The tensile strength of a new synthetic fiber that will be used to make cloth for men’s shirt is to be tested. In particular, the engineer is interested in the percent of cotton used in the blend of the materials for the fiber. “He suspects increasing cotton will increase the strength”. For various levels of cotton use, he measures the strength of “five” shirts. The results are included in the following table:

<table>
<thead>
<tr>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>12</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>17</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>12</td>
<td>18</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>18</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>18</td>
<td>23</td>
<td>11</td>
</tr>
</tbody>
</table>

Rows are observations. Columns are cotton percentages.

**ANOVA Results**

```r
> fabric.data<-c(7,7,15,11,9,12,17,12,18,18,14,18,18,19,19,19,25,22,19,23,7,10,11,15,11)
> fabric.groups<-c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,4,5,5,5,5,5)
> fabric.groups<-factor(fabric.groups)
> length(fabric.data)
[1] 25
> length(fabric.groups)
[1] 25

> fabric.groups
[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 4 5 5 5 5 5
Levels: 1 2 3 4 5

> fabric.lm<-anova(lm(fabric.data~fabric.groups))
> fabric.lm
Analysis of Variance Table

Response: fabric.data
  Df  Sum Sq Mean Sq F value   Pr(>F)
fabric.groups 4 475.76 118.94 14.757 9.128e-06 ***
Residuals   20 161.20   8.06 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Figure 1. Side-by-side boxplots for the tensile strength problem.
Contrasts

This is a rather more complicated example. Here we have to deal with 5 groups. Additionally, we have already learned that the population means of the five groups are different. The problem is to conduct the appropriate contrasts. I would like to add that one might come up with different sets of ”four contrasts” for this problem. The issue of ”what are the right contrasts” really depends on the scientific questions of interest and the overall scheme is planned even before the ANOVA methods are implemented. In this particular case and by looking at the side-by-side boxplots, one might consider the following four contrasts:

\[ H_{01} : \mu_4 = \mu_5 \]
\[ H_{01} : \mu_1 = \mu_3 \]
\[ H_{03} : 1/2 \times \mu_1 + 1/2 \times \mu_3 = 1/2 \times \mu_4 + 1/2 \times \mu_5 \]
\[ H_{04} : \mu_2 = 1/4 \times \mu_1 + 1/4 \times \mu_3 + 1/4 \times \mu_4 + 1/4 \times \mu_5 \]

These contrasts essentially suggest comparing groups 4 and 5, followed by comparing groups 1 and 3, followed by comparing the first two groups with the second two followed by comparing group 2 with all the others! Note in ANOVA table that \( MSE = 8.06 \). Hence, the coefficients table can be written as follow:

<table>
<thead>
<tr>
<th>Contrasts</th>
<th>( \mu_1 )</th>
<th>( \mu_2 )</th>
<th>( \mu_3 )</th>
<th>( \mu_4 )</th>
<th>( \mu_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_1 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>1/2</td>
<td>0</td>
<td>1/2</td>
<td>-1/2</td>
<td>-1/2</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>-1/4</td>
<td>1</td>
<td>-1/4</td>
<td>-1/4</td>
<td>-1/4</td>
</tr>
</tbody>
</table>

**Sample Means**

|         | 9.8 | 15.4 | 17.6 | 21.6 | 10.8 |

Note that the sample means and variances of the 5 groups can be obtained in R using the following two simple commands:

```R
> tapply(fabric.data,fabric.groups,mean)
   1  2  3  4  5
9.8 15.4 17.6 21.6 10.8
> tapply(fabric.data,fabric.groups,var)
   1  2  3  4  5
11.2 9.8 4.3 6.8 8.2
```

Alternatively, we could do the following:

```R
> mean(fabric[,1][fabric[,2]==1])
[1] 9.8
> mean(fabric[,1][fabric[,2]==2])
[1] 15.4
> mean(fabric[,1][fabric[,2]==3])
[1] 17.6
> mean(fabric[,1][fabric[,2]==4])
[1] 21.6
> mean(fabric[,1][fabric[,2]==5])
[1] 10.8
```
AND:

```r
> var(fabric[,1][fabric[,2]==1])
[1] 11.2
> var(fabric[,1][fabric[,2]==2])
[1] 9.8
> var(fabric[,1][fabric[,2]==3])
[1] 4.3
> var(fabric[,1][fabric[,2]==4])
[1] 6.8
> var(fabric[,1][fabric[,2]==5])
[1] 8.2
```

This yields the following results:

<table>
<thead>
<tr>
<th>F-value</th>
<th>Degrees of Freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{C_1}$ = 36.18</td>
<td>1,20</td>
<td>0.000007</td>
</tr>
<tr>
<td>$F_{C_2}$ = 18.87</td>
<td>1,20</td>
<td>0.0003</td>
</tr>
<tr>
<td>$F_{C_3}$ = 3.88</td>
<td>1,20</td>
<td>0.06</td>
</tr>
<tr>
<td>$F_{C_4}$ = 0.10</td>
<td>1,20</td>
<td>0.76</td>
</tr>
</tbody>
</table>

which shows that groups 1 and 3 are significantly different. The same holds when groups 4 and 5 are compared. However, one might be hesitant to reject the hypothesis of the difference between groups 1 and 3 from one side and groups 4 and 5 from the other. Also, we do not have enough evidence to declare the mean population of group 2 different with the population means of the others.

### Checking the ANOVA Model Assumptions

#### The Normality Assumption

We can visually investigate the normality assumption by creating a QQ-plot for the values of the response variable, in this problem the strength-values:

```r
> qqnorm(fabric.data,pch=16)
```

we can also perform the *Shapiro-Wilk* test for assessing the normality of the response values:

```r
> shapiro.test(fabric.data)

Shapiro-Wilk normality test
```
Figure 2. QQ-plot for the response variable of the "fabric strength" problem.
Figure 3. Plot of residuals against time.

data: fabric.data
W = 0.9497, p-value = 0.2470

The large p-value in this test suggests that we can not reject the null hypothesis of "the response variable is normally distributed".

The Independence Assumption

We may plot residuals against time to check potential auto-correlations or serial-correlation in the model. Here, the lack of any existing pattern will confirm the independence assumption.

> plot(seq(1,25,1),residuals(lm(fabric.data~fabric.groups)))
Figure 3. Plot of predictions versus residuals.

**Constant Variance Assumption**

The homogeneity of the group variances can be visually examined by plotting the residuals versus the predicted values. Here, if there is a larger variance for larger values of the prediction, we would obtain a graph that looks like an "outward-opening funnel or megaphone (Montgomery, 1997, pp.85)."

```r
> fabric.pred<-predict(lm(fabric.data~fabric.groups))
> fabric.resi<-residuals(m(fabric.data~fabric.groups))
> plot(fabric.pred,fabric.resi,ylim=c(-10,10))
```

We can also employ Bartlett’s test to assess the homogeneity of the group variances:

```r
> bartlett.test(fabric.data~fabric.groups)
```
Bartlett test for homogeneity of variances

data: fabric.data by fabric.groups
Bartlett’s K-squared = 0.9331, df = 4, p-value = 0.9198

The null hypothesis here is that “there is constant variance among the four groups”. Large
p-value is a sign that the assumption is not violated.

What if the Assumptions Do Not Hold?

In majority of cases, we may normalize the response or we can stabilize the group-variances by
transforming data using the series of transformations known as Box-Cox transformations (also see
the class-handout). The general paradigm is as following:

\[ y^\lambda = \begin{cases} 
\frac{y^{\lambda-1}}{\lambda y^\lambda} & \lambda \neq 0 \\
\ln(y) & \lambda = 0 
\end{cases} \]

Kruskal-Wallis’ Nonparametric Test

This is the nonparametric version of testing the ANOVA hypothesis. It can be implemented using
the \texttt{kruskal.test} command. Here is the application in the ”fabric-strength” problem:

\begin{verbatim}
> kruskal.test(fabric.data~fabric.groups)

Kruskal-Wallis rank sum test
data: fabric.data by fabric.groups
Kruskal-Wallis chi-squared = 19.0637, df = 4, p-value = 0.0007636
\end{verbatim}

The small p-value suggests that the difference between the population means of the four groups
is significant. This result confirms our findings from the one-way ANOVA test.