Inference

German Tanks Activity

Estimating the Maximum from a Sample NAME_____

Seven serial numbers: _____ _____ _____ _____ _____ _____ _____

In your groups, create three statistics that can be calculated from this sample of seven tank numbers that you believe will be a decent estimate of the maximum tank number in the population. Assume the tank numbers are sequential, starting with 1.

1. Write each statistic as a *formula*. Example: MAXIMUM = MEDIAN(data) + (IQR(data))

Formula #1:

Formula #2:

Formula #3:

2. Use the class sample of seven tank numbers to make your estimate:

Estimate using formula #1: _____

Estimate using formula #2: _____

Estimate using formula #3: _____

3. How can we determine which statistic is best for estimating the maximum number of tanks?

4. What are the two main features of a "**good estimator**?" ("good statistic")

5. What is the **parameter** in this problem? What is a **statistic**?

6. Describe the concept of **sampling variability** in the context of this activity.

7. What is a (theoretical) **sampling distribution** in this activity?

8. Choose one of the statistics that was simulated. Describe the **variability of this statistic**.

How is this different than the variability of the data?

From 2010 AP Exam:

2. A local radio station plays 40 rock-and-roll songs during each 4-hour show. The program director at the station needs to know the total amount of airtime for the 40 songs so that time can also be programmed during the show for news and advertisements. The distribution of the lengths of rock-and-roll songs, in minutes, is roughly symmetric with a mean length of 3.9 minutes and a standard deviation of 1.1 minutes.

   (a) Describe the sampling distribution of the sample mean song lengths for random samples of 40 rock-and-roll songs.

****Check out #6 on the 2019 Exam, part (c): (sampling distribution of a median).**

Sampling Distributions:

"The sampling distribution is the basis for inferential statistics, whether one is doing estimation or testing a hypothesis. It is our understanding of the behavior of sample statistics that logically forms the basis for making inferences. Without an understanding of sampling distributions, the process of making inferences is mechanical…" --Chris Olsen in "Special Focus: Sampling Distributions" available at AP Central

"Sampling distributions. The topic that strikes fear into the hearts of introductory statistics teachers everywhere. Clearly this is the most abstract concept that we ask our students to come to terms with in the AP Statistics course. Nonetheless it is critical that students develop an understanding of sampling distributions if they are to comprehend the logic of statistical inference." --Roxy Peck (ibid)

Good resources:

Reese's Pieces applet simulation: http://www.rossmanchance.com/applets.html

Sampling Distribution applet: http://onlinestatbook.com/stat_sim/index.html

Statkey online simulator: lock5stat.com/StatKey/

Special Focus: Sampling Distribution at AP Central:
http://apcentral.collegeboard.com/apc/public/courses/teachers_corner/2151.html

German Tanks Problem (from Special Focus above)

During WWII, Allied spies were asked to estimate the numbers of tanks the Germans had of various types. At about the same time, the Allies were able to capture a number of German tanks, and it was discovered that part numbers on the tanks had coded information that almost certainly indicated serial numbers from the same factories. The part numbers were decoded, and British mathematicians were given the serial numbers and asked to estimate the number of tanks. The mathematicians came up with estimates quite a bit lower than those given by the spies. Long after the war, it was discovered that the spies had been deceived by the Germans repainting their tanks to increase their apparent numbers. The mathematicians were much closer to getting the number of tanks right.

This problem was first introduced to the world in 1947, shortly after many documents concerning WWII became declassified. The original article was *An Empirical Approach to Economic Intelligence in World War II* by Richard Ruggles and Henry Brodie, published in *the Journal of the American Statistical Association,* Vol. 42, No. 237. (Mar., 1947), pp. 72–91. Much has been published about it since then, and information can readily be found on the Web by searching for "German Tank Problem."

A simulation of this problem can be conducted in class—several versions of this activity exist, including a version using Fathom (described in a later article in the Special Focus document).

Reese's Pieces Activity: Sampling Distribution of $\hat{p}$

BIG QUESTION: What percent of Reese's Pieces are orange? Guess: _____

Let me take a sample of 25 pieces from a big bag. My sample proportion (called $\hat{p}$) is _____.

So now we know, right? Well, maybe. How likely is it that MY ONE SAMPLE is a PERFECT representation of the population of ALL Reese's Pieces? Not very, since samples vary…

But if we took LOTS of samples of size 25 and calculated the percent orange, we would begin to get an idea of how *sample proportions* behave.

And if we know how MANY sample proportions behave, we can then begin to understand how confident we can be that our ONE sample taken ONE TIME is near the true proportion of orange pieces.

So let's find out how LOTS of sample proportions behave. Take as many samples of 25 as you can, and calculate the percent orange.

Sample #1: _____% Sample #2: _____% Sample #3: _____%

Make a dot plot of the ***class*** $\hat{p}$'s below:

REMEMBER: This is a distribution of st_____, NOT d_____.

Therefore, it is a very different distribution called a s_____ distribution.

According to mathematicians who studied this a bit more deeply, the mean of a sampling distribution of sample proportions ($\hat{p}$'s) is _____, where _____ is the TRUE population proportion.

The standard deviation of a sampling distribution of $\hat{p}$'s is _____.

Reese's Extension: How far away *in %* is the true proportion of orange from your sample proportion?

First, we need to calculate the standard deviation using the formula. The "official" formula for the standard deviation of a sample proportion is:

But we don't know what *p* is, do we? So instead, we will use the next best thing, _____.

From now on, we will use the formula = _____ (sometimes called standard error)

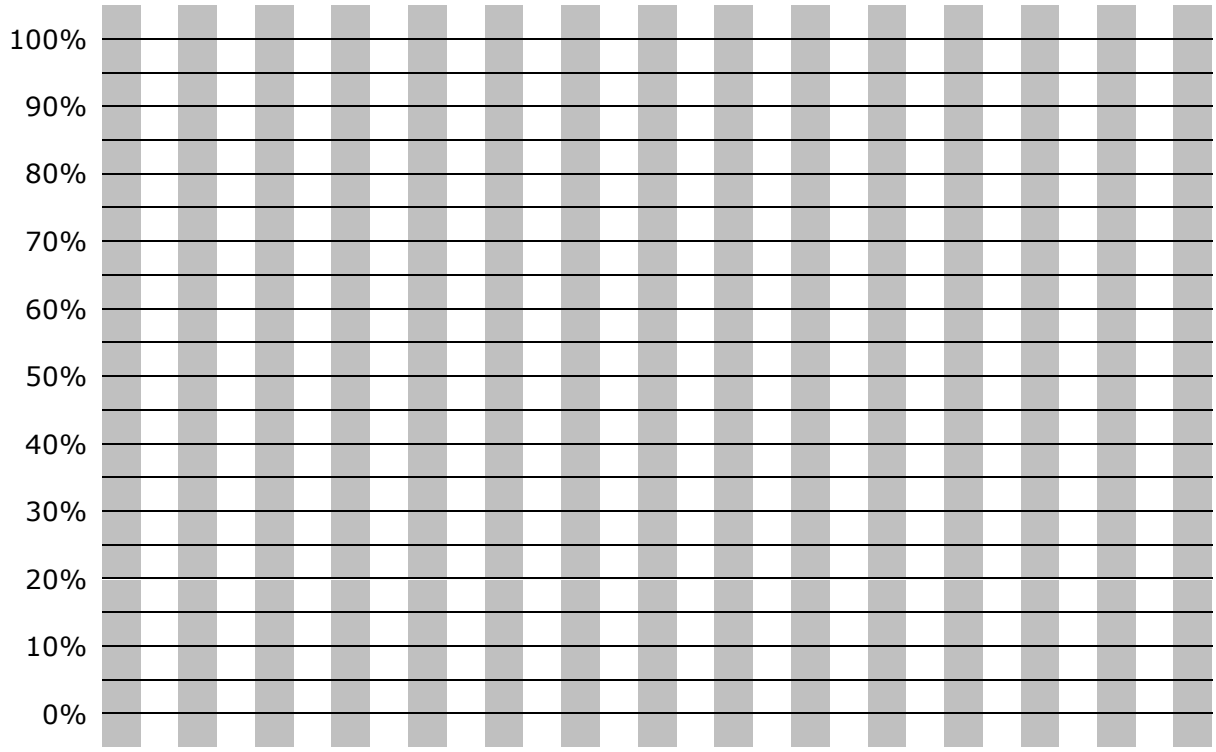
So for our example, we expect the true proportion of orange pieces to be within two standard deviations or _____ from my sample proportion of _____

Thus, we are "pretty sure" that the true proportion lies somewhere between _____% and _____%

This interval is called a ***confidence interval.***

How sure are we? _____ Therefore, this is called a _____ confidence interval. Calculate a 95% confidence interval for all of your sample proportions.

Sample		SD of	95% Conf. Int
1			
2			
3			



(See Rossman Chance applet: http://www.rossmanchance.com/applets/ConfSim.html)

Confidence Intervals:

True or False? (Full document, "Confidence Interval Wording," is at noblestatman.com)

Assume a 95% confidence interval has been calculated for the proportion of orange Reece's Pieces from a bag of 121 candies. The interval is (.32, .49). Forty-nine pieces were orange.

_____1. There is a 95% chance that the true proportion lies between .32 and .49.

_____2. 95% of the time, the true proportion will lie between .32 and .49.

_____3. 95% of the sample proportions lie in this interval.

_____4. There is a 95% chance that the sample proportion lies in this interval.

_____5. If we computed repeated random sample proportions, about 95% of them would lie in this interval.

_____6. In the long run, 95% of all sample proportions will yield an interval that contains the true population parameter.

_____7. The interval (.32, .49) will be correct 95% of the time and wrong 5% of the time.

"Interpret your interval" vs. "Explain the meaning of 95% confidence"

Interpret:

Explain confidence:

Gallup Animal Racing Survey

Gallup conducted a survey of 1,017 national adults, aged 18 and older, on May 8-11, 2008[1]. 387 people said they either strongly support or somewhat support banning animal racing.

a. Create and interpret a 95% confidence interval (and check conditions).

b. Explain "95% confidence" in the context of this survey.

c. What was the margin of error for this survey? _____

[1] The date is important…what happened in the 2008 Kentucky Derby?

Hypothesis Testing

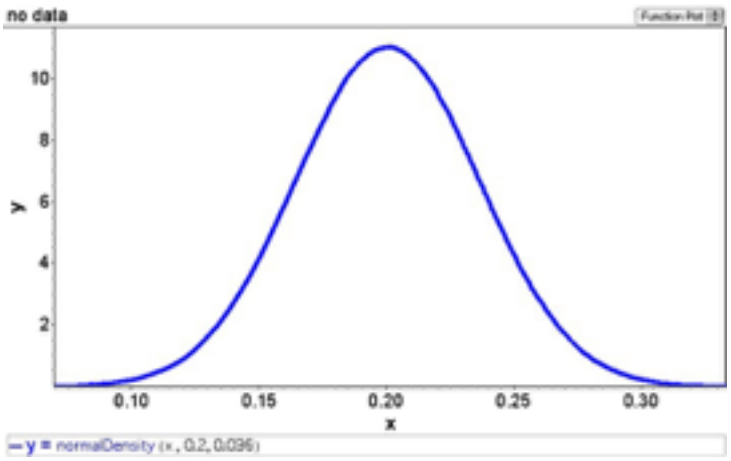
A 1996 report from the U.S. Consumer Product Safety Commission claimed that at least 90% of all American homes have at least one **smoke detector**. A city's fire department has been running a public safety campaign about smoke detectors consisting of posters, billboards, and ads on radio and TV and in the newspaper. The city wonders if this concerted effort has raised the local level above the 90% national rate. Building inspectors visit 400 randomly selected homes and find that 376 have smoke detectors. Is this strong evidence that the local rate is higher than the national rate?

Claim: 20% of all plain **M&M's** in the world are orange. Suppose a bag of 122 has 21 orange ones. Does this contradict the 20% claim?

1. Hypotheses:

2. (Assume all conditions are met.)

3. Mechanics/calculations:



p-value = _____

4. Conclusions:

P-value:

4. A survey organization conducted telephone interviews in December 2008 in which 1,009 randomly selected adults in the United States responded to the following question.

> At the present time, do you think television commercials are an effective way to promote a new product?

Of the 1,009 adults surveyed, 676 responded "yes." In December 2007, 622 of 1,020 randomly selected adults in the United States had responded "yes" to the same question. Do the data provide convincing evidence that the proportion of adults in the United States who would respond "yes" to the question changed from December 2007 to December 2008 ?

Student Sample #2:

$H_0: \hat{p}_1 = \hat{p}_2$ $\quad H_a: \hat{p}_1 \neq \hat{p}_2$

$\hat{p}_1 = \frac{676}{1009} = .6699$

$\hat{p}_2 = \frac{622}{1009} = .6164$

Random = yes
Independent = yes
<10% pop = yes
$\#p_1, \#p_2 > 10$ = yes
$\#q_1, \#q_2$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE(\hat{p}_1 - \hat{p}_2)}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$SE(\hat{p}_1 - \hat{p}_2) = .0213$

$z = \frac{.0535}{.0213}$

$z = 2.51$

p-value = .012

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE(\hat{p}_1 - \hat{p}_2)}$$

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$SE(\hat{p}_1 - \hat{p}_2) = .0213$

I reject the null hypothesis and conclude that the data does support that the proportion of adults in the United States who would respond "yes" to the question changed from December 2007 to December 2008.

PIEP = 2

Type I & II Errors and Power

Bead activity (from Floyd Bullard)

Each group has 200 beads, a mixture of blue and green

1. H_0: The proportion of green beads = 0.50
2. From your cup, take a sample of 20 beads.
3. Reject or do not reject H_0 (use calculator!)
4. Take 15 samples, keeping tally of rejections vs. non-rejections

Rejected H_0:	Did not reject H_0:

Questions:

1. What percent of the time did your group reject H_0? _____

2. What percent of the time did your group NOT reject H_0? _____

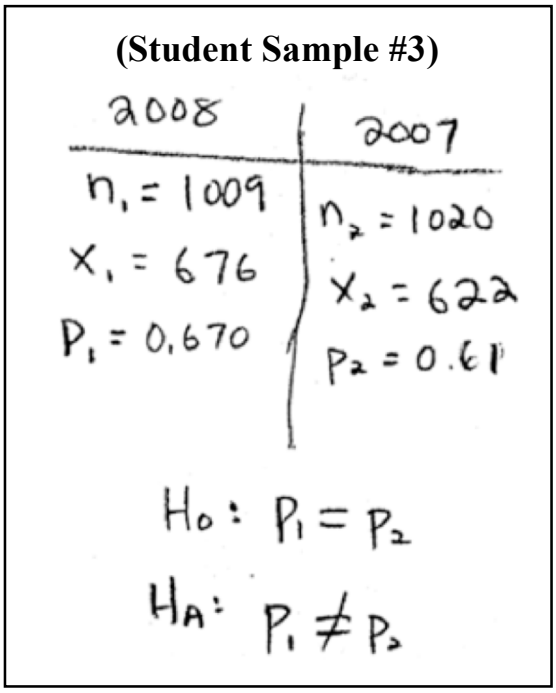
3. What was the true proportion of green beads in your cup? p = _____

4. What type of error did your group make? _____ What was your error rate? _____-

5. What was your power? _____

6. How could you improve your power?

(Student Sample #3)



From the AP Statistics EDG (Electronic Discussion Group):

Yet another method that works well is the criminal justice system. The null hypothesis is not guilty and the alternate hypothesis is guilty. Power then is the ability to reject the null hypothesis when it is false which would mean the ability to find a guilty criminal guilty. Lower alpha level means higher burden of proof. Higher alpha level means lower burden of poof. So it is easier to find a guilty criminal guilty (higher power) if there is lower burden of proof (higher alpha level). But with higher alpha level there is higher chance of Type I error which would be putting an innocent person in jail.

--George Petoff

Talk about this stuff with a concrete example in everyday language, and convince the kids it's just common sense. After they understand the concepts, get them to translate into Stats jargon.

For example, we're testing a new headache medication to see if it's more effective than the current treatment that's known to relieve pain within 15 minutes for 60% of headache suffers. We give the new med to a bunch of volunteers and see what fraction of them report relief. Note that:

1) If the new med actually is 90% effective we'll almost surely notice. If it's only 65% effective we may miss that. (Translation: the greater the effect size, the higher the power.)

2) If we're willing to accept less evidence, we're more likely to notice any effect. (Translation: The higher the alpha level, the greater the power.)

3) If we're easier to convince, the more likely we are to mistakenly think the new med works when it really doesn't. (Translation: The higher the alpha level, the greater the risk of Type I error.)

4) If we're easy to convince, the less likely we are to miss an effective med. (Translation: The higher the alpha level, the lower the risk of Type II error.)

5) If we demand stronger evidence, the more likely we are to overlook the fact that the medication is more effective. (Translation: The lower the alpha level, the greater the risk of Type II error.)

6) If we require a higher standard of proof, the less likely we are to be fooled into thinking an ineffective med works, (Translation: The lower the alpha level, the lower the risk of Type I error.)

7) If we're tougher to convince, the less likely we are to notice that the med really is effective. (Translation: The lower the alpha level, the less power we have.)

Say it as many different ways as you can, get them to explain what common sense suggests, and then challenge them to apply the right vocabulary. The ideas are easy; start there.

--Dave Bock

From *Stats: Modeling the World* (3e, p.503, #34)

A potter has a 40% breakage rate during kiln firing, so she buys more expensive clay. She fires 10 pieces and will decide to use the new clay if at most one of them breaks.

a) Suppose the new clay is no better. What is the probability she is convinced to switch?

b) Suppose the new clay can reduce breakage to 20%. What is the probability that her test will not detect the improvement?

c) How can she improve the power of her test?

Sample MC Problem:

An independent research firm conducted a study of 100 randomly selected children who were participating in a program advertised to improve mathematics skills. The results showed no statistically significant improvement in mathematics skills, using $\alpha = 0.05$. The program sponsors complained that the study had insufficient statistical power. Assuming that the program is effective, which of the following would be an appropriate method for increasing power in this context.

A) Use a two-sided test instead of a one-sided test.
B) Use a one-sided test instead of a two-sided test.
C) Use $\alpha = 0.01$ instead of $\alpha = 0.05$.
D) Decrease the sample size to 50 children.
E) Increase the sample size to 200 children.

Cents and the Central Limit Theorem

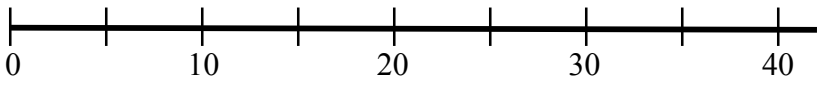
(from *Activity-Based Statistics;* sets up understanding simulations: onlinestatbook.com/stat_sim/sampling_dist/index.html)

1. Using a population of pennies provided by the teacher, you will be taking two random samples of size n = 5, two random sample of size n = 10, and one random sample of size n = 25.

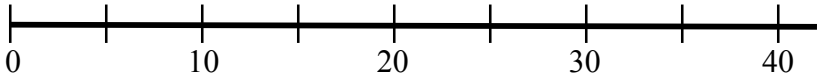
2. Calculate the mean age of each sample, and place a counter on the number line provided in the classroom.

3. Sketch a graph of the means from the entire class on each of the three graphs below.

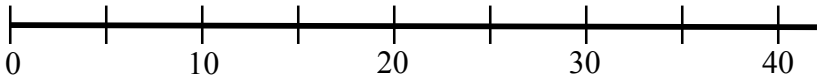
n = 5



n = 10



n = 25



*This is a demonstration of the **Central Limit Theorem**. Describe the CLT below.*

Chi Square Tests

1. Is the distribution of colors of M&M's in my bag the same as the Mars Company claims? Count the data, and let's find out! This will be a Chi-Square Goodness of Fit Test.

	Brown	Yellow	Red	Blue	Orange	Green
M&M Mars Official %'s						
My bag:						
Expected in my bag:						

2. Medical researchers enlisted 108 subjects for an experiment comparing treatments for depression. The subjects were randomly divided into three groups and given pills to take for a period of three months. Unknown to them, one group received a placebo, the second group received the "natural" remedy St. Johnswort, and the third group the prescription drug Paxil. After six months psychologists and physicians (who did not know which treatment each person had received) evaluated the subjects to see if their depression had returned. (This will be a Chi-Square Test for Homogeneity.)

	Treatment			
Diagnosis	Placebo	St J	Paxil	Total
Depression returned	24	22	14	
No sign of depression	6	8	16	
Total				

3. Were class and "survivability" on the Titanic independent? (Chi-Square Test for Independence)

Titanic Passengers

		Survived		Row Summary
		No	Yes	
Class	1st	129	193	322
	2nd	161	119	280
	3rd	574	137	711
Column Summary		864	449	1313

S1 =

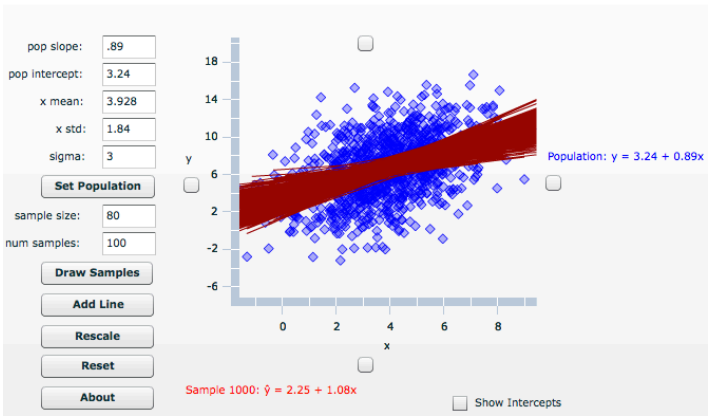
4. How do you tell from the data whether a Chi-Square test for homogeneity or a Chi-Square test for independence is appropriate?

Linear Regression t-tests:

What does your calculator produce using the following program?

```
randInt(1,100,20
)→L1:randInt(1,1
00,20)→L2:LinReg
(a+bx) L1,L2:b
```

Rossman-Chance Applet
Simulating Regression Lines



Conditions for a linear regression t-test:

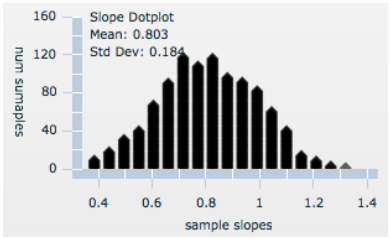
1.

2.

3.

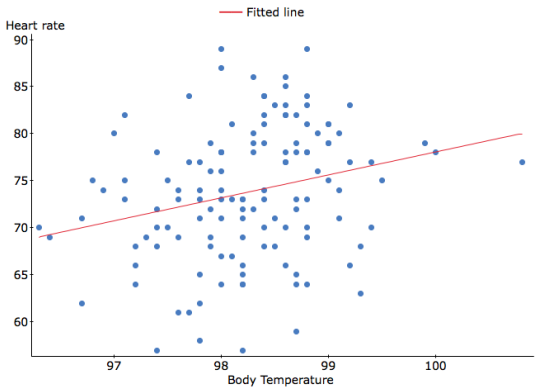
4.

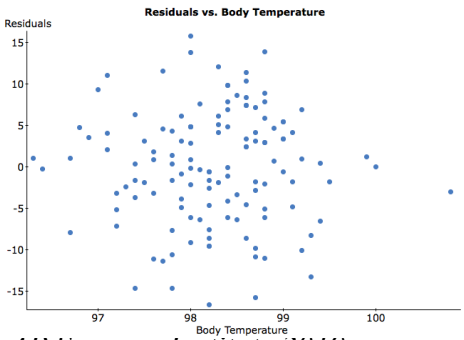
5.

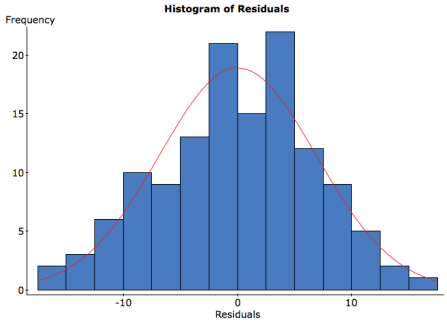


Simple linear regression results:
Dependent Variable: Heart rate
Independent Variable: Body Temperature
Sample size: 130
R-sq = 0.064
S = 6.8577393

Parameter	Estimate	Std. Err.	Alt	DF	T-Stat	P-value
Intercept	-166.28	80.912	≠ 0	128	-2.055	0.0419
Slope	2.44	0.8235	≠ 0	128	2.967	0.0036







2011 #5: Windmills

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

Score the following student response:

(a) Use the computer output above to determine the equation of the least squares regression line. Identify all variables used in the equation.

$\hat{y} = 0.137 + 0.240(x)$

electricity produced = 0.137 + 0.24(wind velocity)

(b) How much more electricity would the windmill be expected to produce on a day when the wind velocity is 25 mph than on a day when the wind velocity is 15 mph? Show how you arrived at ~~your answer~~.

$\hat{y} = 0.137 + 0.24(25) = 6.137$

$\hat{y} = 0.137 + 0.24(15) = 3.737$

6.137
- 3.737
2.4

2.4 more electricity

(c) What proportion of the variation in electricity production is explained by its linear relationship with wind velocity? R^2 - 87.3% of variation in electricity produce can be explained by wind velocity

(d) Is there statistically convincing evidence that electricity production by the windmill is related to wind velocity? Explain. yes, because the p-value is low, there for the null hypothesis should be rejected.

From 2005B #5:

(c) John wants to provide a 98 percent confidence interval for the slope parameter in his final report. Compute the margin of error that John should use. Assume that conditions for inference are satisfied.

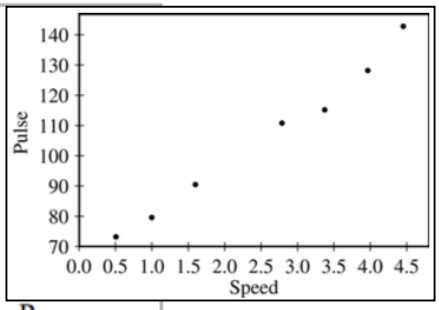
Regression Analysis: Pulse Versus Speed

Predictor	Coef	SE Coef	T	P
Constant	63.457	2.387	26.58	0.000
Speed	16.2809	0.8192	19.88	0.000

S = 3.087 R-Sq = 98.7% R-Sq (adj) = 98.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3763.2	3763.2	396.13	0.000
Residual	5	47.6	9.5		
Total	6	3810.9			



Hint--think of CI formula: _____ ± _____ • _____

2011 #5: EPEP = 3
b) no units, d) no ID of p-value

Tommy John and Errors

Famous pitcher Tommy John once made three errors on a single play: he bobbled a grounder, threw wildly past first base, then cut off the relay throw from right field and threw past the catcher.

In a scientific paper describing a clinical trial comparing a new pain drug with a placebo, the authors wrote something like this: "Although there was no difference in baseline age between the groups (p = 0.458), controls were significantly more likely to be male (p = 0.000)."

This statement is worse than Tommy John's worst day because there are actually four errors in this sentence (or maybe even 4½). See if you can find them.

Garfield and Ben-Zvi's Eight Learning Principles:[2]

1. Students learn by constructing knowledge.
2. Students learn by active involvement in learning activities.
3. Students learn to do well only what they practice doing.
4. It is easy to underestimate the difficulty students have in understanding basic concepts of probability and statistics.
5. It is easy to overestimate how well students understand basic concepts.
6. Learning is enhanced by having students become aware of and confront their errors in reasoning.
7. Technological tools should be used to help students visualize and explore data, not just to follow algorithms to pre-determined ends.
8. Students learn better if they receive consistent and helpful feedback on their performance.

Marzano's Keys to Student Engagement

Students must answer in the affirmative:

How do I feel? Am I interested? Is it important? Can I do it?

[Getting their attention.]…….[Getting them engaged.]

Answers to Tommy John problem:

1. Accepting the null hypothesis
2. Giving a p-value for baseline differences between random groups (p-values test hypotheses).
3. Inappropriate levels of precision (what do the 5 and 8 tell us?)
4. Reporting a p-value = 0.

4½: Why were they measuring baseline ages anyway? All patients will be in the trial over the same time period.

[2] Garfield, Joan, and Dani Ben-Zvi. "How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics." *International Statistical Review* 75.3 (2007): 372-96