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

1. Theory-based inference for regression slope.
2. Regression formulas.
3. ANOVA and F-test, ch9.

NO LECTURE OR Office Hour TUESDAY JUNE 2!!!

Read chapter 9.

The course website is http://www.stat.ucla.edu/~frederic/13/S26 .
HW4 is due Fri, Jun5, 1159pm. 10.1.8, 10.3.14, 10.3.21, and 10.4.11.
These problems are on the next few slides.

The final is Tue Jun9, 1130am-230pm, FOWLER A103b, and
will be on ch 1-7, 10, and 1 question on ch9.
Bring a PENCIL or pen and any books or notes you want.
No electronic devices allowed, not even a calculator.

If you cannot take it because of an emergency or other health reason,
then you will get an incomplete in the course and need to arrange to take the Summer or Fall Stat 13 final.

10.1.8.

10.1.8 Which of the following statements is correct?

A. Changing the units of measurements of the explanatory or response variable does not change the value of the correlation.

B. A negative value for the correlation indicates that there is no relationship between the two variables.

C. The correlation has the same units (e.g., feet or minutes) as the explanatory variable.

D. Correlation between y and x has the same number but opposite sign as the correlation between x and y.

10.3.14.

10.3.12 Reconsider the previous five exercises and the **Legos** data file. The last product listed in the data file has 415 pieces and a price of $49.99.

a. Determine the predicted price for such a product.

b. Determine the residual value for this product.

c. Interpret what this residual value means.

d. Does the product fall above or below the least squares line in the graph? Explain how you can tell, based on its residual value.

10.3.13 Reconsider the previous six exercises and the **Legos** data file. This is very unrealistic, but suppose that one of the products were to be offered at a price of $0.

a. Would you expect this change to affect the least squares line very much? Explain.

b. For which one product would you expect this change to have the greatest impact on the least squares line? Explain how you choose this product.

c. Change the price to $0 for the product that you identified in part (b). Report the (new) equation of the least squares line and the (new) value of r^2. Have these values changed considerably?

Crickets

10.3.14 Consider the following two scatterplots based on data gathered in a study of 30 crickets, with temperature measured in degrees Fahrenheit and chirp frequency measured in chirps per minute.

a. If the goal is to predict temperature based on a cricket's chirps per minute, which is the appropriate scatterplot to examine—the one on the left or the one on the right? Explain briefly.

One of the following is the correct equation of the least squares line for predicting temperature from chirps per minute:

A. predicted temperature = 35.78 + 0.25 chirps per minute

B. predicted temperature = −131.23 + 3.81 chirps per minute

C. predicted temperature = 83.54 − 0.25 chirps per minute

b. Which is the correct equation? Circle your answer and explain briefly.

c. Use the correct equation to predict the temperature when the cricket is chirping at 100 chirps per minute.

d. Interpret the value of the slope coefficient, in this context, for whichever equation you think is the correct one.

Cat jumping*

10.3.15 Harris and Steudel (2002) studied factors that might be associated with the jumping performance of domestic cats. They studied 18 cats, using takeoff velocity (in centimeters per second) as the response variable. They used body mass (in grams), hind limb length (in centimeters), muscle mass (in grams), and percent body fat in addition to sex as potential explanatory variables. The data can be found in the **CatJumping** data file. A scatterplot of takeoff velocity vs. body mass is shown in the figure for Exercise 10.3.15.

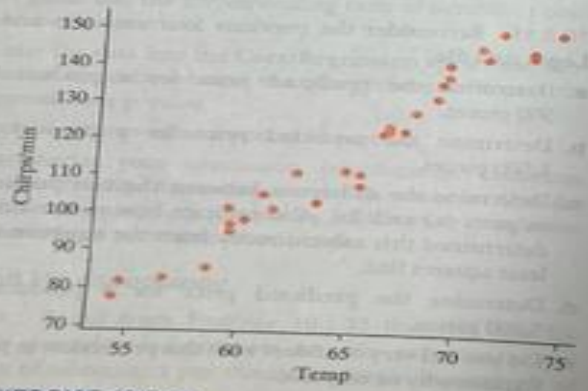
a. Describe the association between these variables.

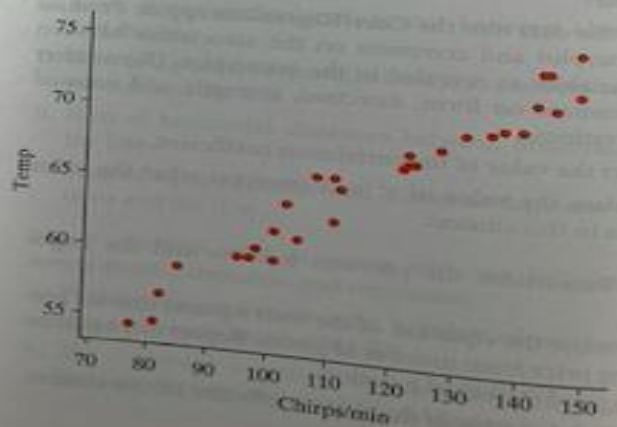
b. Use the **Corr/Regression** applet to determine the equation of the least squares line for predicting a cat's takeoff velocity from its mass.

c. Interpret the value of the slope coefficient in this context.

d. Interpret the value of the intercept coefficient. Is this a context in which the intercept coefficient is meaningful?

e. Determine the proportion of variability in takeoff velocity that is explained by the least squares line with mass.



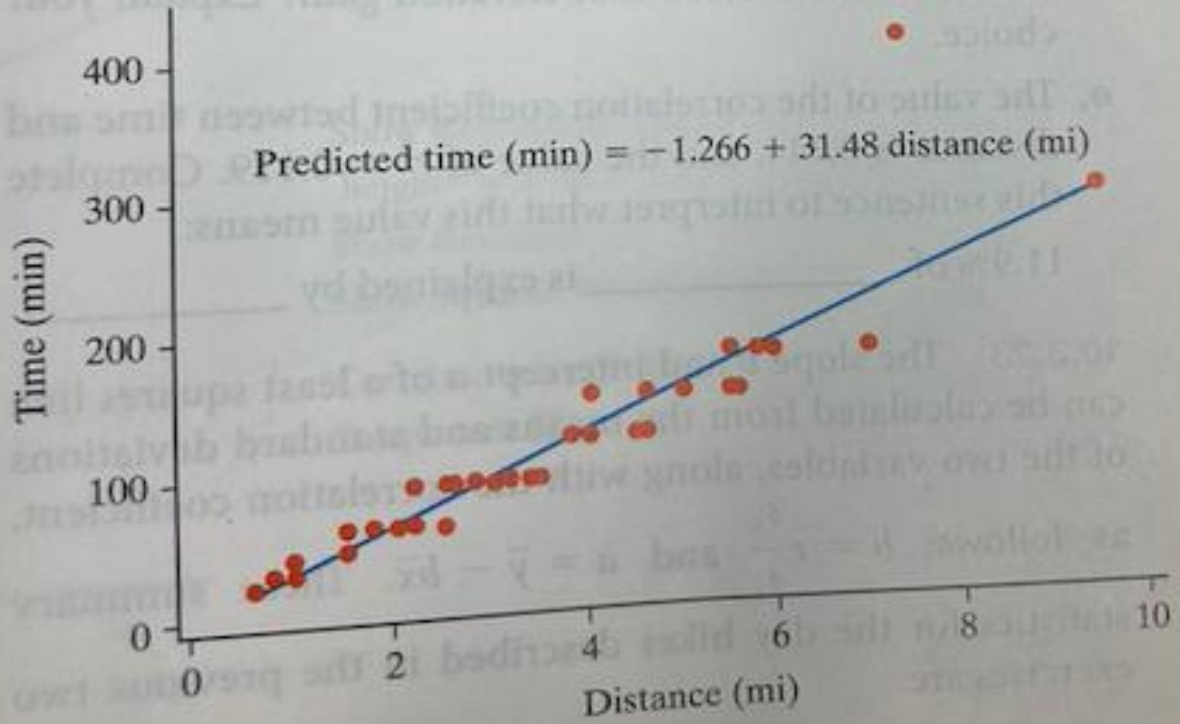


EXERCISE 10.3.14

10.3.21.

Day hikes

10.3.21 The book *Day Hikes in San Luis Obispo County* lists information about 72 hikes, including the distance of the hike (in miles), the elevation gain of the hike (in feet), and the time that the hike is expected to take (in minutes). Consider the scatterplot below, with least squares regression line superimposed:



10.3.21.

a. Report the value of the slope coefficient for predicting time from distance.

b. Write a sentence interpreting the value of the slope coefficient for predicting time from distance.

c. Use the line to predict how long a 4-mile hike will take.

d. Would you feel more comfortable using the line predict the time for a 4-mile hike or for a 12-mile hike? Explain your choice.

e. The value of the correlation coefficient between time and distance is 0.916, and the value of $r^2 = 0.839$. Complete this sentence to interpret what this value means:

83.9% of _____ is explained by _____.

10.3.22 Reconsider the previous exercise. The following

10.4.11.

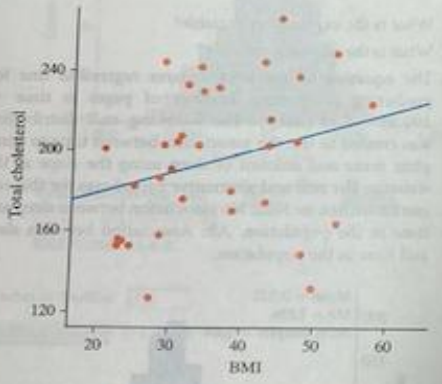
10.4.10 Reconsider the previous exercise about the amount of sleep (in hours) obtained in the previous night and time to complete a paper and pencil maze (in seconds). The equation of the least squares regression line for predicting price from number of pages is time = 190.33 − 7.76 (sleep).

a. Interpret what the slope coefficient means in the context of sleep and time to complete the maze.

b. Interpret the intercept. Is this an example of extrapolation? Why or why not?

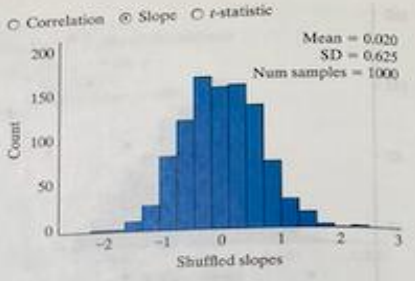
Weight loss and protein

10.4.11 In a study to see if there was an association between weight loss and the amount of a certain protein in a person's body fat, the researchers measured a number of different attributes in their 39 subjects at the beginning of the study. The article reported, "These subjects were clinically and ethnically heterogeneous." Two of the variables they measured were body mass index (BMI) and total cholesterol. The results are shown in the scatterplot along with the regression line.



a. What are the observational units in the study?

b. The equation of the least squares regression line for predicting total cholesterol from BMI is cholesterol = 162.56 − 0.9658 (BMI). The following null distribution was created to test the association between people's total cholesterol number and their BMI using the slope as the statistic. The null and alternative hypotheses for this test can be written as: Null: No association between cholesterol and BMI in the population. Alt: Association between cholesterol and BMI in the population.



i. Based on information shown in the null distribution, how many standard deviations is our observed statistic below the mean of the null distribution? (That is, what is the standardized statistic?)

ii. Based on your standardized statistic, do you have strong evidence of an association between a people's total cholesterol and their BMI? Explain.

10.4.12 Reconsider the previous exercise about the cholesterol and BMI. The equation of the least squares regression line obtained was cholesterol = 162.56 − 0.9658 (BMI).

a. Interpret what the slope coefficient means in the context of cholesterol and BMI.

b. Interpret the intercept. Is this an example of extrapolation? Why or why not?

Honda Civic prices*

10.4.13 The data in the file **UsedHondaCivics** come from a sample of used Honda Civics listed for sale online in July 2006. The variables recorded are the car's age (calculated as 2006 minus year of manufacture) and price. Consider conducting a simulation analysis to test whether the sample data provide strong evidence of an association between a car's price and age in the population in terms of the population slope.

a. State the appropriate null and alternative hypotheses.

b. Conduct a simulation analysis with 1,000 repetitions. Describe how to find your p-value from your simulation results and report this p-value.

c. Summarize your conclusion from this simulation analysis. Also describe the reasoning process by which your conclusion follows from your simulation results.

10.4.14 Reconsider the previous exercise on prices of Honda Civics.

a. Find the regression equation that predicts the price of the car given its age.

b. Interpret the slope and intercept of the regression line.

1. 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.

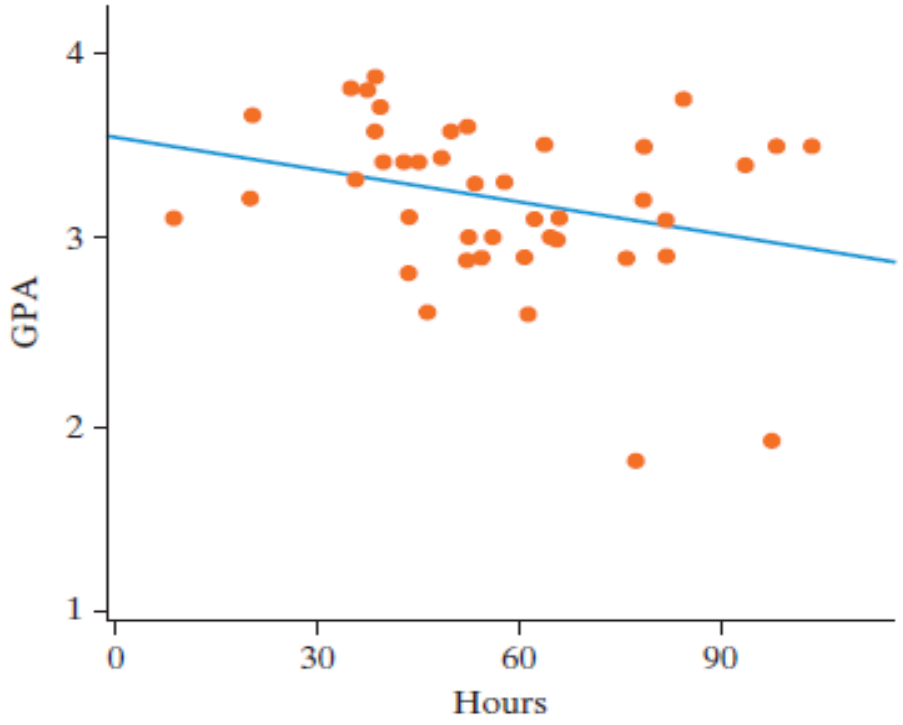
Hypotheses

- 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

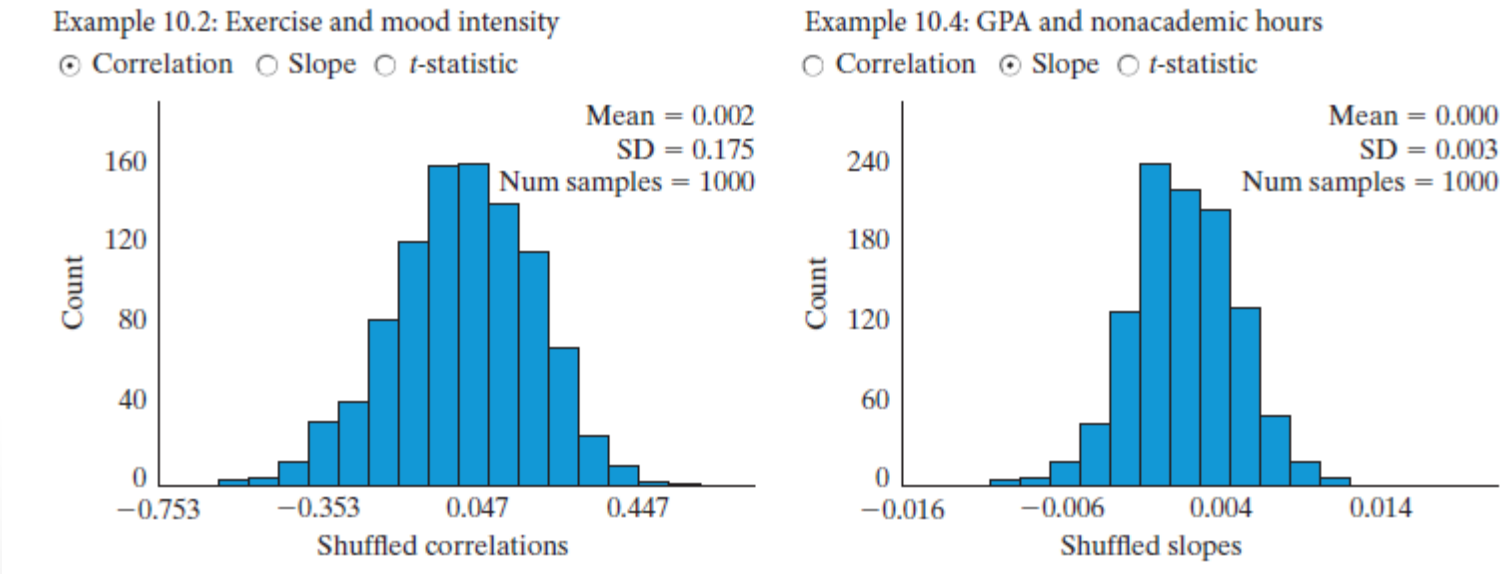
- We are going to shuffle just as we did with correlation to develop a null distribution.
- The only difference is that we will be calculating the slope each time and using that as our statistic.
- **a test of association based on slope is equivalent to a test of association based on a correlation coefficient.**

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

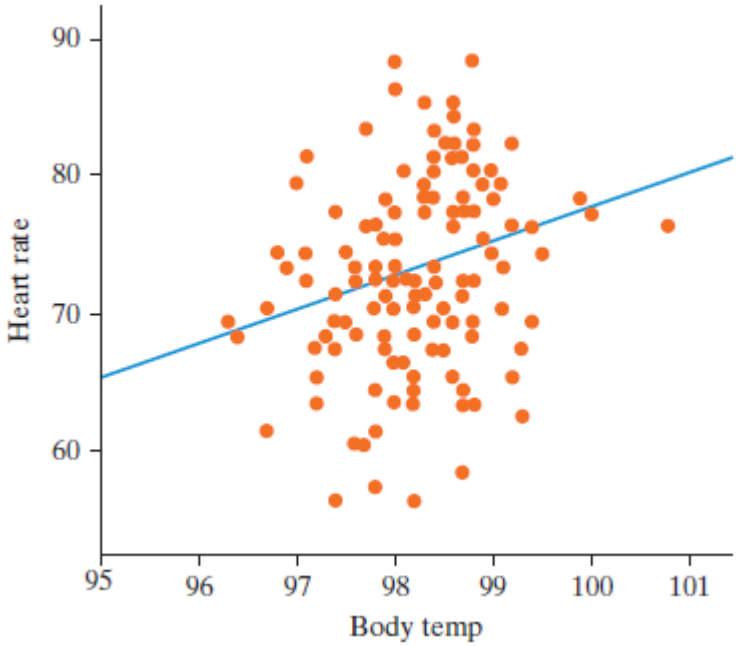
- Under the usual conditions: relationship is linear, observations are iid, both populations are normally distributed, and data are homoskedastic, theory-based inference for correlation or slope of the regression line uses the *t*-distribution.
- We could use simulations or the theory-based methods for 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

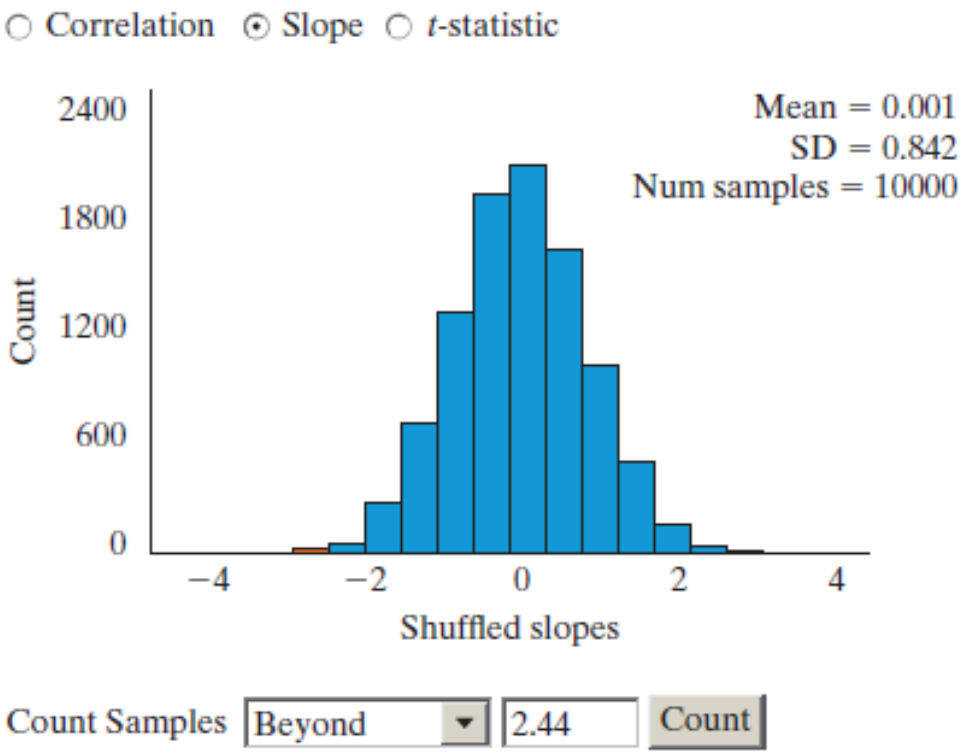


Heart Rate and Body Temp

- 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

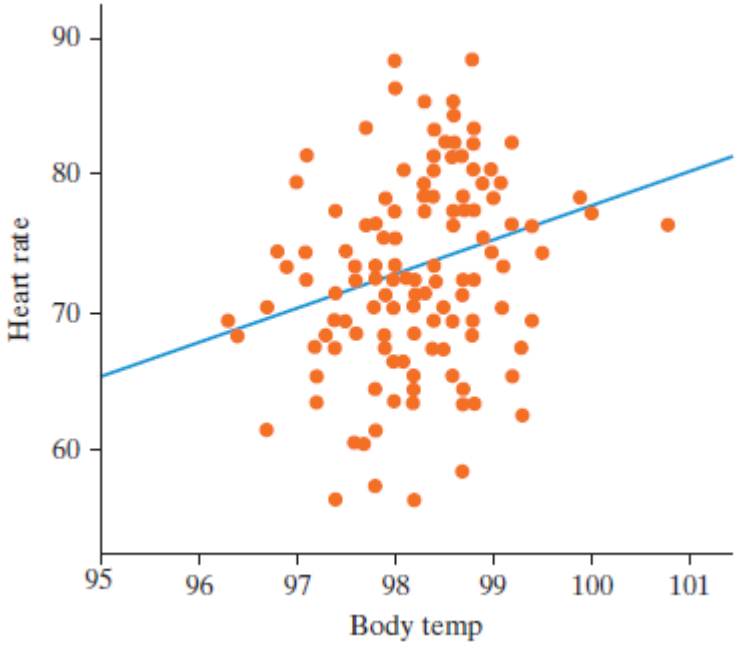
  observed statistic $\pm$ multiplier $\times$ SE

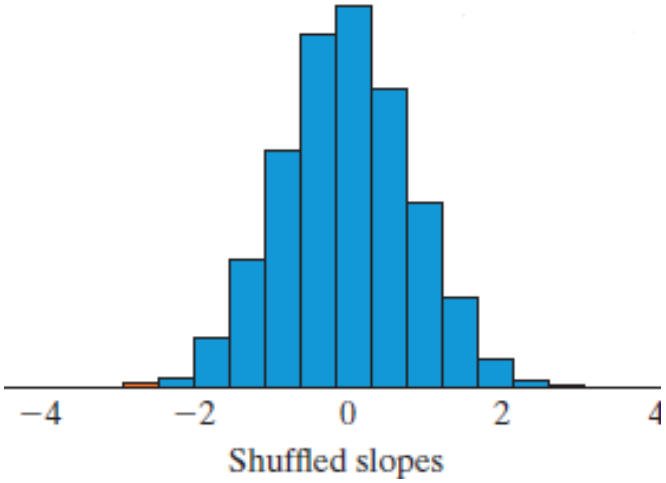
  $2.44 \pm 1.96 \times 0.842 =$ 0.790 to 4.09

- 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.

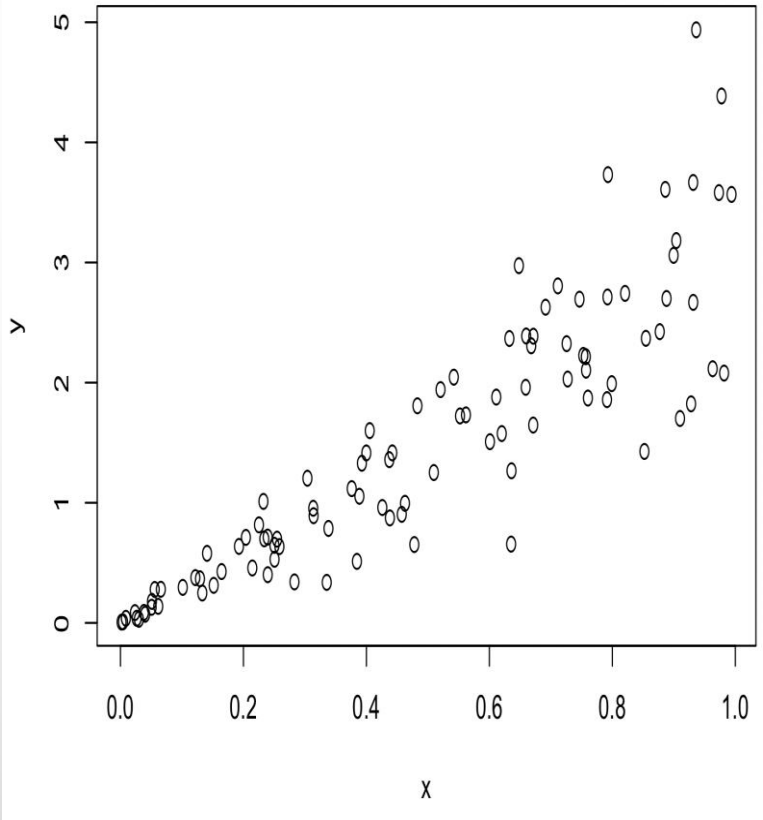
1. The scatterplot should follow a linear trend.
2. The observations should be iid.
3. For the t-test, both variables should be normal.

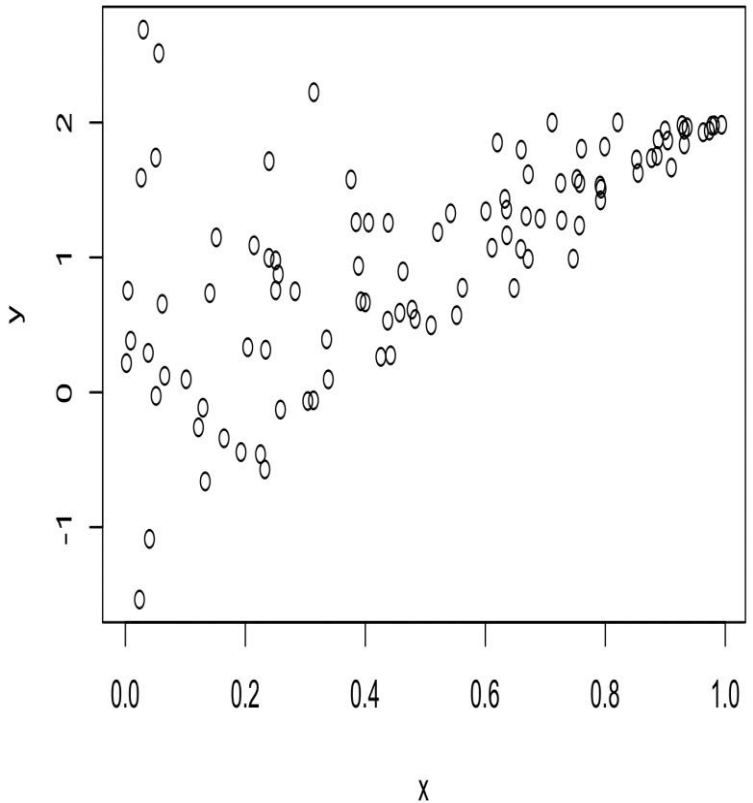
In particular, there should be approximately the same number of points above and below the regression line (symmetry).

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.



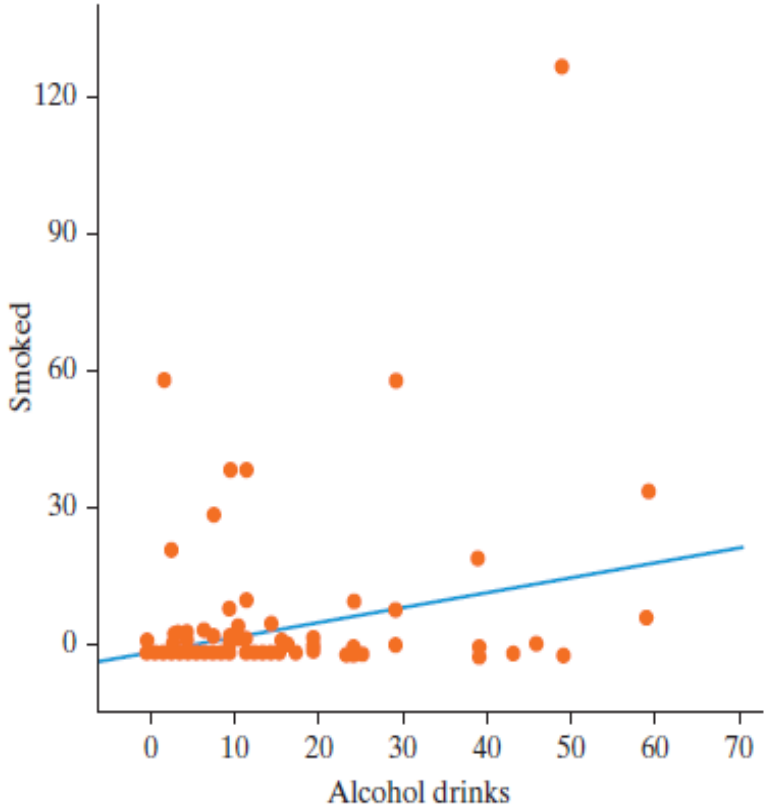


Smoking and Drinking

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

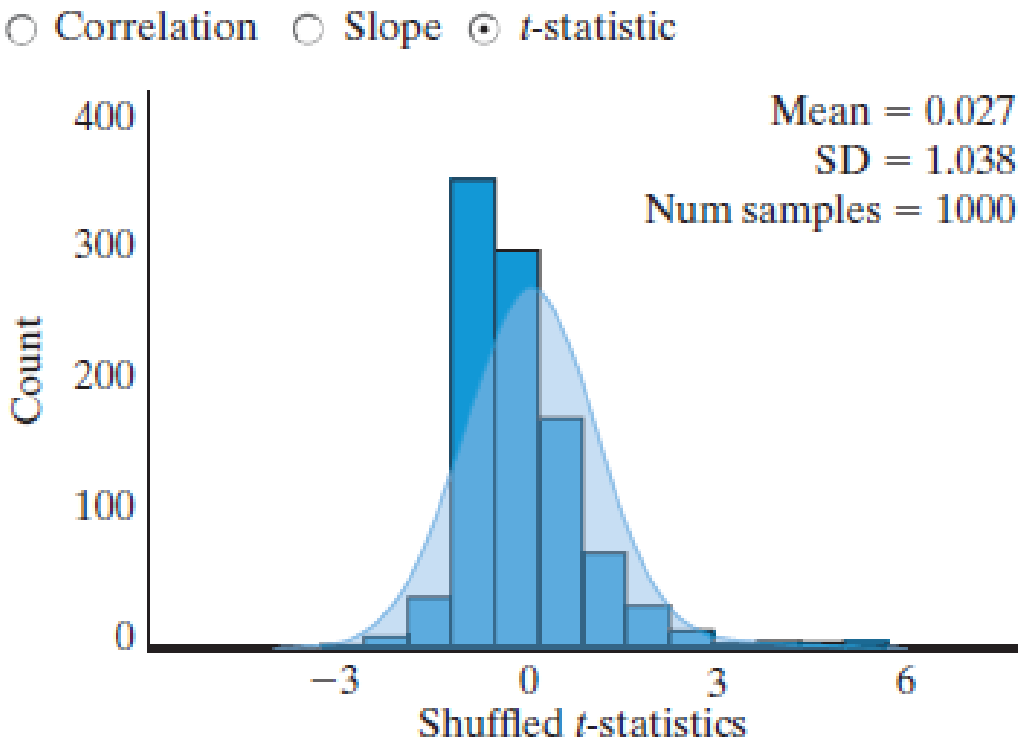
The dot at (0,0) represents 524 students

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.

2. Summary of regression facts.

- Suppose $\hat{y}$ = a + bx is the regression line, i.e. the line with minimum sum of squared residuals.
- Residual = observed y-value minus $\hat{y}$.
- The slope b of the regression line is b = r $\frac{s_y}{s_x}$.
- The intercept a = $\overline{y}$ - b $\overline{x}$.
- The mean of the residuals from regression line is always 0.
- The SD of the residuals is $\sqrt{1-r^2}\ s_y$.

  This is a good estimate of how much the regression predictions will typically be off by.
- The residuals in linear regression always have mean zero.

- When testing the slope or correlation,

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

Tests of slope and correlation are equivalent.

- The SE for r is $\sqrt{\frac{1-r^2}{n-2}}$.
- The SE for b is $\sqrt{\frac{1-r^2}{n-2}}\ \frac{s_y}{s_x}$.

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. $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Recalling Ambiguous Prose

Example 9.2

Comprehension Example

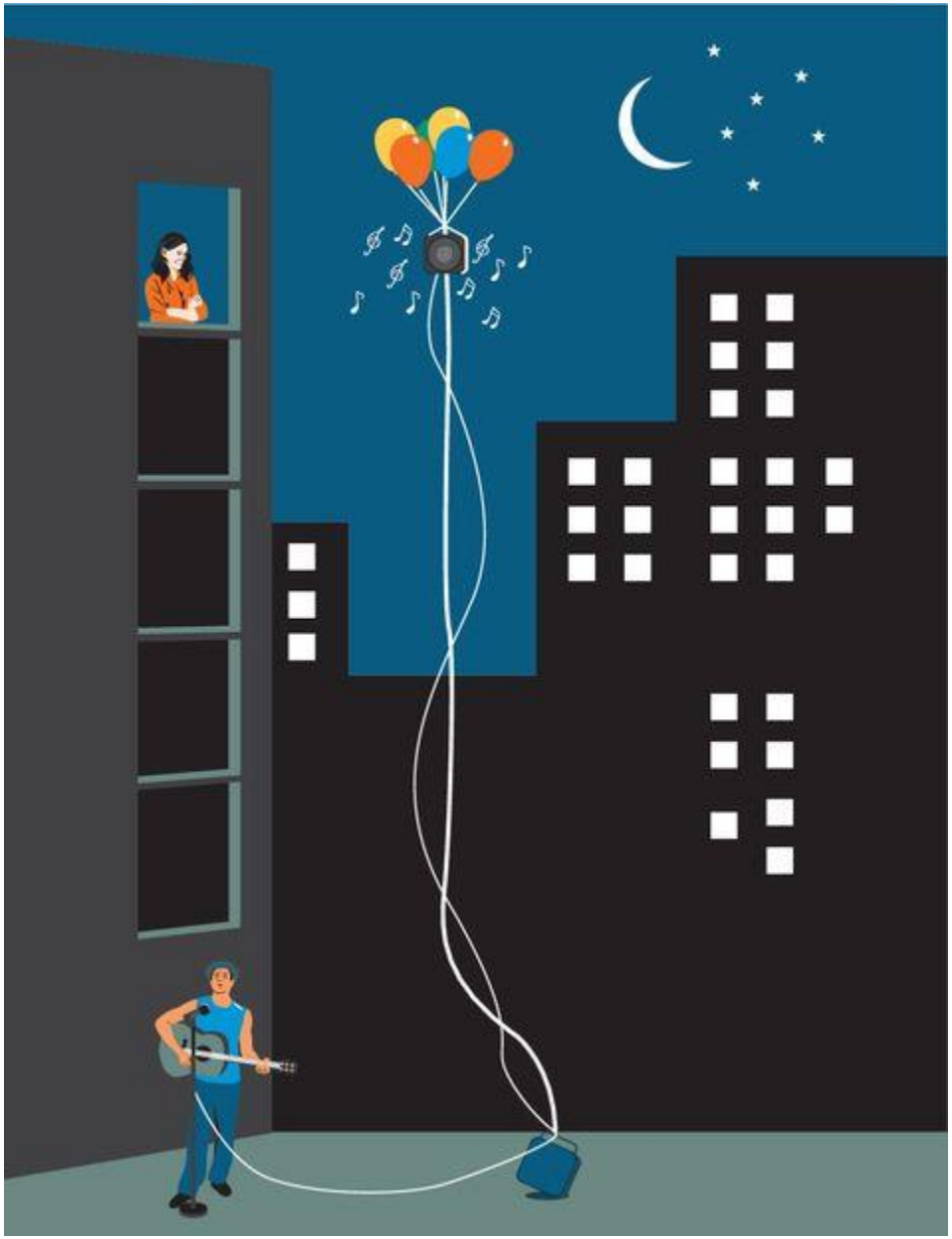
(**Don't** follow along in your book or look ahead on the PowerPoint until after I read you the passage.)

- 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

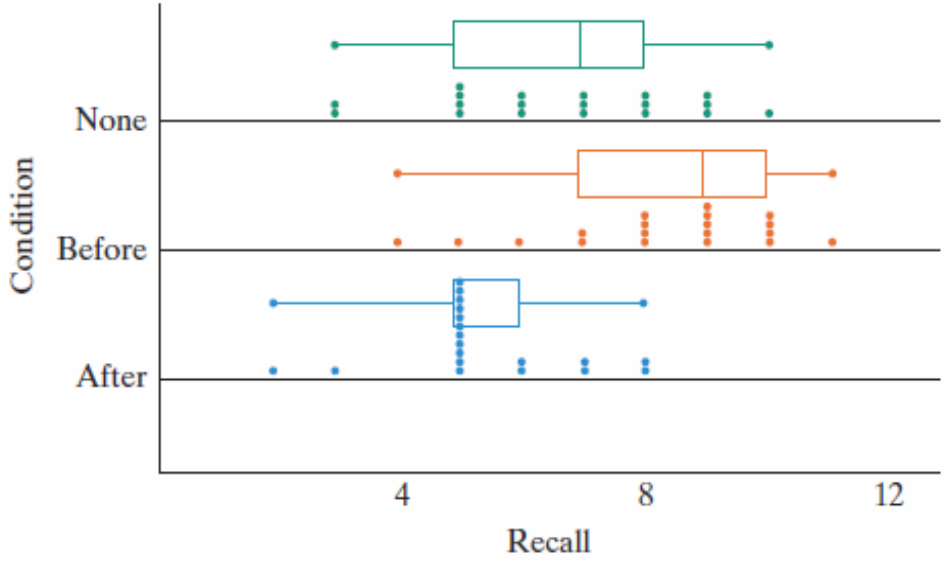
- **Null:** All three of the long term mean comprehension scores are the same.

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

- **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.

Review list.

1. Meaning of SD.
2. Parameters and statistics.
3. Z statistic for proportions.
4. Simulation and meaning of pvalues.
5. SE for proportions.
6. What influences pvalues.
7. CLT and validity conditions for tests.
8. 1-sided and 2-sided tests.
9. Reject the null vs. accept the alternative.
10. Sampling and bias.
11. Significance level.
12. Type I, type II errors, and power.
13. CIs for a proportion.
14. CIs for a mean.
15. Margin of error.
16. Practical significance.
17. Confounding.
18. Observational studies and experiments.
19. Sample size calculations.
20. Random sampling and random assignment.
21. Two proportion CIs and testing.
22. IQR and 5 number summaries.
23. CIs for 2 means and testing.
24. Paired data.
25. Placebo effect, adherer bias, and nonresponse bias.
26. Prediction and causation.
27. Multiple testing and publication bias
28. Regression.
29. Correlation.
30. Calculate & interpret a & b.
31. Goodness of fit for regression.
32. Common regression problems (causation, extrapolation, curvature, heteroskedasticity).
33. ANOVA and F-test.

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

1. What does the correlation of 0.82 imply?

a. 82% of the variation in weight is explained by height.

b. The typical variation in people's heights is 82% as large as the typical variation in their weights.

c. There is strong association between height and weight in this sample.

d. For every inch of increase in one's height, we would predict a 0.82 lb. increase in weight.

e. If a person weighs 100 pounds, then we typically would expect the person to be about 82 inches tall.

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

1. What does the correlation of 0.82 imply?

a. 82% of the variation in weight is explained by height.

b. The typical variation in people's heights is 82% as large as the typical variation in their weights.

c. There is strong association between height and weight in this sample.

d. For every inch of increase in one's height, we would predict a 0.82 lb. increase in weight.

e. If a person weighs 100 pounds, then we typically would expect the person to be about 82 inches tall.

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

2. What is the estimated slope, in lbs/inch, of the regression line for predicting weight from height?

a. 6.56. b. 7.12. c. 8.04. d. 9.92. e. 10.2. f. 11.4.

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

2. What is the estimated slope, in lbs/inch, of the regression line for predicting weight from height?

a. 6.56. b. 7.12. c. 8.04. d. 9.92. e. 10.2. f. 11.4.

$r\, s_y/s_x = .82 \times 40 / 5 = 6.56.$

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

3. How much would a prediction using this regression line typically be off by?

a. 12.7 lbs. b. 13.5 lbs. c. 14.4lbs. d. 20.2 lbs. e. 22.9 lbs.

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

3. How much would a prediction using this regression line typically be off by?

a. 12.7 lbs. b. 13.5 lbs. c. 14.4lbs. d. 20.2 lbs. **e. 22.9 lbs.**

$\sqrt{(1-r^2)}\, s_y = \sqrt{(1-.82^2)} \times 40 = 22.9.$

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

4. If we were to randomly take one adult from this sample, how much would his/her height typically differ from 65 by?

a. 0.05 in. b. 0.1 in. c. 0.5 in. d. 1.0 in. e. 2.5 in. f. 5.0 in.

example problems.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

4. If we were to randomly take one adult from this sample, how much would his/her height typically differ from 65 by?

a. 0.05 in. b. 0.1 in. c. 0.5 in. d. 1.0 in. e. 2.5 in. **f. 5.0 in.**

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the median height is 64.5 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

5. Why shouldn't one trust this regression line to predict the weight of someone who is 25 inches tall?

a. The sample size is insufficiently large.

b. The sample SD of weight is too small.

c. The value of 25 inches is too far outside the range of most observations.

d. The correlation of the ANOVA is a t-test confidence interval with statistical significance.

e. The data come from an observational study, so there may be confounding factors.

f. The height values are heavily right skewed, so the prediction errors are large.

Suppose that among a sample of 100 adults in a given town, the correlation between height (inches) and weight (lbs.) is 0.82, and the mean height is 65 inches, the median height is 64.5 inches, the sd of height is 5 inches, the mean weight is 160 lbs., and the sd of weight is 40 lbs.

5. Why shouldn't one trust this regression line to predict the weight of someone who is 25 inches tall?

a. The sample size is insufficiently large.

b. The sample SD of weight is too small.

c. The value of 25 inches is too far outside the range of most observations.

d. The correlation of the ANOVA is a t-test confidence interval with statistical significance.

e. The data come from an observational study, so there may be confounding factors.

f. The height values are heavily right skewed, so the prediction errors are large.