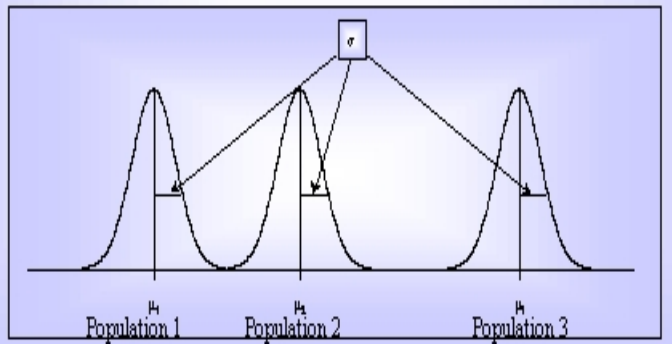
ANOVA

- Stands for ANalysis Of VAriance
- But it's a test of differences in *means*
- The idea:



The Observations y_{ij}

Treatment group

i = 1	i = 2	…	i = k
y_{11}	y_{21}	…	$y_{k,1}$
y_{12}	y_{22}	…	$y_{k,2}$
…	…	…	…
$y_{1,n1}$	$y_{2,n2}$	…	$y_{k,nk}$

means: m_1 m_2 … m_k

The ANOVA table

- The analysis is usually laid out in a table
- For a one-way layout (where the response is assumed to vary according to grouping on one factor):

Source	df	SS	MS	F	p-val
Treatment	k-1	$\Sigma(m_i-m)^2$	SST/(k-1)	MST/MSE	*
Error	n-k	$\Sigma(y_{ij}-m_i)^2$	SSE/(n-k)		
Total	n-1	$\Sigma(y_{ij}-m)^2$			

m = overall mean, m_i = mean within group *i*

Sum of Squares

- The two-sample *t*-test tests for equality of the means of two groups.
- We could express the observations as:

$$X_{ij} = \mu_i + E_{ij} \qquad i = 1,2$$

- Where the E_{ij} are assumed to be $N(0,\sigma^2)$
- H: $\mu_1 = \mu_2$

Sum of Squares

- This can also be written as:

$$X_{ij} = \mu + \alpha_i + E_{ij} \qquad i = 1,2$$

  - μ could be seen as overall mean
  - α_i as deviation from μ in group i
- Model is *overparameterized*
  - Uses more parameters than necessary
  - Necessitates some constraint, *e.g.* $\Sigma_i \alpha_i = 0$

Sum of Squares

- Goal: test difference between means of two (or more) groups
  - *Between SS* measures the difference
- The difference must be measured *relative* to the variance *within* the groups
  - *Within SS*
- *F-test:* considers the ratio of B/W
- The larger *F* is, the more significant the difference

The ANOVA Procedure

- Subdivide observed total sum of squares into several components
- Pick appropriate significance point for a chosen Type I error *a* from an *F* table
- Compare the observed components to test the NULL hypothesis

Comments

- Generalizes to any number of groups
- ANOVAs can be classified in various ways, e.g.
  - *fixed effects* models
  - *mixed effects* models
  - *random effects* model
  - For now we consider *fixed effect models*
    - Parameter a_i is fixed, but unknown, in group *i*

$$X_{ij} = \mu + \alpha_i + E_{ij}$$

One-Way ANOVA

- One-Way fixed-effect ANOVA
- Setup and derivation
  - Like two-sample *t*-test for *g* number of groups
  - Observations (n_i observations, i=1,2,…,g)

$$X_{i1}, X_{i2}, \ldots, X_{in}$$

  - Using overparameterized model for X

$$X_{ij} = \mu + \alpha_i + E_{ij} \qquad j = 1,2,\ldots,n_i \qquad i = 1,2,\ldots,g$$

  - E_{ij} assumed $N(0,\sigma^2)$, $\Sigma n_i a_i = 0$, a_i fixed in group i

One-Way ANOVA

  - Null Hypothesis H_0 is: $a_1 = a_2 = \ldots = a_g = 0$
  - Total sum of squares is

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - \overline{\overline{X}})^2$$

  - This is subdivided into B and W

$$B = \sum_{i=1}^{g} n_i(\overline{X}_i - \overline{\overline{X}})^2 \qquad W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$$

  - with

$$\overline{X}_i = \sum_{j=1}^{n_i}\frac{X_{ij}}{n_i} \qquad \overline{\overline{X}} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\frac{X_{ij}}{N} \qquad N = \sum_{i=1}^{g} n_i$$

One-Way ANOVA

- Total degrees of freedom: $N - 1$
  - Subdivided into $df_B = g - 1$ and $df_W = N - g$
- This gives the test statistic F

$$F = \frac{B}{W} * \frac{N - g}{g - 1}$$

Assumptions

- Have *random samples* from each separate population
- The *variance is the same* in each treatment group
- The samples are *sufficiently large* that the CLT holds for each sample mean (or the individual population distributions are normal)

What does it mean when we reject H?

- The null hypothesis H is a *joint* one: that *all* population means are equal
- When we reject the null, that does *NOT* mean that the means are all different!
- It means that *at least one* is different
- To find out *which* is different, can do *post hoc* testing (pairwise *t*-tests, for example)

Additional aspects

- Why not start off doing separate (z or t) tests for each pair of samples? *...*
- Testing the assumptions
- *Which* mean(s) is/are not equal
  - can do *post hoc* testing (pairwise *t*-tests, for example)
- Multiple comparisons (multiple testing)
- 'Data snooping'

Factorial crossing

- Compare 2 (or more) sets of conditions in the *same experiment*
- Designs with factorial treatment structure allow you to measure *interaction* between two (or more) sets of conditions that influence the response – you will look at this in more detail during the exercises today
- Factorial designs may be either observational or experimental

3 types of 2-factor factorial designs

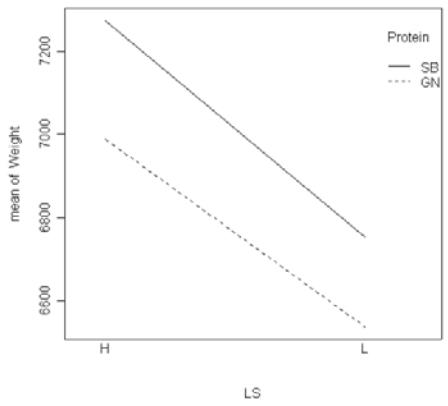
- 2 experimental factors – you *randomize* treatments to each unit
- 2 observational factors – you *cross-classify* your populations into groups and get a sample from each population
- 1 experimental and 1 observational factor – you *get a sample* of units from each population, *then use randomization* to assign levels of the experimental factor (treatments), separately within each sample

Interaction

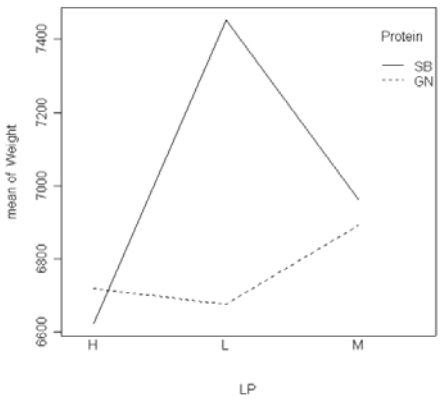
- Interaction is very common (and very important) in science
- Interaction is a *difference of differences*
- Interaction is present if the effect of one factor *is different* for different levels of the other factor
- *Main effects can be difficult to interpret in the presence of interaction*, because the effect of one factor depends on the level of the other factor

Interaction plot

no interaction



interaction



Two-Way ANOVA

- Two-Way Fixed Effects ANOVA
- More complicated setup; example:
  - Expression levels of one gene in lung cancer patients
  - *a* different risk classes
    - E.g.: ultrahigh, very high, intermediate, low
  - *b* different age groups
  - *n* individuals for each risk/age combination

Two-Way ANOVA

- Observations: X_{ijk}
  - *i* is the risk class ($i = 1, 2, \ldots, a$)
  - *j* indicates the age group
  - *k* corresponds to the individual in each group ($k = 1, \ldots, n$)
    - Each group is a possible risk/age combination
  - The number of individuals in each group is the same, *n*
  - This is a "balanced" design (equal numbers in each group
  - Theory for unbalanced designs is a little more complicated

Two-Way ANOVA

- The model for each X_{ijk} is

$$X_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + E_{ijk}$$

$$i = 1,2,\ldots,a \qquad j = 1,2,\ldots,b \qquad k = 1,2,\ldots,n$$

  - E_{ijk} are N(0, σ^2)
  - The mean of X_{ijk} is $\mu + \alpha_i + \beta_i + \delta_{ij}$
  - α_i additive for risk class i
  - β_i additive for age group i
  - δ_{ij} risk/age interaction parameter
    - Should be added if a possible group/group interaction exists

Two-Way ANOVA

- Constraints:
  - $\Sigma_i \alpha_i = \Sigma_i \beta_i = 0$
  - $\Sigma_i \delta_{ij} = 0$ for all j
  - $\Sigma_j \delta_{ij} = 0$ for all i
- The total sum of squares is then subdivided into four groups:
  - Risk class SS
  - Age group SS
  - Interaction SS
  - Within cells ("residual" or "error") SS

Two-Way ANOVA

- Associated with each sum of squares
  - Corresponding *degrees of freedom (df)*
  - Corresponding *mean square (MS)*
    - Sum of squares divided by degrees of freedom
- The mean squares are compared using *F* ratios to test various effects
  - First – test for a significant risk/age *interaction*
  - If there is an interaction, it may not be reasonable to test for significant risk or age differences

Multi-Way ANOVA

- One-way and two-way fixed effects ANOVA can be extended to *multi-way ANOVA*
- Example: four-way ANOVA (*saturated*) model:

$$X_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\varepsilon)_{il} + (\beta\gamma)_{jk} + (\beta\varepsilon)_{jl} + (\gamma\varepsilon)_{kl} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\varepsilon)_{ijl} + (\beta\gamma\varepsilon)_{jkl} + (\alpha\beta\gamma\varepsilon)_{ijkl}$$

- One observation per cell
- In general, interested in *unsaturated* models

Model formulas in R

- A simple *model formula* in R looks something like: `yvar ~ xvar1 + xvar2 + xvar3`
- We could write this model (algebraically) as

  $Y = a + b_1*x_1 + b_2*x_2 + b_3*x_3$
- By default, an intercept is included in the model - you don't have to include a term in the model formula
- If you want to leave the intercept out:

```
yvar ~ -1 + xvar1 + xvar2 + xvar3
```

More on model formulas

- We can also include *interaction terms* in a model formula:

```
yvar ~ xvar1 + xvar2 + xvar3
```

Examples

- `yvar ~ xvar1 + xvar2 + xvar3 + xvar1:xvar2`
- `yvar ~ (xvar1 + xvar2 + xvar3)^2`
- `yvar ~ (xvar1 * xvar2 * xvar3)`

More on model formulas

- The generic form is **`response ~ predictors`**
- The predictors can be **`numeric`** or **`factor`**
- Other symbols to create formulas with *combinations of variables* (e.g. *interactions*)
  - **`+`** to *add* more variables
  - **`-`** to *leave out* variables
  - **`:`** to introduce *interactions* between two terms
  - **`*`** to include *both interactions and the terms*
    (**`a*b`** is the same as **`a+b+a:b`**)
  - **`^n`** *adds all terms* including interactions up to order n
  - **`I()`** treats what's in () as a *mathematical expression*

Interpreting R output

```
> chicks.aov <- aov(Weight ~ House + Protein*LP*LS)
> summary(chicks.aov)
              Df  Sum Sq Mean Sq F value    Pr(>F)
House          1  708297  708297 15.8153 0.0021705 **
Protein        1  373751  373751  8.3454 0.0147366 *
LP             2  636283  318141  7.1037 0.0104535 *
LS             1 1421553 1421553 31.7414 0.0001524 ***
Protein:LP     2  858158  429079  9.5808 0.0038964 **
Protein:LS     1    7176    7176  0.1602 0.6966078
LP:LS          2  308888  154444  3.4485 0.0687641 .
Protein:LP:LS  2   50128   25064  0.5596 0.5868633
Residuals     11  492640   44785
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

Numerical and graphical analysis

- Tables of group means:

		Groundnut	Soybean	Mean
Level of protein	0	6676	7452	7064
	1	6893	6961	7927
	2	6719	6624	6671
Mean		6763	7012	6887

		G-nut	Soy	Level of protein			Mean
				0	1	2	
Level of fish	0	6537	6752	6750	6595	6588	6644
	1	6989	7273	7379	7259	6755	7131
Mean		6763	7012	7064	6927	6671	6887

Numerical and graphical analysis

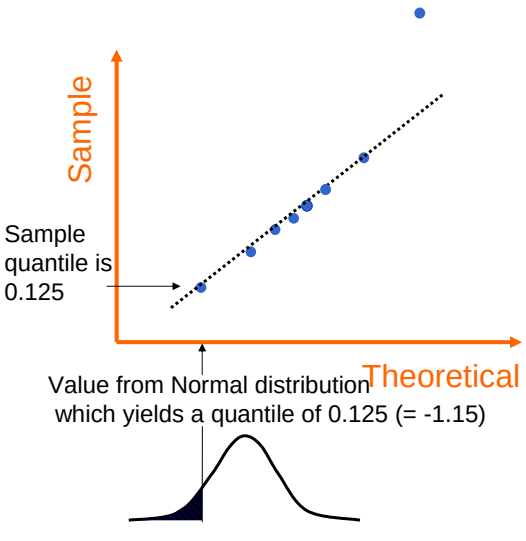
- Design plot
- Boxplots of outcome for each factor
- Interaction plots
- Write out model, assumptions, define all parameters
- anova table
- Plots for assumption checking/model assessment

Model assessment: Normality

- Boxplots of observations (or residuals) should be *symmetric*
- Plot of sample means *vs* sample variances *should not show a pattern*
- *QQ normal plots* of observations (or residuals) should be a *straight line*
- Check for *outliers*

QQ-Plot

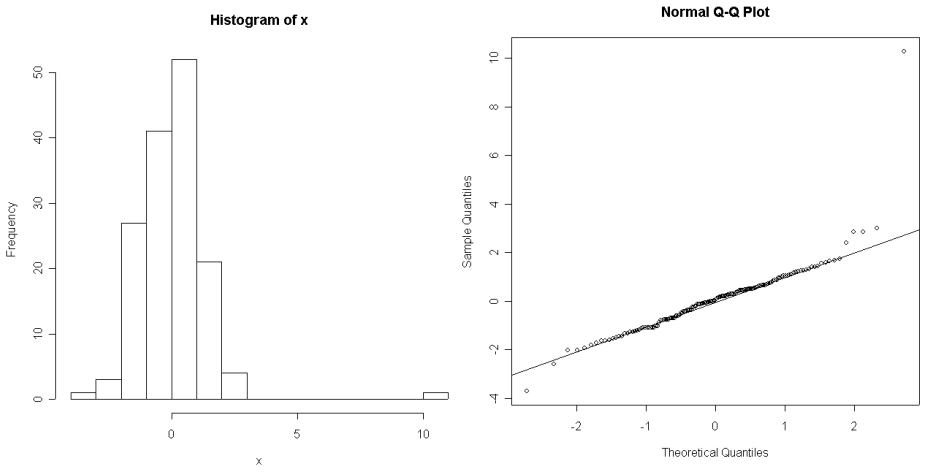
- Quantile-quantile plot
- Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples)
- A method for looking for outliers when data are mostly normal



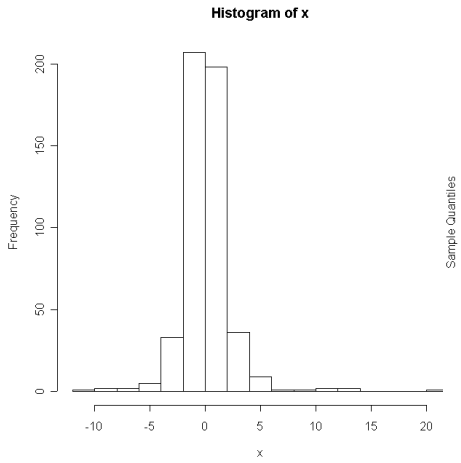
Typical deviations from straight line patterns

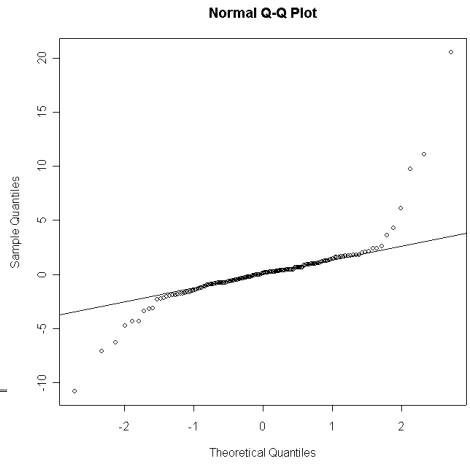
- Outliers
- Curvature at both ends (long or short tails)
- Convex/concave curvature (asymmetry)
- Horizontal segments, plateaus, gaps

Outliers

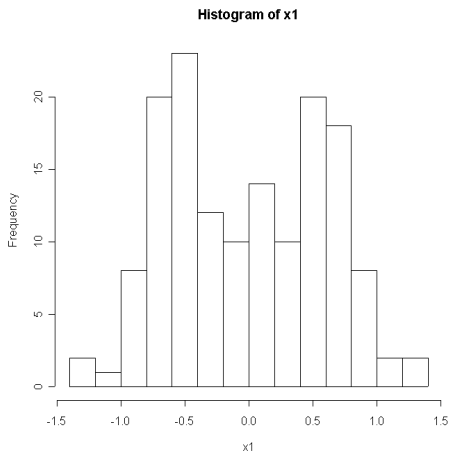


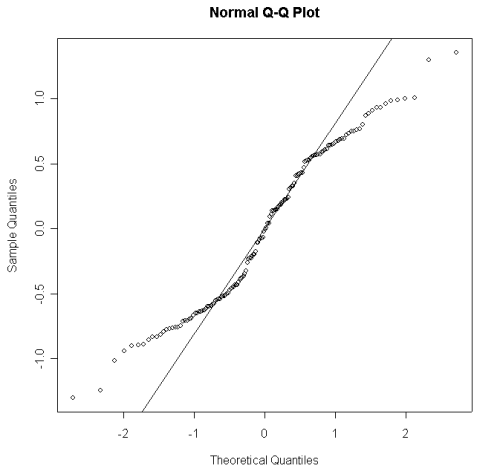
Long Tails



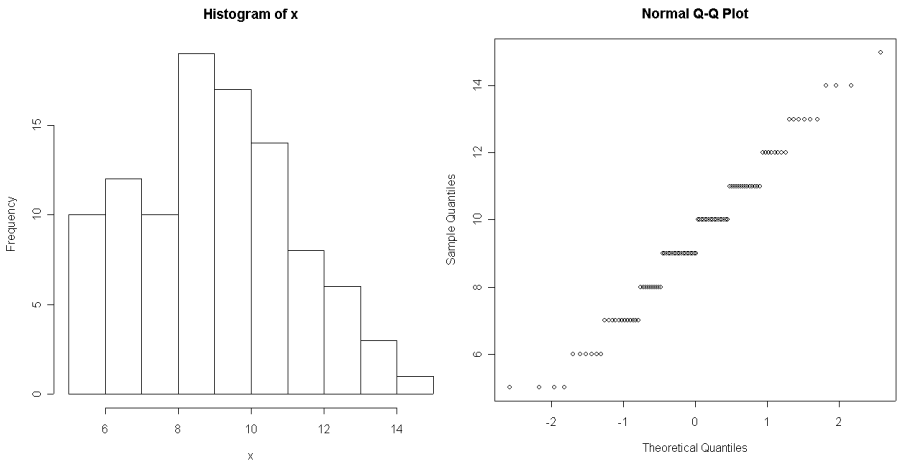


Short Tails





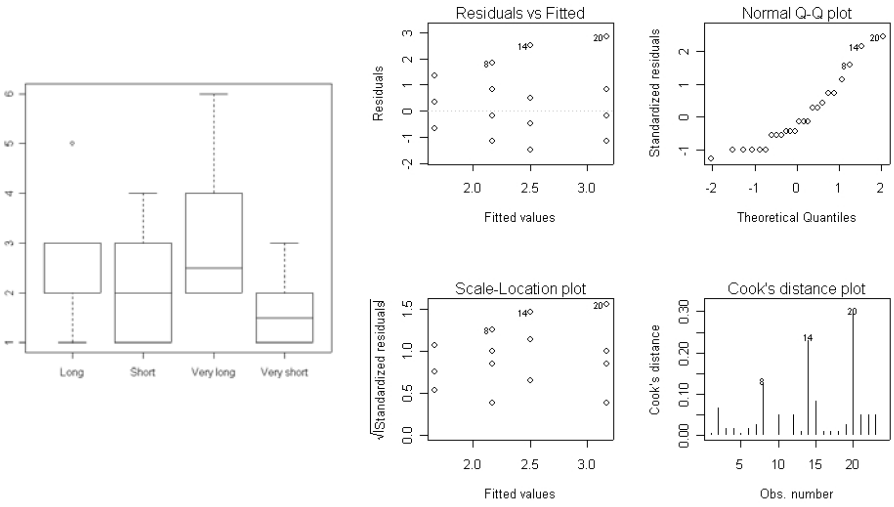
Plateaus/Gaps



Model assessment: Variance homogeneity

- Boxplots of observations should have *similar spread*
- Spread of residuals should be similar when plotted against group means
- There are also *formal tests* (*e.g.*, Bartlett, Levene) but these are not so useful for *diagnosis*

Diagnostic plots



Model assessment: Independence

- Plot residuals against group means, might indicate *e.g.* autocorrelation
- Usually need to deal with the independence issue at the *design stage*, through randomization or other means

chicks.dat Demo

- …