Library Data Collection[1]

Task:

The Math 127 students are charged with the task of obtaining an "almost" random sample of books from the Cecil College Library. Students must obtain one book for this sample and enter the results on StatCrunch.

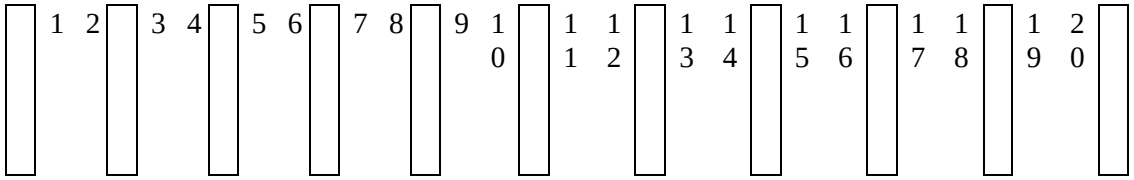
How to find your Book:

There are 10 aisles of "regular" books in the library labeled AC1 to ZA5075. We are not interested in reference books which precede the "regular" books.

The instructor will use StatCrunch to randomly select your "***Aisle***", "***Bookcase***", and "***Book Number***". Then as a class, we will visit the library to obtain the books and record the pertinent information about the book.

A. Starting at the front, where the first book is labeled AC1, there are 10 aisles of books. We want each side of each aisle to be equally likely so we start by picking a random integer between 1 and 20.

The diagram below should help.



The instructor will randomly select your "***Aisle***" number. Write it down. **Aisle: _____**

B. There are generally 8 bookcases in each aisle, and we need to randomly select a bookcase. We will count from the main walkway and then over to the right.

The instructor will randomly select your "***Bookcase***" number. Write it down.

Bookcase Number: _____

C. Now you need to select your book using randomization, but since every bookcase will have a different number of books, we're going to compromise on having a completely random sample here. You will start counting from the top leftmost book and keep counting until you get to your randomly chosen book. We will randomly select a book from #1 to #125

If your shelf has fewer than 125 books, keep counting on the next shelf.

The instructor will randomly select your "***Book Number***". Write it down. **Book: _____**

[1] After completing all the data collections on Day 2, there should be 20-30 minutes remaining for a StatCrunch introduction by your instructor. Don't forget to download the videos and take your flash drive home with you!

Data to Record for your Book:

Once you have your book, you will need to do the following.

1. Write down the Title of the book: _____
2. Write down the complete name of the first author listed: _____
3. Determine the gender of the author listed: Female Male Cannot Determine
4. How many pages are in the book? _____ (Use the last numbered page).
5. What is the most recent copyright, ©, date listed (Use all four digits, e.g. 1998).____
6. Measure the **length** of your book in inches. Use decimals rounded to the nearest hundredth: _____
   For example, 8 1/8" = 8.13". Measure top to bottom of the cover.
7. Measure the **width** of your book in inches. Use decimals rounded to the nearest hundredth: _____
   For example, 5 3/4" = 5.75". Remember the width goes left to right.
8. Measure the **thickness** of your book in inches. Use decimals rounded to the nearest hundredth: _____
   For example, 2 7/8" = 2.88". Lay the book flat and measure up.
9. Now weigh your book on the scale at the front desk. The scale is in grams so record the weight using all of the digits given.

   Weight in Grams_____

Leave the book on the cart near the front desk; library staff will help with reshelving.

You must now visit the www.statcrunch.com Cecil College Math 127 Homepage and enter your responses in the survey near the **bottom** of the page.

Enter your data in the survey "**Calendar Year XXXX Library Data**".

Double check your responses as you input them and be very careful about making egregious typos. THINK about your responses and be certain your values make sense.

Demographic Data Collection

On StatCrunch, complete the "**Calendar Year XXXX Large Survey**" survey.

Answer every question honestly, but do not answer any questions you are uncomfortable answering.

Responses are anonymous.

Armspan vs. Height Data Collection

Let's try to predict your y = "***Arm span***" based on your x = "***Height***".

We will measure "***Armspan***" inches across the back and using the tape measure on the side wall.

We will measure "***Height***" inches with shoes on and using the tape measure in the back of the room.

Enter your measurements, in decimals, on the StatCrunch survey on our group page.

Height: _____ Arm Span: _____

On StatCrunch, complete the "**Calendar Year XXXX Arm Span Height**" survey.

Math 127 Personality Test Data Collection

Visit http://www.humanmetrics.com/cgi-win/JTypes2.asp

or Google "***Take Myers Briggs Test***" and click on the above link.

Answer all 72 questions to the best of your ability. If you know your personality type, take the test today anyhow.

Give your first inclination, don't think too hard about any questions, and answer these questions for how they pertain to your life today (not in the past, not your ideal answer).

There are no right or wrong, no good or bad answers and your results will be anonymous.

When finished, you will receive your 4-letter code. Plenty of interesting information exists on the web about your personality type, and results are typically used to suggest career paths and to explain typical relationships with other people.

_____ My Personality Type

On the StatCrunch group page, complete the "**Calendar Year XXXX Personality Types**" survey.

Introduction to StatCrunch

1. Open up the "**Lego Prices**" dataset on StatCrunch.

1a. Classify each variable as **Q** = Quantitative, **C** = Categorical, or **I** = Identifier.

"***Lego Set #***" _____ "***Lego Set Name***" _____ "***Price***" _____

"***Pieces***" _____ "***Minifigures***" _____

"***Special Minifigures***" _____ "***Min Age***" _____

"***Max Age***" _____ "***Theme***" _____ "***Type***" _____

"***Rating***" _____ "***Exclusive***" _____

"***Hard to Find***" _____ "***Remote***" _____

1b. Explain the "***Who***" for this dataset: _____

1c. Look at the "***Rating***" variable. How many Lego sets are missing a "***Rating***"? _____

Use Data → Validate if you ever need to count up **missing** data values.

1d. **How many** Lego sets have the "***Minecraft***" "***Theme***"? _____

Use Stat → Tables → Frequency to get counts.

Also use Graph → Pie Chart → With Data to get counts.

1e. What percentage of Lego sets are "***Hard to Find***"? Give the fraction, then the decimal, then the percentage.

_____ = _____ = _____

As a rule on categorical variables, go to four decimal places and round properly:

Example: 77 / 115 = 0.6696 = 66.96%

1f. Go again. What percentage of sets have a top "***Rating***" of 5?

_____ = _____ = _____

1g. Make a histogram for "***Pieces***". Certainly we'd all agree that the shape of the distribution is

_____ and _____.

Graph → Histogram

1h. The shape is the same for "***Price***". You can use Options → Edit to quickly change the variable.

1i. Run the summary statistics for "***Price***". Use the correct symbols in your answers below.

Stat → Summary Stats → Columns (always columns)

Sample Size: _____

Sample Mean: _____

Sample Standard Deviation: _____

1j. Now run the summary statistics for "***Price***", but this time we only care about "***Theme***" = "***Star Wars"*** Lego sets.

Stat → Summary Stats → Columns Use Group by → "***Theme***"

Sample Size: _____

Sample Mean: _____

Sample Standard Deviation: _____

1k. Let's use dotplots to answer a few more questions. Graph → Dotplot

Make a dotplot for "***Pieces***" to find the Lego set with the most "***Pieces***".

Set Name: _____ Pieces: _____

Make a dotplot for "***Price***" to count up how many sets cost at least $200: _____

Make a dotplot to find the two Lego sets with the crappiest "***Rating***". "***Rating***" = _____

Set #1: _____

Set #2: _____

StatCrunch Practice

1. Open up the "**Calendar Year XXXX Large Survey**" dataset.

1a. Explain the "***Who***" for this dataset: _____

1b. Draw a diagram of the **population** vs. the **sample** for this dataset:

We believe this sample to be unbiased and representative of the population.

What would happen to the data values in the sample (and the summary statistics and the graphs) if we took another sample of Cecil students?

1c. Make a histogram for Cecil student "***Work Time***". Describe the shape:

1d. Make a histogram for "***Sleep***" hours. Describe the shape:

1e. Make a histogram for "***Living Situation***". Describe the ……. Wait. Hang on. What's up?

1f. Use a dotplot to find the student with the highest "***Number of Tattoos***" and report their "***Marital Status***".

Row # _____ "***Number of Tattoos***" _____ "***Marital Status***" _____

1g. Describe the distribution for the variable "***Self Confidence***".

Shape: _____

Center: _____

Spread: _____

Outliers: (Always use Graph → Boxplot → "Use fences to identify outliers")

1h. Describe the distribution for the variable "***Sleep***".

Shape: _____

Center: _____

Spread: _____

Outliers:

1i. Use dotplots to answer the following questions.

How many students are taking at least 15 "***College Credits***"? _____

How many students have a "***Commute***" of at most 10 minutes? _____

How many of "***Sheppard's***" students are teenagers? _____

Find "***Kupe's***" student who has the most "***Children***". Row # _____ # Kids _____

Find "***Drach's***" student who has the most "***Tattoos***". Row # _____ # Tats _____

1j. How many total "***Tattoos***" do we collectively have? Stat → Summary Stats → Column and add "Sum" to the list.

2. Open up the "**Tuesday Men's Handicap**" dataset. The variable "***Score***" is the team's total bowling score (5 men). "***Lanes***" = 11.5 means the team bowled on lanes 11 and 12 that week.

2a. Make a dotplot for "***Score***" and describe the shape.

2b. Find the highest "***Score***" for the entire league. What was the average score per bowler?

2c. Which pair of lanes yielded the highest mean "***Score***"? _____

2d. Interpret the value of the third quartile for "***Score***". _____

_____

_____

2e. Which game (1, 2, or 3) has the most variation in "***Scores***"? Look at histograms, dotplots, boxplots, standard deviations, IQRs, and ranges.

2f. What was the median "***Score***" for **just Week 9**? Use the **Where Box** this time.

2g. What was the median "***Game 1***" during "***Week 5***"? Need Where Box and Grouping.

2h. What was the median "***Score***" for "***Week 20***" on "***Lanes***" 9 and 10?

3. Open up the "**2010 Movie Revenue**" dataset.

3a. If you were a movie distributor and only cared about selling tickets, which movie rating "***MPAA***" would you encourage your studio to aim for? Why? Hint: Look at means and medians, but also look at the sum.

3b. Which "***Distributor***" had the most "***Tickets Sold***" in 2010? How many tickets did they sell? Hint: Need the sum.

3c. Get the median and IQR for "***Tickets Sold***" for "***Walt Disney Pictures***".

Median: _____ IQR: _____

3d. Find the 20^{th} percentile for "***Tickets Sold***" and explain what the value means.

Explanation: _____

_____

_____

_____

3e. Give the mean "***Tickets Sold***" for "***Universal***" "***R***"-rated pictures.

Answer: _____

3f. Describe the distribution for the variable "***2010 Gross***".

_____

_____

_____

_____

Categorical Variables

We will continue our study of categorical variables using the "**Calendar Year XXXX Large Survey**" dataset.

Professor Kupe has a hunch that a higher percentage of our female students are parents (when compared to the male students).

We would like to create a contingency table for "***Gender***" versus "***Has Children***", except we didn't ask that question in the survey. Rather, we asked for "***Number of Children***" in the survey

1a. By hand, create a contingency table for "***Gender***" vs. "***Has Children***". Use the data from the "**Large Survey**".

1b. What percentage of respondents are parents? Give fraction, decimal, and percentage rounded to the hundredths place.

_____ = _____ = _____

1c. What percentage of females are parents? Give fraction, decimal, and percentage rounded to the hundredths place.

_____ = _____ = _____

1d. What percentage of males are parents? Give fraction, decimal, and percentage rounded to the hundredths place.

_____ = _____ = _____

1e. Does being a parent seem independent of or dependent on gender? Why?

_____

_____

_____

_____

2. Same dataset. Make a contingency table for "***Smoking***" vs. "***Alcohol***". Make pie charts for "***Smoking***" and group it by "***Alcohol***". Using the proper conditional proportions, can we make an argument that the two vices are interconnected or dependent?

_____

_____

_____

_____

3. Open up the "**Calendar Year XXXX Personality Types**" dataset.

3a. Make a pie chart for "***Introvert / Extravert***" grouped by "***Gender***". Does there seem to be a difference? Independent or dependent? Support with proper conditional percentages.

_____

_____

_____

3b. What percentage of "***Females***" are "***Thinking***"? _____

What percentage of "***Males***" are "***Feeling***"? _____

What percentage of "***Thinking***" students are "***Male***"? _____

What percentage of "***Feeling***" students are "***Females***"? _____

What percentage of "***INTJ***" students are "***Male***"? _____

What percentage of students are "***Judging***" and "***Extraverted***"? _____

What percentage of students are "***Male***" and "***Sensing***"? _____

What percentage of students are "***Female***" and "***Introverted***" and "***Feeling***"?

What percentage of students are "***Perceiving***" or "***Extraverted***"? _____

What percentage of students 30 years old or older are "***Thinking***"?

Quantitative Variables

Review Standard Deviation

In words, _____

_____

Formula:

Example: On his recent visit to Maryland, Professor Kupe's Old Man drank 3, 0, 3, 0, 6, and 0 beers during his six-night stay. Calculate the mean and standard deviation by hand. Pop in the data into StatCrunch to verify your answers.

Review Z-Score

In words, _____

_____

Formula:

- Z-scores inside _____ are not unusual.
- Z-scores outside _____ are unusual (very important)
- Z-scores outside _____ are rare.

1a. If Old Man Kupe drinks just one beer tomorrow, what's the *z*-score? Is that unusually low?

1b. If Old Man Kupe drinks seven beers on Friday, what's the *z*-score? Is that unusually high?

1c. Suppose next Sunday, Old Man Kupe had a *z*-score of –0.612. How many beers did he drink?

1d. Give a range of values for "***Number of Beers***" that would be "**unusual**" in terms of *z*-scores. Also give a range of values that would be "**rare**" in terms of *z*-scores.

Review Quartiles and Interquartile Range and the Outlier Rule-of-Thumb

Formulas: IQR = Lower Fence =

Upper Fence =

Example:

2a. Continuing with Kupe's Old Man, determine the quartiles by hand.

2b. Determine the IQR and the fences using the formulas. How many beers would it take to be a high outlier?

Example: Answer the following questions.

3a. The smallest the standard deviation can be is _____. For this to happen:

_____

3b. The smallest the IQR can be is _____. For this to happen:

_____

_____

Example: Fire up the "**Bachelor's Degree Institutions**" dataset.

4a. Convert the "***Kenyon College's***" number of "***Students***" to a *z*-score. Comment if it is an unusually small or large value.

4b. Convert the "***Liberty University's***" number of "***Students***" to a *z*-score. Comment if it is an unusually small or large value.

4c. "***Louisville Bible College***" seems to be missing the "***Students***". Magically we know the *z*-score to be –0.32. Based on that, calculate the "***Students***".

4d. How many "***Public***" schools are official outliers for "***Student-to-Faculty Ratio***"?

# of Low Outliers: _____ # of High Outliers: _____

4e. The "***Atlantic Institute of Oriental Medicine***" is missing its "***Student-to-Faculty Ratio***" but the *z*-score is 2.05. Calculate the "***Student-to-Faculty Ratio***".

4f. Give a range of "***Students***" that would not be considered unusual. Use the idea of *z*-scores to formulate your range.

4g. Calculate the fences by hand for "***Undergrads***".

Sampling Methods

1. For the problem, the population of interest is the entire student body of Towson University. At Towson, there are ~22,285 students, of which ~18,807 are undergraduate students and ~3,478 are graduate students. We will aim to take a sample of ~1,000 students.

Identify the sampling method for each scenario: census, simple random, systematic, cluster, stratified, or convenience.

1a. _____ Using an email list, we just email everyone and take what we can get, because we know that students rarely check their emails. We expect about a 5% reply rate.

1b. _____ Using an email list that is sorted by "***Student ID***" number, we ask StatCrunch to randomly select 1,000 of those numbers and then we email those students. Presume (for whatever reason) most every student replies.

1c. _____ Take a simple random sample of 840 undergrads and a random sample of 160 graduate students to arrive at a total sample of 1000 students. Towson U. is 84% undergraduate, 16% graduate, by the way.

1d. _____ Using an email list that is sorted by "***Student ID***" number, we program a computer to email every 20^{th} student. We presume for some reason most every student replies.

1e. _____ The largest class at Towson might be Psychology 101, with total enrollment of 1,388 when you add up all the sections across one school year. To take our sample, we decide to visit all the PSYC 101 courses over the course of year 2017 and survey those students to get our number of 1,000.

1f. _____ Towson students represent 101 countries of origin. Divide the population into 101 groups, one for each country of origin, and take a random sample from each group. The sample sizes will be dictated by how large each group is (in other words, if 70% of Towson enrollment is Americans, then we will randomly select 700 Americans). Total sample size is still 1,000.

2. Identify the official sampling methodology for each scenario. Answers can be used more than once or not at all.

2a. Administrators at Cecil College wanted to take a sample of students to ask, "Are the restrooms on main campus kept adequately clean and stocked with supplies?". They randomly selected 7 classes from the course schedule, visited those classes in person, and sampled every student present with a pencil and paper survey.
Sampling method:

2b. Starting in Fall 2017, suppose the Cecil College Fitness Center staff requires every visitor to complete a fitness goals survey before they are allowed to use the facility. **Sampling method:**

2c. The US Postal Service is considering canceling Saturday delivery, so to collect data, they leave surveys in 100 randomly selected mailboxes for each and every zip code in the entire country.
Sampling method:

2d. Suppose on the 2nd day of class, instead we did this at the library. We line up at the door of the library, shout in unison, "On your marks! Get set! Go!" and then students run and grab a book off the shelves. Then we record the same variables as we did for our "**Calendar Year XXXX Library Data**" dataset. **Sampling method:**

2e. Suppose on the 2nd day of class, instead we did this at the library. Using the library database on MyCecil, which includes a list of every book in the library (about 35,000 volumes), we use a random number generator to determine each student's book. Then we visit the library and find our book. Then we record the same variables as we did for our "**Calendar Year XXXX Library Data**" dataset. **Sampling method:**

2f. Suppose on the 2nd day of class, instead we did this at the library. There are about 160 bookcases in the library, so we use a random number generator to pick two random bookcases. Then we go collect our sample by taking every book on those two cases. Then we record the same variables as we did for our "**Calendar Year XXXX Library Data**" dataset. **Sampling method:**

StatCrunch University

Activity: Pretend we work for a nonprofit organization that helps people practice good financial habits. In Texas, there is a large university, StatCrunch U, with 46,000 students. Recently, Visa, MasterCard, American Express, and Discover have all been on campus signing up as many students as possible for shiny new credit cards.

We work for the nonprofit. We travel from campus to campus, putting on financial literacy seminars if we feel the students would benefit from it.

We'd like to begin to understand the financial habits of the population, but it is too large to even begin to think about collecting population data. We will rely on sample data to make our decisions.

Our very first task is to take a simple random sample of 32 StatCrunch U students from the population of 46,000 students and collect the following information:

Variables: "***ID***", "***Gender***", "***Class***" (1 = Freshman), "***Hours***" (Credit Hours), "***Work***" (Work Hours), "***Loans***" (Student), "***CC Debt***" (Credit Card)

Diagram the Population and the Sample:

Step 1: We need to number our StatCrunch U students from 1 up to 46,000 and then use the computer to randomly select 32 students. These 32 students will constitute our **simple random sample**. Open up a blank worksheet.

A. Create the numbered list of students. **Data → Sequence Data**
From: 1 To: 46000 By: 1

B. Randomly select 32 students. **Data → Sample**
Sample size: 32

C. Sort the list, smallest to largest. **Data → Sort**

D. Write the 32 numbers in the boxes below, in ascending order.

Step 2: Take the simple random sample. **Make sure** you are logged into StatCrunch!

Type in the web address bar[2] http://www.statcrunch.com/sampler

Select the students at StatCrunch U for your sample using the numbers above.

Once you have selected all 32, hit the Survey! button to create your dataset.

Step 3: Run some preliminary summary statistics.

Mean "***CC Debt***": _____

Median student "***Loans***": _____

Step 4: Put your two values on post it notes and place them on the board. Be careful to get them on the correct graph. Wait for the class to catch up.

[2] Careful – http:// not https://

Step 5: **<u>Sample-to-sample variation</u>** is the variation that exists in the data values and the summary statistics, when taking repeated samples, even when we use an unbiased data collection method like a **<u>simple random sample</u>**.

Did every student get the same mean "***CC Debt***"? ___________

Did every student get the same median student "***Loans***"? ____________

Professor Kupe believes there are about

607,288,378,172,027,500,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

possible samples of size n = _______ taken from population size N = _______________

You just happened to get one of them, your 32 students, how very lucky?!

By the way, with a simple random sample, each of those samples are **<u>exactly equally likely</u>** to occur.

Step 6: Remember we work for the non-profit. We took a sample of SCU students to get a taste of what's happening at StatCrunch U.

Do the students need our services? Should we stay and give financial seminars?

Pretend our rule is this –

> We feel a college needs our services if the mean "***CC Debt***" exceeds $3000.
>
> If it was your sample that was the one we took, would we stay? _____________

Instead, pretend this –

> If the median student "***Loans***" exceeds $1000, we stay.
>
> Would we stay if it was your sample? ________________

Take a look at the board and the post-its.

Do you see we could easily come to different conclusions about SCU based on which particular sample we drew?

Do you see that we could make a mistake making conclusions about a population when we rely on sample data?

Designed Experiment Definitions Practice

Example: A piano player has a recurring gig at a lounge every Friday night and he wants to maximize his tips. He has three repertoires he can play – classical pieces, popular cover tunes, and obscure cover tunes. Since he also dabbles in statistics, he will run a designed experiment to determine which repertoire is the best for generating tips. On each of the next 30 Friday nights, he will randomly select one of three repertoires. Also, every other week, he will bring his female lounge singer to accompany him. Disregard the fact that the tips will be split if she is there for the purposes of this experiment.

Experimental units: _____

Factor 1 and its levels: _____

_____

Factor 2 and its levels: _____

_____

List out the 6 treatments:

Identify the response variable: _____

Example: The good folks at Tide have a new formulation of detergent and they would like to run a designed experiment to determine if water temperature and / or type of stain impacts the detergent's effectiveness.

Forty identical white cloths (10 each) are stained with either ketchup, mustard, relish, or blackberry jam. Half of the ketchup-stained clothes are washed in hot water, the other half in cold water. This is also done for the mustard, relish, and blackberry jam stained clothes. The treatment assigned to each cloth is done randomly using a computer program.

After washing, the white cloths will be graded twice: once using an instrument called a colorimeter and once eyeballed by an expert grading the cloth on a scale of 1 to 10.

Experimental units: _____

Factor 1 and its levels:_____

Factor 2 and its levels: _____

How many different treatments are there? _____ How many cloths are receiving each treatment? _____

Identify the first response variable:_____

Identify the second response variable: _____

Exam I Review

1. Use the "**2016 NFL Players**" dataset on this one.

1a. Classify each variable as **Q** = Quantitative, **C** = Categorical, or **I** = Identifier.

"***Jersey***" _____ "***Name***" _____ "***Position***" _____

"***Age***" _____ "***Height***" _____ "***Weight***" _____

"***Years***" _____ "***College***" _____ "***Team***" _____

"***Position Group***" _____ "***Side***" _____ "***Level***" _____

1b. Explain the "***Who***" for this dataset: _____

1c. Describe the distribution of the variable "***Weight***".

_____

_____

_____

_____

1d. How many players are between 6 foot and 6'5" tall (inclusive)? _____

1e. Calculate the fences by hand for "***Years***" of experience. How many outliers?

1f. Interpret the value of the 45th percentile for "***Age***": _____

_____

_____

1g. Who is the tallest player who went to "***San Diego State***"?

Player: _____ Height: _____

1h. Who is the youngest player on the "***Eagles***"?

Player: _____ Age: _____

2. Open up the "**Calendar Year XXXX Large Survey**" dataset.

2a. What percentage of all students are "***Vegetarian***"? _____

2b. What percentage of the "***Vegetarians***" are "***Male***"? _____

2c. What percentage of respondents are "***Female***" and favor the "***Death Penalty***"? _____

2d. What percentage of respondents are "***Catholic***" or "***Christian***"? _____

2e. What percentage of respondents drink "***Alcohol***" "***Very Often***" or use "***Recreational Drugs***" "***Regularly***"?

2f. Discuss the independence or dependence of "***Gender***" and "***Party Affiliation***". Support with the appropriate conditional proportions / percentages.

_____

_____

_____

_____

2g. How many total "***Children***" do we have? _____

2h. What's the "***Religion***" of the tallest student(s)? _____

2i. Convert Professor Kupe's "***Commute***" to a *z*-score. Is it unusually long? He's in row 1.

2j. Doctor Climent is an incredibly wealthy man with a *z*-score for "***Salary***" of 4.04. Solve backwards for his missing "***Salary***".

3. A Papa John's manager bags up each delivery order for the driver and notes the price, number of pizzas, and the delivery zone. Give the ***Who*** and the ***What*** for this scenario.

4. Circle the only correct expression.

$$\text{Minimum} < Q_1 < \text{Median} < Q_3 < \text{Maximum}$$

$$\text{Minimum} < Q_1 \leq \text{Median} \leq Q_3 < \text{Maximum}$$

$$\text{Minimum} \leq Q_1 \leq \text{Median} \leq Q_3 \leq \text{Maximum}$$

5. Which of the following statistics **<u>could</u>** take on negative values? Circle all that are correct.

z-score	mean	standard deviation	P_{10}	IQR
minimum	Q_1	median	Range	maximum

6. Jack applied for a job as an IT technician and during the interview process, had to complete a programming quiz. His exam score was converted to a *z*-score of 0. How did Jack do? If programming skills are of the utmost importance for this job, should Jack be hired? Why or why not?

7. You just took an economics midterm and you had one of the worst scores in the class. There are hundreds of students in this large lecture class at Penn State. If you translated your exam score to a *z*-score, circle the value of your *z*-score.

-2.91	-0.49	0	+0.49	None of These Make Sense

8. How big can a *z*-score get? How small? _____

9. When we collected the data on the second day of class, was it a designed experiment or an observational study?

"**Large Survey**" _____ "**Library Data**" _____

"**Personality Types**" _____ "**Armspan / Height**" _____

10. True or False.

10a. _____ A negative standard deviation indicates that most of the data values fall below the mean.

10b. _____ Simple random samples are relatively easy to do in the real world.

10c. _____ The types of conclusions you can make using a convenience sample and a designed experiment are essentially the same.

10d. _____ A *z*-score of 0 means the data value is really unusual.

10e. _____ If the IQR is 0, then the upper and lower fences must also be 0.

11. Short Answer.

11a. A *z*-score of –2.73 means our data value was _____

_____

_____

11b. For data that is symmetric, unimodal, and clean, the best measure of the center is the _____ and the best measure of spread is the _____.

11c. Your "***Zip Code***" is certainly a number, but it is not Quantitative. Why not?

_____

12. Use the "**US News National University Rankings**" dataset.

12a. Give a range of values for "***Tuition in-state***" that is not unusual. Think *z*-scores.

12b. How many schools have unusually low "***Tuition in-state***", based on your above interval? _____

12c. How many schools have unusually high "***Tuition in-state***", based on the above interval? _____

12d. Which "***State***" has the most schools? _____

12e. Which "***University***" in "***Ohio***" has the highest "***6yr Graduation rate***"? What is it?

_____

Scatterplots and Correlation

1a. Open up the "**Calendar Year XXXX Library Data**" dataset. Make scatterplots and run correlations to determine which explanatory variable is the best for predicting the *y* = "***Weight***" (grams) of a book.

As a class, address the egregious outliers if needed.

Graph → Scatter plot (overlay polynomial order = 1) Stat → Summary Stats → Correlation

Variable	*y* = "***Weight***"	Possible *x* = "***Pages***"	Possible *x* = "***Copyright***"	Possible *x* = "***Length***"	Possible *x* = "***Width***"	Possible *x* = "***Thickness***"
Relationship	**X**					
Correlation	**X**					

1b. Which *x*-variable is best for linearly predicting *y* = "***Weight***"? Why? _____

_____

_____

1c. Can we improve by combining "***Length***" and "***Width***" and "***Thickness***" into one measurement?

"***Volume***" = _____ with units _____

Create a new variable using Data → Compute → Expression

Run a scatterplot and get the correlation for *x* = "***Volume***" to predict *y* = "***Weight***".

Relationship: _____

Correlation: _____

2. Run Applets → Correlation by Eye → Randomly Generated Data.

Main point: Learn what types of scatterplots go with certain values of the correlation.

3a. Open the "**Calendar Year XXXX Arm Span Height**" dataset. Address any egregious outliers first. Write a brief description of the relationship between x = "***Height***" and y = "***Armspan***". Comment on form, direction, strength, and unusual features. Include a measure of strength in your write up.

_____

_____

_____

_____

3b. Since we are working with two variables now, we **describe the relationship** using form, direction, strength, and unusual features.

Instead if you were asked to **describe the distribution** of, say, "***Height***", what would you need to report? Do it.

_____

_____

_____

_____

4. Sometimes we attach words to certain values of correlation. Complete the table so that we are all speaking the same language.

Correlation	**Descriptive Words**

5. Data was collected on Cecil College students and we recorded x = "***High School GPA"*** and y = "***College GPA"***. The correlation was $r = 0.941$.

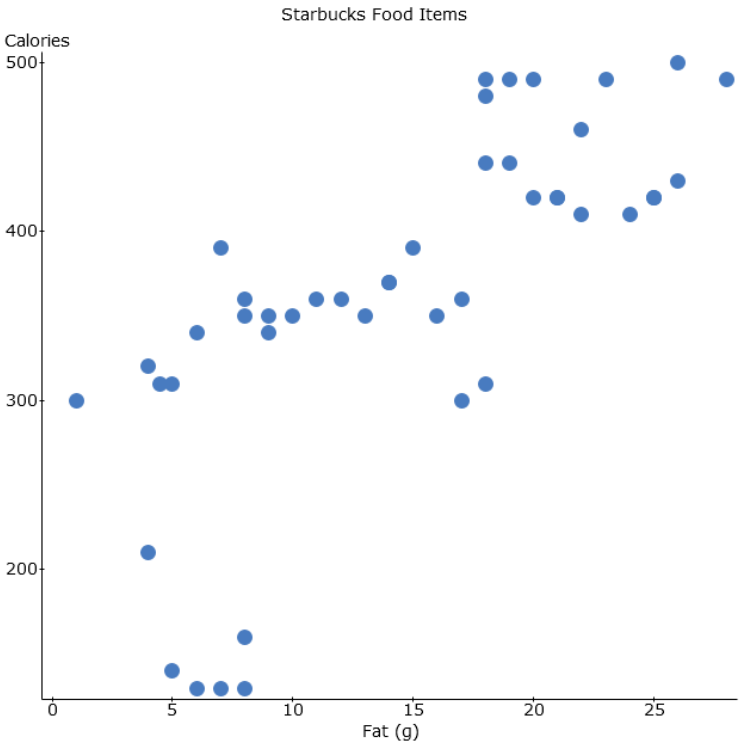
5a. The relationship was linear. Absolutely True Not Necessarily True

5b. As ***High School GPA*** increases, ***College GPA*** increases.

Absolutely True Not Necessarily True

5c. There were no outliers in the dataset. Absolutely True Not Necessarily True

6. Circle all the words that **describe the relationship** between "***Fat***" grams and "***Calories***" for the food items found at Starbucks. We know that $r = 0.732$.



Skewed Right Skewed Left

Linear Symmetric

Positive Negative

Center of About 15 Grams

Moderate Strength

Center of About 350 Calories

Uniform

Regression I

1a. Open the "**Calendar Year XXXX Library Data**" dataset. Using x = "***Pages***" to predict y = "***Weight***", determine the linear regression equation.

Stat → Regression → Simple Linear.

Address any outliers before running the regression. Start with a scatterplot to refresh your memory (always checking for linearity before proceeding).

Equation:_____

1b. Interpret the value of the slope in the context of the problem. **Slope = _____.**

Interpretation:_____

_____

_____

1c. Explain why the y-intercept does not have a meaningful interpretation. ***y*-intercept = _____.**

Interpretation:_____

_____

_____

1d. Review the three requirements for the y-intercept to be a meaningful point on your regression line:

1f. Predict the "***Weight***" of a 495-page book. Show the calculation:

Now cheat using StatCrunch.

1g. Run another linear regression but use x = "***Length***" to predict y = "***Weight***".

Equation: _____

Slope Interpretation: _____

_____

_____

_____

1h. Predict the "***Weight***" of a book that has "***Length***" = 10 inches

1i. Give a range of values for x = "***Length***" for which you would be comfortable predicting the y = "***Weight***".

_____

1j. If you had to pick one, which x-variable ("***Pages***" or "***Length***") are you more comfortable using to predict the y = "***Weight***" of a book? Why?

_____

_____

1k. What is the population for this problem? _____

1l. What would happen to your regression equation if we took another sample of books?

_____

_____

2a. Pull up the "**Calendar Year XXXX Arm Span Height**" dataset. Create the linear equation to predict "***Arm Span***" based on "***Height***". Click on "**Save Residuals**".

Equation: _____

2b. Interpret the slope of the line in the context of the problem.

_____

_____

_____

2c. Explain why the *y*-intercept does not have a meaningful interpretation.

_____

_____

_____

2d. Predict the "***Arm Span***" for 5' tall, 5'8" tall, and 6'2" tall. Use StatCrunch for the most accuracy.

2e. **Review:** In words, what is a residual? _____

_____

_____

2f. **Review:** Give the formula for a residual:

2g. **Review:** In general, _____ residuals mean our regression equation will have **more** predictive value.

2h. The residual for the person in row 30 is _____.

Interpret this residual with a sentence in context: _____

_____

_____

2i. Find the largest negative residual in the dataset. Interpret its value.

_____

_____

_____

2j. Calculate by hand the residual for the person in row 5.

3. Open up "**WalMart Supermarket**" and run a linear regression to predict y = "***Supermarket Price***" using x = "***Walmart Price***".

3a. Regression equation: _____

3b. Interpret the slope with a sentence in context: _____

_____

_____

3c. Now remove the egregious outlier. The students were supposed to get the everyday price, but this item was clearly on sale at Food Lion. Re-run your regression. Look at the slope now. Interpret.

3d. New regression equation: _____

3e. Determine the residual for "***Totino's Pizza Rolls – 40 count***" and interpret its value with a sentence in context.

_____

_____

_____

Regression II

1. For Professor Kupe's ex-"**Neighborhood**" dataset, run the linear regression to predict "***Assessed***" value using "***Square Footage***".

1a. Linear equation: _____

S_e: _____ R^2: _____

Review: The value of S_e is an estimate of the standard deviation of the data points about the linear regression line. The smallest S_e can be is _____ and this means the data values are exactly on the regression line. The largest S_e can be is _____ and this means that the linear regression has no predictive value.

Interpretation: The value of S_e is interpreted as approximately the average amount we expect our predicted *y*-values to be off by using our linear regression equation.

1b. Interpret the value of s_e: _____

_____

_____

Review: R^2 is the percentage of variation in the *y*-variable accounted for by the linear regression using the *x*-variable. The smallest R^2 can be is _____. This means none of the variation in *y* is explained by knowing *x* and the linear regression is predicatively **useless**. The highest R^2 can be is _____ and this means that the data points fall exactly on a straight line. This means that knowing *x* tells us exactly the value of *y*. With real data, you will be somewhere in between.

1c. Interpret the value of R^2: _____

_____

_____

_____

1d. Since _____% of the variation in *y* = "***Assessed***" value **is still unaccounted for**, make a list of potential explanatory *x*-variables that might help explain what's missing.

Conditions for Linear Regression

1e. Is the **linearity** condition met? Check a scatterplot.

_____

Why do we check for linearity? _____

_____

1f. Is the **equal spread** condition met? Check the scatterplot.

_____

Why do we check for equal spread about the regression line? _____

_____

_____

1g. Are there any egregious outliers? Let's formalized the process of unusual data values in the linear regression setting.

Rather than eyeballing residuals to see if they're "large", let's take it a step further. **Studentized residuals** are essentially like *z*-scores. Since this is Introductory Statistics, just take our word for it that what we are doing is taking our residuals and essentially standardizing them.

Remember, when we standardized data values into *z*-scores, any *z*-score outside _____ standard deviations was a flag for "**unusualness**". Let's treat **Studentized residuals** in the same way.

Rerun the linear regression and check the "**Studentized residuals**" box. Check for any Studentized residuals more than ±2. In the real world, investigate these values – possibly re-measure each variable, check for accuracy, and so on.

Any houses with large Studentized residuals? Make a list.

Address	Residual	Studentized Residual

1h. In linear regression, an **<u>influential point</u>** is a data point that significantly changes the value of the slope, y-intercept, correlation, R^2, and S_e if omitted. Identifying these points is important because it is yet another flag to find points that might warrant further investigation.

<u>Influential points</u> are typically far from the center in the x-direction.

We can run the regression and checkbox the "**Cook's distance**" button. This statistic measures the influence of each data point (reasons beyond the scope of Math 127).

Rule for Large Cook's Distance: Any Cook's distance exceeding $\frac{4}{\text{Sample Size}}$ should be looked at for accuracy and other reasons.

For the dataset, determine if we have any large Cook's distances. Investigate those points.

Address	Cook's

2. Open the "**Roller Coasters**" dataset and predict y = "***Speed***" based on x = "***Drop***".

2a. Check all conditions for regression by checking the scatterplot. Address the "***Xcelerator***". Remove this data point and re-run the regression.

_____

_____

_____

_____

2b. Regression equation: _____

2c. Interpret the slope with a sentence in context: _____

_____

_____

2d. Briefly explain why the *y*-intercept is just a point on the regression line: _____

_____

_____

2e. Check for unusual residuals by analyzing the Studentized residuals. Make a list.

_____

_____

_____

2f. Check for influential points by analyzing Cook's distances.

_____

_____

_____

2g. Interpret the value of *R*-sq with a sentence in context: _____

_____

_____

2h. Interpret the value of S_e with a sentence in context: _____

_____

_____

2i. A new rollercoaster is being built and the "***Drop***" will be 175 feet. What is the predicted maximum "***Speed***"?

2j. Interpret the residual for the "***Alpengeist***" with a sentence in context: _____

_____

_____

_____

Probability I

Review Probability Rules:

Rule 1: $P(A^c)$ = _____

Rule 2: P(A or B) = _____ if A and B are disjoint.

Rule 3: P(A and B) = _____ if A and B are independent.

1. The following table lists the prizes and probabilities for a $1 scratch-off lottery ticket.

Prize	$0	$1	$2	$10	$100	$1000
Probability	0.50		0.05	0.01	0.001	0.0001

1a. P(Win $1) = _____

1b. P(Win $1 or $2) = _____

1c. P(Win at least a dollar) = _____

1d. P(Win at most $2) = _____

1e. Buy two tickets, P(Both Winners) = _____

1f. Buy three tickets, P(Win $1 on all three tickets) =

2. You live in Delaware and take Route 40 to school five days each week. You estimate the probability of getting stopped by the train to be 12%. The train seems to come in an unpredictable way, so we will conclude that hitting the train on one particular day is independent of hitting the train on any other day.

2a. P(No train all five days) =

2b. P(First train is on Thursday) =

2c. P(Hit at least one train during your week) =

3a. Use the "**Calendar Year XXXX Large Survey**" to estimate

P(Cecil Student is Christian) = _____

3b. In a random grouping of 6 Cecil students, determine the P(at least one non-Christian).

4a. Use the "**Calendar Year XXXX Library Data**" to estimate

P(Random library book has over 1000 "***Pages***") = _____

4a. If you select 10 books at random, determine P(at least one book has over 1000 "***Pages***").

4b. What if you took 100 books? Same probability.

4c. What if you took 1,000 books at random? P(at least one book has over 1000 "***Pages***") =

Review More Probability Rules:

Rule 4: P(A|B) = _____

Rule 5: P(A or B) = _____ for all A, B.

5. Draw a Venn Diagram if we know that 15% of Cecil students have been in the military, 10% already have a college degree, and 5% have been in the military and already have a degree.

5a. P(Have a degree | Been in military) =

5b. P(Have a degree | Not been in military) =

5c. Are having a degree and being in the military independent or dependent? Why?

6. At Cecil College, 67% of students are female. Also, 10% of all students are business majors. Presume that gender is independent of whether or not a student is a business major.

6a. Determine the probability that a randomly selected student is a female business major.

6b. Determine the probability that a randomly selected student is a male business major.

6c. Draw a well-labeled Venn diagram.

6d. Determine the probability that in a sample of 10 students, we don't get one single business major.

6e. Determine the probability that in a sample of 10 students, we get at least one business major.

Probability II

When we generate trials from a probability experiment, the values can either be **continuous** or **discrete**.

1. **Continuous probability distributions** generate an infinite number of possible values, sometimes bounded with an absolute minimum or maximum value. Very often, things we measure are continuous.

For example, "***Men's Heights***" follow a continuous Normal model (Lesson 12 upcoming). The possible heights of men are infinite, only constrained by how accurate of a measuring tape or ruler that we have.

Today we will briefly look at _____ distributions and _____ distributions. These are both continuous distributions, along with our Lesson 12 Normal distributions upcoming.

When working with a continuous probability distribution, the _____ under the curve corresponds to the probability.

Uniform Distribution

2. Suppose a college professor always runs classes late. He never finishes on time and always finishes within 5 minutes after the class is supposed to end. The variable X = "***Number of Minutes Running Late***" follows a Uniform distribution (in other words, all possible times are equally likely).

2a. Draw the graph of this Uniform probability model. Give the probability function.

2b. Determine the probability that class goes at least 3 minutes late. Draw the shaded distribution.

2c. Determine the probability that today's and tomorrow's classes both go at least 3 minutes late.

2d. How many minutes do we expect class to go over?

2e. What's the chance today's class goes at most one minute over? Draw the shaded distribution.

2f. What's the chance class goes between 90 seconds and 200 seconds late? Draw and shade.

2g. Find the 39^{th} percentile. Explain.

3. Suppose we program a computer to spit out randomly generated real numbers on the interval from –7 to +10. Don't ask me why. All possibilities are equally likely and we are dealing with real numbers, not just the nice integers. Then the variable *X* follows a Uniform distribution.

3a. Draw the graph of this Uniform probability model. Give the probability function.

3b. What is the expected value or mean for this random variable? What is the interpretation of this value?

3c. Determine the probability that we get a negative number. Draw and shade.

3d. Now determine the probability we get 7 negative numbers in a row.

3e. Determine the probability that we generate a number within 0.5 units of the mean. Draw and shade.

Exponential Distribution

An Exponential distribution is a probability model often used to understand the times between events, such as time between customers pulling up to the Taco Bell drive through, time between hits on a webpage, or time until an electrical component fails.

Though the function is a bit beyond the scope of Math 127, we can use the StatCrunch calculator and emphasize our understanding of the rules of probability while introducing this new probability model.

4. Suppose we work in a lab that has a piece of equipment with a mean lifetime of 2000 hours. We know that the equipment's lifetime follows an **Exponential** probability model.

4a. Determine the probability that our equipment lasts longer than average. Draw and shade.

4b. Suppose we just bought two new machines. Determine the probability that both of them last at least 3000 hours.

4c. If the manufacturer replaces the machine free of charge if it dies within the first 500 hours, what percentage of machines do they end up replacing?

4d. The best 5% of machines last how long?

4e. The worst 10% of machines last how long?

5. During the hours of 11:30 p.m. to 1:30 a.m., the time between customers pulling up to the Taco Bell drive-thru follows an Exponential distribution with a mean 180 seconds.

5a. Determine the probability that the next car doesn't pull up for at least 15 minutes.

5b. Determine the probability that the next car arrives within the next minute.

5c. Suppose a 2nd Taco Bell down the road follows the same Exponential model. What's the chance that both get a customer in the next two minutes?

5d. The 90th percentile of this distribution is _____. This means

_____

_____

_____

_____

Normal Models

1. The webpage www.tvbythenumber.com reports that ***The Walking Dead*** averages 13.8 million viewers per episode. Suppose the standard deviation is 1.54 million viewers and that a Normal model applies. **For every problem, draw the Normal model**

1a. What percentage of the time will fewer than 11 million people tune in?

1b. What's the probability that a randomly selected show has between 13 and 15 million viewers?

1c. Find the number of viewers that represents the 80^{th} percentile.

1d. Determine the first, second, and third quartiles for viewership. Determine the IQR.

1e. For the next four episodes, what's the probability that all four of them have at least 14.5 million viewers?

1f. For the next four episodes, what's the chance at least one episode has at least 16 million viewers?

1g. Give a range of values for viewership that would give the cutoffs for what is not unusual. Think *z*-scores.

1h. Don't ask why, but we know that last night's episode had a *z*-score of –0.83. How many viewers?

2. A certain interior house paint comes in 1-gallon cans. The color "***Midnight White***" requires the paint machine to add a bit of blue dye. If the machine adds blue dye with $\mu = 5$ ml and $\sigma = 0.03$ ml following a Normal model, answer the following questions.

2a. Cans with more than 5.08 ml or less than 4.91 ml of blue dye can be detected by the eye as a color mismatch. What percentage of cans sold will show a color difference?

2b. If you buy six gallons to paint your living room, give the probability that all six cans color match.

2c. The bluest 2% of cans will have this much blue dye added: _____

2d. Determine the 19th percentile for the amount of blue dye: _____

2e. On a particular can, the machine spits out 4.929 ml of blue dye. Convert that value to a *z*-score:

3a. **Standard** Normal Model. The mean is _____ and the standard deviation is _____.

3b. P(*Z* exceeds 3) = _____ P(*Z* < 0) = _____

P(–3 < *Z* < 3) = _____ 95th Percentile: _____

Give the two *Z* values that capture the central 95% of the area: _____

4. Pretend we don't know the mean IQ at Harford Community College, but are willing to presume the standard deviation is 15. We give a number of IQ tests, and 17.5% of students scored 90 or below. Determine the mean IQ presuming a Normal model applies.

5. Women's height in the Midwest follows a Normal model with a mean of 64.5 inches. We collect data from a random sample of women and learn that 20% of women were 5'7" or taller. Determine the standard deviation.

Discrete Models (With Tables)

Discrete probability distributions generate a fixed or countable number of possible values. The most important one for us is the Binomial distribution. We will briefly touch on discrete distributions given in table form, and then move on to the Binomial distribution.

1. For insurance purposes, Professor Kupe's old house had a replacement value of $275,000, even though the house was not worth nearly that much. Each year, he had to pay the home insurance premium, which ran about $800.

Pretend we live in a simplified world where there are only three outcomes each year for Kupe's home – ***Total Destruction*** (burn to the ground, tornado sweeps it away), ***Medium Damages*** (Ted's tree falls through roof, hurricane Irene floods the basement), or ***Neither***. A table with the outcomes, payouts, and probabilities are listed next. The random variable X = the payment.

Kupe's Outcome	**Payment x**	**Probability $P(X = x)$**
Total Destruction	$275,000	1 / 1000
Medium Damages	$10,000	40 / 1000
Neither	$0	959 / 1000

1a. What's the chance that the insurance company pays something out this year?

1b. What's the chance that the insurance company pays out nothing for the next 10 years? Assume years are _____.

1c. What's the chance, in the next 50 years, that the insurance company pays out something at least one time?

1d. What is the mean, or expected payout, each year? Is the insurance company making money?

2. Draw a card from a standard deck. If you get a black card, you win nothing. If you get a diamond, you win $100. If you get a heart, you get $200 plus a bonus of $1000 for the queen of hearts.

2a. Set up a probability table for the prize you win.

2b. Find the expected amount won. What should you be willing to spend to play the game?

2c. P(Win at least $100) = _____ P(Win at most $100) = _____

Play twice, shuffle up each time, P(Win exactly $1200 total) =

3. The following table lists salaries and probabilities for a very large company's IT Department:

Salary	**$50,000**	**$60,000**	**$75,000**	**$90,000**	**$125,000**	**$150,000**	**$200,000**
Probability	**0.61**	**0.20**	**0.08**	**0.05**	**0.03**	**0.02**	**0.01**

3a. P(Random IT employee makes under $100,000) = _____

3b. P(Random IT employee makes at least $60,000) = _____

3c. P(Two random employees both make $50,000) = _____

3d. Expected salary in this department:

Binomial Models

1. The website ***Science Daily*** performed a study a few years back and concluded that 38.5 percent of college students have smoked pot in the last year. Here at Cecil, we will randomly select $n = 11$ students and count up the number who have smoked pot in the last year.

1a. Check the conditions required to use a Binomial model.

Condition 1: _____

Condition 2: _____

Condition 3: _____

Condition 4: _____

1b. Give the parameters and give the Binomial probability function.

1c. Determine P(Exactly 3 have smoked pot). Show the formula calculation. Then show the StatCrunch shortcut.

1d. Determine P(majority have smoked pot).

1e. Determine P(At most 2 have smoked pot).

1f. Determine the mean and standard deviation of this Binomial distribution.

2. Use the "**Calendar Year XXXX Large Survey**" on this one.

2a. Estimate P(Cecil Student is Republican) by using the "***Party Affiliation***" variable.

Answer: _____

2b. If we take a new unbiased sample of $n = 25$ students, then $X =$ "***Number of Republicans***" will follow a Binomial probability model. How many Republicans do we expect to get?

2c. Give a range of values for X that would not be unusual. Think z-scores and our rule of thumb.

2d. P(At least 10 Republicans) = _____

P(At most 8 Republicans) = _____

P(Only 4 or 5 Republicans) = _____

P(Majority are Republicans) = _____

3. **Recall:** A certain interior house paint comes in 1-gallon cans. The color "***Midnight White***" requires the paint machine to add a bit of blue dye. If the machine adds blue dye with $\mu = 5$ ml and $\sigma = 0.03$ ml following a Normal model, answer the following questions.

Cans with more than 5.08 ml or less than 4.91 ml of blue dye can be detected by the eye as a color mismatch.

What percentage of cans sold will show a color difference? **Answer:** 0.0052 = 0.52%
What percentage of cans sold will be OK? **Answer:** 0.9948 = 99.48%

3a. Working as summer painters, we buy a whole pallet of this paint, $n = 24$ cans. Give the Binomial probability function for the number of cans out of 24 that are OK.

3b. Determine the probability that exactly one can is **no** good (out of 24).

3c. Determine the probability that at least 22 of the cans are OK (out of 24).

3d. How many cans do we expect to be OK?

Exam II Review

I. Linear Regression

1. Open up the "**Audi A5**" dataset on StatCrunch. In the market for a new car (well, new to him), Professor Kupe began investigating Audi A5s on www.autotrader.com and recorded some information about $n = 40$ used cars for sale. He took an unbiased sample and noted the "***Year***", "***Mileage***", and "***Price***" for the 40 closest cars to his house.

1a. Use "***Mileage***" as the explanatory variable. Describe the relationship with the response variable "***Price***". Hit all points, and include a measure of strength in your write up.

1b. Run the least squares regression equation, and save residuals, Studentized residuals, predicted values, and Cook's distances. **Give the equation of line of best fit**.

1c. Using the scatter plot, discuss all the required conditions that need to be met to have a valid linear regression equation.

1d. Professor Kupe's buddy has an Audi A5 with 70,000 miles on it. He wants $30,000 for it. Is it a good value? Determine the predicted "***Price***" and the value of the residual and make an assessment.

1e. Interpret the value of the slope with a sentence in context.

1f. Analyze the Studentized residuals. Are there any values that might need further investigation? Discuss.

1g. Explain the meaning of the Studentized residual for the car in row 20.

1h. Now explain the meaning of the residual for the car in row 20.

1i. Interpret the value of R^2 with a sentence in context.

1j. Though it's a bit of extrapolation, use the model to predict the price of a new Audi A5.

1k. What's a big Cook's distance for this dataset? _____

Which rows have large Cook's D? _____

What does Cook's D measure? _____

1l. Interpret the value of S_e with a sentence in context: _____

II. Probability

2. Time to grade Math 127 quizzes follows a Uniform distribution with endpoints of 3 minutes and 5.5 minutes.

2a. The average time to grade a Math 127 quiz is _____.

2b. Draw the probability model and give the probability function:

2c. Determine the probability that the next quiz takes longer than 5 minutes to grade:

2d. Determine the 99^{th} percentile for grading time and interpret its value:

2e. In a class with 30 students, what is the probability that no quizzes takes longer than 5 minutes to grade?

3. Gas prices in Cecil County follow a Normal distribution with a mean of \$2.64 and a standard deviation of \$0.25. Gas prices were current at the time of writing.

3a. What's the chance a randomly selected gas station has gas under \$2.50?

3b. If we randomly select 15 gas stations, determine the probability that at least one of them has gas over \$3.00.

3c. Give the value of the 90^{th} percentile for gas prices: _____

4. In New Castle County, we don't know the mean gas price. Presume the standard deviation is 25¢ just like in Cecil County. We know that 20% of gas stations are at \$3.00 or over. If a Normal model applies, determine the mean gas price in New Castle County.

5. P(Cecil student is late for class) = 0.10. We have a class with 30 students.

5a. If we are counting up the number of students that are late on any given day, which probability model do we use? Give the parameters.

5b. How many students do we expect to be late?

5c. What's the probability no students are late?

5d. What's the probability that at least 20% of the class is late?

6. Professor Kupe's least favorite hockey player of all time is Sydney Crosby. Suppose the following table estimates the probabilities for the "***Number of Goals***" scored each game.

Goals	**0**	**1**	**2**	**3**	**4**	**5**
Probability	**0.56**	**0.39**	**0.03**	**0.01**	**0.006**	**0.004**

6a. P(At most one goal) = _____

6b. P(At least one goal) = _____

6c. For any given game, how many goals do we expect Crosby to score? Show calculation.

7. Draw the Venn Diagram for the following facts: 45% of Cecil students have taken Western Civilization I, 35% have taken Intro to Sociology, and 25% have taken both courses.

7a. Draw and label the Venn Diagram:

7b. P(Taken Western Civ | Intro Sociology) =

7c. P(Taken Western Civ | Not Taken Intro Sociology) =

7d. P(Taken Sociology | Taken Western Civ) =

7e. P(Taken Sociology | Not Taken Western Civ) =

8. Your McDonald's is running a contest on McMuffins. "***1 in 5 wins***" it says.

To be nice, you buy your whole 8:00 English 101 class of $n = 13$ students a McMuffin, with the express agreement that if anyone wins, it's your prize.

Determine the probability that you win at least one time. Show calculation.

9. iPhones have a mean lifetime of 3 years. Let's suppose iPhone lifetimes follow an Exponential model. Presume 12 months in a year and that we just divide a year up into 12 equal chunks.

9a. P(Random iPhone lasts at least 4 years) = _____

9b. P(Random iPhone dies within the first year) = _____

9c. P(Random iPhone dies within the first month) = _____

9d. The best 10% of iPhones last at least _____ years.

9e. The worst 1% of iPhones die within the first _____ years.

9f. Kourtney, Kim, Khloe, Kendall, and Kylie all get new iPhones. Presuming independence, what is the probability that all five iPhones last at least one year? Show calculation:

10. 16-pound bowling balls have weights that follow a Normal model with a mean of 15.97 pounds and a standard deviation of 0.12 pounds.

10a. Determine the probability that a randomly selected ball weighs over 16 pounds.

Answer: _____

10b. Determine the probability a ball weighs in within 0.05 pounds of the stated 16-pounds.

Answer: _____

10c. Determine the probability that a ball weighs in within two standard deviations of the mean of 15.97 pounds.

Answer: _____

10d. Four random bowling balls are pulled off the rack. Calculate the probability that at least one of them is over 16 pounds.

Sampling Distributions for $\hat{p}$

Lesson 14 Formulas:

Lesson 14 StatCrunch Moves: **Stat → Calculators → Normal**

1. The American Association of Community Colleges states on their website that 36% of community college students are the first generation in their immediate family to attend college.

At Cecil College, we will run a simulation **as if** we had access to the whole population of $N = 2861$ students. Lesson 14 deals with hypothetical scenarios so that we can learn the theory behind a sampling distribution. **In practice**, we wouldn't have access to a whole population of data, and we wouldn't take many, many random samples "just to see what happens". In Lesson 14, this is what we do.

1a. Open up the "**Cecil College First Generation**" dataset StatCrunch.

This dataset would **never** be available to us. It is the entire population, it is made up by Professor Kupe, but should be realistic for the purposes of this investigation.

We will repeatedly sample $n = 200$ students and ask those students if they are in the first generation to go on to college. Then we will repeatedly calculate the sample proportion of first generation college students, $\hat{p}$.

The Whole Point: We need to see how $\hat{p}$ behaves. We need $\hat{p}'s$ probability model, its mean, and its standard deviation. Lesson 14 is the theory Lesson. Then, in the next 3 Lessons, we will harness this information to do statistical inference.

We have coded those who said "***Yes***" = 1 and those who said "***No***" = 0.

First, determine the **population proportion** of Cecil College students who are the first to go on to college. We would never know this in practice! In today's class we do, because today we are playing in make-believe land, running a simulation for a hypothetical scenario.

True Population Proportion: _____

Diagram Population vs. Sample

1b. Applets → Sampling Distributions → From Data Table → Values In: "Coded"

Each student will individually take <u>up to</u> 10 separate samples of size *n* = 200 students and determine <u>up to</u> 10 separate sample proportions. The instructor will demonstrate first because this is a bit tricky.

Click the "**1 Time**" button once. The middle section, it says "**Mean**". This is the sample proportion since we coded the data values as 1 and 0. Record the sample proportions below.

$\hat{p}_1$ = _____ $\hat{p}_2$ = _____ $\hat{p}_3$ = _____ $\hat{p}_4$ = _____

$\hat{p}_5$ = _____ $\hat{p}_6$ = _____ $\hat{p}_7$ = _____ $\hat{p}_8$ = _____

$\hat{p}_9$ = _____ $\hat{p}_{10}$ = _____

1c. Transfer your sample proportions onto your post it notes. Stick them on the board. Watch the sampling distribution come alive. Wait for the entire class to finish.

1d. Even after one-hundred or so sample proportions, the shape of the sampling distribution for the sample proportion is clearly _____ and _____.

We can correctly use a _____ model.

1e. We need the mean and standard deviation of the sampling distribution for $\hat{p}$:

Theoretical Mean: _____

For this example:

Theoretical Standard Deviation: _____

For this example:

1f. Now, return to the applet, "**Reset**", and hammer on that "**1000 times**" button until you have at least 100,000 samples.

i. Compare the theoretical mean to your mean after 100,000+ samples. Are they close?

ii. Compare the theoretical standard deviation to your standard deviation after 100,000+ samples. Are they close?

iii. At the bottom, click the "**+**" next to "**Sample Means**". Discuss.

1g. In practice, we won't have access to the population data, but many times we will start our investigations by making an assumption about a population parameter.

Suppose instead, we were willing to **assume** that at Cecil College, $p = 36\%$ first generation college students. With a sample size of $n = 200$, which Normal model would we use under this **assumption**? Draw it. Understand the difference between our simulation and this model in 1g.

1h. Suppose at Cecil College, we did not have access to the population data, we did not just run a simulation, and we began our investigation by assuming the 36% figure holds at our college. We take one sample of size $n = 200$. What values of $\hat{p}$ would lead you to change your mind about Cecil College?

2. Suppose in Cecil County, a certain candidate for Sheriff has in fact 53% of the entire county's support. The election is tomorrow, but the Cecil Whig runs a poll by interviewing a random sample of $n = 150$ residents during the week before voting day. A majority is needed to win the election.

2a. Will this candidate win on voting day? _____

2b. Why? _____

_____

2c. Determine the probability that the Cecil Whig incorrectly predicts the winner of the election. Get the model for the sample proportion based on knowing the fact that 53% of residents support this guy (PS, we wouldn't know this in practice, today is make-believe, hypothetical day). Then shade your model in the appropriate direction to determine the probability the Whig is wrong.

2d. What if this candidate actually had 65% of the vote. Determine the chance that the Whig makes a mistake.

3. Let's pretend that the information collected by students for the "**Calendar Year 2017 Cell Phone Addiction**" project is the honest-to-goodness truth in Cecil County. Let us presume that 19.26% of all teenagers on their phones are on "***SnapChat***". Let's pretend we will take an unbiased sample of n = 300 Cecil County teenagers.

3a. Give the mean and standard deviation for the Normal model used for modeling $\hat{p}$. Show the standard deviation calculation.

3b. P(A random sample of 300 teenagers has more than one-quarter on "***SnapChat***") = _____

3c. The expected number on SnapChat (out of 300) is _____.

3d. Say 51 teenagers are on "***SnapChat***". Convert the raw number to a proportion. Convert the proportion to a z-score. Show it.

3e. Maybe our model is wrong? If you took one sample of 300 teens, what values of the sample proportion would lead you to believe the true proportion using "***SnapChat***" is higher than 19.26%? Lower than 19.26%? Justify.

3f. Presuming the $\hat{p}$ model is correct, give the values of the three quartiles of the model:

Q_1: _____ Q_2: _____ Q_3: _____

An Introductory Example For Confidence Intervals

Example: Consider the people in the classroom a population. The mean age is unknown. We are going to estimate the mean age with a best guess, and then cook up a confidence interval.

Draw a diagram of the population with the unknown mean age:

All confidence intervals are of the form:	**Best Guess ± Margin of Error**

Look around the room and guess the average "***Age***", everyone included: _____

Now centered at your best guess, add and subtract your **margin of error** – you should add and subtract the exact amount so that **you** are 95% confident that your interval will capture the actual true average "***Age***" of the people in the classroom today.

You **must** add and subtract the same amount.

You **are not** trying to create an interval that goes from the youngest "***Age***" in the room to the oldest "***Age***" in the room.

You **are** trying to create an interval that you are 95% confident will capture the classroom mean "***Age***".

Points to consider: You want to be 95% confident you hit the target, but at the same time, a super wide interval will be useless.

For Example: If Professor Kupe needed to estimate his mean "***Commute***" to Cecil College, he would guess 50 minutes. If he needed to cook up an interval so that he was 95% confident he captured the true mean "***Commute***", he might say something like -

$$50 \pm 6 = (44 \text{ minutes up to } 56 \text{ minutes})$$

There are no wrong answers.

My mean "***Age***" interval: _____ to _____

The instructor will put the intervals on the board. Record them on the next page:

15 20 25 30 35 40 45

---

15 20 25 30 35 40 45

1a. What if we were asked to be 99% confident? What would happen to your interval? What if there were a $1,000 prize for hitting the target?

1b. Look around the room at the student "***Ages***". What if there were more variation in "***Ages***" (typical night classes)? What if there were less (typical day classes)? What would happen to your interval?

1c. **Totally Optional:** Calculate the mean "***Age***" for the entire classroom to see if your interval hit the target. This would never happen in practice.

Confidence Intervals for *p*

Lesson 15 Formulas:

Lesson 15 StatCrunch Moves:

Get $\hat{p} = \frac{\#\text{ Successes}}{\#\text{ Observations}} = \frac{x}{n}$ using Stat → Tables → Frequency or Graph → Pie Chart

Get *z* values using Stat → Calculators → Normal

Get confidence intervals for a proportion using Stat → Proportion Stats → One Sample → Summary

1. Recall the confidence interval formula for a population proportion:

2. Recall the common critical values for the following confidence levels. Draw the Normal model for the 95% confidence interval.

90% → *z* = _____ 95% → *z* = _____

98% → *z* = _____ 99% → *z* = _____

3. The more confident we need to be in actually capturing the true population proportion, *p*, the larger the critical value _____. This number tells us how many _____ to jump out from the sample proportion.

4. In data collected at the college, 34.09% of students sampled play the "***Lottery***" at least once per year. Altogether, 132 students were sampled from our population of over 2,800 students.

4a. Check the conditions to proceed with the confidence interval for the true proportion of all Cecil College students who play the "***Lottery***".

Condition 1: _____

Condition 2: _____

Condition 3: _____

Condition 4: _____

4b. Use the formula and then StatCrunch to determine a 97% confidence for the true proportion.

4c. Interpret your interval with a sentence in context: _____

_____

_____

_____

4d. Statistically, are we convinced more than 1/3 of our students play the "***Lottery***"? Why or why not?

_____

_____

_____

4e. We have the margin of error from **4b** already. Demonstrate an alternative method to find the margin of error. Show the calculation.

5. Let's investigate **confidence**. If we took many repeated random / unbiased samples from the population we are investigating, the confidence level is the proportion of intervals that would contain the true population proportion.

Example: Recall the "**Cecil College First Generation**" dataset has population data for every Cecil student. We are investigating if students are the in the first family generation to go on to college. For the whole school, $p = 0.4289$, but in a real world setting, we would not know this!

Applets → Confidence Intervals → For a Proportion

Using data table: Values in "**First Generation**". Success: **Yes**

We will simulate many 95% confidence intervals pretending to take repeated samples of size $n = 200$ Cecil College students. (In practice, we take just one sample and compute just one confidence interval).

Discuss what happens to the interval with various sample sizes:

6. Sample size calculations.

6a. Suppose we would like to perform an analysis to determine the proportion of county residents who have a "***Credit Card***". We'd like to be 99% confident and will use a 4.5% margin of error. Show the calculation for the sample size required.

6b. Suppose we want to perform an analysis to determine the proportion of county residents who have experimented with "***Heroin***". We'd like to be 95% confident and we can live with a 3% margin of error. Determine the required sample size[3].

[3] Use the internet to determine a good estimate for the proportion who have used "***Heroin***".

7. Load up the "**Calendar Year XXXX Large Survey**" dataset. Determine a 95% confidence interval for the proportion of all Cecil College students who "**Favor**" the "***Death Penalty***".

7a. Think about the conditions.

7b. Give the sample proportion: _____

7c. Give the confidence interval: _____

7d. Calculate the margin of error:

7e. What two things could we do to decrease margin of error? _____

_____

7f. What's the one thing we can't change that affects margin of error? _____

7g. Interpret your interval with a sentence in context: _____

_____

_____

7h. The Washington Post claims that 42% of all Americans prefer the death penalty for punishment of convicted murderers. Do we have statistical evidence that Cecil College students have a difference in opinion? Why? (PS – The Post is left-leaning. What kind of sample might they have taken?)

_____

_____

_____

7i. Suppose we wanted to collect a larger sample so as to reduce the margin of error down to 3%. What is the required sample size? Use the value of $\hat{p}$ from the dataset and retain the 95% confidence.

Hypothesis Tests for *p* Part I

Lesson 16 Formulas:

Lesson 16 StatCrunch Moves:

Get $\hat{p} = \frac{\#\text{ Successes}}{\#\text{ Observations}} = \frac{x}{n}$ using Stat → Tables → Frequency or Graph → Pie Chart

Get P-values using Stat → Calculators → Normal

Run one-sample proportions tests using Stat → Proportion Stats → One Sample → Summary

1. Open up the "**Calendar Year XXXX Large Survey**" dataset on StatCrunch.

Run the test to determine if a majority of all Cecil College students carry a "***Credit Card***".

1a. Summarize the data with a sample proportion:

1b. Check all the conditions required to proceed with the one-sample test for a proportion.

Condition 1: _____

Condition 2: _____

Condition 3: _____

Condition 4: _____

1c. Write the hypotheses (first in words, then in symbols):

1d. Diagram the situation with a Population vs. Sample picture:

1e. Run the test[4] using the formula and by shading the Normal model properly. Then run the test on StatCrunch → Stat → Proportions → One Sample → With Summary.

[4] Determine the test statistic and the P-Value.

1f. Based on the P-value diagram, make a decision to reject H_0 or fail to reject H_0.

1g. Write a concluding remark in the context of the problem. _____

_____

_____

_____

_____

2. Same dataset. Using the "***Marital Status***" variable, do we have evidence that less than one-quarter of all Cecil College have been married at some point in their lives? Conditions should be met.

2a. Summarize the data with a sample proportion.

2b. Run the test, all steps, using StatCrunch features.

2c. Interpret the value of the test statistic with a sentence. _____

_____

_____

_____

2d. Interpret the P-value with a sentence. _____

_____

_____

_____

2e. If we made a mistake, what kind? What would it mean in the context of the problem?

_____

_____

_____

_____

3a. Same dataset. Test if more than 40% of all Cecil College "***Females***" have "***Tattoos***". Show all steps. Presume the conditions are met.

3b. Support your conclusion with a 95% confidence interval for the true proportion of all Cecil College "***Females***" who have "***Tattoos***". Interpret your interval with a sentence in context.

_____

_____

_____

3c. Interpret the standard error of $\hat{p}$ (from the confidence interval StatCrunch window) with a sentence in context.

_____

_____

_____

3d. What's the margin of error of the interval? If we wanted to get it down to 4%, what is the required sample size? Use 95% confidence and the value of $\hat{p}$ from the data.

Hypothesis Tests for *p* Part II

Two-Tailed Hypothesis Tests

1. Some studies suggest that half of all people are "***Introverted***". Do we have reason to believe that at Cecil College, the proportion **differs**? Crack open that "**Calendar Year XXXX Personality Types**" dataset.

Treat this sample as an unbiased representative sample of all Cecil College students (it may not be for personality type, considering the class you are taking).

Run a two-tailed hypothesis test. Show all steps. Draw a shaded Normal model to show the P-value.

Using a Confidence Interval To Run A Hypothesis Test

2. You can use a confidence interval to run a hypothesis test. The author is sweeping some nitpicky details under the rug, but essentially it goes like this:

A. Set up your hypotheses as per usual.
B. Cook up the required confidence interval.
C. If your interval contains the hypothesized value, fail to reject H_0.
D. If your interval is self-contained in the direction of H_A, reject H_0.

2a. Use the "**Calendar Year XXXX Library Data**" to test if more than one-quarter of our books are from this century. Show all steps for the hypothesis test and reconcile the result with a 99% confidence interval. Conditions are met.

2b. Use the "**Calendar Year XXXX Large Survey**" to test if less than one-third of all Cecil College students are "***Catholic***". Show all steps for the hypothesis test and reconcile the result with a 95% confidence interval. Conditions are met.

Fixed Significance Level Tests

3. We have been using the P-value diagram rather fluidly. If the P-value is under 0.05, we typically reject the null. If it's under 0.01, even better, because we have even more evidence for the alternative hypothesis. If the P-value is over 0.10, we are assuredly failing to reject the null.

In that GRAY AREA, we might have two different people come to opposite conclusions.

In practice, typically the researcher publishes results and the P-value and it's up to the readers to determine if they are convinced that the evidence is there.

At times, we can specify **exactly** how much evidence we need to reject the null hypothesis.

That line in the sand is _____ or the _____ level.

The two most common values are _____ or _____.

3a. Use the "**Calendar Year XXXX Large Survey**" to test if more than 12% of Cecil students think "***Trump***" is doing a "***Good***" job?

Show all steps. Conditions are met. Run the test with a 5% level of significance.

3b. Gallup reports that 62% of Americans think "***Global Warming***" is real. Do we have evidence that Cecil College students have a difference in opinion? Run the test, show all steps, conditions are met. Use a 1% level of significance. "**Calendar Year XXXX Large Survey**" dataset.

Tests and Intervals for the Difference in Two Proportions

Lesson 17 Formulas:

Lesson 17 StatCrunch Moves (Probably Will Need Grouping):

Get $\hat{p} = \frac{\#\text{ Successes}}{\#\text{ Observations}} = \frac{x}{n}$ using Stat → Tables → Frequency or Graph → Pie Chart

Get P-values using Stat → Calculators → Normal

Run two-sample proportions intervals and tests using

Stat → Proportion Stats → Two Sample → Summary

Confidence Intervals for the Difference in Two Proportions

1. Use the "**Calendar Year XXXX Large Survey**" to determine if a higher proportion of "***Females***" at Cecil College have "***Tattoos***" when compared to the "***Males***".

We will answer this question with a 95% confidence interval for the true difference in proportions between these two independent groups.

Recall: From Lesson 16, we can run a test using an interval, so that is what we will do on this first example.

1a. The key number in this Lesson 17 is _____.

_____ difference means **no** difference.

If "***Females***" and "***Males***" were **the same** with respect to "***Tattoos***", the difference in the proportions having "***Tattoos***" would be _____.

If a higher proportion of "***Females***" have "***Tattoos***", the difference will become _____ than zero. In our sample data, if that difference becomes "*large enough*", we will reject H_0.

1b. Write the hypotheses for this test:

1c. Summarize the data with a couple of sample proportions.

1d. Determine the **difference** in sample proportions. Understand that the confidence interval is centered at this value.

1e. Show the calculation for a 95% confidence interval for the true difference in proportions. Verify using Stat → Proportions → Two Samples → With Summary.

1f. Interpret your interval with a sentence in context: _____

_____

_____

_____

1g. Can we conclude that there is a statistically significant difference in the proportions of each gender who have "***Tattoos***"? Why or why not?

_____

_____

_____

2. Run the complete hypothesis test. We'd like to know if there is a difference between "***Republicans***" and "***Democrats***" at Cecil College with respect to "***Religion***" = "***Christian***". Use the "**Calendar Year XXXX Large Survey**".

2a. Summarize the data and get the difference in proportions:

2b. Don't forget to check (think about) the conditions.

2c. Hypotheses:

2d. Show the calculation for the test statistic:

2e. Show the shaded Normal model to get the P-value:

2f. Now verify your results using Stat → Proportion Stats → Two Sample → With Summary.

2g. Decision: _____

2h. Concluding remark. _____

_____

_____

_____

2i. Support your results with a 99% confidence interval for the true difference in proportions:

_____

3. Same dataset, "**Calendar Year XXXX Large Survey**". Test if a higher proportion of "***Teenagers***" use "***Instagram***" compared to students in their "***20s***". Don't forget we have the "***Age Category***" variable and let's lump all "***Instagram***" users together. Conditions should be met.

3a. Hypotheses: _____

3b. Summarized data plus the difference in proportions:

3c. Test Statistic: _____ P-value: _____

3d. Decision: _____

3e. Concluding remark: _____

_____

_____

3f. 95% CI for the true difference in proportions: _____

3g. If we made an error in our test, what kind? Comment on the implications of making the error.

_____

_____

_____

_____

3h. On StatCrunch, the Confidence Interval output window will show a standard error. Interpret its value.

_____

_____

_____

_____

3i. Also, interpret the value of the test statistic with a sentence in the context of the problem.

_____

_____

_____

_____

3j. Also, interpret the P-value with a sentence in the context of the problem.

_____

_____

_____

_____

Lesson 18 Formulas:

Lesson 18 StatCrunch Moves: **Stat → Calculators → Normal**

Sample Mean Distribution

1. We switch to **quantitative** variables and we will begin running **tests** and calculating **confidence intervals** for the **population mean**.

The theory behind all of this is that the sample mean, _____, follows an approximate Normal model as long as one of two conditions is met:

1. _____

2. _____

The theoretical mean of the sample mean is: _____

The theoretical standard deviation of the sample mean is:

2. Load up the fake population dataset "**Cecil College Student Ages**" on StatCrunch.

2a. Summarize the **population** with a sentence or two. (In other words, describe the distribution, shape, center, spread, unusual features, outliers, remember?)

_____

_____

_____

_____

_____

2b. Diagram Population vs. Sample

2c. Let's build the sampling distribution model for the sample mean with a simulation.

Fire up Applets → Sampling Distributions → From data table → Values in Age

We are sampling multiple random samples of size $n = 68$ from our population of size $N = 2861$.

The mean of the population is $\mu \approx 25.95$ and the standard deviation of the population is $\sigma \approx 8.25$.

2d. Hit "Sample 1 Time", record your sample mean on the post-it. **Round to one decimal!**

2e. Repeat a total of up to 10 times, so that you have up to 10 sample means on post-it notes. Your instructor will let you know how many to generate.

$\bar{y}_1 =$ _____ $\bar{y}_2 =$ _____ $\bar{y}_3 =$ _____ $\bar{y}_4 =$ _____
$\bar{y}_5 =$ _____

$\bar{y}_6 =$ _____ $\bar{y}_7 =$ _____ $\bar{y}_8 =$ _____ $\bar{y}_9 =$ _____ $\bar{y}_{10} =$ _____

2f. Build the sampling distribution on the white board as we did for sample proportions.

2g. With repeated samples of size $n = 68$, we expect the mean of the sampling distribution to be _____.

2h. With repeated samples of size $n = 68$, we expect the standard deviation of the sampling distribution to be:

2h. Since the sample size exceeds _____, the distribution of the sample mean is _____.

2i. You can hammer on the "1000 times" button now. Notice the histogram is not perfectly Normal, but the model should be close enough to Normal to allow us to compute confidence intervals and run hypothesis tests. (The reason is that our population is very skewed right and our sample size of $n = 68$ might actually not be large enough for the approximate Normality of $\bar{y}$ to kick in.)

2j. "Reset" and switch to a sample size of $n = 2$. Run 1000s of samples and notice that the Central Limit Theorem does not kick in to give us Normality in the sample mean. The sample size is too small. Explore the distribution of sample means.

2k. Above, $n = 68$ had some issues in the far tails with the shape of the distribution of sample means. Probably not enough to really impact any inference we would do on this population.
Switch to $n = 1000$ and notice that with such a large sample size, the distribution of sample means is now very, very close to being Normal.

3. Go again. This time, use the fake population dataset "**Cecil College Male Heights (Fake Data)**".

3a. Run a histogram to assess the shape of the distribution of "***Height***".

Clearly "***Height***" is _____ and _____ and very likely to be _____.

3b. Run the Applets → Sampling Distributions → From data table with various sample sizes.

No matter if $n = 2$, $n = 5$, $n = 29$, or $n = 330$, the distribution of the sample mean is _____.

This is because our population is _____.

4. **"*IQ Scores*"** are known to follow a Normal distribution with a population mean of 100 and a population standard deviation of 15.

4a. What's the chance that **a** randomly selected person has an IQ under 95?

4b. What's the chance that **two** randomly selected people **both** have an IQ under 95?

4c. What's the chance that **two** randomly selected people have a mean IQ under 95? Notice the important difference between 3b and 3c. In 3b, we are using the multiplication rule for two independent events. In 3c, and for the rest of the course, we are taking a sample and then determining probabilities associated with the sample mean from that sample. Big difference!

4d. What's the chance that **five** randomly selected people have a mean IQ under 95?

4e. What's the chance that **thirty** randomly selected people have a mean IQ under 95?

4f. What's the chance that **<u>one-hundred</u>** randomly selected people have a mean IQ under 95?

4g. What's the chance that **<u>one-thousand</u>** random people have a mean IQ under 95? Remember, we are assuming we are randomly pulling from a population with a mean IQ of 100!

4h. Suppose $n = 64$ Clevelanders are randomly selected and their mean IQ is 95. Begin by assuming that Clevelanders are typical, have a mean IQ of 100, and so on. What does this random sample and resulting sample mean tell us about Clevelanders? One of two things is true. Which is more likely? Do you see this is prequel to hypothesis testing?

5. Suppose a certain rope bridge in a national park is built to hold a maximum of 10,000 pounds of people. We're not saying it will break if the weight exceeds 10,000 pounds, but I wouldn't be caught on the bridge in that instance.

Suppose the average American weighs 177.8 pounds with a standard deviation of 30.9 pounds. Pretend only full-grown adults visit this bridge (totally unrealistic, but go with it).

5a. On a moderately busy day, as many as 50 people can be on the bridge at once. What is the probability that the sum weight exceeds 10,000 pounds?

5b. Although signs exist restricting the maximum number of people on the bridge to 50, one 4th of July weekend, 55 people were on the bridge. What's the probability that the weight of the people exceeds 10,000 pounds?

Lesson 19 Formulas:

Lesson 19 StatCrunch Moves:

Get the summary statistics *n*, $\bar{y}$, and *s* using Stat → Summary Stat→ Column

Get *t* using Stat → Calculators → T

Run one-sample confidence interval for a mean using Stat → T Stats → One Sample → With Data

Confidence Interval for Means

1. NFL games are lasting longer and longer due to rule changes, excessive instant replay, and TV timeouts. The TV networks used to give a 3-hour time slot for an NFL game, but are considering increasing the time-slot to 3.5 hours to accommodate the ever-longer game times.

Also, we are interested if attendance figures are changing. The league is very dynamic, yet many polarizing news stories have broken recently and perhaps this is affecting fan interest.

We have taken an unbiased sample of NFL games from last season.
You will find it here: "**NFL Games**" dataset on StatCrunch.

If there is evidence that the **mean** game duration of all NFL games now exceeds 180 minutes, the TV networks will consider slotting NFL games at 3.5 hours = 210 minutes starting next season.

1a. What conditions are required to be met to run a one-sample *t*-interval for the true mean game "***Duration***"? Always check a graph.

1b. Use StatCrunch to obtain the appropriate summary statistics.

1c. Use the formula first and then demonstrate the StatCrunch features to determine the 95% confidence interval for the true mean "***Duration***". Show the *t* distribution.

1d. Interpret the interval with a sentence in context: _____

_____

_____

_____

1e. What should the TV networks do about the time slot? Should it be increased to 3.5 hours?

_____

_____

_____

1f. **More important than it might seem**: If we had a couple of bored summer interns with nothing to do, explain a strategy for the TV networks so that they would not have to rely on confidence intervals or hypothesis tests.

_____

_____

_____

2. The average NFL attendance back in 2013 was 67,101 per game. Is there evidence that the mean attendance is now higher?

2a. Even though we haven't started hypothesis testing for means yet, give the hypotheses:

_____

2b. Summarize the data with the appropriate summary statistics:

2b. Run a 99% confidence interval for the true mean NFL attendance: _____

2c. Make a concluding remark in the context of the problem. _____

_____

_____

2d. Recall the easy margin of error formula is: _____

Calculate the margin of error:

2e. What is the required sample size if we'd like to reduce margin of error down to 1,500 people?

3. Professor Kupe would like to investigate the true mean salary of community college math professors. He wants to be within $2,000 of the true mean and to be 95% confident. He will survey other math professors across the country and ask for salary information. What sample size should he target?[5]

[5] Rule of Thumb: Standard Deviation ≈ Range / 6, but please, this is just one rule of thumb for estimating.

Lesson 20 Formulas:

Lesson 20 StatCrunch Moves:

Get the summary statistics *n*, $\bar{y}$, and *s* using Stat → Summary Stat→ Column

Get P-values using Stat → Calculators → T

Run one-sample T-Test using Stat → T Stats → One Sample → With Data

T-Tests

1. Let's run the T test to determine if the mean NFL game "***Duration***" exceeds 180 minutes (3 hours). Use the NFL data collected in class during the previous meeting. Show all steps and shade under the *t* distribution to show the P-value. Keep in mind we know the result from last time. Show all steps, formulas, and then verify on StatCrunch. Use the "**NFL Games**" dataset.

2. Use the "**Calendar Year XXXX Library Data**" to determine if the mean "***Copyright***" is newer than 1980.

2a. Run the test, show all steps.

2b. If we made a mistake, it would be Type _____. This means _____

_____

_____

2c. Interpret the test statistic with a sentence in context: _____

_____

_____

_____

2d. Interpret the standard error with a sentence in context: _____

_____

_____

_____

2e. Interpret the P-value with a sentence in context: _____

_____

_____

_____

3. Professional golfers must use golf balls with a maximum initial velocity of less than 250 feet per second. Surely, golfers are always trying to get an edge by using the very best equipment, and equipment manufacturers are trying to push the limits of what equipment can qualify as "legal" for tournament play.

3a. Suppose Nike designs a new top-performance golf ball and submits a random sample of $n = 20$ golf balls for testing to the USGA. The USGA will run a hypothesis test to determine if the mean maximum initial velocity exceeds 250 feet per second. Set up the hypotheses.

3b. Suppose the hypothesis test is run and the P-value is 0.3058. Will the new golf balls be approved for play?

3c. Give all possible details you can about the value of the sample mean.

3d. What kind of error could have been made at the conclusion of this hypothesis test? What are the implications?

Lesson 21 Formulas (Independent Samples):

Lesson 21 StatCrunch Moves (Probably need Grouping) (Independent Samples):

Get the summary statistics n_1, n_2, $\bar{y}_1$ $\bar{y}_2$, s_1, s_2 using Stat → Summary Stat→ Column

Get P-values using Stat → Calculators → T

Run two-sample T-Tests using Stat → T Stats → Two Sample → With Data

Lesson 21 StatCrunch Moves (Dependent Samples):

First, compute a column of differences using Data → Compute Expression

Then, get the summary statistics for the new columns using Stat → Summary Stat→ Column

Get P-values using Stat → Calculators → T

Run a one-sample T-Tests using Stat → T Stats → One Sample → With Data

Independent Samples: Testing for the Difference in Two Means

1. Let's test if there is a gender-specific difference between "***Males***" and "***Females***" at Cecil College for what is thought to be the ideal "***Marriage Age***". Specifically, test if "***Males***" want to push off marriage longer than "***Females***". Run a two-sample *t*-test for the difference in two means. Use the "**Calendar Year XXXX Large Survey**".

1a. What are the relevant summary statistics for this analysis?

1b. Conditions met?

1c. Write the hypotheses: _____

1d. Convert the summary statistics to the test statistic using the formula. Then confirm on StatCrunch with Stat → T → Two Sample → Data or Summary.

Uncheck "Pool Variances". Always.

1e. P-value: _____ Decision: _____

1f. Write a concluding remark in the context of the problem: _____

_____

_____

_____

1g. Show the calculation for a 95% CI for the difference in mean "***Marriage Ages***".

2. Using the "**IMDB Movie Ratings**" sample dataset, let's test if movies from 2004 rated higher on average than movies from 2003. Check conditions. Run the test at the 5% significance level.

2a. Hypotheses: _____

2b. Proper summary statistics:

2c. Test statistic: _____

2d. P-value: _____Decision: _____

2e. Make a concluding remark: _____

_____

_____

_____

2f. Remember this dataset with $n = 2410$ movies is just a sample of the population of $N = 30{,}681$ movies. In the population dataset, we have the following:

$$\mu_{2003} = 6.267 \quad \text{and} \quad \mu_{2004} = 6.252$$

Did we make the correct decision and conclusion? _____

If we didn't, what type of error did we commit? _____

With a P-value of _____, the chance of committing this error was small, but after revealing the population data, we learn that **we actually did commit this error**. Always be aware that with hypothesis testing, we can make mistakes in our conclusions, though the chances of making mistakes is typically small.

2g. Explain with a sentence the meaning of $\text{SE}\left(\bar{y}_{2004} - \bar{y}_{2003}\right) = 0.207$.

_____

_____

_____

Dependent Samples: Testing for the Mean Difference

We move to dependent samples. With means, sometimes the two samples are dependent on each other. In this case, we take the difference from each pair of data values and run a one-sample *t*-test on the column of differences.

Essentially, you are taking two columns of dependent data values, subtracting one from the other, and running a one-sample *t*-test on the column of differences. This boils down to Lesson 19 and 20 material.

3. Use our "**Calendar Year XXXX Arm Span Height**" data from earlier in the semester.

3a. These are paired / dependent data values – same student, two measurements. Create a column of differences using Data → Compute Expression. Take "***Height***" – "***Arm Span***". Check conditions and give the summary statistics on the column of differences.

3b. Run the correct hypothesis test to determine if the mean difference is greater than zero. It should be zero, right? We shouldn't expect that on average, our "***Height***" is greater than our "***Armspan***". Show all steps.

3c. If we made an error, what kind, and what does that mean? _____

_____

_____

3d. Interpret the P-value with a sentence in context: _____

_____

_____

_____

Exam III Review

1a. A premier restaurant down by the Chesapeake offers lobster as the special on Saturday nights. Over time, the owner has realized that about 28% of customers order the lobster.

A typical Saturday night expects 305 customers. Determine the correct model for the proportion of customers that will order the lobster. What conditions must be met? No need to diagram, but we could if we wanted to.

1b. How many lobsters does the owner expect to sell? _____

1c. If the owner wants to be 90% sure he doesn't run out of lobster, how many should he have on hand this Saturday?

1d. What if the owner wanted to be 99% sure?

1e. Give an interval of values that will contain the central 95% of Saturday nights for number of lobsters ordered.

2a. A 95% confidence interval for the true mean "***Weight***" of NCAA linemen, based on a random sample of $n = 50$ players, was (254.2 lbs., 275.8 lbs.). What was the sample mean?

2b. What was the sample standard deviation?

2c. The margin of error was 10.8 pounds. If an NFL scout wants to get that down to 5 pounds, what sample size should be taken? Keep 95% confidence.

3. Open up the "**Titanic**" dataset. Explain why we do not need to make any confidence intervals or run any hypothesis tests on this dataset to make conclusions about the fate of the passengers.

4a. Open up the "**StatCrunch U**" dataset. This school has over 40,000 students. Is there evidence that the students at this college have a mean credit hour load exceeding 12 (full-time status). Run the appropriate test, start to finish. Use the variable credit "***Hours***".

4b. Explain the meaning of the P-Value = 0.0273 in the context of the problem.

5. We ran a hypothesis test and failed to reject at the $\alpha = 0.01$ significance level.

5a. Would we reject at the 10% significance level? Yes No Can't Tell

5b. Would we reject at the 5% significance level? Yes No Can't Tell

5c. Would we reject at the 2% significance level? Yes No Can't Tell

5d. Would we reject at the ½% significance level? Yes No Can't Tell

6. Use the "**Calendar Year XXXX Large Survey**" dataset. Run the complete test to determine if "***Males***" at Cecil College take more "***College Credits***" than "***Females***", on average. Show all steps.

7. The reason we use the Student's *t* model instead of the normal model for inference about means is because _____ is unknown.

8. _____ _____ means that the collected data are too unusual to attribute to chance (or that we rejected the null).

_____ _____ data means that we are actually going to take action or that the results are meaningful in the context of the problem.

9. ABC News reports that a certain candidate has 54% of the votes, based on exit polling. The interval had a margin of error of 5%. Explain why the poll is inconclusive if a majority is required to win the election.

10. Which hypothesis is presumed true before you collect your data ? _____

11. We are going to collect some data in the county – we'd like to know what proportion of households have high-speed internet. There are 42,113 households total, far too many to take a census. A survey done by the National Telecommunications and Information Administration done in 2013 claims that 72.4% of all households in the USA have high-speed at home.

If we'd like to be 99% confident and can live with a 4% margin of error, determine the required number of households we will need to survey.

12. Open up the "**Walmart Supermarket**" dataset on StatCrunch. A few semesters ago, Math 127 students visited Wal-Mart and a supermarket of their choosing and collected the price on a single food item at both stores. We'd like to test if prices at the supermarket are more expensive.

12a. This is a two-sample mean problem with dependent samples. Create a column of differences in StatCrunch.

12b. Run the one-sample *t* test on the column of differences to test if the mean difference is positive, indicating that supermarket prices are higher than Wal-Mart prices. Show all steps.

12c. Give an interval of plausible values for exactly how much more expensive food items are at the supermarket. Use 95% confidence.

12d. Making a mistake after running this hypothesis test would imply what?

13. Data collected from a national poll showed that 46% of college-educated people and 40% non-college educated people approved of the Obama presidential administration. If these proportions were based on samples of size 250 from each group, run the appropriate hypothesis test to determine if this is evidence that a higher proportion of people with college degrees approve of the former president. Show all steps.

14. A two-sample *t* test was run to test for a difference in means. The test statistic was $t = -0.8$. What were the hypotheses and what was the P-value if $df = 112$? Draw the P-value shaded model. What would the decision be?

15. Give a range of plausible values at the 99% confidence level for the percentage of all colleges in the USA that are in "***Large Cities***". Use the "**Bachelor's Degree Institutions**" sample dataset.

16. Suppose the average waist size of American men in their 40s is 37.9 inches with a true standard deviation of 4.1 inches. The distribution is moderately skewed right.

16a. If we want to model the sample mean with a Normal model, we need a sample size of _____.

16b. Determine the mean and standard deviation of the model for the sample mean for sample size of $n = 48$.

16c. What's the probability that a sample of 48 men have an average waist size under 36"?

16d. What's the probability the mean waist size is within one inch of 37.9"?

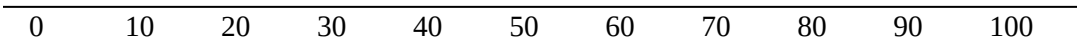
Understanding the Construction of a Histogram

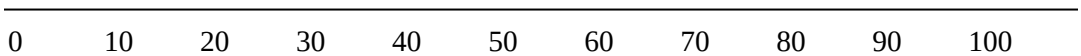
Example: Rate yourself on a scale of 0 to 100 on your skill as a driver. Let 0 indicate you are the worst driver on the road, let 100 indicate you are the best driver on the road, and if you don't drive, then you'll just have to sit this one out. Write your rating here _____. As a class, we will make a dotplot, then a histogram, and then a stem-and-leaf plot.

Write the class ratings here:

For small class sizes, you can add these data values from a previous class:

91 50 85 90 95 70 75 88 81 100 93 75 74



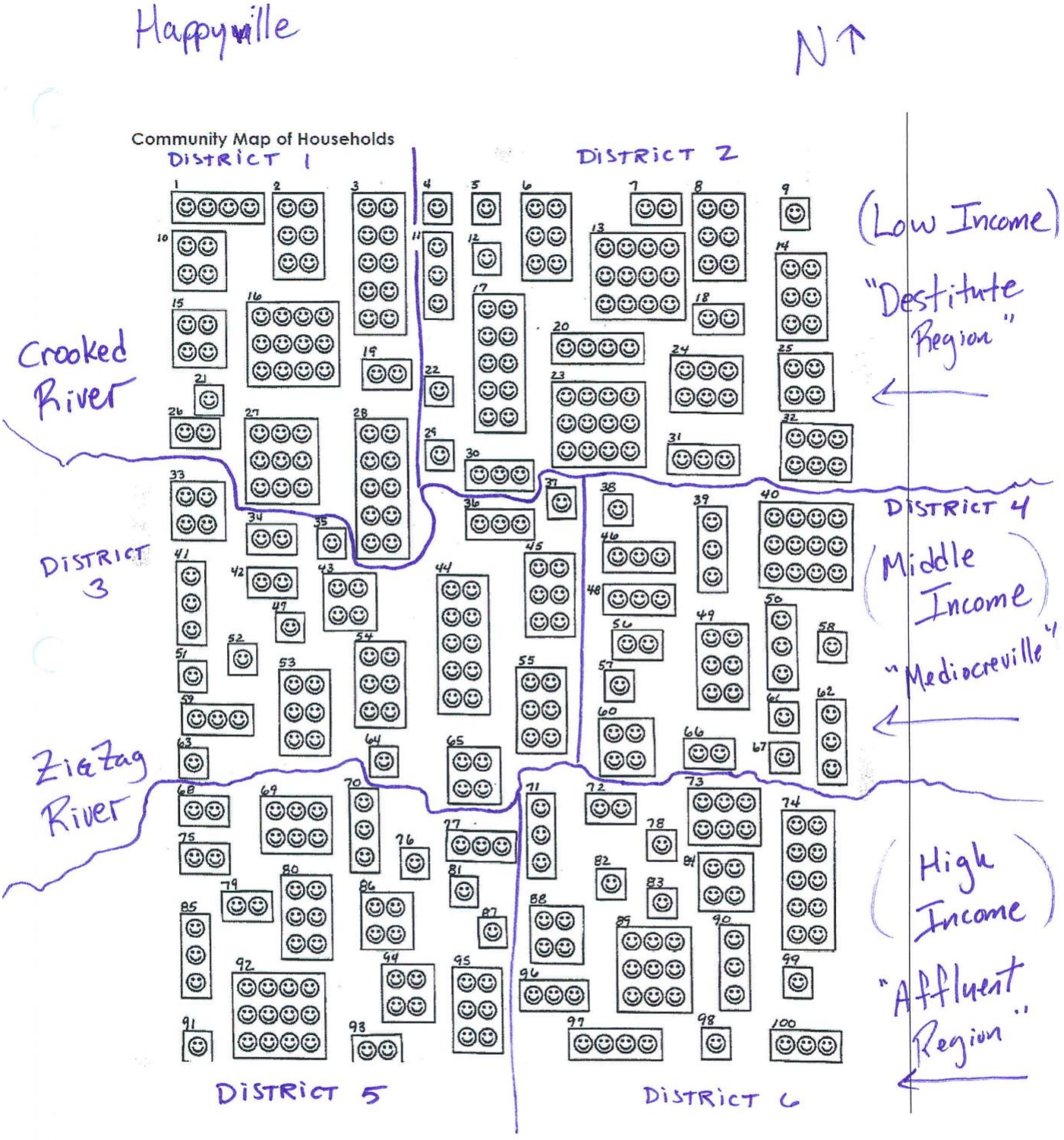


Happyville

A lot of happy people live in Happyville. We'd like to investigate the average household size.

The community is split into three distinct regions, up north is the "***Destitute Region***", which is primarily comprised of low-income families. In between the two rivers is "***Mediocreville***", with middle-income families. Down south, all the rich folks live in the "***Affluent Region***".

The town is also split into 6 voting districts. Districts 1 and 2 are in "***Destitute Region***", Districts 3 and 4 are in "***Mediocreville***", and Districts 5 and 6 are in the "***Affluent Region***".


Happyville
N
Community Map of Households
DISTRICT 1
DISTRICT 2
DISTRICT 3
DISTRICT 4
DISTRICT 5
DISTRICT 6
Crooked River
ZigZag River
(Low Income)
"Destitute Region"
(Middle Income)
"Mediocreville"
(High Income)
"Affluent Region"

1. Use StatCrunch to take a **simple random sample** of $n = 15$ households. Determine the sample mean for number of residents per household.

Create the list 1, 2, …, 100 in a column in StatCrunch (Data → Sequence).
Randomly select 15 households (Data → Sample)

Household Number															
Number of Residents															

Determine the sample mean number of residents: _____

2. Use StatCrunch to take a **stratified sample** of $n = 5$ residents from each of the three regions.

Create the list 1, 2, …, 32 (Data → Sequence) and then randomly select 5 (Data → Sample) for "***Destitute Region***".

Household Number					
Number of Residents					

Determine the sample mean for "***Destitute Region***": _____

Create the list 33, 34, … 67 (Data → Sequence) and then randomly select 5 (Data → Sample) for "***Mediocreville***".

Household Number					
Number of Residents					

Determine the sample mean for "***Mediocreville***": _____

Create the list 68, 69, … 100 (Data → Sequence) and then randomly select 5 (Data → Sample) for "***Affluent Region***".

Household Number					
Number of Residents					

Determine the sample mean for "***Affluent Region***": _____

Does it appear that the three regions have different average household sizes?

3. Use StatCrunch to take a **systematic sample** of $n = 15$ households. We will select every 6^{th} house, but we need to randomly pick a starting point to start counting.

Pick a random starting point:
(Data → Simulate → Discrete Uniform, Row: 1, Column: 1, Minimum: 1, Maximum: 10)

Household Number															
Number of Residents															

Determine the sample mean number of residents: _____

4. Happyville is also broken down into voting districts. Take a cluster sample by selecting one voting district at random and sampling every household in that voting district. Use StatCrunch to randomly select your district.

Voting district: _____ Sample mean: _____

5. Now it's census time. Too bad you don't have a team of interns.

Determine the population mean number of residents: _____

Randomization Tests For The Difference In Two Means

1. Scenario:

- Back in Spring 2011, a Math 127 professor gave two versions of the midterm to the twenty students – here are the data.
- As occasionally happens, the students began to complain that the Green Version was harder and therefore "unfair" to those who took that test.

Green	70	78	62	93	51	72	58	84	91	62
Yellow	60	53	88	75	96	85	68	90	90	76

Calculate the Mean for Green: _____

Calculate the Mean for Yellow: _____

Determine the Difference: _____

2. The main question – "Was our difference of 6 points between the midterm versions because the green version was harder or was the difference due to the _____ of test versions when the instructor distributed the exams?"

3. We have competing theories:

Theory A (The Null Hypothesis) (The Professor's Theory):

_____

Theory B (The Alternative Hypothesis) (The Students' Theory):

_____

4. Question: Which theory is presumed true before we run the test? _____

5. Question: Who is the burden of proof on? _____

6. Task: Put this scenario in the context of a **randomized designed experiment**:

Response Variable: _____

Factor and Levels: _____

Blind or Double Blind: _____ Replication: _____

Activity: Run a Randomization Test

7. Take the 20 cards (each has a test score) and shuffle them up very thoroughly. Deal out the cards into two piles (at random, of course). List the scores below.

Green										
Yellow										

8. Calculate the mean score for each group:

Mean score for the Green Version: _____

Mean score for the Yellow Version: _____

Subtract Green – Yellow: _____ (sign matters, plus or minus!)

9. Write your difference in means on the post-it note and place it on the white board.

10. Now look at the post-it notes on the board.

The biggest difference in mean scores was _____.

The number of differences that were at least 6 points was _____.

The proportion of differences that were at least 6 points was _____.

11. Are we convinced that **Theory A** or **Theory B** is the best choice? What should we do?

_____

12. We can run a simulation on StatCrunch. Load up the dataset "**Green Versus Yellow Midterm**".

Click Applets → Resampling → Randomization Test For Two Means.

Altogether, we will shuffle the exam scores and compute the difference at least 10,000 times.

13. Using the simulation data, the proportion of differences that were at least 6 points was _____. We call this a _____.

14. P-Value Definition (for this example):

If the versions of the midterm are no different (null hypothesis true), the P-Value is the proportion of post-it notes that would have a difference in means of at least 6 points when we randomize test versions.

When we randomized 10,000+ times, we got a P-Value of _____.

Is this proportion unusual? _____.

What is unusual?

- Said another way, **<u>if the versions of the test are no different</u>**, we would get a difference in means of at least 6 points _____% of the time if we randomize the midterm version scores.

P-Value Definition (in general):

The P-Value is the probability or proportion of times we would collect data like ours, or even more unusual data, if the null hypothesis is true.

15. If the students had a valid complaint, it should be relatively difficult to randomize and get a difference in mean scores exceeding _____.

We were able to beat what the students were complaining about _____ of the time, just by randomizing the scores.

16. Which competing theory should we go with?

Theory A (Versions of the midterm are no different)

Theory B (The Green Version is harder)

17. Did we just prove conclusively that the versions are equal in difficulty? _____

What did we show?

18. Statistical Significance:

If your results / data are surprising, assuming the skeptical theory or null hypothesis is true, we say the data are **statistically significant**.

For this example, the actual midterm scores are / are not statistically significant.

For this example, the professor should / should not feel satisfied that the versions were probably fair.

19. The kicker: _____

20. Well, nothing's ever easy. Along came the final exam. Here we go again.

Green	51	55	62	50	61	62	60	81	91	88
Yellow	63	66	74	75	90	88	90	96	90	96

Run the randomization test as before:

Mean for Green: _____ Mean for Yellow: _____

Difference: _____

Null Hypothesis: Way 1: _____

Null Hypothesis: Way 2: _____

Alternative Hypothesis: Way 1: _____

Alternative Hypothesis: Way 2: _____

P-Value After 10,000 Randomizations: _____

Decision: _____

_____

_____

Are these results statistically significant? _____

_____

_____