Section VI

Comparing means & analysis of variance

VII – Analysis of variance

Analysis of variance (ANOVA) refers to methods for comparing **means**. It can also be thought of as a special case of linear regression where all of the predictor variables (Xs) take on categorical values. Gender, diagnosis and ethnicity are examples of categorical predictors. The outcome, Y, is continuous. In comparing means with ANOVA, as opposed to doing lots of t tests, the results (SEs, p values) are based on a pooled SD, not the individual SDs.

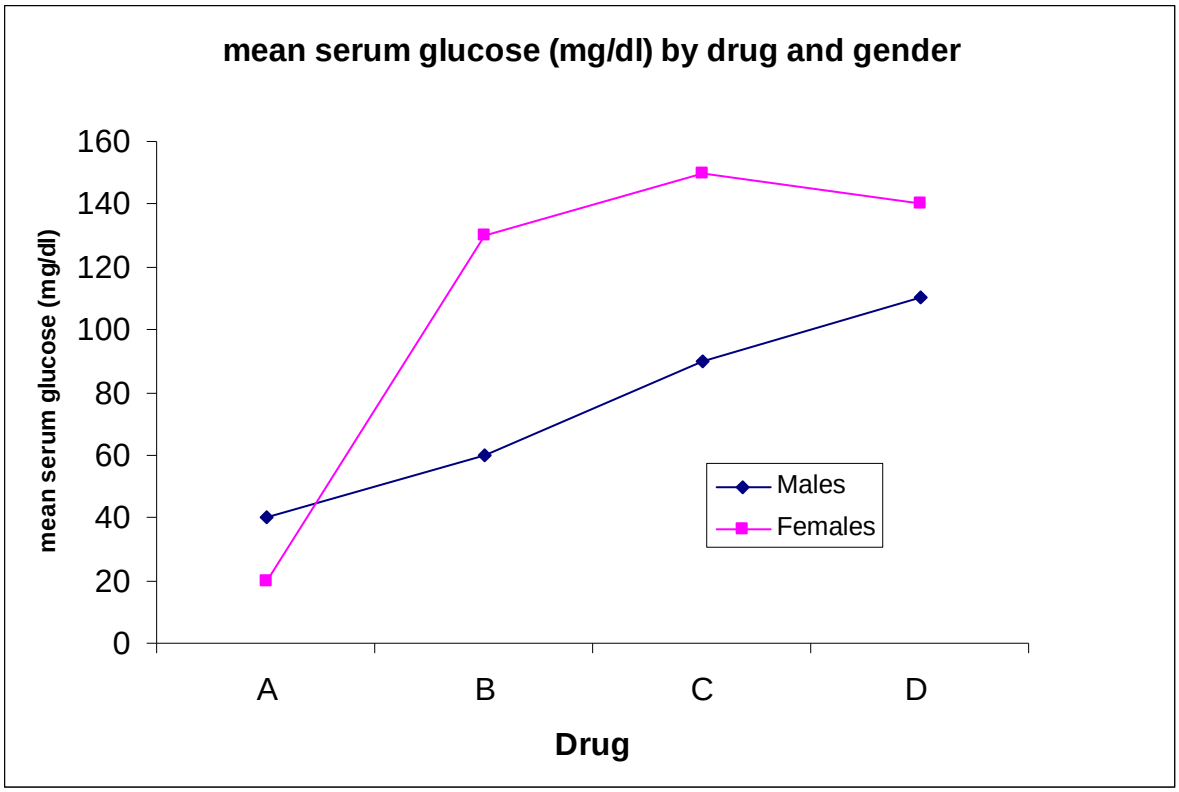
In a study design with several factors, each cross classified combination of the factors (predictors) form a "cell". For example if the predictors are sex (male or female) and dementia (yes or no), there are four possible cross classified categories or four "cells" (males without dementia, males with dementia, females without dementia, females with dementia) each with their cell mean for Y. In balanced ANOVA, the sample size is **the same** for every cross classified combination of the X predictors. That is, the sample size is the same in every cell.

When properly coded, predictor variables in balanced ANOVA models are all uncorrelated. The uncorrelated variables are called "orthogonal" since this is an "artificially" induced zero correlation.

	Males	Females	Overall
Dementia	Cell	Cell	Margin
No dementia	Cell	Cell	Margin
Overall	Margin	Margin	

Presenting means - ANOVA data

Since all of the factors are discrete, it is often easy and strongly desirable to make graphs of the means as a function of the factors.



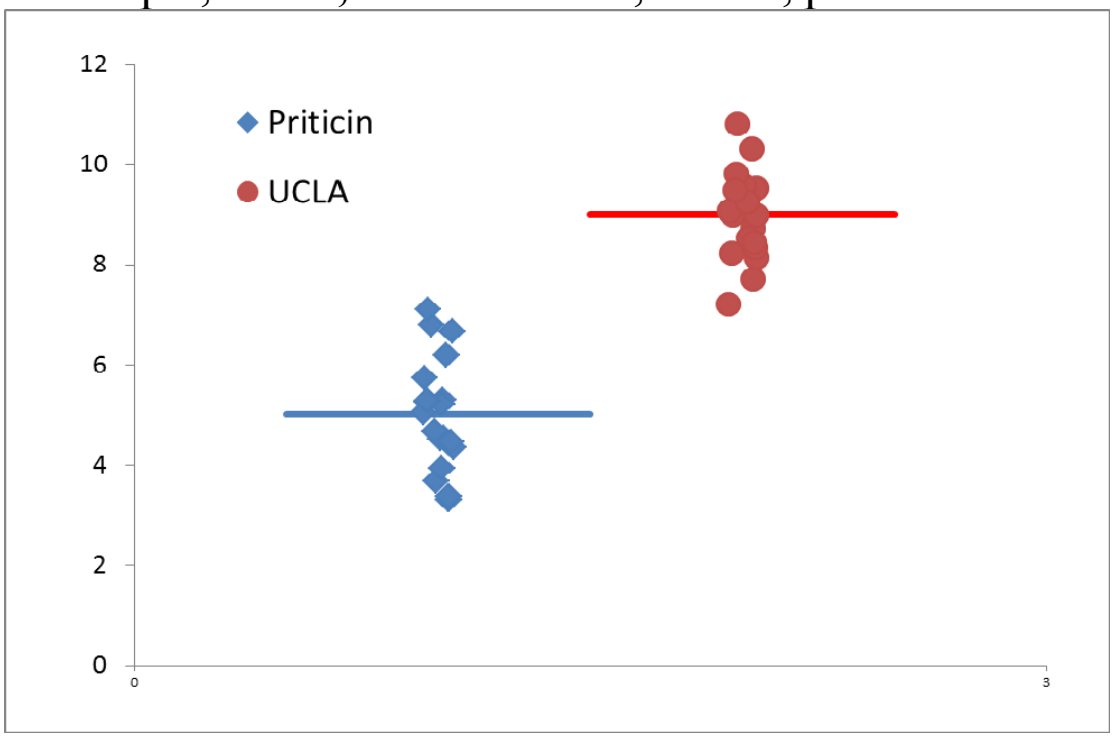
One can also add "error bars" to these means. In analysis of variance, these error bars are based on the sample size and the **pooled** standard deviation, SD_e. This SD_e is the same residual SD_e as in regression.

When comparing means, the "yardstick" is critical.

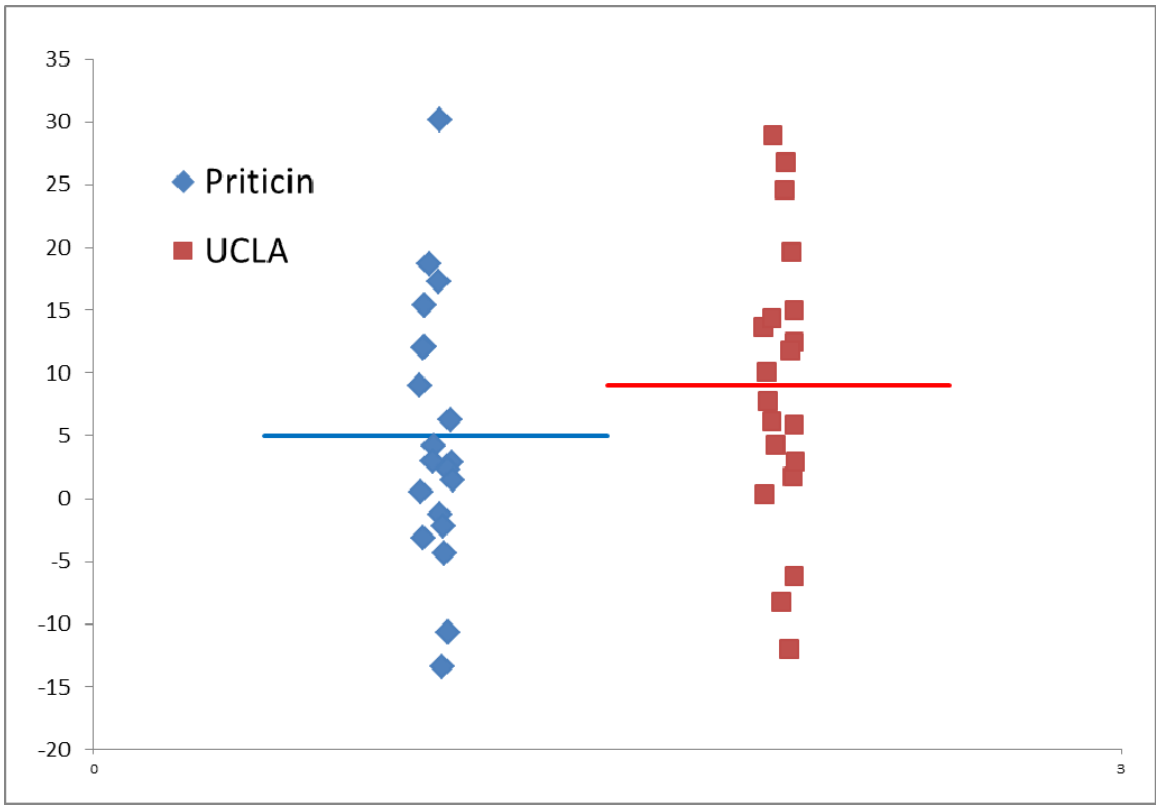
In the weight loss comparison below, is a 4 lb difference "big"? Compared to what? What is the "yardstick"?

Diet	mean weight loss (lbs)	n
Pritikin	5.0	20
UCLA GS	9.0	20
mean difference	4.0	

Example, **SD=1,** SE diff = 0.32, t=12.6, p value < 0.001



What if the SD changes? In the example below, the means are still 4 lbs apart and have not changed but **SD=5**, SE diff=1.58, t=2.5, p value=0.02



The p value is smaller since the SD is larger. A larger SD makes the SE diff larger and the p value larger.

Comparing Means
Two groups – t test (review)

Mean differences are judged "statistically significant" (different beyond chance) relative to their **standard error** (SE_d), a measure of mean variability ("noise").

$$t = \frac{(\overline{Y_1} - \overline{Y_2})}{SE_d} = \frac{\text{"signal"}}{\text{"noise"}}$$

$\overline{Y}_i$ = mean of group i, SE_d=standard error of mean difference

The t statistic is the mean difference in SE_d units. The **p value** is a function of the t statistic. As |t| increases, p value gets smaller.

Rule of thumb: statistical significance, defined as p < 0.05, is achieved if |t| > 2

$$\text{or } |\overline{Y}_1 - \overline{Y}_2| > t_{cr}\, SE_d = 2\, SE_d = \text{LSD}$$

$t_{cr}\, SE_d = 2\, SE_d$ is sometimes referred to as the critical distance or the LSD=least significant difference.

So, getting the correct SE_d is crucial!!

The SE is the "yardstick" for significance and depends on:
a) the mean difference, b) the SD=the individual variability, c) the sample size.

How to compute SE_d (review)

SE_d depends on n, SD and study design.
(study design example: factorial or repeated measures)

For a single mean, if n=sample size

$$SEM = SD/\sqrt{n} \quad = \quad \sqrt{SD^2/n}$$

For a mean <u>difference</u> $(\overline{Y_1} - \overline{Y_2})$

The SE of the mean difference, SE_d is given by

$$SE_d = \sqrt{[\, SD_1^2/n_1 \; + SD_2^2/n_2 \,]} \quad \text{or}$$

$$SE_d = \sqrt{[SEM_1^2 \; + SEM_2^2]}$$

If data is <u>paired</u> (before-after), first compute differences ($d_i=Y_{2i}-Y_{1i}$) for each person

For paired data: $SE_d = SD(d_i)/\sqrt{n}$

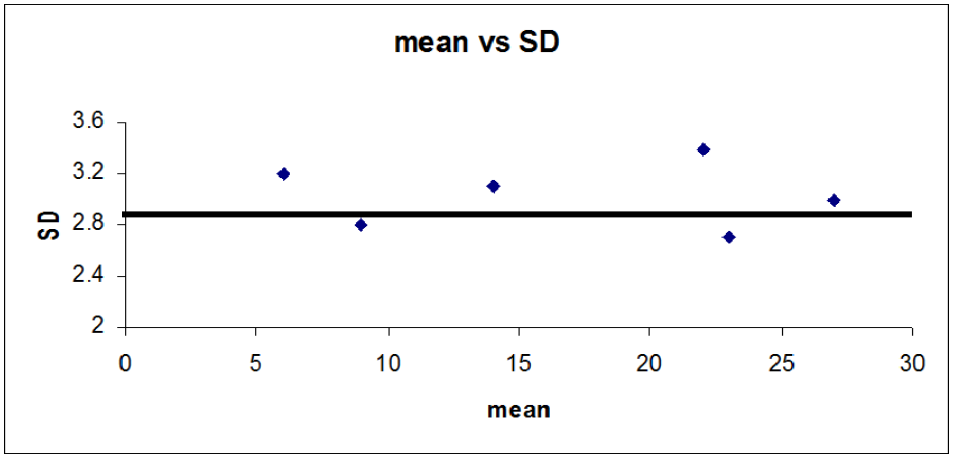
3 or more groups-analysis of variance (ANOVA)
Pooled SDs

What if we have many treatment groups, each with its own mean and SD?

Group	Mean	SD	sample size (n)
A	$\bar{Y}_1$	SD_1	n_1
B	Y_2	SD_2	n_2
C	Y_3	SD_3	n_3
…			
k	$\overline{Y_k}$	SD_k	n_k

Usually, at least on some scale, (perhaps the log scale) there is a single true underlying SD, called σ, for <u>all</u> groups. The observed SDs, SD_1, SD_2, … SD_k vary around σ due to "chance" (random variation). This will be true when the variability is caused by the equipment or the experimental conditions, not by the treatments / groups.

One can check visually for "variance homogeneity" by plotting "X"=means versus "Y"=SDs. Should get a scatter about a horizontal line at σ if SD homogeneity is true.



If the constant variance assumption is reasonable, then the best thing to do is to **pool** ALL of the sample SDs into a single common estimate, the pooled SD_e. **This is the main idea for an analysis of variance (as opposed to a bunch of t tests).**

When the individual SDs only vary randomly, the SD_e is more accurate than any of the individual SDs and thus gives more accurate standard errors for means and mean differences. It also provides a common "yardstick".

$$SD^2_{\text{pooled error}} = SD^2_e =$$

$$\frac{(n_1-1)\,SD_1^2 + (n_2-1)\,SD_2^2 + \ldots (n_k-1)\,SD_k^2}{(n_1-1) + (n_2-1) + \ldots (n_k-1)}$$

$$\text{so, } SD_e = \sqrt{SD^2_e}$$

In ANOVA - we use this pooled SD_e to compute SE_d and to compute "post hoc" (post pooling) t statistics and p values.

$$SE_d = \sqrt{[\,SD_1^2/n_1 + SD_2^2/n_2\,]}$$

$$= SD_e \sqrt{(1/n_1) + (1/n_2)}$$

since SD_1 and SD_2 are replaced by SD_e a "common yardstick".

Note: If $n_1=n_2=n$, then $SE_d = SD_e\sqrt{2/n}$=constant

Comparing means – post hoc t test under ANOVA

The usual t test is: $\mathbf{t} = (\overline{Y}_1 - \overline{Y}_2)/SE_d$

where $SE_d = \sqrt{[\ SD_1^2/n_1 + SD_2^2/n_2\ ]}$

Under ANOVA , $SD_1=SD_2=\ldots\ SD_k = SD_e = SD_{pooled}$

So, under ANOVA, **t** is as above except SD_e, is used in place of SD_1 & SD_2 for any comparison between two means.

So, $SE_d = \sqrt{SD_e^2/n_1 + SD_e^2/n_2} = SD_e\sqrt{(1/n_1+1/n_2)}$.

If $n_1=n_2=\ldots n_k=n$, $\mathbf{SE_d = SD_e\sqrt{2/n}}$ – a constant for all k mean comparisons.

Multiplicity & F tests

Multiple testing can create "false positives". That is, we can incorrectly declare that means are "significantly" different as an artifact of doing many tests even if none of the means are truly different beyond chance.

Imagine we have k=four groups: A, B, C and D.

There are six possible mean comparisons:

A vs B
A vs C
A vs D
B vs C
B vs D
C vs D

If we use $p < 0.05$ as our "significance" criterion, we have a 5% chance of a "false positive" mistake for any one of the six comparisons, assuming that **none** of the groups are really different from each other. We have a 95% chance of no false positives if none of the groups are really different. So, the chance of a "false positive" in **any** of the six comparisons is $1 - (0.95)^6 = 0.26$ or 26%.

To guard against this we first compute the "overall" F statistic and its p value, a "screening" test.

The overall (omnibus) F statistic compares all k group means to the overall mean (M).

$$F = \frac{\Sigma\, n_i(\overline{Y_i} - M)^2/(k-1)}{(SD_e)^2} = \frac{MS_x}{MS_{error}} = \frac{\text{between group var}}{\text{within group var}}$$

$$= \frac{[n_1(\overline{Y_1} - M)^2 + n_2(\overline{Y_2} - M)^2 + \ldots n_k(\overline{Y_k} - M)^2]/(k-1)}{(SD_e)^2}$$

MS=mean square

If the "overall" p value corresponding to F is $p > 0.05$, we conclude that none of the mean differences are beyond what would be expected by chance (not "significant") and so we do not have to examine all of the individual mean comparisons. Only if the overall $p < 0.05$ will some or all of the individual comparisons have no more than an **overall** 5% chance of a "false positive".

This criterion was suggested by RA Fisher and is called the Fisher LSD (least significant difference) criterion. It is less conservative (has fewer false negatives) than the very conservative Bonferroni criterion. Bonferroni criterion: if making "m" comparisons, declare significant only if $p < 0.05/m$.

F is the ratio of between group variation to (pooled) within group variation. This is why this method is called "analysis of variance".

$$\frac{\text{Between group variation}}{\text{Within group variation}}$$
$$\text{Total variation}$$

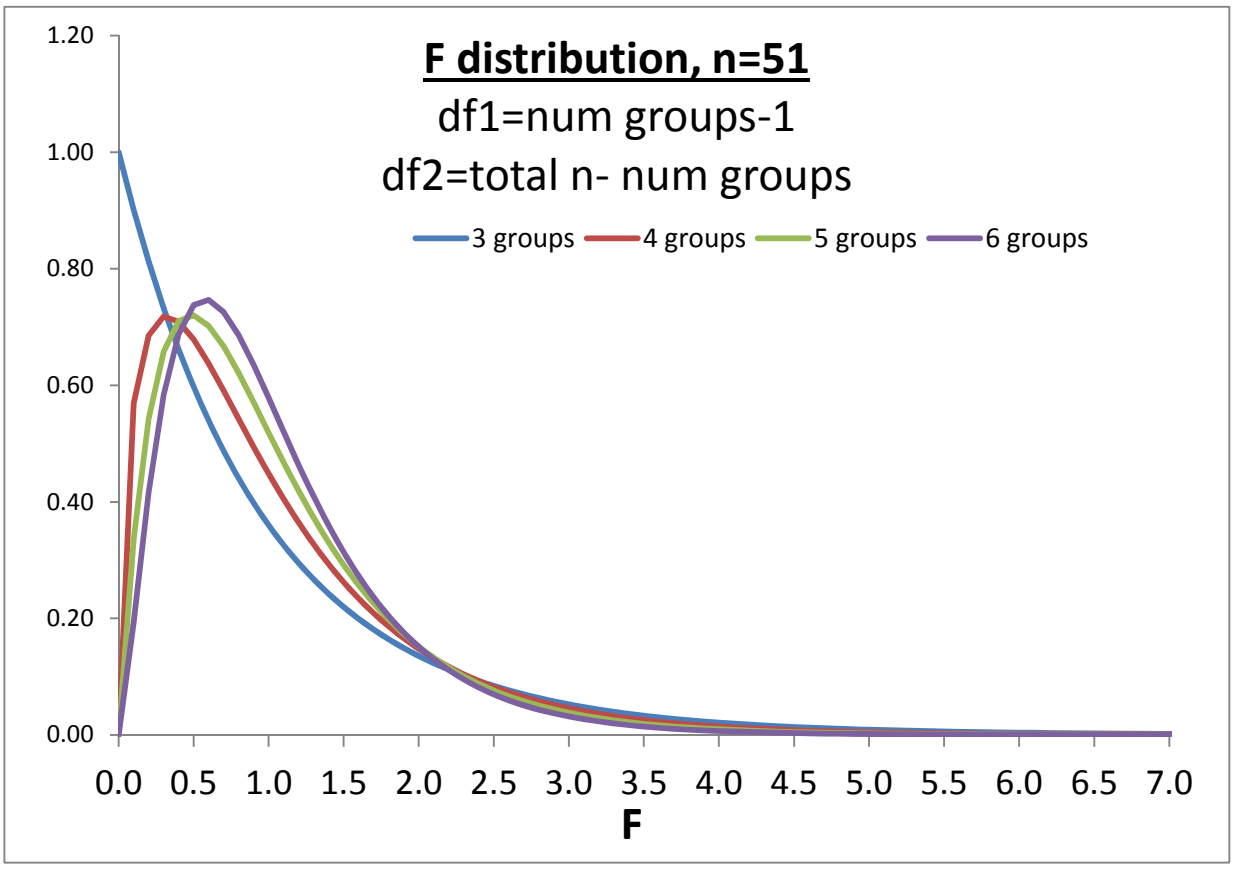
F = Between var / Within var

$F \approx 1$ when all of the means are about the same.

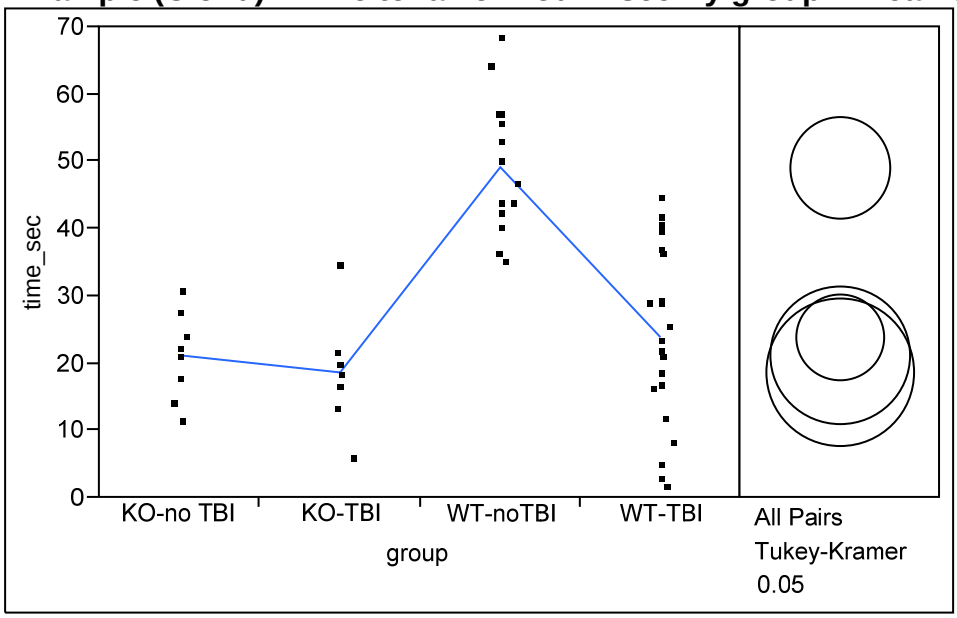
The F distribution –

The distribution of F under the null hypothesis depends on:

df 1 = number of groups -1
df 2 = total sample size – number of groups



Example (Clond) - Time to fall off rod in sec By group – mean comparisons



Means and Std Deviations – *no ANOVA model*

group	n	Mean	median	Std Dev	Std Err Mean
KO-no TBI	8	21.1963	21.65	6.4598	2.2839
KO-TBI	7	18.6586	18.47	8.7316	3.3002
WT-noTBI	15	49.1973	46.93	9.9232	2.5622
WT-TBI	21	23.9019	23.33	13.3124	2.9050

Oneway ANOVA model

R square	0.579811	
Adj Rsquare	0.55299	
Root Mean Square Error=SD_e	10.98601	← Pooled SD_e
Mean of Response	30.19765	
Observations (or Sum Wgts)	51	

Analysis of Variance Table

Source	DF	Sum of Squares	Mean Square	F Ratio	p value
group	3	7827.438	2609.15	**21.6181**	**<.0001***
Error	47	5672.546	120.69		
Total	50	13499.984			

Means for Oneway ANOVA - Means same but SEs are not

group	N	Mean	Std Error Mean
KO-no TBI	8	21.1963	3.8841
KO-TBI	7	18.6586	4.1523
WT-noTBI	15	49.1973	2.8366
WT-TBI	21	23.9019	2.3973

Standard errors computed using a pooled SD_e.

Means are the same whether an ANOVA model is used or not but the SD_e and standard errors of the mean (Std Error) are not the same.

Multiple mean comparisons
Tukey's "studentized" range – q

Using the F test as a screen and the Fisher LSD criterion is sometimes not optimal for all pairwise comparisons among k means.

For "k" means, Tukey computed the distribution of $\mathbf{q}=[\max(Y_1, Y_2,\ldots Y_k) - \min(Y_1, Y_2, \ldots Y_k)]/SE_d =$ $\mathbf{q}$= mean **range**/SE_d under the null hypothesis that all k population means are equal. The percentile q is called the "studentized range" percentile.

The mean difference is significant if it is larger than $\mathbf{q}\ \mathbf{SE_d}$ (instead of t SE_d or Z SE_d). This criterion keeps the overall false pos level $\leq \alpha$ for all pairwise comparisons. When the sample size is the same for all groups, this is the best criterion (the "exact" solution) assuming the means follow the Gaussian.

Percentiles for t vs q for $p < 0.05$ (97.5th percentile), large n

k = num means	t≈Z	q*
2	1.96	1.96
3	1.96	2.34
4	1.96	2.59
5	1.96	2.73
6	1.96	2.85

*some tables give q for SE, not SE_d, so must multiply q by $\sqrt{2}$.

One looks up "t" on the q table instead of the t table.

Means Comparisons for all pairs using Tukey-Kramer HSD

group			Mean
WT-noTBI	A		49.197333
WT-TBI		B	23.901905
KO-no TBI		B	21.196250
KO-TBI		B	18.658571

Levels not connected by same letter are significantly different.

In this example, the WT-no TBI group mean is significantly higher than the means of the other three groups.

group	vs group	Mean Difference	Std Err Dif	p-Value-Tukey
WT-noTBI	KO-TBI	30.53876	5.028712	<.0001*
WT-noTBI	KO-no TBI	28.00108	4.809649	<.0001*
WT-noTBI	WT-TBI	25.29543	3.713950	<.0001*
WT-TBI	KO-TBI	5.24333	4.794689	0.6952
WT-TBI	KO-no TBI	2.70565	4.564408	0.9338
KO-no TBI	KO-TBI	2.53768	5.685802	0.9700

group	vs group	Mean Difference	SE diff	t	p-Value- no correction	p-Value-Tukey
WT-noTBI	KO-TBI	30.54	5.03	6.073	<.0001*	<.0001*
WT-noTBI	KO-no TBI	28.00	4.81	5.822	<.0001*	<.0001*
WT-noTBI	WT-TBI	25.30	3.71	6.811	<.0001*	<.0001*
WT-TBI	KO-TBI	5.24	4.79	1.094	0.2797	0.6952
WT-TBI	KO-no TBI	2.71	4.56	0.593	0.5562	0.9338
KO-no TBI	KO-TBI	2.54	5.69	0.446	0.6574	0.9700

The Tukey p values are larger than the p values without the Tukey correction. The Tukey p values are computed such that the chance of a type I (alpha) error for all six possible comparisons of the four means is no more than alpha=0.05 (two sided). The uncorrected p values do not take this into account.

Sums of Squares (SS) and summary ANOVA table

Most ANOVA & regression software prints a table like the one below, the "summary" analysis of variance table.

Analysis of Variance table for time on rod data

Source	DF	Sum of Squares	Mean Square	F Ratio	p value
Model	3	7827.438	2609.15	21.618	< 0.0001
Error	47	5672.546	120.69		
Total	50	13499.984			

SD_e^2

This table shows how much of the variation in the outcome Y (time on rod) is accounted for by the "model", that is, by the groups (WT or KO strain & TBI/no TBI in this example), and how much variation in Y is not accounted for, the "error" sum of squares (SS) variation.

For a given dataset, the total SD of Y (SD_y) and the variance of Y ($=SD_y^2$) is fixed. The total sum of squares (SS_{total}) for Y, defined as $SS_{total}= (n-1)SD_y^2$, is also fixed and is a measure of how much Y varies, the total "information" available.

The table below is for k groups. In the example above, k=4 since there are four groups (WT-no TBI, WT-TBI, KO-no TBI, KO-TBI)

	df	**Sum of Squares=SS**	**Mean Square=MS=SS/df**
Model	**k-1**	$\sum n_i(\bar{y}_i - \bar{y})^2$	
Error	**n-k**	$\sum e^2 = (n-k)\, SD_e^2$	SD_e^2
Total	**n-1**	$\sum (y - \bar{y})^2 = (n-1)SD_y^2$	SD_y^2

Above, $\bar{y}$ is the grand mean Y and $\bar{y}_i$ is the mean in the i^{th} group, i=1,2,…k.

R^2 = Model SS /Total SS = 0.5798

F = Model Mean Square/ Error Mean Square=21.61.

The p value from this "overall" F tests the null hypothesis that all of the true population group means are the same as the population grand mean and therefore are all the same as each other.

Transformations

There are two main requirements for the analysis of variance (ANOVA) model.

1. Within any treatment group, the mean should be the middle value. That is, the mean should be about the same as the median. When this is true, the data can usually be reasonably modeled by a Gaussian ("normal") distribution around the mean in each group (each cell).

2. The SDs should be similar (variance homogeneity) from group to group.

One can plot means vs medians & residual errors to check #1 and means versus SDs to check #2. But what if either requirement is not true? In this case there are two options:

a. Find a transformed scale where it is true.

b. Don't use the usual ANOVA model. Use a non constant variance ANOVA model or non parametric models that do not assume the data follows a normal distribution.

Option "a" is better if possible as it provides more statistical power. The most common transform is the **log** transformation. It usually works for

1. Radioactive count data
2. Titration data (titers), serial dilution data
3. Cell, bacterial, viral growth, CFUs
4. Steroids & hormones (E2, Testosterone, …)
5. Power data (decibels, earthquakes)
6. Acidity data (pH), …
7. Liver enzymes (bilirubin, Creatinine)

In general, log transforms works when a multiplicative phenomena is transformed to an additive phenomena. One can compute stats on the log

scale & "back transform" results to original scale for final report. Since log(A)–log(B)=log(A/B), mean differences on the log scale correspond to mean **ratios** on the original scale. Remember

$$10^{\text{mean(log data)}} = \text{geometric mean} < \text{arithmetic mean}$$

When the log transformation does not work, only can try any of the transformations below. If none work, then non parametric methods are needed.

$$\underline{\text{monotone transformation ladder}}$$

$$Y^2,\ Y^{1.5},\ Y^1,\ Y^{0.5}=\sqrt{Y},$$

$$Y^0=\log(Y),$$

$$Y^{-0.5}=1/\sqrt{Y},\ Y^{-1}=1/Y, Y^{-1.5},\ Y^{-2}$$

Section VIb
Multiway ANOVA

Brain weight data example

Two factors may influence brain weight:

Dementia (yes or no)
Sex (male or female)

2 x 2 = four groups, 7 subjects per group

Dementia	Sex	Brain Weight (gm)
No	F	1223
No	F	1228
No	F	1222
No	F	1204
No	F	1234
No	F	1211
No	F	1217
…	…	…

Mean brain weights (gms) in Males & Females with and without dementia

A "balanced" 2 x 2 ANOVA design

7 persons per cell, 7 x 4 = 28 persons total

cell

Dementia	Males (1)	Female (-1)	Margin
Yes (1)	1321.14	1201.71	1261.43
No (-1)	1333.43	1219.86	1276.64
Margin	1327.29	1210.79	1269.04

Difference in marginal sex means (Male – Female)
1327.29 - 1210.79 = 116.50, 116.50/**2** = 58.25

Difference in marginal dementia means (Yes – No)
1261.43 - 1276.64 = -15.21, -15.21/**2** = -7.61

Differences in cell mean differences=**interaction**
(1321.14 - 1333.43) – (1201.71 - 1219.86) = 5.86
(1321.14 - 1201.71) – (1333.43 - 1219.86) = 5.86
note: 5.86/(**2x2**) = 1.46

balance = same sample size in every cell

ANOVA table for 2 x 2 design - sex x dementia

Effect coding (-1,1) - balanced design

MODEL: brain wt = sex dementia sex*dementia

Class	Levels	Values	
sex	2	-1 1	
dementia	2	-1 1	28 observations

Source	DF	Sum of Squares	Mean Square	F Value	p value
Model	3	96686	32228.70	451.05	<.0001
Error	24	1715	71.45 = SD^2_e		
C Total	27	98402			

R-Square	Coeff Var	Root MSE	Mean brain wt
0.9826	0.666092	8.453=SD_e	1269.04

ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	p value
sex	1	95005.75	95005.75	1329.64	<.0001
dementia	1	1620.32	1620.32	22.68	<.0001
sex*dementia	1	60.04	60.04	0.84	0.3685

The "sums of squares" are squared functions of the differences among the means for a given factor. If all means are the same, all mean differences are zero and the sum of squares (SS) is zero.

	Sum of squares= SS= n(mean diff)^2
Sex	58.25^2 x 28 = 95005.75
Dementia	7.61^2 x 28 = 1620.32
Sex-dementia	1.46^2 x 28 = 60.04

$$\text{mean of (k) means} = \sum \text{mean}_i / k$$

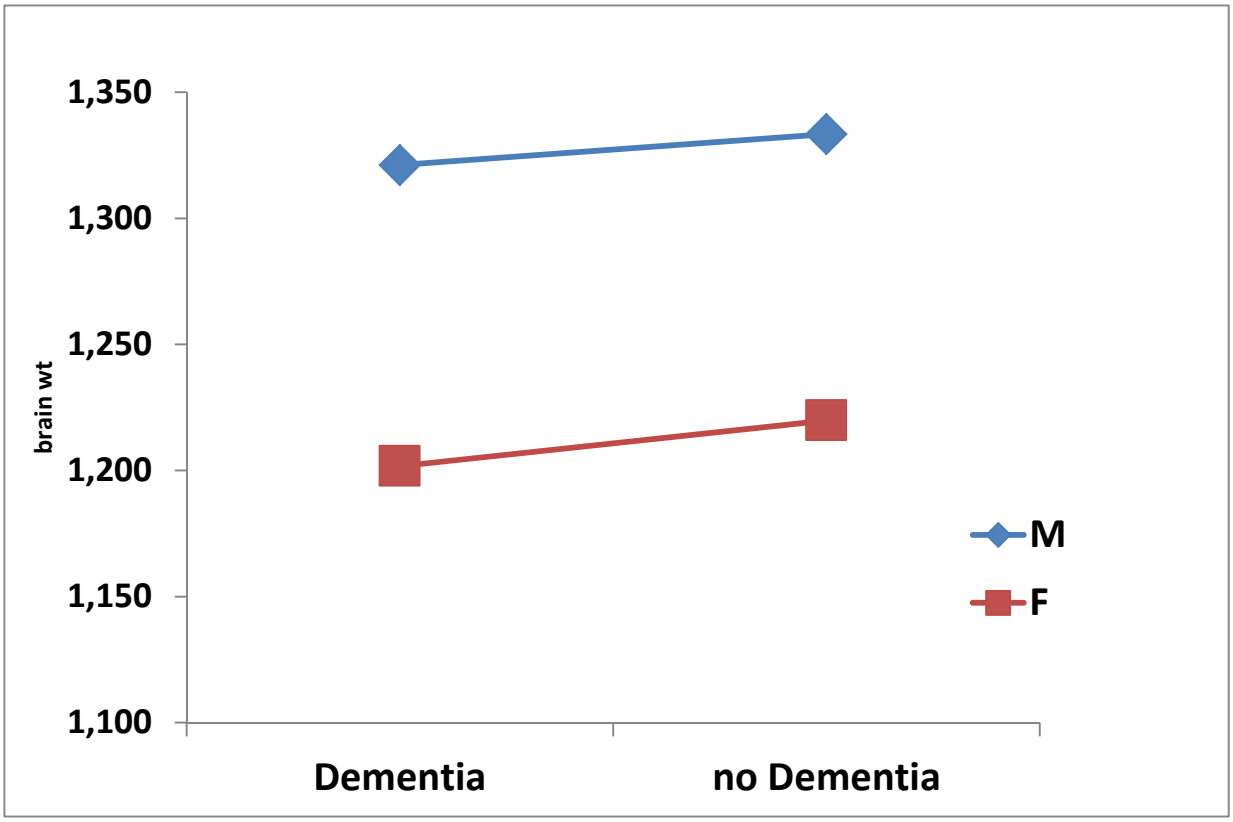
$$\text{Sum of squares=SS=} \sum(\text{mean}_i - \text{mean of means})^2$$

Mean square= MS = SS/(k-1) ,

df=k-1

Mean Brain weight by dementia and sex.

In the above, there is no significant dementia x sex interaction. This implies that the simultaneous effects of dementia and sex on brain weight are ADDITIVE. Graphically, there is parallelism.



ANOVA tables as a compact summary

In general, if factor A has "a" levels (and "a" means), there are "a" differences from the "grand" mean. However, since the sum of these mean differences must add to zero, only a-1 of them are free to vary. Thus we have "a-1" (not "a") degrees of freedom (df) for a factor with "a" levels.

$$SS_a = constant\ (b_1 + b_2 + b_3)^2 = \sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2$$

If factor A is NOT significant in the ANOVA table, we can conclude that all "a" means are about the same <u>without</u> looking at each one individually or making all the comparisons, a major simplification.

If factor A has "a" levels and factor B has "b" levels, there are a x b possible combinations (cells) of A and B and df= (a-1)(b-1).

ANOVA table – summarizes effects in three lines

Factor	df	Sum Squares (SS)	Mean square=SS/df
A	a-1	SS_a	$SS_a/(a-1)$
B	b-1	SS_b	$SS_b/(b-1)$
AB	(a-1)(b-1)	SS_{ab}	$SS_{ab}/(a-1)(b-1)$

More details of the ANOVA table

SS = sum of squares
MS = mean square = SS/df

F for factor "X" = MS_x/MS_{error}
where $MS_{error} = (SD_e)^2$ =pooled SD_e^2

If F gets larger, the corresponding p value gets smaller. If factor "X" has k levels, its F statistic and corresponding p value is testing the null hypothesis that the mean response for all k levels of X is the same.

Factor	df	SS	MS	F	p value
A	**a-1**	SS_a	MS_a	MS_a/MS_e	p_a
B	**b-1**	SS_b	MS_b	MS_b/MS_e	p_b
AB	**a-1(b-1)**	SS_{ab}	MS_{ab}	MS_{ab}/MS_e	p_{ab}

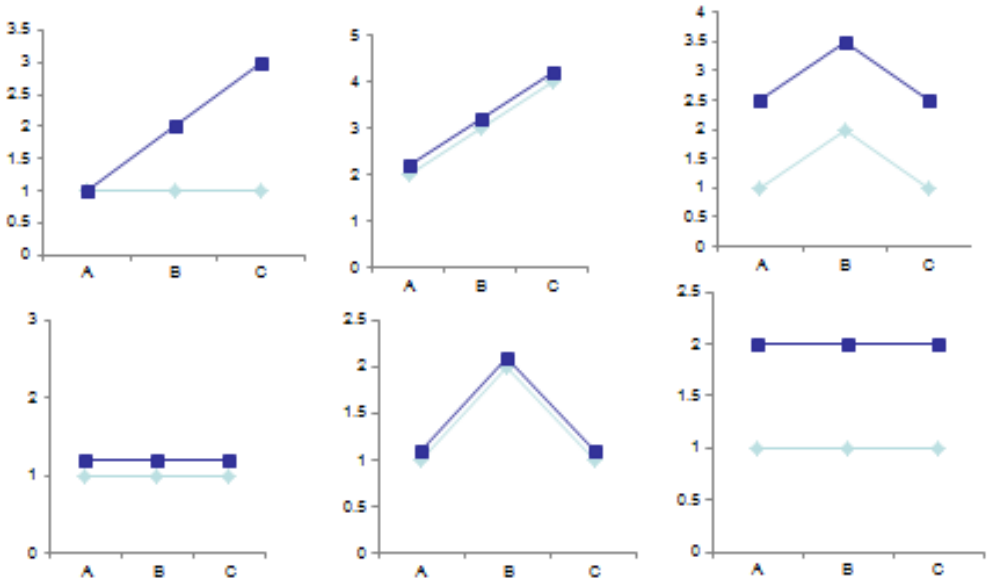
The "AB" factor is the "interaction" of A with B. This is a test that the differences among the levels of A are the same for any fixed B or, equivalently, that the differences among the levels of B are the same for any fixed A**. This is a test for "parallelism" or "additivity".**

(To be done in class)

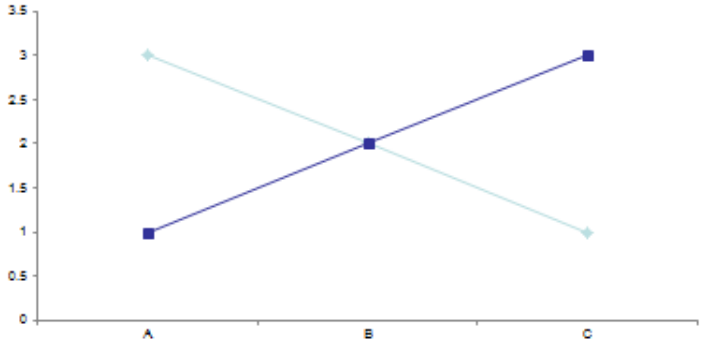
ANOVA intuition

Y may depend on group (A,B,C), sex & their interaction.

Which is significant in each example?



ANOVA intuition (cont)



Why is the ANOVA table useful?

When there are few variables and interactions and the variables have few levels, the usual regression analysis is more than adequate and an ANOVA presentation is not useful. But when there are many variables and interactions and/or the variables have many discrete levels, then the ANOVA table is a compact "congealed" summary of a regression model.

As a conceptual example, imagine we are investigating the effects of gender, race (B, W, H, A), education (no HS, HS, BA, MA, PhD) and occupation (laborer, office worker, manager, scientist, health worker) on Y=**depression** score. There are 2 x 4 x 5 x 5 = 200 cells for these four factors and all of their possible interactions. Rather than show a regression equation with 200 terms, we summarize the results into the table below which has 15 rows, not 200.The 15 rows are **orthogonal** only in the balanced case.

Dependent Variable: depression score

Sum of all SS in the table

Source	DF	SS	Mean Square	F Value	p value
Model	199	**3387.414863**	17.022185	4.42	<.0001
Error	400	1540.177404	3.850444 $=SD_e^2$		
Corrected Total	599	4927.592267			

R-Square	Coeff Var	Root MSE	y Mean
0.687438	21.13799	1.962255 $=SD_e$	9.283069

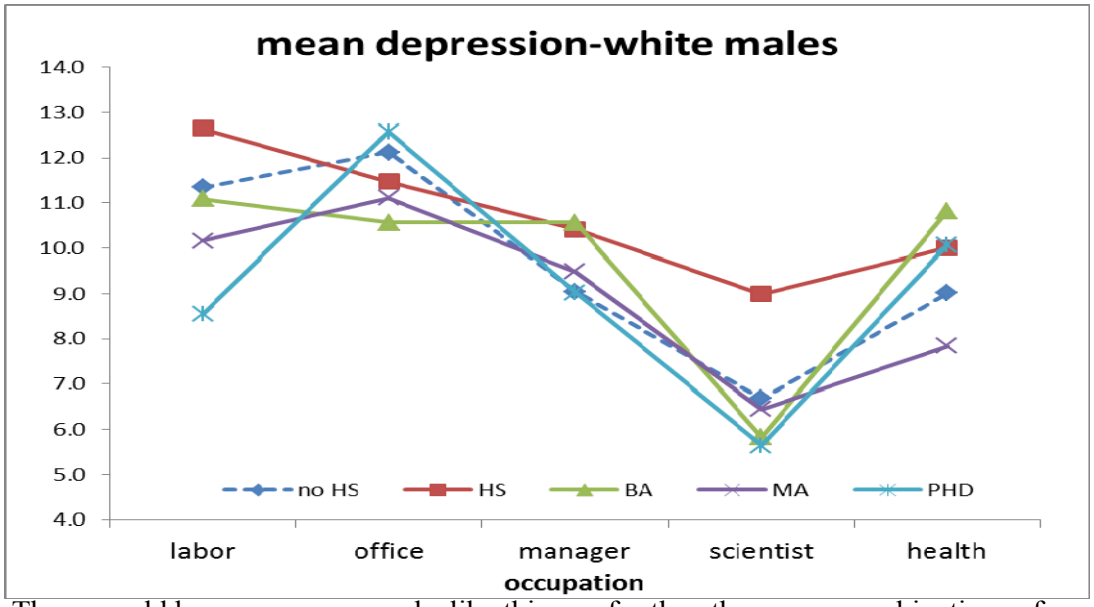
Source	DF	SS	Mean Square	F Value	p value
gender	1	778.084	778.084	202.08	<.0001
race	3	229.689	76.563	19.88	<.0001
educ	4	104.838	26.209	6.81	<.0001
occ	4	1531.371	382.843	99.43	<.0001
gender*race	3	1.879	0.626	0.16	0.9215
gender*educ	4	3.575	0.894	0.23	0.9203
gender*occ	4	8.907	2.227	0.58	0.6785
race*educ	12	69.064	5.755	1.49	0.1230
race*occ	12	62.825	5.235	1.36	0.1826
educ*occ	16	60.568	3.786	0.98	0.4743
gender*race*educ	12	77.742	6.479	1.68	0.0682
gender*race*occ	12	59.705	4.975	1.29	0.2202
gender*educ*occ	16	100.920	6.308	1.64	0.0565
race*educ*occ	48	206.880	4.310	1.12	0.2792
gender*race*educ*occ	48	91.368	1.903	0.49	0.9982

When removing non significant factors from the model, one must abide by the hierarchically well formulated (HWF) rule. An factor cannot be removed from the model unless all the higher order interactions containing this factor are not significant. For example, gender cannot be removed if any interaction involving gender is significant.

Eight graphs of 200 depresssion means. Y=depr,

X=occupation, X=strata by educ, separate graph for each gender/race

Males	Females
W	W
B	B
H	H
A	A



There would be seven more graphs like this one for the other seven combinations of gender and race.

Final ANOVA results for depression score

Source	DF	Sum of Squares	Mean Square	F Value	overall p
Model	12	2643.981859	220.331822	56.64	<.0001
Error	587	2283.610408	3.890307		
Corrected Total	599	4927.592267			

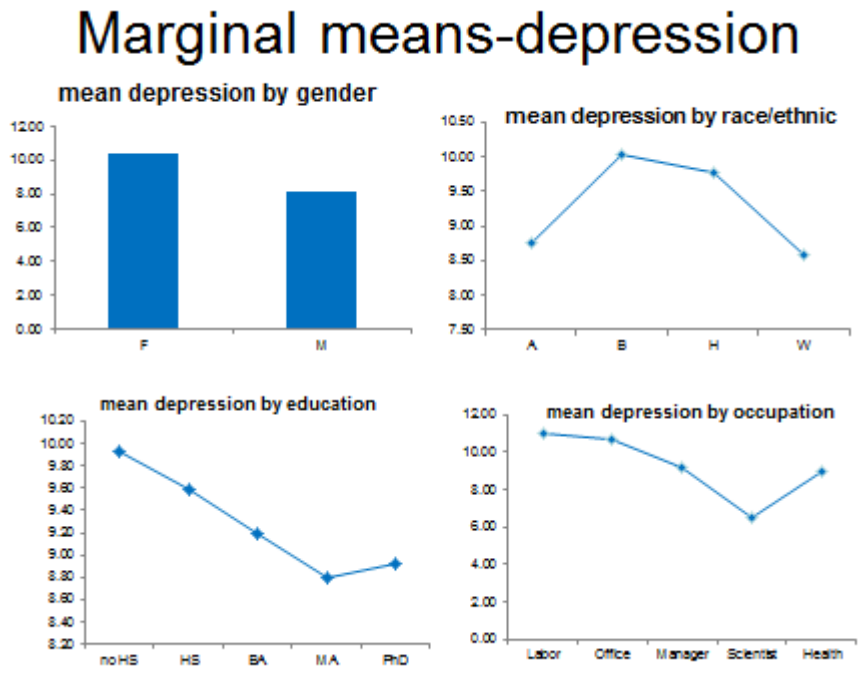
R-Square	Coeff Var	Root MSE	y Mean
0.536567	21.24713	1.972386	9.283069

Source	DF	SS	Mean Square	F Value	p value
gender	1	778.084257	778.084257	200.01	<.0001
race	3	229.688698	76.562899	19.68	<.0001
educ	4	104.837607	26.209402	6.74	<.0001
occ	4	1531.371296	382.842824	98.41	<.0001

The "sum of squares" table is a condensed table that is useful for screening, particularly screening interactions. It allows one to test "chunks" of the model.

If we also have **balance**, then all the rows of the ANOVA table are orthogonal (not correlated) so the assessment of one factor or interaction is not affected if another factor or interaction is significant or not. This is an ideal analysis situation.

In the example above, none of the interactions are significant. Thus we learn that gender, race, education and occupation all have a significant and simultaneous effect on depression score. However, since none of the interactions in this example are significant, the effect of each of the four factors is ADDITIVE. Therefore, it is ok to only report the marginal means. The effect of each factor is the same for all levels of the other factors.



Balanced versus unbalanced ANOVA

below "n_c" denotes the sample size in each cell

Cell and marginal mean amygdala volumes in cc
Not balanced

	Male	**Female**	adj marg mean	Obs marg mean
Dementia	0.5 (n_c=10)	0.5 (n_c=90)	0.5	0.5 (n=100)
No Dementia	1.5 (n_c=190)	1.5 (n_c=10)	1.5	1.5 (n=200)
adjusted marg. Means	1.0	1.0		
Observed marg. means	1.45 (n=200)	0.6 (n=100)		n=300

If we compare the averages of males versus females ignoring dementia, we apparently see that the marginal mean of 1.45 cc in the 200 males is larger than the 0.60 cc mean value in the 100 females, an apparent "sex" effect. However, if we **control** for dementia and look at **the marginal means we would have obtained if the design had been balanced**, then the marginal mean for males is 1.0 and the marginal means for female is 1.0, imply no effect of sex at all.

The marginal means for one factor ignoring the other factors is denoted the **observed** marginal means. The marginal means for a given factor computed under a model controlling for other factors as if the model was balanced is denoted the **adjusted** marginal means.

So, in this example, even though there is no sex difference in those with dementia and no sex difference in those without dementia, and therefore no significant interaction (actually, the interaction here is exactly zero), looking at the "wrong" analysis by not controlling for dementia can mistakenly lead one to conclude that there is a sex difference (effect of sex on the outcome) when there is not.

Only the results corresponding to the **simultaneous** assessment of gender and dementia correctly shows that sex is not significant after controlling for dementia. That is, the simultaneous analysis generates means and p values according to the model

$$Y = \beta_0 + \beta_1 \text{ sex} + \beta_2 \text{ dementia} + \beta_3 \text{ sex*dementia}$$

When there is balance (equal sample size in all cells), then there will be **orthogonality** and looking at sex ignoring dementia or dementia ignoring sex will not be misleading. However, in the unbalanced case, only the simultaneous analysis gives estimates that does not depend on sample size. Sample size should only affect confidence bounds and power, not the size of the mean differences.

The effect of gender is not the same ignoring dementia versus controlling for dementia.

Repeated Measures Analysis of Variance

When the same patient is measured more than once, this is called a repeated mesure design and the analysis of variance method needed is called the repeated measure analysis of variance model. This is a generalization of the paired comparison (paired t test).

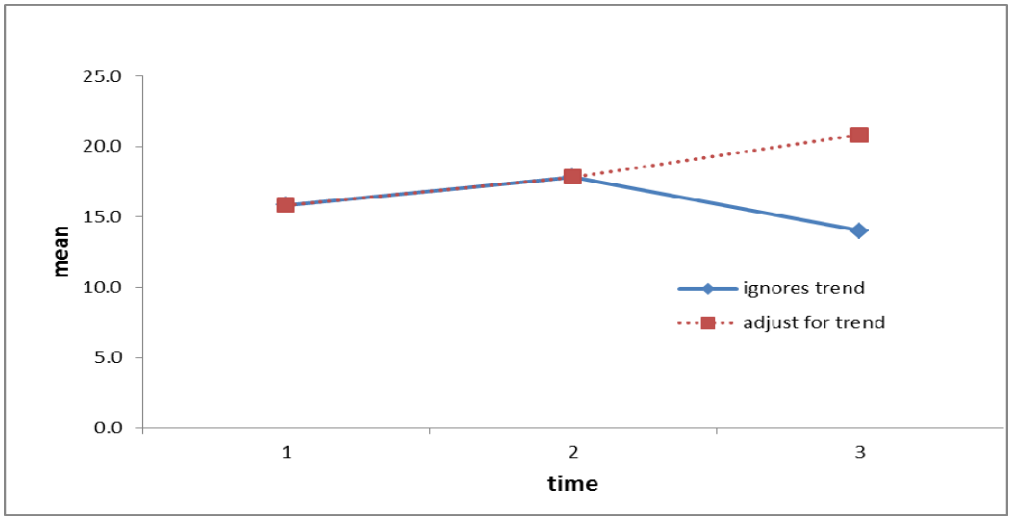
In the artificial example below, the correlation between measurements at time 1 and time 2 or time 2 and time 3 is r=1.0. Every patient increases 2 units from time 1 to time 2 and increases 3 units from time 2 to time 3. The value for patient F is missing at time 3.

patient	time 1	time 2	time 3
A	5	7	10
B	8	10	13
C	9	11	14
D	12	14	17
E	11	13	16
F	50	52	
unadjusted means	15.8	17.8	14.0
adjusted means	15.8	17.8	20.8

If one computes means only using the observed data, the mean at time 3 is 14.0, lower than the means at time 1 and time 2. But this is misleading since the values are increasing in every patient!

The repeated measure model, in contrast, uses the correlation and change to estimate what the mean would have been at time 3 if the data for patient F had been observed. Under the repeated measure model, the estimated mean is 20.8, not 14. The 20.8 is 3 points higher than 17.8 at time 2, consistent with every patient increasing 3 points from time 2 to time 3.

Mean profiles



The adjusted means are based on the repeated measure ANOVA model.

The factorial vs repeated measure models give different standard errors and p values for the means and mean differences, particularly the mean differences.

	Factorial		Repeated measure	
time	**Mean**	**SEM**	**mean**	**SEM**
1	15.83	5.8672483	15.83	4.1272113
2	17.83	5.8672483	17.83	4.1272113
3	14.00	6.4272485	20.83	4.1272216

		Factorial			Repeated measure		
time	**vs time**	**Mean Difference**	**Std Error**	**p value**	**Mean Difference**	**Std Error**	**p value**
1	2	2.00	8.297542	0.8130	2.00	0.0238283	<.0001*
1	3	1.83	8.702536	0.8362	5.00	0.0255530	<.0001*
2	3	3.83	8.702536	0.6663	3.00	0.0255530	<.0001*

The standard errors for the mean differences are MUCH larger under the factorial model since this model is assuming there is a different group of subjects at each time, not the same subjects measured 3 times.

So, SEs for mean differences are much larger if the repeated measure method is ignored.