
Analysis of Variance - ANOVA

Eleisa Heron

30/09/09

Introduction

- Analysis of variance (ANOVA) is a method for testing the hypothesis that there is no difference between two or more population means (usually at least three)
- Often used for testing the hypothesis that there is no difference between a number of treatments

Independent Two Sample t-test

- Recall the independent two sample t-test which is used to test the null hypothesis that the population means of two groups are the same
- Let \bar{x}_1 and \bar{x}_2 be the sample means of the two groups, then the test statistic for the independent t-test is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)}$$

with

$$se(\bar{x}_1 - \bar{x}_2) = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

n_1, n_2 sample sizes,

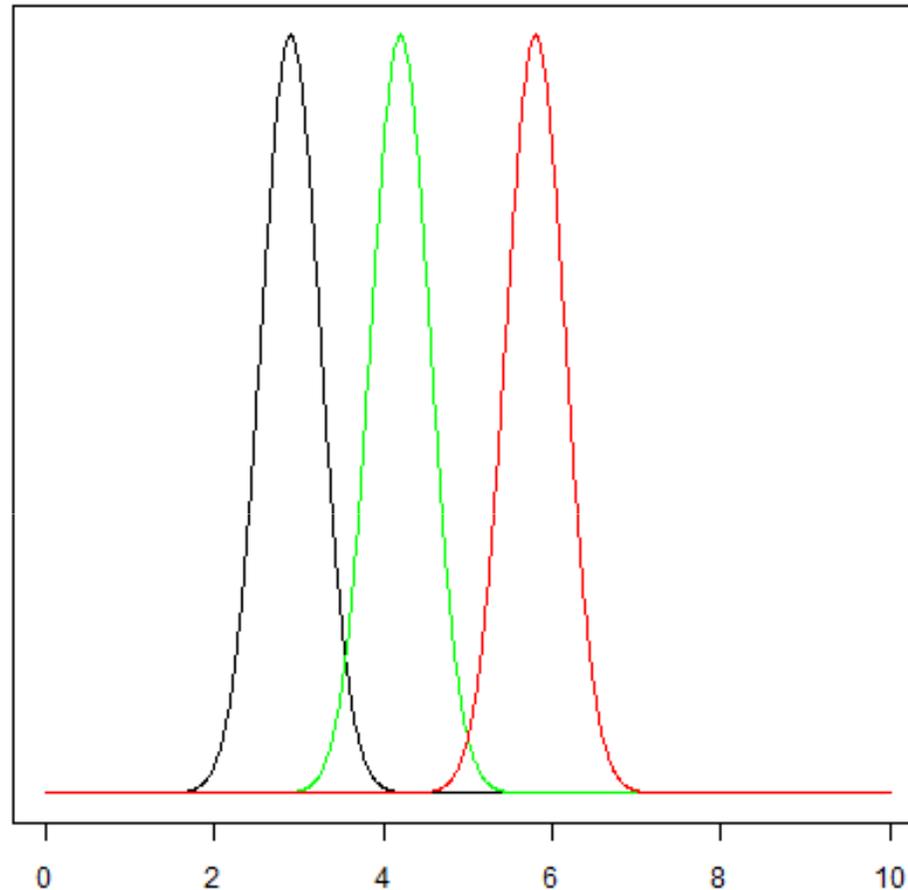
s_1^2, s_2^2 standard deviations

- The test statistic is compared with the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom (df)

Why Not Use t-test Repeatedly?

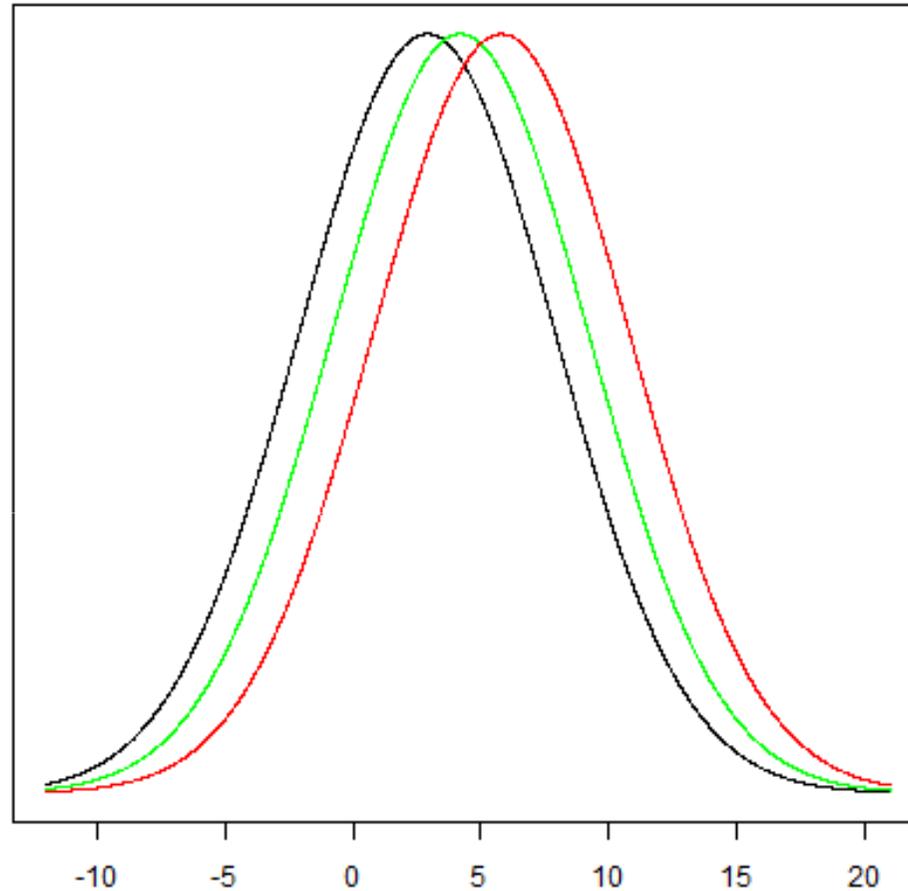
- The t-test, which is based on the standard error of the difference between two means, can only be used to test differences between two means
- With more than two means, could compare each mean with each other mean using t-tests
- Conducting multiple t-tests can lead to severe inflation of the Type I error rate (false positives) and is NOT RECOMMENDED
- ANOVA is used to test for differences among several means without increasing the Type I error rate
- The ANOVA uses data from all groups to estimate standard errors, which can increase the power of the analysis

Why Look at Variance When Interested in Means?



- Three groups tightly spread about their respective means, the variability within each group is relatively small
- Easy to see that there is a difference between the means of the three groups

Why Look at Variance When Interested in Means?



- Three groups have the same means as in previous figure but the variability within each group is much larger
- Not so easy to see that there is a difference between the means of the three groups

Why Look at Variance When Interested in Means?

- To distinguish between the groups, the variability between (or among) the groups must be greater than the variability of, or within, the groups
- If the within-groups variability is large compared with the between-groups variability, any difference between the groups is difficult to detect
- To determine whether or not the group means are significantly different, the variability between groups and the variability within groups are compared

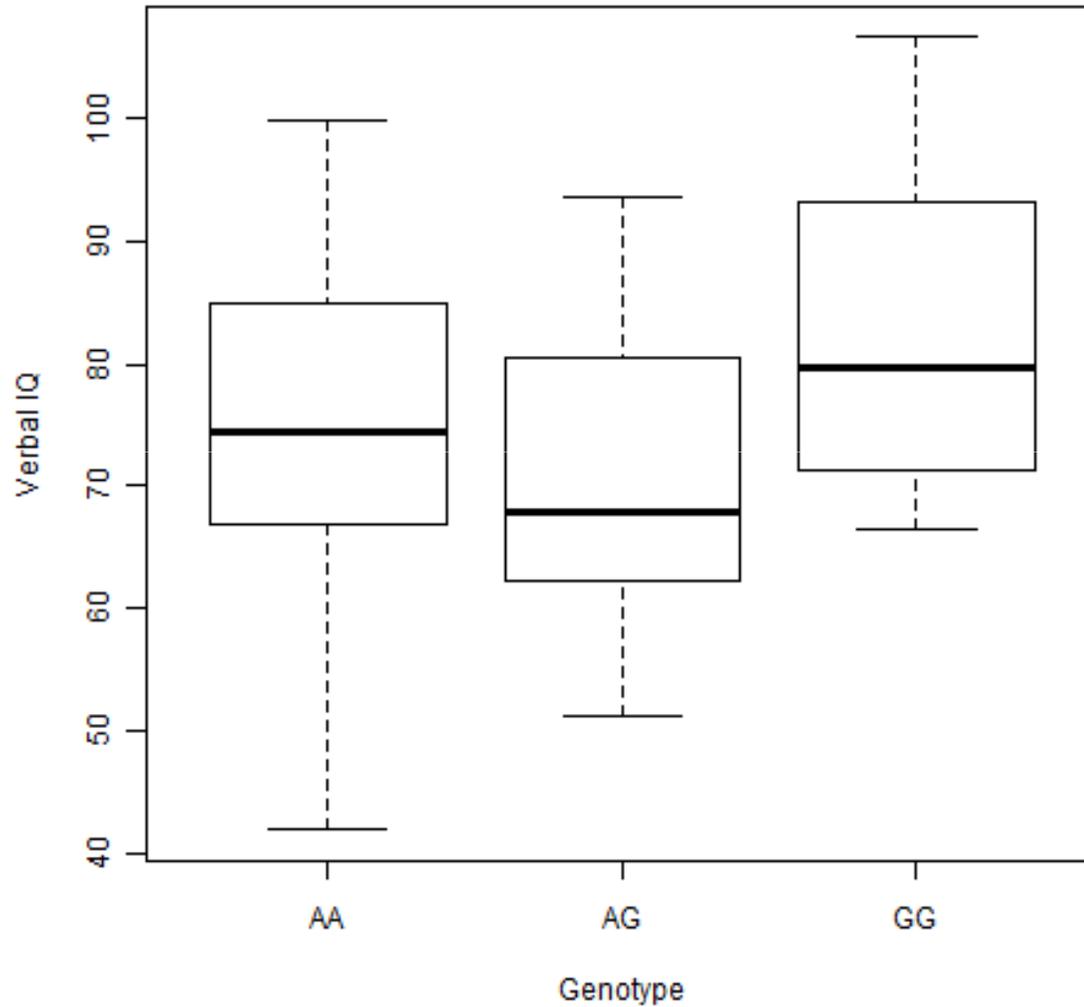
One-Way ANOVA and Assumptions

- One-Way ANOVA
 - When there is only one qualitative variable which denotes the groups and only one measurement variable (quantitative), a one-way ANOVA is carried out
 - For a one-way ANOVA the observations are divided into I mutually exclusive categories, giving the one-way classification

- ASSUMPTIONS
 - Each of the populations is Normally distributed with the same variance (homogeneity of variance)
 - The observations are sampled independently, the groups under consideration are independent

ANOVA is robust to moderate violations of its assumptions, meaning that the probability values (P -values) computed in an ANOVA are sufficiently accurate even if the assumptions are violated

Simulated Data Example



- 46 observations
- 15 AA observations
mean IQ for AA = 75.68
- 18 AG observations
mean IQ for AG = 69.8
- 13 GG observations
mean IQ for GG = 85.4

Introduction of Notation

- Consider I groups, whose means we want to compare
- Let n_i , $i = 1, 2, \dots, I$ be the sample size of group i
- For the simulated verbal IQ and genotype data, ($I = 3$), representing the three possible genotypes at the particular locus of interest. Each person in this data set, as well as having a genotype, also has a verbal IQ score
- Want to examine if the mean verbal IQ score is the same across the 3 genotype groups
 - Null hypothesis is that the mean verbal IQ is the same in the three genotype groups

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

x_{ij} = observation j in group i

$\bar{x}_1, \bar{x}_2, \bar{x}_3$ = sample means of the genotype groups AA, AG, GG

\bar{x} = mean of all the observations

N = $n_1 + n_2 + n_3$ = total sample size

Within-Groups Variance

- Remember assumption that the population variances of the three groups is the same
- Under this assumption, the three variances of the three groups all estimate this common value
 - True population variance = σ^2
 - Within-groups variance = within-groups mean square = error mean square = s_w^2
- For groups with equal sample size this is given by the average of the variances of the groups

$$s_w^2 = \frac{1}{I} \sum_{i=1}^I s_i^2 = \frac{1}{I} \sum_{i=1}^I \sum_{j=1}^{n_i} \left(\frac{(x_{ij} - \bar{x}_i)^2}{n_i - 1} \right)$$

- For unequal sample sizes, the variances are weighted by their degrees of freedom

$$s_w^2 = \sum_{i=1}^I \frac{(n_i - 1) s_i^2}{N - I}$$

Within-Groups Variance

- In our example data

$$s_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}$$

- Since the population variances are assumed to be equal, the estimate of the population variance, derived from the separate within-group estimates, is valid whether or not the null hypothesis is true

Between-Groups Variance

- If the null hypothesis is true, the three groups can be considered as random samples from the same population
(assumed equal variances, because the null hypothesis is true, then the population means are equal)
- The three means are three observations from the same sampling distribution of the mean
- The sampling distribution of the mean has variance σ^2/n
- This gives a second method of obtaining an estimate of the population variance
- The observed variance of the treatment means is an estimate of σ^2/n and is given by

$$\sigma^2/n = \sum_{i=1}^I \frac{(\bar{x}_i - \bar{x})^2}{I - 1}$$

Between-Groups Variance

- For equal sample sizes, the between-groups variance is then given by:

$$s_b^2 = n \sum_{i=1}^I \frac{(\bar{x}_i - \bar{x})^2}{I - 1}$$

- For unequal sample sizes, the between-groups variance is given by:

$$s_b^2 = \sum_{i=1}^I n_i \frac{(\bar{x}_i - \bar{x})^2}{I - 1}$$

Testing the Null Hypothesis, F-test

- If the null hypothesis is true then the between-groups variance s_b^2 and the within-groups variance s_w^2 are both estimates of the population variance σ^2
- If the null hypothesis is not true, the population means are not all equal, then s_b^2 will be greater than the population variance, σ^2 it will be increased by the treatment differences
- To test the null hypothesis we compare the ratio of s_b^2 and s_w^2 to 1 using an F-test
- F statistic is given by:

$$F = \frac{s_b^2}{s_w^2}$$

with $I - 1$ and $I(n - 1)$ degrees of freedom

Testing the Null Hypothesis, F-test

- Another way of thinking about this ratio: $F = \frac{s_b^2}{s_w^2}$

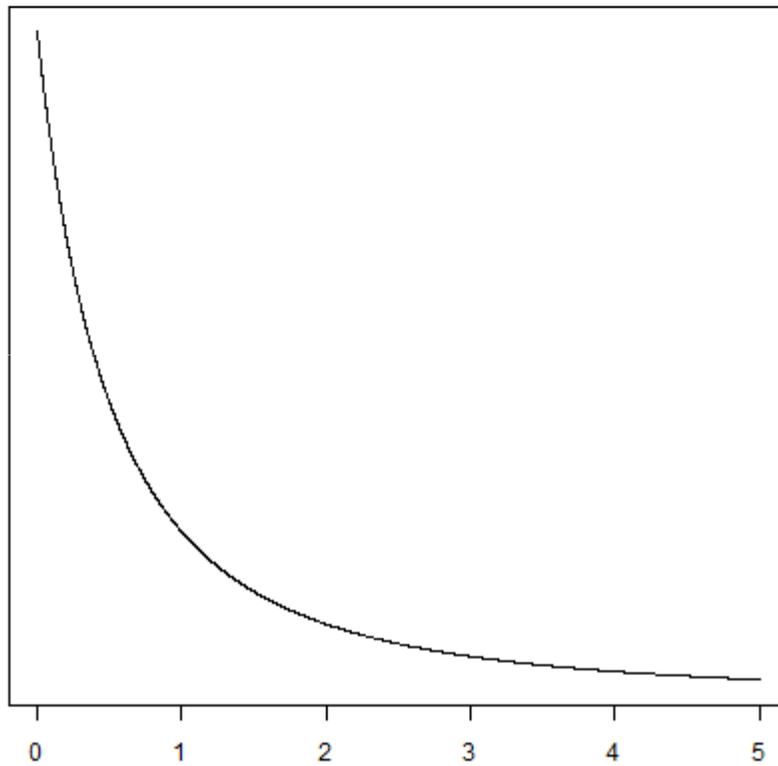
$$F = \frac{\text{variability due to treatment effect and variability due to chance}}{\text{variability due to chance}}$$

F Distribution and F-test

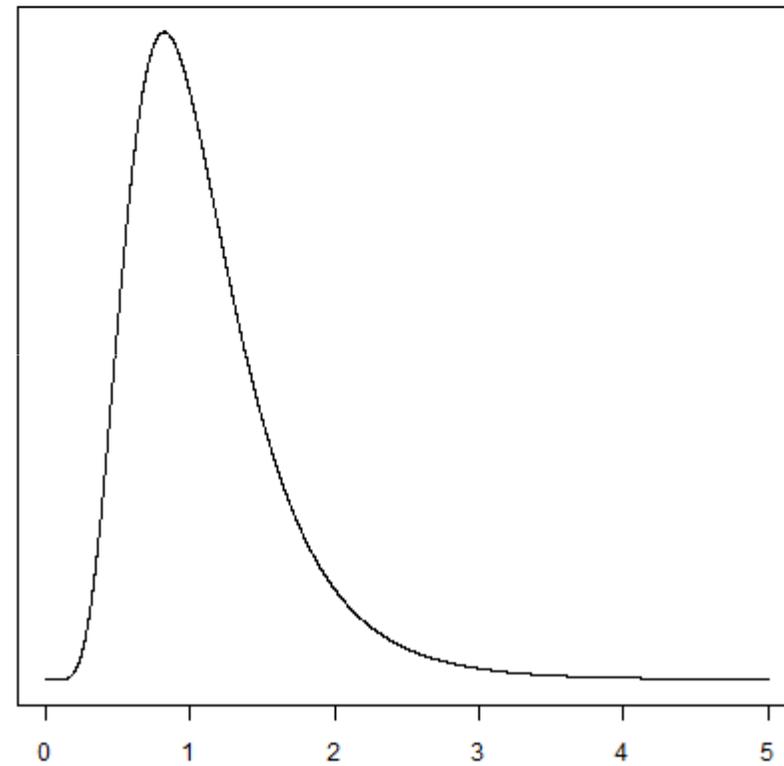
- The F distribution is the continuous distribution of the ratio of two estimates of variance
- The F distribution has two parameters: degrees of freedom numerator (top) and degrees of freedom denominator (bottom)
- The F-test is used to test the hypothesis that two variances are equal
- The validity of the F-test is based on the requirement that the populations from which the variances were taken are Normal
- In the ANOVA, a one-sided F-test is used, why not two-sided in this case?

F Distribution

F Distribution with $df1 = df2 = 2$



F Distribution with $df1 = df2 = 20$



ANOVA Table

Source of variation	Df	Sum Sq	Mean Sq
Between groups (between treatments)	I-1	$n \sum_{i=1}^I (\bar{x}_i - \bar{x})^2$	s_b^2
Within groups (residual variation)	I(n-1)	$\sum_{i=1}^I \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2$	s_w^2
Total variation	In-1	$\sum_{i=1}^I \sum_{j=1}^n (x_{ij} - \bar{x})^2$	

I-1 = 3-1 = 2, since 3 genotype groups, AA, AG, GG

Slightly more complicated as the sample sizes are not all equal (15-1 + 18-1+13-1) = 43

- For the simulated genotype, verbal IQ data:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Genotype_IQ[, 1]	2	1372.58	686.29	3.68	0.0336
Residuals	43	8026.77	186.67		

One-sided P value, "statistically significant" at 0.05 level

Sum Sq / Df = Mean Sq

Assumption Checking

- Homogeneity of variance = homoscedasticity
 - The dependent variable (quantitative measurement) should have the same variance in each category of the independent variable (qualitative variable)
 - Needed since the denominator of the F-ratio is the within-group mean square, which is the average of the group variances
 - ANOVA is robust for small to moderate departures from homogeneity of variance, especially with equal sample sizes for the groups
 - Rule of thumb: the ratio of the largest to the smallest group variance should be 3:1 or less, but be careful, the more unequal the sample sizes the smaller the differences in variances which are acceptable

Assumption Checking

- Testing for homogeneity of variance
 - Levene's test of homogeneity of variance
 - Bartlett's test of homogeneity of variance (Chi-square test)
 - Examine boxplots of the data by group, will highlight visually if there is a large difference in variability between the groups
 - Plot residuals versus fitted values and examine scatter around zero, residuals = observations – group mean, group mean = fitted value
- Bartlett's test: null hypothesis is that the variances in each group are the same
 - For the simulated genotype, IQ data:

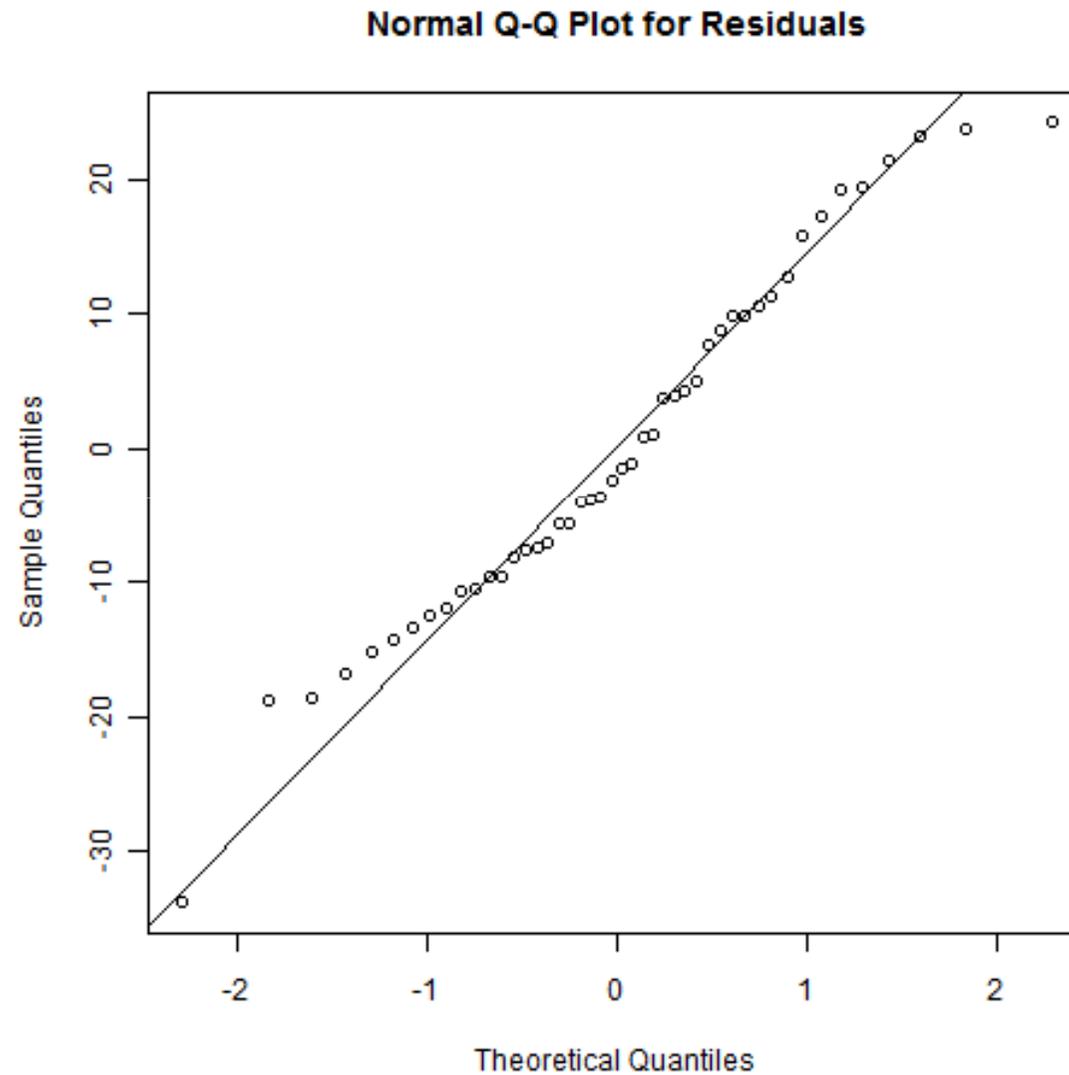
Bartlett's K-squared = 1.5203, df = 2, p-value = 0.4676

Assumption Checking

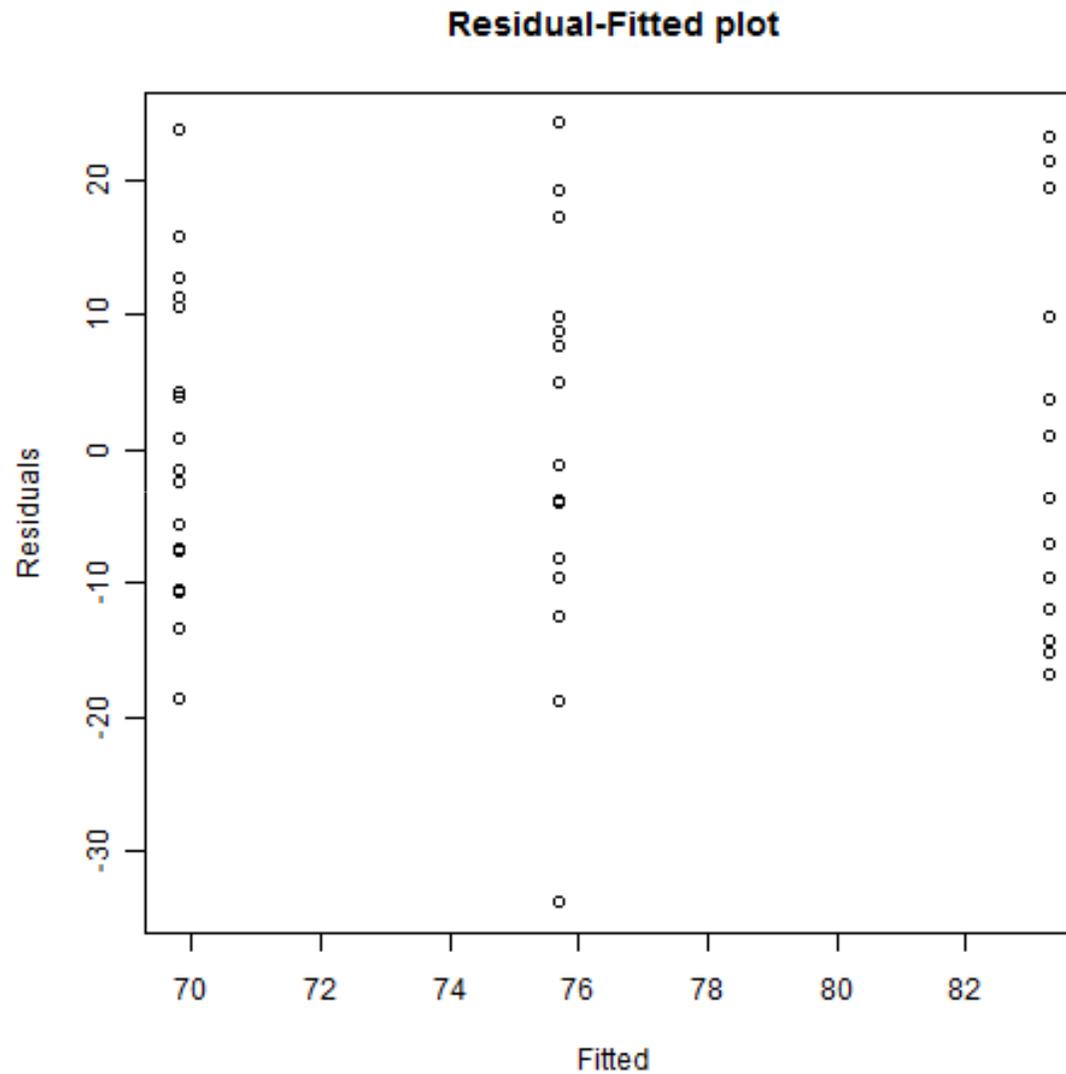
Normality Assumption

- The dependent variable (measurement, quantitative variable) should be Normally distributed in each category of the independent variable (qualitative variable)
- Again ANOVA is robust to moderate departures from Normality
- Checking the Normality assumption
 - Boxplots of the data by group allows for the detection of skewed distributions
 - Quantile-Quantile plots (QQ plots) of the residuals, which should give a 45-degree line on a plot of observed versus expected values,
 - Usual tests for Normality may not be adequate with small sample sizes, insufficient power to detect deviations from Normality

Residuals for Simulated Genotype, IQ Data



Residuals vs Fitted for Simulated Genotype, IQ Data



Assumption Checking

- If data are not Normal or don't have equal variances, can consider
 - transforming the data, (eg. log, square root)
 - non-parametric alternatives

R Code for ANOVA

- In R, due to the way an ANOVA is carried out the data needs to be in a specific format
- If the factor is numeric, R will read the analysis as if it were carrying out a regression, and not a comparison of means, (coerce the numeric into factor with `as.factor()`)

"Genotype"	" IQ"
AA	94.87134
AA	80.74305
AA	71.70750
.	.
.	.
.	.
GG	84.30119
GG	93.08468
GG	69.08003
GG	66.46150

```
# Loading in the data contained in the file "ANOVA_data.txt"  
> Genotype_IQ <- read.table("ANOVA_data.txt")
```

```
# Creating a boxplot of the data  
> boxplot(Genotype_IQ[,2]~Genotype_IQ[,1], xlab = "Genotype", ylab = "Verbal IQ")
```

R Code for ANOVA

```
# ANOVA is carried out in R using the "lm" linear model function  
> g<- lm(Genotype_IQ[,2]~Genotype_IQ[,1])
```

```
# "anova" function computes the ANOVA table for the fitted model object (in this case g)  
> anova(g)
```

Analysis of Variance Table

Response: Genotype_IQ[, 2]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Genotype_IQ[, 1]	2	1372.6	686.3	3.6765	0.03358 *
Residuals	43	8026.8	186.7		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Summary So Far

- Within-Groups Mean Square
 - This is the average of the variances in each group
 - This estimates the population variance regardless of whether or not the null hypothesis is true
 - This is also called the error mean square
- Between-Groups Mean Square
 - This is calculated from the variance between groups
 - This estimates the population variance, if the null hypothesis is true
- Use an F-test to compare these two estimates of variance
- Relationship with t-test
 - If the one-way ANOVA is used for the comparison of two groups only, the analysis is exactly and mathematically equivalent to the use of the independent t-test, (the F statistic is exactly the square of the corresponding t statistic)

What to do with a Significant ANOVA Result (F-test)

- If the ANOVA is significant and the null hypothesis is rejected, the only valid inference that can be made is that at least one population mean is different from at least one other population mean
- The ANOVA does not reveal which population means differ from which others
- Some questions we might want to ask:
 - Is one treatment much better than all the other treatments?
 - Is one treatment much worse than all the other treatments?
- Only think about investigating differences between individual groups when the overall comparison of groups (ANOVA) is significant, or that you had intended particular comparisons at the outset
- Need to consider whether the groups are ordered or not
- There are many approaches for *after the ANOVA*, here are some

What to do with a Significant ANOVA Result (F-test)

- Modified t-test:
 - To compare groups after the ANOVA, use a modified t-test
 - Modified t-test is based on the pooled estimate of variance from all the groups (within groups, residual variance in the ANOVA table), not just the pair being considered

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x}_1 - \bar{x}_2)} \quad se(\bar{x}_1 - \bar{x}_2) = s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Need to think about multiple testing corrections
- Least significant difference:
 - The least difference between two means which is significant is given by:

$$s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{I(n-1)}$$

- Arrange the treatment means in order of magnitude and the difference between any pair of means can be compared with the least significant difference

What to do with a Significant ANOVA Result (F-test)

- Tukey's honest significance test
 - The test compares the means of every group to the means of every other group
 - The test corrects for the multiple comparisons that are made
- Linear Trend
 - When the groups are ordered we don't want to compare each pair of groups, but rather investigate if there is a trend across the groups (linear trend)
 - Between groups variability s_b^2 can be partitioned into a component due to a linear trend and a non-linear component
- There are many other tests available...

Two-Way ANOVA, Factorial Designs

- One dependent variable (quantitative variable), two independent variables (classifying variables = factors)
- Advantage is that factorial designs can provide information on how the factors interact or combine in the effect that they have on the dependent variable

Factor B	Factor A				
	1	2	3	c
1	X....X	X....X	X....X	X....X
2	X....X	X....X	X....X	X....X
3	X....X	X....X	X....X	X....X
:	:	:	;		:
:	;	:	;		;
:	:	;	:		:
r	X....X	X....X	X....X	X....X

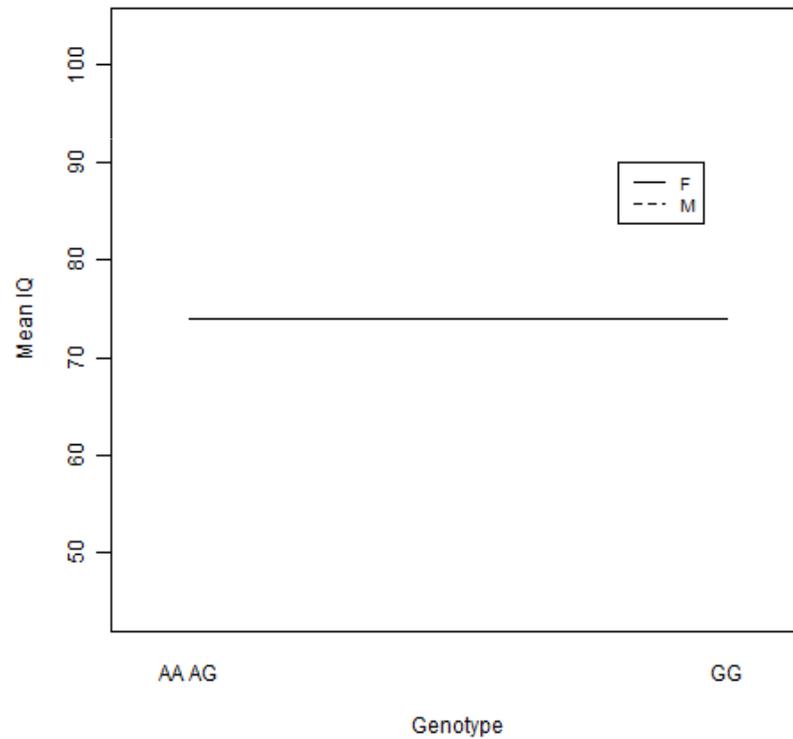
Factorial ANOVAs

- Different types of effects are possible, the treatment effect: a difference in population means, can be of two types:
 - Main effect: a difference in population means for a factor collapsed over the levels of all other factors in the design
 - Interaction effect: when the effect on one factor is not the same at the levels of another
- Interaction: whenever the effect of one factor depends on the levels of another factor
- For a 2-way ANOVA, 8 possible outcomes that could have occurred:
 - Nothing
 - Main effect of factor A
 - Main effect of factor B
 - Both main effects (factor A and factor B)
 - AxB interaction
 - AxB interaction and main effect of factor A
 - AxB interaction and main effect of factor B
 - AxB interaction and both main effects (factors A and B)

Interactions in ANOVA

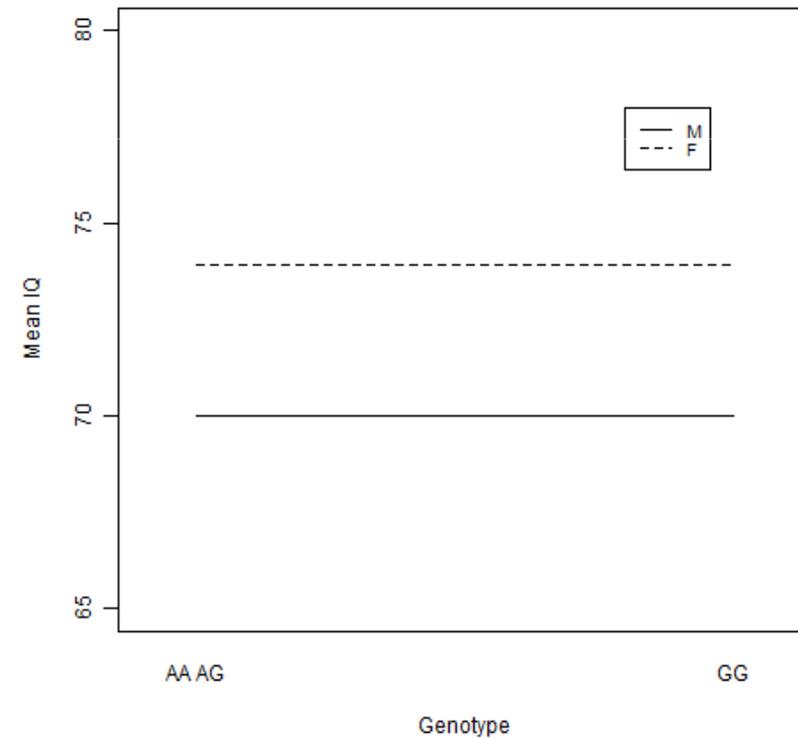
		Factor 1		
Factor 2		AA	AG	GG
	F	74	74	74
M	74	74	74	
	74	74		

No Effect



		Factor 1		
Factor 2		AA	AG	GG
	F	74	74	74
M	70	70	70	
	72	72		

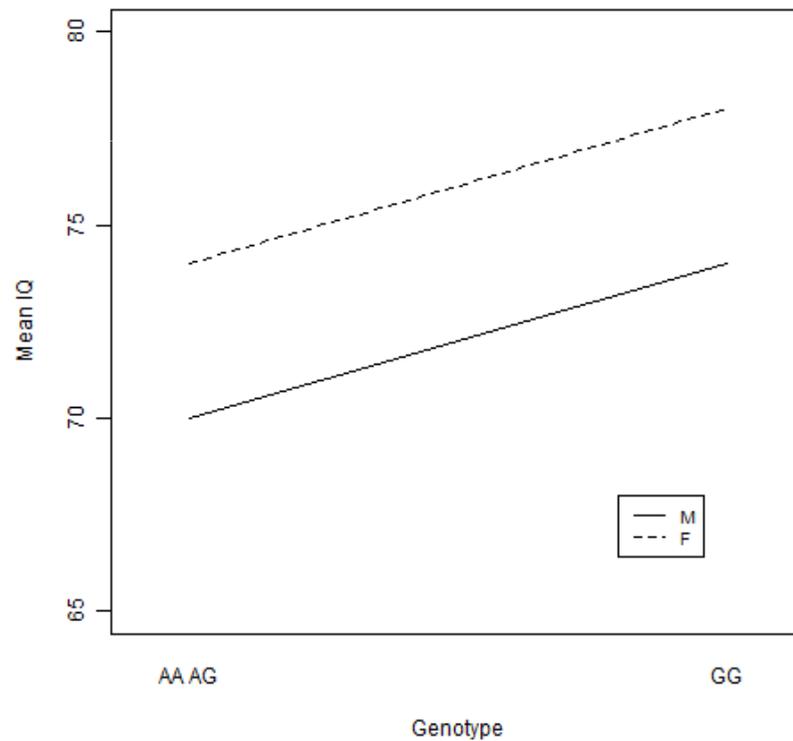
Factor 2 Main Effect



Interactions in ANOVA

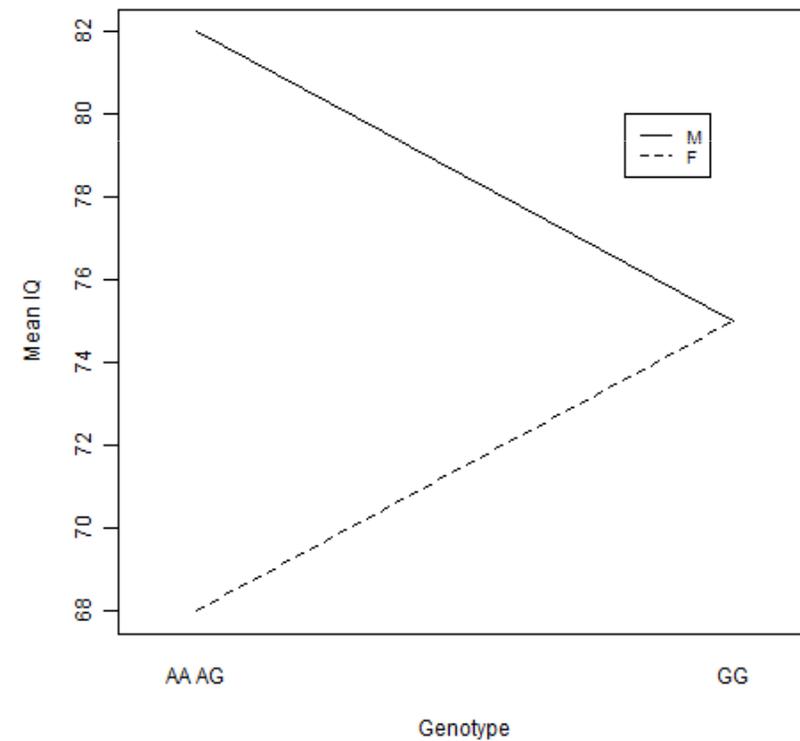
		Factor 1		
Factor 2		AA	AG	GG
	F	74	78	76
	M	70	74	72
		72	76	

Factor 1 and Factor 2 Main Effects



		Factor 1		
Factor 2		AA	AG	GG
	F	68	75	71.5
	M	82	75	78.5
		75	75	

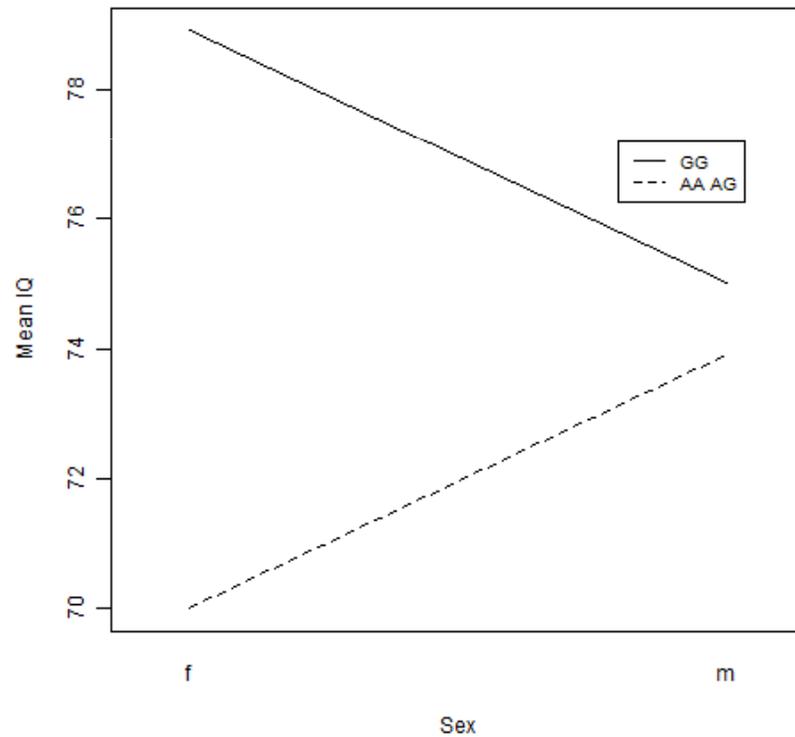
Factor 2 Main Effect and Interaction



Interactions in ANOVA

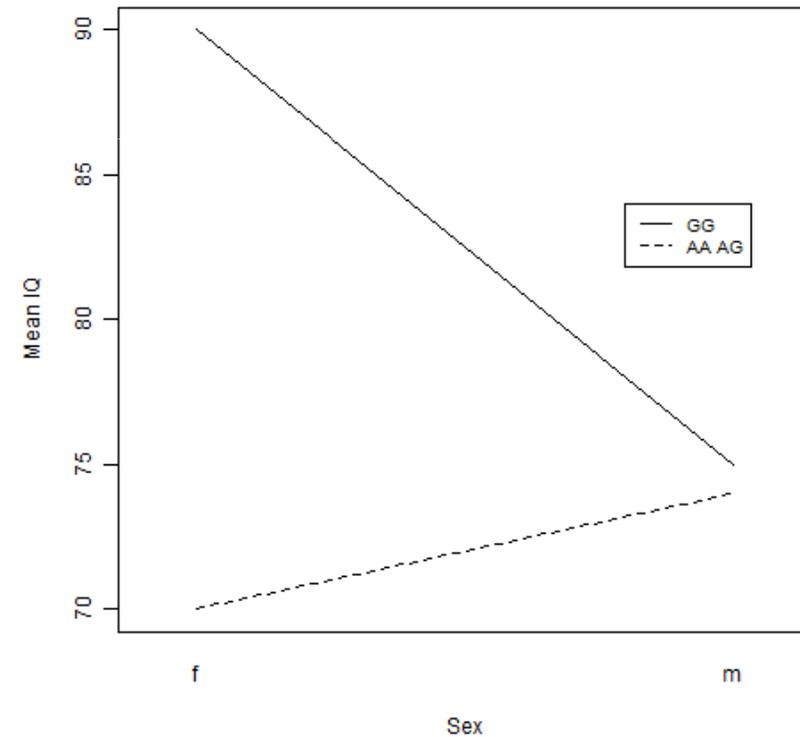
		Factor 1		
Factor 2		AA AG	GG	
	F	70	79	74.5
	M	74	75	74.5
		72	77	

Factor 1 Main Effect and Interaction



		Factor 1		
Factor 2		AA AG	GG	
	F	70	90	80
	M	74	75	74.5
		72	82.5	

Factors 1,2 Main Effects and Interaction



Repeated Measures ANOVA

- Repeated measures ANOVA is used when all members of a random sample are measured under a number of different conditions
- A standard ANOVA would treat the data as independent and fail to model the correlation between the repeated measures

	Time (mins)				
Subject	0	10	20	60
1	X	X	X	X
2	X	X	X	X
3	X	X	X	X
:	:	:	:		:
:	:	:	:		:
:	:	:	:		:
20	X	X		X

ANCOVA

- ANCOVA: Analysis of Covariance
 - Extend the ANOVA to include a qualitative independent variable (covariate)
 - Used to reduce the within group error variance
 - Used to eliminate confounders
 - Most useful when the covariate is linearly related to the dependent variable and is not related to the factors (independent qualitative variables)
 - Similar assumptions to the ANOVA

Non Parametric ANOVA – Kruskal-Wallis Test

- ANOVA is the more general form of the t-test
- Kruskal-Wallis test is the more general form of the Mann-Whitney test
- Kruskal-Wallis test, doesn't assume normality, compares medians:
 - Rank the complete set of N observations, 1 to N (regardless of the group they belong to)
 - For each group, calculate the sum of the ranks
 - Let the sum of the ranks of the n_i observations in group i be R_i , then the average rank in each group is given by \bar{R}_i
 - The H statistic is defined as:

$$H = \frac{12 \sum n_i (\bar{R}_i - \bar{R})^2}{N(N + 1)}$$

\bar{R} average of all the ranks

- When the null hypothesis is true, H follows a Chi squared distribution $I - 1$ degrees of freedom
- Any variation among the groups will increase the test statistic (upper tail test)

Non-Parametric Two-Way ANOVA

- Friedman's two way analysis of variance non-parametric hypothesis test
 - It's based on ranking the data in each row of the table from low to high
 - Each row is ranked separately
 - The ranks are then summed in each column (group)
 - The test is based on a Chi squared distribution
 - Just like with the ANOVA the Friedman test will only indicate a difference but won't say where the difference lies

Literature Example

- ***“The MTHFR 677C-T Polymorphism and Behaviors in Children With Autism: Exploratory Genotype–Phenotype Correlations”***

Robin P. Goin-Kochel, Anne E. Porter, Sarika U. Peters, Marwan Shinawi, Trilochan Sahoo, and Arthur L. Beaudet

Autism Research 2: 98–108, 2009

Table II. Means and One-Way ANOVAs for PPVT-III, RCPM, and VABS Scores by *MTHFR* 677C → T Genotype

Instrument	<i>MTHFR</i> 677C → T genotype			F(df)	P
	CC M(SD)	CT M(SD)	TT M(SD)		
ADI-R behavior composite ^a	6.12(2.14)	6.13(2.13)	5.6(2.21)	0.52(2, 144)	0.60
PPVT-III	69.04(27.64)	63.68(26.73)	62.40(27.84)	0.38(2, 69)	0.69
RCPM	24.26(7.93)	22.39(7.12)	23.45(7.00)	0.47(2, 72)	0.63
VABS					
Communication	45.14(24.26)	42.64(19.84)	45.18(24.56)	0.16(2, 109)	0.85
Socialization	43.74(20.02)	44.58(17.72)	48.29(13.23)	0.40(2, 109)	0.67
Daily living	35.94(22.22)	32.89(19.45)	37.76(20.09)	0.43(2, 109)	0.65
Motor	72.80(20.97)	71.28(24.20)	79.53(20.07)	0.76(2, 96)	0.47
Composite	38.98(18.88)	36.96(15.89)	39.88(16.76)	0.24(2, 109)	0.79

Literature Example

- ***“No Evidence for an Effect of COMT Val158Met Genotype on Executive Function in Patients With 22q11 Deletion Syndrome”***

Bronwyn Glaser, Martin Debbane, Christine Hinard, Michael A. Morris, Sophie P. Dahoun, Stylianos E., Antonarakis, Stephan Eliez
 Am J Psychiatry 2006; 163:537–539

TABLE 1. Scores on Cognitive Measures From 34 Patients With 22q11.2 Deletion Syndrome With the Met or Val Allele at the Val158Met Locus of the COMT Gene

Measure	Met-Hemizygous Group (N=20)		Val-Hemizygous Group (N=14)		Analysis of Variance (df=1, 32)	
	Mean	SD	Mean	SD	F	p
Executive function						
Verbal fluency in animal naming	13.90	4.80	13.29	4.05	0.153	0.698
Stroop task interference	50.05	6.57	51.07	7.73	0.172	0.681
Arithmetic	5.15	1.81	4.71	2.56	0.339	0.564
Digit Span	7.50	2.50	6.79	3.66	0.458	0.503
Executive function composite	15.39	2.28	15.20	1.91	0.062	0.805
General						
IQ						
Full-scale	70.45	9.82	69.57	14.06	0.046	0.831
Performance	72.10	11.55	70.86	10.83	0.100	0.754
Verbal	73.90	11.94	75.57	17.33	0.111	0.741
Memory						
Visual						
Immediate	77.85	15.22	73.43	15.22	0.695	0.411
Delayed	82.55	12.77	78.07	11.39	1.105	0.301
Verbal						
Immediate	91.70	10.65	85.86	14.60	1.826	0.186
Delayed	91.70	11.21	82.57	17.32	3.493	0.071

Further Reading

BOOKS:

Interpretation and Uses of Medical Statistics

by Leslie E Daly and Geoffrey J. Bourke, Blackwell science, 2000

Practical Statistics for Medical Research

by Douglas G. Altman, Chapman and Hall, 1991