0231
```

Der *p*-Wert für den Faktor "Labor" ist wieder sehr klein; der Faktor "Labor" beeinflusst somit deutlich den ermittelten Fettgehalt. Die beiden restlichen *p*-Werte sind hingegen groß; der Faktor "Technik" besitzt damit keinen Einfluss auf den ermittelten Fettgehalt, außerdem besteht zwischen den beiden Faktoren keine Wechselwirkung.

Man beachte, dass in allen drei Fällen der SSS-Wert (0.443025) für den Faktor "Labor" gleich ist und sich die als Schätzwerte für $s\,t\,n\,\sigma^2$ verwendeten Werte SSE + SS(ST) + SST bzw SSE + SS(ST) bzw SSE mit 0.580075 bzw

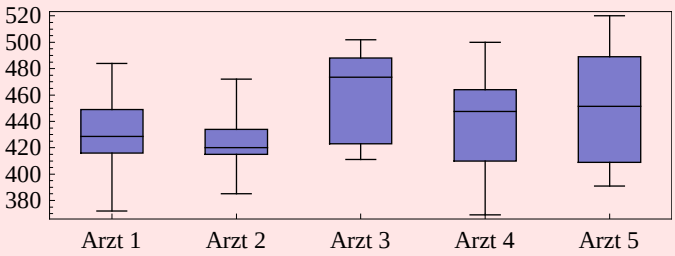
0.579667 bzw 0.549 nur geringfügig voneinander unterscheiden. Man erhält somit in allen drei Fällen annähernd den gleichen *p*-Wert und gelangt somit zur gleichen Entscheidung.

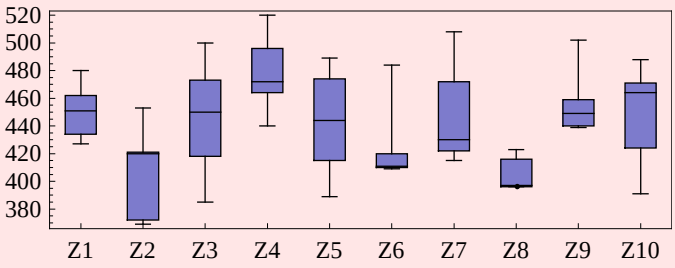
11.2.8 Beispiel: Fünf Ärzten wurde die Aufgabe gestellt, das Blut eines Patienten an jeder der 10 Zähleinrichtungen einer Klinik auf die Anzahl der roten Blutkörperchen hin zu untersuchen (man vergleiche dazu das Datenmaterial blutkörperchen). Es soll geprüft werden, ob die Ärzte signifikant unterschiedlich zählen bzw ob die Zähleinrichtungen signifikant unterschiedlich arbeiten.

▼

Lösung: Wir veranschaulichen den Einfluss der Faktoren "Arzt" bzw "Zähleinrichtung" graphisch durch Box-Plots:

```
blutkörperchen = Rest[<< "blutkörperchenfile"];
arzt = Table[Select[blutkörperchen, #1[[1]] == i &][[All, 3]], {i, 1, 5}];
zähl = Table[Select[blutkörperchen, #1[[2]] == i &][[All, 3]], {i, 1, 10}];
BoxWhiskerPlot[Transpose[arzt], BoxLabels → {"Arzt 1", "Arzt 2", "Arzt 3", "Arzt 4", "Arzt 5"},
  ImageSize → {250, 100}]
BoxWhiskerPlot[Transpose[zähl],
  BoxLabels → {"Z1", "Z2", "Z3", "Z4", "Z5", "Z6", "Z7", "Z8", "Z9", "Z10"}, AspectRatio → 0.35,
  ImageSize → {250, 100}]
Clear[arzt, zähl]
```





An Hand dieser Graphiken vermuten wir, dass beide Faktoren einen signifikanten Einfluss auf das Zählergebnis haben dürften. Diese Vermutung soll wieder durch eine Varianzanalyse überprüft werden:

a) Dazu prüfen wir zuerst mit Hilfe einer **einfachen** Varianzanalyse, ob der Faktor "Arzt" bzw der Faktor "Zähleinrichtung" einen signifikanten Einfluss auf das Zählergebnis hat:

```
ANOVA[blutkörperchen, {Arzt}, {Arzt, Zähleinrichtung}, CellMeans → False]
```

```
ANOVA→
          DF    SumOfSq     MeanSq     FRatio     PValue
Arzt      4     11609.5     2902.37    2.30444    0.0728826
Error     45    56676.2     1259.47
Total     49    68285.7
```

```
ANOVA[blutkörperchen, {Zähleinrichtung}, {Arzt, Zähleinrichtung}, CellMeans → False]
```

		DF	SumOfSq	MeanSq	FRatio	PValue
ANOVA →	Zähleinrichtung	9	22 604.9	2511.65	2.19931	0.0426526
	Error	40	45 680.8	1142.02		
	Total	49	68 285.7			

Während der Faktor "Arzt" keinen signifikanten Einfluss auf das Zählergebnis hat ($p = 0.0728826$), beeinflusst der Faktor "Zähleinrichtung" das Zählergebnis in ziemlich signifikanter Weise ($p = 0.0426526$).

b) Nun prüfen wir mit Hilfe einer **zweifachen** Varianzanalyse **ohne** Wechselwirkung, ob der Faktor "Arzt" bzw der Faktor "Zähleinrichtung" einen Einfluss auf das Zählergebnis hat (da in jeder Zelle lediglich ein einziger Messwert vorhanden ist, lässt sich der Wert SSE nicht ermitteln; wir müssen deshalb annehmen, dass eine Wechselwirkung zwischen diesen beiden Faktoren prinzipiell unmöglich ist):

```
ANOVA[blutkörperchen, {Arzt, Zähleinrichtung}, {Arzt, Zähleinrichtung}, CellMeans → False]
Clear[blutkörperchen]
```

		DF	SumOfSq	MeanSq	FRatio	PValue
	Arzt	4	11 609.5	2902.37	3.06666	0.0284451
ANOVA →	Zähleinrichtung	9	22 604.9	2511.65	2.65383	0.0180066
	Error	36	34 071.3	946.426		
	Total	49	68 285.7			

An Hand der p-Werte erkennt man, dass nun beide Faktoren einen signifikanten Einfluss auf das Zählergebnis haben. Der Unterschied hinsichtlich des Faktors "Arzt" lässt sich dabei folgendermaßen erklären: Während in beiden Fällen der SSS-Wert (11 609.5) für den Faktor "Arzt" gleich ist wird im ersten Fall (einfache Varianzanalyse - keine Berücksichtigung des Faktors "Zähleinrichtung") als Schätzwert für $s\,t\,\sigma^2$ der Wert SSE (56 676.2) verwendet und im zweiten Fall (zweifache Varianzanalyse - Berücksichtigung des Faktors "Zähleinrichtung") als Schätzwert für $s\,t\,\sigma^2$ der Wert SSE' (34 071.3) herangezogen.

11.3 Zweifache Varianzanalyse (allgemeine Versuchspläne)

Wir befassen uns abschließend noch kurz mit nicht-balanzierten Versuchplänen, also jenen Fällen, in denen nicht alle $n_{i\,j}$ gleich sind und verwenden die naheliegenden Bezeichnungen

$$n_{\bullet\,\bullet} = \sum_{i=1}^{s}\sum_{j=1}^{t} n_{i\,j} \qquad \text{bzw} \qquad n_{i\,\bullet} = \sum_{j=1}^{t} n_{i\,j} \qquad \text{bzw} \qquad n_{\bullet\,j} = \sum_{i=1}^{s} n_{i\,j}$$

Im Falle eines nicht-balanzierten Versuchsplan verwendet man die Reparametrisierung $\mu_{i\,j} = \mu + \xi_i + \eta_j + \zeta_{i\,j}$ mit

$$\mu = \frac{1}{n_{\bullet\,\bullet}} \sum_{i=1}^{s}\sum_{j=1}^{t} n_{i\,j}\,\mu_{i\,j}; \qquad \xi_i = \frac{1}{n_{i\,\bullet}} \sum_{j=1}^{t} n_{i\,j}\,\mu_{i\,j} - \mu;$$

$$\eta_j = \frac{1}{n_{\bullet\,j}} \sum_{i=1}^{s} n_{i\,j}\,\mu_{i\,j} - \mu; \qquad \zeta_{i\,j} = \mu_{i\,j} - \mu - \xi_i - \eta_j$$

wobei offenbar sowohl $\sum_{i=1}^{s} n_{i\,\bullet}\,\xi_i = 0$ also auch $\sum_{j=1}^{t} n_{\bullet\,j}\,\eta_j = 0$ ist und außerdem für alle $1 \le i \le s$ und alle $1 \le j \le t$ einerseits $\sum_{j=1}^{t} n_{i\,j}\,\zeta_{i\,j} = 0$ und andererseits $\sum_{i=1}^{s} n_{i\,j}\,\zeta_{i\,j} = 0$ gilt. Damit beschreiben die Größen ξ_i wieder den Einfluss des Faktors S, die Größen η_j den Einfluss des Faktors T und die Größen $\zeta_{i\,j}$ die Wechselwirkung zwischen den beiden Faktoren S und T.

Auch wenn die Formeln für $\overline{X}_{i\,\bullet\,\bullet}$, $\overline{X}_{\bullet\,j\,\bullet}$, $\overline{X}_{i\,j\,\bullet}$, $\overline{X}_{\bullet\,\bullet\,\bullet}$ und damit für SSS, SST, SS(ST), SSE und SStot entsprechend modifiziert werden (man also den Index k jeweils von 1 bis $n_{i\,j}$ laufen lässt und man an Stelle von $n\,s\,t$ bzw $n\,t$ bzw $n\,s$ die Größen $n_{\bullet\,\bullet}$ bzw $n_{i\,\bullet}$ bzw $n_{\bullet\,j}$ verwendet), gilt nun aber die Zerlegung der Varianz

$$\text{SStotal} = \text{SSS} + \text{SST} + \text{SS(ST)} + \text{SSE} = \text{SSS} + \text{SST} + \text{SSE}'$$

nicht! Für die Behandlung nicht-balanzierter Stichprobenpläne benötigt man daher andere Methoden.

Wir gehen kurz auf jene Methode ein, welche die Varianzanalyse als Design-Modell der Regressionsanalyse auffasst und dabei die im Rahmen der Regressionsanalyse hergeleiteten Formeln verwendet. Mit dieser Methode kann gezeigt werden, dass sich die Totalvarianz SStot stets in der Form

$$\text{SStotal} = \text{SSS} + {}^{S}\text{SST}_{\text{reg}} + {}^{ST}\text{SS}\left(\text{ST}\right)_{\text{reg}} + \text{SSE}_{\text{reg}} = \text{SSS} + {}^{S}\text{SST}_{\text{reg}} + \text{SSE}'_{\text{reg}}$$

darstellen lässt (weil dabei die Reihenfolge der Faktoren *S* und *T* wesentlich ist, spricht man in diesem Zusammenhang von einer sequentielle Zerlegung der Varianz). Mit der Statistik SSS wird dabei wie bisher der Einfluss des Faktors *S* beschrieben. Die Statistik ${}^{S}\text{SST}_{\text{reg}}$ beschreibt den Einfluss des Faktors *T*, wobei der Einfluss des Faktors *S* bereits berücksichtigt ist. Die Statistik ${}^{ST}\text{SS}\left(\text{ST}\right)_{\text{reg}}$ beschreibt den Einfluss der Wechselwirkung zwischen den beiden Faktoren *S* und *T*, wobei die individuellen Einflüsse dieser beiden Faktoren bereits berücksichtigt sind. Die Statistiken SSE_{reg} bzw SSE'_{reg} werden wieder dazu verwendet, den unbekannten Wert $n_{\bullet\bullet}\sigma^2$ zu schätzen. Eine genaue Definition dieser Statistiken findet sich in Bemerkung 12.6.2. Falls es sich beim zu untersuchenden Datenmaterial um einen balanzierten Versuchsplan handelt, so stimmen diese modifizierten Statistiken mit den ursprünglichen Statistiken überein!

Verwendet man nun an Stelle der Statistiken SST, SS(ST) und SSE bzw SSE' die Statistiken ${}^{S}\text{SST}_{\text{reg}}$, ${}^{ST}\text{SS}\left(\text{ST}\right)_{\text{reg}}$ und SSE_{reg} bzw SSE'_{reg} so kann mit den in 11.2.4 bzw 11.2.5 angeführten Tests geprüft werden, ob

- der Faktor *S* einen Einfluss auf die Mittelwerte besitzt;
- der Faktor *T* einen Einfluss auf die Mittelwerte besitzt, wenn der Einfluss des Faktors *S* bereits berücksichtigt ist;
- zwischen den beiden Faktoren *S* und *T* eine Wechselwirkung besteht, wenn die individuellen Einflüsse dieser beiden Faktoren bereits berücksichtigt sind.

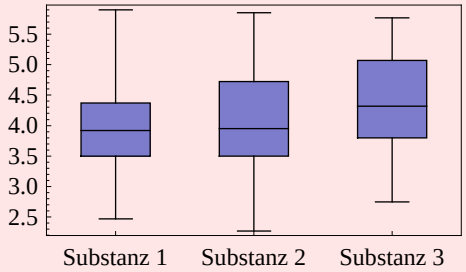
Wir erläutern die zweifache (mehrfache) Varianzanalyse für nicht-balanzierte Stichprobenplänen an einigen Beispielen. Man beachte dabei, dass der Befehl ANOVA sowohl für balanzierte Stichprobenpläne als auch für nicht-balanzierte Stichprobenpläne verwendet werden kann:

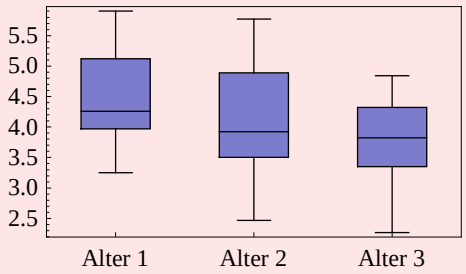
11.3.1 Beispiel: Um die Schädigung der Lunge durch toxische Substanzen zu ermitteln, wurde von insgesamt 98 zufällig ausgewählten Arbeitern, welche ständig einer von $s = 3$ toxischen Substanzen ausgesetzt waren und $t = 3$ möglichen Altersklassen angehören, das Atemvolumen ermittelt (man vergleiche dazu das Datenmaterial gesundheit). Es soll geprüft werden, ob die Art der toxischen Substanz bzw die Altersklasse des Arbeiters einen Einfluss auf sein Atemvolumen hat und ob zwischen diesen beiden Faktoren hinsichtlich einer Schädigung der Lunge (geringeres Atemvolumen) eine Wechselwirkung besteht.

▼

Lösung: Wir veranschaulichen die möglichen Einflüsse der beiden Faktoren "Substanz" und "Altersklasse" auf das Atemvolumen graphisch durch Box-Plots:

```
gesundheit = Rest[<< "gesundheitfile"];
sub1 = Select[gesundheit, #1[[1]] == 1 &][[All, 3]];
sub2 = Select[gesundheit, #1[[1]] == 2 &][[All, 3]];
sub3 = Select[gesundheit, #1[[1]] == 3 &][[All, 3]];
BoxWhiskerPlot[sub1, sub2, sub3, BoxLabels → {"Substanz 1", "Substanz 2", "Substanz 3"},
  ImageSize → {170, 100}]
alt1 = Select[gesundheit, #1[[2]] == 1 &][[All, 3]];
alt2 = Select[gesundheit, #1[[2]] == 2 &][[All, 3]];
alt3 = Select[gesundheit, #1[[2]] == 3 &][[All, 3]];
BoxWhiskerPlot[alt1, alt2, alt3, BoxLabels → {"Alter 1", "Alter 2", "Alter 3"}, ImageSize → {170, 100}]
Clear[sub1, sub2, sub3, alt1, alt2, alt3]
```





An Hand dieser Graphiken vermutet man, dass sowohl der Faktor "Substanz" als auch der Faktor "Altersklasse" einen Einfluss auf das Atemvolumen haben könnten. Diese Vermutung soll nun durch eine Varianzanalyse überprüft werden:

a) Wir prüfen zuerst mit einer **einfachen** Varianzanalyse, ob die Faktoren "Substanz" bzw "Altersklasse" einen Einfluss auf das Atemvolumen haben:

```
ANOVA[gesundheit, {Substanz}, {Substanz, Altersklasse}, CellMeans → False]
ANOVA[gesundheit, {Altersklasse}, {Substanz, Altersklasse}, CellMeans → False]
```

```
ANOVA→
                DF    SumOfSq    MeanSq      FRatio     PValue
Substanz        2     2.39113    1.19557     1.99077    0.142249
Error           95    57.0529    0.600556
Total           97    59.444
```

```
ANOVA→
                DF    SumOfSq    MeanSq      FRatio     PValue
Altersklasse    2     7.13512    3.56756     6.47917    0.00230251
Error           95    52.3089    0.55062
Total           97    59.444
```

Der Faktor "Substanz" hat damit praktisch keinen Einfluss auf das Atemvolumen, der Faktor "Altersklasse" beeinflusst das Atemvolumen hingegen sehr signifikant.

b) Wir prüfen nun mit einer **zweifachen** Varianzanalyse **ohne** Wechselwirkung, ob die beiden Faktoren "Substanz" bzw "Altersklasse" einen Einfluss auf das Atemvolumen haben (da ein nicht-balanzierten Stichprobenplan vorliegt, spielt die Reihenfolge der Faktoren eine Rolle):

```
ANOVA[gesundheit, {Substanz, Altersklasse}, {Substanz, Altersklasse}, CellMeans → False]
ANOVA[gesundheit, {Altersklasse, Substanz}, {Substanz, Altersklasse}, CellMeans → False]
```

		DF	SumOfSq	MeanSq	FRatio	PValue
ANOVA →	Substanz	2	2.39113	1.19557	2.2189	0.114453
	Altersklasse	2	6.94353	3.47176	6.44339	0.00239506
	Error	93	50.1093	0.53881		
	Total	97	59.444			

		DF	SumOfSq	MeanSq	FRatio	PValue
ANOVA →	Altersklasse	2	7.13512	3.56756	6.62118	0.00204934
	Substanz	2	2.19954	1.09977	2.04111	0.135663
	Error	93	50.1093	0.53881		
	Total	97	59.444			

Inhaltlich hat sich das Ergebnis nicht verändert: Während der Faktor "Altersklasse" nach wie vor in beiden Fällen einen sehr signifikanten Einfluss auf das Atemvolumen hat, beeinflusst der Faktor "Substanz" das Atemvolumen in beiden Fällen nicht signifikant. Man erkennt aber auch, dass sich sowohl die Größe ${}^{S}\mathrm{SST}_{\mathrm{reg}}$ (6.94353) von der Größe SST (7.13512) als auch die Größe ${}^{T}\mathrm{SSS}_{\mathrm{reg}}$ (2.19954) von der Größe SSS (2.39113) unterscheidet.

c) Wir prüfen nun mit einer **zweifachen** Varianzanalyse **mit** Wechselwirkung, ob die beiden Faktoren "Substanz" bzw "Altersklasse" einen Einfluss auf das Atemvolumen haben und ob zwischen diesen beiden Faktoren eine Wechselwirkung besteht (man beachte wieder, dass wir es mit einem nicht-balanzierten Stichprobenplan zu tun haben und daher die Reihenfolge der Faktoren eine Rolle spielt):

```
ANOVA[gesundheit, {Substanz, Altersklasse, Substanz Altersklasse}, {Substanz, Altersklasse},
  CellMeans → False]
ANOVA[gesundheit, {Altersklasse, Substanz, Substanz Altersklasse}, {Substanz, Altersklasse},
  CellMeans → False]
```

			DF	SumOfSq	MeanSq	FRatio	PValue
ANOVA →	Substanz		2	2.39113	1.19557	3.65378	0.0298516
	Altersklasse		2	6.94353	3.47176	10.6101	0.0000736611
	Altersklasse	Substanz	4	20.9873	5.24683	16.0349	6.37621×10^{-10}
	Error		89	29.122	0.327214		
	Total		97	59.444			

			DF	SumOfSq	MeanSq	FRatio	PValue
ANOVA →	Altersklasse		2	7.13512	3.56756	10.9028	0.0000581893
	Substanz		2	2.19954	1.09977	3.36102	0.0391583
	Altersklasse	Substanz	4	20.9873	5.24683	16.0349	6.37621×10^{-10}
	Error		89	29.122	0.327214		
	Total		97	59.444			

Dieses Ergebnis ist folgendermaßen zu interpretieren:

i) Der Faktor "Substanz" hat nun in beiden Fällen einen signifikanten Einfluss auf das Atemvolumen. Dieser Widerspruch zu den in a) bzw b) erzielten Ergebnissen kommt dadurch zustande, dass dort der Faktor "Altersklasse" bzw eine mögliche Wechselwirkung zwischen den beiden Faktoren nicht berücksichtigt wird und man deshalb als Schätzwert für den unbekannten Parameter $n_{\bullet\bullet}\,\sigma^2$ größere Werte (57.0529 bzw 50.1093) erhält als in jener Situation, bei der sowohl der Faktor Altersklasse als auch eine mögliche Wechselwirkung berücksichtigt wird (29.122).

ii) Auch wenn man den Einfluss berücksichtigt, den ein Faktor auf das Atemvolumen besitzt, hat auch der andere Faktor einen signifikanten Einfluss auf das Atemvolumen.

iii) Zwischen den beiden Faktoren "Substanz" und "Altersklasse" besteht eine sehr signifikante Wechselwirkung. Diese Wechselwirkung wird deutlich, wenn man die Schätzwerte für die Mittelwerte $\mu_{i\bullet}$ bzw $\mu_{\bullet j}$ mit den Schätzwerten für die Mittelwerte $\mu_{i\ j}$ vergleicht:

```
ANOVA[gesundheit, {Substanz, Altersklasse, Substanz Altersklasse}, {Substanz, Altersklasse}][[2]]
Clear[gesundheit]
```

```
                                                    All                                4.14143
                                                    Substanz[1]                        3.98576
                                                    Substanz[2]                        4.07903
                                                    Substanz[3]                        4.34941
                                                    Altersklasse[1]                    4.4797
                                                    Altersklasse[2]                    4.11636
                                                    Altersklasse[3]                    3.81844
CellMeans → Altersklasse[1] Substanz[1] 4.44636
            Altersklasse[1] Substanz[2] 4.936
            Altersklasse[1] Substanz[3] 4.13
            Altersklasse[2] Substanz[1] 3.394
            Altersklasse[2] Substanz[2] 3.7775
            Altersklasse[2] Substanz[3] 5.14273
            Altersklasse[3] Substanz[1] 4.05667
            Altersklasse[3] Substanz[2] 3.52889
            Altersklasse[3] Substanz[3] 3.79545
```

iv) Ändert man die Reihenfolge der beiden Faktoren, so ändern sich zwar die beiden *p*-Werte geringfügig, das Ergebnis bleibt jedoch insgesamt gleich.

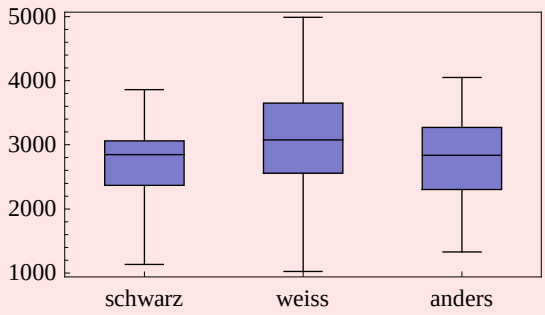
v) Außerdem beachte man, dass die Werte von ${}^{S}\mathrm{SST}_{\mathrm{reg}}$ (6.94353) und ${}^{T}\mathrm{SSS}_{\mathrm{reg}}$ (2.19954) der Varianzanalyse **mit** Wechselwirkung mit den entsprechenden Werten der Varianzanalyse **ohne** Wechselwirkung übereinstimmen.

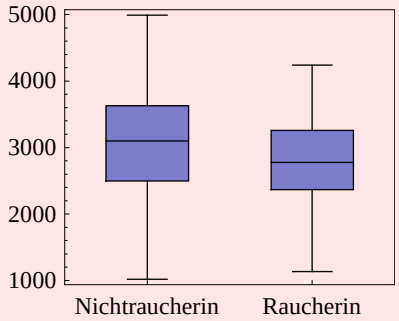
11.3.2 Beispiel: Am Baystate Medical Center wurde das Geburtsgewicht von Neugeborenen, die Rasse der Mutter sowie die Tatsache, ob die Mutter Raucherin ist bzw während der Schwangerschaft Probleme mit der Gebärmutter hatte, erhoben (man vergleiche dazu das Datenmaterial geburtsgewicht). Geprüft soll werden, ob die Faktoren "Rasse", "Raucherin" bzw "Probleme" das Geburtsgewicht signifikant beeinflussen und zwischen welchen dieser Faktoren eine Wechselwirkung besteht.

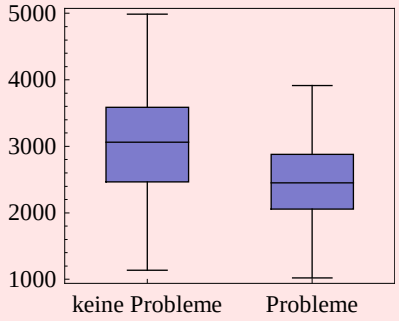
▼

Lösung: Wir veranschaulichen die möglichen Einflüsse der drei Faktoren "Rasse", "Raucherin" und "Probleme" graphisch durch Box-Plots:

```
geburtsgewicht = Rest[<< "geburtsgewichtfile"];
ras1 = Select[geburtsgewicht, #1[[1]] == 1 &][[All, 4]];
ras2 = Select[geburtsgewicht, #1[[1]] == 2 &][[All, 4]];
ras3 = Select[geburtsgewicht, #1[[1]] == 3 &][[All, 4]];
BoxWhiskerPlot[ras1, ras2, ras3, BoxLabels → {"schwarz", "weiss", "anders"}, ImageSize → {200, 120}]
rau0 = Select[geburtsgewicht, #1[[2]] == 0 &][[All, 4]];
rau1 = Select[geburtsgewicht, #1[[2]] == 1 &][[All, 4]];
BoxWhiskerPlot[rau0, rau1, BoxLabels → {"Nichtraucherin", "Raucherin"}, ImageSize → {145, 120}]
pro0 = Select[geburtsgewicht, #1[[3]] == 0 &][[All, 4]];
pro1 = Select[geburtsgewicht, #1[[3]] == 1 &][[All, 4]];
BoxWhiskerPlot[pro0, pro1, BoxLabels → {"keine Probleme", "Probleme"}, ImageSize → {145, 120}]
Clear[ras1, ras2, ras3, rau0, rau1, pro0, pro1]
```







Diese Graphiken lassen vermuten, dass alle drei Faktoren einen Einfluss auf das Geburtsgewicht haben dürften.

a) Wir überprüfen diese Vermutung durch eine dreifache Varianzanalyse **ohne** Wechselwirkung:

```
ANOVA[geburtsgewicht, {Rasse, Raucherin, Probleme}, {Rasse, Raucherin, Probleme},
  CellMeans → False]
```

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Rasse	2	4.79896×10^6	2.39948×10^6	5.56734	0.0044923
Raucherin	1	6.58661×10^6	6.58661×10^6	15.2824	0.000130034
Probleme	1	5.75235×10^6	5.75235×10^6	13.3468	0.000337452
Error	184	7.93025×10^7	430 992.		
Total	188	9.64404×10^7			

Da alle drei *p*-Werte sehr klein sind, haben alle drei Faktoren "Rasse", "Raucherin" und "Probleme" einen sehr signifikanten Einfluss auf das Geburtsgewicht (dass dieser Einfluss derart signifikant ist, lässt sich an Hand der Box-Plots nicht erkennen). Wir wollen nun untersuchen, ob zwischen den einzelnen Faktoren eine paarweise Wechselwirkung besteht. Dazu unterziehen wir unser Datenmaterial einer dreifachen Varianzanalyse **mit paarweiser** Wechselwirkung:

```
ANOVA[geburtsgewicht, {Rasse, Raucherin, Probleme, Probleme Rasse, Probleme Raucherin, Rasse Raucherin},
  {Rasse, Raucherin, Probleme}, CellMeans → False]
Clear[geburtsgewicht]
```

ANOVA →

	DF	SumOfSq	MeanSq	FRatio	PValue
Rasse	2	4.79896×10^6	2.39948×10^6	5.6761	0.00407298
Raucherin	1	6.58661×10^6	6.58661×10^6	15.581	0.000113457
Probleme	1	5.75235×10^6	5.75235×10^6	13.6075	0.000298666
Probleme Rasse	2	180 864.	90 432.2	0.213922	0.807617
Probleme Raucherin	1	411 506.	411 506.	0.973442	0.325155
Rasse Raucherin	2	3.04085×10^6	1.52043×10^6	3.59665	0.0294146
Error	179	7.56693×10^7	422 734.		
Total	188	9.64404×10^7			

An Hand der *p*-Werte erkennt man, dass lediglich zwischen den Faktoren "Rasse" und "Raucherin" eine signifikante Wechselwirkung besteht. Außerdem beachte man, dass sich die *p*-Werte der drei Faktoren "Rasse", "Raucherin" und "Probleme" gegenüber der Varianzanalyse ohne Wechselwirkung etwas verändert haben (die *p*-Werte der einzelnen Faktoren hängen vom jeweils gewählten Modell ab!)