General Outline from the midterm to the exam

Oct 23 not in TEXT Useful stats exercise (pages 85-86) HW: Paper 1 due Oct 30 (page 87-88)	Oct 25 TEXT Chapter 12 Linear regression (page 89-90) HW: finish Paper 1
Oct 30 TEXT chapter 8 Confidence intervals (pages 91-94) HW: Paper 2 due Nov 6 (page 95-96) Finish confidence intervals (pages 91-94) Read research article (pages 97-102)	Nov 1 Discuss article HW: finish Paper 2 Correct paper 1 if needed Re-read research article
Nov 6 TEXT Chapter 9 Finish discussing article Begin Hypothesis testing (pages 107-114) HW: Paper 3 due Nov 13 (page 103-106)	Nov 8 TEXT Chapter 9 Hypothesis testing (pages 107-114) HW: finish Paper 3 Correct papers 1 and 2 if needed Finish Hypothesis testing (pages 107-114)
Nov 13 TEXT Chapter 10 Sample to sample comparisons (pages 115-125) HW artcle analysis homework	Nov 15 TEXT Chapter 10 Sample to sample comparisons (pages 115-125) HW: Correct papers 1,2 and 3 if needed Finish sample to sample comparison (pg 115-125)
Nov 20 not in text Probability of groups (pages 139-144) HW: Paper 4 due Nov 29 (page 137-138) Finish Probability of groups (pages 139-144)	Nov 22 No Class Thanksgiving
Nov 27 Work day- Homework due Friday???	Nov 29 Review sheet workday HW: Exam Review A (pages 145-148) due May 1
Dec 4 Questions on Exam Review A HW: Exam Review B (pages 149-154)	Dec 6 Questions on Exam Review B HW: Exam review sheet C (155-156)
Dec 11 Questions on Exam Review C HW: Exam review sheet D (pages 157-162)	Dec 13 Questions on Exam Review D HW: study for the final exam

Groups

I am still number ________

Shapes

Square	1	2	3	4
Triangle	5	6	7	8
Hexagon	9	10	11	12
Octagon	13	14	15	16
Rectangle	17	18	19	20
Rhombus	21	22	23	24
Diamond	25	26	27	28

Trucks

Dump	1	5	10	14
Mixer	3	7	12	16
Recycling	2	8	9	13
Semi	4	11	15	18
Pick-up	6	17	21	26
Logging	19	22	25	28
Bucket	20	23	24	27

Kayak Words

Kayak	1	7	22	28
Paddle	3	9	14	18
Tandem	2	5	11	20
Rudder	4	10	13	27
Roof rack	6	12	15	23
Life vest	16	17	25	26
Dry bag	8	19	21	24

Veggies

Broccoli	1	6	9	16
Peppers	3	5	22	26
Spinach	7	11	14	20
Lettuce	12	13	17	18
Carrots	2	15	24	27
Asparagus	4	10	21	25
Eggplant	8	19	23	28

Fruits

Apple	1	8	11	16
Orange	3	6	13	22
Kiwi	4	5	24	25
Pear	7	10	15	26
Mango	2	14	17	23
Peach	12	18	19	27
Banana	9	20	21	28

Pets

Hamster	1	13	20	26
Snake	3	6	11	18
Hermit crab	5	9	12	22
Bird	7	17	24	28
Fish	2	14	15	19
Cat	4	8	21	27
Dog	10	16	23	25

Trees

Elm	1	2	6	25
Oak	3	14	27	28
Maple	4	5	23	26
Pine	7	9	18	24
Dogwood	8	10	11	17
Aspen	12	15	20	21
Willow	13	16	19	22

Olympic Sports

Bobsled	1	12	26	27
Archery	3	8	10	15
Curling	5	16	17	20
Handball	2	4	7	19
Ice dancing	6	9	23	28
Steeple chase	11	22	24	25
Biathlon	13	14	18	21

Insects

Ant	1	15	24	28
Bee	2	3	21	26
Caterpillar	5	6	10	19
Dragonfly	7	18	20	23
Earwig	4	9	14	17
Firefly	11	16	22	27
Grasshopper	8	12	13	25

Finding useful statistics from survey data

	Age	Height (inches)	Cups of Coffee (per week)	Drink Alcohol (days per week on avg)	Number of Supplements Taken per workout	Days Use a Gym (weekly)	How Long Have Been Working Out at the Gym (years)
mean	20.48	67.55	1.2	1.466666	0.8	1.283	2.083
standard deviation	3.34	2.317	1.5709	1.255046	1.021796358	1.474	1.7202
median	19.5	68	0	1	0	1	2
max	36	72	5	5	3	6	7
min	17	63	0	0	0	0	0
mode	19	68	0	2	0	0	2
Q1	19	66	0	0	0	0	0
Q3	21	69	3	2	2	2	3

	Asthmatic	Diabetic	Caffeine Allergy	Exercise at all	Eat Healthy	Use Bartley Center	Kinds of Workouts
Yes	6	2	0	50	15	13	Endurance 28
No	54	58	60	10	45	47	Strength 23

1. What is a broad overall claim you can make about all HCC students using this data?

2. Find at least 5 statistics (preferably more) that support this claim. State in a full sentence how they support the claim.

3. Find at least 5 statistics (preferably more) that contradict this claim. State in a full sentence how they contradict the claim.

49 full time students and 11 part time students were interviewed

	Asthmatic	Diabetic	Caffine Allegy	Exercise at all	Eat Healthy	Use the Bartley Center	Kinds of Workouts
Yes	FT PT 6 0	FT PT 1 1	FT PT 0 0	FT PT 40 10	FT PT 8 7	FT PT 9 4	Endurance FT=24, PT=4
No	FT PT 43 11	FT PT 48 10	FT PT 49 11	FT PT 9 1	FT PT 41 4	FT PT 40 7	Strength FT=16, PT=7

4. Find at least two examples of where using the full time or part time data only would be more persuasive than using all the survey data to again draw conclusions about the entire HCC population.

Paper 1: Descriptive Statistics Only

Part I: Choose one of the following data sets and topics (direct links on Google site) Remember you must Make A Copy of the file before you can edit it.

Titanic- Prove that who died was not random.
https://docs.google.com/spreadsheets/d/1FFeQtqlJMP5u3LHbEXiA8XD2ycOC-HB1TJ4fkscrLVM/edit?usp=sharing

HCC pay- Prove that HCC is biased in their employee pay.
https://docs.google.com/spreadsheets/d/1EbflUL1h0O8VjXXMAMNR7vTlvWRXLtb6egieCYfPHEo/edit?usp=sharing

Best college- Prove the best region and type of college to attend along with best major
https://docs.google.com/spreadsheets/d/14CHrk7z2JzrBBf7AbU4oo0QJsguGnWQ2FpoU87V1TiY/edit?usp=sharing

Part II: Analyze the data.
Use descriptive statistics to analyze the data. Find statistics that support your argument. Make sure you look at subsets of the data as that data might be more convincing. The paper must include at least 7 stats and they must be at least 3 different kinds (i.e. they can not all be percentages or all means.)

Part III: Create at least 1 graph
Create at least 1 graph to accompany your article. This graph must be created on a computer. I recommend using Google sheets since you already know how to use it from the graphing assignment earlier this semester. The graph must include a title, axis titles, axis scales, and a legend. The point of the graph must be referenced in the article even if you do not mention the graph itself.

Part IV: Write the article.
The article must be 500 to 1000 words longs (most word processors have a word count feature.) It must be well-written and proof read for errors. Use only the data provided in the spreadsheets. The paper must include at least 7 statistics and they must be at least 3 different kinds (i.e. they can not be all percentage or all means.)

Part V: Submit the article.
The article must be submitted in paper form with this piece of paper and it must also be submitted electronically by sharing a Google document with cdillard@hcc.edu or by emailing a copy of the file to cdillard@hcc.edu.

Rubric for paper 1

Checklist prior to grading. Paper will not be graded until all 3 requirements are met.
__________Article is typed
__________Article is 500-1000 words
__________Article is submitted electronically and a paper copy was handed in with this grading sheet

	Points
There is a clear and relevant main point.	/5 points
At least 7 (but hopefully more) stats were used in the article and they were at least 3 different kinds. The stats used in the article supported the point and were included appropriately. Each statistic used in the article was explained or interpreted clearly and in depth.	/50 points
At least one computer generated graph was used in this article. The graph(s) supports the point of the article. The take-away from the graph is referenced in the article (even if the graph itself is not mentioned).	/30 points
The article was well written and easy to understand.	/15 points
	Total /100 points

Linear Regression Worksheet

Use this data for this worksheet (live link on google site for this course)
https://docs.google.com/spreadsheets/d/1mah0fXb0pPG9W_PUBJ6qIjj97YK_DHxDWLU1y52yhiY/edit?usp=sharing

1. Create four scatterplots and linear regression models.
a) poverty percents to teen births linear model equation and correlation coefficient

b) poverty percents to violent crimes rates linear model equation and correlation coefficient

c) poverty percents to birth rates for 15-17 year olds linear model equation and correlation coefficient

d) poverty percents to birth rates for 18-19 year olds linear model equation and correlation coefficient

2-9. Use the linear regression models to answer the following questions.
2. Which correlation is stronger poverty to crime or poverty to teen births? Justify your answer.

#3-#9 use the linear regression models to predict certain data points.

3. Using only data from 15 to 17 year olds, what is the predicted teen birth rate for states with 20% poverty?   Show work.

4. Using only data from 18-19 year olds, what is the predicted teen birth rate for state with 20% poverty? Show work.

5. Using the teen birth data, what is the predicted teen birth rate for states with 20% poverty? Show work.

6. What is the predicted teen birth rate for states with 0% poverty? Show work.

7. What is the predicted teen birth rate for states with 40% poverty? Show work.

8. If a state has a teen birth rate of 32, what is the expected percent poverty in that state?

9. If a state has a teen birth rate of 10, what is the expected percent poverty in that state?

Confidence Intervals

When using a sample to estimate characteristics of a population, Confidence intervals are one way to discuss error

When collecting data, we rarely have the opportunity to gather information from an entire population, so we assume the sample is representative of the population. This is only possible if the sample chosen and the data gathered was random and unbiased (to the best of our ability.)

Let's say in my research I found that 45% of the first year HCC students I spoke to were planning to take statistics as their college level math class. So the best estimate I have is that 45% of ALL first year HCC students plan to take stats. But how good is that estimate? How confident am I in that statistic?

Statisticians use something called a ***confidence interval***. A confidence interval is an interval that you are confident contains the estimated information.

I'll repeat to make sure you got that: A confidence interval is an interval that you are confident contains the estimated information.

The most common confidence intervals are 90%, 95% and 99%. We will only work with these, but you can have any level confidence level.

Back to my research:

After doing the computations, I found that a 95% confidence interval gave me an error of 1.5%

So my interval is 45-1.5 to 45+1.5 which is 43.5% to 46.5%

This means I am 95% confident that the population percent lies in the interval 43.5% to 46.5%

Or in context, it means that I am 95% confident that the actual percentage of all HCC first year students who plan to take statistics is between 43.5% and 46.5%

Note: This does **NOT** mean there is a 95% probability that it lies in the interval.

1. My brother is doing research on Americans with tattoos. As part of his survey, he found that 17% of his sample had at least one tattoo. He found using a 99% confidence level that his error is 0.024

a) What is his 99% confidence interval?

b) What does this mean in a sentence?

c) According to: Pew Research Center, Tattoo Finder, Vanishing Tattoo. 14% of Americans have at least one tattoo. Does my brother's research *disprove* the Pew Center research?

Estimating Percents of a Population

The sample percent is the best estimate of the population percent

Confidence interval for a percent (call it p) is $p \pm E$ where E is the estimated error

<table>
<tr>
<td>

$p \pm E$

$E = z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$

</td>
<td>

E = estimated error

p = sample percent as a decimal

n = sample size

<table>
<tr>
<td>If want a 90% confidence interval $z_{\alpha/2}$ = 1.645</td>
<td>If want a 95% confidence interval $z_{\alpha/2}$ = 1.96</td>
<td>If want a 99% confidence interval $z_{\alpha/2}$ = 2.575</td>
</tr>
</table>

</td>
</tr>
</table>

2. a) Find a 99% confidence interval for my 45% of all first year HCC students plan to take stats, if I spoke to 68 students.

b) Explain what this means.

3. a) Find a 90% confidence interval for my 45% of all first year HCC students plan to take stats, if I spoke to 68 students.

b) Explain what this means.

c) Why is this interval smaller?

Estimating the Mean of a Population with unknown σ

The sample mean is the best estimate of the population mean

The following works ONLY IF n>30 or POPULATION IS NORMAL

Confidence interval for the mean is $\mu \pm E$ where E is the estimated error

$\mu \pm E$ $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$	E = estimated error s = sample standard deviation n = sample size degrees of freedom = n-1 $t_{\alpha/2}$ see below

Finding $t_{\alpha/2}$

Use an online calculator:

http://www.statdistributions.com/t/

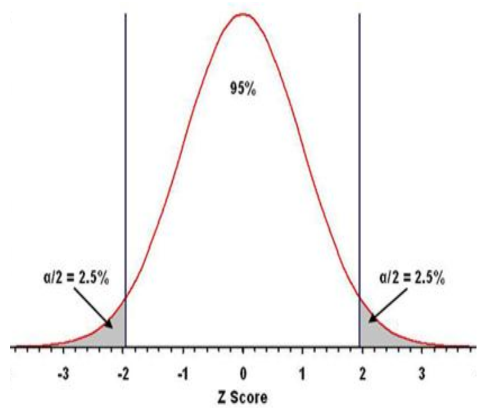
p-value (choose from table)

% interval	wo tailed p-value
90%	0.05
95%	0 025
99%	0.005

t-value is the $t_{\alpha/2}$ so let the computer fill that in

d.f. = degrees of freedom

choose two-tailed because we want the middle interval



It is two tailed because we want the middle interval

4. Data was gathered on the city miles per gallon of 32 different car models. The mean is 21.19 miles per gallon with a standard deviation of 3.477 miles per gallon.

   a) Estimate the mean of all car models with a 90% confidence interval.

   b) Explain what this means.

Estimating the Standard Deviation of a Population

The sample standard deviation is the best estimate of the population standard deviation

The following works ONLY IF THE POPULATION IS NORMAL

Confidence interval for the standard deviation is below

$$\sqrt{\frac{(n-1)\cdot s^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)\cdot s^2}{\chi_L^2}}$$

s = sample standard deviation

n = sample size

degrees of freedom = n-1

χ_R^2 and χ_L^2 can be found using an online calculator

http://www.statdistributions.com/chisquare/

p-value (choose from table)

% interval	p-value
90%	0.05
95%	0.025
99%	0.005

χ^2 value is what the computer fills in

d.f. = degrees of freedom

Click right tail for χ_R^2

Cilck left tail for χ_L^2

5. The height of Old Faithful's geyser was measured 40 times throughout a week and the data had a mean of 127.2 feet with a standard deviation of 13.15 feet.

a) Estimate the standard deviation of the height of the geyser throughout the year with a 95% confidence interval.

b) Explain what this means.

Paper 2: Linear regression

The World Moderators of International Climate have asked you to determine the two best indicators and the two worst indicators of carbon dioxide emissions from the following list.

Population has basic needs met (required to be a developed country)
Widespread communications technology (required to be a developed country)
Country is considered wealthy (required to be a developed country)
Size and type of land in the country
Size and location of the population
Energy use

Write a 500-1000 word data driven argument for your two best and two worst indicators. Include numerous statistics, numerous types of statistics and several graphs. Each of the above indicators have 3-8 columns of data associated with them so make sure you use all the columns of data that help your argument and do not limit yourself to one data set per indicator. Check the rubric for more details.
https://docs.google.com/spreadsheets/d/1TrC6SLG4gvaUJx5Kh3SCmYajG3hsguQpOQ0qHaJK1Ys/edit?usp=sharing (direct link on google site)

Data Column Names

Access to electricity (% of population)	Merchandise imports by the reporting economy (current US$)
Agricultural land (% of land area)	Mobile cellular subscriptions
Agricultural land (sq. km)	Mobile cellular subscriptions (per 100 people)
Capture fisheries production (metric tons)	People using at least basic drinking water services (% of population)
CO2 emissions (kt)	People using at least basic sanitation services (% of population)
CO2 emissions (metric tons per capita)	Population density (people per sq. km of land area)
Energy use (kg of oil equivalent per capita)	Population, total
Energy use (kg of oil equivalent) per $1,000 GDP (constant 2011 PPP)	Rural population
Fossil fuel energy consumption (% of total)	Rural population (% of total population)
Fixed telephone subscriptions	Secure Internet servers
Fixed telephone subscriptions (per 100 people)	Secure Internet servers (per 1 million people)
Forest area (% of land area)	Surface area (sq. km)
Forest area (sq. km)	Total fisheries production (metric tons)
GDP (current US$)	Urban population
GDP per capita (current US$)	Urban population (% of total population)
GDP per person employed (constant 2011 PPP $)	
Land area (sq. km)	
Merchandise exports by the reporting economy (current US$)	

Rubric for paper 2

Checklist prior to grading. Paper will not be graded until all 3 requirements are met.
__________Article is typed
__________Article is 500-1000 words
__________Article is submitted electronically and a paper copy was handed in with this grading sheet

	Points
There is a clear and relevant main point.	/5 points
Enough statistics were used in this article to make a solid argument. The stats used in the article supported the point and were included appropriately. Each statistic used in the article was explained or interpreted clearly and in depth. A variety of stats are included. Linear regression is included.	/50 points
At least 3 graphs accompany this article. The take-aways from the graphs are referenced in the article (even if the graphs themselves are not mentioned). Must include at least one linear regression graph.	/30 points
The article was well written and easy to understand.	/15 points
	Total /100 points

Paper 3- Confidence Intervals

Part I: Look over the student survey data. (direct link on Google site)
https://docs.google.com/spreadsheets/d/1LxFHnFUwJOrZcrlS6GEsWXGbzXAO6WPe-iHt4nVogs0/edit?usp=sharing

There is a lot of data here so you will need to carefully choose which questions you would like to analyze. Look over the questions (on the next page) and pick a topic that stands out to you or is interesting to you and find 5-10 questions that are directly related to that topic.

Come up with a name for the university represented in this data.

Part II: Analyze the data
Analyze and compare the descriptive statistics for each of the 5-10 questions you chose. Come up with a thesis statement- a persuasive statement you will prove in your article that is about ***every student*** at this university, not just the ones who took the survey. You are NOT comparing groups but rather telling the reader about ALL the students at this university.

Part III: Confidence Intervals
As you are analyzing the data, find several confidence intervals that might be helpful in your argument. You will needs to include at least 4 confidence intervals that are analyzing the at least 2 different types of descriptive statistics.

Part IV: Write an article
The article must be a persuasive article trying to convince the reader of your thesis statement about the students at this university.
It must be 500-1000 words, include at least 7 statistics and at least 4 confidence intervals (that are analyzing at least two different types of statistics), and include at least 1 graph.

Rubric for Paper 3

Checklist prior to grading. Paper will not be graded until all 3 requirements are met.
__________Article is typed
__________Article is 500-1000 words
__________Article is submitted electronically and a paper copy was handed in with this grading sheet

	Points
There is a clear and relevant main point.	/5 points
At least 7 (but hopefully more) stats were used in the article and they were at least 3 different kinds. There were at least 4 confidence intervals of at least 2 different kinds. The stats used in the article supported the point and were included appropriately. Each statistic used in the article was explained or interpreted clearly and in depth.	/50 points
At least one computer generated graph was used in this article. The graph(s) supports the point of the article. The take-away from the graph is referenced in the article (even if the graph itself is not mentioned).	/30 points
The article was well written and easy to understand.	/15 points
	Total /100 points

Questions on the survey:

for q1-q51 note that 5 is the best ranking and 1 is the worst ranking

q1	Department provides comprehensive guidelines to the students in advance
q2	Department ensures a conducive learning environment
q3	Academic decisions are made with fairness and transparency
q4	Academic calendar is maintained properly
q5	Results are published timely in compliance with the ordinance
q6	Students' opinion regarding academic and extra-academic matters are addressed properly
q7	Student feedback process is in practice
q8	Website is informative and updated properly
q9	Curriculum load is optimum and induces no pressure
q10	Courses in the curriculum from lower level to higher are properly arranged
q11	Teaching strategies are clearly stated in the curriculum
q12	Assessment strategies are clearly stated in the curriculum
q13	Teaching-learning is interactive and supportive
q14	Class size is optimum for interactive teaching learning
q15	Modern devices are used to improve teaching-learning process
q16	Diverse methods are used to achieve learning objectives
q17	Lesson plans/course outlines are provided in advance to the students
q18	Assessment information is communicated to students at the beginning of the semester
q19	Assessment system meets the objectives of the course
q20	Diverse methods and tools are used for assessment.
q21	Assessment feedback is provided to the students immediately.
q22	The questions of examinations reflect the content of the course.
q23	Both formative (quizzes, assignments, term papers, continuous assessments, presentations etc.) and summative assessment (final examination only) strategies are followed.
q24	Admission policy ensures entry of quality students.
q25	Admission procedure is quite fair
q26	Sincerity and commitment of the students exist to ensure desired progress and achievement.
q27	Overall classroom facilities are suitable for ensuring effective learning.
q28	Laboratories facilities are suitable for practical teaching-learning and research
q29	The library has adequate up-to-date reading and reference materials
q30	Internet facilities with sufficient speed are available
q31	Adequate indoor and outdoor medical facilities are available
q32	Adequate indoor and outdoor sport facilities are available
q33	Existing gymnasium facilities are good enough
q34	Adequate safety measures are available
q35	There is an arrangement to provide guidance and counseling.
q36	Mentoring is done to take care of the students
q37	Scholarships/ grants available to students in case of hardship
q38	Students are encouraged to involve in co- curricular and extra-curricular activities
q39	Alumni are organized and supportive.

q40	Supporting staff are adequate and co-operative
q41	There are opportunities to get involve with community services
q42	The department has a research and development policy
q43	Mechanism exists for engaging the students in research and development
q44	Research findings are properly used in current teaching-learning
q45	The department has a community service policy
q46	What was your expectation about the University as related to quality of education?
q47	What was your expectation about the University as related to quality of Faculty?
q48	What was your expectation about the University as related to quality of resources?
q49	What was your expectation about the University as related to quality of learning environment?
q50	To what extent was your expectation met?
q51	What are the best aspects of the program?
q52	In your opinion, the best aspect of the program is
q53	In your opinion, the next best aspect of the program is
q54	What aspects of the program could be improved?

Hypothesis Testing

First Use: given a sample, estimate something about the population

We already have an error method for this- confidence intervals

Second Use: given information about the population AND the sample, did something affect the sample?

We will focus on these since confidence intervals are no help here

We are going to make a hypothesis and then see if the data is strong enough to support the hypothesis.

Essentially what we are finding is: What is the probability of obtaining a sample outcome?
Is it likely that the sample outcome was a coincidence?
Or is it likely that the sample outcome had an outside influence?

The variable associated with this probability is called the **p-value**.

Hypothesis testing DOES NOT PROVE that anything is true or false! Just like probabilities do not prove anything will or will not happen (assuming it is not 0% or 100% which most probabilities are not.) Much like the confidence intervals, they only provide a certain degree of certainty.

Note: There are several methods used to test hypotheses. We will focus on the statistical method using p-values with $\propto = 0.05$. This alpha is the most common and is similar to a 95% confidence interval.

The basic idea: P-values are an attempt to quantify the chance that the result was by coincidence.

Choosing the two hypotheses (when we know information about the population):

In statistics, the hypothesis and the "other" have very specific rules.

One hypothesis must be of the form: sample characteristic = population characteristis (i.e. μ=3.2)

The second hypothesis is one of these: Sample characteristic > population # (i.e. μ>3.2)
Sample characteristic < population # (i.e. μ<3.2)
Sample characteristic ≠ population # (i.e. $\mu \neq$3.2)

The hypothesis with the = is always called the null hypothesis, denoted H_0. The other hypothesis is called the alternate hypothesis, denoted H_1. The naming is not important for the calculations but it is very important for wording the conclusion.

Interpreting the results:

The p-value is the probability of obtaining a sample outcome, given that the value stated in the null hypothesis is true. If this probability is small, we will reject the null hypothesis. If the probability is not small, then we will not reject it.

Fail to reject H_0 means H_0 could be true. There is evidence to support H_0.

For us this means the sample results have a could have been coincidence and not affected by an outside influence. If p>0.05 there is more than a 5% of randomly getting this sample result. This is too big to convince us that the sample was affected by something.

Reject H_0 means H_1 could be true. There is evidence to support H_1.

For us this means the sample was probably affected by an outside influence. If p<0.05 then there is less than a 5% chance of randomly getting this result, so we assume the sample was affected.

Example A: I stood at a doorway to Frost building at some random time and took the gender of the first 300 people who passed through the doors.

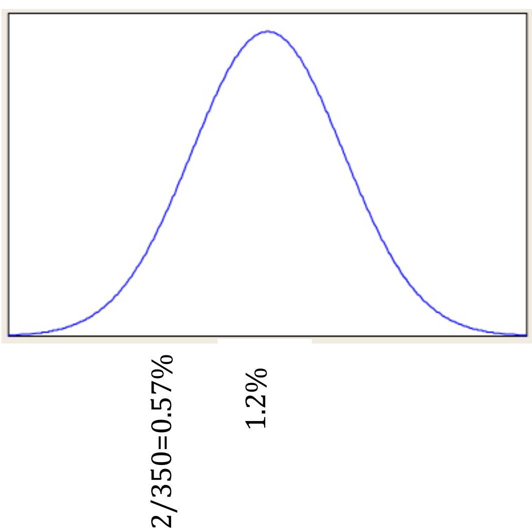
1. What if all of them were male? Statistically speaking is that something out of the ordinary?

2. What if 135 of them were male? Statistically speaking is that something out of the ordinary?

Example B:

A pharmaceutical company is claiming that their drugs will help prevent a certain type of disease. The disease currently affects 1.2% of the population. The research of those using the drug showed only 2 out of 350 people got the disease. Is the claim statistically significantly justified?

Sketch- not drawn to scale



Is the 0.57% rare?

P-value question: What is the chance that a random group of 350 people would have only 2 people with the disease?

If there is a good chance that it was an accident, then I do *not* want to buy that drug.

If there was a tiny chance that it was by accident, then I am going to pay attention to the drug.

IN THIS CASE,
THE DRUG COMPANY WANTS THE SAMPLE TO BE RARE.
THEN THEY CAN ASSUME SOMETHING AFFECTED THE SAMPLE, WHICH THEY HOPE WAS TAKING THE DRUG.

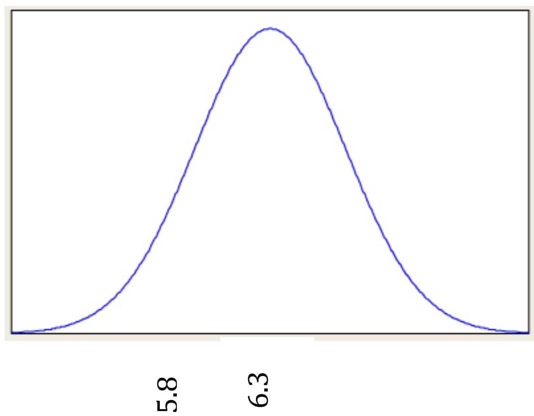
Worksheet for example B on next page!

General Steps for a percentage			This example
Step 1: Determine H_0 and H_1			H_0: H_1:
Step 2: Find z using the formula	$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$	$\hat{p}$ = sample percent p = population percent n = size of sample	z=
Step 3: Find the P-value from z	Use a table (use z to probability table from before)	If z<0 p-value= probability If z>0 p-value= 1-probability	p-value=
Step 4: Draw your conclusion	Reject H_0 if p-value ≤ 0.05 Fail to reject H_0 if p-value > 0.05	Conclusion: (state using the context of the example)	

Example C: According to the US Department of Education, it takes on average 6.3 years to earn a bachelor's degree.

81 recent college graduates from UMass Amherst were asked how long it took them to finish their bachelor's degree. The sample has a mean of 5.8 years and a standard deviation of 2.2 years. The newspaper claimed that it takes less time on average to finish a bachelor's degree at UMass Amherst than other US 4-year institutions. Is that claim statistically significantly justified?

Sketch- not drawn to scale



Is 5.8 years rare?

P-value question: If it takes an average of 6.3 years for all college students, then what are the chances that a randomly selected group of 81 recent graduates would average 5.8 years?

If there is a small chance, then we would assume something influenced the sample, hopefully going to UMass Amherst.

If there is a big chance, then randomly getting a group to average 5.8 is likely and thus UMass Amherst does not seem to graduate students more quickly.

THE NEWSPAPER WANTS 5.8 TO BE RARE. SO THE CLAIM WOULD BE JUSTIFIED

Worksheet for example C on the following page.

General Steps for a mean (population σ unknown) ONLY IF n>30 OR THE POPULATION IS NORMAL			This example
Step 1: Determine H_0 and H_1			H_0: H_1:
Step 2: Find t using the formula	$t = \frac{\hat{x} - \mu}{\frac{s}{\sqrt{n}}}$	$\hat{x}$ = sample mean μ = population mean s =sample standard deviation n = size of sample Degrees of freedom $= n - 1$	t=
Step 3: Find the p-value from t http://www.socsciststistics.com/pvalues/tdistribution.aspx	Use an online calculator	One-tailed if $H_1 > \#$ or if $H_1 < \#$ Two-tailed if $H_1 \neq \#$ Type in the t value and the degrees of freedom	p-value=
Step 4: Draw your conclusion	Reject H_0 if p-value ≤ 0.05 Fail to reject H_0 if p-value > 0.05		Conclusion: (state using the context of the example)

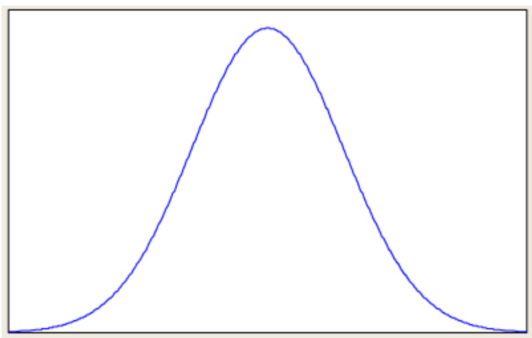
Example D:

Claim: the height of supermodels varies less than the general female population.

The female population has a standard deviation of 2.5 inches in their height.

Random sample information: 9 super model heights have a mean of 70.0 inches with a standard deviation of 1.5 inches.

Sketch- not drawn to scale



1.5 2.5

Is 1.5 inches rare?

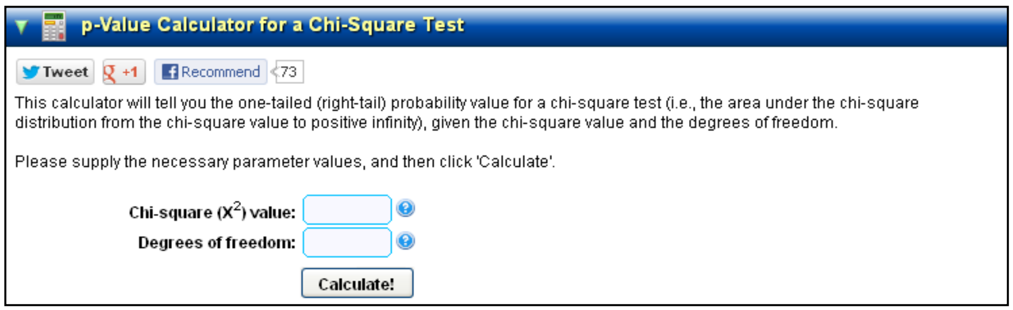
P-value question: What are the chances of randomly selecting 9 females who have a standard deviation of their height of 1.5 inches?

If the chance is small, then I would assume there is less variation in supermodel heights.

If the chance is large, then there is a good chance that the lower σ was by coincidence.

THE PEOPLE MAKING THE CLAIM WANT 1.5 TO BE RARE. THEN THEY CAN ASSUME SOMETHING AFFECTED THE SAMPLE, IN THIS CASE, THE FACT THAT THEY ARE SUPER MODELS AFFECTED THE SAMPLE.

Worksheet for example D on the following page.

General Steps for standard deviation - THE POPULATION MUST BE NORMAL	This example
Step 1: Determine H_0 and H_1	H_0: H_1:
Step 2: Find χ^2 using the formula {Pronounced KAI-squared Hard C like in cat then EYE} $\chi^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ σ = population standard deviation s =sample standard deviation n = size of sample Degrees of freedom = $n - 1$	χ^2=
Step 3: Find the p-value from χ^2 Use an online calculator One-tailed if $H_1 >$ # or if $H_1 <$ # Two-tailed if $H_1 \neq$ #	p-value=
if $H_1 >$ # or if $H_1 <$ # http://www.danielsoper.com/statcalc3/calc.aspx?id=11  Type in the chi-squared value and the degrees of freedom and click on Calculate! If $H_1 <$ # then your P-value is (1 – probability) If $H_1 >$ # then your P-value is the probability	Conclusion: (state using the context of the example)
Step 4: Draw your conclusion Reject H_0 if p-value ≤ 0.05 Fail to reject H_0 if p-value > 0.05	

Errors in Hypothesis Testing

Statements you must hear in a statistics class, but will only use if you take future statistics courses.

Type I error: reject H_0 when H_0 is true (also known as a false positive)

Type II error: fail to reject H_0 when H_0 is false (also known as a false negative)

Practical applications of these *ideas* are very, very important. Probably the most relevant application for you will be with percentages.

Example E: A certain test is 99.7% accurate. (Seems really good, right?) Of the inaccurate results, 70% are false positives and 30% are false negatives. If 5,000,000 people are tested, how many people are told they have the disease but do not? And how many people are told they do not have the disease but do?

Fill in the blanks with the numbers of people who fit each category.

Tested______________

Test was accurate _____________

Test was wrong _______________

Test said have it but don't

_______________ (false positive)

Test said don't have it but do

______________ (false negative)

Consequences:

False positives (told they have it but don't) may cause added stress and more testing.

False negatives (told they do not have it but do) may lead to missed early treatment.

Comparing two samples – Notes

Inference includes confidence intervals and hypothesis testing

ALL PERCENTAGES MUST BE WRITTEN IN DECIMAL FORM IN THE FORMULAS

Hypothesis testing for a percent (a proportion) using two samples

SAMPLES MUST BE INDEPENDENT AND HAVE AT LEAST 5 "YES" AND 5 "NOT YES" FROM EACH SAMPLE

Step 1: Determine H_0 and H_1	$H_0\colon\ p_1 = p_2$ $H_1\colon p_1 > p_2 \ \ or\ \ p_1 < p_2$ H_1 is whatever your data supports	
Step 2: Find z using the formula	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\bar{p}(1-\bar{p})}{n_1} + \dfrac{\bar{p}(1-\bar{p})}{n_2}}}$	$\hat{p}_1$ = sample percent of one sample $\hat{p}_2$ = sample percent of other sample n_1 = size of one sample n_2 = size of other sample $\bar{p} = \dfrac{true\ in\ sample\ 1 + true\ in\ sample\ 2}{n_1 + n_2}$
Step 3: Find the P-value from z	Use a table If z is positive then P=1-probability If z is negative then P=probability	Technically, if z>2 or z<-2 then the event is rare and you do not need to find the P-value (this is because we are using a 95% test)
Step 4: Draw your conclusion	Reject H_0 if p-value ≤ 0.05 Fail to reject H_0 if p-value > 0.05	

Interpreting the results:

If $p_1 = p_2$ is a rare event then we can assume that one of them is bigger than the other.
The bigger one in your data is assumed to be the bigger one in the population.

If $p_1 = p_2$ is NOT a rare event then we canNOT assume anything. They might be equal, or one might be bigger than the other, but we don't know which one is bigger.

Confidence Interval Estimate of $p_1 - p_2$ for comparing two samples

Confidence interval for a difference of two sample percents is $(\hat{p}_1 - \hat{p}_2) \pm E$ where E is the estimated error

$(\hat{p}_1 - \hat{p}_2) \pm E$ $E = z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$\hat{p}_1$ = sample percent of one sample $\hat{p}_2$ = sample percent of other sample n_1 = size of one sample n_2 = size of other sample E = estimated error

If want a 90% confidence interval $z_{\alpha/2}$ = 1.645	If want a 95% confidence interval $z_{\alpha/2}$ = 1.96	If want a 99% confidence interval $z_{\alpha/2}$ = 2.575

Interpreting the results: If the error is bigger than the difference, then we cannot assume anything.
If the error is smaller than the difference, then we can assume one of them is bigger.

Hypothesis testing for two sample means that are independent (population σ's unknown)

ONLY IF n>30 OR THE POPULATION IS NORMAL FOR BOTH SAMPLES

The process for independent means:

Step 1: Determine H_0 and H_1	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2 \quad or \quad \mu_1 < \mu_2$ whatever your data supports	μ_1 =population mean for sample 1 μ_2 =population mean for sample 2
Step 2: Find t using the formula	$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\overline{x_1}$ =sample 1 mean $\overline{x_2}$ =sample 2 mean s_1 =sample 1 standard deviation s_2 =sample 2 standard deviation n_1 = size of sample 1 n_2 = size of sample 2 Degrees of freedom is the smaller of $n_1 - 1$ and $n_2 - 1$
Step 3: Find the P-value from t	Use an online calculator See example C on page 119	One-tailed if $H_1 > \#$ or if $H_1 < \#$ Two-tailed if $H_1 \neq \#$
Step 4: Draw your conclusion	Reject H_0 if p-value ≤ 0.05 Fail to reject H_0 if p-value > 0.05	

Interpreting the results:

If $\mu_1 = \mu_2$ is a rare event then we can assume that one of them is bigger than the other. The bigger one in your data is assumed to be the bigger one in the population.

If $\mu_1 = \mu_2$ is NOT a rare event then we canNOT assume anything. They might be equal, or one might be bigger than the other, but we don't know which one is bigger.

Confidence Interval Estimate of $\mu_1 - \mu_2$ for comparing two independent samples (population σ's unknown)

Confidence interval for a difference of two sample means is $(\bar{x}_1 - \bar{x}_2) \pm E$ where E is the estimated error

$(\bar{x}_1 - \bar{x}_2) \pm E$ $E = t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\overline{x_1}$ =sample 1 mean $\overline{x_2}$ =sample 2 mean s_1 =sample 1 standard deviation s_2 =sample 2 standard deviation n_1 = size of sample 1 n_2 = size of sample 2 Degrees of freedom is the smaller of $n_1 - 1$ and $n_2 - 1$ To find $t_{\propto/2}$ see page 101

Interpreting the results: If the error is bigger than the difference, then we cannot assume anything.

If the error is smaller than the difference, then we can assume one of them is bigger.

Hypothesis testing for two sample means that are dependent (population σ's unknown)

ONLY IF n>30 OR THE POPULATION IS NORMAL FOR BOTH SAMPLES

We will use this for paired data – namely change over time.

The process for dependent means:

Step 1: Determine H_0 and H_1	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2 \;\; or \;\; \mu_1 < \mu_2$ whatever your data supports	μ_1 =population mean for sample 1 μ_2 =population mean for sample 2
Step 2: Find t using the formula	$t = \dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$	$\bar{d}$ =the mean difference Find the difference between the two numbers in each pair. Average those differences. s_d =the standard deviation of the difference n = the number of PAIRS of data (30 in part A and 30 in part B means 60 data points but only 30 PAIRS) Degrees of freedom = $n - 1$
Step 3: Find the P-value from t	Use an online calculator See example C on page 119	One-tailed if $H_1 > \#$ or if $H_1 < \#$ Two-tailed if $H_1 \neq \#$
Step 4: Draw your conclusion	Reject H_0 if p-value ≤ 0.05 Fail to reject H_0 if p-value > 0.05	

Interpreting the results:

If $\mu_1 = \mu_2$ is a rare event then we can assume that one of them is bigger than the other.
Use bigger one in your data is assumed to be the bigger one in the population.

If $\mu_1 = \mu_2$ is NOT a rare event then we canNOT assume anything. They might be equal, or one might be bigger than the other, but we don't know which one is bigger if they are not equal.

Confidence Interval Estimate of $\mu_1 - \mu_2$ for comparing two dependent samples (population σ's unknown)

Confidence interval for a difference of two sample means is $\bar{d} \pm E$ where E is the estimated error

$\bar{d} \pm E$ $E = t_{\alpha/2} \dfrac{s_d}{\sqrt{n}}$	$\bar{d}$ =the mean difference Find the difference between the two numbers in each pair. Average those differences. s_d =the standard deviation of the difference n = the number of PAIRS of data Degrees of freedom = $n - 1$ To find $t_{\propto/2}$ see page 101

Interpreting the results: If the error is bigger than the difference, then we cannot assume anything.
If the error is smaller than the difference, then we can assume one of them is bigger.

Comparing two samples – Worksheet

Inference includes confidence intervals and hypothesis testing

All examples use the data provided on page 133 referring to a weight loss study.

Hypothesis testing for a percent (a proportion) using two samples

Example: Percentage of women who lost weight versus the percentage of men who lost weight.
In the sample, more women than men lost weight so we are trying to say that at all college campuses, more freshman women lose weight than freshman men.

SAMPLES MUST BE INDEPENDENT AND HAVE AT LEAST 5 "YES" AND 5 "NOT YES" FROM EACH SAMPLE

Step 1: Determine H_0 and H_1 (no numbers)	H_0: $p_1 = p_2$ H_1: circle one $p_1 > p_2$ $p_1 < p_2$	In words, explain what each of the following mean in this context? p_1 p_2 H_0: $p_1 = p_2$ Your H_1
Step 2: Find z using the formula	$z =$	$\hat{p}_1 =$ $\hat{p}_2 =$ $n_1 =$ $n_2 =$ $\bar{p} =$ $z =$
Step 3: Find the p-value from z	p=	
Step 4: Draw your conclusion and interpret your results in the context of the example		

Confidence Interval Estimate of $p_1 - p_2$ for comparing two samples

Example: Percentage of women who lost weight versus the percentage of men who lost weight.
In the sample, more women than men lost weight so we are trying to say that at all college campuses, more freshman women lose weight than freshman men. Let's use a 99% confidence interval.

Confidence interval for a difference of two sample percents is $(\hat{p}_1 - \hat{p}_2) \pm E$ where E is the estimated error

In words, what do the following variables represent? $\hat{p}_1 =$ $\hat{p}_2 =$ $n_1 =$ $n_2 =$	
Find the error in the difference $E =$	Numerical values $\hat{p}_1 =$ $\hat{p}_2 =$ $n_1 =$ $n_2 =$ E =

Interpret the result using the context of the example.

<u>Hypothesis testing for two sample means that are independent (population σ's unknown)</u>

Example: Amount of weight women gained on average versus the amount of weight men gained on average

In the sample, women gained more weight on average than the men so we are trying to say that at all college campuses, freshman women gain more weight than freshman men.

ONLY IF n>30 OR THE POPULATION IS NORMAL FOR BOTH SAMPLES Is this true?_________

Step 1: Determine H_0 and H_1 (no numbers)	$H_0: \mu_1 = \mu_2$ H_1: circle one $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	In words explain what each of the following mean in this context? μ_1 μ_2 $H_0: \mu_1 = \mu_2$ Your H_1
Step 2: Find t using the formula	$t =$	$\overline{x_1} =$ $\overline{x_2} =$ $s_1 =$ $s_2 =$ $n_1 =$ $n_2 =$ Degrees of freedom=
Step 3: Find the P-value from t	p=	
Step 4: Draw your conclusion and interpret your results in the context of the example		

Confidence Interval Estimate of $\mu_1 - \mu_2$ for comparing two independent samples (population σ's unknown)

Example: Amount of weight women gained on average versus the amount of weight men gained on average

In the sample, women gained more weight on average than the men so we are trying to say that at all college campuses, freshman women gain more weight than freshman men. Let's use a 90% confidence interval.

Confidence interval for a difference of two sample means is $(\bar{x}_1 - \bar{x}_2) \pm E$ where E is the estimated error

In words, what do the following variables represent? $\overline{x_1} =$ $\overline{x_2} =$ $s_1 =$ $s_2 =$ $n_1 =$ $n_2 =$	
Find the error in the difference $E =$	Numerical values $\overline{x_1} =$ $\overline{x_2} =$ $s_1 =$ $s_2 =$ $n_1 =$ $n_2 =$ Degrees of freedom =

Interpret the result using the context of the example.

<u>Hypothesis testing for two sample means that are dependent (population σ's unknown)</u>

We will use this for paired data – namely *change over time*.

Example: The average weight in September versus the average weight in April for both genders together.
The data says the average weight went up, but is the data strong enough to say that it is expected that college freshman will gain weight?

ONLY IF n>30 OR THE POPULATION IS NORMAL FOR BOTH SAMPLES Is this true?_________

Step 1: Determine H_0 and H_1 (no numbers)	H_0: $\mu_1 = \mu_2$ H_1: circle one $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	In words explain what each of the following mean in this context? μ_1 μ_2 H_0: $\mu_1 = \mu_2$ Your H_1
Step 2: Find t using the formula	$t =$	$\bar{d} =$ $s_d =$ $n =$ Degrees of freedom =
Step 3: Find the P-value from t	p=	
Step 4: Draw your conclusion and interpret your results in the context of this example.		

Confidence Interval Estimate of $\mu_1 - \mu_2$ for comparing two dependent samples (population σ's unknown)

We will use this for paired data – namely *change over time*.

Example: The average weight in September versus the average weight in April for both genders together. The data says the average weight went up, but is the data strong enough to say that it is expected that college freshman will gain weight? Let's use a 95% confidence interval.

Confidence interval for a difference of two sample means is $\bar{d} \pm E$ where E is the estimated error

In words, what do the following variables represent? $\bar{d} =$ $s_d =$ $n =$	
Find the error $E =$	Numerical values $\bar{d} =$ $s_d =$ $n =$ Degrees of freedom =

Interpret the result using the context of the example.

Comparing Two Samples Worksheet Data

The Freshman 15 is the theory that on average college freshman gain 15 pounds in their first year of college. Researchers were interesting in finding out if the "The Freshman 15" is true. This is a partial summary of the data they collected. Weight is in pounds.

ALL	**Weight - SEPTEMBER**	**Weight - APRIL**	difference
mean	143.43	146.03	2.60
median	141.10	145.50	4.41
mode	141.10	149.91	2.20
st dev	24.88	24.88	8.57
co var	17.35	17.04	329.77
WOMEN	**Weight - SEPTEMBER**	**Weight - APRIL**	difference
mean	127.99	130.64	2.65
median	125.66	127.87	4.41
mode	123.46	149.91	4.41
st dev	14.02	12.93	5.97
co var	10.96	9.89	225.79
MEN	**Weight - SEPTEMBER**	**Weight - APRIL**	difference
mean	160.32	162.87	2.55
median	156.53	156.53	3.31
mode	163.14	152.12	2.20
st dev	23.21	23.97	10.83
co var	14.48	14.72	424.90
	st dev= sample standard deviation		
	co var=coefficient of variation		

Number of participants: 67 Number of female participants: 35 Number of male participants: 32	Number of people who lost weight: 17 Number of females who lost weight: 10 Number of males who lost weight: 7

Paper 4: Sample to sample comparison

Part I: Look over the student survey data. This is the same data as paper 3 (direct link on Google site) https://docs.google.com/spreadsheets/d/1LxFHnFUwJOrZcrlS6GEsWXGbzXAO6WPe-iHt4nVogs0/edit?usp=sharing

There is a lot of data here so you will need to carefully choose which questions you would like to analyze. Look over the questions and pick a topic that stands out to you or is interesting to you and find 5-10 questions that are directly related to that topic. This may be the same topic and/or groups as paper 3 or paper 4.

Choose two or more groups within this data to compare.

Sort and separate the data into these two groups and delete the data for the students who do not fall into either group and the questions that are not related to your topic.

Come up with a name for the university represented in this data.

Part II: Analyze the data
Analyze and compare the descriptive statistics for each group for each of the 5-10 questions you chose. Come up with a thesis statement- a persuasive statement you will prove in your article that is about every student at this university who fits into the two groups, not just the ones who took the survey.

Part III: Sample to sample tests
As you are analyzing the data, do several sample to sample tests that might be helpful in your argument. You will needs to include at least 4 sample to sample tests that are analyzing at least 2 different types of descriptive statistics.

Part IV: Write an article
The article must be a persuasive article trying to convince the reader of your thesis statement about comparing two groups of students at this university.
It must be 500-1000 words, include at least 7 statistics and at least 4 sample to sample tests with appropriate accompanying statistics (that are analyzing at least two different types of statistics), and include at least 1 graph.

Rubric for Paper 4

Checklist prior to grading. Paper will not be graded until all 3 requirements are met.
__________Article is typed
__________Article is 500-1000 words
__________Article is submitted electronically and a paper copy was handed in with this grading sheet

	Points
There is a clear and relevant main point.	/5 points
At least 7 (but hopefully more) stats were used in the article and they were at least 3 different kinds. The result of at least 4 sample to sample tests are included (of at least 2 different kinds). The stats used in the article supported the point and were included appropriately. Each statistic used in the article was explained or interpreted clearly and in depth.	/50 points
At least one computer generated graph was used in this article. The graph(s) supports the point of the article. The take-away from the graph is referenced in the article (even if the graph itself is not mentioned).	/30 points
The article was well written and easy to understand.	/15 points
	Total /100 points

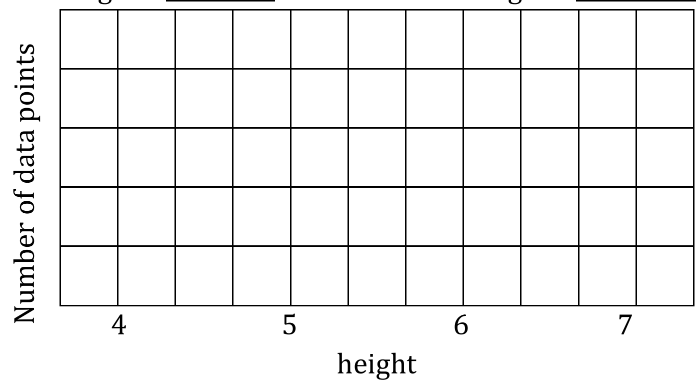
Probabilities of Groups Worksheet – conceptual prep

There are 5 people in this population. Their names and heights were recorded in the chart.

Amy is 4 feet tall	Brad is 5 feet tall	Carlos is 5 feet tall	Diamond is 6 feet tall	Elmer is 7 feet tall

Part I: looking only at individuals

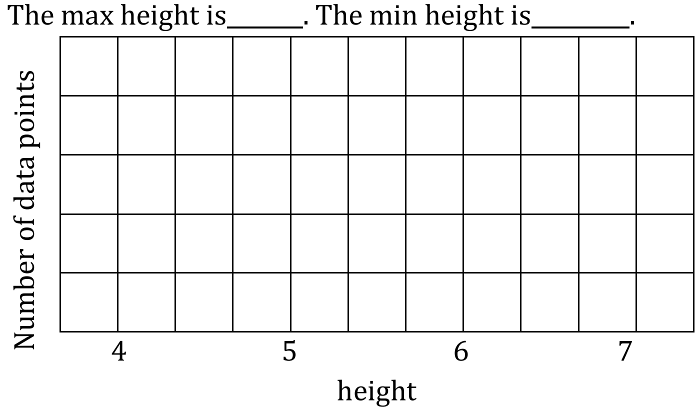
1. The mean height is __________. The maximum height is_________. The minimum height is__________
2. Create a frequency graph



Part II: Looking at groups of size three. Fill in the chart for every possible group of three

Names (initials)										
Heights										
Average height of the group of 3										

1. The mean height for a group is _________. The max height is______. The min height is________.
2. Create a frequency graph



Comparing the individual to the group:

The mean of the groups is ____________________________ the standard deviation of the individuals.

The standard deviation of the groups is __________________________ the standard deviation of the individuals.

The maximum of the groups is _____________________________ the maximum of the individuals.

The minimum of the groups is ______________________________ the minimum of the individuals.

Probabilities of Groups of People

Up to this point we have talked a lot about probabilities of individuals obtaining a certain value.

For example: The probability that a US male will be 5'11"
The probability that an HCC student will take statistics
The probability that a NASA flight lasts between 250 and 300 hours
The probability that a pregnancy with twins lasts between 32 and 35 weeks

All of these examples are talking about the probability that a single randomly selected member of the population will have a given characteristic.

But what about this example?

A ferry (for people only) has a weight limit of 10,000 pounds, but since it would be bad for business to weigh every passenger before boarding, ferry limits are often converted to a rule about the maximum number of people allowed on board. How many people should be allowed on the ferry?

Now we are not talking about one individual, but rather we are talking about a group of people. We want to the know the probability that a randomly selected group of people will weigh more than 10,000 pounds.

This is called a sampling distribution. And the **Central Limit Theorem** tells us: (7.1 in TEXT)

The technical wording	What it means for us
If we take all of the samples all of the same size from a population, assuming the samples are large enough, then	ONLY IF n>30 n=the size of the group or THE POPULATION IS NORMAL
1. The means of those samples a) will follow a normal curve b) will have a mean that will approach the population mean	We will use the population mean as an approximation the sample of groups mean. $\mu_{population} \approx \mu_{sample\ of\ groups}$ If the population mean is unknown, we will use the mean of our largest sample.
2. The standard deviations of those samples a) are skewed b) will have a standard deviation that is a slightly biased approximation of the population standard deviation.	We will use the standard deviation of the population to approximate the standard deviation of the sample of the groups using the following formula. $\sigma_{sample\ of\ groups} \approx \frac{\sigma_{population}}{\sqrt{n}}$ where n is the size of the group If the population standard deviation is unknown, we will use the standard deviation of our largest sample.

Example 1: What is the probability that using a maximum of 56 people for the 10,000 pound limit will result in a sinking ship?

Note: the weight of adult males has a mean of 172 pounds and a standard deviation of 29 pounds. It is common in situations like this to use the male information since the average weight of males is higher than the average weight for females.

Step 1: Does this example fit our restrictions? ________

(Is the data normal AND/OR is the group size larger than 30?)

Step 2: What is the sample of groups mean and sample of groups standard deviation?

$\mu =$__________ $\sigma =$___________ (note: the standard deviation is not 29)

Step 3: For the boat to sink carrying 56 people,

the average weight per person must be at least__________

Step 4: What is the z-score associated with that average?

Step 5: What is the probability of the average weight of a randomly selected group of 56 people being at or above the average found in step 3?

Example 2: Redwood trees in California grow very quickly. The trees that are approximately 2000 years old can have a circumference as large as 63 feet. A tour guide for an all women's hiking company nicknamed King likes to take her group on a hike that ends at one of these huge trees. How large of a group should she take if she wants to make sure that the vast majority of the time, the group of women cannot group-hug the tree and reach all the way around, but at the same time, the more people per group, the money she makes?
Assume wingspan and height are the same. The mean height for females is 63.2 inches with a standard deviation of 2.74 inches. Note: the height of women is normal data

a) Find the probability that a group of 11 women can reach around the tree.

b) Find the probability that a group of 12 women can reach around the tree.

c) Would you recommend King take a group of 11 women or 12 women? Explain your answer.

Example 3: The amount of baggage a plane passenger checks is random, with a mean of 20 lbs and a standard deviation of 30 pounds. A plane which carries 100 passengers can handle 3000 pounds of checked baggage.

What is the probability that a random load of 100 passengers will check too much baggage for the plane to handle?

Example 4: From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for a soccer game between Boston College and Boston University follows a distribution that has a mean of 2.4 and standard deviation of 2.

Suppose that few hours before the start of one of these games there are 100 eager students standing in line to purchase tickets. If only 250 tickets remain, what is the probability that all 100 students will be able to purchase tickets?

Math 142 Final Exam Review Sheet A

Question numbers correspond to questions on review sheet B

1. Explain what the *mean* tells you.
2. Explain what the *median* tells you.
3. Explain what the *mode* tells you.
4. Explain what the *standard deviation* tells you.
5. Explain what *quartiles* tell you.

6. Explain what the *range* tells you.
7. Explain what the *coefficient of variation* tells you.
8. Explain what *outliers* are.
9. Explain what *percentiles* tell you.
10. Explain the *law of large numbers*.

11. Explain the relationship between a *population* and a *sample*.

12. Explain what a *z-score* is.

13. Explain what *linear regression* is and what it tells you.

14. When finding the characteristic of a group of people, why does the standard deviation decrease?

15. Give an example of a *biased* question. Why is it biased?

16. Explain what it means to have a *random sample*.

17. Explain what a 95% *confidence interval* is.

18. Explain what a *Type I error* is. Also known as a *false-positive* result.

19. Explain what a Type II error is. Also known as a *false-negative* result.

Math 142 Final Exam Review Sheet B

Concepts in Context

Question numbers correspond to review sheet A

1. Data was collected recording the number of cars on campus at 2:30 in the afternoon every day of the week from Feb 1st until May 15th. The mean was 89.43 Explain what this means in the context provided?

2. MCAS scores from across the state were collected and the median of the data was 239.5 Explain what this means in the context provided. Below is an explanation of the scores.

Advanced (260-280)	Students at this level demonstrate a comprehensive and in-depth understanding of rigorous subject matter and provide sophisticated solutions to complex problems.
Proficient (240-259)	Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems.
Needs Improvement (220-239)	Students at this level demonstrate a partial understanding of subject matter and solve some simple problems.
Warning/Failing (200-219)	Students at this level demonstrate a minimal understanding of subject matter and do not solve simple problems.

3. Data was collected recording the number of students in class on Fridays at 3:30 for the entire spring semester. The mode was 736. Explain what this means in the context provided?

4. Refer to the MCAS chart on previous page. For a certain school district the mean was 257 and the standard deviation was 20. Explain what this means in the context provided.

5. The number of earned runs for J. Raymond by inning was recorded and the following quartiles were computed. 0-0 0-0 1-2 3-9
Explain what this means in the context provided. Note: pitchers want a low number of earned runs.

6. When counting the number of traffic stops per county last Saturday it was found that the minimum was 1 and the maximum was 237. Explain what the range means in the context provided.

7. Without finding the coefficient of variation, explain which set of data has a higher coefficient of variation and how do you know?
Data: The weight of baby boys. Mean is 8 pounds with a standard deviation of 1.9 pounds.
Data: The cost per cubic yard of mulch. Mean is $35 with a standard deviation of $3.

8. The number of ice cream cones sold per day over a two week period is as follows:
   5 5 8 6 4 5 39 12 7 0 6 8 4 9
   Would you consider any of these outliers? What might explain the difference in the number of cones sold per day?

9. Look at the table below and answer the following questions: Note: quintile means 5 groups.
   a) Explain the line "Lowest quintile 4.3 percent"

   b) Explain the last line "Top 0.01 percentile 31.5%"

   c) What is the tax rate for the people in the 87th percentile of income?

Total Effective Federal Tax Rates for 2005 by income level

INCOME LEVEL	TAX RATE
Lowest quintile	4.3 percent
Second quintile	9.9 percent
Middle quintile	14.2 percent
Fourth quintile	17.4 percent
Percentiles 81-90	20.3 percent
Percentiles 91-95	22.4 percent
Percentiles 96-99	25.7 percent
Percentiles 99.0-99.5	29.7 percent
Percentiles 99.5-99.9	31.2 percent
Percentiles 99.9-99.99	32.1 percent
Top 0.01 Percentile	31.5 percent

10. The law of large numbers essentially says the bigger the sample the better. Give an example of two samples where the bigger sample is not as good as the smaller sample.

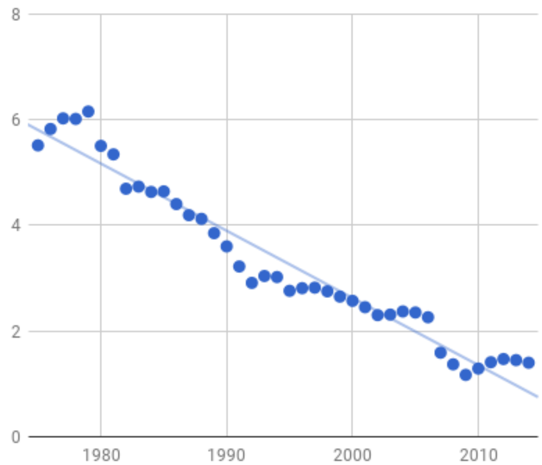
11. Fill in the blanks. Note: there are numerous answers for each blank.

a) sample is all male HCC students, population is ________________ or ________________

b) population is all MA residents, sample is ________________ or ________________

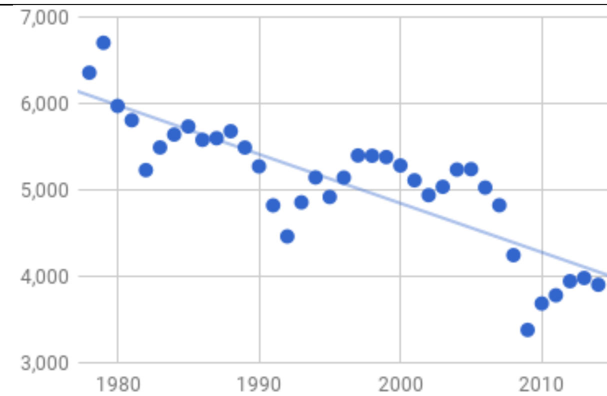
12. The age of the dolphin in a Florida aquarium is 37 years old. The z-score for that data point is 2.4 Explain what this means in the context provided.

13. Linear regression models



Year versus rate of fatalities in crashes involving large trucks per 100 million miles driven

$y=-0.127x+257$ $r^2=0.949$



Year versus number of fatalities in crashes involving large trucks

$y=-56.7x+118274$ $r^2=0.0689$

a) What is the difference in the information displayed in these two graphs?

b) Which line is a better model for the given data and why?

c) Use both models to predict what happened in 1997, what will happen in 2015 and what will happen in 2030. Are these good predictions? Why or why not?

14. NO CONTEXT QUESTION FOR THIS TOPIC

15. You want to find out which item people like more, ice cream or frozen yogurt.
   a) Write a question that is biased towards ice cream.

   b) Write a question that is biased towards frozen yogurt.

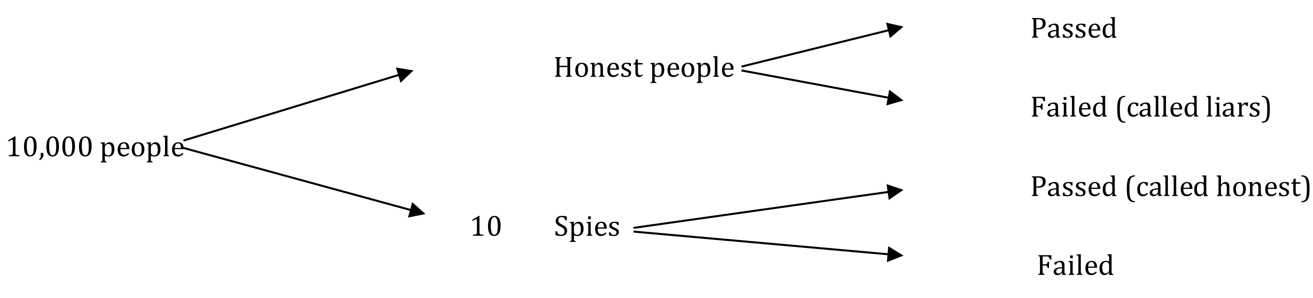
   c) Write a question that is not biased.

16. A company wants to evaluate an employee who works with clients. a) What would be a fair way to select which clients to speak with about their experience with the employee?

   b) Why are optional comment cards an unfair way to evaluate employees?

17. NO CONTEXT QUESTION FOR THIS TOPIC

18. and 19. Fill in the following chart if the Type I error: false-positive rate for honest people is 16% and the Type II error: false-negative rate for spies is 20% on a standard lie detector test.



What do these results mean in the context provided?

Math 142 Final Exam Review Sheet C

Use the data from the Uniform Crime Report published by the FBI for reporting colleges and universities in Massachusetts and Connecticut for 2011. Data starts on page 163.
Use the calculations done for you and use other easier calculations you do yourself (mainly percentages.)

1. a) Which school appears to be the most dangerous school on the list?
b) Find at least 6 statistics that would support your choice.

2. a) Which school appears to be the safest school on the list?
b) Find at least 6 statistics that would support your choice.

3. a) Do community colleges or non-community colleges appear to be more dangerous?
b) Find at least 6 statistics that would support your choice.

4. a) Do Massachusetts colleges and universities or Connecticut colleges and universities appear to be more dangerous?
b) Find at least 6 statistics that would support your choice.

Math 142 Exam Review Sheet D

Use the data from the Uniform Crime Report published by the FBI for reporting colleges and universities in Massachusetts and Connecticut for 2011. Data start on page 163.
Use the calculations done for you and use other easier calculations you do yourself (mainly percentages.)

Decide if each of the following 15 statements is a correct interpretation of the data or an incorrect interpretation. If the statement is an incorrect interpretation, say why it is incorrect or say what the correct interpretation is.

1. According to the data, the enrollment of all MA colleges/universities are 8,094 $\pm$ 7500.

2. According to the data, the larceny-theft crime rate for US community colleges is 0 to 47.

3. According to the data, when drawing conclusions about US colleges/universities, Harvard, Northeastern and UMass Amherst should not be included because they are too big.

4. According to the data, there is a 28% chance that MA colleges/universities have the same property crime rate as CT colleges/universities.
5. According to the data, the average number of robberies expected on a 10,000 student campus is between 0 and 2.
6. According to the data, one-fourth of colleges/universities have no violent crime.

7. According to the data, the probability of a college having arson and rape in the same year is 12.5%.

8. 32.5% of colleges/universities have robbery and 60% have aggravated assault. According to the data, 92.5% of them have robbery or aggravated assault.

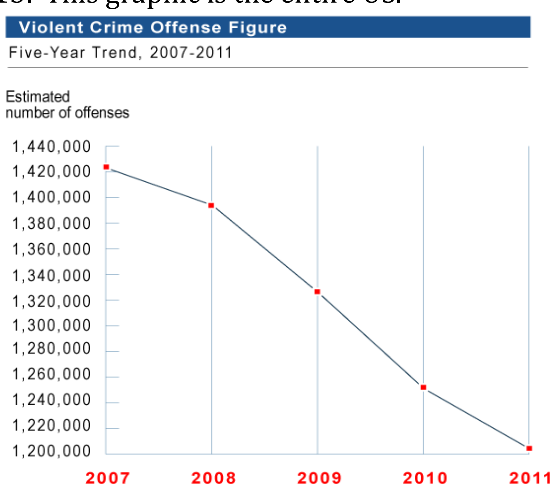
9. According to the data, only ½ of all community colleges have violent crime as compared to 80.6% of non-community colleges.

10. The z-score for a rape rate of 8 per year is 2.0814 which has the associated probability of 98.13% According to the data, a rape rate of 8 happens at 98% of the colleges/universities in the US.
11. According to the data, it is rare for a motor vehicle theft rate to be above 3.4.
12. In state ranking, MA is higher than 12 of the 50 states for their violent crime rate. According to the data, MA is in the 26th percentile of the states in the US for violent crime rate.

13. 507 out of 4782 MA property crimes occurred at Harvard. According to the data, Harvard ranks in the 11th percentile in the US for property crimes.

14. According to the data and because the z-score is negative, it is very unlikely that 3 randomly chosen MA colleges/universities will have a total of 300 property crimes.

15. This graphic is the entire US.



According to the graph, the number of violent crimes is greatly decreasing and we can expect almost complete elimination of violent crime in the US in the next few years.

16. List limitations of the crime data used in this worksheet.

17. List some possible reasons for why the crime rates vary so much from school to school.

CRIME DATA

	Student enrollment[1]	**Violent crime rate**	Murder rate	Forcible rape rate	Robbery rate	Aggravated assault rate	**Property crime rate**	Burglary rate	Larceny-theft rate	Motor vehicle theft rate	Arson rate[2]
MASSACHUSETTS											
mean	8,094	6	0	2	1	3	145	14	130	1	1
median	5,684	5	0	1	0	2	122	11	111	0	0
mode	#N/A	0	0	0	0	0	#N/A	0	#N/A	0	0
standard deviation	7500.452837	5.898515	0	2.88263948	1.66766543	3.9256048	96.045179	13.1484252	90.653736	1.20129802	4.377037
range	28,253	24	0	12	8	16	421	51	408	4	18
95% confidence interval (mean)	8063 to 8124	3.6 to 8.4	0	0.82 to 3.18	0.32 to 1.68	1.39 to 4.6	106 to 184	8.6 to 19.4	93 to 167	0.51 to 1.49	0.8 to 2.8
MA COMMUNITY COLl.											
mean	9,138	1	0	0	0	1	43	0	41	1	0
median	8,439	1	0	0	0	1	44	0	43	1	0
mode	#N/A	0	0	0	0	0	#N/A	0	#N/A	#N/A	#VALUE!
standard deviation	2176.532315	1.324580	0	0	0	1.3245806	21.531546	0	22.132483	1.04563157	0
range	4,867	3	0	0	0	3	46	0	46	3	0
95% confidence interval (mean)	28125 to -9849	3.8 to -1.8	0	0	0	3.8 to -1.8	88 to -2	0	87 to -5	3.2 to -1.2	0
MA NON-COMMUNITY COLLEGES											
mean	7,950	7	0	2	1	4	159	16	142	1	1
median	5,364	6	0	2	0	2	134	12	119	0	0
mode	#N/A	0	0	0	0	0	#N/A	#N/A	#N/A	0	#VALUE!
standard deviation	7975.505952	5.942521	0	2.96086689	1.75495455	4.0768944	93.756735	12.9065014	89.748126	1.22840732	4.625802
range	28,253	24	0	12	8	16	409	51	408	4	18
95% confidence interval (mean)	4443 to 11457	4.4 to 9.6	0	0.7 to 3.3	0.23 to 1.8	2.2 to 5.8	118 to 200	10.3 to 21.7	103 to 182	0.46 to 1.54	3 to -1.03

	Student enrollment[1]	**Violent crime rate**	Murder and nonnegligent manslaughter rate	Forcible rape rate	Robbery rate	Aggravated assault rate	**Property crime rate**	Burglary rate	Larceny-theft rate	Motor vehicle theft rate	Arson rate[2]
community college vs non MA											
hypothesis testing (mean=mean)											
t	0.64646	-4.5269		-4.3856	-2.7186	-2.5140	-5.6889	-6.5566	-5.0570	0.75278	-1.7410
p-value	0.28201	0.01008		0.01097	0.03631	0.04331	0.00538	0.00360	0.00746	0.25313	0.09001
MA vs CT											
hypothesis testing (mean=mean)											
t	-0.8716	3.07763		1.58729	-0.3263	4.48716	0.61149	-0.0925	0.71195	-0.9627	0.17830
p-value	0.20845	0.01086		0.08177	-0.8199	0.00208	0.28165	0.46463	0.25161	0.18643	0.43217
Groups of MA colleges so that the group has 300 property crimes (actual numbers not rates)											
group of 1 college z-score							1.568				
group of 2 colleges z-score							0.5204				
group of 3 colleges z-score							-0.0554				
group of 4 colleges z-score							-0.464				
group of 5 colleges z-score							-0.7871				

Exam Review Solution Hints for Sheet B

1. There is an average of 89.43 cars on campus at 2:30 pm.
   If every day at 2:30 pm between Feb 1st and May 15th had the same number of cars on campus, that number would be 89.43 in order to get the same number of total cars on campus at 2:30 pm from Feb 1st to May 15th.
2. 50% of the students were advanced or proficient and 50% were needs improvement or warning/failing.
   Half of the students are preformed at or above the expected level. Half of the students are performing below the expected level.
3. The most common number of students on campus Fridays at 3:30 during the spring semester was 736.
4. The majority of students fall between 237 and 277 but the vast majority fall between 217 and 297. This means the vast majority are proficient or advanced, with only a few performing below the expected level. This school district is doing very well on their MCAS scores.
5. In half of the innings Raycroft pitched, there were no earned runs. In one fourth of the innings Raycroft pitched, there were 1 or 2 earned runs. In another fourth of the innings Raycroft pitched there were 3-9 earned runs scored. The lowest inning was 0 earned runs. The highest inning was 9 earned runs.
6. The number of traffic stops in each county last Saturday was between 1 and 237. Every county had 1 or more traffic stops. At least one county had only 1 traffic stop. At least one county had 237 traffic stops.
7. The weight has more variation. The weight of the baby boys ranging from 4.2 pounds to 11.8 pounds is more significant than the cost of mulch ranging from $29 to $41.
8. 39 appears to be an outlier. Differences might come from weather, weekend or not, special event in the area, sales, etc.
9. a) People whose income is in the bottom 20% of income earners, pay a tax rate of 4.3%
   b) The top 1% of income earners pay a tax rate of 31.5%
   c) People in the 87th percentile pay a 20.3% tax rate.
10. There are many answers. One way is the have a small, random, unbiased sample versus a large not random biased one.
11. a) There are many answers. The population must include all male HCC students but must be a larger group. For example, all HCC students or all male community college students in the US or all college students in the world.

b) There are many answer. The sample must be residents of MA, but not all the residents. For example, Holyoke residents or female MA residents or MA residents at the mall on Friday willing to take the questionnaire.

12. It is rare for a dolphin to live to be 37 years old.
13. a) raw data versus rate of fatalities per 100 million miles driven. b) the rate because r^2 is closer to 1
    c) 1997: 3.381 & 5044.1 (both good), 2015: 1.095 & 4023.5 (both seem a little off),
    2030: -0.81 (impossible so way off) & 3173 (seems okay but hard to tell)
15. Numerous answers.
16. a) Numerous answers. Essentially it needs to be random and unbiased.
    b) Numerous answers. For example, usually only really happy or really mad people fill it out. Or some employees do not encourage the cards while others do.
18. and 19. 9990 honest people, 1598 honest people called liars, 2 spies were believed, and 8 spies failed and were caught lying. So lots of honest people are called liars and 20% of spies are not caught.

Exam Review Solution Hints for Sheet D

1. ERROR The vast majority (not all) fall in the interval $8094 \pm 2*7500$

2. ERROR A range of 47 does not imply 0 to 47. It could be any interval that spans an interval of 47. In this case, the larceny-theft crime rates are 16 to 63.

3. ERROR They should be included because other US college/universities are that big.

4. ERROR A p-value of 28% or 0.28 means you fail to reject the null hypothesis. In this case, that means that the MA mean and the CT mean might be the same.

5. TRUE Looking at the confidence interval of the mean, the population mean (i.e. the mean of all 10,000 student colleges and universities) is between 0.32 and 1.68

6. ERROR 9/40=0.225=22.5% One-fourth would be 10/40 or 25%

7. TRUE There are 5 schools have both arson and rape. 5/40=0.125=12.5%

8. ERROR Some schools have been counted twice, namely those that have robbery and aggravated assault.

9. TRUE

10. ERROR The probability associated with a z-score is always the probability of that result or lower. So in this case, 98.13% of schools have a rape rate of 8 or less. Also, this interpretation doesn't make sense because we know a z-score of over 2 or below -2 implies a rare event. Also, looking at the data, most schools have less than 8 rapes per year.

11. TRUE Looking at the mean and the standard deviation for all MA schools, the vast majority of schools have a motor vehicle theft rate between 0 and 3.4

12. TRUE MA is 13th out of 50. That puts them in the 26th percentile.

13. ERROR 507/4782=0.11=11% 11% of the property crimes committed at MA colleges/universities occurred at Harvard.

14. ERROR A negative z-score does not mean unlikely. A negative z-score just says the data point lies below the mean. Turns out the z-score is -0.05 which is very likely.

15. ERROR The scale does not go to 0. The number of offences is still at 1,200,000

16. There are numerous limitations including: only MA and CT, not all the MA and CT schools reported, these are only crimes reported to law enforcement, this is only one year of data, non-school people being involved etc.

17. There are numerous possible reasons including: location (state, urban, rural, surrounding community, etc.) number of residential students, size of enrollment, cost of tuition, number of part time versus number of full time students, male to female ratio, on campus reporting guidelines, a good year versus a bad year, etc.