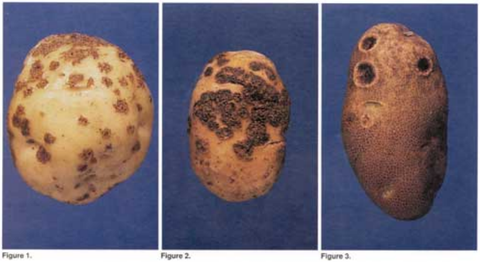
Single Factor experiments

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

1. Comparison of more than 2 groups
2. Analysis of Variance
3. F test

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

Plots, subjects

Randomisation

$\swarrow \downarrow \searrow$

Group 1	Group 2	...	Group 7
$\times$	$\times$		$\times$
$\times$	$\times$		$\times$
$\times$	$\times$	...	$\times$
$\times$	$\times$		$\times$
$\times$	$\times$		$\times$

Complete Randomisation

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

in R:

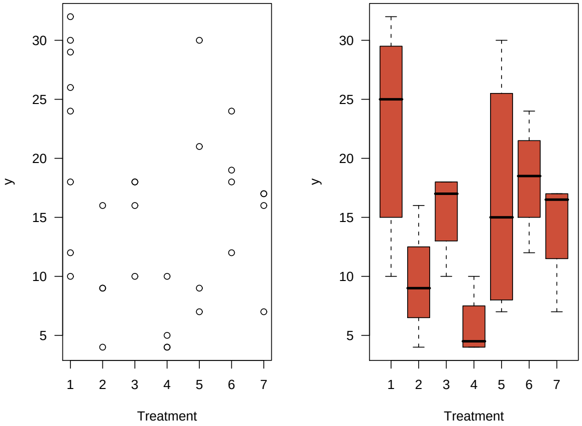
```
> treatment=factor(c(rep(1,8),rep(2:7,each=4)))
> treatment
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7
> sample(treatment)
 [1] 6 4 3 4 7 3 1 2 3 5 5 6 1 7 1 1 2 1 3 2 1 5 7 4 2 1 7 6 6 1
```

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

Question: How to plot the data?
Histogram? Bar chart? Boxplot? Pie chart? Scatter plot?

Graphical display



Why t tests don't work?

$$\begin{aligned}
\text{Group 1} &\ - \text{Group 2} &: \quad H_0: \mu_1 &= \mu_2 \\
\text{Group 1} &\ - \text{Group 3} &: \quad H_0: \mu_1 &= \mu_3 \\
\text{Group 1} &\ - \text{Group 4} &: \quad H_0: \mu_1 &= \mu_4 \\
\text{Group 1} &\ - \text{Group 5} &: \quad H_0: \mu_1 &= \mu_5 \\
\text{Group 1} &\ - \text{Group 6} &: \quad H_0: \mu_1 &= \mu_6 \\
\text{Group 1} &\ - \text{Group 7} &: \quad H_0: \mu_1 &= \mu_7
\end{aligned}$$

...

$\alpha = 5\%$, $P($ 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. $|H_0) = 0.66$ $\quad 1-(1-\alpha)^n$

more realistic: 0.42

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.

Example: 0.05/21=0.0024

Overcorrection, not very efficient.

Terminology

- 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)

What is 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
		σ^2+ treatment effect		σ^2

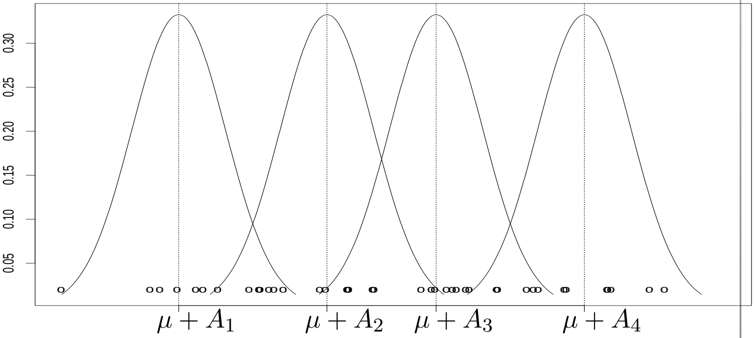
Anova model

Model:

$$Y_{ij} = \mu + A_i + \epsilon_{ij}, \quad i = 1, \ldots, I;\ j = 1, \ldots, J_i$$

Y_{ij} = response of the jth replicate in group i
μ = overall mean
A_i = ith treatment effect
ϵ_{ij} = random error, $\mathcal{N}(0, \sigma^2)$ iid.

Illustration of the model



Decomposition of the deviation of a response from the overall mean

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

$y_{i.} = \frac{1}{J_i} \sum_j y_{ij}$ 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}

Total and Residual mean squares

- Total mean square:

$$MS_{tot} = SS_{tot}/(N-1)$$

- Residual mean square:

$$MS_{res} = SS_{res}/(N-I)$$

$$s_i^2 = \frac{\sum_j (y_{ij} - y_{i.})^2}{J_i - 1} \quad \text{is an estimate of } \sigma^2$$

Pooled estimate of σ^2:

$$\frac{\sum_i (J_i - 1) S_i^2}{\sum_i (J_i - 1)} = \frac{SS_{res}}{N-I} = MS_{res}$$

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

Treatment mean square

- Treatment mean square:

$$MS_{treat} = SS_{treat}/(I-1)$$

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

$$\begin{aligned} df_{tot} &= df_{treat} + df_{res} \\ N-1 &= I-1 + N-I \end{aligned}$$

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 an F distribution with $I - 1$ and $N - I$ degrees of freedom under H_0.

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

Anova table

Source	SS	df	MS=SS/df	F	p
Treatment	SS_{treat}	$I-1$	MS_{treat}	MS_{treat}/MS_{res}	
Residual	SS_{res}	$N-I$	MS_{res}		
Total	SS_{tot}	$N-1$			

in R:

```
> 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.