5. Concepts of ANOVA

• The $t$-test studied in the previous chapter does not apply directly when there are more than two means to be compared, but the basic ideas extend in a natural way.

• Example: We are to investigate the formulation of a new synthetic fibre that will be used to make cloth for shirts. The cotton content varies from 10% - 40% by weight (the one factor is cotton content) and the experimenter chooses 5 levels of this factor: 15%, 20%, 25%, 30%, 35%. The response variable is $Y = \text{tensile strength}$. There are 5 replicates (complete repetitions of the experiment). In a replicate five shirts, each with a different cotton content, are randomly chosen from the five populations of shirts. The 25 tensile strengths are measured, in random order. This is then a Completely Randomized Design (CRD). (Why random order?)
• Data on web site and below. The levels 15%, ..., 35% are labelled 1, ..., 5, or A, ..., E; their numerical values are not important for the analysis. Write $y_{ij}$ for the $j^{th}$ observation at level $i$ ($i = 1, ..., a = \# \text{ of levels}$). Thus, e.g., $y_{23} = 12$. Totals and averages at level $i$ are $y_i$ and $\bar{y}_i = y_i/n$ ($n = 5 = \# \text{ of replicates}$).

<table>
<thead>
<tr>
<th>Tensile strength data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Level</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
</tbody>
</table>

$y_{..} = 376 \quad \bar{y}_{..} = 15.04$

<table>
<thead>
<tr>
<th>Measurement order</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 22 2 19 7</td>
</tr>
<tr>
<td>17 9 11 12 16</td>
</tr>
<tr>
<td>13 1 24 14 3</td>
</tr>
<tr>
<td>6 8 15 20 25</td>
</tr>
<tr>
<td>4 10 21 23 5</td>
</tr>
</tbody>
</table>
• Questions: Does changing the cotton content (level) change the mean strength? If so, is there a level which results in the maximum mean strength? From the following ‘stripchart’ we suspect that the answers are ‘yes’ and ‘D’.

Fig. 3.1
To answer the first question, we rephrase as: is the variation between the means at the 5 levels large enough, relative to the underlying random variation, that we can conclude that it is not arising purely by chance?

We carry out an ‘Analysis of Variance’ (ANOVA); in this example the numerical output (commands on web site) is

```r
> g <- lm(strength~content)
> anova(g)
Analysis of Variance Table
Response: strength
   Df  Sum Sq  Mean Sq  F value  Pr(>F)
content   4  475.76  118.94  14.757  9.128e-06
Residuals 20  161.20   8.06
```
> A <- c(7, 7, 15, 11, 9)
  etc.
> data <- rbind(A, B, C, D, E)
> data
A  7   7  15  11  9
B 12  17  12  18  18
C 14  18  18  19  19
D 19  25  22  19  23
E  7  10  11  15  11

# The sample mean for each treatment
> apply(data, 1, mean)
     A    B    C    D    E
9.8 15.4 17.6 21.6 10.8

# The overall mean
> mean(data)
[1] 15.04

# Total SS:
> (25-1)*var(strength)
[1] 636.96
# Draw a scatterplot of 'data' as a data frame
# with treatments as columns
> stripchart(data.frame(t(data)), vertical = TRUE)

# Arrange responses by column
> strength <- c(data)
> strength
   [1]  7 12 14 19  7  7 17 18 25 10 15 ... 
# Set the factor at 5 levels
> d <- rep(c("A", "B", "C", "D", "E"), times=5)
> content <- as.factor(d)
> content
   [1] ABCDEABCDEABCDEABCDE... 
Levels: A B C D E
# Perform one-way ANOVA
# lm stands for 'linear model'
> g <- lm(strength~content)
> anova(g) # Produces the ANOVA table
• Interpretation: The Total Sum of Squares ($SS_T$) measures the total variability around the overall average:

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = 636.96.$$  

How much of this is attributable to differences between the level (treatment) means?

$$SS_{Treatments} = \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 = 475.76.$$  

How much is attributable to random variation of the $y_{ij}$ around these treatment means?

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 = 161.20.$$  

Degrees of freedom of the SS’s: Note that $SS_T$ is the sum of $N = an$ squares, but the sum of the unsquared terms is $\sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..}) = 0$. Thus one of them cannot vary and there are $N - 1 = 24$ d.f. Similarly $SS_{Treatments}$ has
\[ a - 1 = 4 \text{ d.f.} \] The error sum of squares has \[ \sum_{i=1}^{a} (n - 1) = N-a \text{ d.f., since } \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.}) = 0 \text{ for } i = 1, \ldots, a. \]

- The theoretical ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>( F_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>( SS_{Tr} )</td>
<td>( a - 1 )</td>
<td>( MS_{Tr} = \frac{SS_{Tr}}{a-1} )</td>
<td>( F_0 = \frac{MS_{Tr}}{MS_E} )</td>
</tr>
<tr>
<td>Error</td>
<td>( SS_E )</td>
<td>( N - a )</td>
<td>( MS_E = \frac{SS_E}{N-a} )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( SS_T )</td>
<td>( N - 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It can be shown that if the mean strengths are the same at all levels (i.e. ‘under the hypothesis of no treatment effects’), so that \( \bar{y}_1, \ldots, \bar{y}_a \) and \( \bar{y}_. \) are all estimating the same thing (the ‘overall mean’), then (assuming as well that the observations are independently and normally distributed, with common variance \( \sigma^2 \))

\[
\frac{SS_{Tr}}{\sigma^2} \sim \chi^2_{a-1}, \text{ ind. of } \frac{SS_E}{\sigma^2} \sim \chi^2_{N-a},
\]
so that

\[
\frac{SS_{Tr}}{\sigma^2} / (a - 1) = \frac{MS_{Tr}}{MS_E} = F_0 \sim F_{a-1, N-a}.
\]

If the mean strengths are not equal we expect larger values of \( MS_{Tr} \) than otherwise, hence larger values of \( F_0 \); thus large values of \( F_0 \) indicate that the hypothesis of no treatment effects should be rejected. The corresponding \( p \)-value is

\[
P\left(F_{N-a}^{a-1} > F_0 \right),
\]

which in our example is \( P \left(F_{20}^{4} > 14.757 \right) = 9.128e-06 \); certainly very small! It appears that (at least some of) the treatment means are significantly different.