Assumptions of the two-sample t-test

- observations chosen at random from the two treatment levels for each of the two samples.
- data are continuous
- populations are distributed normally
- the variances in the two treatment levels are equal

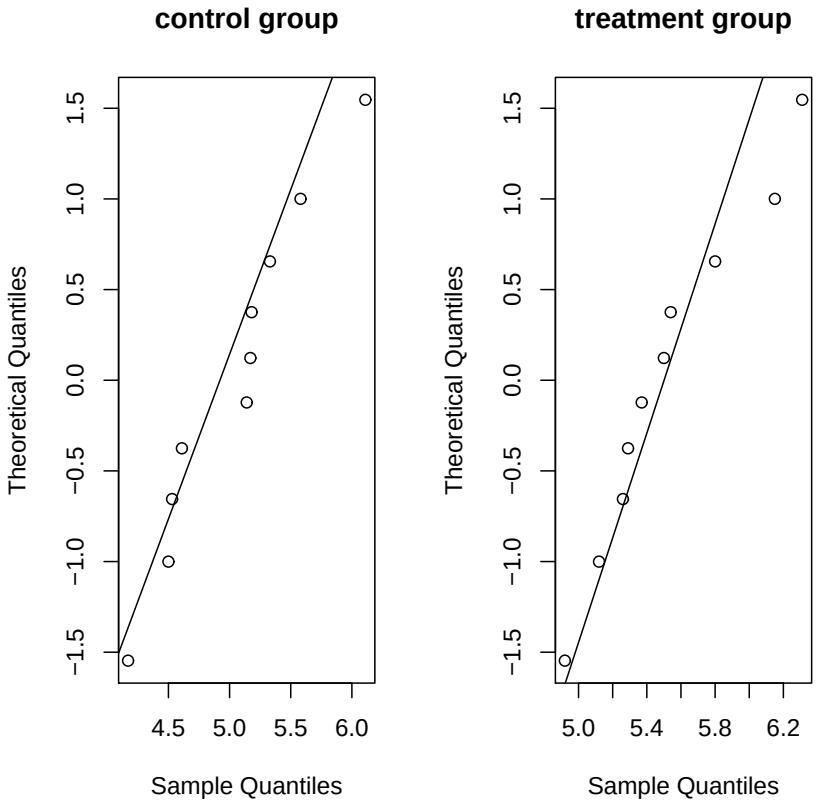
Normality

How do you test for normality?

- graphically
- statistical tests
  - Shapiro-Wilks (fairly powerful relative to others)
  - Kolmogorov-Smirnov (mostly historic)
  - several others

Unfortunately all statistical tests have a high typeII error rate if n<10

Normal QQ



Shapiro-Wilks

Measures how well the normal QQ plot adheres to a predicted line

```
> shapiro.test(X)

        Shapiro-Wilk normality test

data:  X
W = 0.9567, p-value = 0.7475

> shapiro.test(Y)

        Shapiro-Wilk normality test

data:  Y
W = 0.941, p-value = 0.5643
```

0.1 More t-tests

Paired t-test

So far we have considered two randomly drawn samples when performing a t-test.
In certain circumstances, the two samples are related as pairs, such as the same animals before (X) and after (Y) exposure to hypoxic conditions.

$$\hat{X} = (X_i - \bar{X}), \hat{Y} = (Y_i - \bar{Y})$$

$$t = (\bar{X} - \bar{Y}) \sqrt{\frac{n(n-1)}{\Sigma(\hat{Y} - \hat{X})^2}} \tag{1}$$

This only works with paired data. Of course paired data should not be analyzed by a normal Students t-test, the two samples are not independent.

Cedar apple rust

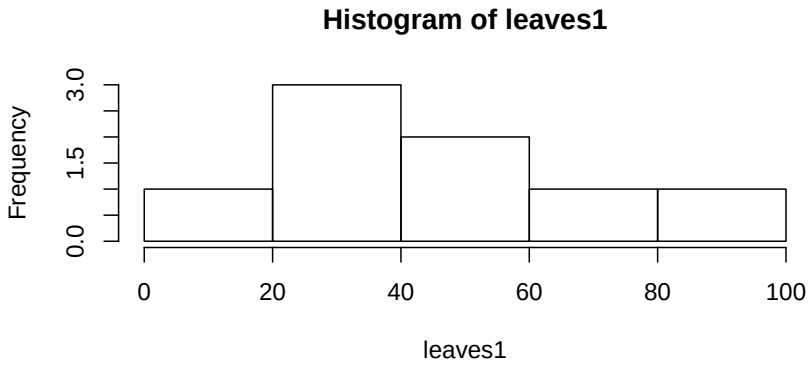
Paired data

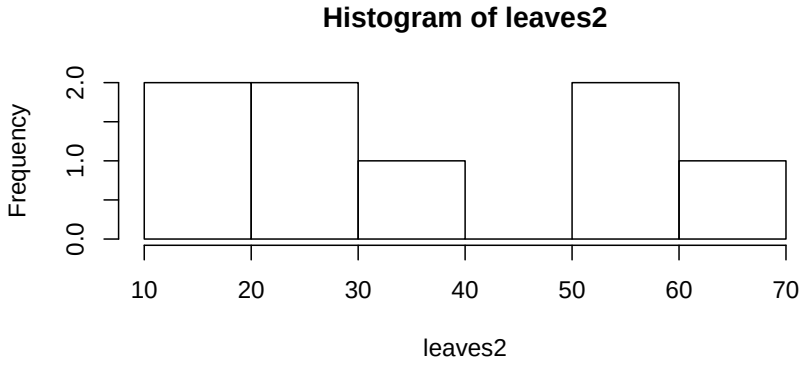
These data come from cedar apple rust experiments. The variables leaves1 and leaves2 refer to the number of infected leaves on apple trees the year before and after cedar trees are removed from the vicinity.

```
> cedar <- read.table("cedar.dat", header = T)
> cedar
```

	tree	leaves1	leaves2	diff
1	1	38	32	6
2	2	10	16	-6
3	3	84	57	27
4	4	36	28	8
5	5	50	55	-5
6	6	35	12	23
7	7	73	61	12
8	8	48	29	19

Paired t-test





Paired t-test (regular t-test)

```
> attach(cedar)
> t <- (mean(leaves1) - mean(leaves2))/((var(leaves1)/length(leaves1)) +
```

```
+     (var(leaves2)/length(leaves2)))^0.5
> df <- 2 * (length(leaves1) - 1)
> alpha <- 2 * (1 - pt(abs(t), df))
> t

[1] 0.989871

> alpha

[1] 0.3390378

> detach(cedar)
```

Paired t-test, execution

$$\hat{X} = (X_i - \bar{X}), \hat{Y} = (Y_i - \bar{Y})$$

$$t = (\bar{X} - \bar{Y}) \sqrt{\frac{n(n-1)}{\Sigma(\hat{Y} - \hat{X})^2}}$$

```
> attach(cedar)
> xhat <- leaves1 - mean(leaves1)
> yhat <- leaves2 - mean(leaves2)
> df <- length(leaves1) - 1
> meandif <- mean(leaves1) - mean(leaves2)
> t <- meandif * ((df * (df + 1))/sum((yhat - xhat)^2))^(0.5)
> alpha <- 2 * (1 - pt(abs(t), df))
> t

[1] 2.434162

> alpha

[1] 0.04514399

> detach(cedar)
```

Paired t-test

```
> attach(cedar)
> t.test(leaves1, leaves2, paired = T)
```

```
        Paired t-test

data:  leaves1 and leaves2
t = 2.4342, df = 7, p-value = 0.04514
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.2999574 20.7000426
sample estimates:
mean of the differences
                  10.5
```

```
> detach(cedar)
```

1 Introduction to Anova

Anova defined

Anova stands for **An**alysis **o**f **va**riance.
In general, Anova is used to detect the effect of some environment or experimental factor on the average response of a dependent variable.
The Anova framework is very flexible and can be applied to a large variety of experimental situations

1.1 Applications

Uses

- Analyzing experimental results where there are:
  - multiple levels of treatment per factor
    - ordered treatments (levels of copper in *Artemia* experiment)
    - unordered treatments (different types of webworm control on *Amaranthus*
  - multiple factors (independent variables)
    Polymerase chain reaction optimization (Mg++ and Temp)

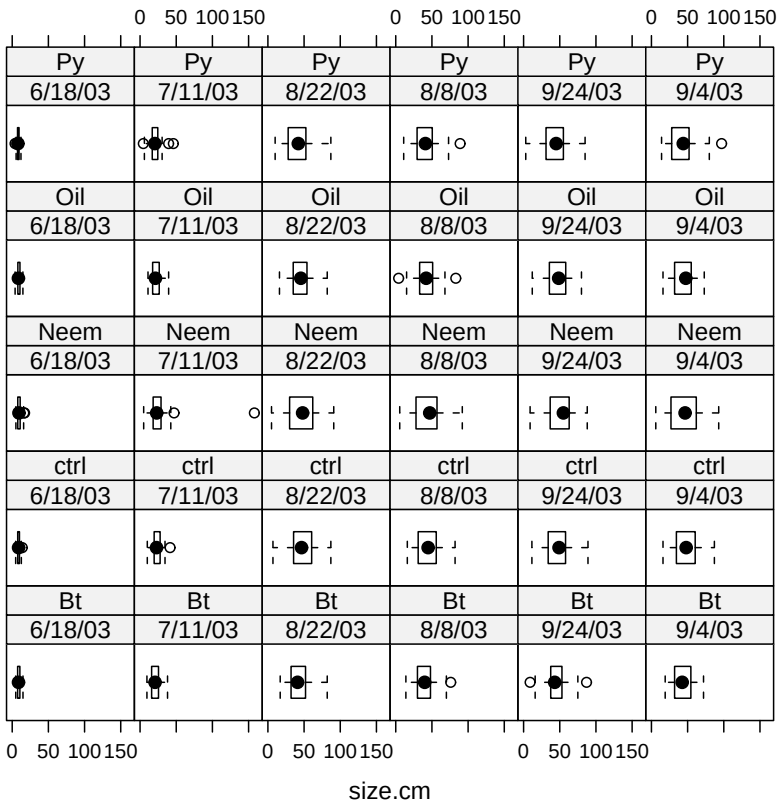
One-way Anova, two levels

Probably the simplest anova has a single factor with two levels. Call the levels treat and control.
In this case:

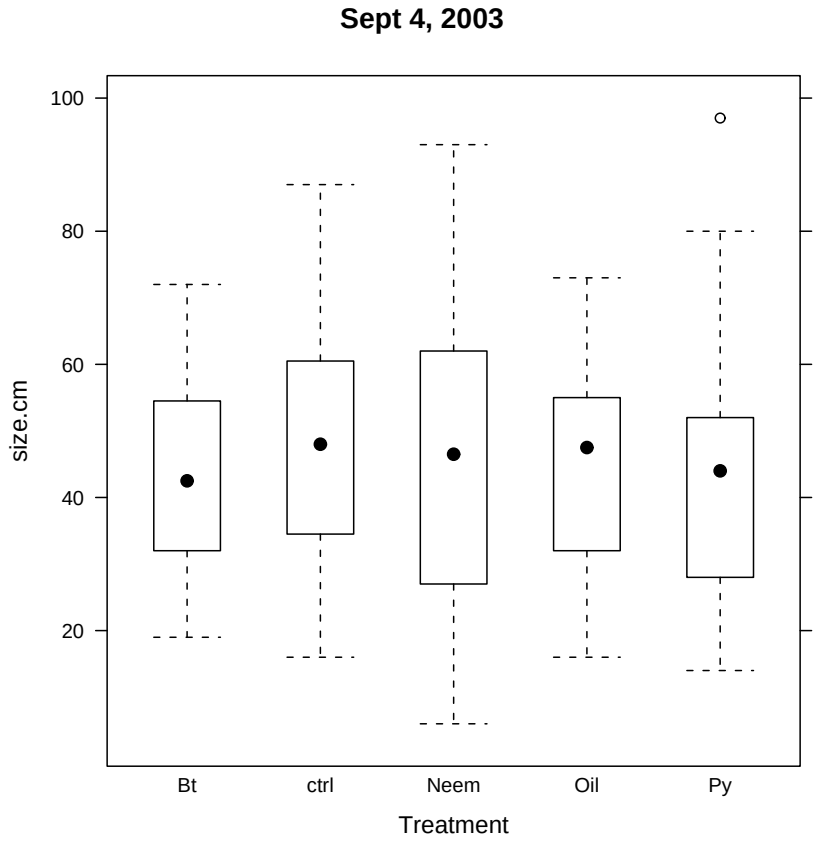
- $H_0 : \mu_{\text{treat}} = \mu_{\text{control}}$

- $H_1 : \mu_{\text{treat}} \neq \mu_{\text{control}}$

One-way Anova, multiple levels



One-way Anova, multiple levels



One-way Anova, multiple levels

One-way Anova can also be used with many levels of a factor. Here the multiple levels are: control, bt, neem, py, oil
In this case:

- $H_0 : \mu_{\text{bt}} = \mu_{\text{neem}} = \mu_{\text{py}} = \mu_{\text{oil}} = \mu_{\text{control}}$
- $H_1 : H_0^C$ If any mean differs H_0 rejected.

1.1.1 Why not use t-tests

Why not t-tests

The two-level one-way Anova is really testing the same hypothesis as the t-test.
It would certainly be possible to do many t-tests among all the treatments.
However as the number of levels increases, the number of pairwise comparisons increases more rapidly. ($\frac{k^2-k}{2}$)

- 2 levels = 1 test

- 3 levels = 3 tests
- 4 levels = 6 tests
- 8 levels = 28 tests

1.1.2 Multiple tests

Multiple test

As the number of tests increases the likelihood of type I error increases.
For example, if you perform 20 tests and use an α of 0.05 in each, you can expect one of these tests to reject the null even if it was true.

Linear model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{2}$$

- Y_{ij} is the observed value for observation j in treatment level i.
- μ is a grand mean for the entire experiment
- τ_i is an estimate of the deviation from the grand mean for a group
- ϵ_{ij} measures error for observation j in treatment level i. ($\epsilon\ N(0, \sigma^2)$)

Linear model-less formally

$$Y_{ij} = \bar{Y_{\cdot}} + (\bar{Y_{i-}} - \bar{Y_{\cdot}}) + (Y_{ij} - \bar{Y_{i-}}) \tag{3}$$

- Y_{ij} is the observed value for observation j in treatment level i.
- $\bar{Y_{\cdot}}$ is a grand mean for the entire experiment
- $(\bar{Y_{i-}} - \bar{Y_{\cdot}})$ is an estimate of the deviation from the grand mean for a group or treatment
- $(Y_{ij} - \bar{Y_{i-}})$ measures error for observation j in treatment level i.

1.2 Assumptions

Assumptions of Anova

There are several assumptions that Anova makes. Some are related to the experimental design, some assumptions apply to the data themselves

- type of effect (fixed or random)
- random association between experimental units and treatments
- ϵ independent among experimental units
- variance homogeneous across treatments.
- ϵ normally distributed with mean 0 and sd σ_ϵ
- terms in the model are additive

Some of these assumptions can be tested before analysis others must wait until analyses are performed.

1.2.1 Fixed versus random effects

Fixed versus random effects

If the levels of the treatment are chosen *a priori*, any effect observed for that factor is known as a *fixed effect* .
If the levels are chosen at random, the effects observed for that factor are known as *random effects*
There is little difference in how the data are analyzed. There is a major difference in how the results are interpreted.

1.2.2 Independence

Random assignment and Independence

The assumption that experimental units are randomly assigned to the treatments and that experimental units are maintained as independent entities falls under experimental design:
Things that can affect independence

- physical proximity in a greenhouse, field, culturing facility
- choosing experimental units in a non-random fashion (choosing robust plants for an experiment)
- Choosing only smokers to participate in an effects of smoking study.

1.2.3 Normality

Testing for normality

Same approaches as used in t-test. Same problems with those approaches. Anova may be relatively *robust* to violations of the normality assumption. Especially, normality of the ϵ is the distribution of concern not the original data.

- normal QQ plots
- Shapiro-Wilks

2 Calculations

Sums of Squares

$$Y_{ij} = \bar{Y_{\cdot}} + \underbrace{(\bar{Y_{i-}} - \bar{Y_{\cdot}})} + \underbrace{(\bar{Y_{ij}} - \bar{Y_{i-}})} \tag{4}$$

- $\text{SSD}_W = \Sigma_i \Sigma_j (Y_{ij} - \bar{Y_{i-}})^2$
- $\text{SSD}_B = \Sigma_i n_i (\bar{Y_{i-}} - \bar{Y})^2$
- $\text{SSD}_{\text{total}} = \text{SSD}_B + \text{SSD}_W = \Sigma_i \Sigma_j (Y_{ij} - \bar{Y})^2$

Mean Square

The Sums of Squares can be normalized by degrees of freedom. The resulting quantities estimate variances and components of variance and are called *mean square* terms.

- $\text{MS}_W = \frac{\text{SSD}_W}{N-k}$
- $\text{MS}_B = \frac{\text{SSD}_B}{k-1}$

Degrees of freedom

The degrees of freedom depend on the number of experimental units and the levels of the treatment.

- For MS_W, $df = N - k$
- For MS_B, $df = k - 1$

$$Y_{ij} = \bar{Y_{\cdot}} + \underbrace{(\bar{Y_{i-}} - \bar{Y_{\cdot}})} + \underbrace{(\bar{Y_{ij}} - \bar{Y_{i-}})} \tag{5}$$

The MS_W term estimates the variance in the trait (σ^2) in a population.

- If there is no effect for the treatment, the MS_B also equals (σ^2).
- If there is an effect for the treatment, the MS_B is greater than (σ^2).

The F-ratio is the test-statistic ($\mathrm{MS}_B/\mathrm{MS}_W$)

The F-ratio can then be used to estimate Type I error with 'pf()'.

```
> pf(1, df1 = 2, df2 = 10)

[1] 0.5981224

> pf(2, df1 = 2, df2 = 10)

[1] 0.8140656

> pf(3, df1 = 2, df2 = 10)

[1] 0.9046326

> pf(4, df1 = 2, df2 = 10)

[1] 0.9470779

> pf(5, df1 = 2, df2 = 10)

[1] 0.96875

> pf(6, df1 = 2, df2 = 10)

[1] 0.9805962
```