Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Hand in HW4.
2. Testing the slope or correlation, continued.
3. Regression conditions.
4. Comparing more than two means using MAD and simulations.
5. ANOVA and comparing more than two means using a theory based approach.

Read ch9.

The final is Fri Dec14, 8-11am.

Bring a PENCIL and CALCULATOR and any books or notes you want. No computers.

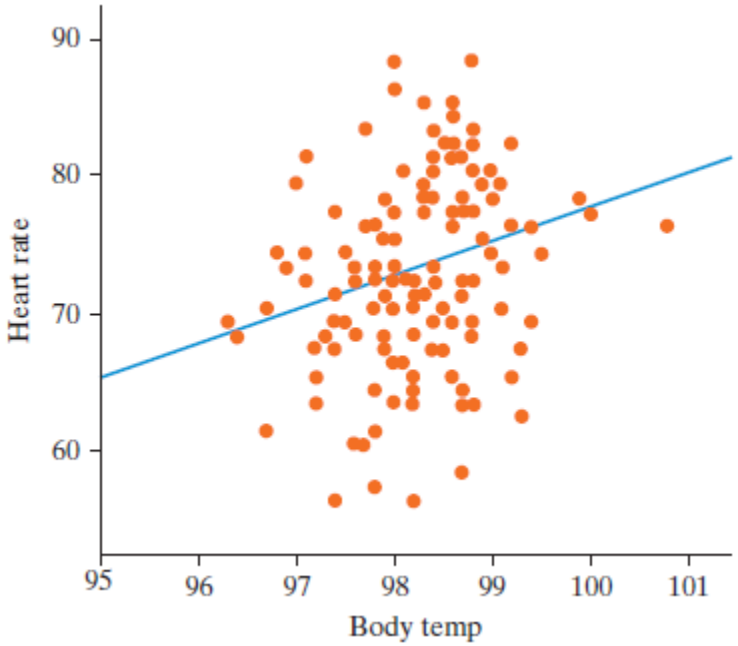
http://www.stat.ucla.edu/~frederic/13/F18.

1. Hand in HW4.
2. Predicting Heart Rate from Body Temperature

Example 10.5A

Heart Rate and Body Temp

- Earlier we looked at the relationship between heart rate and body temperature with 130 healthy adults
- $\text{Predicted Heart Rate} = -166.3 + 2.44(\text{Temp})$
- r = 0.257

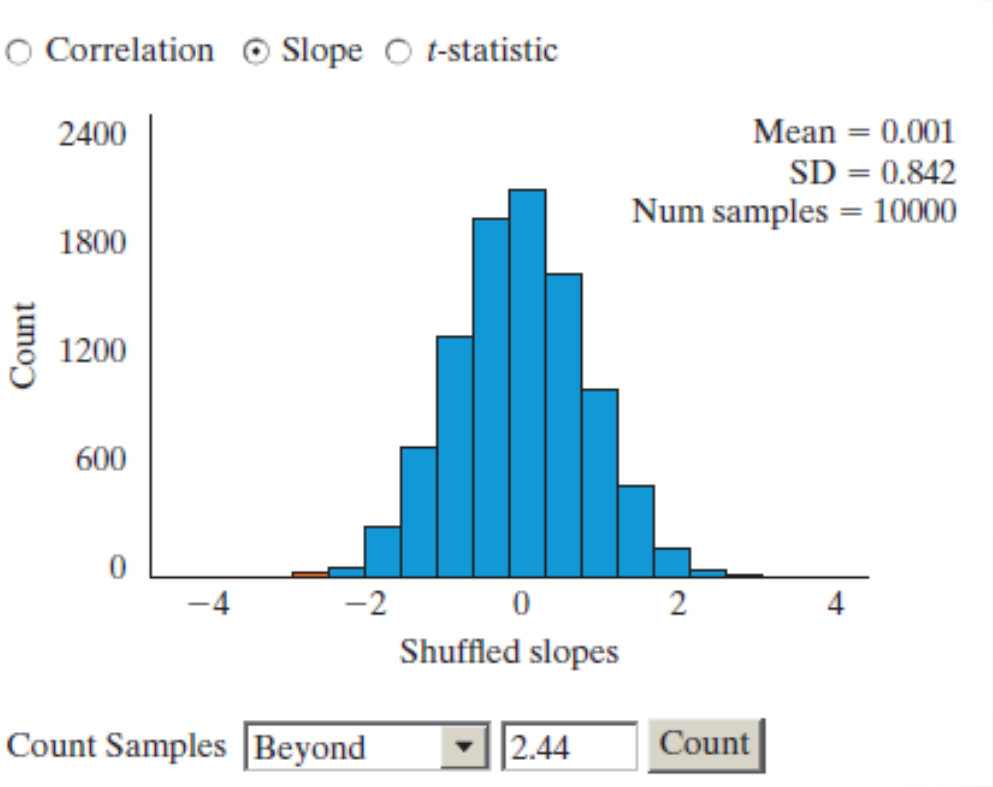


Heart Rate and Body Temp

- We tested to see if we had convincing evidence that there is a positive association between heart rate and body temperature in the population using a simulation-based approach. (We will make it 2-sided this time.)
- **Null Hypothesis**: There is no association between heart rate and body temperature in the population. $\beta = 0$
- **Alternative Hypothesis**: There is an association between heart rate and body temperature in the population. $\beta \neq 0$

Heart Rate and Body Temp

We get a very small p-value (0.0036). Anything as extreme as our observed slope of 2.44 happening by chance is very rare



Heart Rate and Body Temp

- We can also approximate a 95% confidence interval

  observed statistic $\pm$ 1.96 SE

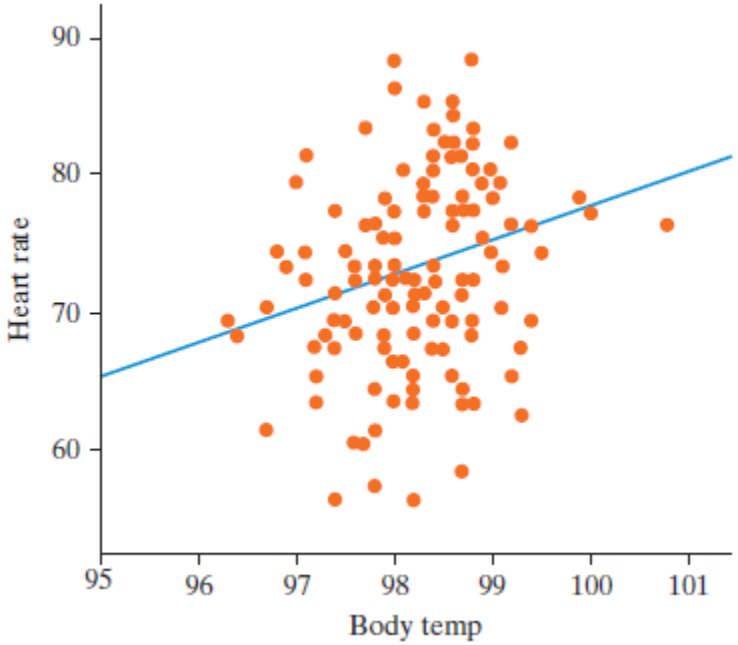
  $2.44 \pm 1.96(0.842) = 0.790$ to 4.09

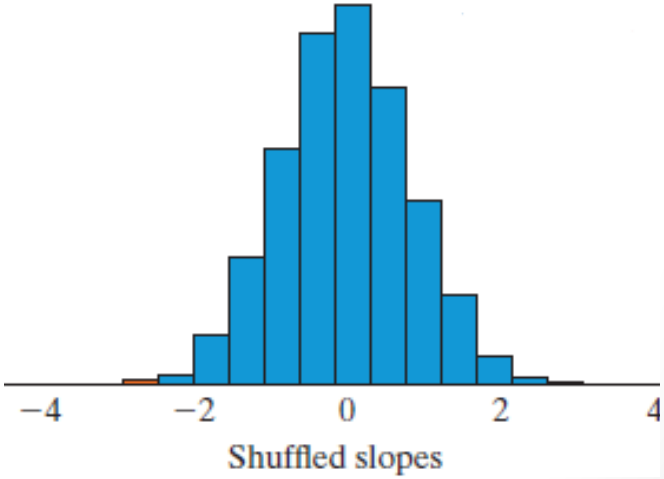
- What does this mean?

  We're 95% confident that, in the population of healthy adults, each 1° increase in body temp is associated with an increase in heart rate of between 0.790 to 4.09 beats per minute

Heart Rate and Body Temp

- The theory-based approach should work well since the distribution has a nice bell shape
- Also check the scatterplot





Heart Rate and Body Temp

- We will use the t-statistic to get our theory-based p-value.
- We will find a theory-based confidence interval for the slope.
- On p554, the book notes the formula
- $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$.
- Here the t statistic is 2.97.
- The p-value is .36%. So the correlation is statistically significantly greater than zero.

Smoking and Drinking

Example 10.5B

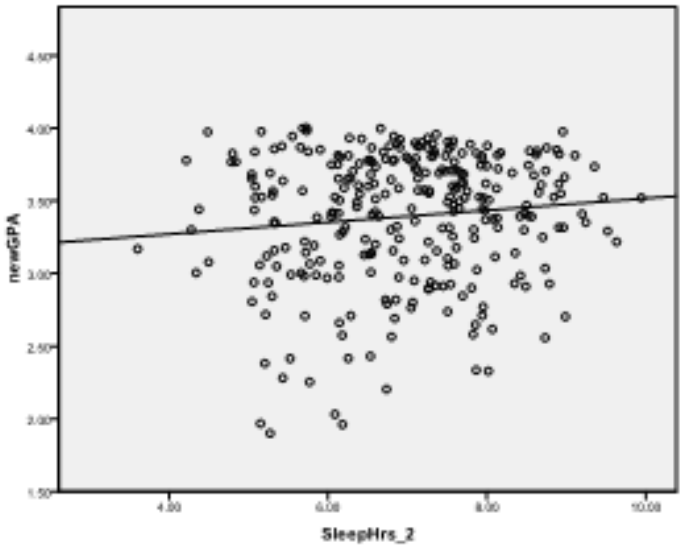
Validity Conditions

Remember our validity conditions for theory-based inference for slope of the regression equation.

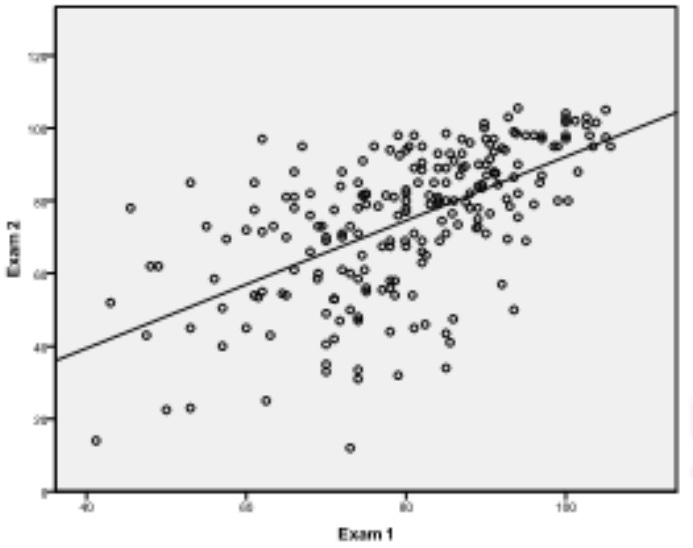
1. The scatterplot should follow a linear trend.
2. There should be approximately the same number of points above and below the regression line (symmetry).
3. The variability of vertical slices of the points should be similar. This is called homoskedasticity.

Validity Conditions

- Let's look at some scatterplots that do not meet the requirements.



(a)



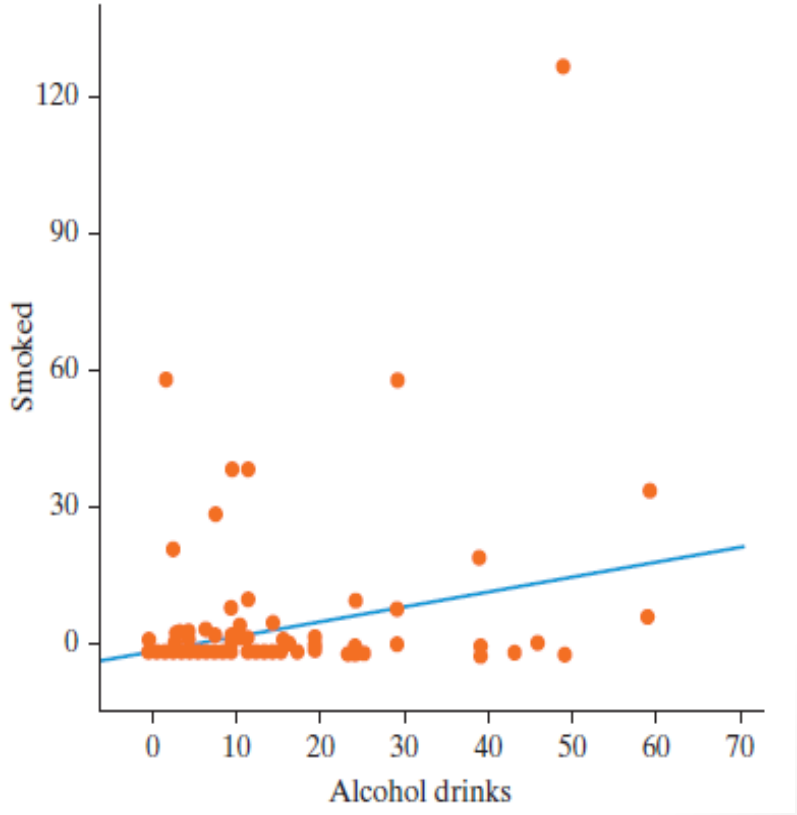
(b)

Smoking and Drinking

The relationship between number of drinks and cigarettes per week for a random sample of students at Hope College.

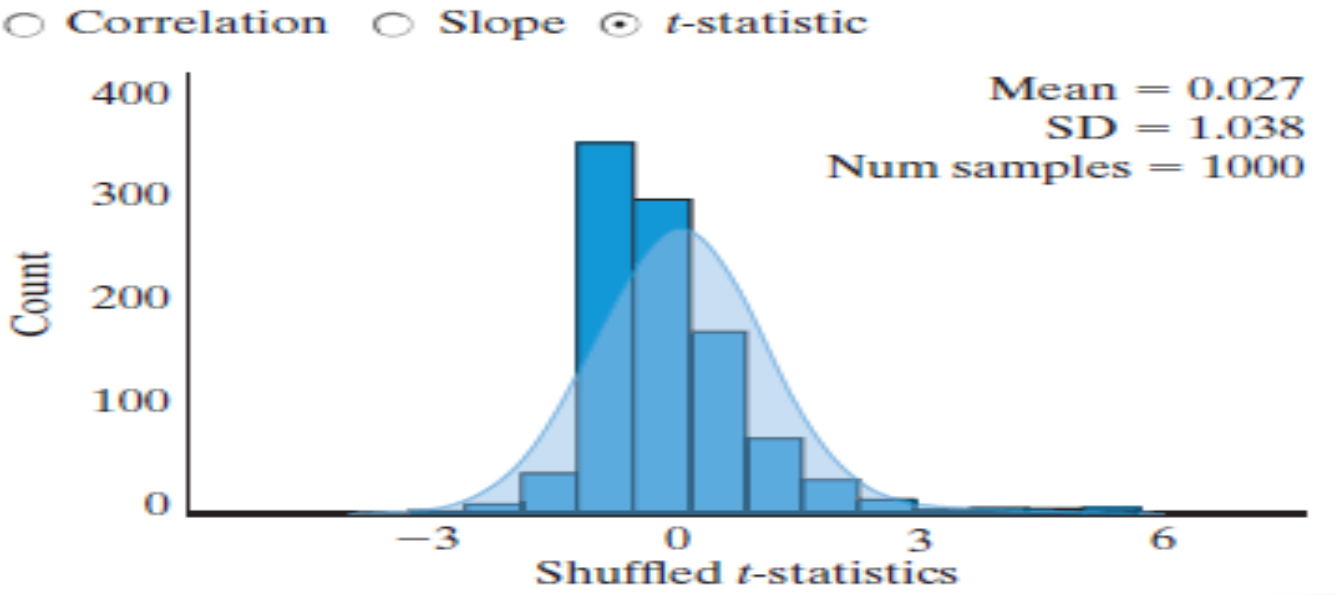
The dot at (0,0) represents 524 students

Are the conditions met? Hard to say. The book says no.



Smoking and Drinking

- When the conditions are not met, applying simulation-based inference is preferable to theory-based t-tests and CIs.
- One can also fit a curve, instead of a line.



Comparing more than 2 means.

Review of Simulation-Based Tests

- One proportion:
- We created a null distribution by flipping a coin, rolling a die, or some computer simulation.
- We then found where our sample proportion was in this null distribution.

Simulation-Based Tests

- Comparing two proportions:
  - Assuming there was no association between explanatory and response variables (the difference in proportions is zero), we shuffled cards and dealt them into two piles. (This essentially scrambled the response variable.)
  - We then calculated the difference in proportions.
  - We repeated this process many times and built a null distribution.
  - We finally found where the observed difference in sample proportions was located in the null distribution.

Simulation-Based Tests

- Comparing two means:
  - Assuming there was no association between explanatory and response variables (the difference in means is zero), we shuffled cards and dealt them into two piles. (This time the cards had numbers on them, the response, instead of words)
  - We then calculated the difference in means
  - We repeated this process many times and built a null distribution.
  - We finally found where the observed difference in sample means was located in the null distribution.

Simulation-Based Tests

- Paired Test:
  - Assuming there was no relationship between the explanatory and response variables (so the mean difference should be zero), we randomly switched some of the pairs and calculated the mean of the differences
  - We repeated this many times and built a null distribution.
  - We then found where the original mean of the differences from the sample was located in the null distribution.

Two more types of tests

- We now want to compare multiple (more than two) means.
- In chapter 10 we looked at the association between two quantitative variables using correlation and regression.
- Both of these processes are basically the same as most of the simulation-based tests we did in chapters 1-7. Just the data types (or number of categories) and the statistic we use is different.

Comparing Multiple Means: Simulation-Based Approach

Section 9.1

Pre-Example

- Suppose we wanted to compare how much various energy drinks increased people's pulses.
- We would end up with a number of means.

Controlling for Type I Error

- We could do this with multiple tests where we compared two means at a time, but
  - If we were comparing 3 means, we would have to use 3 two-sample tests to compare these three means. (A vs B, B vs C, and A vs C)
- If each test has a 5% significance level, there's a 5% chance of making a **Type I Error.**
  - We can call this a false alarm.
  - This is the error in rejecting the null when it is true.
  - With 5% probability, if there really is no difference between our groups, we will get a result out in the tail just by chance alone.

Controlling for Type I Error

- These type I errors “accumulate” when we do more tests on the same data.
  - At the 5% significance level, the probability of making at least one Type I error for three tests would be 14%.
  - Comparing 4 means (6 tests), this probability of at least one Type I error jumps to 26%.
  - Comparing 5 means (10 tests), this jumps to 40%.
- An alternative approach uses one overall test that compares all means at once.

Overall Test

- If I have two means to compare, we just need to look at their difference to measure how far apart they are.
- Suppose we wanted to compare three means. How could I create something that would measure how different all three means are?

MAD Statistic

- We can use the MAD statistic.
- MAD = (|avg1 – avg2|+|avg1 – avg3|+ |avg2 – avg3|)/3.
- Let’s try this on an example.

Comprehending Ambiguous Prose

Example 9.1

Comprehension Example

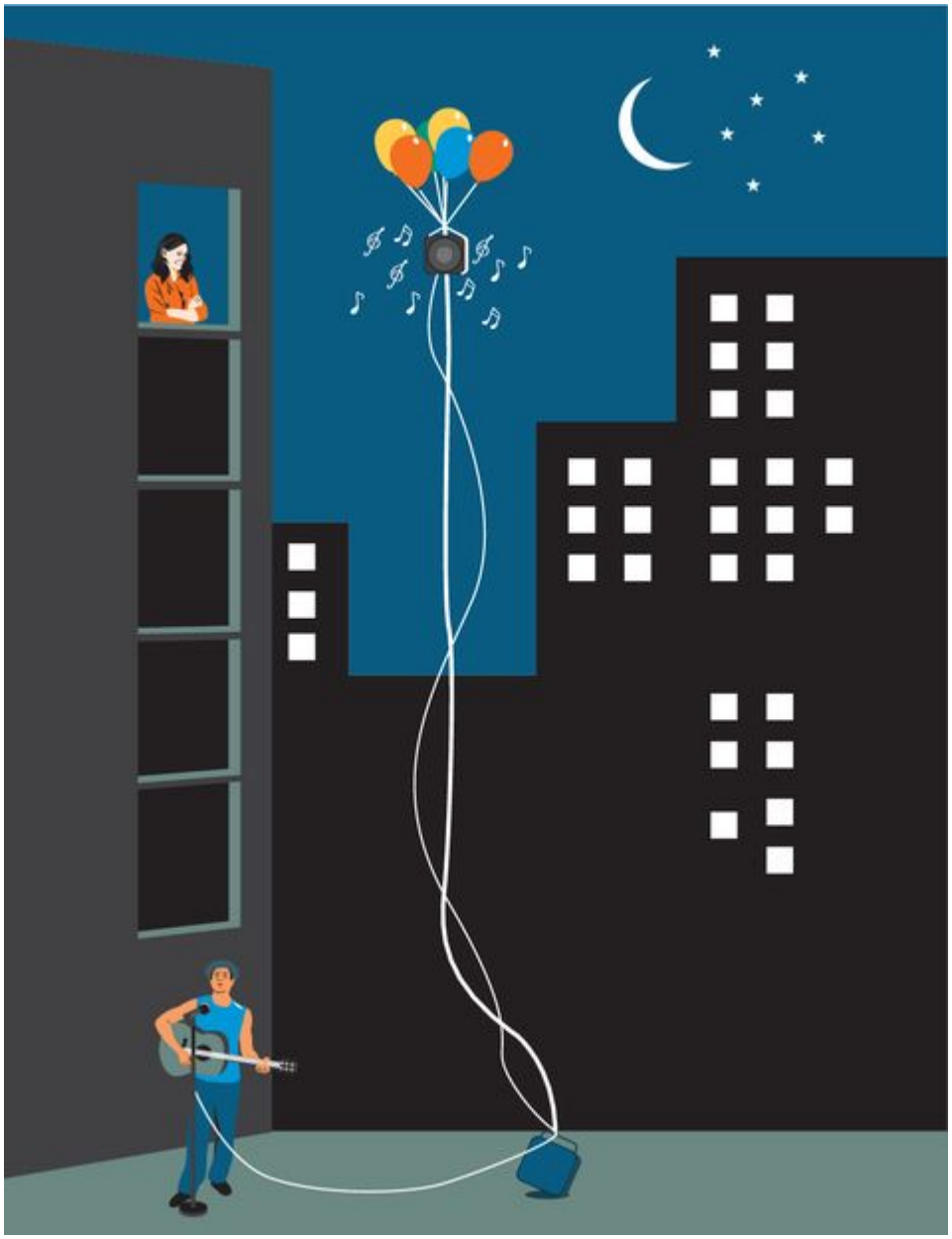
(**Don't** follow along in your book or look ahead on the PowerPoint until after I read you the passage.)

- Students were read an ambiguous prose passage under one of the following conditions:
  - Students were given a picture that could help them interpret the passage **before** they heard it.
  - Students were given the picture **after** they heard the passage.
  - Students were **not** shown any picture before or after hearing the passage.
- They were then asked to evaluate their comprehension of the passage on a 1 to 7 scale.

Comprehension Example

- This experiment is a partial replication done at Hope College of a study done by Bransford and Johnson (1972).
- Students were randomly assigned to one of the 3 groups.
- Listen to the passage and see if it makes sense. Would a picture help?

If the balloons popped, the sound wouldn't be able to carry since everything would be too far away from the correct floor. A closed window would also prevent the sound from carrying, since most buildings tend to be well insulated. Since the whole operation depends on a steady flow of electricity, a break in the middle of the wire would also cause problems. Of course, the fellow
could shout, but the human voice is not loud enough to carry that far. An additional problem is that a string could break on the instrument. Then there could be no accompaniment to the message. It is clear that the best situation would involve less distance. Then there would be fewer potential problems. With face to face contact, the least number of things could go wrong.



Hypotheses

- **Null:** In the population there is no association between whether or when a picture was shown and comprehension of the passage
- **Alternative:** In the population there is an association between whether and when a picture was shown and comprehension of the passage

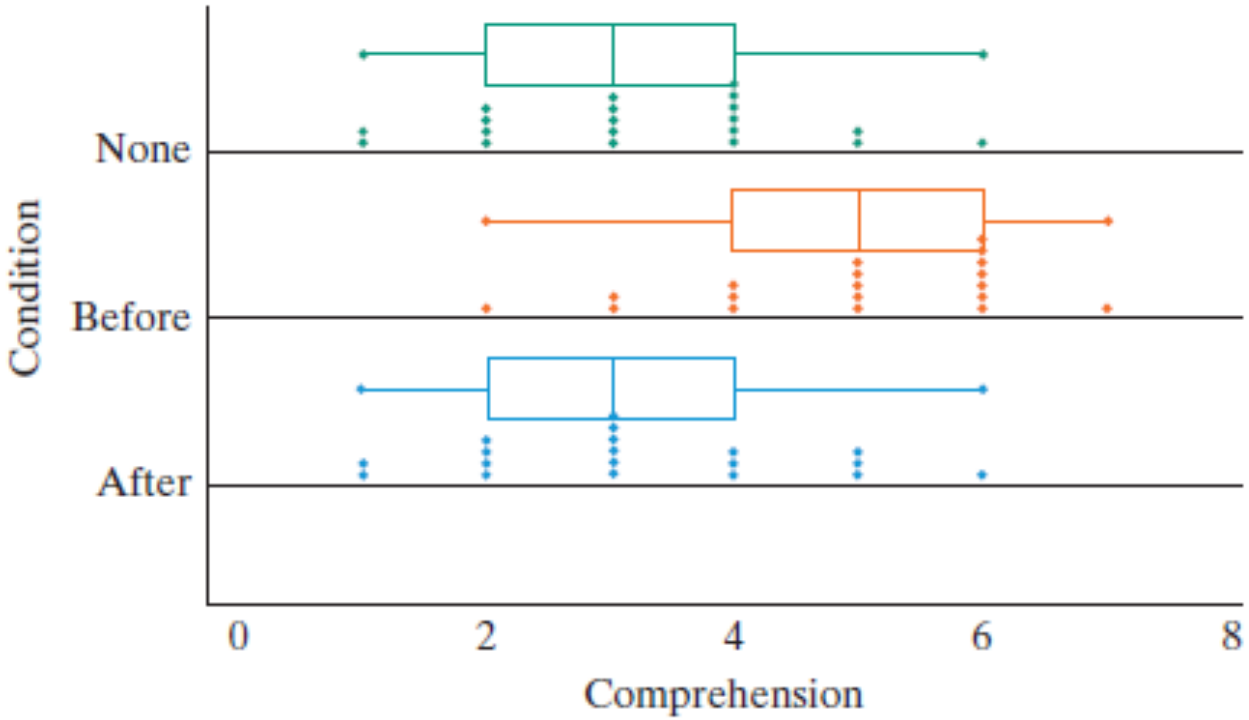
Hypotheses

- **Null:** All three of the long term mean comprehension scores are the same.

$$\mu_{\text{no picture}} = \mu_{\text{picture before}} = \mu_{\text{picture after}}$$

- **Alternative:** At least one of the mean comprehension scores is different.

Results



Means

3.37

4.95

3.21

Calculating the MAD

$$MAD = (|3.21-4.95|+|3.21-3.37|+|4.95-3.37|)/3$$
$$= (1.74 + 0.16 + 1.58)/3$$
$$= 3.48/3$$
$$= 1.16.$$

- What is the likelihood of getting a statistic as large as (or larger than) this by chance if there were really no difference in comprehension between the three groups?
- What types of values (e.g., large, small, positive, negative) of this statistic will give evidence against the null hypothesis?

Simulation

- Similar to testing two means we can shuffle values of the response variable (the comprehension scores) and randomly place them into piles representing the categories of the explanatory variable (the picture condition).
- This time we have three piles instead of two.
- After each shuffle, we calculate the MAD statistic of the shuffled data and that will be a point in the null distribution.

None

2	5	3	4
2	3	2	4
2	4	4	6
4	3	3	1
5	3	4	

mean = 3.74

Before

5	4	2	6
6	3	6	6
6	5	4	6
5	6	4	3
7	5	5	

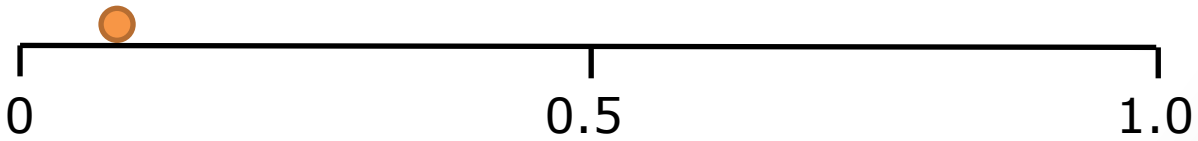
mean = 4.84

After

6	5	4	2
1	3	3	5
3	2	2	1
4	4	3	2
5	3	3	

mean = 3.95

MAD = (|3.74 – 3.84| + |3.74 – 3.95| + |3.84 – 3.95|)/3 = 0.14

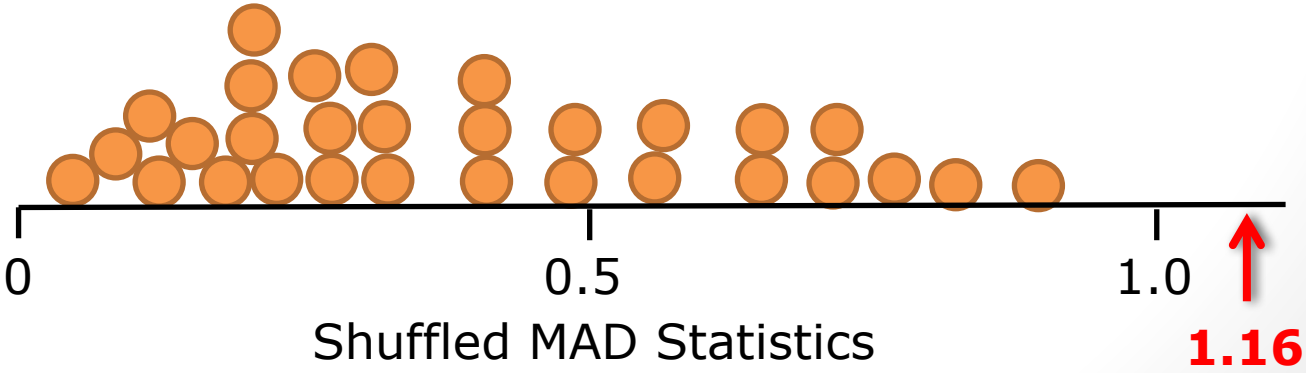


Shuffled MAD Statistics

More Simulations

0.05 0.60 0.77 0.18 0.32
0.21 0.53 0.46 0.60
0.39 0.21 0.42 0.49
0.11
0.07 0.32 0.11
0.07 0.18 0.39
0.63
0.32 0.74 0.53 0.07

With 30 repetitions of creating simulated MAD statistics, we did not get any that were as large as 1.16.



Conclusion

- The results of many simulations are shown on p487 of the book. For comprehension, a MAD of 1.16 corresponds to a p-value of < 0.001.
- Since we have a small p-value we can conclude that the differences between the mean comprehension scores are statistically significant. We reject the null hypothesis that the 3 population means would actually be the same.
- Can we tell which one or ones are different?
- Go back to the dotplots and take a look.
- We can do pairwise confidence intervals to find which means are significantly different than the other means.

Learning Objectives for Section 9.1

- Be able to calculate the MAD statistic given a data set (or set of means).
- Understand how a simulation-based test would work using cards and shuffling for comparing multiple means.
- Understand that we do an overall test when comparing multiple means or proportions instead of pairwise tests to control for the probability of making a type I error.
- Use the MAD statistic to compare multiple means.

3. Comparing Multiple Means: Theory-Based Approach (**AN**alysis **O**f **Va**riance ANOVA)

Section 9.2

ANOVA

- The convention is to use a statistic slightly different from the MAD for theory based tests.
- This new statistic is called an F statistic and the theory-based distribution that estimates our null distribution is called an F distribution.
- Unlike the MAD statistic, the F statistic takes into account the variability within each group.

F test statistic

- The analysis of variance *F* test statistic is:

$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

- This is similar to the t-statistic when we were comparing just two means. $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}}$

F test statistic

$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

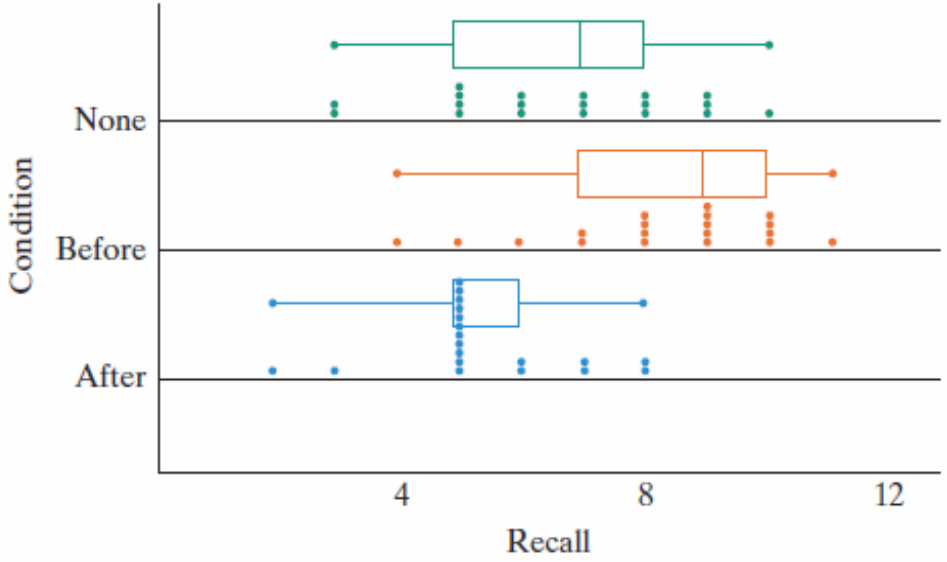
- Remember measures of variation are always non-negative. (Our measure of variation can be zero when all values in the data set are the same.)
- So, our F statistic will be non-negative.

Recalling Ambiguous Prose

Example 9.2

Recall Score

- Going back to the ambiguous prose example, the students rated their comprehension, and the researchers also had the students recall as many ideas from the passage as they could. They were then graded on what they could recall and the results are shown.

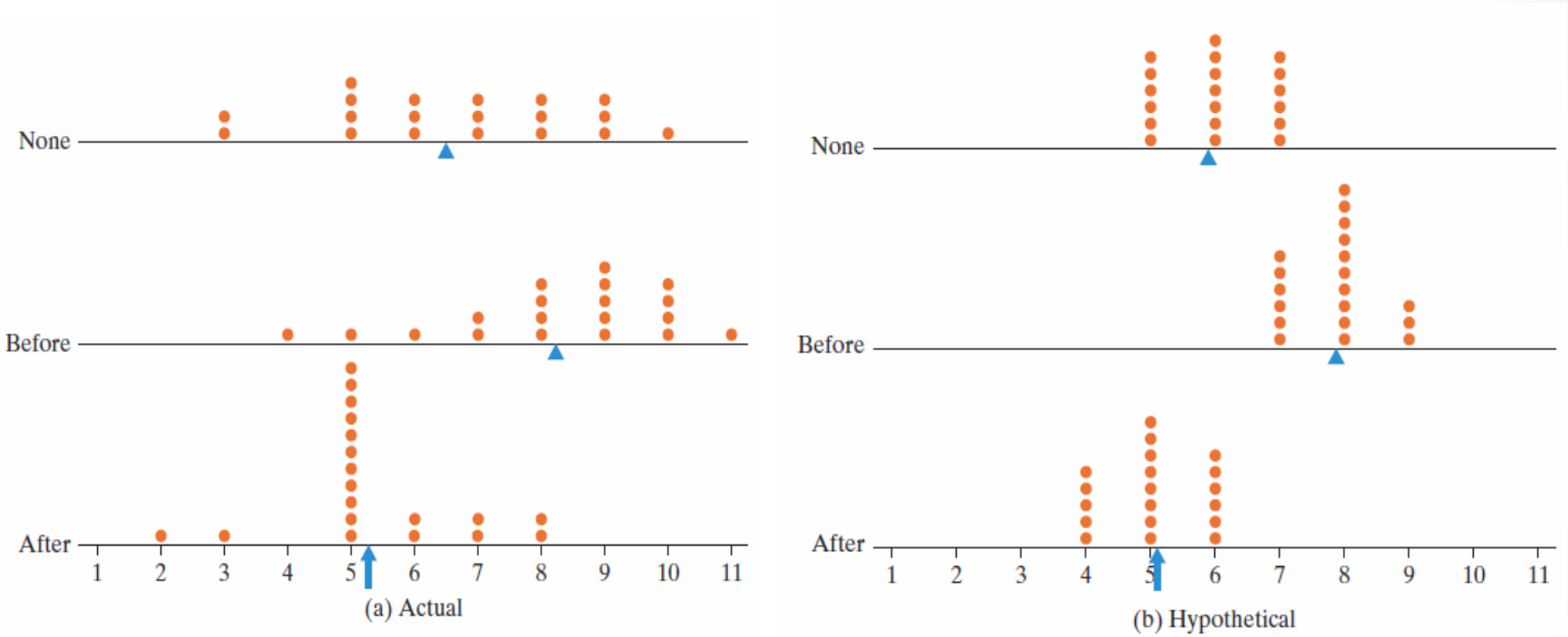


Summary Statistics:

	n	Mean	SD
None	19	6.63	2.01
Before	19	8.26	1.82
After	19	5.37	1.46
Pooled	57	6.75	1.78

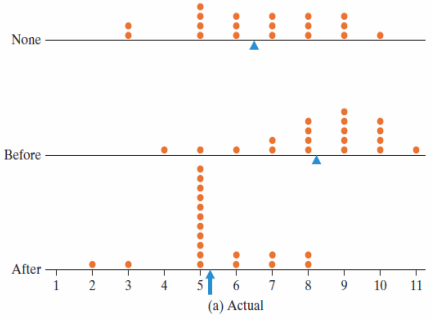
Observed MAD = 1.930

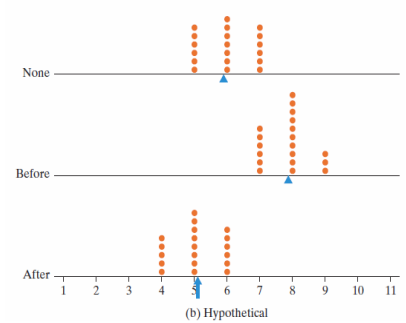
The difference in means matters and so does the individual group's variation



Original recall data are on the left, and hypothetical (fake) recall data are on the right. Variation **between** groups is the same, but variation **within** groups is different. How will this affect the F test statistic?

The difference in means matters and so does the individual group's variation





$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

- The variability between the groups hasn't changed.
- The variability within the groups is smaller on the new data.
- This makes the denominator smaller making the F statistic larger.
- A larger F statistic shows stronger evidence of a difference.
- This should make intuitive sense as well.

Hypotheses

- Null: All three of the long-run mean recall scores for students under the different conditions are the same. (No association)

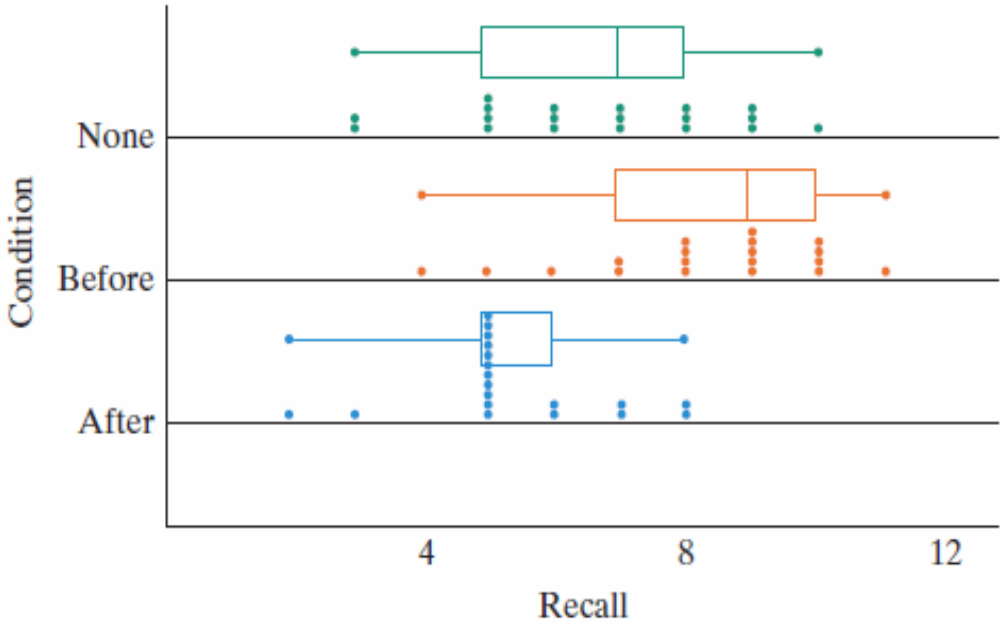
- Alternative: At least one of the long-run mean recall scores for students under the different conditions is different. (Association)

Validity Conditions

- Just as with the simulation-based method, we are assuming we have independent observations.
- Two extra conditions must be met to use traditional ANOVA:
  - Normality: If sample sizes are small within each group, data shouldn't be very skewed. If it is, use simulation approach. (Sample sizes of at least 30 is a good guideline for being considered large.)
  - Equal variation: Largest standard deviation should be no more than twice the value of the smallest.

Validity Conditions

- Are these conditions met for our recall data?



Summary Statistics:

	n	Mean	SD
None	19	6.63	2.01
Before	19	8.26	1.82
After	19	5.37	1.46
Pooled	57	6.75	1.78

Observed MAD = 1.930

ANOVA Output

- This is the kind of output you would see in most other statistics packages when doing ANOVA.
- The variability between the groups is measured by the mean square treatment (40.02).
- The variability within the groups is measured by the mean square error (3.16).
- The F statistic is 40.02/3.16 = 12.67.

Source	df	SS	MS	F	p-value
Treatment	2	80.04	40.02	12.67	0.0000
Error	54	170.53	3.16		
Total	56	250.56			

Conclusion

- Since we have a small p-value we have strong evidence against the null and can conclude at least one of the long-run mean recall scores is different.
- If we were to make separate 95% confidence intervals,
  - After - Before: (-4.05, -1.74)*
  - After - None: (-2.42, -0.11)*
  - Before - None: (0.4756, 2.7875)*
- We can see that each is significant so, with 95% confidence,
  - $\mu_{\text{picture after}} \neq \mu_{\text{picture before}}$
  - $\mu_{\text{pictureafter}} \neq \mu_{\text{no picture}}$
  - $\mu_{\text{picture before}} \neq \mu_{\text{no picture}}$

Strength of Evidence

- As sample size increases, strength of evidence increases.
- As the means move farther apart, strength of evidence increases. (This is the variability between groups.)
- As the standard deviations increase, strength of evidence decreases. (This is the variability within groups.)
- These are all exactly the same as when we compared two means.

Learning Objectives for Section 9.2

- Recognize that larger values of the F statistic mean more evidence against the null hypothesis.
- Identify whether or not an ANOVA (F) test meets appropriate validity conditions.
- Know how to interpret the results of an ANOVA.