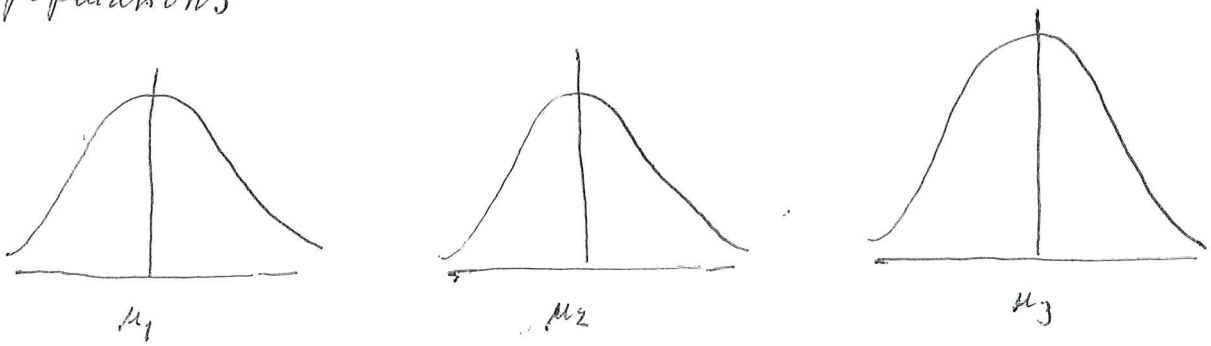
13.3 Analysis of variance (One-Way Anova)

How to make inference about more than two populations



General model

$$\text{Let } Y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} \sim N(0,\sigma^2) & i = 1,2,\dots,k \\ \text{and independent} & j = 1,2,\dots,n_i \end{cases}$$

More common Let $\mu = \sum_{i=1}^{k} n_i \mu_i \Big/ \sum_{i=1}^{k} n_i$

We have $\mu_i = \mu + \underbrace{\mu_i - \mu}_{\alpha_i} = \mu + \alpha_i$

where $\sum_{i=1}^{k} n_i \alpha_i = \sum_{i=1}^{k} n_i(\mu_i - \mu) = \sum_{i=1}^{k} n_i\mu_i - \sum_{i=1}^{k} n_i\mu$

$$= \sum_{i=1}^{k} n_i\mu_i - \sum_{i=1}^{k} n_i\mu_i = 0$$

Which gives $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \begin{cases} \sim N(0,\sigma^2), & i = 1,2,\dots,k \\ \text{and independent} & j = 1,2,\dots,n_i \end{cases}$

Decomposing the variation

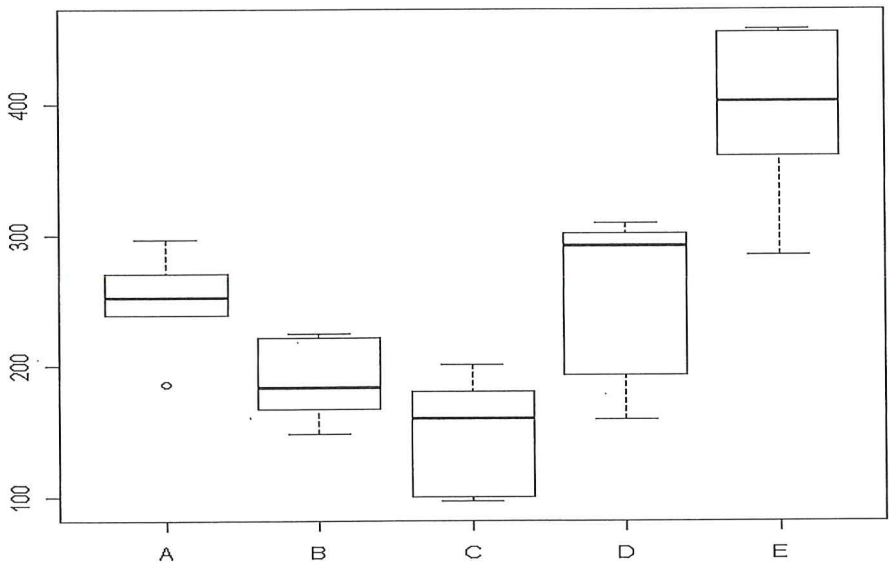
$$\text{Let } \bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad \bar{Y}_{\cdot\cdot} = \frac{1}{\sum_{i=1}^{k} n_i}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$$

The Filter Example

Filters used to remove solid pollutants must be replaced as soon as they fail due to cracking or holes in the filter. An experiment was conducted to test five types of filters, A,B,C,D and E made from different fabrics. Six filters of each type were used under the same conditions and the number of hours until failure were recorded for each. Unfortunately one of the observations for filter type E was corrupted and had to be taken out. The data are given in the table below:

Filters				
A	B	C	D	E
261.1	221.9	201.4	300.9	360.6
186.2	188.7	146.1	301.2	285.0
239.1	167.7	173.9	308.9	455.1
243.3	224.9	280.8	283.3	403.3
296.8	178.8	96.8	193.3	457.9
270.5	147.9	100.3	159.4	

Boxplot




```
      Estimate Std. Error t value Pr(>|t|)
typeA   249.50      20.88  11.950 1.36e-11 ***
typeB   188.30      20.88   9.019 3.55e-09 ***
typeC   149.88      20.88   7.179 2.03e-07 ***
typeD   257.83      20.88  12.349 6.89e-12 ***
typeE   392.18      22.87  17.147 5.70e-15 ***
```

We then have: $y_{ij} = \bar{y}_{..} + \underbrace{\bar{y}_{i.} - \bar{y}_{..}}_{\text{estimator for } \alpha_i} + y_{ij} - \bar{y}_{i.}$

We get $\sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}_{..})^2$

$$= \sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{m_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})$$

Variation within groups — Variation between groups

Some manipulations:

$$\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})\underbrace{\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i.})}_{0}$$

Such that $\sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{m_i}(\bar{y}_{i.} - \bar{y}_{..})^2$

i.e. $SS_T = SS_E + SS_A$

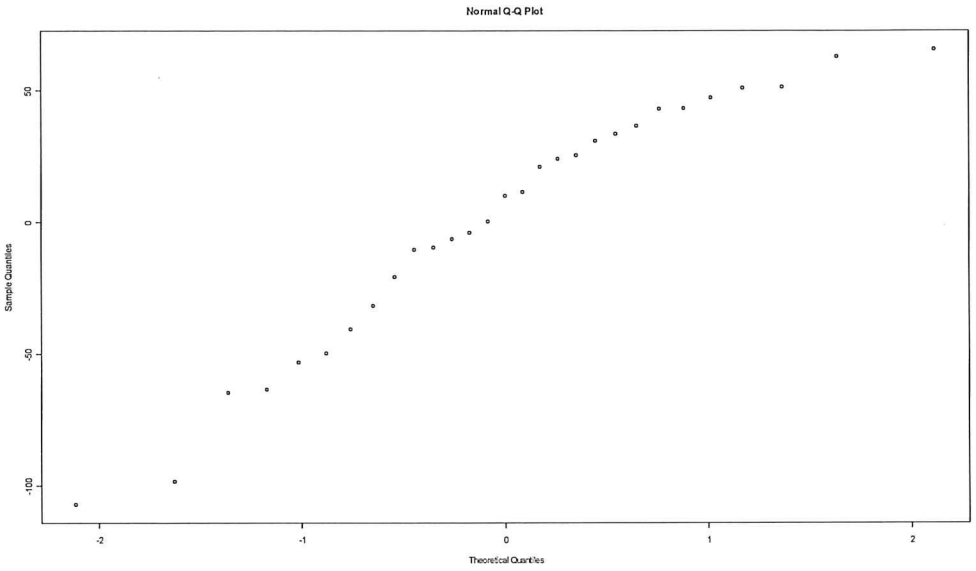
Checking the model

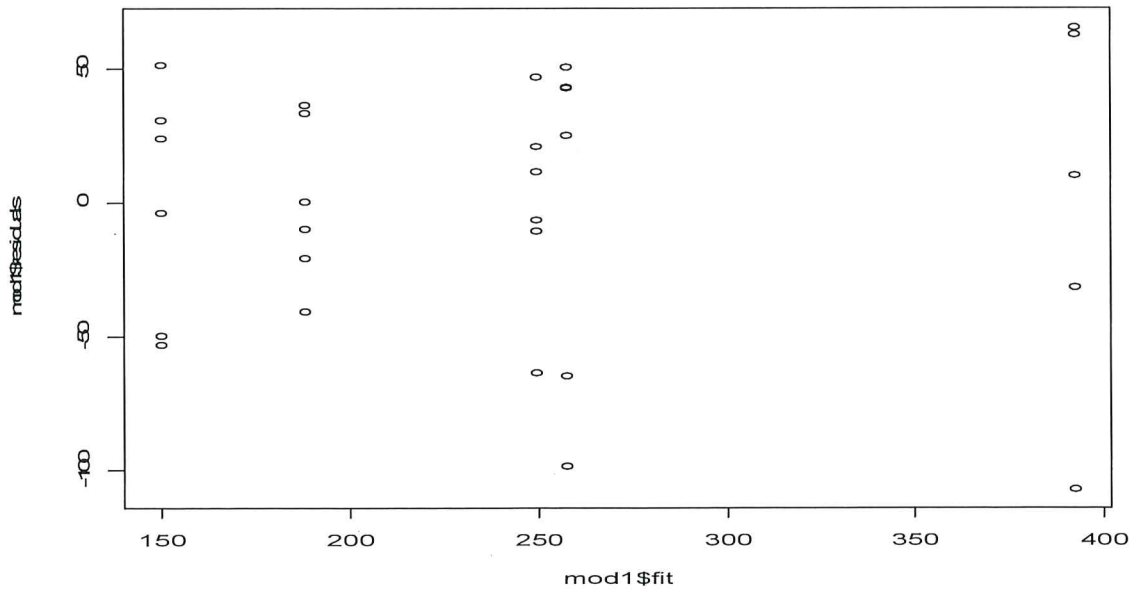
Let $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $\begin{cases} N(0,\sigma^2) \text{ and independent} \\ j = 1,2,\dots,m_i \quad i = 1,2,\dots,k \end{cases}$

Deviation between the model and the data can be revealed by checking the residuals given by

$$y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$$

```
> summary(aov(y~type,data=filter))
            Df Sum Sq Mean Sq F value    Pr(>F)
type         4 182818   45705  17.474 7.724e-07 ***
Residuals   24  62774    2616
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





The following should be performed.

1. Dot-diagram (Check for deviations from the normal distribution). normalplot.
2. Groupwise dot-diagram (Check for deviations within groups)
3. Plot of residuals against $\bar{y}_{i\cdot}$ (Is the variance dependent on the expected value.
4. Plot of residuals against the order the experiments are conducted (look for pattern, correlation)

Test of hypothesis.

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

H_1: at least two are different

at least one $\alpha_i \neq 0$.

Under H_0: $\dfrac{SS_A}{\sigma^2(k-1)} \Big/ \dfrac{SS_E}{\sigma^2(n-k)} \sim F_{k-1,\, n-k}$

Reject H_0 if $F_{obs} \geq f_{\alpha, k-1, n-k}$.

The analysis of variance table

Source	SS	DF	MS	F
Treatment	$SS_A = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$k-1$	$SS_A/k-1$	$\frac{SS_A}{k-1} \Big/ \frac{SS_E}{\sum_{i=1}^{k} n_i - k}$
Error	$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	$\sum_{i=1}^{k} n_i - k$	$SS_E/\sum n_i - k$	
Total	$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$n-1$		

13.6 Multiple comparisons

If $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ is rejected we have found differences between expected values, but not how they differ from each other. A useful tool for this is pairwise comparisons.

LSD-method (Least significant difference)

Make confidence intervals for $\mu_i - \mu_j$.

$$c_{ij} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t_{\frac{\alpha}{2},\nu} s \sqrt{\frac{1}{m_1} + \frac{1}{m_2}}$$

There are $\binom{k}{2}$ such intervals. If c_{ij} does not cover 0, we have proven that μ_i and μ_j differ.

Bonferroni's method

The method above has proven to be useful with the F-test, but one should be aware of the following.

Let $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$. H_1: at least two are different

Let $\bar{c}_{ij}$ be the event that c_{ij} covers 0. $P(\bar{c}_{ij}) = 1-\alpha$ if H_0 is true. Then

$P(\bar{c}_{12} \cap \bar{c}_{13} \cap \cdots \cap \bar{c}_{k-1,k}) = (1-\alpha)^{\binom{k}{2}}$ if we can assume independence between the intervals.

and $P(\text{reject } H_0) = 1 - (1-\alpha)^{\binom{k}{2}} \approx \binom{k}{2}\alpha \geq \alpha$.

Therefore the significance level in pairwise comparisons is often