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

1. Common problems in regression, continued.
2. Testing significance of correlation or slope.
3. ANOVA and F-test, ch9.

Read ch9.

http://www.stat.ucla.edu/~frederic/13/W23 .

The final is Fri in class and will be on ch 1-7, 10, and at most 1 question on ch9.
Bring a PENCIL or pen and CALCULATOR and any books or notes you want.
No computers.

If you cannot take it because of Covid or other health reasons,
then email me to arrange a time to take an oral, in-person, 10-15 minute final exam in my office.
You can alternatively get an incomplete in the course and take the Spring stat 13 final, but don't come if you have covid.

Common problems with regression.

- c. Curvature.

The best fitting line might fit poorly. Port et al. (2005).

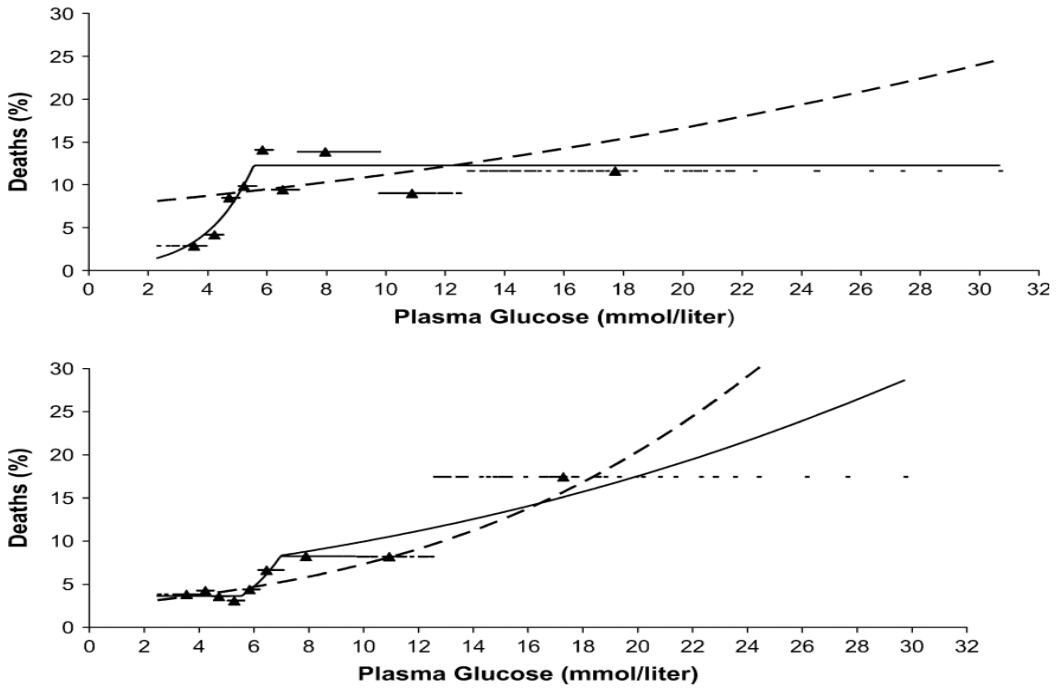
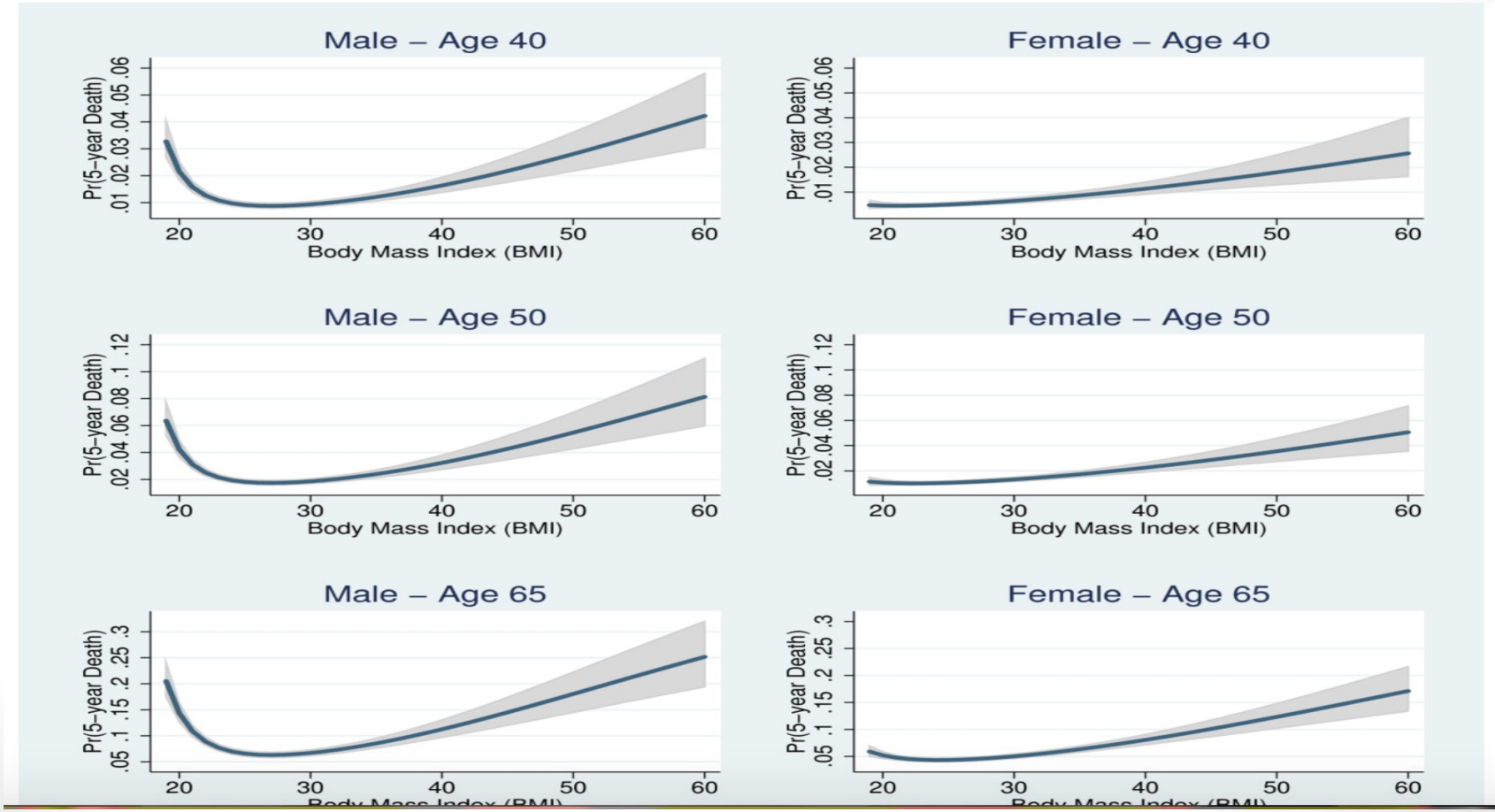


FIGURE 4. Adjusted 2-year rates of death from all causes for men (upper panel) and women (lower panel) separately, by glucose level, predicted by three models, Framingham Heart Study, 1948–1978. Linear model (dashed curve); optimal spline models (solid curve). The horizontal dashed

Common problems with regression.

- c. Curvature.

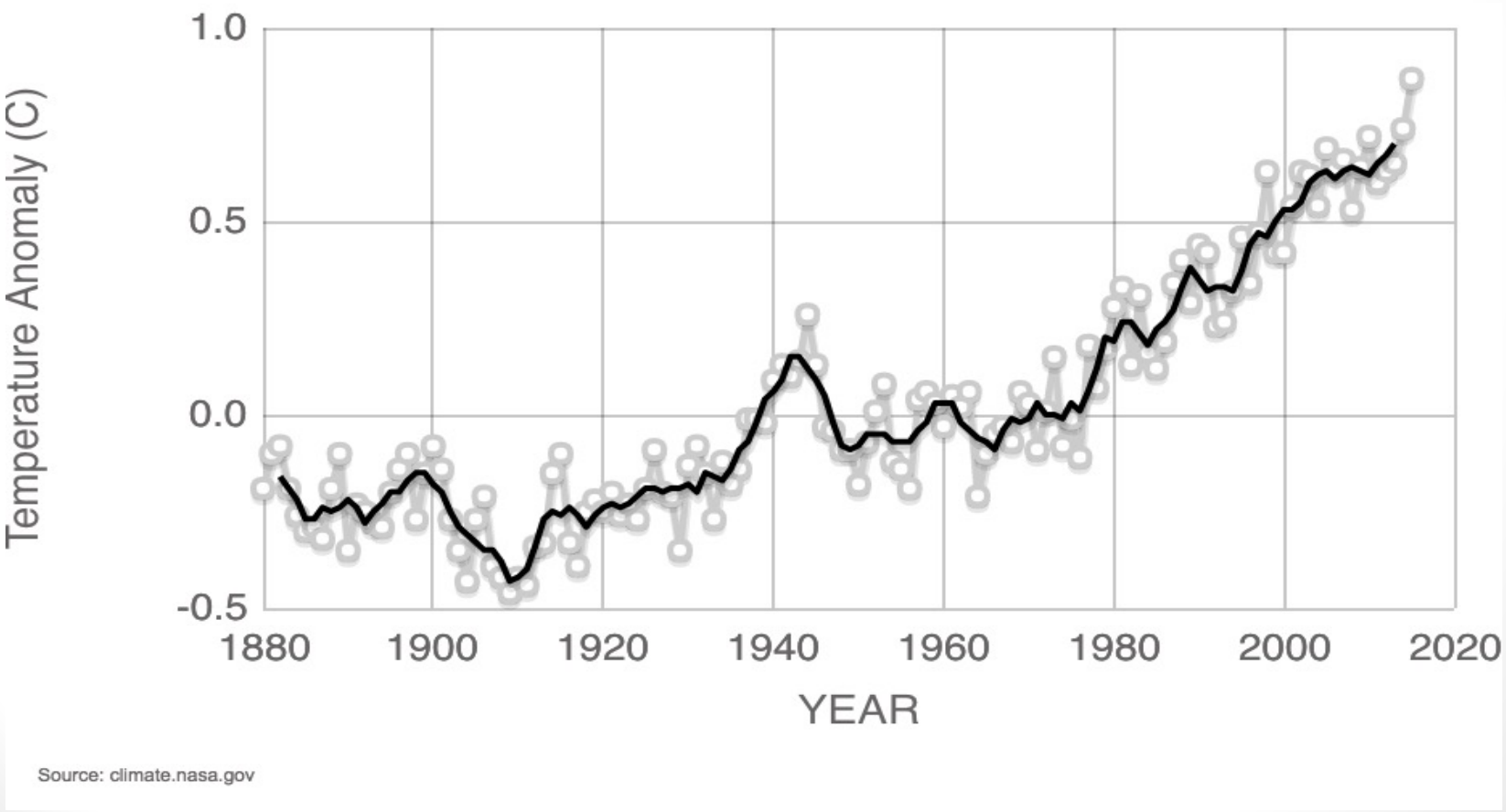
The best fitting line might fit poorly. Wong et al. (2011).



Common problems with regression.

- d. Statistical significance.

Could the observed correlation just be due to chance alone?



2. Inference for the Regression Slope: Theory-Based Approach

Section 10.5

Do students who spend more time in non-academic activities tend to have lower GPAs?

Example 10.4

Do students who spend more time in non-academic activities tend to have lower GPAs?

- The subjects were 34 undergraduate students from the University of Minnesota.
- They were asked questions about how much time they spent in activities like work, watching TV, exercising, non-academic computer use, etc. as well as what their current GPA was.
- We are going to test to see if there is a significant negative association between the number of hours per week spent on nonacademic activities and GPA.

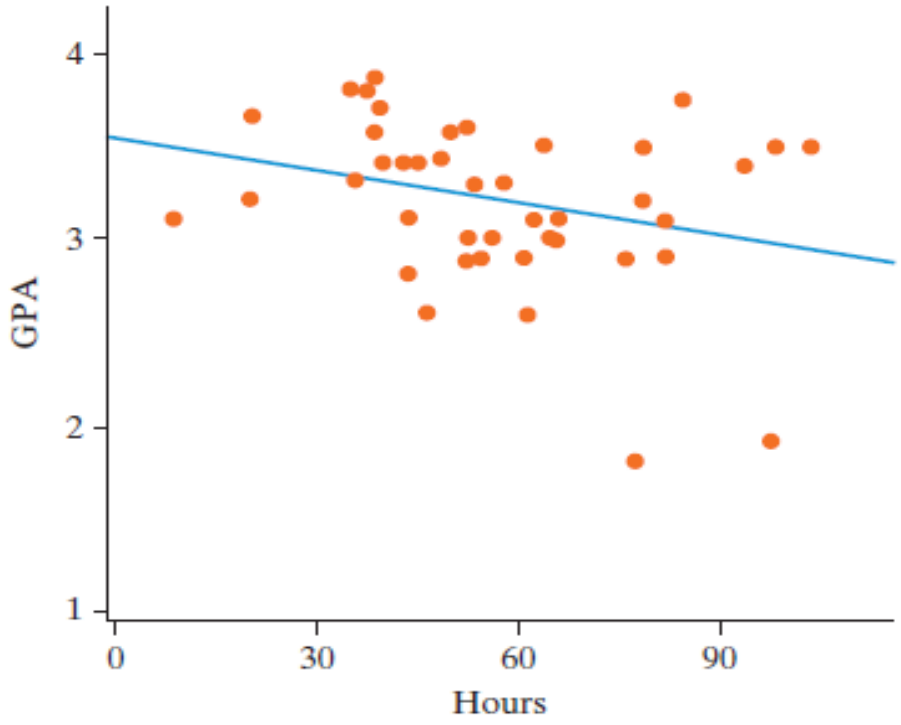
Hypotheses

- Null Hypothesis: There is no association between the number of hours students spend on nonacademic activities and student GPA in the population.

- Alternative Hypothesis: There is a negative association between the number of hours students spend on nonacademic activities and student GPA in the population.

Descriptive Statistics

- $\widehat{\text{GPA}} = 3.60 - 0.0059(\text{nonacademic hours})$.
- Is the slope significantly different from 0?



Shuffle to Develop Null Distribution

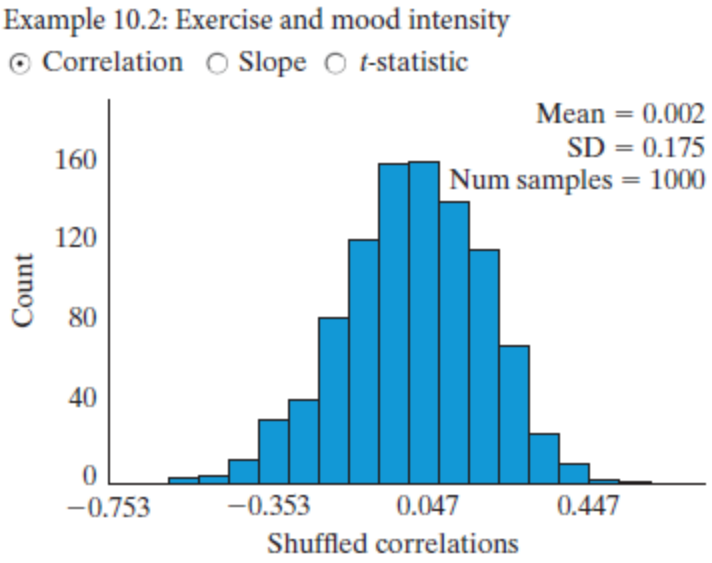
- We are going to shuffle just as we did with correlation to develop a null distribution.
- The only difference is that we will be calculating the slope each time and using that as our statistic.
- **a test of association based on slope is equivalent to a test of association based on a correlation coefficient.**

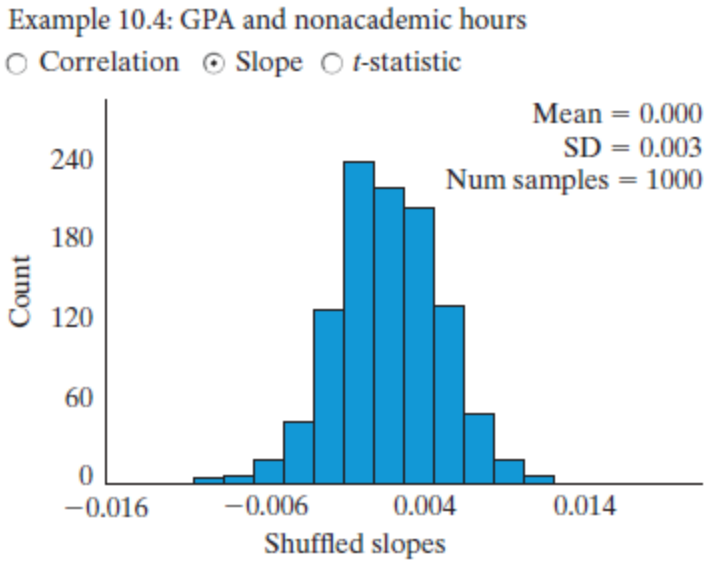
Beta vs Rho

- Testing the slope of the regression line is equivalent to testing the correlation (same p-value, but obviously different confidence intervals since the statistics are different)
- Hence these hypotheses are equivalent.
  - H_o: $\beta = 0$ H_a: $\beta < 0$ (Slope)
  - H_o: $\rho = 0$ H_a: $\rho < 0$ (Correlation)
- Sample slope (b) Population (β: beta)
- Sample correlation (r) Population (ρ: rho)
- When we do the theory based test, we will be using the *t*-statistic which can be calculated from either the slope or correlation.

Introduction

- Our null distributions are again bell-shaped and centered at 0 (for either correlation or slope as our statistic).





The book on p549 finds a p value of 3.3% by simulation.

Validity Conditions

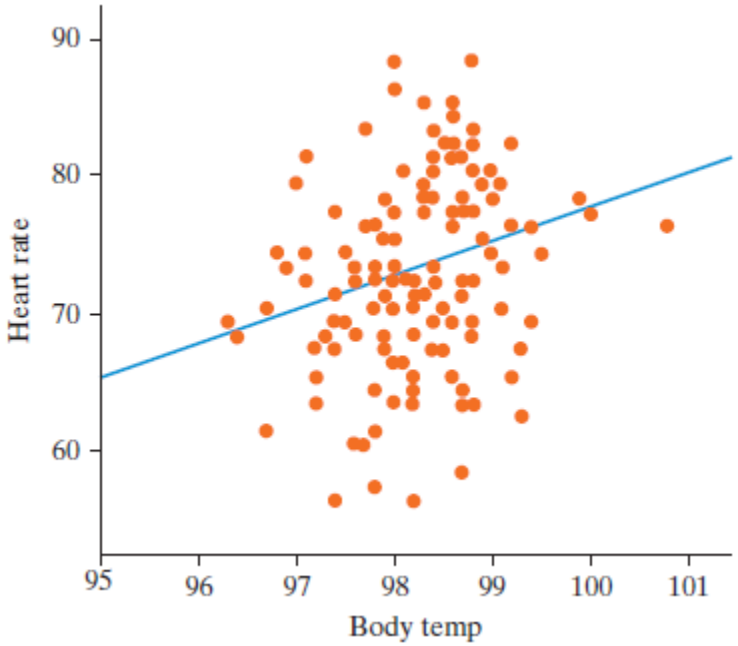
- Under the usual conditions: relationship is linear, observations are iid, both variables are normally distributed, and data are homoskedastic, theory-based inference for correlation or slope of the regression line uses the *t*-distribution.
- We could use simulations or the theory-based methods for the slope of the regression line.
- We would get exactly the same p-value if we used correlation as our statistic.

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

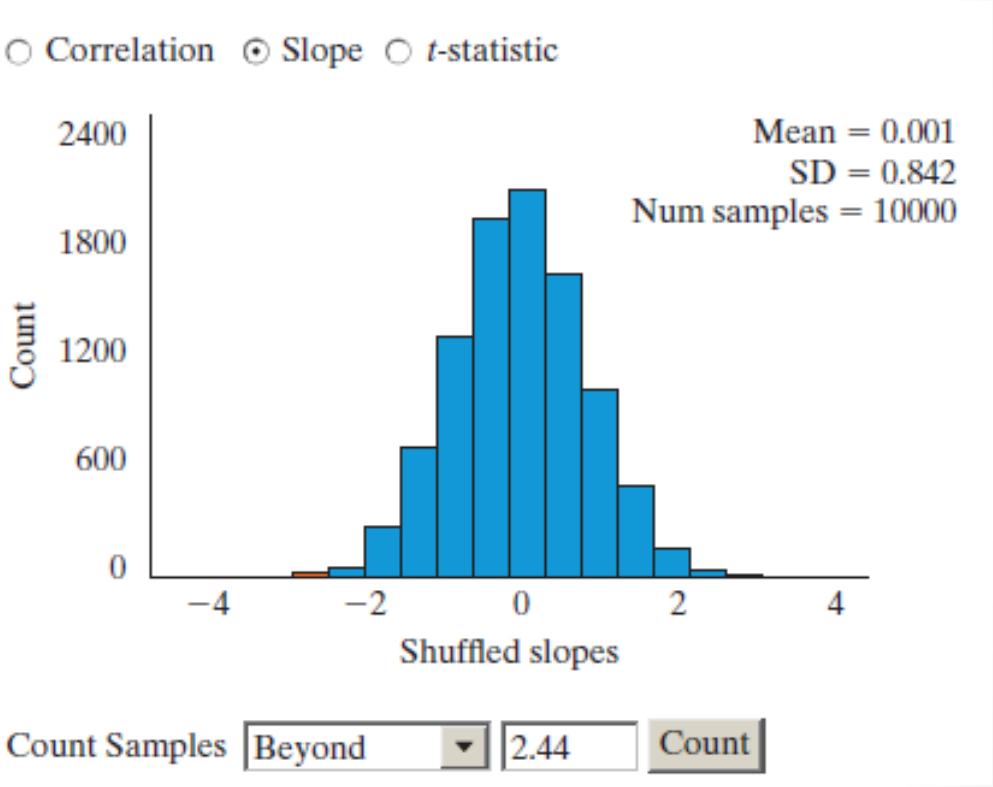


Heart Rate and Body Temp

- 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

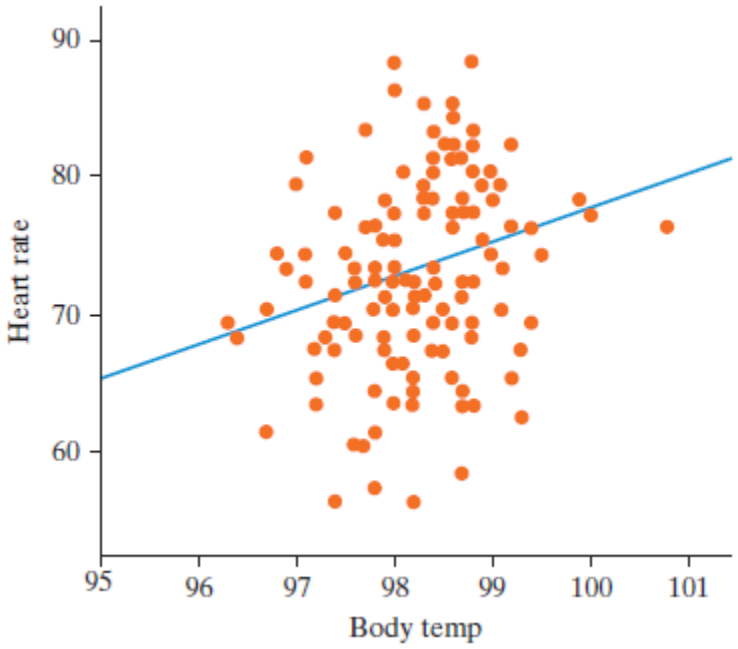
- We can also approximate a 95% confidence interval

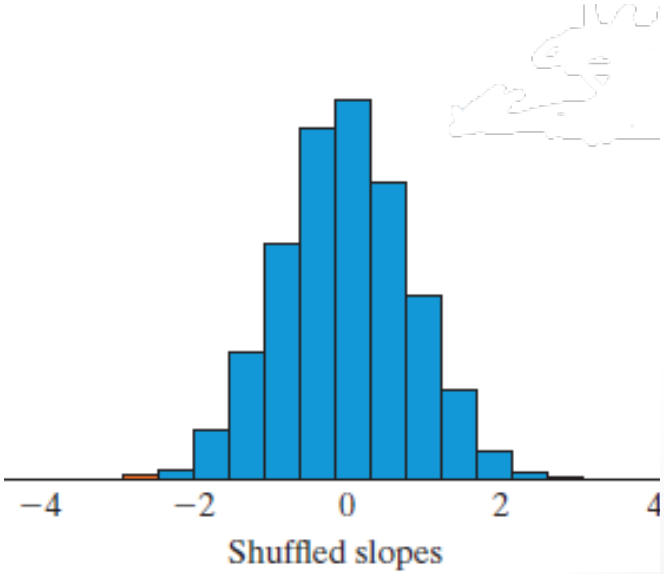
  observed statistic $\pm$ multiplier $\times$ SE

  $2.44 \pm 1.96 \times 0.842 =$ 0.790 to 4.09
- When both variables are normally distributed (scatterplot is elliptical), use the t-multiplier instead of 1.96, but when n is large it makes very little difference.
- This means we are 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 of the slopes 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 0.36%. 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.

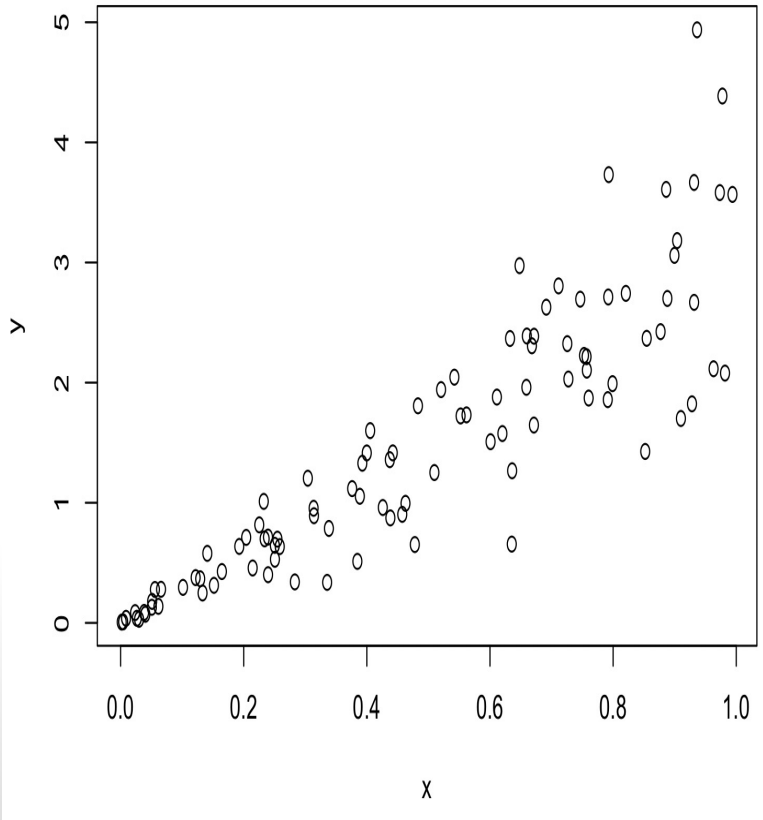
1. The scatterplot should follow a linear trend.
2. The observations should be iid.
3. For the t-test, both variables should be normal.

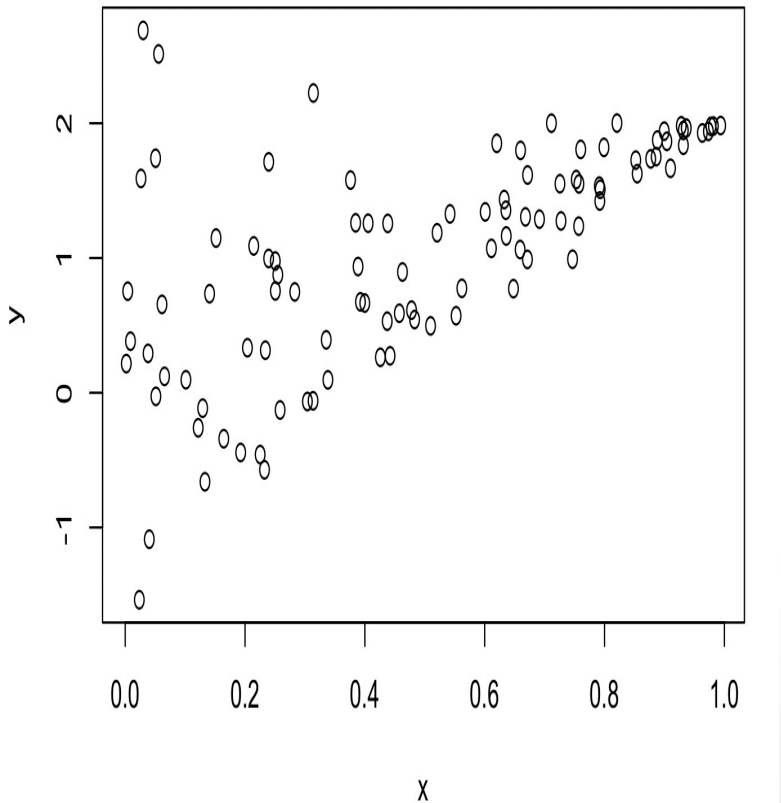
In particular, there should be approximately the same number of points above and below the regression line (symmetry).

4. 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.



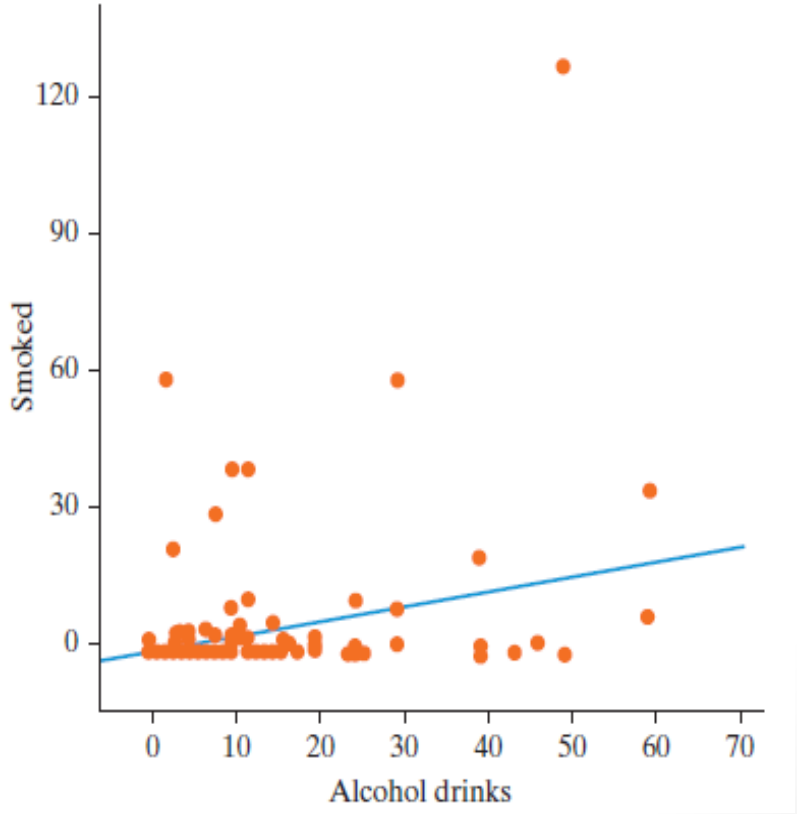


Smoking and Drinking

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

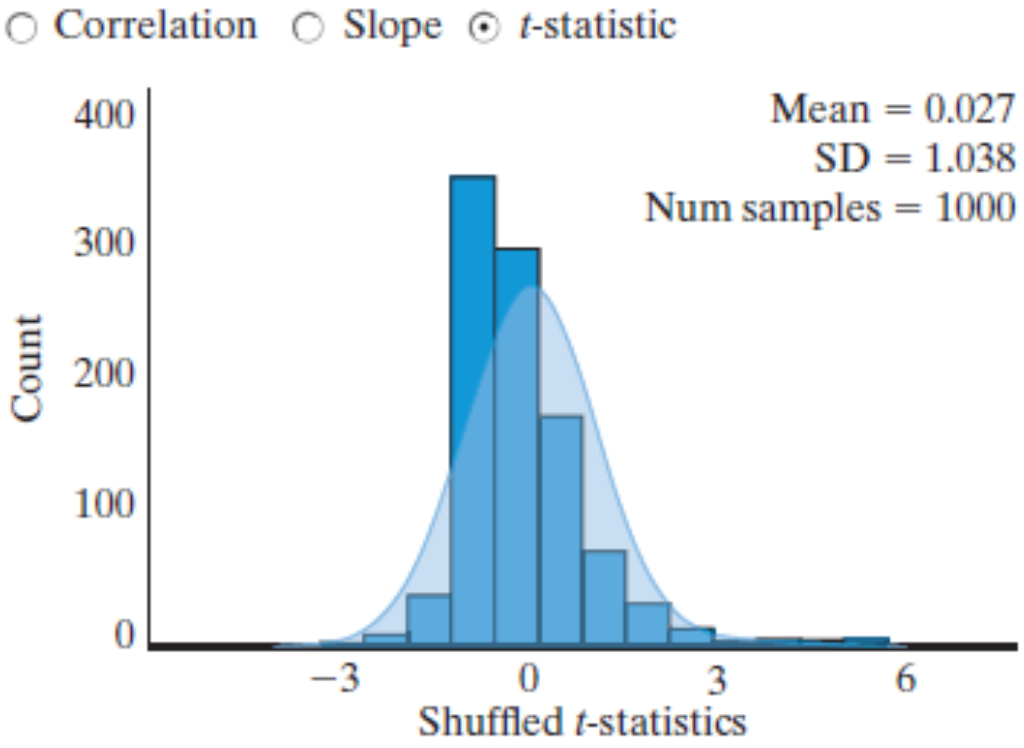
The dot at (0,0) represents 524 students

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.



Validity Conditions

- What do you do when validity conditions aren't met for theory-based inference?
  - Use the simulated-based approach.
- Another strategy is to "transform" the data on a different scale so conditions are met.
  - The logarithmic scale is common.
- One can also fit a different curve, not necessarily a line.

3. ANOVA and F-test.

Section 9.2

ANOVA

- ANOVA stands for ANalysis Of VAriance.
- Useful when comparing more than 2 means.
- If I have 2 means to compare, I just look at their difference to measure how far apart they are.
- Suppose I wanted to compare three means. I have the mean for group A, the mean for group B, and the mean for group C.

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. $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}}$

Recalling Ambiguous Prose

Example 9.2

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

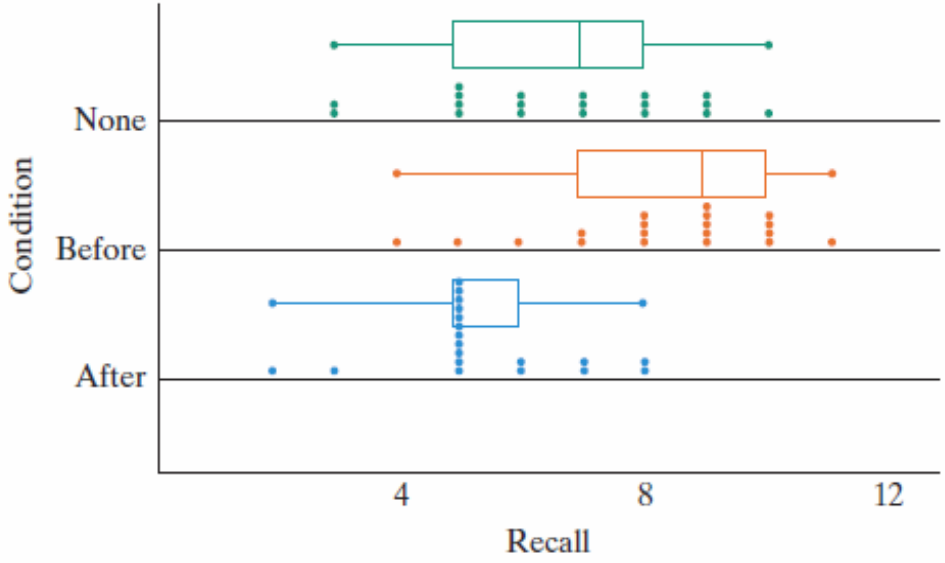
- **Null:** All three of the long term mean comprehension scores are the same.

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

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

Recall Score

- 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

Validity Conditions

- Just as with the simulation-based method, we are assuming we have independent groups.
- 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.)
  - Equal variation: Largest standard deviation should be no more than twice the value of the smallest.

ANOVA Output

- This is the kind of output you would see in most 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 significantly different from the others.