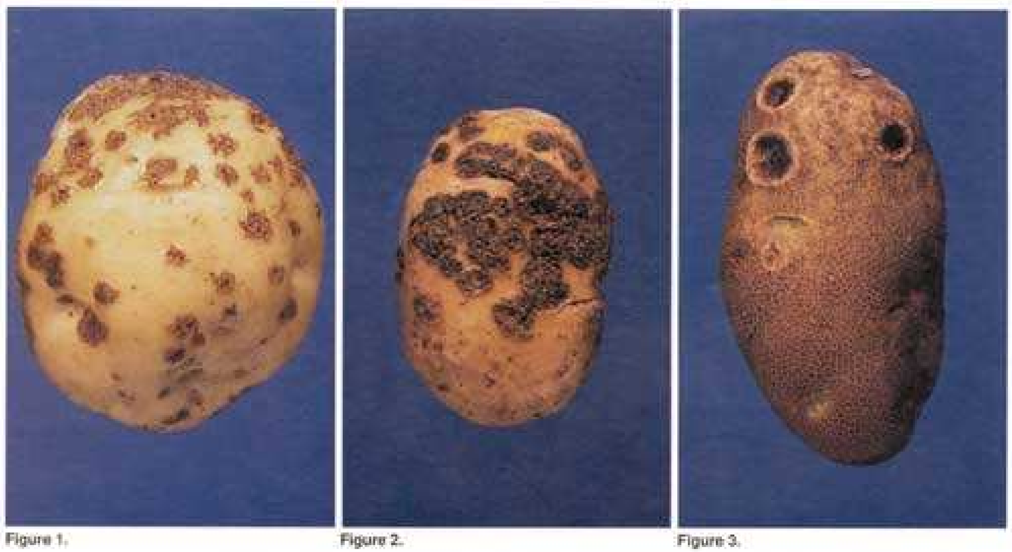
Single Factor Experiments

- Topic:
  - Comparison of more than 2 groups
  - One-Way Analysis of Variance
  - F test
- Learning Aims:
  - Understand model parametrization
  - Carry out an anova
- Reason: Multiple t tests won't do!

Potatoe scab



- widespread disease
- causes economic loss
- known factors: variety, soil condition

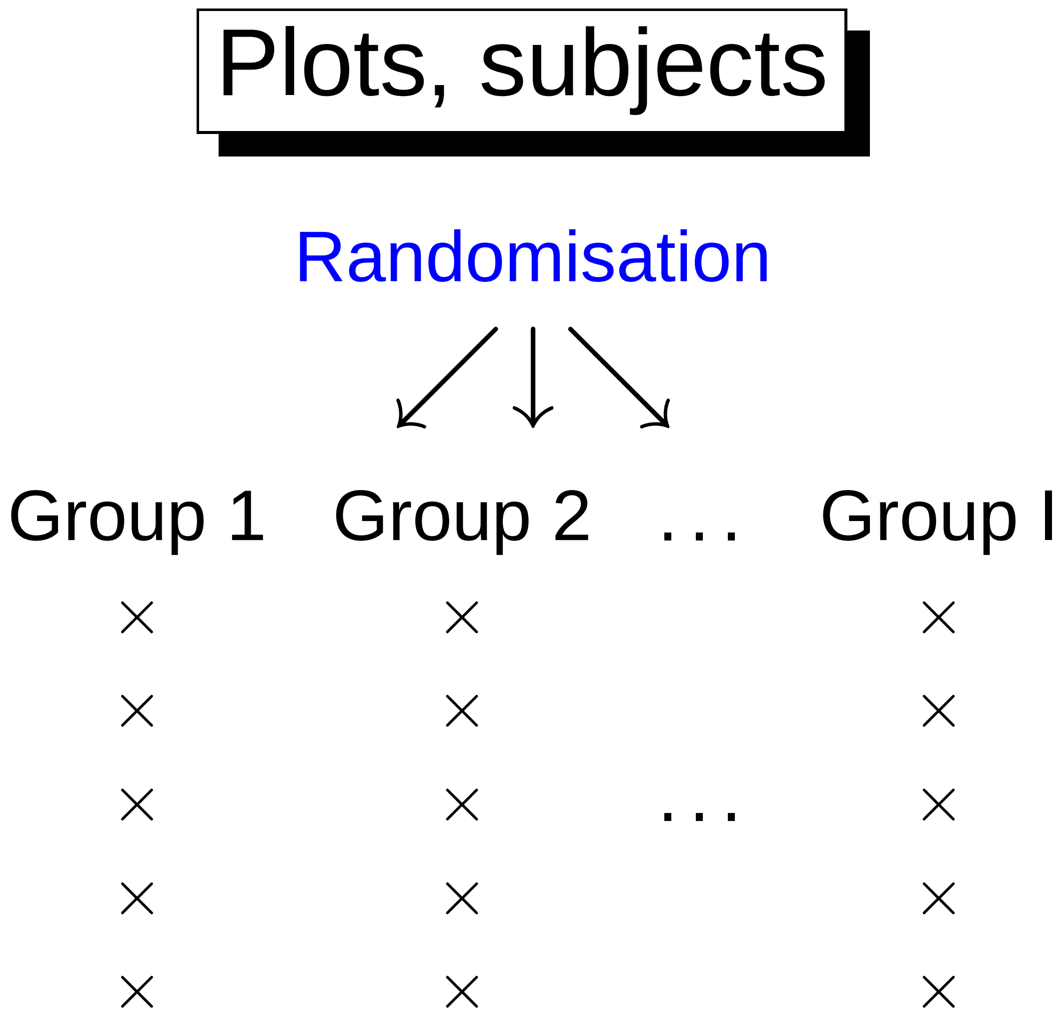
Experiment with different treatments

- Compare 7 treatments for effectiveness in reducing scab
- Field with 32 plots, 100 potatoes are randomly sampled from each plot
- For each potatoe the percentage of the surface area affected was recorded. Response variable is the average of the 100 percentages.

Field plan and data

2 9	1 12	6 18	4 10	6 24	7 17	5 30	3 16
1 10	5 7	4 4	3 10	5 21	1 24	1 29	6 12
2 9	7 7	3 18	1 30	3 18	7 16	2 16	4 4
5 9	1 18	7 17	6 19	1 32	4 5	1 26	2 4

1-Factor Design



Complete Randomisation

a) number the plots 1, . . ., 32.
b) construct a vector with 8 replicates of 1 and 4 replicates of 2 to 7.
c) choose a random permutation and apply it to the vector in b).

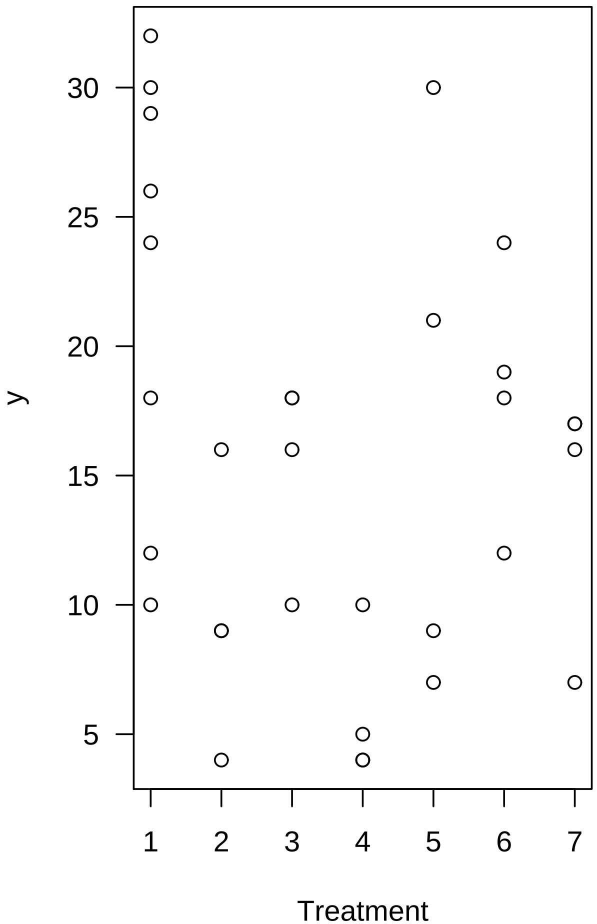
in R:

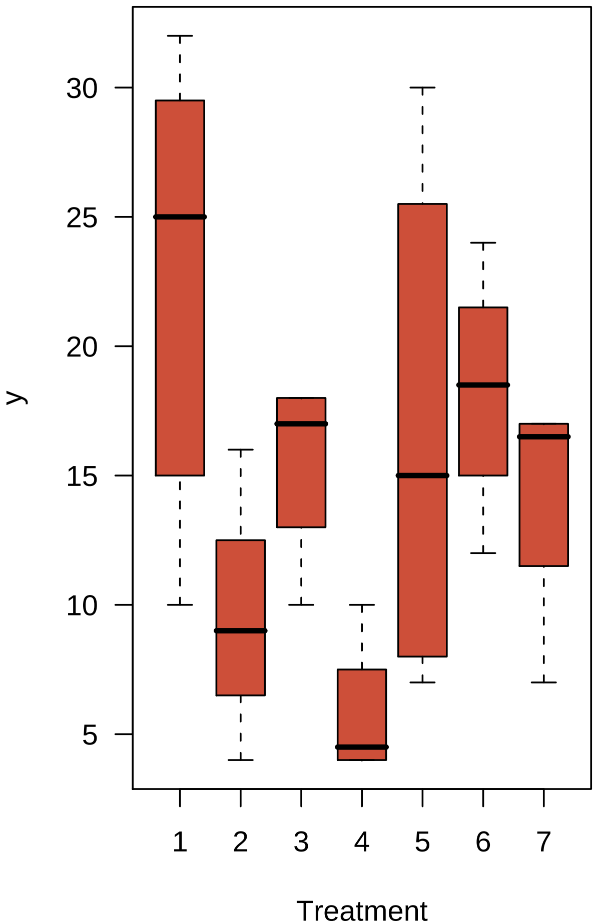
```
> treatment=factor(c(rep(1,8),rep(2:7,each=4)))
> treatment
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7
> sample(treatment)
 [1] 6 4 3 4 7 3 1 2 3 5 5 6 1 7 1 1 2 1 3 2 1 5 7 4 2 1 7 6 6 1 5 4
```

Exploratory data analysis

Group	y								$\bar{y}$
1	12	10	24	29	30	18	32	26	22.625
2	9	9	16	4					9.5
3	16	10	18	18					15.5
4	10	4	4	5					5.75
5	30	7	21	9					16.75
6	18	24	12	19					18.25
7	17	7	16	17					14.25

Graphical display





Two sample t tests

Group 1 – Group 2 : $H_0 : \mu_1 = \mu_2$
Group 1 – Group 3 : $H_0 : \mu_1 = \mu_3$
Group 1 – Group 4 : $H_0 : \mu_1 = \mu_4$
Group 1 – Group 5 : $H_0 : \mu_1 = \mu_5$
Group 1 – Group 6 : $H_0 : \mu_1 = \mu_6$
Group 1 – Group 7 : $H_0 : \mu_1 = \mu_7$

$\ldots$

$\alpha = 5\%$, $P($ Test not significant $|H_0) = 95\%$

7 groups, 21 independent tests:

$P($ none of the tests sign. $|H_0) = 0.95^{21} = 0.34$

$P($ at least one test sign. $|H_0) = 0.66$ (more realistic: 0.42)

$$1 - (1 - \alpha)^n$$

Bonferroni correction

Choose α_T such that

$$1 - (1 - \alpha_T)^n = \alpha_E = 5\%$$

($\alpha_T = \alpha$ „testwise“, $\alpha_E = \alpha$ „experimentwise“)

Since $1 - (1 - \frac{\alpha}{n})^n \approx \alpha$, the significance level for a single test has to be divided by the number of tests.

Overcorrection, not very efficient.

Analysis of variance

- Comparison of more than 2 groups
- for more complex designs
- global F test

Idea:

total variability in data	=	source of variation 1	+	source of variation 2	+ ...

Comparison of components

total	=	treatment	+	experimental error
total	=	variability of plots with different treatments	+	variability of plots with the same treatment
		$\sigma^2 +$ treatment effect		σ^2

Definitions

- Factor: categorical, explanatory variable
  Level: value of a factor
  Ex 1: Factor= soil treatment, 7 levels 1 – 7.
  $\Longrightarrow$ One-way analysis of variance
  Ex 2: 3 varieties with 4 quantities of fertilizer
  $\Longrightarrow$ Two-way analysis of variance
- Treatment: combination of factor levels
- Plot, experimental unit: smallest unit to which a treatment can be applied
  Ex: feeding (chicken, chicken-houses), dental medicine (families, people, teeth)

One-way analysis of variance

Model:

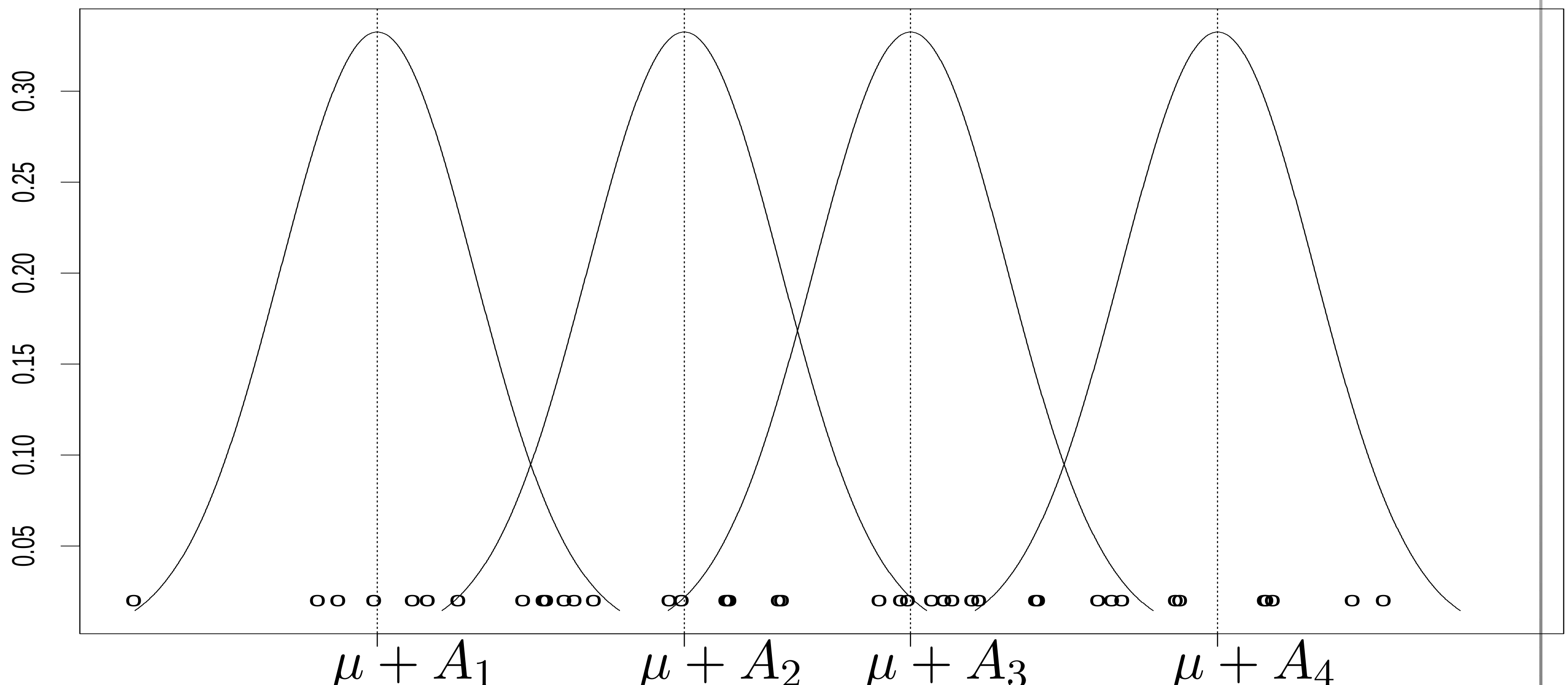
$$\text{response} \quad = \quad \text{treatment} \; + \; \text{error (Plot)}$$

$$Y_{ij} \quad = \quad \mu + A_i \quad + \quad \epsilon_{ij} \qquad (1)$$

$$i{=}1,...,I;\ j{=}1,...,J_i$$

$\mu =$ overall mean
$A_i =$ ith treatment effect
$\epsilon_{ij} =$ random error, $\mathcal{N}(0, \sigma^2)$ iid.

Illustration of model (1)



Necessary constraint

Model (1) is overparametrized, a restriction is needed.

- usual constraint:
  $\sum J_i A_i = 0, \sum A_i = 0$ if $J_i = J$ for all i
  A_i denotes the deviation from overall mean.
- $A_1 = 0$, resp. $A_I = 0$
  First (or last) group is reference group.

Decomposition of the deviation of a response from the overall mean

$$y_{ij} - y_{..} = \underbrace{y_{i.} - y_{..}}_{\text{deviation of the group mean}} + \underbrace{y_{ij} - y_{i.}}_{\text{deviation from the group mean}}$$

$y_{i.} = \frac{1}{J_i} \sum_j y_{ij}$ mean of group i,
$y_{..} = \frac{1}{N} \sum_i \sum_j y_{ij}$ overall mean, $N = \sum J_i$.

Analysis of variance identity

$$\underbrace{\sum_i \sum_j (y_{ij} - y_{..})^2}_{\text{total variability}} = \underbrace{\sum_i \sum_j (y_{i.} - y_{..})^2}_{\text{variability between groups}} + \underbrace{\sum_i \sum_j (y_{ij} - yi.)^2}_{\text{variability within groups}}$$

total sum of squares	=	treatment sum of squares	+	residual sum of squares
SS_{tot}	=	SS_{treat}	+	SS_{res}

Mean squares

Total mean square: $MS_{tot} = SS_{tot}/(N-1)$
Residual mean square: $MS_{res} = SS_{res}/(N-I)$

$$\frac{SS_{res}}{N-I} = \frac{\sum_i (J_i - 1) S_i^2}{\sum_i (J_i - 1)}, \quad S_i^2 = \frac{\sum_j (y_{ij} - y_{i.})^2}{J_i - 1}$$

$$MS_{res} = \hat{\sigma}^2 = \widehat{Var(Y_{ij})}, \quad E(MS_{res}) = \sigma^2$$

Treatment mean square: $MS_{treat} = SS_{treat}/(I-1)$

$$E(MS_{treat}) = \sigma^2 + \sum J_i A_i^2/(I-1)$$

$$df_{tot} = df_{treat} + df_{res}, \quad N - 1 = I - 1 + N - I$$

F test

$$H_0 : \quad \text{all } A_i = 0$$
$$H_A : \quad \text{at least one } A_i \neq 0$$

Since $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $F = \frac{MS_{treat}}{MS_{res}}$ has under H_0 an F distribution with $I - 1$ and $N - I$ degrees of freedom.

one-sided test:
reject H_0 if $F > F_{95\%, I-1, N-I}$

Chisquare and t distribution

- Let $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1), iid$. Then

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

  has a χ^2 distribution with n df, $X \sim \chi_n^2$

- Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi_n^2$ be independent random variables. The distribution of

$$T = \frac{Z}{\sqrt{X/n}}$$

  is called the t distribution with n df, $T \sim t_n$

F distribution

- Let $X_1 \sim \chi^2_n$ and $X_2 \sim \chi^2_m$ be independent random variables. The distribution of

$$F = \frac{X_1/n}{X_2/m}$$

is called the F distribution with n and m df, $F \sim F_{n,m}$

Properties: $F_{1,m} = t^2_m$
$E(F_{n,m}) = \frac{m}{m-2}$

R: anova table

```
> mod1=aov(y~treatment,data=scab)
> summary(mod1)
          Df  Sum Sq  Mean Sq F value Pr(>F)
treatment  6  972.34  162.06  3.608  0.0103 *
Residuals 25 1122.88   44.92
```

F test is significant, there are significant treatment differences.