Design of Experiments

We will cover the following topics:

- One way ANOVA
- Two way ANOVA

Experimental (vs.) Observational Data

A important issue in data analysis is **causality**. If we detect correlation between two variables can we conclude that change in one variable (the independent variable) causes change in the second (dependent) variable?

Note that the independent variable is often referred to as the **explanatory** variable and the dependent variable is often referred to as the **response** variable.

Examples

- It was observed that the number of storks per year nesting in small villages in Denmark was positively correlated with the number of babies born per year. Do storks bring newborns?
- In a study of people aged 50 to 71 years of age it was found that coffee drinkers had increased mortality rates. Is drinking coffee bad for your health?
- Male drivers are correlated with more accidents.

N.B. Sometimes correlation is enough. Since male drivers are correlated with more accidents, insurance companies (used to!) charge them more.

To try to establish causality between variables researchers collect **experimental** data. In an experiment the independent/explanatory variable is changed while holding all other variables constant and the effect of this on the dependant/response variable is examined. It is not always possible to conduct an experiment to collect data, in which case, **observational** data is collected. An observational study draws inferences about the possible relationships between variables where the manipulation of the explanatory variable(s) is outside the control of the investigator. In an observational study no variables are manipulated, the information is simply recorded.

Examples

- Sample surveys are observational studies.
- Study of the relationship between smoking during pregnancy and child's subsequent IQ a few years after birth.
- Google Flu Trends where the spread of the flu virus is predicted by analysing google search queries.

Large data sets are more likely to fall into the latter category since experiments are often expensive to run and therefore performed on a small scale.

Design of Experiments

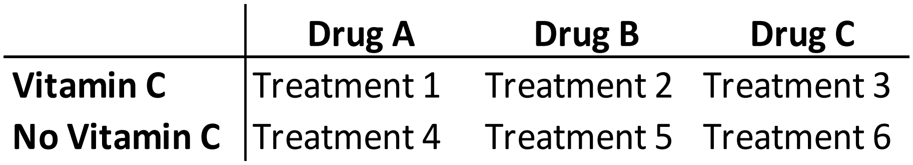
The aim of an experiment is to gain understanding of a subject through hypothesis testing. Once a testable hypothesis has been constructed, an experiment must be designed to gather information to test the hypothesis. The data is analysed statistically and conclusions drawn, additionally, inferences from the new data can be used to inform new hypotheses. An experiment should be designed carefully to ensure that the right type of data, and enough of it, is available to test the hypothesis in question.

Definition 1 *A **factor** of an experiment is a controlled **explanatory variable**; a variable whose **levels** are set by the experimenter.*

Definition 2 *In an experiment, a **treatment** is something that researchers administer to **experimental units/subjects.** Different treatments constitute different levels of a factor, if an experiment has more than one **factor** then each combination of factors is a **treatment.***

Example

Six groups of patients are each treated with one of three different drugs either in combination with a supplement of vitamin C or alone. The experimental design is shown below. In this example, the patients are the experimental units and there are two factors : ‘type of drug’ and ‘vitamin C’. The factor ‘type of drug’ has three levels: ‘A’, ‘B’ , ‘C’ and the factor ‘vitamin C’ has two levels: ‘presence’ and ‘absence’. There are six treatments, one for each combination of drug and vitamin C presence/absence.



	Drug A	**Drug B**	**Drug C**
Vitamin C	Treatment 1	Treatment 2	Treatment 3
No Vitamin C	Treatment 4	Treatment 5	Treatment 6

Experimental control

When designing an experiment we should try to identify known or expected sources of variability in the experimental units since one of the main aims of a designed experiment is to reduce the effect of these sources of variability on the answers to questions of interest. Control involves making the experiment as similar as possible for experimental units in each treatment.

Example

Suppose a pharmaceutical company wishes to compare the effectiveness of a new drug for preventing osteoporosis. The experiment is run on a group of 100 people. The new drug is given to a group of 50 men and a placebo is given to a group of 50 women. At the end of the experiment fewer men report osteoporosis symptoms. Can we conclude that the new drug is effective?

The experiment did not control for the effect of the differences in gender. This leads to experimental **bias**, the favoring of certain outcomes over others.

Randomized design

To avoid **bias** in experiments **randomized** experimental design is used where experimental units are randomly assigned (by chance) to a treatment group. If the sample size is large enough then a randomized design ensures that the background characteristics of each treatment group are sufficiently similar so that comparisons of the groups’ outcome variables measure primarily differences in the effects of the treatments.

If there are specific differences among experimental units that a researcher wishes to control for then a **randomized block design** may be used where experimental subjects units are first divided into homogeneous blocks before they are randomly assigned to a treatment group. In the example above, the researcher could have split the group into male and female subgroups and then randomly assigned patients from each subgroup into the two treatment groups.

Example

A researcher is carrying out a study of the effectiveness of 4 different fertilizers. He has 80 plots located on 10 different farms and plans to divide them into 4 treatment groups of 20 plots each. Using a randomized block design, the plots are put in blocks of 8 according to which farm they are on; the plots from farm 1 are the first block, the plots from farm 2 are the second block, the plots from farm 3 are the third block and so on. The 8 members of each block are then randomly assigned, two to each of the four treatment groups (see Fig. 1).

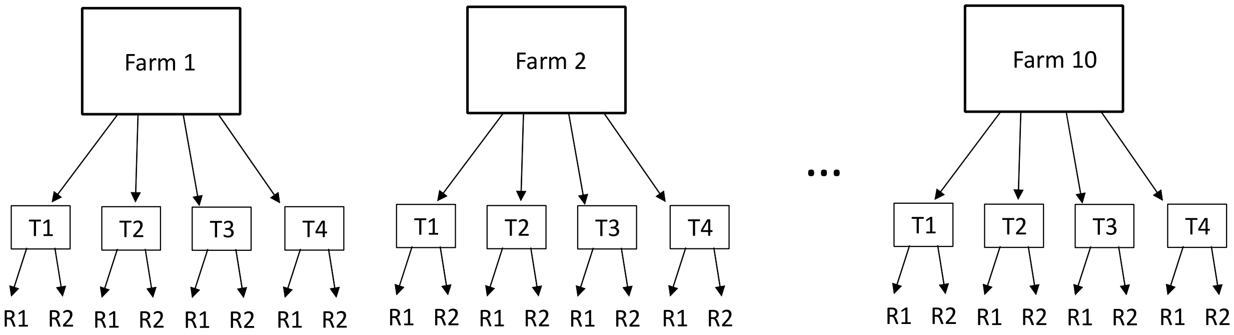


Fig. 1

Replication

Randomization ensures that treatment groups are as similar as possible, provided the sample size is large enough. If the sample size is too small there may be enough background variation in the treatment groups to bias the results. To improve the significance of an experimental result, **replication**, the repetition of an experiment on a large group of subjects, is required. If a treatment is truly effective, the long-term averaging effect of replication will reflect its experimental worth. If it is not effective, then the few members of the experimental population who may have reacted to the treatment will be negated by the large numbers of subjects who were unaffected by it. Replication reduces variability in experimental results, increasing their **significance** and the **confidence level** with which a researcher can draw conclusions about an experimental factor.

Confounding Variables

A **confounding variable** is an extraneous variable that correlates with both the response and explanatory variables being studied so that the results you get do not reflect the actual relationship between the variables under investigation. A perceived relationship between an explanatory variable and a response variable that has been misestimated due to the failure to account for a confounding factor is termed a spurious relationship.
Example 1 above shows a spurious relationship between the number of storks and the number of newborns: the number of storks per year nesting in small villages in Denmark was positively correlated with the number of roof tops in a village suitable for nesting . The larger the town, the more rooftops for storks to nest in but additionally, the larger the town, the more babies born per year.

ANOVA

Analysis of variance (ANOVA) is a statistical method used when we need to compare three or more sample means. ANOVA tests the null hypothesis that samples in two or more groups are drawn from populations with the same mean values. In the simple case where we are comparing just two sample means a t-test can be used and in this case the t-test and an ANOVA are identical. ANOVA is used in cases where the explanatory variable (the factor) has a discrete number of levels. and is analogous to regression analysis where the explanatory variable is continuous. For experiments with both continuous and discrete explanatory variables regression and ANOVA are combined to fit a series of regression lines known as analysis of covariance (ANCOVA).

One Way ANOVA

A one way ANOVA is used when we are investigating a single factor. The variation between experimental units within treatments is compared with the variation between treatments using the F-distribution. There are several ways to define a one way ANOVA model, we begin by considering the **effects** model. The effects model for a one way ANOVA with a levels of a single factor and n replicates within each level of the factor is of the form:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where μ represents the overall mean
τ_i,represents the effect of treatment $i = 1...a$
ε_{ij} represents the error term.
It is assumed that ε_{ij} are independent, identically distributed (i.i.d.) random variables that follow a Normal distribution $\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$.

Example 1 *An agricultural researcher wishes to investigate the effect of soil type (sand, clay or loam) on crop yield. We have an experiment in which crop yields per unit area were measured from 10 randomly selected fields on each of the soil types. All fields were sown with the same variety of seed and provided with the same fertilizer and pest control inputs The question is whether soil type significantly affects crop yield, and if so, to what extent.*

First we state the hypothesis to be tested:

$$\begin{aligned} H_0 &: \text{ the mean crop yields on each soil type are equal i.e. } \mu_1 = \mu_2 = \mu_3 \\ H_A &: \text{ at least two of the mean crop yields on each soil type are not equal} \end{aligned}$$

Note that an alternative way to state this hypothesis is

$$\begin{aligned} H_0 &: \text{ the mean crop yields on each soil type are equal i.e. } \tau_1 = \tau_2 = \tau_3 = 0 \\ H_A &: \tau_1,\ \tau_2,\ \tau_3 \text{ not all equal to } 0 \end{aligned}$$

We will test this assumption at the 5% significance level so $\alpha = 0.05$.
The yield data is shown in the table below.

	Sand	Clay	Loam
	6	17	13
	10	15	16
	8	3	9
	6	11	12
	14	14	15
	17	12	16
	9	12	17
	11	8	13
	7	10	18
	11	13	14
$\frac{\sum}{n}$	9.9	11.5	14.3

We start by examining the data using a boxplot (Fig. 2)

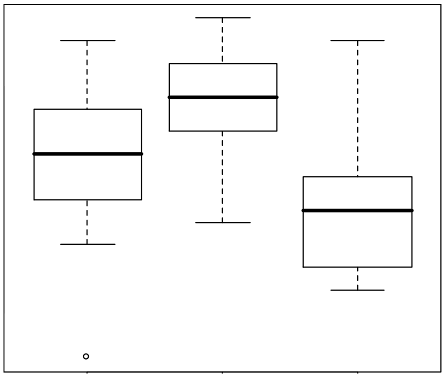


Fig. 2

To compare the means of each treatment we compare the **variation** of yields **between** treatments to the variation in yield **within** each treatment. To calculate the variance we first calculate the **sums of squares** (SS). To do this we first calculate the overall mean (or grand mean) denoted $\overline{\overline{y}}$ and the treatment means, let $\overline{y}_1$ represent the average yield in sand soil, $\overline{y}_2$ represent the average yield in clay soil and let $\overline{y}_3$ represent the average yield in loam soil. In this example $\overline{\overline{y}} = 11.9$, $\overline{y}_1 = 9.9, \overline{y}_2 = 11.5$ and $\overline{y}_3 = 14.3$.

Sums of Squares (SS)

We calculate three different sums of squares:

- the total sum of squares denoted $\mathbf{SS_T}$ (see Fig 3)
- the factor sum of squares, denoted $\mathbf{SS_F}$ (see Fig. 4)
- the error sum of squares, denoted $\mathbf{SS_E}$ (see Fig. 4)

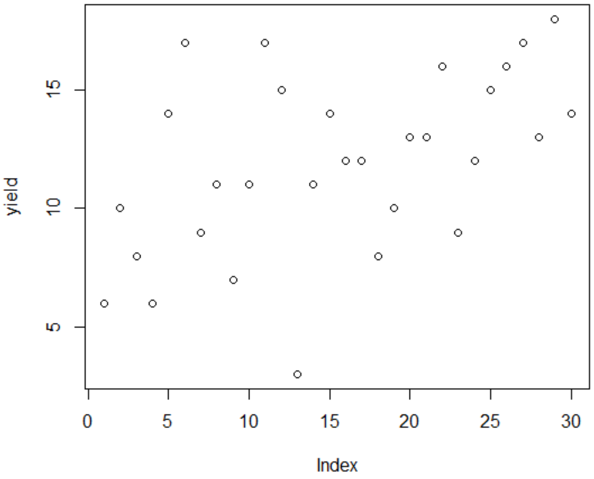


Figure 3

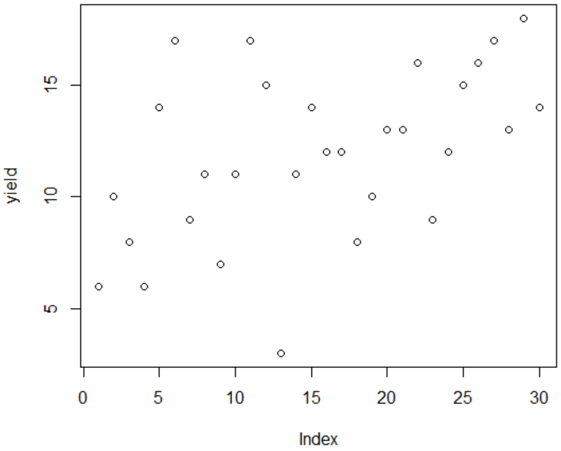


Figure 4

We denote each measure of yield by y_{ij}
where $i = 1, ..., a$ represents the treatment.
where $j = 1, ..., n$ represents the replicate within each treatment
We define:

$$\mathbf{SS_T} = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(y_{i,j} - \overline{\overline{y}}\right)^2$$

where $\overline{\overline{y}}$ is the overall mean (sometimes referred to as the grand mean).

$$\mathbf{SS_F} = n \sum_{i=1}^{a} \left(\overline{y}_i - \overline{\overline{y}}\right)^2$$

Factor SS gives a measure of the variation **between** each treatment.

$$\mathbf{SS_E} = \mathbf{SS_T} - \mathbf{SS_F}$$

Error SS gives a measure of the variation **within** each treatment.
Note that we need only calculate two of the sums of squares since

$$\mathbf{SS_T} = \mathbf{SS_F} + \mathbf{SS_E}$$

When the difference between the means of treatment groups is large, $\mathbf{SS_F}$ will be large.

For the soil example:
$i = 1, 2, 3$ represents the treatment (1 represents sand, 2 represents clay and 3 represents loam)
$j = 1, ..., 10$ represents the replicate within each treatment (there are 10 replicates within each treatment).

Aside: to get a better understanding of this consider the case when:

1. $\mathbf{SS_T} = \mathbf{SS_E}$

2. $\mathbf{SS_E} = 0$

Now

$$\mathbf{SS_T} = \sum_{i=1}^{3}\sum_{j=1}^{10}\left(y_{i,j} - \overline{\overline{y}}\right)^2 = 414.7$$

Recall, $\overline{y}_1$ represents the average yield in sand soil, $\overline{y}_2$ represents the average yield in clay soil and $\overline{y}_3$ represents the average yield in loam soil. Then

$$\mathbf{SS_F} = n\sum_{i=1}^{a}\left(\overline{y}_i - \overline{\overline{y}}\right)^2 = 99.2$$

$$\begin{aligned}\mathbf{SS_E} &= \mathbf{SS_T}-\mathbf{SS_F}\\ &= 414.7 - 99.2\\ &= 315.5\end{aligned}$$

Mean Squares

Recall, our aim is to determine whether the type of soil has an effect on the yield. Our null hypothesis is that mean crop yields on each soil type are equal (i.e. there is no effect of soil type on yield). To compare the means of each treatment we compare the variation of yield **between** treatments (the **Factor SS**) to the variation of yields **within** each treatment (the **Error SS**). If there is a large effect of soil type, we expect the **Factor SS** to be large. We need to determine whether the **Factor SS** are **significantly** large, to do this we examine the ratio of the **Factor Mean Square** with the **Error Mean Square** using the F-test.

The Mean Squares are determined by dividing the sums of squares by their associated **degrees of freedom**.
To calculate the degrees of freedom for each sum of squares we consider the number of treatments (or levels of the factor) and the number of replicates for each treatment.
Let a represent the number of treatments and let n represent the number of replicates for each treatment then the associated degrees of freedom are:

- $\mathbf{SS_F} = a - 1$
- $\mathbf{SS_E} = a(n - 1)$
- $\mathbf{SS_T} = an - 1$

In our example there are 3 levels of the factor *type of soil* so there are $3 - 1 = 2$ degrees of freedom associated with $\mathbf{SS_F}$.

There are 10 replicates within each of the 3 soil treatments so there are $3(10-1)=27$ degrees of freedom associated with $\mathbf{SS_E}$.
There are $10\times 3-1=29$ degrees of freedom associated with $\mathbf{SS_T}$.

Therefore

$$\begin{aligned}\textbf{Factor MS} &= \frac{\mathbf{SS_F}}{(a-1)}=\frac{99.2}{2}\\ \textbf{Error MS} &= \frac{\mathbf{SS_E}}{a(n-1)}=\frac{315.5}{27}\end{aligned}$$

Once the mean squares have been calculated they are usually compared using an ANOVA table of the form:

Source	Sum of Squares (SS)	*df*	Mean Square (MS)	F-ratio	Critical F
Treatments	$\mathbf{SS_F}$	$a-1$	Factor MS $=\frac{\textbf{Factor SS}}{a-1}$	$\frac{\text{Factor MS}}{\textbf{Error MS}}$	$\mathrm{F}_{a-1,a(n-1)}$
Error	$\mathbf{SS_E}$	$a(n-1)$	Error MS$=\frac{\textbf{Error SS}}{a(n-1)}$		
Total	$\mathbf{SS_T}$	$na-1$			

The ratio $\frac{\text{Factor MS}}{\textbf{Error MS}}$ is known as the F-ratio and the significance of this is checked using the F-distribution. If the F-ratio is greater than the associated critical F value then the treatments are considered significantly different at the α% level

For our soil example the table is:

Source	Sum of Squares (SS)	*df*	Mean Square (MS)	F-ratio	Critical F
Treatments	$\mathbf{SS_F}$	2	$\frac{99.2}{2}=49.6$	$\frac{49.6}{11.69}=4.24$	3.37
Error	$\mathbf{SS_E}$	27	$\frac{315.5}{27}=11.69$		
Total	$\mathbf{SS_T}$	29			

Our F-ratio is greater than F-Crit so we reject the null hypothesis and conclude that the mean crop yields on each soil type are not equal i.e. soil type affects yield. At least one soil type has a mean yield that is significantly different from the others at the 5% level.

The ANOVA Model

Recall, the effects model for a one way ANOVA with a levels of a single factor and n replicates within each level of the factor is of the form:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

y_{ij} is the j^{th} observation $(j = 1...n)$ from the i^{th} treatment $(i = 1...a)$
μ represents the overall mean
τ_i represents the effect of treatment $i = 1...a$,
ε_{ij} represents the random error for the j^{th} observation from the i^{th} treatment $\varepsilon_{ij} \sim \mathrm{NID}(0, \sigma^2)$

For the soil example, we know that the value of the overall mean, μ is 11.9.The effect of the different soil types is denoted as τ_i and this represents the difference between the treatment mean and the overall mean. In this example:
the mean for sand was $\overline{y}_1 = 9.9$, therefore $\tau_1 = 9.9 - 11.9 = -2$
the mean for clay was$\overline{y}_2 = 11.5$, therefore $\tau_2 = 11.5 - 11.9 = -0.4$
the mean for loam $\overline{y}_3 = 14.3$, therefore $\tau_3 = 14.3 - 11.9 = 2.4$
Thus the fitted model for sand is:

$$\widehat{y}_{1j} = 11.9 - 2 + e_{1j}$$

the fitted model for clay is:

$$\widehat{y}_{2j} = 11.9 - 0.4 + e_{2j}$$

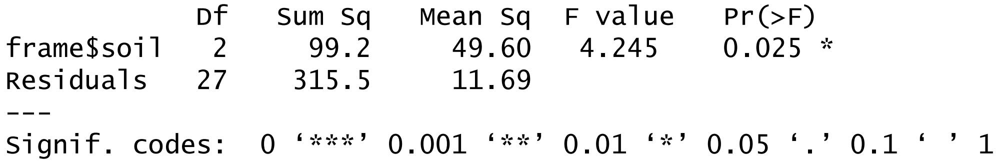
the fitted model for loam is:

$$\widehat{y}_{3j} = 11.9 + 2.4 + e_{2j}$$

The value of τ_i is the **effect** of treatment i, the stronger the effect of a treatment the larger τ_i will be. In the soil example, both sand and clay had a negative effect on the yield but sand had the most negative effect. Loam had a large positive effect on the yield.

ANOVA using R

If we fit an ANOVA on the yield data using R, we obtain the following ANOVA table:



```
             Df   Sum Sq   Mean Sq  F value   Pr(>F)
frame$soil    2     99.2     49.60   4.245    0.025 *
Residuals    27    315.5     11.69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Notice that the output includes a p-value of 0.025 which tells us there is a significant difference (at the 5% level) between the yield for different soil types.
We can also obtain a summary of the ANOVA model:

```
Coefficients:
               Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)      11.500        1.081    10.638   3.7e-11 ***
frame$soilLoam    2.800        1.529     1.832   0.0781 .
frame$soilSand   -1.600        1.529    -1.047   0.3046
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-squared:  0.2392,  Adjusted R-squared:  0.1829
F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.02495
```

Note that the default output appears to be different to the effects model we wrote down above. The model summary produced by R has set the intercept term to be the average yield for Clay, this can be thought of as a baseline. For experiments where there is a ’control’ treatment it is helpful to set the control treatment as the baseline, then treatments can be compared to the control.
The coefficients for Loam and Sand are the estimates for the difference between the means of treatment groups and the baseline group (in this case Clay).
The standard error shown the standard error of the difference between these means
Note that:

- the estimate for the baseline group has been compared to 0 - not interesting
- comparisons have been made between treatment groups and baseline group only - no comparisons between treatment groups (i.e. Loam and Sand).

Dummy Variables (ANOVA using a Linear Regression Model)

An alternative way to define the ANOVA model is to use Dummy Variables. For the Yield example, if we set Clay to be the baseline group, Loam to be group 1 and Sand to be group 2 then we can write :

$$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_{ij}$$

- β_0 is the mean value of the baseline group
- β_1 is the difference between the mean value of the baseline group and the mean value of group 1
- β_2 is the difference between the mean value of the baseline group and the mean value of group 2
- x_1 is a dummy (or indicator) variable for group 1, it takes on the value 1 if the observation is in group 1 and 0 otherwise
- x_2 is a dummy (or indicator) variable for group 2, it takes on the value 2 if the observation is in group 2 and 0 otherwise

The fitted model for the Yield example is:

$$\begin{aligned} \widehat{y}_{ij} &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + e_{ij} \\ \widehat{y}_{ij} &= 11.5 + 2.8 x_1 - 1.6 x_2 + e_{ij} \end{aligned}$$

Remark: It is possible to obtain the effect estimates for the effects model ($y_{ij} = \mu + \tau_i + \varepsilon_{ij}$) by changing the contrast settings in R.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  11.9000     0.6241  19.067   <2e-16 ***
frame$soil1  -0.4000     0.8826  -0.453   0.6540    
frame$soil2   2.4000     0.8826   2.719   0.0113 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-squared:  0.2392,  Adjusted R-squared:  0.1829 
F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.02495
```

Note that since the treatment effect estimates are not unique (we obtain different coefficients depending on the model specified), we focus on the difference between treatment means which are unique.

Effect Size

To measure the effect size associated with a one way ANOVA, we use η^2 (Eta squared).

$$\eta^2 = \frac{\mathbf{SS_F}}{\mathbf{SS_T}}$$

Eta squared measures the proportion of the total variance in the response variable that is associated with changes in the explanatory variable or factor.
For the example above

$$\eta^2 = \frac{\mathbf{SS_F}}{\mathbf{SS_T}} = \frac{99.2}{414.7} = 0.239 \text{ or } 23.9\%$$

We can say that 23.9% of the variation in yield was caused by soil type.

Assumptions of ANOVA

Before using ANOVA on a data set we should check that certain assumptions hold. If the assumptions are violated then the results of the ANOVA are not reliable.
The results of an ANOVA are only valid if the following assumptions are met:

- random sampling
- variances of different treatments are equal (homogeneity of variance)
- the error terms are independent from observation to observation and are normally distributed with zero mean and the same variance i.e. the error terms are i.i.d. random variables.

Variances Before running an ANOVA, a boxplot can show the spread of data within each treatment group, this will indicate whether the variances are the same across treatment groups. This can be tested formally with Bartlett's test, Levene's test or the Fligner - Killeen test.

Residual Analysis The assumptions of ANOVA can be checked by analysing the **residuals** of the fitted model. The residual is defined to be the difference between an actual observation y_{ij} and the fitted value $\overline{y}_i$. Let e_{ij} represent a residual, then

$$e_{ij} = y_{ij} - \overline{y}_i$$

Example

For the soil example the residuals are:

Sand	6	10	8	6	14	17	9	11	7	11
$\overline{y}_1$	9.9	9.9	9.9	9.9	9.9	9.9	9.9	9.9	9.9	9.9
$\mathbf{e}_{1j}$	-3.9	0.1	-1.9	-3.9	4.1	7.1	-0.9	1.1	-2.9	1.1

Clay	17	15	3	11	14	12	12	8	10	13
$\overline{y}_2$	11.5	11.5	11.5	11.5	11.5	11.5	11.5	11.5	11.5	11.5
$\mathbf{e}_{2j}$	5.5	3.5	-8.5	-0.5	2.5	0.5	0.5	-3.5	-1.5	1.5

Loam	13	16	9	12	15	16	17	13	18	14
$\overline{y}_3$	14.3	14.3	14.3	14.3	14.3	14.3	14.3	14.3	14.3	14.3
$\mathbf{e}_{3j}$	-1.3	1.7	-5.3	-2.3	0.7	1.7	2.7	-1.3	3.7	-0.3

The residuals (vs) fitted values for the soil example are shown below in Fig. 5:

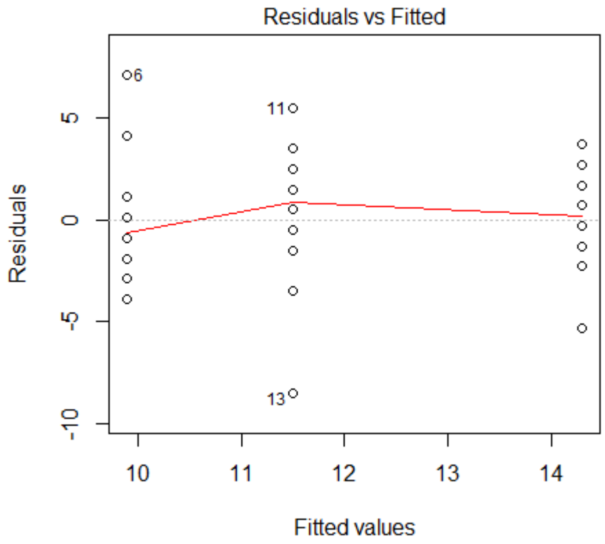


Fig. 5

We see that the spread or variation of the residuals is similar across the three groups so the data does not violate the assumption that the variances of the different treatments are equal.

Normal Probability Plots The assumption that the error terms follow a normal distribution can be checked by constructing a normal probability plot of the residuals.

A normal probability plot is a graphical way of checking whether sample data follow a normal distribution. In general, the basic idea is to compute the theoretically expected value for each data point based on the normal distribution and plot these against the observed data points. If the data does follow a normal distribution, the points in the normal probability plot will approximately lie on the line $y = x$.

To construct a probability plot, the observations are first ranked from smallest to largest. Next, the cumulative frequency of each observation is calculated and this value is converted into a Z-score. These theoretical Z-scores are then plotted against the observed data and the fit is checked.

Example

To check the residuals for the soil example, we must first order the residuals and then calculate their cumulative frequency using the formula $(j-0.5)/n$. The cumulative frequency is then converted to a $z-$score using the normal distribution tables (or a statistical software package).

$$\frac{j-0.5}{n} = P\left(Z \leq z_j\right)$$

We use the inverse normal table to find z_j. The table below shows the calculation of the theoretical Z-scores for the first 10 residuals only (there are 30 in the data set).

order j	**Residual**	$(j-0.5)/30$	z_j
1	-8.5	0.017	-2.13
2	-5.3	0.05	-1.64
3	-3.9	0.083	-1.38
4	-3.9	0.117	-1.19
5	-3.5	0.15	-1.04
6	-2.9	0.183	-0.90
7	-2.3	0.217	-0.78
8	-1.9	0.25	-0.67
9	-1.5	0.283	-0.67
10	-1.3	0.317	-0.48

The theoretical z scores are then plotted against the residuals (Fig. 6):

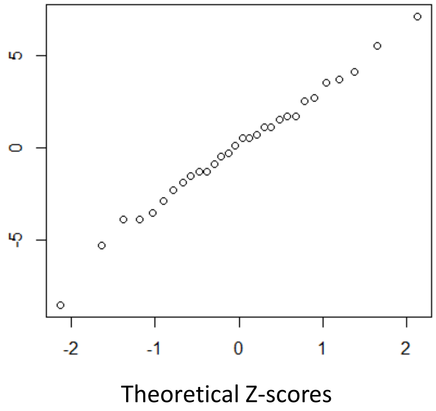


Fig. 6

The normal probability plot shows that the residuals (vs) the theoretical z-scores approximates a straight line so we may assume that the assumption of normality has not been violated.

Pairwise Comparisons

For the yield example, the ANOVA found that there was a significant difference between the three soil types but did not tell us which types of soil had significantly different yields. It is possible to compare each pair of soil types by calculating the difference between the treatment means and testing whether this difference is significantly different to zero. A popular method for performing pairwise comparisons is Tukey's honest significant difference (HSD). The output for the yield example is shown below.

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = frame$yield ~ frame$soil)

$`frame$soil`
              diff        lwr        upr     p adj
Loam-Clay      2.8  -0.9903777   6.5903777  0.1785489
Sand-Clay     -1.6  -5.3903777   2.1903777  0.5546301
Sand-Loam     -4.4  -8.1903777  -0.6096223  0.0204414
```

The `diff` coefficients report the difference between the mean yields for each of the pairwise comparisons, the `lwr` and `upr` coefficients report the 95% confidence interval around the difference and the p-value tells us whether this difference is statistically different to zero. Here we see that the mean yield of the Sand and Loam treatments were significantly different. This information is summarised graphically below in Fig. 7.

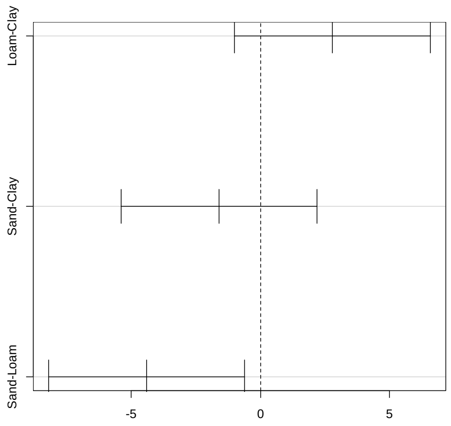


Fig. 7

The confidence interval around the Sand - Loam difference does not contain zero, indicating that there is a significant difference between the yield of Sand and Loam soil types.

Large k (multiple comparisons)

If a data set contains many variables and these are analysed simultaneously, then it becomes more likely that a statistically significant relationship will be found by random chance alone i.e. it becomes more likely that we will make a type I error. For example, if we test a null hypothesis which is in fact true, using 0.05 as the critical significance level, we have a probability of 0.95 of coming to a 'not significant' (i.e. correct) conclusion. If we test two independent true null hypotheses, the probability that neither test will be significant is $0.95 \times 0.95 = 0.90$. If we test twenty such hypotheses the probability that none will be significant is $0.95^{20} = 0.36$. This gives a probability of $1 - 0.36 = 0.64$ of getting at least one significant result; we are more likely to get one than not. The expected number of spurious significant results is $20 \times 0.05 = 1$.

If the probability of making a type I error in a single test is α then the probability of not making a type I error is $(1 - \alpha)$. Scaling up to m tests, the probability of not making a type I error in m tests is $(1 - \alpha)^m$, therefore, the probability of making at least one error in m tests is $1 - (1 - \alpha)^m$. If we plot $1 - (1 - \alpha)^m$ for α set to 0.05 we see that as m gets large, the probability of making at least one type I error (or false positive) tends to 1 (Fig 8.).

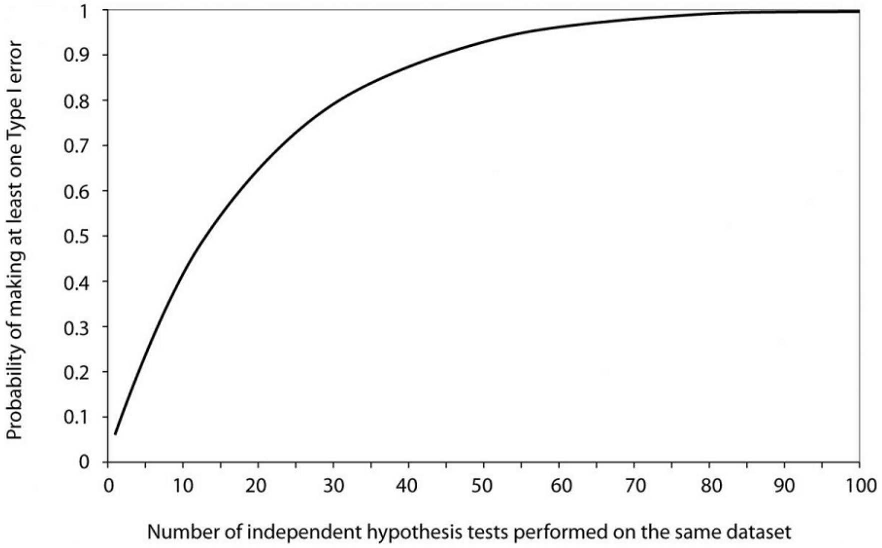


Figure 8

The probability of getting a significant result simply due to chance (a type I error) increases with the number of relationships tested.

To deal with the problem of spurious significant results when performing multiple testing, the critical significance level α is often adjusted in some way so that the probability of observing at least one significant result due to chance alone remains below the significance level required. There are numerous ways of dealing with multiple comparisons, the details depend on the choice of statistical model.

The Bonferroni correction The Bonferroni correction is the most well known method for dealing with multiple comparisons but it is criticized for being too conservative i.e. it greatly increases the chances of making a type II error or false negative. To guarantee that the **overall** significance test is still at the α level, we have to adapt the significance level α' of the **individual** tests. The Bonferroni correction sets the significance level α' of the individual tests to $\frac{\alpha}{m}$ where m represents the number of tests performed. In the example above, with 20 tests and $\alpha = 0.05$, you'd only reject a null hypothesis if the p-value is less than 0.0025. The Bonferroni correction tends to be a bit too conservative. To demonstrate this, let's calculate the probability of observing at least one significant result out of the twenty hypothesis tests when using the Bonferroni correction:

Example 2

$$\begin{aligned} P\left(\textit{at least one significant result}\right) &= 1-(1-0.0025)^{20} \\ &= 0.0488 \end{aligned}$$

Which is under the required 0.05 *level.*

Family Wise Error Rate

A common way of dealing with multiple comparisons is to consider the Family Wise Error Rate. If we are testing m hypothesis where for each case we can either accept or reject the null hypothesis then the outcomes of the m tests will be of the form:

	H_0True	H_A True
Reject H_0	V	S
Fail to reject H_0	U	T

where V represents the number of type I errors (false positives) and T represents the number of type II errors (false negatives).

Definition 3 *The probability of obtaining at least one type I error ($P(V \geq 1)$), is known as the Family Wise Error Rate.(FWER)*

The family refers to the set of inferences (or tests) being analysed simultaneously.

To avoid making type I errors when making multiple comparisons the FWER is set to some level α and the significance of the individual tests α' is adjusted to ensure that:

$$\text{FWER} = P(V \geq 1) < \alpha.$$

The Bonferroni correction is an example of this technique but there are many other methods used to adjust the significance of the individual tests α'. The method used will depend on choice of the statistical model.

False Discovery Rate (FDR) An alternative way of dealing with multiple comparisons it to use a False Discovery Rate (FDR) procedure. As in the case for FWER, the FDR is a way of controlling the number of type I errors in null hypothesis testing when conducting multiple comparisons. Using the notation above where V represents the number of type I errors/false positives, and S represents the number of correct rejections of the null hypotheses (true positives), let $R = V + S$ represent the total number of discoveries.
The False Discovery Rate is defined to be

$$E\left[\frac{V}{R}\right]$$

A FDR-controlling procedure is designed to keep the FDR, $E\left[\frac{V}{R}\right]$ under a given threshold. FDR-controlling procedures are less conservative compared to familywise error rate (FWER) controlling procedures (such as the Bonferroni correction), which control the probability of **at least one** Type I error. Thus, FDR-controlling procedures have greater power, at the cost of increased rates of Type I errors.

Large data sets and multiple comparisons

Despite adjusting significance levels for multiple comparisons, if a dataset has a large number of observations (large n), it is still highly likely that many significant relationships will be found. Therefore, as discussed above, it is particularly important to consider effect size when making inferences from large data sets.