Analysis of Variance
ANOVA

Overview

- We've used the *t* -test to compare the means from two independent groups.
- Now we've come to the final topic of the course: how to compare means from *more than two* populations.
- When we're comparing the means from *two* independent samples we usually asked: "Is one mean different than the other?"

Overview (cont)

- Fundamentally, we're going to proceed in this section exactly as we did in "Hypothesis testing on two means," up to a point.
- It's more complicated when we have more than two groups. So, we'll need to address those issues.

Overview (cont)

- First, we'll need to have a clear understanding of the data and what we're testing.
- Both the *t*- test situation and the correlation/ regression situation will help us understand the analysis of variance (ANOVA).
- The theory behind ANOVA is more complex than the two means situation, and so before we go through the step-by-step approach of doing ANOVA, let's get an intuitive feel for what's happening.
- How does ANOVA compare means?

Model Comparisons

- First we'll review what we learned when considering the correlation/regression question.
- We fit a straight line through the data in a scatterplot (when it was appropriate).
- There's an intuitive link between this situation and what we do when we compare means.
- This leads us to comparing models.

Models

- Recall that in the correlation/ regression situation there were three ways to phrase the question:

1. Non-zero correlation?
2. Non-zero slope?
3. Is the straight-line model better?

Models

- In the case of simple regression, the last form of this question: Is the straight-line model better? is it more complicated that we really need?
- However, this form of the simple regression question helps us understand the analysis of variance.

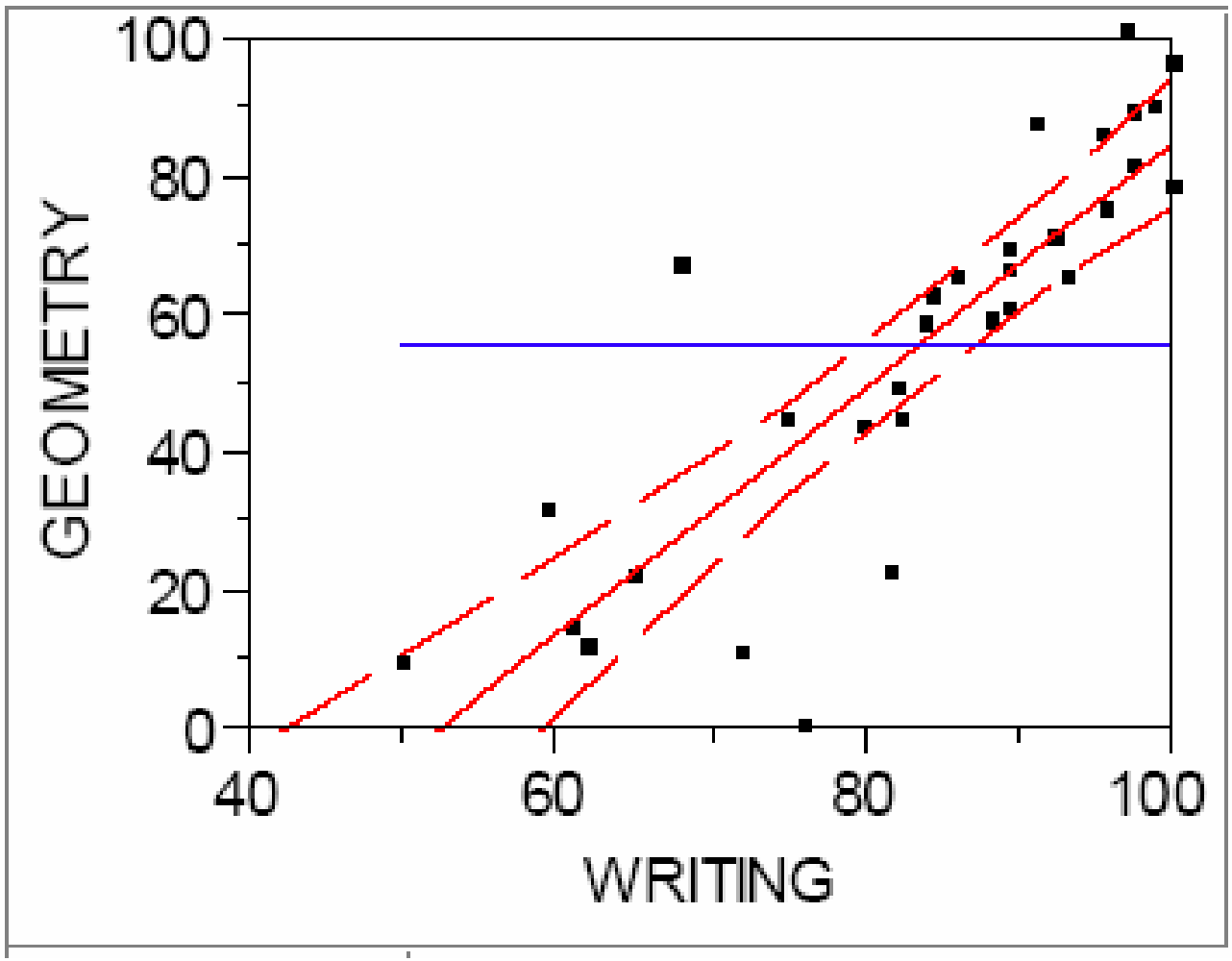
Fitting a straight line

- In a previous section, we talked about comparing two models of the data.

  - H_0: $Y = \bar{Y}$ + error
  - H_A: Y = intercept + slope • X + error

Models

- What we were doing was looking at the average Y response and wondering:
  - Is the average Y constant across all values of X? or
  - Does the average Y change with X?
- The way we visualized this comparison was to look at the confidence band around our straight-line model and compare it to the horizontal line drawn at the mean of Y


100
80
60
40
20
0
GEOMETRY
40
60
80
100
WRITING

Comparing group means

- When we're comparing the average *Y* response in different *groups*, we're asking a similar question.
- Say we have four groups (*X* = a, b, c, or d) then the two models would be:

  - H_0: $Y = \bar{Y}$ + error, or.
  - H_A: $Y = \bar{Y}_a$ + error, (if X = a),

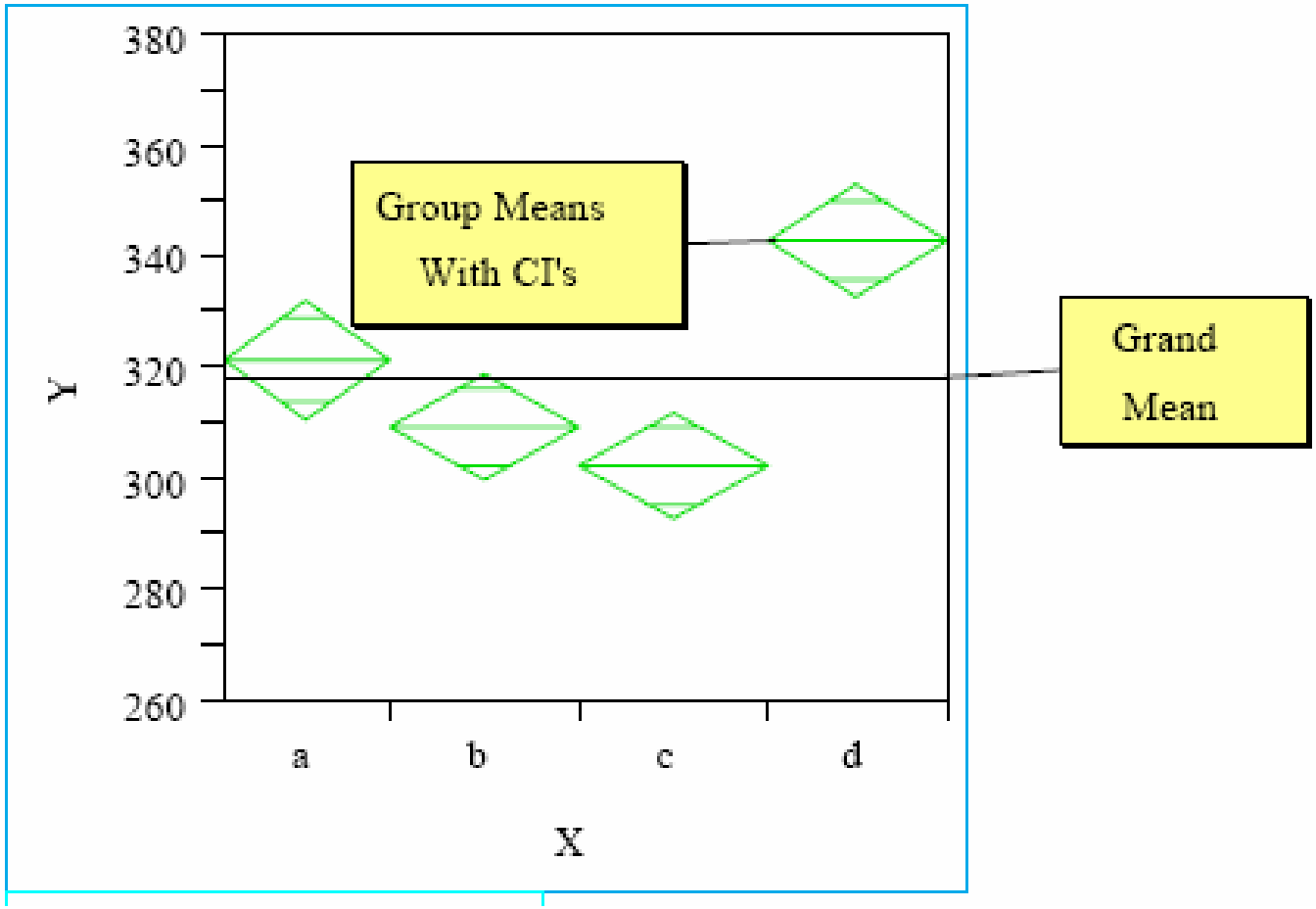
    $Y = \bar{Y}_b$ + error, (if X = b),

    $Y = \bar{Y}_c$ + error, (if X = c),

    $Y = \bar{Y}_d$ + error, (if X = d).

Comparing group means

- What we are doing is looking at the average Y response and wondering:
  - Is the average Y constant across all values of X? or
  - Does the average Y change with X?
- The way we visualize this comparison is to look at the confidence bounds around each mean and compare it to the horizontal line drawn at the grand mean of Y


380
360
340
320
300
280
260
Y
Group Means
With CI's
Grand
Mean
a
b
c
d
X

Example

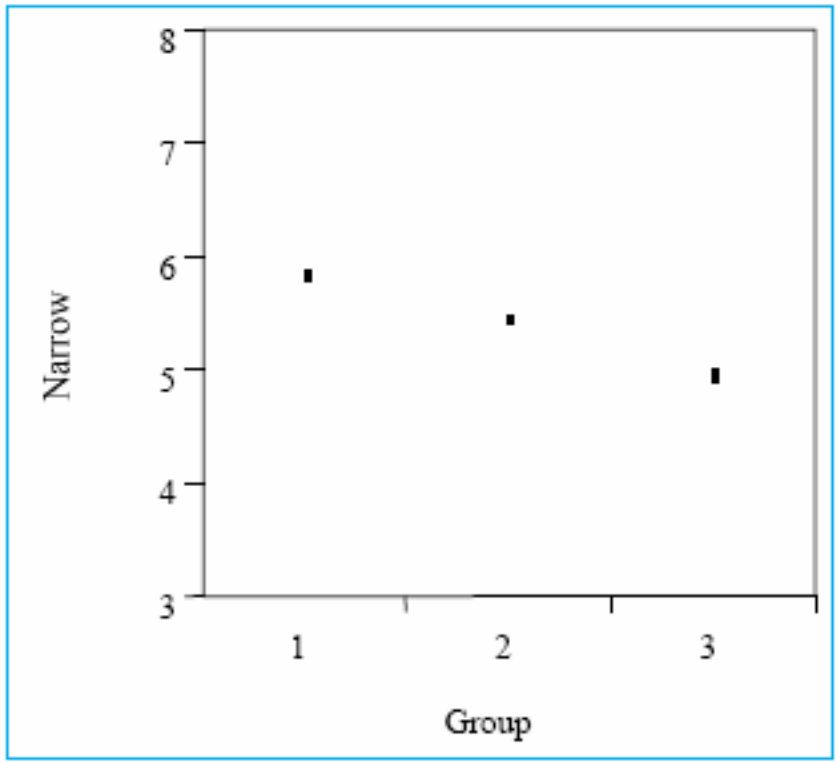
- Consider two different experiments designed to compare three different treatment groups.
- In each experiment, five subjects are randomly allocated to one of three groups.
- After the experimental protocol is completed, their response is measured.
- The two experiments use different measuring devices, Narrow and Wide

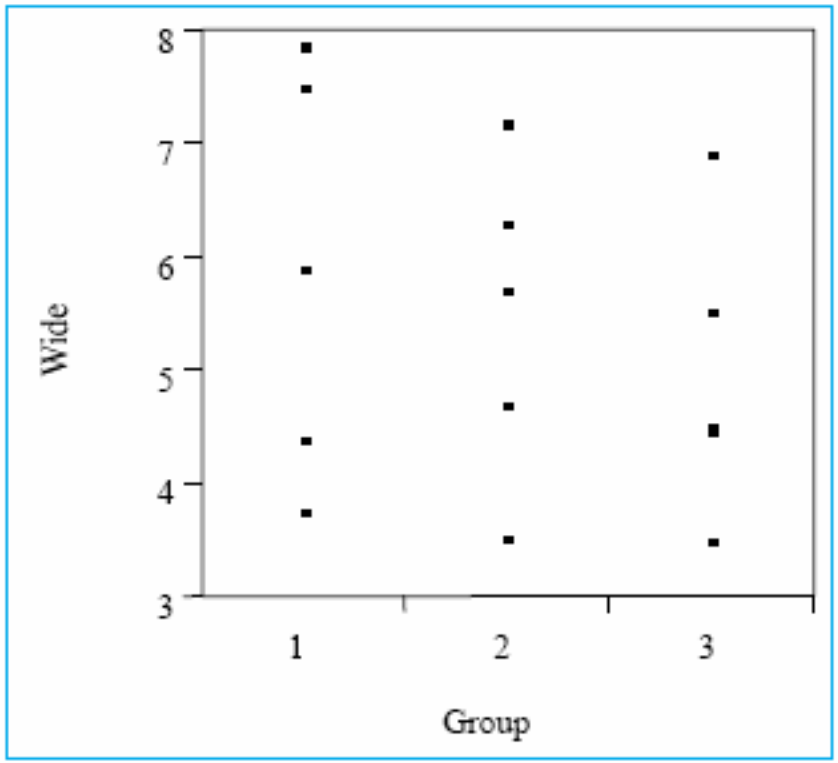
Response = Narrow

Group		
1	2	3
5.90	5.51	5.01
5.92	5.50	5.00
5.91	5.50	4.99
5.89	5.49	4.98
5.88	5.50	5.02

Response = Wide

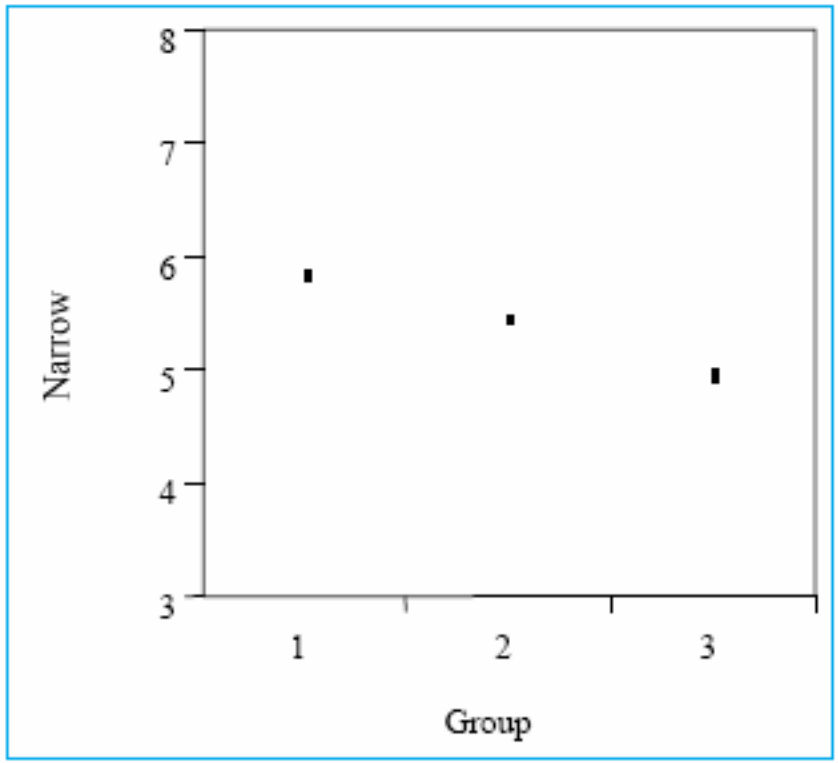
Group		
1	2	3
5.90	6.31	4.52
4.42	3.54	6.93
7.51	4.73	4.48
7.89	7.20	5.55
3.78	5.72	3.52

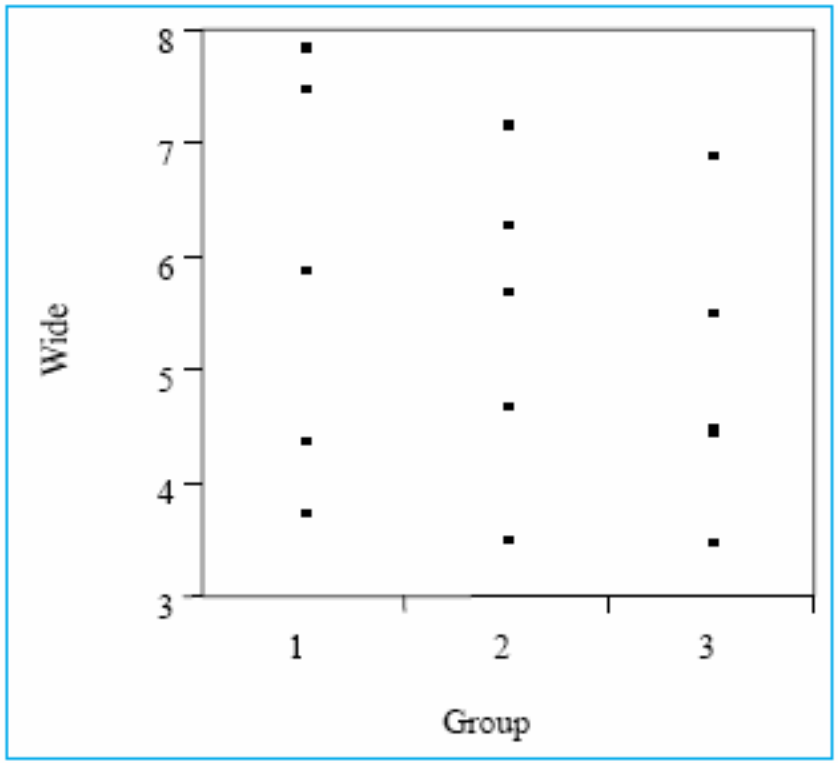

Narrow
8
7
6
5
4
3
1
2
3
Group


Wide
8
7
6
5
4
3
1
2
3
Group

Comparing Means

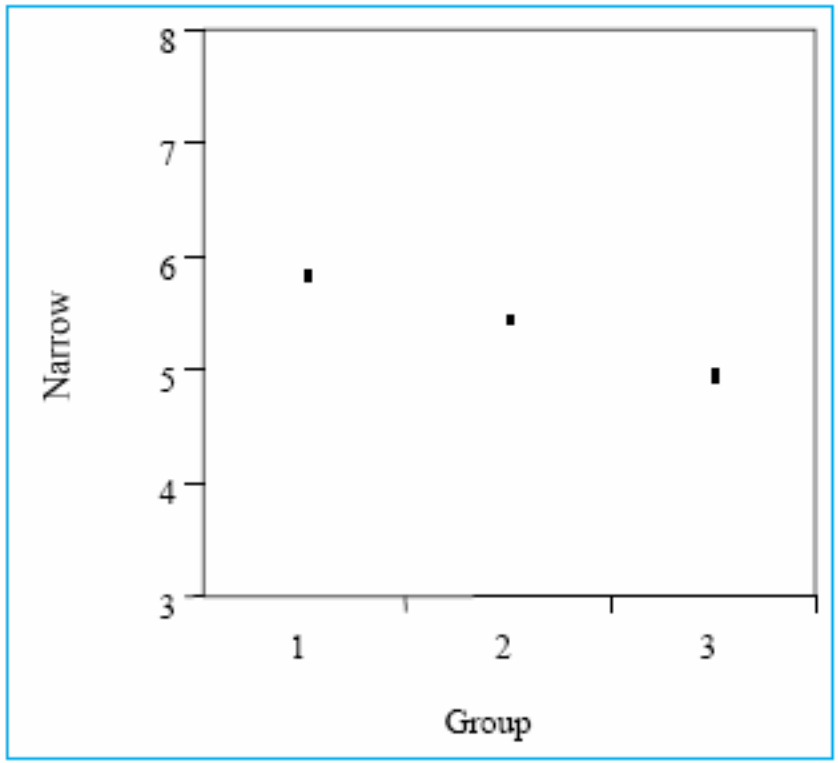
- In *both* experiments the mean response for Group 1 is 5.9, for Group 2 is 5.5, and for Group 3 is 5.0.
- In the first experiment—the one on the left measured by the variable Narrow— are the three groups different?
- In the second experiment—the one on the right measured by the variable Wide—are the three groups different?

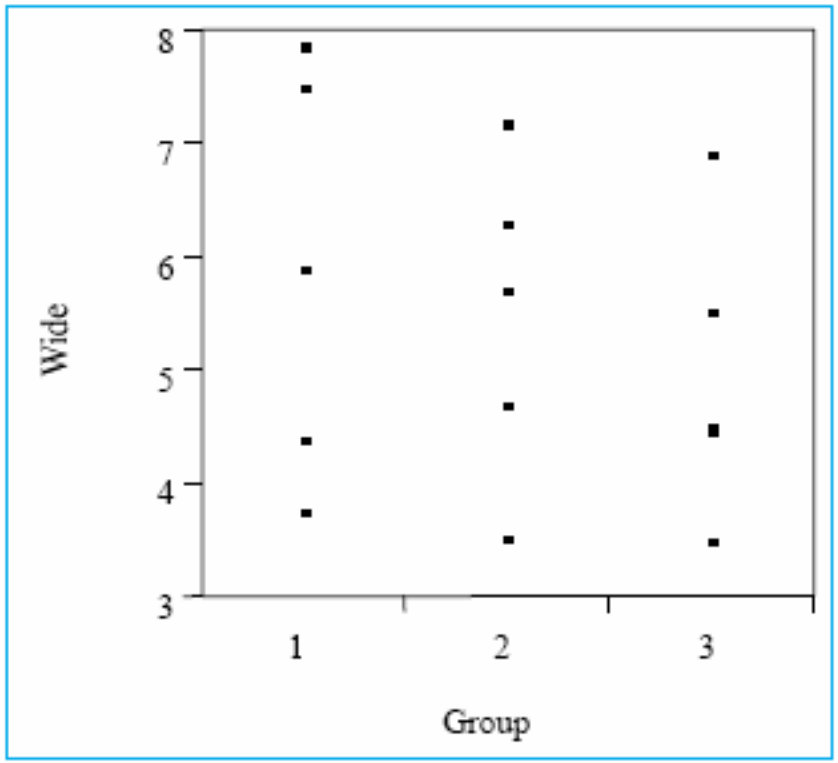

Narrow
8
7
6
5
4
3
1
2
3
Group


Wide
8
7
6
5
4
3
1
2
3
Group

The interoccular-trauma test

- Again, your eye goes to the ink: the three groups seem to have a different mean in the first experiment.
- Whereas in the second experiment, the best explanation would probably be that any apparent differences in the means could have come about through chance variation.


Narrow
8
7
6
5
4
3
1
2
3
Group


Wide
8
7
6
5
4
3
1
2
3
Group

- In fact, what your eye is doing is comparing the differences *between* the means in each group to the differences *within* each group.
- That is, in the Narrow experiment a summary of the results would include a table of means and SD’s

Narrow Group

Group	n	Mean	SD	SE	95%CI	
1	5	5.9	0.016	0.0071	5.887	5.913
2	5	5.5	0.007	0.0032	5.487	5.513
3	5	5.0	0.016	0.0071	4.987	5.013
Average		5.467	0.013	0.0058		

Testing differences

- So, a difference as large as 5 vs. 5.9 is considered in the context of very small SD's (on the order of 0.013).
- The 0.9 unit difference is compared to standard errors of approximately 0.0058.

t-test

- Consider comparing the two groups with a *t*-test, we'd calculate something like:

$$t = \frac{difference}{SE_{\text{difference}}} = \frac{5.9 - 5.0}{2 \times 0.0058} \approx 78$$

Wide Group

- But, in the Wide experiment a summary of the results would include a table of means and SD's as below

Group	n	Mean	SD	SE	95%CI	
1	5	5.9	1.819	0.8135	4.412	7.388
2	5	5.5	1.417	0.6336	4.012	6.988
3	5	5.0	1.296	0.5796	3.512	6.488
Average		5.467	1.511	0.6756		

Comparing the means

- So, a difference as large as 5 vs. 5.9 is considered in the context of large SD's (on the order of 1.5).
- The 0.9 unit difference is compared to standard errors of approximately 0.68.

t-test

$$t = \frac{difference}{SE_{\text{difference}}} = \frac{5.9 - 5.0}{2 \times 0.6756} \approx 0.66$$

Understanding the F-test

- What ANOVA does is compare the differences *between* the means in each group to the differences *within* each group's observations.
- This is an extremely important concept because it's key to your understanding of the statistical test we use.

F-test

$$F = \frac{SSR / df_{\text{model}}}{SSE / df_{\text{error}}}$$

F-test for ANOVA

- It's a ratio of the mean square of the model (which compares the straight line predicted-values to the mean predicted-values) to the mean square error (which compares the straight-line predicted-values to the observed values).
- This is exactly what we do when we compare means.
- We use an F test that compares the differences *between* the means in each group to the differences *within* each group.

F-test for ANOVA

$$F = \frac{SS_{\text{model}} / df_{\text{model}}}{SSE_{\text{error}} / df_{\text{error}}}$$

SS model

- First, consider the differences between the means. The SSmodel is:

	value	deviation	squared	n	n*squared
	5.9	0.4333	0.1878	5	0.939
	5.5	0.0333	0.0011	5	0.006
	5.0	-0.4667	0.2178	5	1.089
Average=	5.4667			Sum=	2.0333

SS Model

- So the differences between the means in each group is just the sum of the weighted squared deviations (SSmodel = 2.0333) divided by the number of groups minus one (*df* = groups – 1).
- So the numerator for the F test is SSmodel/dfmodel = 2.033/2 = 1.0667 = MSmodel.

SS error

- In the Narrow experiment, we consider the differences within each group.
- We take each observation and compare it to the group mean:

Values	Group Mean	deviation	squared
5.90	5.90	0.00	0.0000
5.92	5.90	0.02	0.0004
5.91	5.90	0.01	0.0001
5.89	5.90	-0.01	0.0001
5.88	5.90	-0.02	0.0004
5.51	5.50	0.01	0.0001
5.50	5.50	0.00	0.0000
5.50	5.50	0.00	0.0000
5.49	5.50	-0.01	0.0001
5.50	5.50	0.00	0.0000
5.01	5.00	0.01	0.0001
5.00	5.00	0.00	0.0000
4.99	5.00	-0.01	0.0001
4.98	5.00	-0.02	0.0004
5.02	5.00	0.02	0.0004
		Sum=	0.0022

SS Error

- So the differences within all groups is just the sum of the squared deviations (SSerror = 0.0022) divided by the number of values minus the number of groups (*df* = *n* – groups = 15 – 3 = 12).
- So the denominator for the F test is SSerror/dferror = 0.0022/12 = 0.00018 = MSerror.
- Again what we're doing is comparing the differences *between* the means in each group to the differences *within* each group.

F-test Result for Narrow

$$F = \frac{SS_{\text{model}} / df_{\text{model}}}{SSE_{\text{error}} / df_{\text{error}}} \text{ or } F = \frac{2.033 / 2}{.0022 / 12} = \frac{1.067}{.0018} \approx 5545$$

SS error in the Wide experiment

Values	Group Mean	deviation	squared
5.90	5.90	0.00	0.0000
4.42	5.90	-1.48	2.1904
7.51	5.90	1.61	2.5921
7.89	5.90	1.99	3.9601
3.78	5.90	-2.12	4.4944
6.31	5.50	0.81	0.6561
3.54	5.50	-1.96	3.8416
4.73	5.50	-0.77	0.5929
7.20	5.50	1.70	2.8900
5.72	5.50	0.22	0.0484
4.52	5.00	-0.48	0.2304
6.93	5.00	1.93	3.7249
4.48	5.00	-0.52	0.2704
5.55	5.00	0.55	0.3025
3.52	5.00	-1.48	2.1904
		Sum=	27.9846

F-test Result in Wide

$$F = \frac{2.033 / 2}{27.98 / 12} = \frac{1.067}{2.332} \approx 0.44$$

Understanding ANOVA

- ANOVA is applicable when the response variable is continuous and we have more than two groups to compare.
- Our two intuitive understanding of the analysis of variance are as follows:
  1. What ANOVA does is compares two models: One overall grand mean, vs. Different means for each group.
  2. It does this by comparing the differences *between* the means in each group to the differences of the individual values *within* each group.

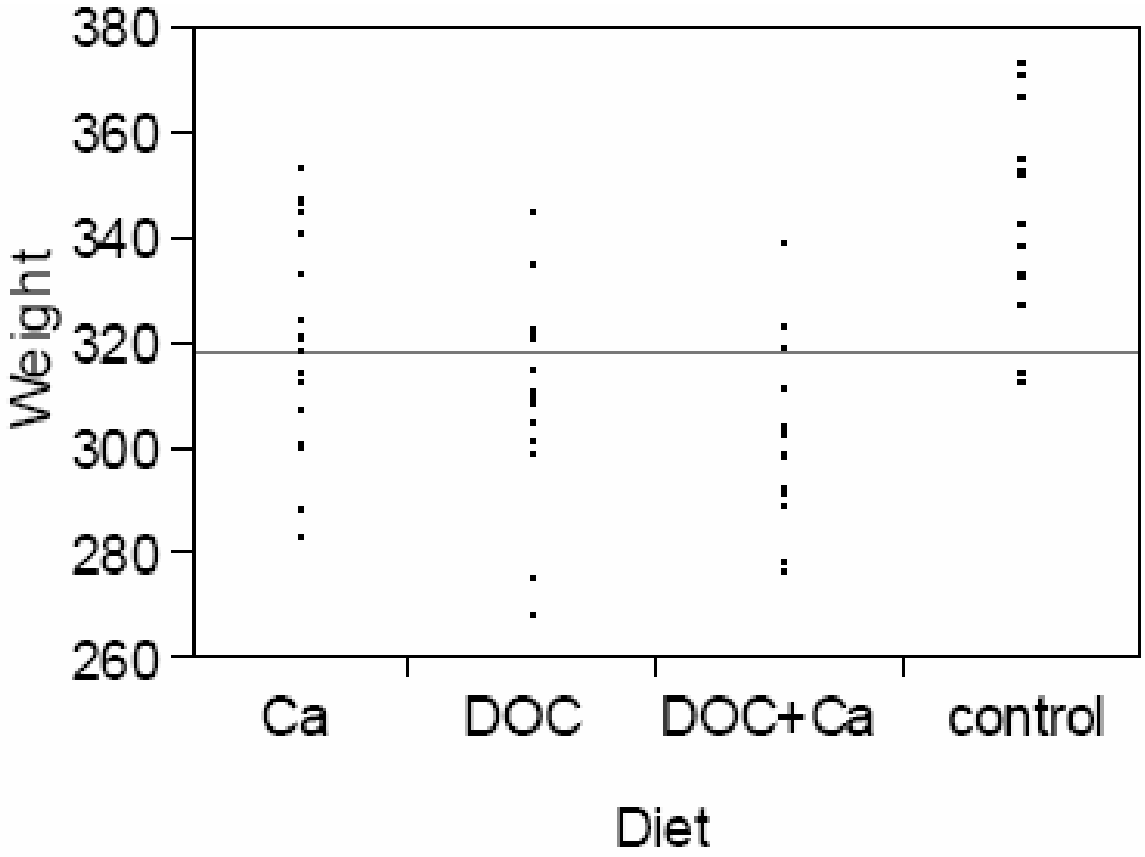
ANOVA: More than Two Sample Means

- **Calcium and Weight:** Researchers (randomly?) divided 7-week old rats into four groups on different diets:
  1. untreated controls,
  2. high calcium diet (Ca),
  3. deoxycortiosterone-NaCl treated rats (DOC), and
  4. rats receiving both dietary supplements (DOC+Ca).
- The question is: do the four conditions have differing effects on the mean weight of Wistar-Kyoto rats?

Phase 1: State the Question

- **1. Evaluate and describe the data**
- Recall our first two questions:
  1. Where did this data come from?
  2. What are the observed statistics?
- The first step in any data analysis is evaluating and describing the data.
- **Preliminary Analysis:** What are the observed statistics?

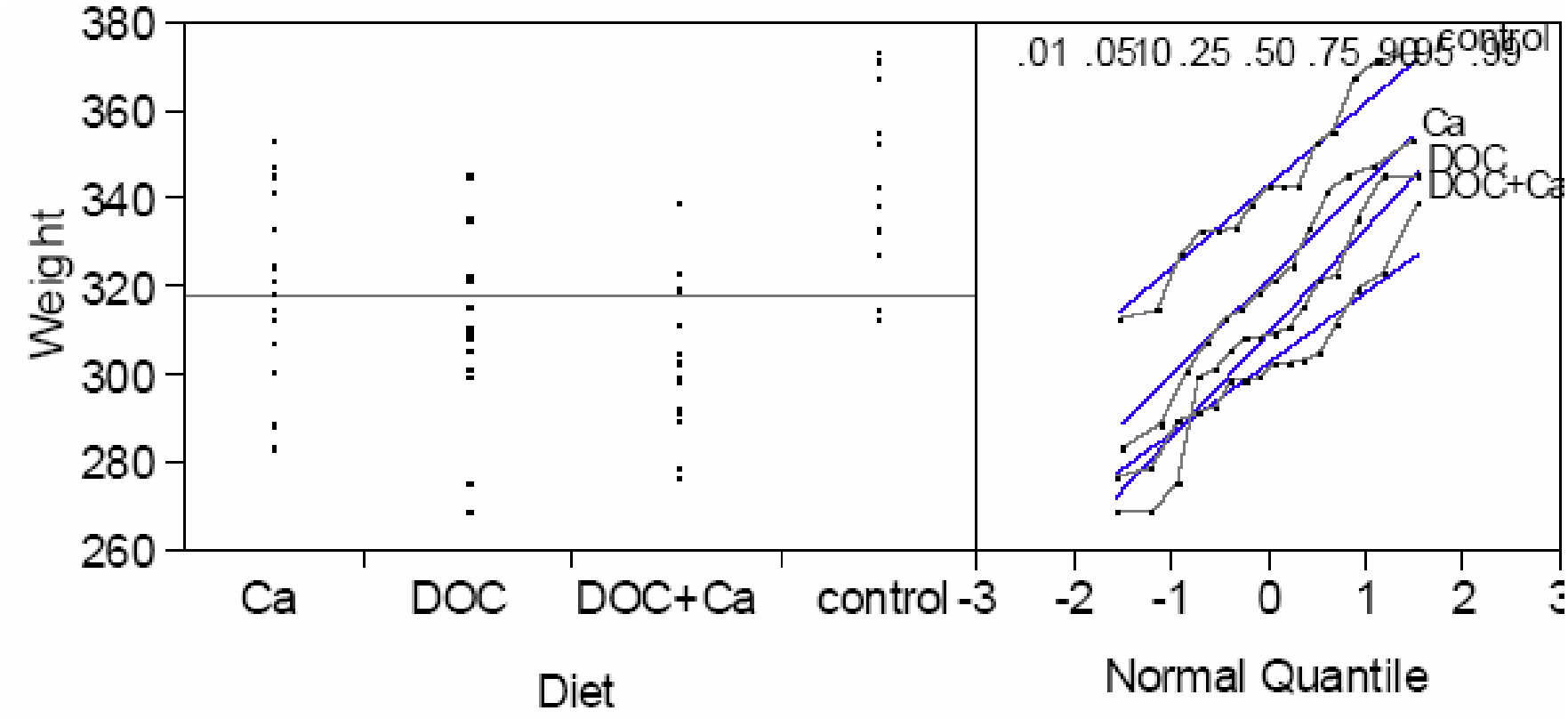
Fit Y by X



Oneway Analysis of Weight By Diet

- The dot plots show some differences between groups and spread within each group.
- Think back to what we'd do next if we were comparing *two* means.
- We know we should concern ourselves with two assumptions:
  - equal variance within each group, and
  - normality.
- In the two-group situation, how did we assess these assumptions?

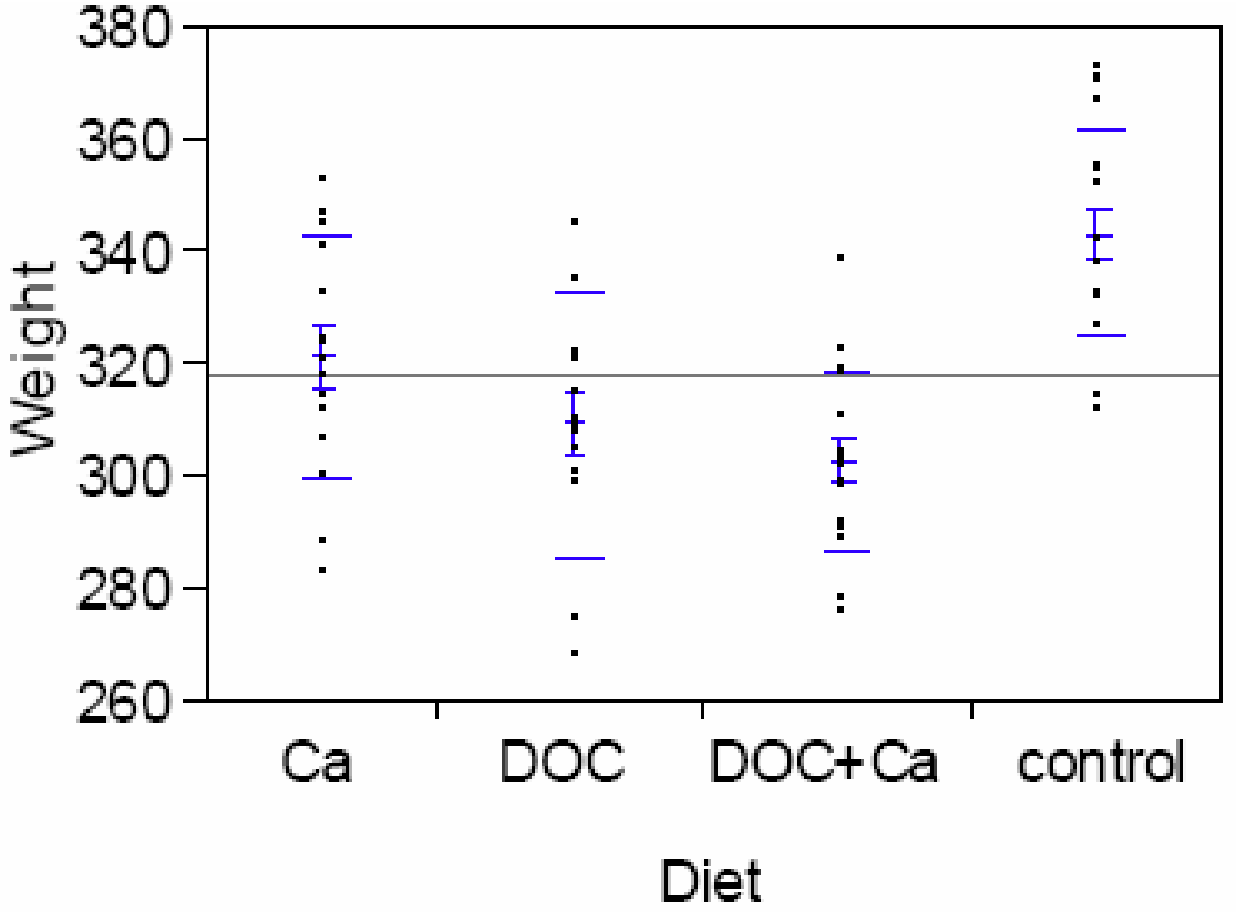
Normal Quantile Plots



Preliminary analysis, showing means

- If the data is normally distributed then means and SDs make sense. (If these distributional assumptions are unwarranted, then we should consider nonparametric methods.)
- The next thing to do in our preliminary analysis is to show the means and standard deviations calculated within each group.

Means and SDs



Table

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Ca	14	321.429	21.7210	5.8052	308.89	333.97
DOC	16	309.375	23.5397	5.8849	296.83	321.92
DOC+Ca	16	302.500	16.0499	4.0125	293.95	311.05
control	15	343.133	18.7116	4.8313	332.77	353.50

Notes

- In the figure, the dashed lines above and below the means are one standard deviation away from their respective mean.
- In the table, note that the 95% CI on the means are shown and recall that these CIs do NOT assume equal variability.

Summing up Step 1

- What have we learned about the data?
- We have not found any errors in the data.
- We’re comfortable with the assumptions of normality and equal variance.
- We’ve obtained descriptive statistics for each of the group we’re comparing.
- It’s our guess that there is a significant difference.

2. Review assumptions

- As always there are three questions to consider.
  1. Is the process used in this study likely to yield data that is *representative* of each of the two populations?
  2. Is each animal in the samples *independent* of the others?
  3. Is the *sample size within each group sufficient*?

Assumptions

- **Bottom line:** We have to be comfortable that the first two assumptions are met before we can proceed at all.
- If we're comfortable with the normality assumption, then we proceed, as below.
- In a following section, we'll discuss what to do when normality can not be safely assumed.

3. State the question—in the form of hypotheses

- Presuming that we're OK with normality, here are the hypotheses (using group names as subscripts):
  - The null hypothesis is Ca = DOC = DOC+Ca = control (the means of *all* populations are the *same*),
  - The alternative hypothesis is that not all the means are equal (there is at least one difference).

Phase 2: Decide How to Answer the Question

- **4. Decide on a summary statistic that reflects the question**
  - We can not use a *t*-test because there are more than two means.
  - Note also, that it is completely inappropriate to use multiple *t*-tests!
  - We use the F-test discussed above to compare the differences between the means to the differences within each group.

F-test

- As in the case when two groups are compared, we need to concern ourselves with whether the variances are the same in each group.
- There are, as before, two possibilities.
- The two possibilities depend upon the standard deviations within each group.
  - Are they the same?
  - Or do the groups have different standard deviations?

Equal Variance

- If the SDs (or variances) within the populations are equal than the “average” standard error in the denominator of the *F*-test is appropriate.
- If there is equal variance, then we calculate the *p*-value using the distribution of F.
- The distribution is complex, but recall that there are two *df* needed:
  1. the *df*-numerator (which is the number of groups minus one), and
  2. the *df*-denominator (the number of subjects minus the number of groups).

Unequal Variance

- If the variances are not equal, the calculation is more complicated.
- However, like the multiple proportions example, JMP handles the calculation details.

Deciding on the correct test

- Which test should we use?
- We may not need to choose; if the all sample sizes are equal (termed a *balanced* design) the two methods give identical results.
- It's even pretty close if the *n*'s are slightly different.
- If one *n* is more than 1.5 times the other, you'll have to decide which *t*-test to use.

Choosing a Test

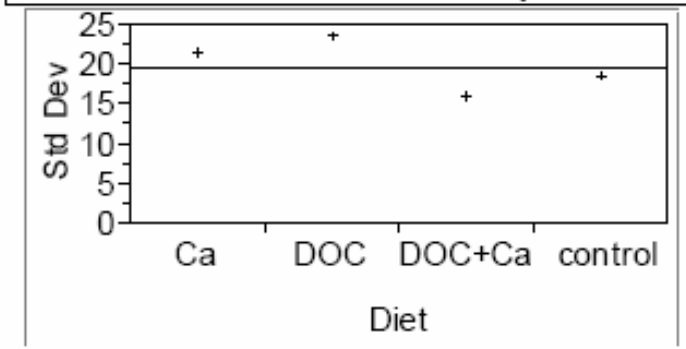
1. Decide whether the standard deviations are different.
2. Use the equal variance *F*-test if the SDs appear the same, or Use the unequal variance Welch ANOVA if the SDs appear different.

Options

- As before, look at the normal quantile plot. If the lines are in that “gray area” between clearly parallel and clearly not parallel, what do we do?
- Four possibilities come to mind:
  1. Ignore the problem and be risky: use the equal variance *F*-test.
  2. Ignore the problem and be conservative: use the unequal variance *F*-test.
  3. Make a formal test of unequal variability in the two groups.
  4. Compare the means using nonparametric methods.

Equal Variance Test

Tests that the Variances are Equal



Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
Ca	14	21.72101	17.28571	17.28571
DOC	16	23.53968	16.87500	16.87500
DOC+Ca	16	16.04992	11.37500	11.37500
control	15	18.71160	14.31111	14.26667

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	0.9420	3	57	0.4264
Brown-Forsythe	0.7029	3	57	0.5542
Levene	0.7049	3	57	0.5530
Bartlett	0.7929	3	.	0.4976

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
14.2137	3	31.004	<.0001

Which Test?

- If the Prob>F value for the Brown-Forsythe test is < 0.05, then you have unequal variances.
- This report also shows the result for the *F*-test to compare the means, allowing the standard deviations to be unequal.
- This is the “Welch ANOVA” Here is a written summarization of the results using this method:

  “The four groups were compared using an unequal variance *F*-test and found to be significantly different ($F(3, 31) = 14.2$, *p*-value < .0001). The means were found to be different ….”

Steps 5 & 6

- **5. How could random variation affect that statistic?**
  - Recall the rough interpretation that *F*'s larger than 4 are remarkable.
- **6. State a decision rule, using the statistic, to answer the question**
  - The universal decision rule: Reject Ho: if *p*-value < α.

Phase 3: Answer the Question

- **7. Calculate the statistic**
- There are three possible statistics that may be appropriate:
  1. an equal variance F-test,
  2. an unequal variance F-test, or
  3. the nonparametric Wilcoxon rank-sums test.

Next Time

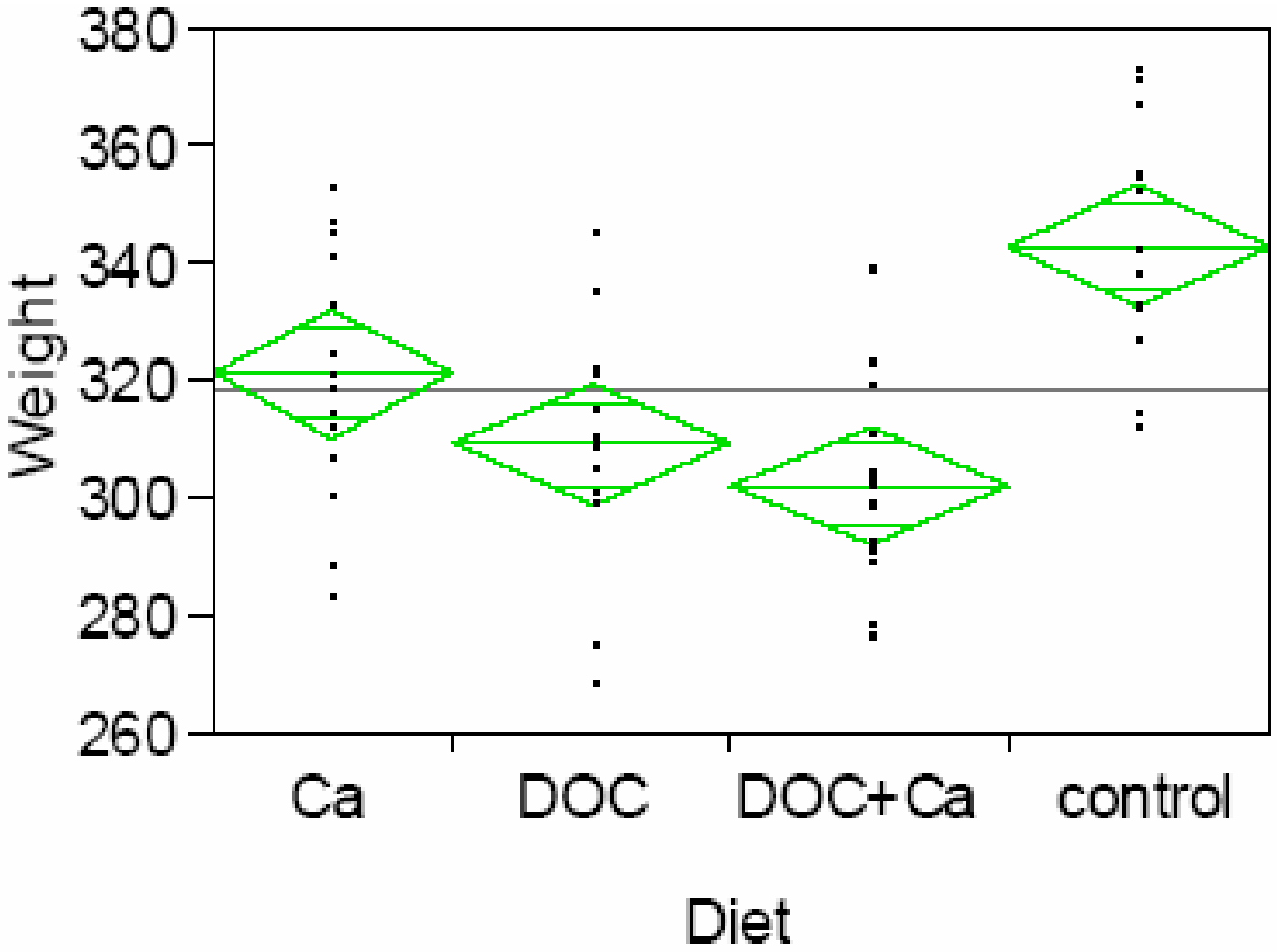
- We’ll look at each of the three options
- We’ll use JMP to work through and example

Which Statistic?

- **7. Calculate the statistic**
- There are three possible statistics that may be appropriate:
  1. an equal variance F-test,
  2. an unequal variance F-test, or
  3. the nonparametric Wilcoxon rank-sums test.

Equal variance

- If the equal variance assumption is tenable then the standard *F*-test is appropriate.
- Note: When reporting a *F*-test it's assumed that, unless you specify otherwise, it's the equal-variance *F*-test.
- We've seen the means diamonds in the situation with a two-group *t*-test.
- As the JMP Help shows, they represent the averages and a 95% confidence interval for each group.


380
360
340
320
300
280
260
Weight
Ca
DOC
DOC+Ca
control
Diet

- Recall that we talked about the relationship between confidence intervals and the two-group *t*-test.
- We said that you can interpret confidence intervals as follows: If two confidence intervals do not overlap (vertically) then the two groups are different (in large samples).

JMP Output

Oneway Anova

Summary of Fit

Rsquare	0.386929
Adj Rsquare	0.354662
Root Mean Square Error	20.17942
Mean of Response	318.6393
Observations (or Sum Wgts)	61

Analysis of Variance

Source	DF	Sum of Square	Mean Square	F Ratio	Prob > F
Diet	3	14649.154	4883.05	11.9915	<.0001
Error	57	23210.912	407.21		
C. Total	60	37860.066			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Ca	14	321.429	5.3932	310.63	332.23
DOC	16	309.375	5.0449	299.27	319.48
DOC+Ca	16	302.500	5.0449	292.40	312.60
control	15	343.133	5.2103	332.70	353.57

Std Error uses a pooled estimate of error variance

Interpretation

- So, the F is “large” ($F = 11.99$, with $df = 3, 57$), and the p-value is “small” (p-value < 0.0001).
- Note that the n’s and means in the report are the same as the means SD report.
- However, the standard errors are different. As the note says, these standard errors use the pooled estimate of variance; simply calculate the standard deviations within each group and divide by the square root of each n.
- Recall that the F-test is two-tailed; there is no direction of the difference. Since the null hypothesis specified a test for equality, this is the p-value we want.

Unequal Variance

- Use the Welch ANOVA results under the test for equal var.
- Report the denominator degrees of freedom to 1 decimal place
- Make sure you say “Welch ANOVA” or “unequal variance F”
- Otherwise, the interpretation is the same as the Equal Variance approach

Nonparametric comparison of the means

- If we wish to compare the *medians* we don't have to make any normality assumptions.
- So, we use a nonparametric test based solely on the ranks of the values of the Y-variable.
- The Wilcoxon rank-sum test (also called the Kruskal-Wallis test) simply ranks all the Y-values and then compares the sum of the ranks in each group.
- If the median of the first group is, in fact equal to the median of the second group, to the third group, etc, then the sum of the ranks should be equal.

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Score Mean	(Mean-Mean0)/Std0
Ca	14	473	33.7857	0.660
DOC	16	410	25.6250	-1.402
DOC+Ca	16	292.5	18.2813	-3.329
control	15	715.5	47.7000	4.188

1-way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
23.3083	3	<.0001

Interpretation

- When reporting the results of a nonparametric test, it's usual to only report the *p*-value, although reporting the chi-square value is also appropriate.
- A summary sentence: "The groups were compared using the nonparametric Kruskal-Wallis rank sum test and found to be different (chi-square = 23.3, *df* = 3, *p*-value < 0.0001)."

Steps 8 & 9

- **8. Make a statistical decision**
  - Using all three tests, the groups are different. (All *p*-values are less than 0.05.)
- **9. State the substantive conclusion**
  - The four diets have different means.

Phase 4: Communicate the Answer to the Question

- **10. Document our understanding with text, tables, or figures**
- A total of *n* = 61 Wistar-Kyoto (WKY) rats were assigned to one of four dietary groups: untreated controls, high calcium diet (Ca), deoxycortioterone-NaCl treated rats (DOC), and rats receiving both dietary supplements (DOC+Ca). The ANOVA test indicated that the groups were significantly different ($F(3, 57) = 12$, $p < .0001$). (Show a table of summary statistics and a plot of the means.)

So Far: Multiple Independent Means

- Briefly, here is how we proceeded when comparing the means obtained from multiple independent samples.
  - Describe the groups and the values in each group. What summary statistics are appropriate? Are there missing values? (why?)
  - Assess the assumptions, including normality and equal variance. If normality is warranted, then it may be useful to determine confidence intervals on each of the means.
  - Perform the appropriate statistical test. Determine the *p*-value that corresponds to your hypothesis.
  - Reject or fail to reject? State your substantive conclusion.
- Is that it? What about where the means are different – which groups are the same?

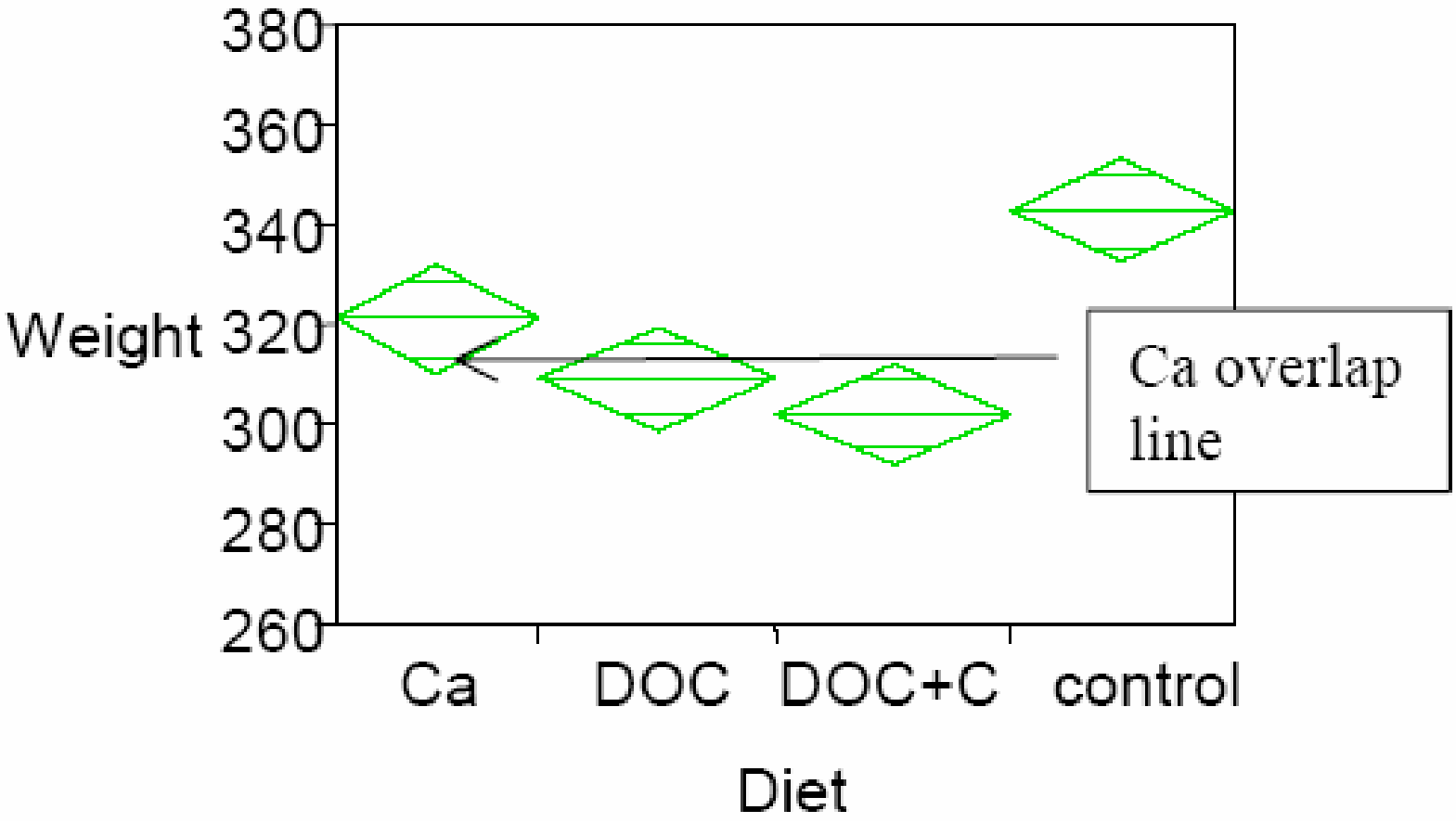
Considering the means

- If we look at the ordered means, it would appear that—from lightest to heaviest—the four groups are: DOC+Ca (303 gm), DOC (309 gm), Ca (321 gm), and control (343 gm).
- Questions:
  - Is DOC+Ca significantly different than DOC?
  - Is DOC+Ca significantly different than Ca?
  - Is DOC+Ca significantly different than control?
  - Is DOC significantly different than Ca?
  - Is DOC significantly different than control?
  - Is Ca significantly different than control?

- All we’ve decided is that there is a difference somewhere; that all four of the means are not equal.
- But we’re far from explaining where the difference(s) lie.
- We have an impression that the control group is higher than the others, but are there any other differences?

Graphical comparison

- Recall the interpretation of means diamonds. Since they are confidence intervals, if they don’t overlap then the groups are different.
- However, what if they don’t overlap? Are the groups different? The answer is, if they overlap “a little” then the groups are different? How much is “a little”?
- The smaller horizontal lines within the means diamonds are one way to tell. These are called significance overlap lines.
- If we pay attention to whether the area within the overlap lines in two groups separate, then we can (roughly) see whether the groups are different

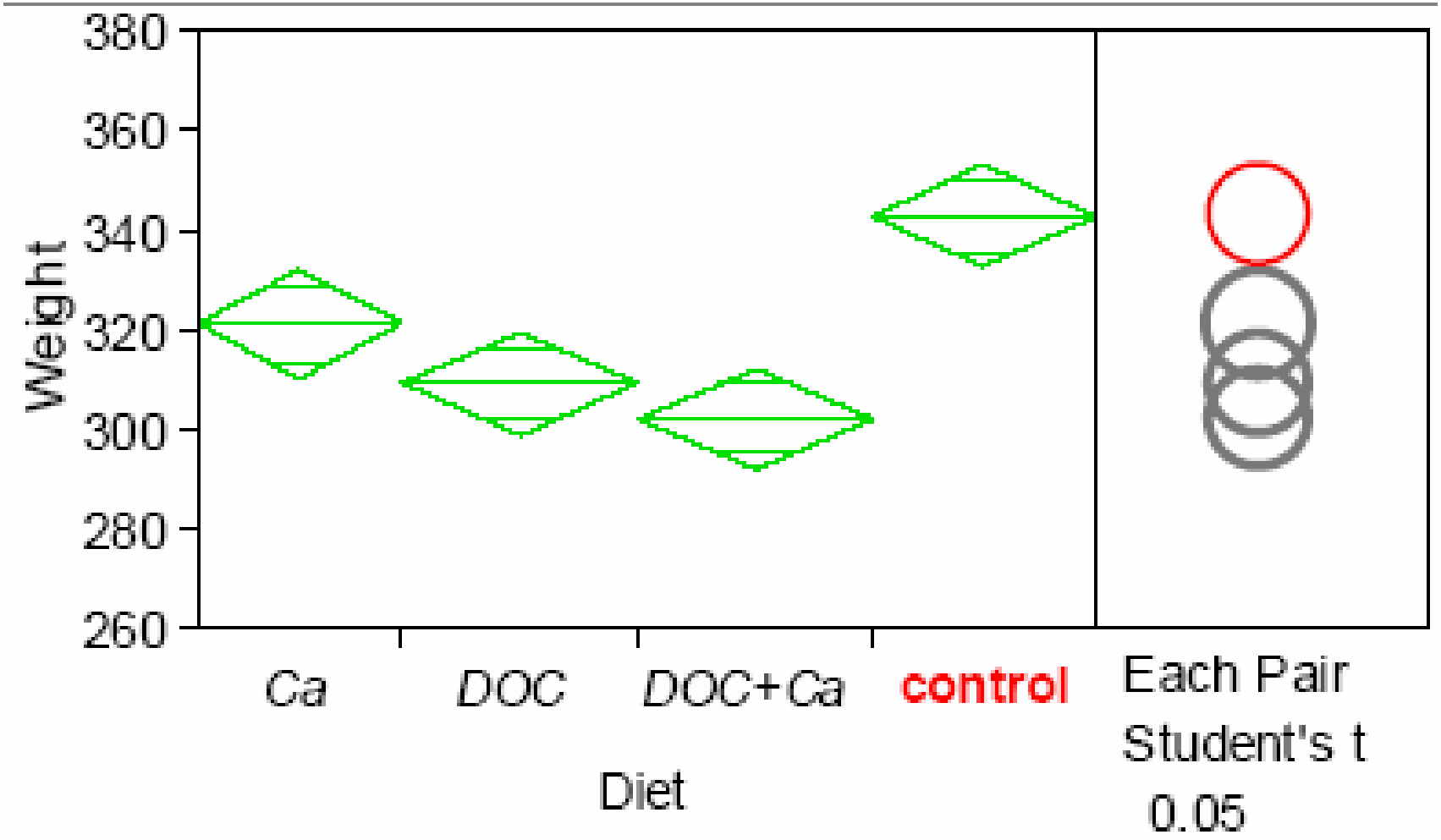

380
360
340
320
300
280
260
Weight
Ca
DOC
DOC+C
control
Diet
Ca overlap line

Interpretation

- In the above figure, we show the lower limit of the Ca group's overlap line.
- Extending it across to the right we see that the DOC group's overlap lines are within the lower Ca limit.
- We also see that the lower Ca limit is above the DOC+Ca group's overlap lines.
- From this we gain the impression that the Ca and DOC groups may not be different but the Ca and DOC+Ca groups may be different.
- But we need a definitive answer.

All possible t-tests

- One solution to this problem is to do all possible *t*-tests.
- We could answer each of the six questions above as though we had done a series of two-group studies.
- Note: we **do not do this in practice**. It is NOT a good idea.


380
360
340
320
300
280
260
Weight
Ca
DOC
DOC+Ca
control
Each Pair
Student's t
0.05
Diet

Multiple comparisons

- The more statistical tests we do, eventually one will come out significant.
- We didn't have this problem in the two-group *t*-test situation.
- We just had two groups and there were two alternative: the groups were not different or the groups were different.
- There was only one *t*-test—and it was controlled by a Type I error rate, a = 0.05.
- We only say "significant difference!" when it's *not* true, 5% of the time.

Multiple Comparisons

- With three groups, there are three possible *t*-tests to do.
- If *each* test has α = 0.05, then the probability of saying “significant difference!” at least once when it’s *not* true, is *more* than 5%
- Doing each test at a = 0.05 does *not* yield an experiment whose overall type I error rate is a = .05.

Number of groups	Number of comparisons	alpha
2	1	0.050
3	3	0.143
4	6	0.265
5	10	0.401
6	15	0.537
7	21	0.659
8	28	0.762
9	36	0.842
10	45	0.901
11	55	0.940
12	66	0.966
13	78	0.982
14	91	0.991
15	105	0.995

- This is what is called *the multiple comparison problem.*
- So, even if there really is no difference, if you do the three *t*-tests required to compare three groups, at least one of them will appear “significant”—by chance alone—over 14% of the time.
- If we have 6 groups, this error rate is over 50% and by the time we’re up to 10 groups we’re virtually assured of “finding” a “significant” difference.
- We need to “correct” for the number of tests we are performing so that the error rate stays at 5%.

The Bonferroni correction

- The Bonferroni correction says this: If we want to operate with $\alpha = 0.05$, then count up the number of comparisons we're doing—the second column in the table above, call this number *k*—and do each *t*-test at *p*-value < (0.05/*k*).
- So, with three groups there are three comparisons (*k* = 3), so compare the *p*-values to < 0.01667.

Bonferroni (cont)

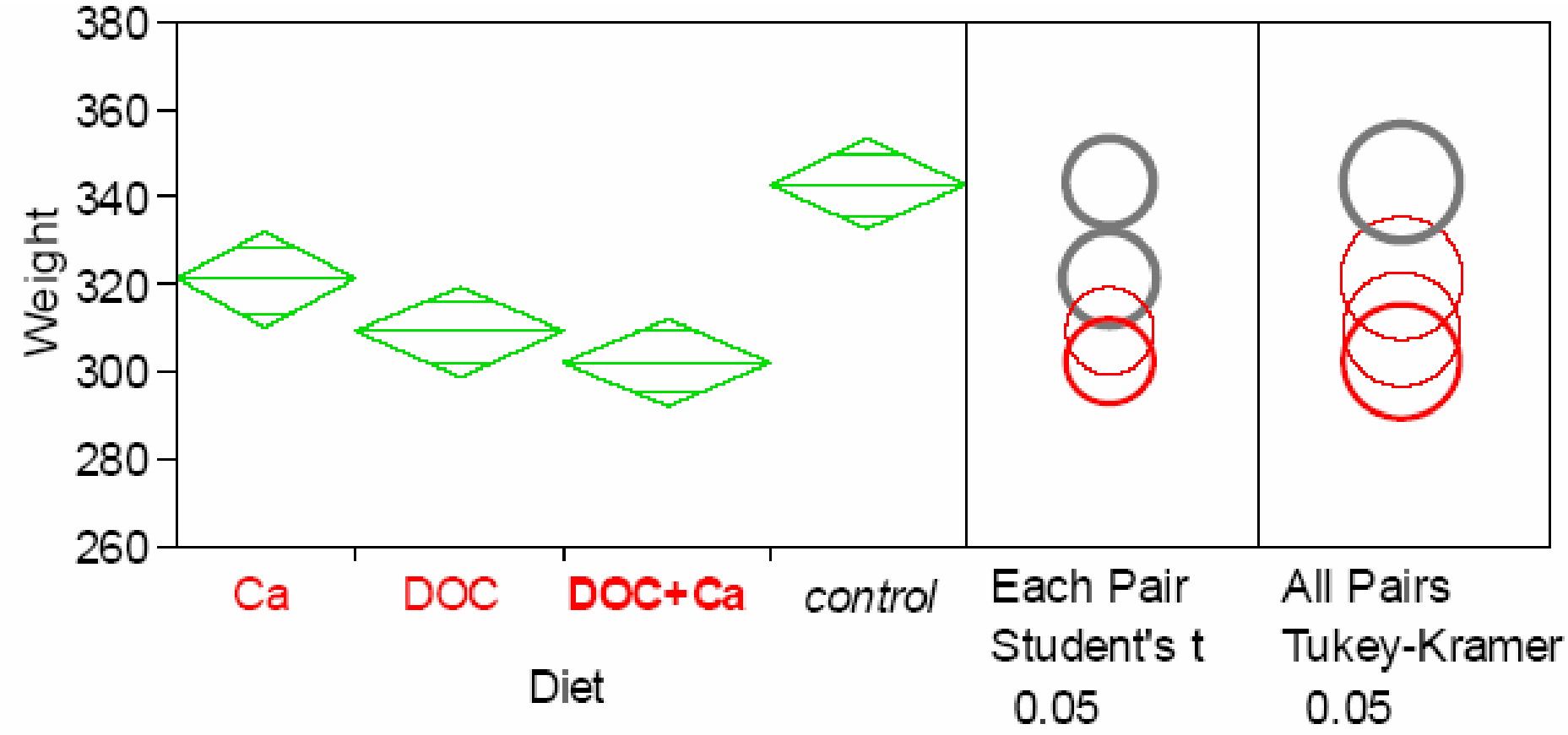
- This approximation is simple and works fairly well as long as you don't mind how conservative it is.
- It's very hard to find a significant difference using Bonferroni-corrected *p*-values.
- On the other hand, if you can declare a difference with this severe a penalty, it's believable.
- There is a better way to handle the multiple comparison problem.

Tukey-Kramer

- Doing all possible *t*-tests without some sort of modification of significance level is a bad idea.
- Then what should we do?
- Use Tukey-Kramer *honestly significant difference* test.
- The HSD takes a much more sophisticated approach to this problem.
- The HSD looks at the distribution of the “biggest” difference.
- That is, the test comparing the smallest mean to the largest mean.

Tukey

- Important things to know:
  - The observed differences are compared to an appropriate standard that's larger than the *t*-test standard.
  - The standard is arrived at so that the overall experiment-wise error rate is 0.05.


380
360
340
320
300
280
260
Weight
Ca
DOC
DOC+Ca
control
Diet
Each Pair
Student's t
0.05
All Pairs
Tukey-Kramer
0.05

Comparison

- Notice how—by an uncorrected t-test—the Ca and DOC+Ca groups *are* declared "significantly different."
- By the HSD, the Ca and DOC+Ca groups are *not* different.
- Notice also that the comparison circles for HSD are larger than the Student's *t* comparison circles.
- This is a reflection of the higher standard for calling a difference "significant."

Tukey

- You can believe the HSD results.
- **Bottom line:** If we're interested in all possible differences between the groups, use **Compare all pairs, Tukey's HSD** to compare means.

Means Comparisons

Dif=Mean[i]-Mean[j]

	control	Ca	DOC	DOC+Ca
control	0.000	21.705	33.758	40.633
Ca	-21.705	0.000	12.054	18.929
DOC	-33.758	-12.054	0.000	6.875
DOC+Ca	-40.633	-18.929	-6.875	0.000

Alpha=0.05

Comparisons for all pairs using Tukey-Kramer HSD

q*
2.64647

Abs(Dif)-LSD

	control	Ca	DOC	DOC+Ca
control	-19.500	1.859	14.565	21.440
Ca	1.859	-20.185	-7.490	-0.615
DOC	14.565	-7.490	-18.881	-12.006
DOC+Ca	21.440	-0.615	-12.006	-18.881

Positive values show pairs of means that are significantly different.

Means Comparisons

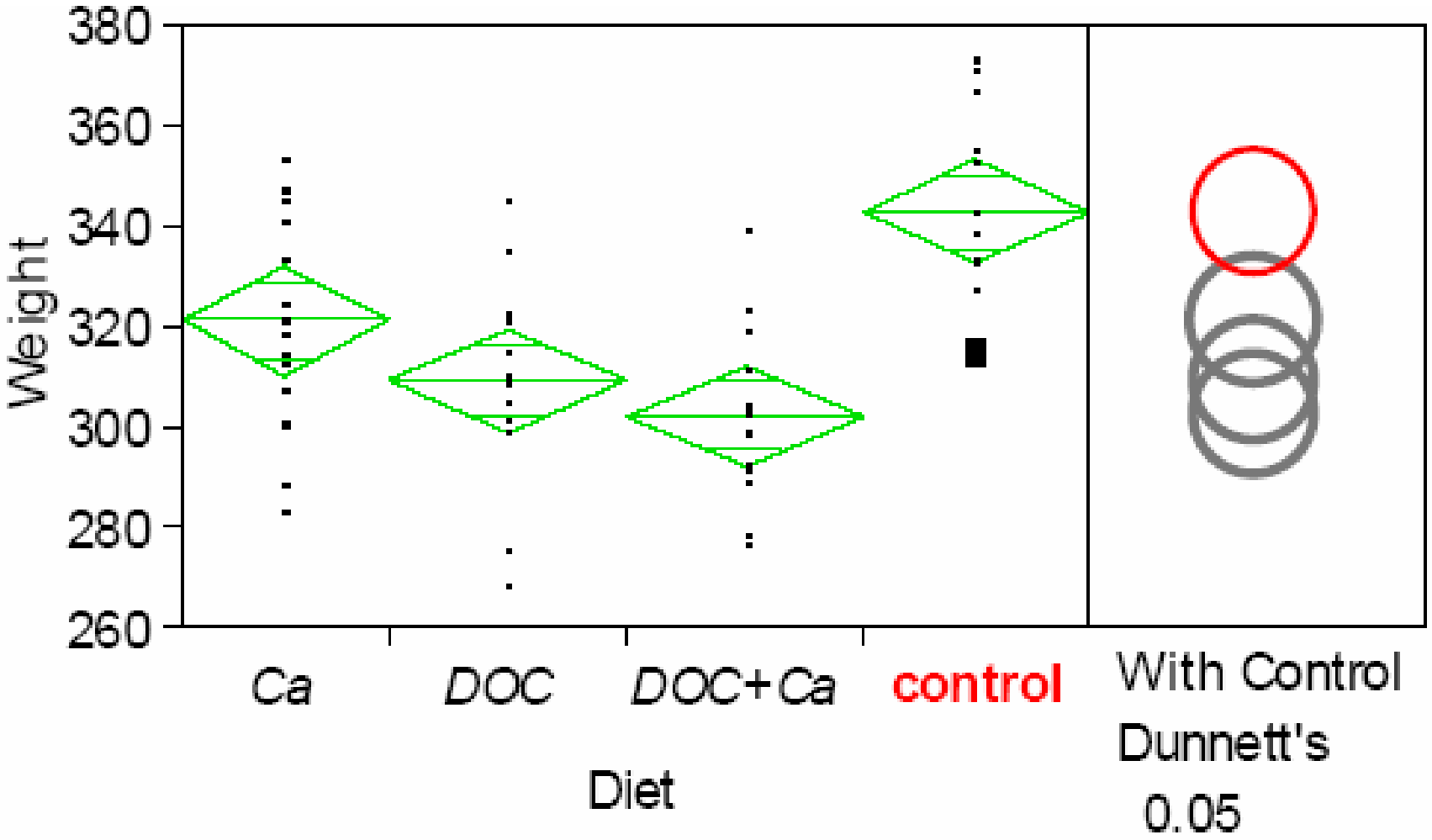
- The Means Comparisons report first shows a table of all possible ordered differences between the means.
- It’s arranged so that the smallest differences are just off the diagonal and the largest difference is in the upper right-hand (or lower left-hand) corner.
- Positive values [in the lower report] show pairs of means that are significantly different

JMP Update

- JMP now provides the 95% CIs on the difference between each pair
- There is also an additional report that separates out the groups by giving all similar groups the same letter.

Comparison with a control

- If we're not interested in all the possible mean differences then the HSD is too conservative.
- The most common situation when this occurs is when the experiment is only interested in whether a mean is different than a pre-planned control group.
- That is *if* these are the *only* questions of interest:
  - Is DOC+Ca significantly different than control?
  - Is DOC significantly different than control?
  - Is Ca significantly different than control?


380
360
340
320
300
280
260
Weight
Ca
DOC
DOC+Ca
control
Diet
With Control
Dunnett's
0.05

Phase 4: Communicate the Answer to the Question

- **10. Document our understanding with text, tables, or figures**
- Now we complete the description of our conclusions begun earlier:
- Using Tukey's HSD, it was determined that the control diet had significantly higher weight (mean = 343 gm) than each of the special-diet groups. Within the three special diet groups: DOC+Ca, DOC, and Ca (combined mean* 311 gm, SD = 21.6), there was no significant difference.

Means and Plots

- The earlier summary table of means, SD's and CI's is probably sufficient.
- However, if we wanted to show a figure, the best depiction using JMP would be the dot plot and CI's represented by the means diamonds.
- ***Note:*** There are any number of ways to calculate the combined mean and SD.

Summary: Comparing Multiple Independent Means

- Describe the groups and the values in each group. What summary statistics are appropriate? Are there missing values? (why?)
- Assess the assumptions, including normality and equal variance. If normality is warranted, then it may be useful to determine confidence intervals on each of the means.
- Perform the appropriate statistical test.
- If Normality is unwarranted then use the Rank-Sum test (or consider transforming the Y-variable to make it more normal).

Summary (cont)

- If Normality and equal variance are apparent, then use the F-test.
- If Normality and unequal variance are apparent, then use the Welch ANOVA F-test (or consider transforming the Y-variable to equalize variance). Determine the *p*-value that corresponds to your hypothesis.
- Null hypothesis: No difference. Reject or fail to reject?

Summary (cont)

- If we fail to reject: There is no evidence for more than one mean. Report the single mean & etc.
- If we reject: Then use the appropriate multiple comparison test to determine which groups are significantly different. Report means & etc. that reflect the pattern that is evident.

Summary (cont)

- **Indeterminate results**
- Note: The following scenarios are possible.

1. The *F*-test is *not* significant but one or more of the group comparisons *is* significant.
   - No fair; you weren't supposed to look at the group comparisons if the overall test was not significant. "Fishing" is not allowed.
2. The *F*-test *is* significant but none of the group comparisons is significant. In other words, the F-test says there is a difference but we can't find it. This will sometimes happen (and it's irritating when it does). All you *should* do is report it (or redo the study with larger *n*). What people really do is "fish" until they find a plausible conclusion (be careful how you report this).

3. The F-test is significant. When the group comparisons are considered, the pattern of the means is difficult to interpret. It's even possible—especially when the *n*'s in each group are different—that groups that "should" be different, aren't.
   - For example, with a three-group study, it's possible that the multiple comparison tests indicate that:
     - Group A > Group B, and
     - Group B > Group C, but
     - Group A is not > Group C (!?)
   - All you can do is report the overall F-test and the results or just the F-test and re-run the study in a balanced fashion with larger n.

Next Semester

- We'll pick up with ANOVA
- Look at:
  - Logistic Regression
  - Dependant Means
  - Multiple Time measures
  - Multiple Random Variables
  - Multiple Outcomes
  - And more!