Introduction to ANOVA

Example 1: Advertisement Experiment

In an experiment to investigate the effect of types of advertisement (A1, A2, A3, A4, A5) on rating of a new product, 30 representative individuals were selected and each AD was assigned at random to the 6 groups. The response is the combined ratings of 3 characteristics of the product (each rated 1 to 7 with the combined rating ranging from 3 to 21).

- Advertisement is a qualitative explanatory variable and rating of new product is a quantitative response variable.
- 5 Advertisement Types (A1, A2, A3, A4, A5) – 6 subjects on each Ad Types.
- Goal: compare the advertisement types in terms of their effectiveness

Example 2: Kenton Food Company – KNNL Textbook p685

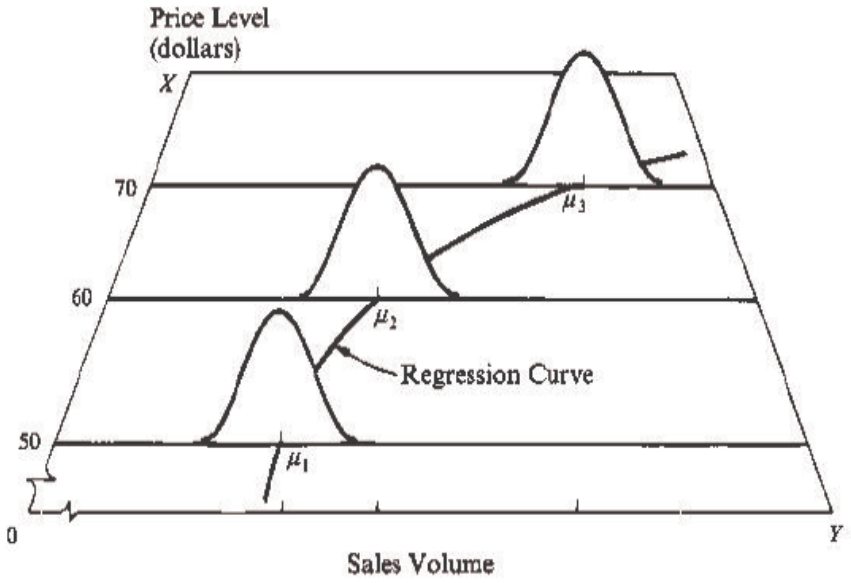
Kenton Food Company (KFC) – Experiment with 4 different package designs for new breakfast cereal; 20 stores with similar characteristics were selected. Each store randomly assigned one of the designs, with each package design to 5 stores; one store that caught fire during the study was dropped. The number of was recorded over the study period.

- Package design is a qualitative explanatory variable and cases of cereals sold is a quantitative response variable.
- 4 package designs (D1, D2, D3, D4)
- Goal: compare the package designs in terms of their effectiveness

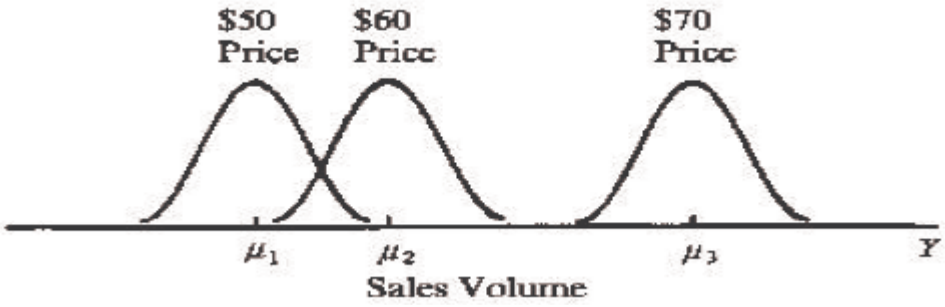
ANOVA

- ANOVA = Analysis of Variance
- Compare means among treatment groups, without assuming any parametric relationships (regression does assume such a relationship).
- Example: Price vs. Sales Volume

Regression Model



ANOVA Model



- Important difference: No assumption is made about the manner in which Price and Sales Volume are related.

Similarities to Regression

- Assumptions on errors identical as to regression.
- We assume each population is normal and the variances are identical. We also assume independence.
- Can get "predicted values" for each group, as well as confidence intervals.

Differences

- No specific relationship is assumed.
- Goal becomes: looking for differences among the groups.

Terminology

- We may refer to any qualitative predictor variable as a factor.
- Each factor has a certain number of levels.
- Experimental factors are "set" or "assigned" to the experimental units; observational factors are characteristics of the experimental units that cannot be assigned.
- Factors are qualitative if they represent traits that could not be placed in some logical numerical order.
  _ GENDER, BRAND, DRUG
- Factors are quantitative if levels are described by numerical quantities on an equal interval scale.
  _ AGE, TEMPERATURE
- A Treatment is a specific experimental condition (determined by factors and levels of each factor).
- The Experimental Unit (Basic Unit of Study) is the smallest unit to which a treatment can be assigned.
- A design is called balanced if each treatment is replicated the same number of times (i.e. same number of experimental units per treatment).

Example 1: ADVERTISEMNT Experiment

- Five advertisements – each used for 6 subjects
- Advertisement is an experimental factor; Experimental Unit is the subject (person) exposed to or receiving the advertisement.
- There are five advertisements (treatments), which may or may not have any logical "ordering"
- Design is balanced (generally) since we are able to assign the advertisements (treatments).
- Six subjects per AD group – 30 subjects

		Subject (j)								No. of
		1	2	3	4	5	6	Total	Mean	Subjects
		Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}	Y_{i6}	$Y_{i.}$	$\bar{Y}_i$	n_i
Advertisement Group (i)	A1	6	8	9	7	3	5			6
	A2	16	13	12	18	19	19			6
	A3	21	20	20	19	20	18			6
	A4	05	10	13	18	09	18			6
	A5	13	17	12	18	15	13			6
All Ads										30

Example 2: Kenton Food Company – KNNL Textbook p685

Kenton Food Company (KFC) – Experiment with 4 different package designs for new breakfast cereal; 20 stores with similar characteristics were selected. Each store randomly assigned one of the designs, with each package design to 5 stores; one store that caught fire during the study was dropped.

Four package designs – each sold in 5 stores

Package design is an experimental factor; Experimental Unit is a store selling a design of cereal package.
There are 4 package designs (treatments), which may or may not have any logical "ordering"

Experimental design is balanced (generally by arrangement) since we are able to assign the package designs (treatments). In this case, however, a fire occurred in one store and it was dropped from the study. Experimental design is unbalanced.

Five store groups – 20 stores, but one store was dropped due to fire. One of the designs is tested in only 4 stores

		Store (j)							No. of
		1	2	3	4	5	Total	Mean	Stores
		Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}	$Y_{i.}$	$\bar{Y}_i$	n_i
Package Design (i)	D1	11	17	16	14	15	73	14.6	5
	D2	12	10	15	19	11	67	13.4	5
	D3	23	20	18	17		78	19.5	4
	D4	27	33	22	26	28	136	27.2	5
All designs							$Y_{..}$=354	$\bar{Y} = 18.63$	19

Multiple Factors

- With two or more factors, each combination of levels is generally called a treatment combination
- Can treat as single variable if desired
- Example: Package design * Package size

  _ 4 Designs (D1, D2, D3, D4)

  _ 3 Sizes (Small, Medium, Large)

  _ 12 treatment combinations

Crossed Factors

- Two factors are crossed if **all factor combinations** are represented.
- Example:
  - Package design (all)* Package size (all)

	S1	S2	S3
D1	xx	xx	xx
D2	xx	xx	xx
D3	xx	xx	xx
D4	xx	xx	xx

  - Blood Type (A, B, AB, O) * Medication (1, 2, 3, 4, 5)

Nested Factors

- One factor has levels that are unique to a given level of another factor
- Example: Plant * Operator

Plant # 1	Plant # 2	Plant # 3	Plant # 4
Operator # 1	Operator # 5	Operator # 9	Operator # 13
Operator # 2	Operator # 6	Operator # 10	Operator # 14
Operator # 3	Operator # 7	Operator # 11	Operator # 15
Operator # 4	Operator # 8	Operator # 12	Operator # 16

- Operators are nested within manufacturing plants

Control Groups

- Often a control or placebo treatment is used. This treatment is more of a "standard" than a treatment, as it is the case of no treatment at all.

- Comparing treatments to controls can be a very effective way of showing that a treatment is effective.

Fixed vs. Random Factors

- For the most part, we will consider only fixed effect models in this class. A factor is called fixed because the levels are chosen in advance of the experiment and we were interested in differences in response among those specific levels.

- A factor is called random because the levels are chosen randomly from a large population of possible levels.

Randomization

- Completely separate concept from random effects.
- In an experimental study, generally want to avoid any potential bias in the design by randomizing treatments to experimental units whenever possible.
- Randomization may be constrained.

  Example: Have 100 people, 50 men and 50 women. Randomly assign each of the 5 treatments to 10 men and 10 women.

Experimental Designs

- Completely Randomized Design
- Randomized Complete Block Designs
- Factorial Experiments
- Nested Designs
- Repeated Measures Designs
- Incomplete Block Designs

Single factor ANOVA

- The response variable *Y* is continuous (same as in regression)
- There are two key differences regarding the explanatory variable *X:*
  - it is a qualitative variable (business type, investment plans, advertisement types, gender, location). Instead of calling it an *explanatory variable,* we now refer to it as a *factor.*
  - No assumption (i.e. linear relationship) is made about the nature of the relationship between X and *Y.* Rather we attempt to determine whether the response differ significantly at different levels of X. This is a generalization of the *two-independent-sample t-est.*

- We will have several different ways of parameterizing the model:
  - the Cell Means Model
  - the Factor Effects Model (with two different possible constraint systems)

Notation for Single-factor /One-way ANOVA

X is the qualitative factor
- r is the number of *levels*
- we often refer to these as *groups* or *treatments*

Y is the continuous response variable
- Y_{ij} is the value of the response in the j-th trial for the i-th factor level i = 1, 2, ..., r levels of the factor X, j = 1, 2, . . . , n_i observations at factor level i

Cell Means Model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where

Y_{ij} is thevalue of the response in the j-th trial for the i-th factor level or treatment

μ_i are parameters

ε_{ij} are error terms that are independent and N(0, σ^2)

$i = 1, 2, ..., r; j = i, 2, ..., n_i$

Model Assumptions

- Response variable is normally distributed
- Mean may depend on the level of the factor
- Variance is constant
- All observations are independent

Note

- Model may be stated as Y_{ij} : N(μ_i, σ^2) and may be used for data from observational or experimental studies based on a CRD

- Model is without "intercept" term but with a potentially *different* mean for each level of *X*. Also, the mean does not depend numerically on the actual value of *X* (unlike the linear regression model)

Parameters

The parameters of the model are $\mu_1, \mu_2, \ldots, \mu_r, \sigma^2$.

Objective: Does the explanatory variable help to explain the mean of Y?

Equivalently, does mean of Y, μ_i, depend on the factor level ***i***? This leads to the testing of the hypothesis

$$\begin{cases} H_o: \mu_1 = \mu_2 = \ldots = \mu_r \\ H_a: \text{not all } \mu_i \text{ are equal} \end{cases}$$

Data layout for Single-factor (CRD)

		Observations						n_i	T_i	$\bar{Y}_i$	S_i^2
Treatment	1	Y_{11}	Y_{12}		…		$Y_{1,n1}$	n_1	T_1	$\bar{Y}_1$	S_1^2
	2	Y_{21}	Y_{22}		…		$Y_{2,n2}$	n_2	T_2	$\bar{Y}_2$	S_2^2
	…										
	i				Y_{ij}						
	…	…	…				…		…	…	…
	r	Y_{r1}	Y_{r2}			…	$Y_{r,nr}$	n_r	T_r	$\bar{Y}_r$	S_r^2
								$n_T = \sum_{i=1}^{r} n_j$	$T = \sum_{i=1}^{r} T_j$	$\bar{Y} = T / n_T$	$S_P^2 = MSE$

Estimates

- Estimate μ_i by the mean of the observations at factor level *i*:

$$\hat{\mu}_i = \bar{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij} / n_i$$

- Overall or grand mean: $\bar{Y}_{..} = \sum\sum Y_{ij} / n_T$

- Pooled Estimate of σ^2:

$$S_P^2 = \frac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{\sum_{i=1}^{r}(n_i - 1)} = \frac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{n_T - r} = \frac{\sum_{i=1}^{r}\sum_{j}^{n_i}(Y_{ij} - \bar{Y}_{i.})}{n_T - r} = MSE$$

ANOVA Table for Single-Factor study

Source	SS	df	MS	F
Between Treatments	$SSTR = \sum n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	r - 1	MSTR=SSTR/(r - 1)	$F = \frac{MSTR}{MSE}$
Error (within treatments)	$SSE = \sum\sum(Y_{ij} - \bar{Y}_{i.})^2$	n_T - r	MSE = SSE/(n_T - r)	
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y}_{..})^2$	n_T - 1		

F-test

$$\begin{cases} H_o: \mu_1 = \mu_2 = ... = \mu_r \\ H_a: \text{not all } \mu_i \text{ are equal} \end{cases}$$

$$F = \frac{MSR}{MSE}$$

- Under H_0, $F \sim F_{(r-1, n_T - r)}$
- Reject H_0 when F is large.
- Report the p-value

CRD Example: ADVERTISEMNT Experiment (advert.txt)

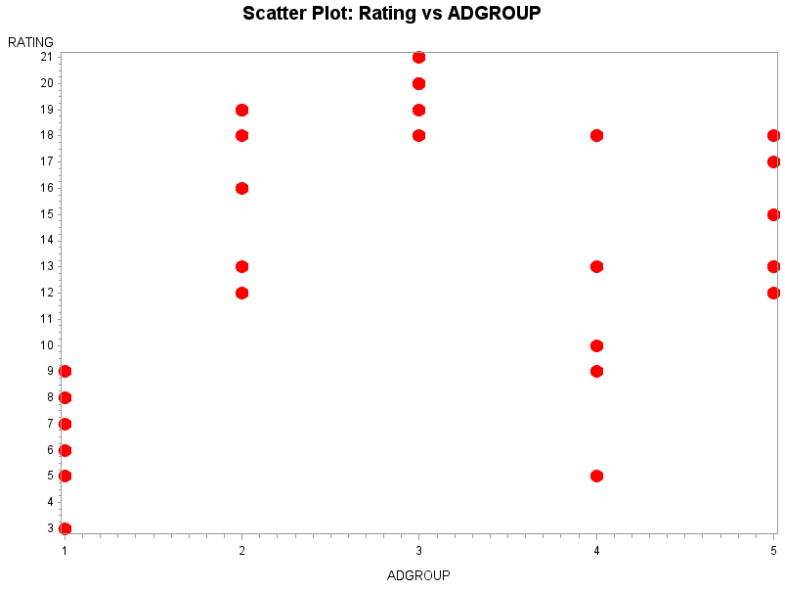
```
PROC GLM data=advert;
title 'One-way CRD - Pairwise Comparisons';
   class adgroup;
     model rating = adgroup;
          means adgroup / cldiff lines lsd tukey bon scheffe;
run;
quit;
title;
```

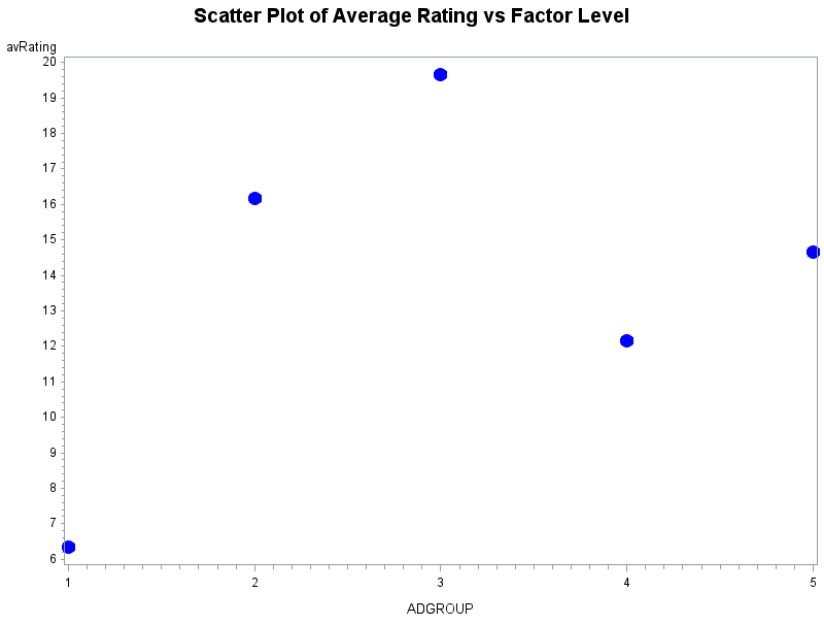
proc glm is used to do analysis:

- **glm** stands for "General Linear Model"
- The **class** statement tells **proc glm** that **ADGROUP** is a classification/categorical variable. The class statement defines variables which are qualitative in nature.

  The **means** statement requests sample means and standard deviations for each factor level.

OUTPUT





OUTPUT

Dependent Variable: RATING

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	595.1333333	148.7833333	15.52	<.0001
Error	25	239.6666667	9.5866667		
Corrected Total	29	834.8000000			

R-Square	Coeff Var	Root MSE	RATING Mean
0.712905	22.43648	3.096234	13.80000

Bonferroni (Dunn) t Tests for RATING

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	25
Error Mean Square	9.586667
Critical Value of t	3.07820
Minimum Significant Difference	5.5026

Comparisons significant at the 0.05 level are indicated by *.**

Comparisons significant at the 0.05 level are indicated by *.**

ADGROUP Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 4	7.500	1.997	13.003	***
3 - 1	13.333	7.831	18.836	***
2 - 1	9.833	4.331	15.336	***
5 - 1	8.333	2.831	13.836	***
4 - 1	5.833	0.331	11.336	***

Means with the same letter are not significantly different.

Bon Grouping		Mean	N	ADGROUP
	A	19.667	6	3
	A			
B	A	16.167	6	2
B	A			
B	A	14.667	6	5
B				
B		12.167	6	4
	C	6.333	6	1

Factor Effects Model

- This is a re-parameterization of the cell means model and a useful way of looking at more complicated models.
- Often the null hypotheses are easier to interpret with the factor effects model. Re-parameterizing the cell means as $\mu_i = \mu + \tau_i$, the cell model becomes

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \text{ where } \tau_i = (\mu_i - \mu)$$

- Here μ is the overall or grand mean and τ_i represents the difference between the overall mean and the mean for factor level *i*.

- The cell means model looks at the amount by which the mean at each level deviates from some "standard".

- The parameters of the factor effects model are $\mu, \tau_1, \tau_2, \ldots, \tau_r, \sigma^2$. There are $r + 2$ of these. One of the τ_i's is redundant (if you know the grand mean and (r – 1) of them, you can compute the rest). To avoid redundancy and make the models equivalent, we assume $\sum_{i=1}^{r} \tau_i = 0$. The factor effects model then is stated as

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} \sim iid \ \ \mathrm{N}(0, \sigma^2), \ \sum \tau_i = 0$$

Hypotheses Tests

- The cell means model hypothesis for the F-test was

$$\begin{cases} H_o: \mu_1 = \mu_2 = ... = \mu_r \\ H_a: \text{not all } \mu_i \text{ are equal} \end{cases}$$

- For the factor effects model, the hypothesis for the F-test is stated as

$$\begin{cases} H_o: \tau_1 = \tau_2 = ... = \tau_r = 0 \\ H_a: \text{at least one of the } \tau_i \text{ is different fro zero} \end{cases}$$

- **Parameter estimates are**

$$\hat{\mu} = \bar{Y}_{..} \text{ and } \hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$$

- **Constraints**
  - The constraint $\sum_{i=1}^{r} \tau_i = 0$ is needed for the parameters to have unique solutions. It also means that the τ_i's represent differences from the grand mean.

    Example: assume that r=3 and $\mu_1 = 20,\ \mu_2 = 30,\ \mu_3 = 40$. If there is no constraint, any set of values is possible for the parameters of the factor effects model:

$$if\ \mu = 0,\ \tau_1 = 20,\ \tau_2 = 30,\ \tau_3 = 40$$
$$if\ \mu = 50,\ \tau_1 = -30,\ \tau_2 = -20,\ \tau_3 = -10$$
$$if\ \mu = 100,\ \tau_1 = -80,\ \tau_2 = -70,\ \tau_3 = -60$$

  - **Remarks on SAS for proc glm**: SAS uses instead $\tau_r = 0$, which means that μ will be the mean for the r^{th} level instead of the grand mean. So in SAS treatments are all

compared to the r^{th} level.

Example: Advertisement Experiment

- Objective: Estimate the parameters for the cell means and for the factor effects model
- model using the constraint $\sum_{i=1}^{r} \tau_i = 0$.
- Easiest way to get the cell-means estimates is to use PROC MEANS. Alternatively, one can use the MEANS statement and put things together from PROC GLM.

- The cell-means estimates are then used to produce the factor effects estimates.

- Using CLASS statement causes means to be produced for each level of the class variable(s).

- Rating score (Factor Effects)
  - Grand mean
  - Factor effects

- **Model using SAS**

```
proc glm data=advert;
     class adgroup;
     model rating=adgroup /solution;
```

  - SAS Estimates are "biased" by the choice of constraint; in this case the intercept represents the cell mean for A5 (since it is alphabetically last)
  - A1 to A4 levels are compared to A5
  - We could reproduce estimates for the textbook parameterizations from the SAS estimates:

- **Whatever parameterization is used, we will still be looking to determine answers to the questions**:
  - **Is there a difference among the levels of the factor? (F-test)**
  - **Where do the differences lie? How big are the differences? (multiple comparisons)**

Analysis of Factor Level Means

- F-test is significant; there exist differences among the means. Now what?
- Want to determine which means are different. Form groups of means that are statistically the same.

Visual Assessment

- Can often get an idea by looking at plots
  - Side-by-side Box Plots
  - Plots of the factor level means (called Main Effects Plots)
  - Bar Graphs
- These plots do not give any information about the precision of the estimates. Need to consider standard errors.

ANOVA Models

- **Cell Means Model**

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{where} \quad \varepsilon_{ij} \sim iid \quad \mathrm{N}(0, \sigma^2)$$

- **Factor Effects Model**

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \text{where} \quad \varepsilon_{ij} \sim iid \quad \mathrm{N}(0, \sigma^2), \ \sum \tau_i = 0$$

Estimates

- Overall or grand mean: $\bar{Y}_{..} = \sum\sum Y_{ij} / n_T$
- Mean for factor level i: $\hat{\mu}_i = \bar{Y}_{i.} = \sum Y_{ij} / n_i$
- Factor Effects estimated by $\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$

Variances and Standard Errors

- All come from the fact that $Var(Y_{ij}) = \sigma^2$
- For the grand mean:

$$Var(\bar{Y}_{..}) = Var\left(\sum\sum Y_{ij} / n_T\right) = \sum\sum Var(Y_{ij}) / n_T^2 = \sigma^2 / n_T$$

- Pooled Estimate of σ^2:

$$S_P^2 = \frac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{\sum_{i=1}^{r}(n_i - 1)} = \frac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{n_T - r} = \frac{\sum_{i=1}^{r}\sum_{j}^{n_i}(Y_{ij} - \bar{Y}_{i.})}{n_T - r} = MSE$$

- Estimate by plugging in $\hat{\sigma}^2 = \mathrm{MSE}$

- **Cell Mean Variances**

  - For cell means (fixed level i):

$$Var(\bar{Y}_{i.}) = Var\left(\sum Y_{ij} / n_i\right) = \sum Var(Y_{ij}) / n_i^2 = \sigma^2 / n_i$$

  - Plug in $\hat{\sigma}^2$ = MSE to get the estimate

- **Standard Error of the Mean** (used to develop confidence intervals)

  - $\widehat{SE}(\bar{Y}_{..}) = \sqrt{MSE / n_T}$

  - $\widehat{SE}(\bar{Y}_{i.}) = \sqrt{MSE / n_i}$

  - **SAS**: Means, Standard Errors, CLM obtained using **glm procedure**

```
proc glm data=advert;
class adgroup;
model rating=adgroup;
means adgroup /clm t bon;
```

  - Output for (1- α)100% CLM for mean of factor level i, μ_i :

Differences Between Levels

- Consider the general pairwise comparison (difference between two means):

$$D = \mu_i - \mu_k$$

- Estimate D by

$$\hat{D} = \bar{Y}_{i.} - \bar{Y}_{k.}$$

- Since $\bar{Y}_i$ *and* $\bar{Y}_k$ are independent,

$$Var(\hat{D}) = \sigma^2 / n_i + \sigma^2 / n_k$$

- Standard Error for a difference

$$\widehat{SE}(\hat{D}) = \sqrt{MSE(1/n_i + 1/n_k)}$$

- Sampling distribution of the difference between the two sample means $\hat{D}$ is

$$\hat{D} \sim N\left(D,\ \sigma^2(1/n_i + 1/n_k)\right)$$

- The test statistic for testing $H_0: D = D_0$ vs $H_a: D \neq D_0$ is

$$\frac{\hat{D} - D_0}{\widehat{SE}(\hat{D})} \sim t_{n_T - r}$$

- Develop Hypotheses testing and CI for Differences between levels $D = \mu_i - \mu_k$
  - *Hypotheses stated as* : $\begin{cases} H_0: D=0 \\ H_a: D\neq 0 \end{cases}$ or $\begin{cases} H_0: \mu_i=\mu_k \\ H_a: \mu_i\neq\mu_k \end{cases}$
  - Test statistic: $t^* = \hat{D} / \widehat{SE}(\hat{D})$
  - CI given as: $\hat{D} \pm t_{crit}\widehat{SE}(\hat{D})$
- SAS

```
proc glm data=advert;
class adgroup;
model rating= adgroup;
means adgroup /t bon cldiff lines;
run;
```

  - "t", "bon" and "cldiff" requests confidence limits for the pairwise differences, based on the t-distribution and Bonferroni
  - "lines" requests a plot of the groupings
  - Output

Multiple Comparisons

- If we are only interested in doing one test, or looking at one confidence interval, all is well.

- Usually we are interested in looking at ALL pairwise comparisons. So again there are issues with Family Type I Error Rates.

- When there are r levels for the factor, there are $\frac{r(r-1)}{2}$ comparisons to be made. Determining the critical values involve dealing with multiplicity problem by adjusting t_{crit} .

○ Least Significant Differences (LSD)

- Least conservative of all the procedures, no adjustment made for multiple comparisons (Type I errors more likely)
- LSD – controls the test-alpha: $t_{crit} = t_{1-\alpha/2}(n_T - r)$
- The least significant difference is the minimum amount by which two means must differ in order to be considered statistically different. LSD is given by $LSD = t_{1-\alpha/2}(n_T - r)\sqrt{MSE(1/n_i + 1/n_k)}$
- Strongly significant F-test (say p-value < 0.01 or 0.005) preferred before looking at LSD comparisons.
- If the F-test indicates that a factor is significant, then any pair of means that differ by at least LSD are considered to be different.
- Use to develop hypothesis tests and confidence intervals
  - ✓ For any difference in means D, testing $H_0 : D = 0$ vs $H_a : D \neq 0$
  - ✓ (1-α)100 % CI is given by $\left(\bar{Y}_{i.} - \bar{Y}_{k.}\right) \pm t_{1-\alpha/2}(n_T - r)\sqrt{MSE(1/n_i + 1/n_k)}$
- **SAS**: Use either 't' or 'lsd' in MEANS statement in proc glm

○ Tukey

- Simple pair-wise comparisons can be accomplished all at once using Tukey adjustments. Specifies an EXACT family significance level for comparing ALL PAIRS OF TREATMENT MEANS.
- Tukey (based on studentized range distribution): $t_{crit} = \frac{q_{1-\alpha}(r,\ n_T - r)}{\sqrt{2}}$
- Is more conservative (and generally more appropriate than LSD)
- Minimum significant difference is given by $\frac{q_{1-\alpha}(r,\ n_T - r)}{\sqrt{2}}\sqrt{MSE(1/n_i + 1/n_k)}$
- Use to develop hypothesis tests and confidence intervals
  - ✓ For any difference in means D, testing $H_0 : D = 0$ vs $H_a : D \neq 0$
  - ✓ $(1\text{-}\alpha)100$ % CI is given by $\left(\bar{Y}_{i.} - \bar{Y}_{k.}\right) \pm \frac{q_{1-\alpha}(r,\ n_T - r)}{\sqrt{2}}\sqrt{MSE(1/n_i + 1/n_k)}$
- **SAS**: Use 'tukey' in MEANS statement in proc glm

○ Bonferroni

- Sacrifices slightly more power than TUKEY, but can be applied to any set of contrasts or linear combinations (useful in more situations than Tukey).
- Usually better than Tukey for a small number of planned comparisons
- For g=r(r-1)/2 pairwise comparisons, Bonferroni adjustment is given by

  $t_{crit} = t_{1-\alpha/2g}(n_T - r)$
- A Bonferroni adjustment works well when r is small (2, 3, or 4). For r > 4, starts to get much more conservative than necessary.

- For all g = r(r-1)/2 pairwise comparisons, minimum significant difference is given by $t_{1-\alpha/2g}(n_T - r)\sqrt{MSE(1/n_i + 1/n_k)}$
- Use to develop hypothesis tests and confidence intervals
  - ✓ For any difference in means D, testing $H_0 : D = 0$ vs $H_a : D \neq 0$
  - ✓ $(1\text{-}\alpha)100$ % CI is given by $\left(\bar{Y}_{i.} - \bar{Y}_{k.}\right) \pm t_{1-\alpha/2g}(n_T - r)\sqrt{MSE(1/n_i + 1/n_k)}$
- **SAS**: Use 'bon' in MEANS statement in proc glm

○ Scheffé

- Most conservative (least powerful) of all tests. Protects against data snooping!
- Controls the family alpha level for testing all possible contrasts
- Should be used if you have not planned contrasts in advance.
- For testing pairs of treatment means it is too conservative (should use Tukey or Bonferroni)
- Based on F-distribution; Critical value is $t_{crit} = \sqrt{(r-1)F_{1-\alpha}(r-1,\ n_T - r)}$
- Minimum significant difference is given by
  $\sqrt{(r-1)F_{1-\alpha}(r-1,\ n_T - r)}\sqrt{MSE(1/n_i + 1/n_k)}$
- Use to develop hypothesis tests and confidence intervals
  - ✓ For any difference in means D, testing $H_0 : D = 0$ vs $H_a : D \neq 0$
  - ✓ (1-α)100 % CI is given by

$$\left(\bar{Y}_{i.} - \bar{Y}_{k.}\right) \pm \sqrt{(r-1)F_{1-\alpha}(r-1,\ n_T - r)}\sqrt{MSE(1/n_i + 1/n_k)}$$

- **SAS**: Use 'scheffee' in MEANS statement in proc glm

○ Summary

- LSD is too liberal (get Type I errors / CI's too narrow).
- Scheffe is conservative (no power for certain comparisons / CI's wide).
- Bonferroni is OK for small *r* (but conservative for large *r*).
- Tukey is recommended for general use.

Linear Combinations of Means

- Often we test hypotheses about or make CI's for particular linear combinations of the group means that come from research questions, not from an examination of the data.
- A linear combination of means is any quantity of the form $L = \sum c_i \mu_i$ for any constants c_i. L is estimated by

  $\hat{L} = \sum c_i \bar{Y}_i$ and with $E(\hat{L}) = L$ and $Var(\hat{L}) = \sigma^2 \sum (c_i^2 / n_i)$
- The variance is estimated by $\widehat{\text{Var}}(\hat{L}) = s^2 \sum (c_i^2 / n_i)$
- Tests of hypotheses and CI can be done for any L

Contrasts

- A contrast is a special case of a linear combination with $\sum_i c_i = 0$. These turn out to be particularly useful because the interesting hypothesis tests are of the form $H_0 : L = 0$.
- Example 1: $\mu_1 - \mu_2$ $\quad (c_1 = 1, c_2 = -1)$
- Used to test whether levels 1 and 2 have equal means.
- Example 2: $\mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$ $\quad (1, -0.5, -0.5)$
- Used to test whether level 1 has the same mean as the combination of levels 2/3.
- Example 3: $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$ $\quad (0.5, 0.5, -0.5, -0.5)$
- Used to test whether the first two levels have the same mean as the last two (think $1, 2$ = men; $3, 4$ = women and $1, 3$ = diet A; $2, 4$ = diet B - this would then test for gender differences)

Multiple Contrasts

We can simultaneously test a collection of contrasts (1 df each contrast)
Example 1, $H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$
The F statistic for this test will have an $F_{1, n_T - r}$ distribution
Example 2, $H_0 : \mu_2 = \mu_3 = \mu_4$.
The F statistic for this test will have an $F_{2, n_T - r}$ distribution
We do this by setting up one contrast for each comparison and doing them simultaneously.

```
PROC GLM data=advert;
title 'One-way CRD - Linear Contrasts';
     CLASS ADGROUP;
       MODEL RATING = ADGROUP;
        means ADGROUP / alpha=0.05 cldiff lines bon tukey;
                     CONTRAST 'AD_D VS OTHERS'
                          ADGROUP -.25 -0.25 -0.25 1 -0.25;
                     CONTRAST 'AD_A VS AD_E'
                          ADGROUP 1 0 0 0 -1;
RUN;
quit;
title;
```

Dependent Variable: RATING

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
AD_D VS OTHERS	1	20.0083333	20.0083333	2.09	0.1610
AD_A VS AD_E	1	208.3333333	208.3333333	21.73	<.0001

Residuals

- Predicted values are cell means: $\hat{Y}_{i,j} = \bar{Y}_{i.}$.
- Residuals are the differences between the observed values and the cell means $e_{i,j} = Y_{i,j} - \bar{Y}_{i.}$.

Basic plots

- Plot the data vs the factor levels (the values of the explanatory variables)
- Plot the residuals vs the factor levels
- Construct a normal quantile plot of the residuals

Notice that we are no longer checking for linearity since this is not an assumption in ANOVA.

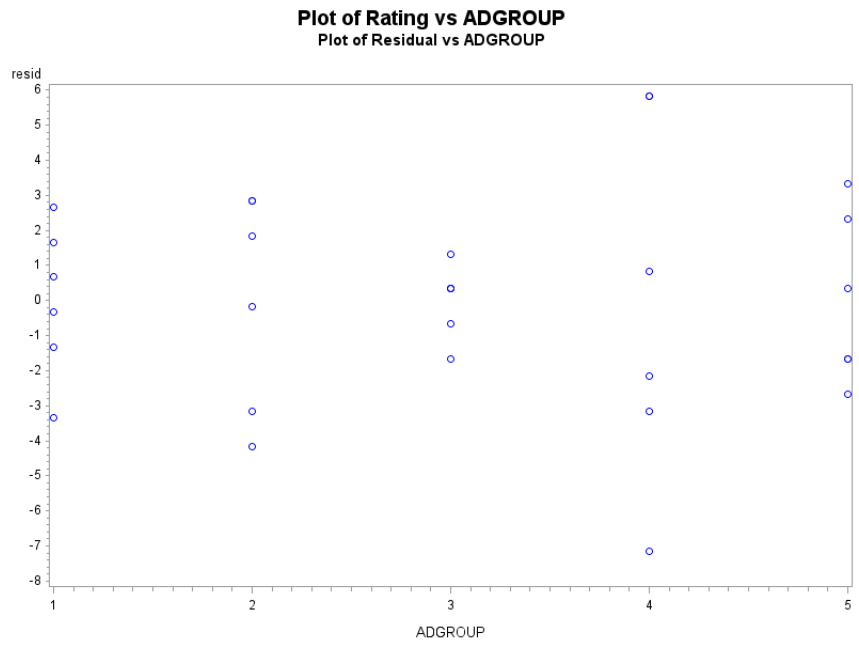
```
proc glm data=advert noprint;
     class ADGROUP;
          model RATING = ADGROUP;
               output out=resout r=resid;
```

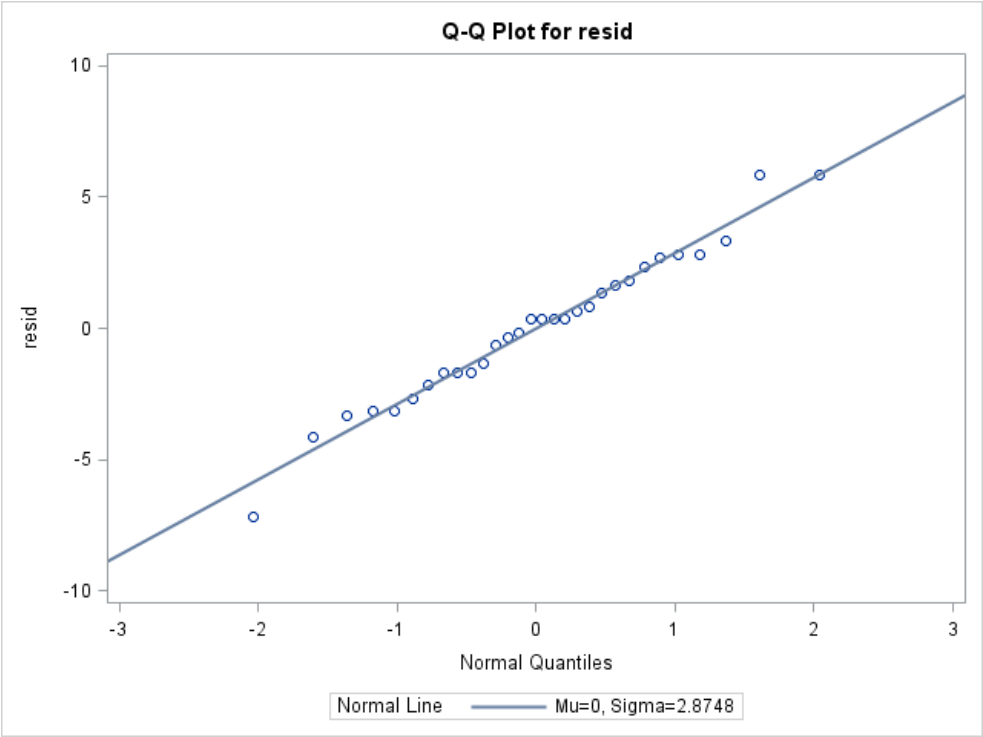
```
run;
quit;


symbol1 v=circle h=1 c=blue i=none;
proc gplot data=resout;
title1 'Plot of Rating vs ADGROUP';
title2 'Plot of Residual vs ADGROUP';
     plot (rating resid)*adgroup;
run;
quit;
title;


/** Normal quantile plot of the residuals  **/

proc univariate data = resout;
title 'Q-Q Plot Residuals';
qqplot resid / normal (L=1 mu=est sigma=est);
run;
quit;
title;
```





Summary of plot diagnostics

Look for

- Outliers
- Variance that depends on level
- Non-normal errors

Plot residuals vs time and other variables

Homogeneity tests

Homogeneity of variance (homoscedasticity) is assumed in the model. We can test for that.

$$\begin{aligned} \mathrm{H}_0: &\quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2 \quad \text{(constant variance)} \\ \mathrm{H}_A: &\quad \text{not all } \sigma_i^2 \text{ are equal} \quad \text{(non-constant variance)} \end{aligned}$$

- Several significance tests are available. Note that this was also available in regression if each X has multiple Y observations (this is usually true in ANOVA): see section 3.6.
- Text discusses Hartley, modified Levene.
- SAS has several including Bartlett's (essentially the likelihood ratio test) and several versions of Levene.

```
/** constant variance - Modified Levene's HOVTest **/

proc glm data=advert;
title "Levene's HOVTEST";
     class ADGROUP;
          MODEL RATING = ADGROUP;
               means adgroup/hovtest=levene(type=abs);
run;
quit;
title;
```

Levene's Test for Homogeneity of RATING Variance
ANOVA of Absolute Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ADGROUP	4	37.8148	9.4537	4.40	0.0079
Error	25	53.7037	2.1481		

Remedies

- Delete outliers – Is their removal important?
- Use weights (weighted regression)
- Transformations
- Nonparametric procedures

Transformation Guides

Transformations can also be used to solve constant variance problems, as well as normality.

- When σ_i^2 is proportional to μ_i, use $\sqrt{Y}$.
- When σ_i is proportional to μ_i, use $\log(Y)$.
- When σ_i is proportional to μ_i^2, use $1/Y$.
- When Y is a proportion, use $2\arcsin(\sqrt{Y})$; this is `2*arsin(sqrt(y))` in a SAS `data` step.
- Can also use Box-Cox procedure.

Issues of Unequal Sample Sizes

- Encountered for a variety of reasons including:
  - Convenience – usually if we have an observational study, we have very little control over the cell sizes.
  - Cost Effectiveness – sometimes the cost of samples is different, and we may use larger sample sizes when the cost is less
  - In experimental studies, you may start with a balanced design, but lose that balance if problems occur.
- What changes?
  - Loss of balance brings “intercorrelation” among the predictors (i.e., variables are no longer orthogonal)
  - Type I and III SS will be different; typically Type III SS should be used for testing
  - LSMeans should be used for testing
  - Standard errors for cell means and for multiple comparisons will be different
  - Confidence intervals will have different widths