$y_{ij} = \mu_i + \epsilon_{ij}, j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k$ $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$.

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{k1} \\ \vdots \\ y_{kn_k} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} + \epsilon.$$

$$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix}$$

RSS $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum \sum \hat{\epsilon}_{ij}^2 = \sum \sum (y_{ij} - \hat{\mu}_i)^2$.

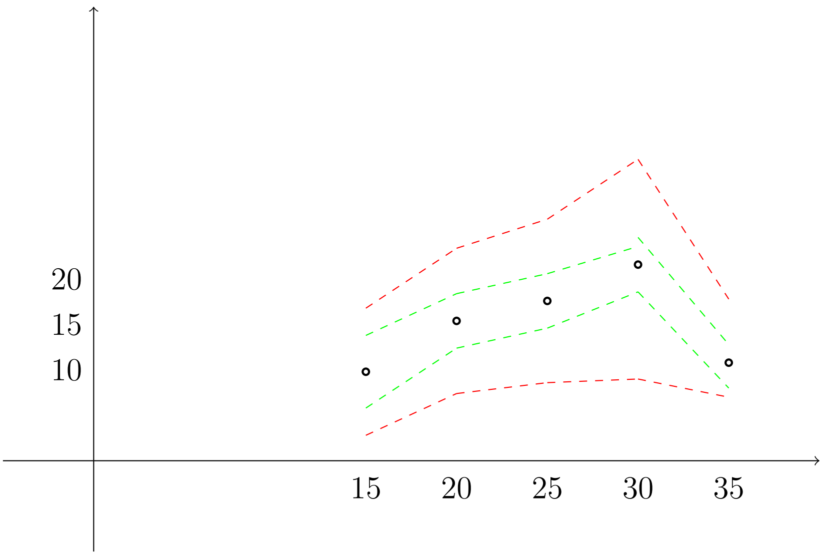
**Questions.**

(i) Are the group means $\mu_i$ equal? i.e., test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.
(ii) If not, how are they different?

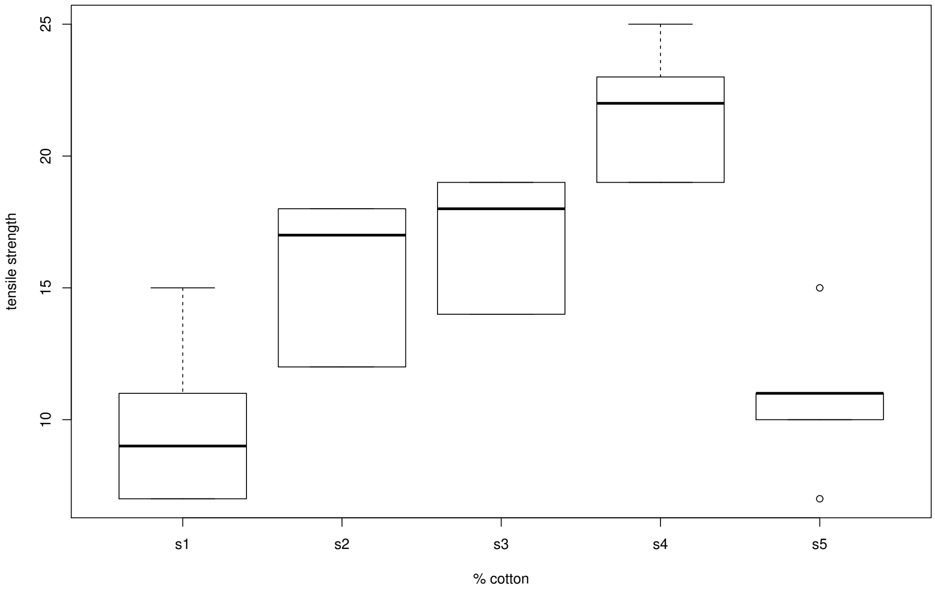
**Example.** It is believed that the tensile (breaking) strength of synthetic fibre is affected by the %age of cotton in fibre:

| % cotton | tensile strength ($lb/inch^2$) | sample mean |
|---|---|---|
| 15 | 7, 7, 15, 11, 9 | $\bar{y}_1 = 9.8$ |
| 20 | 12, 17, 12, 18, 18 | $\bar{y}_2 = 15.4$ |
| 25 | 14, 18, 18, 19, 19 | $\bar{y}_3 = 17.6$ |
| 30 | 19, 25, 22, 19, 23 | $\bar{y}_4 = 21.6$ |
| 35 | 7, 10, 11, 15, 11 | $\bar{y}_5 = 10.8$ |

Are there substantial differences in the mean breaking strength?
(i) Plot the sample means:



But sample means do not tell the whole story, especially for small samples. One must look at variation within samples and between samples. In the plot above, the conclusions would be different according to whether the error bands are green or red.



It is easier to do this investigation of variations using box-plots, as shown above. Variation within samples is not too large or different, but between sample variation is large. Note that, if within sample variation is large compared to between sample variation (like the red error bands in the plot), then the different samples can be considered to be from a single population. However, if within sample variation is small compared to between sample variation (like the green error bands in the plot, i.e., $|\bar{y}_i - \bar{y}_j|$ are large compared to the error) then there is reason to believe that the groups differ.

To formalize this, we return to linear models:
$y_{ij} = \mu_i + \epsilon_{ij}$, $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, $\epsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. Are the group means different?

$\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_k \end{pmatrix}$ so that RSS $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$.

To test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, consider
$A_{(k-1)\times k} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$. Then we test $H_0 : A\mu = 0$ where $A$ has rank $k-1$. To test $H_0$, we obtain $\hat{\mu}_{H_0}$, $\text{RSS}_{H_0}$ and consider

$$F = \frac{(\text{RSS}_{H_0} - \text{RSS})/(k-1)}{\text{RSS}/(\sum_{i=1}^{k} n_i - k)}, \text{ which} \sim F_{k-1, \sum_{i=1}^{k} n_i - k} \text{ under } H_0.$$

To find $\hat{\mu}_{H_0}$, $\text{RSS}_{H_0}$, note that, under $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, these means are equal, and so it is enough to find

$$\min_{\mu_1 = \mu_2 = \cdots = \mu_k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \min_{\mu} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2.$$

Therefore,

$$\hat{\mu}_{H_0} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \equiv \bar{y}_{..}, \text{ and hence } \text{RSS}_{H_0} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2.$$