Two Group Comparisons and Beyond

Rick White

February 3, 2016

Outline

- Some Theory
- Testing the location parameter (t-tests)
- More than two groups (ANOVA)
- Questions

Our data

We assume our data is drawn at random from a probability distribution.

The data have a mean and a variance and variables can be correlated.

If X and Y are our samples from 2 groups .

- The means are denoted by μ_x and μ_y
- The variances are denoted by σ_x^2 and σ_y^2
- The covariance is denoted by σ_{xy}
- The correlation is $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ is a number between -1 and 1.

Properties of the mean and variance

The mean of the sum is the sum of the mean.
$\mu_{x+y} = \mu_x + \mu_y$

The mean of the difference is the difference of the mean.
$\mu_{x-y} = \mu_x - \mu_y$

The variance of the sum is usually **not** the sum of the variance.
$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$

The variance of the difference is **not** the difference of the variances.
$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}$

If X and Y are independent then the covariance (σ_{xy}) and correlation (ρ_{xy}) $= 0$

The 2 group experimental setup

We have a random sample of data from 2 populations (observational study) or from a population randomized into 2 groups (controlled experiment).

We measure a variable of interest on each member of the sample and want to determine if the mean of that variable is different in the two groups.

Group 1: $X = x_1, \ldots, x_n$ are iid $F_1(\mu_x, \sigma_x^2)$

Group 2: $Y = y_1, \ldots, y_m$ are iid $F_2(\mu_y, \sigma_y^2)$

iid means **Independent** Identically Distributed

Estimating the parameters from the sample

We use the sample average to estimate the group mean.

$\hat{\mu}_x = \bar{x} = \sum_1^n x_i/n$

We use the sample variance to estimate the group variances.

$\hat{\sigma}_x^2 = s_x^2 = \sum_1^n (x_i - \bar{x})^2/(n-1)$

Usually the groups are independent samples which means the data are independent.

If the data are paired ($n = m$), we can estimate the covariance between the variables which allows us to compute the correlation.

$\hat{\sigma}_{xy} = s_{xy} = \sum_1^n (x_i - \bar{x})(y_i - \bar{y})/(n-1)$

Significance testing

We compare the means of the two group by a hypothesis test.

A hypothesis test consists of a Null hypothesis (H_0) and an alternative hypothesis (H_1).

We compute a test statistic and a p-value based on our data. If our p-value is less than α we reject the Null hypothesis in favour of the alternative (H_1).

α specifies the chance of a Type 1 error or a false positive result. If we set $\alpha = 0$, we will never reject H_0.

Sample Size and Power calculations

α (or significance) is the probability of rejecting H_0 when it is true. It does not depend on the sample size.

$1 - \beta$ (or power) is the probability of rejecting H_0 when it is false. As N increases so does the power.

Power or sample size calculations require you to fully specify the true parameters of the model.

2 sample t-test with equal variance
$n = (2\sigma^2/\Delta^2) * (z_{(1-\alpha/2)} + z_{(1-\beta)})^2$

Software to calculate sample size or power is available. For a list of software see Wikipedia (**Statistical Power**).

The distribution of the sample mean.

With a large sample size the sample average will converge to a normal distribution (bell curve) for almost any distribution of the original data.

$\sqrt{n}(\bar{x} - \mu_x) \xrightarrow{d} N(0, \sigma_x^2)$

How quickly it converges depends on the distribution of data.

On line Example

Since the t-test is based on sample averages, this property makes it very robust to the normality assumption if the sample size is reasonably large.

t-test for two uncorrelated samples

Stated assumptions:

- Data are normally distributed
- equal variance in the two groups
- data are independent

The Null Hypothesis is the group means are equal ($H_0 : \mu_x = \mu_y$).

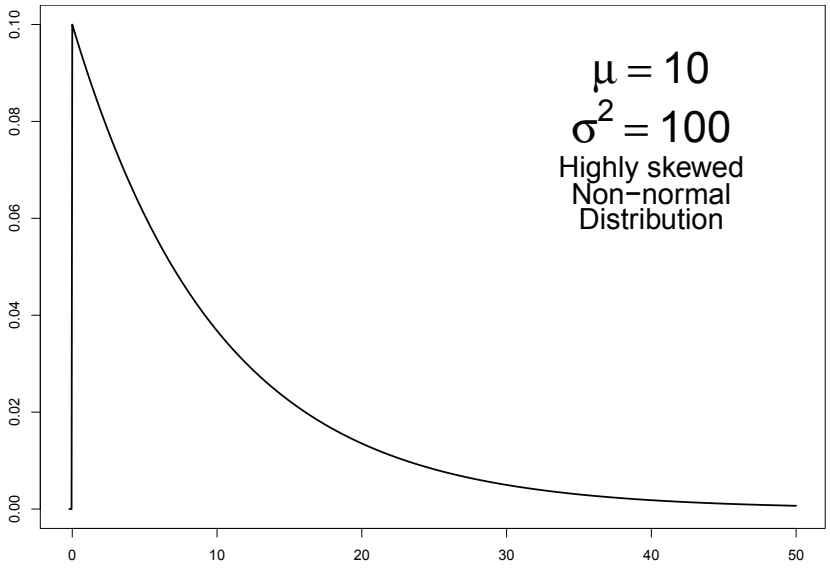
The alternative is usually $H_1 : \mu_x \neq \mu_y$.

What happens if the assumptions are violated?

We can test this by simulating data and computing the p-value.

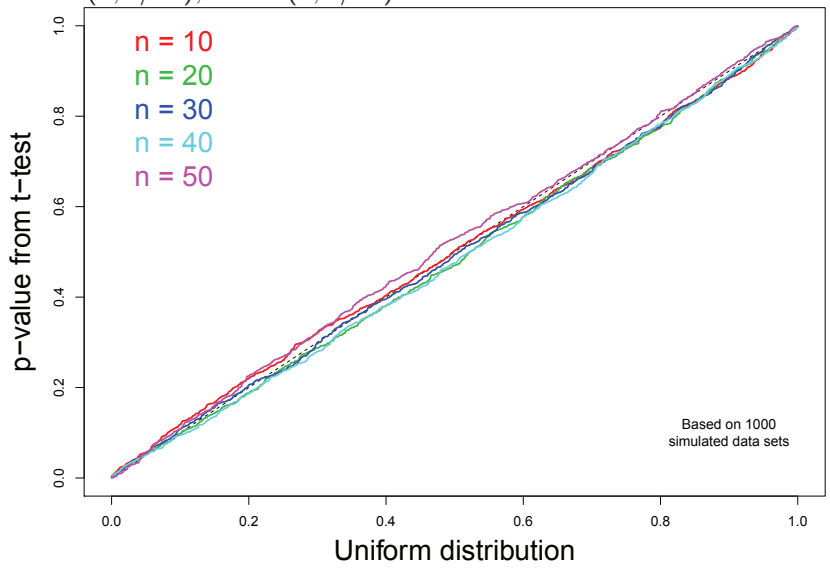
If we plot the p-values against the quantiles of a uniform distribution we should get a straight line.

This is a $\Gamma(1, 1/10)$ distribution



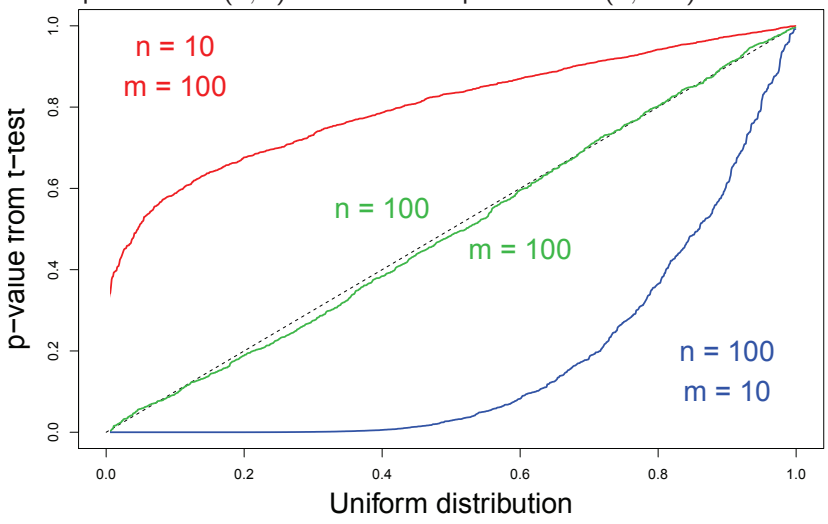
Normality assumption violated

$X \sim \Gamma(1, 1/10), Y \sim \Gamma(1, 1/10)$



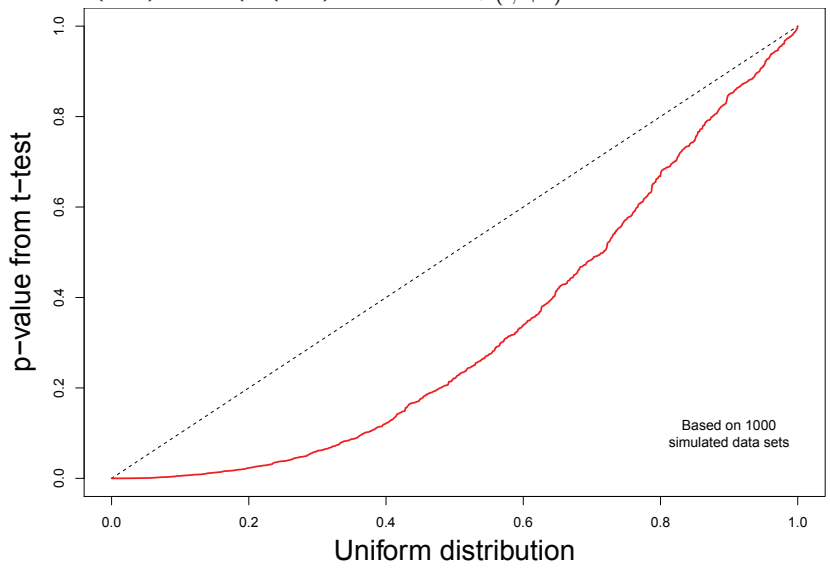
Equal variance assumption violated

n samples $X \sim N(0,1)$ versus m samples $Y \sim N(0,100)$



Independence assumption violated

$X \sim N(0,1)$, $Y \sim (N(0,1)$ correlation $\rho_{(i,i+1)} = 0.5$



Conclusions about 2 sample t-tests.

The independence assumption is critical for the t-test to be valid.

- If the data within a group are not independent then the dependence must be estimated and adjusted for.

The equal variance assumption is not critical if the sample size in each group is similar.

- If the variances and the sample sizes in the two groups are different, the **Welch's t-test** can used instead.

The normality assumption is not critical for the t-test and can essentially be ignored.

- Violation of this assumption can affect the power of the test.
- If data is skewed either transform or use an alternative test.

Paired t-test

Paired data means for each x there is a specific y related to it.

This usually means there is correlation ($\rho_{xy} \neq 0$).

We need to adjust for the correlation by using a paired t-test.

This should be a standard test in most statistical software.

If we take the difference in the observed values in each pair, a paired t-test becomes a one sample t-test.

We compute $z_i = x_i - y_i$ then test if $\mu_z = 0$.

Non-Parametric tests

Non-parametric tests can be used as an alternative to a two-sample or paired t-test.

- Two-sample t-test -> **Wilcoxon rank sum test**

Wilcoxon test is the same as **Mann-Whitney U test**

- Paired t-test -> **Wilcoxon signed rank test**

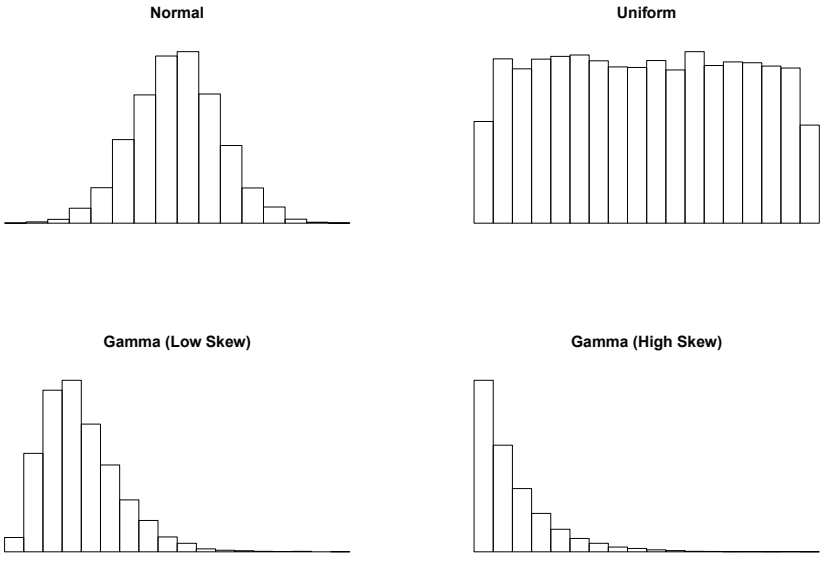
Wilcoxon tests makes similar assumptions as the t-test except for normality.

Two sample t-test versus wilcoxon rank sum test

The key factor in choosing between a t-test and a Mann-Whitney test is the statistical power under the alternative hypothesis.

- For symmetric data, including the normal distribution, the t-test is slightly more powerful than the Mann-Whitney test.
- If the data are skewed, the Mann-Whitney test can be substantially more powerful than the t-test.
- t-test only reject H_0 if the group means are different.
- The Mann-Whitney test can reject the Null for reasons other than a difference in the group means.

Sample distributions



Power of the two sample t-test vs Mann-Whitney

We can compare the power of the two tests by simulation.

In all distributions $\mu_1 - \mu_2 = 2$, $\sigma^2 = 100$ and $n = 250$ in each group.

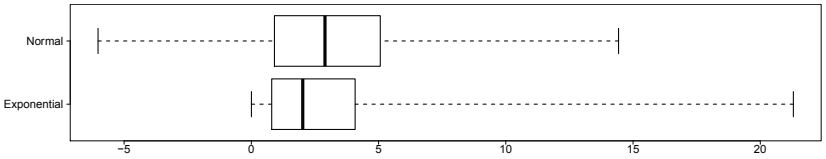
```
##      Dist     Shape T.test MW.test
## 1  Normal symmetric  0.598   0.571
## 2 Uniform symmetric  0.612   0.582
## 3   Gamma  low skew  0.628   0.694
## 4   Gamma high skew  0.616   0.886
```

Regardless of the distribution, the power of the t-test is about 60%,

The power of the Mann-Whitney test increases as the data become more skewed.

More about the Mann-Whitney test

Mann-Whitney may reject H_0 if the data come from different distributions even if the means are the same.




```
##          Dist     mean variance   median
## 1      Normal 2.965056 9.639457 2.894027
## 2 Exponential 2.982511 9.005277 2.020666
```

```
## t.test = 0.8982921 ,   wilcoxon = 0.004800793
```

Paired t-test versus wilcoxon signed rank test

Paired t-test compares the mean of the difference in the paired data.

Signed rank test compares the median of the difference in the paired data.

If the data are skewed, the mean is different from the median.

```
## mean(x) = -0.0744075, median(x) = 0
## mean(y) = 0.9284962, median(y) = 0
```

```
## If z = x-y then:
## mean(z) = -1.002904, median(z) = -0.1748501
## p-value from a paired t-test = 0.01515364
## signed rank test = 0.08973196
```

Testing the difference in the scale parameter

$H_0 : \sigma_x^2 = \sigma_y^2$ versus $H_1 : \sigma_x^2 \neq \sigma_y^2$

The **F test** is the ratio of two variance estimates.

Bartlett's test can test equality of variances in many groups.

The two tests above rely heavily on the normality assumption.

Levene's Test is less susceptible to the normality assumption and can test many groups.

Brown–Forsythe test is similar to Levene's test but is even more robust to the distributional assumptions.

Non-parametric tests are also available **Fligner-Killeen test**, **Ansari-Bradley test**, and **Mood test**.

One-Way ANOVA

One-way **AN**alysis **O**f **VA**riance (ANOVA) is a way to compare more than2 groups.

Y is a continuous response variable.

A is a categorical variable that indicates K distinct groups.

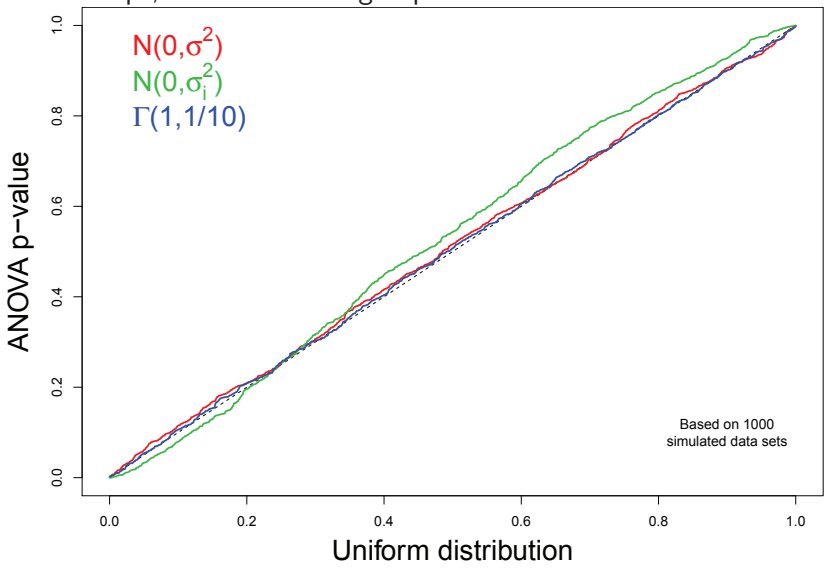
We assume Y_{ij} are independent $N(\mu + \alpha_i, \sigma^2)$

H_0 : all $\alpha_i = 0, H_1$: at least one $\alpha_i \neq 0$

Like the two sample t-test.
ANOVA is robust to the Normality assumption.
Balanced ANOVA is robust to unequal variance.
Independence is a very important assumption.

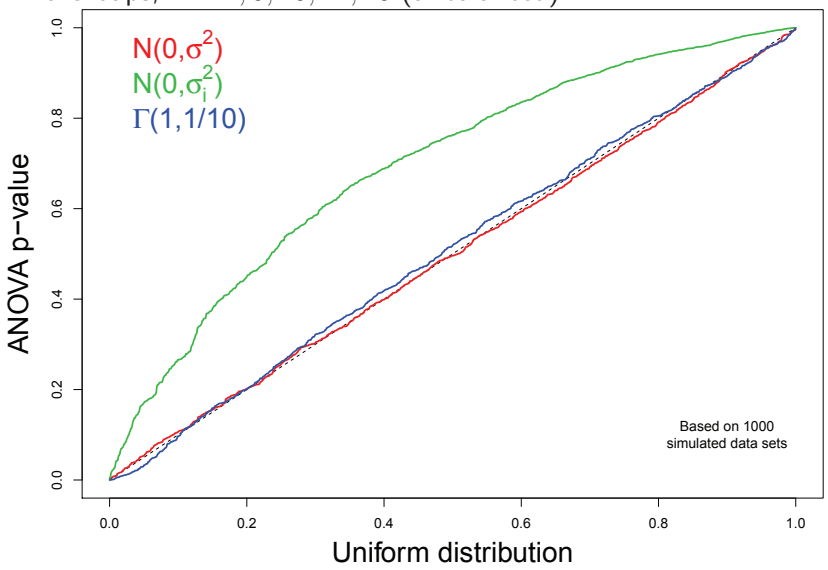
Normality and equal variance (balanced case)

Five Groups, $n = 10$ in each group



Normality and equal variance (unbalanced case)

Five Groups, $n = 2, 6, 10, 12, 20$ (unbalanced)



ANOVA parametrization

ANOVA requires a parametrization for the model.

The usual method is to select one of the groups to be the reference.

- All other levels of the factor are compared to this group.

With only 2 groups, the model computes

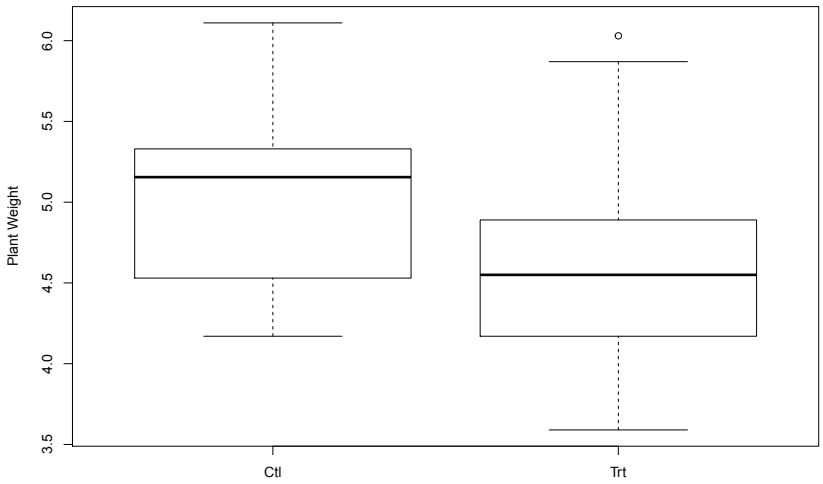
- the mean for the reference group
- the difference between the means of the treatment and reference groups.

We are usually only interested in the difference.

If there are more than 2 groups, none of the non-reference groups are compared to each other.

Example Plant Weight Data

With any analysis you should try plotting the data first. Here are boxplots of the data in each group.



One-Way ANOVA

ANOVA with 2 groups is equivalent to a two sample t-test.

```
## Response = Plant Weight
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        1  0.688  0.6882   1.419  0.249
## Residuals 18  8.729  0.4850
```

```
## Two sample t-test
## t = 1.19126 , df = 18 , p value = 0.2490232
```

Note: $\sqrt{1.419} = \pm 1.191$
In fact a $t_\nu^2 \sim F_{1,\nu}$.

Quantifying the difference

The model can fit the mean of each group.

```
##     Estimate Std.Error   2.5%  97.5%
## Ctl    5.032    0.2202 4.5693 5.4947
## Trt    4.661    0.2202 4.1983 5.1237
```

However the parameters the model uses are below.

```
##             Estimate Std.Error  tvalue Pr(>|t|)
## Ctl            5.032    0.2202 22.8501    0.000
## Trt - Ctl     -0.371    0.3114 -1.1913    0.249
```

Usually we only care about the difference in the group means.

Selecting parameters for the model

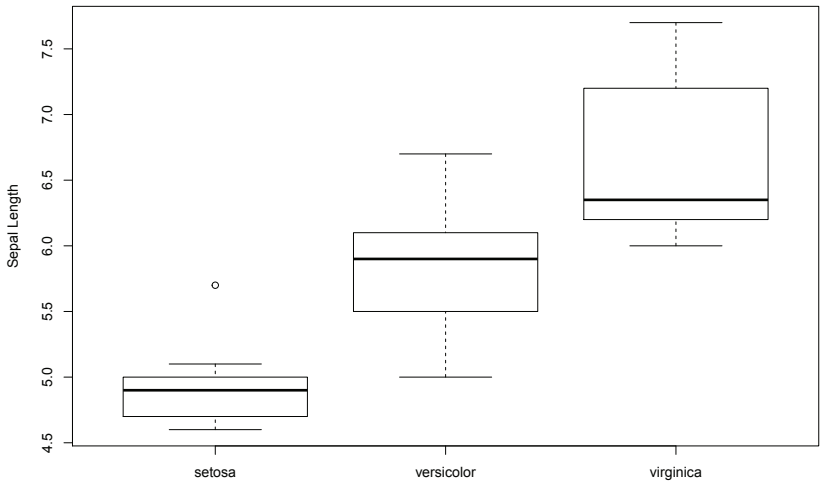
There are many possible parameter choices for an ANOVA model.

Important: Make sure you know which parametrization is used by your software before you interpret the output from the model.

- Most common method is to set a specific level of a factor as a reference.
- Alternately set the average effect over all the levels of the factor to be zero.

The model will have a single parameter to reflect a "typical" value and $K - 1$ parameters that model the deviations from this value.

ANOVA Example: Sepal length in 3 species of Iris



There is a clear indication of a difference in the sepal length between the species.

ANOVA results for the iris data

```
## ANOVA Table
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Species    2 13.789  6.8943  29.314 1.71e-07
## Residuals 27  6.350  0.2352
```

The ANOVA test for the significant differences between the 3 species has only 2 degrees of freedom.

This is the number of variables you need to estimate to compare all the level of the factor.

The residual term represents the error in the model.

The degrees of freedom for error mainly determines the power of the ANOVA test.

Estimated coefficients for Iris data

```
## Setting a reference group

##      Estimate Std. Error t value Pr(>|t|)
## MU       6.59      0.153   42.97 2.13e-26
## Var1    -1.66      0.217   -7.65 3.12e-08
## Var2    -0.79      0.217   -3.64 1.13e-03

## Making the sum of the group effects = 0

##      Estimate Std. Error t value Pr(>|t|)
## MU     5.7733     0.0885  65.205 3.05e-31
## Var1  -0.8433     0.1252  -6.735 3.13e-07
## Var2   0.0267     0.1252   0.213 8.33e-01
```

Without more information these parameters don't mean much.

Posthoc tests

If we look at the pairwise comparisons of the groups in either model the results are the same.

```
##                          Diff    SE Tstat
## versicolor - setosa      0.87 0.217  4.01
## virginica - setosa       1.66 0.217  7.65
## virginica - versicolor 0.79 0.217  3.64
```

We can compute p-values for these comparisons but need to be adjusted for multiple comparisons.

The amount of adjustment increases with the number of pairwise comparisons.

ANOVA with a blocking factor

ANOVAs may contain more than one factor.

Factors that are not of interest are considered blocking factors.

Blocking factors are primarily used to control other sources of error that otherwise might hide significant effects in our factor of interest.

Y is a continuous response, A is a factor with I levels, B is the blocking factor with J levels.

Y_{ijk} are independent $N(\mu_{ij}, \sigma^2)$

$\mu_{ij} = \mu + \alpha_i + \beta_j$

Our hypothesis of interest is
H_0 : all $\alpha_i = 0$, H_1 : at least one $\alpha_i \neq 0$

Example Students's Sleep data

The data shows the effect of two drugs (group) on the amount of extra sleep hours. The blocks are subjects (ID) who act as their own control.

```
##           Df Sum Sq Mean Sq F value  Pr(>F)
## group      1  12.48  12.482  16.501 0.00283
## ID         9  58.08   6.453   8.531 0.00190
## Residuals  9   6.81   0.756
```

If we ignore ID then we do not see a significant effect for group.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## group      1  12.48  12.482   3.463 0.0792
## Residuals 18  64.89   3.605
```

More on the block design

ANOVA with a blocking factor and 2 groups and only 1 observation per group in each block is equivalent to a paired t-test.

```
## Paired t-test
## t = -4.062128 , df = 9 , p value = 0.00283289
```

We can quantify the difference in the group means.

```
##              Estimate Std. Error t value Pr(>|t|)
## Block Design     1.58     0.3890   4.062 0.002833
## One-Way          1.58     0.8491   1.861 0.079187
```

Note the estimated difference is the same for the block design and the One-Way ANOVA but the standard error is different.

Non-parametric generalizations

Kruskal-Wallis is a rank based version of a One-Way ANOVA.

A Kruskal-Wallis test with only 2 groups is identical to a Mann-Whitney test.

Friedman rank-sum test is a non-parametric way to analyse unreplicated ($n = 1$) complete blocked data.

- If there are only 2 groups, it is not the same as a signed rank test.

Once you are outside these 2 special cases there are very few non-parametric methods available.

However we have seen an ANOVA model is still valid even if the normality assumption is violated.

Two-Way ANOVA

Two-Way ANOVA is similar to a block design except both factors are of interest and can interact with each other.

Y is a continuous response, A is a factor with K levels, B is a factor with J levels.

Y_{ijk} are independent $N(\mu_{ij}, \sigma^2)$

$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$

We are interested in testing
H_0 : all $\gamma_{ij} = 0, H_1$: at least one $\gamma_{ij} \neq 0$

If there is no interaction then we are interested in
H_0 : all $\alpha_i = 0, H_1$: at least one $\alpha_i \neq 0$
H_0 : all $\beta_j = 0, H_1$: at least one $\beta_j \neq 0$

Meaning of the interaction

An interaction between factor A and B means the effect of A depends on the level of B and the effect of B depends on the level of A.

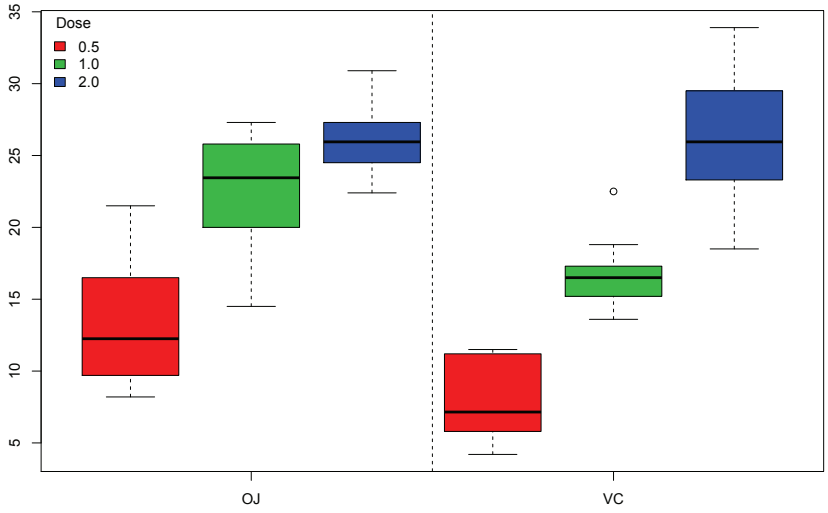
If an interaction between A and B is in the model then we cannot interpret the main effects of either factor.

A main effect need not be significant in the presence of an interaction term.

But the main effects must remain in the model otherwise we cannot interpret any of the interaction terms that involve that factor.

Example Tooth Growth in Guinea Pigs.

Treatments are Vitamin C dose and delivery method.



Two-Way ANOVA

```
## ANOVA Table
##             Df Sum Sq Mean Sq F value   Pr(>F)
## supp         1  205.4   205.4  15.572 0.000231
## dose         2 2426.4  1213.2  92.000  < 2e-16
## supp:dose    2  108.3    54.2   4.107 0.021860
## Residuals 54  712.1    13.2
```

Setting a reference group, we can see the estimated effects.

```
##               Estimate Std. Error t value Pr(>|t|)
## Ref(OJ:0.5)     13.230      1.148  11.521 3.60e-16
## VC              -5.250      1.624  -3.233  0.00209
## 1.0              9.470      1.624   5.831 3.18e-07
## 2.0             12.830      1.624   7.900 1.43e-10
## VC:1.0          -0.680      2.297  -0.296  0.76831
## VC:2.0           5.330      2.297   2.321  0.02411
```

Posthoc comparisons in a Two-Way ANOVA

If there is an interaction between the factors posthoc comparisons must be done within the levels of the other factor.

In our example, there are 6 groups which means 15 pairwise comparisons are possible.

- Each of the 3 levels of Dose contain a single comparison of delivery methods (VC-OJ).
- Each of the 2 delivery methods have 3 possible dose comparisons (1.0-0.5, 2.0-0.5, 2.0-1.0).
- The other 6 pairwise comparisons are usually not of interest because both the delivery method and the dose change between the 2 groups being compared.

```
## Within Dose Level

##                   Diff   SE   Tstat
## VC:0.5 - OJ:0.5 -5.25 1.62 -3.2327
## VC:1.0 - OJ:1.0 -5.93 1.62 -3.6514
## VC:2.0 - OJ:2.0  0.08 1.62  0.0493

## Within delivery method

##                   Diff   SE Tstat
## OJ:1.0 - OJ:0.5  9.47 1.62  5.83
## OJ:2.0 - OJ:0.5 12.83 1.62  7.90
## OJ:2.0 - OJ:1.0  3.36 1.62  2.07
## VC:1.0 - VC:0.5  8.79 1.62  5.41
## VC:2.0 - VC:0.5 18.16 1.62 11.18
## VC:2.0 - VC:1.0  9.37 1.62  5.77
```

More Complicated ANOVA

ANOVA models can have any number of factors.

- As the number of factors of interest increases the number of factor level combinations increases dramatically.
- This is not true for blocking factors because they do not interact with the other factors.

When the number of factors is large it may become impossible to observe every combination of factor levels possible.

We can reduce the sample size by assuming certain higher order interactions are negligible and design an experiment that confounds these effects.

This leads to incomplete block designs, fractional factorial designs, Latin square designs and others.

Example: 3 binary factors in 6 blocks.

Response is growth of peas

factors: nitrogen (N), phosphate (P), potassium (K).

blocks: 6 plots of land each subdivided into 4 sections.

We have 3 factors each with 2 levels (present/absent) so there are 8 groups in total.

We can only observe 4 groups in each plot.

In order to maximize our statistical power when estimating the main effects and two-way interaction between the 3 elements, we use a fractional factorial design in each plot that confounds the three way interaction between the 3 elements.

```
## Response: yield
##           Sum Sq Df F value   Pr(>F)
## block     343.29  5  4.4467 0.015939
## N         194.04  1 12.5672 0.004033
## P           0.35  1  0.0225 0.883288
## K           7.35  1  0.4758 0.503435
## N:P        21.28  1  1.3783 0.263165
## N:K        33.13  1  2.1460 0.168648
## P:K         0.48  1  0.0312 0.862752
## N:P:K             0
## Residuals 185.29 12
```

The model contains a main effect for the block and three elements factors, and 3 two way interactions between the 3 element factors.

Block is significant but none of the two way interactions are significant.

Since we see no evidence of interactions, we refit the model excluding interactions so we can interpret the main effects.

```
## Response: yield
##             Sum Sq Df F value  Pr(>F)
## N          189.282  1 11.8210 0.00366
## P            8.402  1  0.5247 0.47999
## K           95.202  1  5.9455 0.02767
## Residuals 240.185 15
```

Now we have significant effects for both N and K.

```
##           Estimate Std. Error t value Pr(>|t|)
## N Present   5.6167     1.6336  3.4382  0.00366
## P Present  -1.1833     1.6336 -0.7244  0.47999
## K Present  -3.9833     1.6336 -2.4383  0.02767
```

Summary

ANOVA are used to compare numeric responses by categorical predictors.

Predictors can be of interest (factors) or not (blocks)

ANOVA is robust to the normality assumption.

Balanced ANOVA is robust to the equal variance assumption.

Independent observations is an important assumption.

Questions?

- www.stat.ubc.ca/SCARL
- STAT 551 - Stat grad students taking this course offer free statistical advice. Fall semester every academic year.
- SOS Program - An hour of free consulting to UBC graduate students. Funded by the Provost and VP Research.
- Short Term Consulting Service - Advice from Stat grad students. Fee-for-service on small projects (less than 15 hours).
- Hourly Projects - SCARL professional staff. Fee-for-service consulting.

The End