MA240 Elementary Statistics Problems

Sterling Math Department

Welcome to Elementary Statistics! The flow of this class will likely be different from most other math classes you have had. In this class, there will be only limited lecture. The bulk of the class time will be spent on discussion and going over Problems from the list on the following pages. Allow me to prepare you for what lies ahead by explaining the expectations for class time:

- First of all, make sure you do the assigned problems BEFORE class. This ensures that we can adequately discuss and talk about them during class time. Forgetting once in a while is one thing, but success in this class is dependent on doing the assigned questions before each class. This is your homework in the class.
- Another quirk you'll notice is that there are no chapters, no sections, no headings, and no titles of what we are talking about. It is just problem after problem after problem. This is intentional. We want you to approach these problems without the aid of a title to clue you in or artificial breaks like chapter and section numbers.
- You will typically have 12 problems due before each class, but this may vary between 10 and 15. In a 3-day-a-week section, you will typically have 8 problems that could vary between 6 and 10.
- At the start of class, students will go to the board, place their name and problem number on the board, and write out their proposed solution. This will be done for all of the problems. You will receive 1 point for Class Discourse for each problem you put on the board. Each student is limited to one problem a day on the board.
- During each class, one class member will be designated the Leader of that day's discussion. It will be up to them to go through the questions aloud with the class and call on classmates to participate.
- The professor will serve as facilitator, asking questions and guiding the conversation as appropriate. However, the professor will NOT largely lecture and lead the class, except when needed.
- The expectation is that EVERY student will participate each day. You will receive a grade for Class Discourse each week that is largely based on your participation in these discussions. You also receive 1 point in Class Discourse for putting a problem solution on the board. Over the course of the semester, you can earn up to 50 points for Class Discourse.
- It is okay not to be correct! It is OKAY to struggle! True learning often involves a productive struggle. The best learning often happens when we really don't know how to proceed. Embrace this.

- It is the expectation that every student will be taking notes as we discuss in class. This is not required and will not be graded, but since there is no lecture for this class, the large base of notes will be created by YOU. So make sure to spend effort and time creating good notes each class period. This may take some practice.
- When we do reach the part in the course where statistical tests of inference are done, the Appendix B in the back of this document has extra information and some practice problems that we will go over in class to supplement the Problems. Make use of this.
- Throughout the text are bold words - these are important vocabulary. You can also find these words in the Glossary in Appendix A.
- Especially in the latter half of the course, work on the TI84 calculator is emphasized. In class, we will also introduce the R programming language to solve statistical problems also. Appendix C lists some of the common TI84 commands. More info on these can be found by an Internet search.
- Tests will occur roughly once a month, and there is no final exam. Instead, you will complete a Culminating Project worth two test grades.
- The Culminating Project is detailed to some extent in Appendix D. This will be your chance to show mastery of the material.
- Above all, we hope that this journey into Elementary Statistics proves to be a potent investigation into the world of statistics and that you leave with a deep appreciation for it.

Build-Your-Own-Textbook!

Surprise - there is no textbook for this class! Or rather, there is no textbook for this class...yet! Over the course of this semester, you will build your own textbook for this class. This will count for a test grade at the end of the semester.

- As we answer Problems in class, take notes on the answers and about the concepts we discuss. This is NOT your textbook.
- You will notice that many of the questions have themes that tie them together - condense these into different topics that will be in your textbook. For example, you might have a topic for "Ethics" and a topic for "Confidence Intervals".
- Organize your notes and SUMMARIZE them into coherent topics. It is important that you do not simply copy down what we talk about in class for your textbook, but rather contribute your own summarized ideas that you synthesize from what we talked about. I want this to be YOURS, not a regurgitation of what I say.
- Type your results into a document. For formulas and math, you can either use the Formula feature in Word, or you can take a picture and upload the picture into the document.
- Give your sections headings (titles), and number your chapters/sections in a meaningful way.
- Make edits and adjustments, as necessary, over the course of the semester.
- NOTE: Do this from Day 1! Do NOT wait until the week of Finals to start on this, or you will be in a really tough position. If you put together a poor effort, you will get a poor grade.
- Save your textbook! You never know when you may need it again...

Course Outline

- Weeks 1-2: Syllabus, Questions 1-36
- Weeks 3-4: Questions 37-72, Review for Test 1
- Weeks 5-6: Test 1, Questions 73 - 108
- Weeks 7-8: Questions 109-155
- Weeks 9-10: Review for Test 2, Test 2, Questions 156-179
- Weeks 11-12: Questions 180-228
- Weeks 13-14: Questions 229- 266, Review for Test 3, Test 3
- Weeks 15-16: Culminating Project Proposals, Work on Culminating Project, Turn in Culminating Project

Grade Breakdown:

Class Discourse	50 points
Test 1	100 points
Test 2	100 points
Test 3	100 points
BYO Textbook	100 points
Homework	150 points
Culminating Project	200 points
Total:	**800 points**

In order to receive a certain grade, you must earn the following number of points:

A	744 points or more
A-	720-743 points
B+	696 - 719 points
B	664 - 695 points
B-	640 - 663 points
C+	616 - 639 points
C	584 - 615 points
C-	560 - 583 points
D+	536 - 559 points
D	504 - 535 points
D-	480 - 503 points
F	479 points or below

You must receive at least a C- to get credit in the class if it is required for your degree.

Dedicated to Sterling College

We may throw the dice, but the Lord determines how they fall. - Proverbs 16:33

Again I saw that under the sun the race is not to the swift, nor the battle to the strong, nor bread to the wise, nor riches to the intelligent, nor favor to those with knowledge, but time and chance happen to them all. - Ecclesiastes 9:11

1. **Statistical significance** means that an event is unlikely to occur by random chance alone. A famous study from the 1960s explored two dolphins to see if they could communicate abstract ideas. In this experiment, one dolphin was shown one of two lights, and the dolphin had to communicate to their partner (who has behind a barrier) to press the matching light on their side of the pool. In the study, 15 out of 16 trials resulted in success. Do you think this is statistically significant? What if the study had 9 out of 16 successes? Or 12 out of 16 successes? Why? How many trials would have to be correct before you could claim such? Support your decision.

2. Consider a survey done at Sterling College. In this survey, 50 students were randomly surveyed about the average amount of money spent each year on textbooks and related supplies. Three of the students spent \$150, \$275, and \$300, respectively. What aspects of this survey are important to know?

3. A class has their height measured, and the average height is 67 inches. Worldwide, the average height is 64 inches. What is important to know here? What is the difference between the two results?

4. In a study done to determine reading habits of students, the gender of each student and the hours read each week are recorded. Compare and contrast the gender in this case with hours read each week. For what purposes do you imagine they might be used in the study?

5. (Continuation) A **variable** is a characteristic or measurement that can be determined for each member of a population. In a study meant to investigate study habits of students and how they may impact their performance in school, what might some of the variables in this study be? Why do you consider these variables important? Are there differences among the variables? It'd be nice if you could come up with at least 4 variables. Explain

each of them.

6. (Continuation from #2) The **population** is a collection of persons, things, or objects under study, while the **sample** is a portion of the larger population obtained to gain information about the population. Give an example of a population and a sample of that population. Think about the size of a sample versus the population from which it comes. In a population of 500, do you think a sample of 50 or 200 would be better? Why?

7. A survey is conducted to gather information about sentiment towards law enforcement and police brutality. One of the questions asks, on a scale of 1-5 (with 1 being very unsafe and 5 being very safe), how safe do you feel around police? If some type of graphical display is desired for showing this result, how might that be done? What sort of graphic can you come up with?

8. The information in the following list can be separated into three groups - can you figure it out? {The number of pairs of shoes you own, the type of car you drive, the distance it is from your home to school, the number of classes you take each year, the type of calculator you use, weights of classmates, ACT Scores, the number of correct answers on a quiz }

9. (Continuation from #3) A **parameter** is a (usually unknown) number describing the population. A **statistic** is a number that describes a sample. The hope obviously is that a statistic will tell us information about the population. What are some practical considerations to make when it concerns a parameter? Why do you think a parameter is usually unknown? Why would a statistic be useful, if the parameter is usually unknown and unknowable?

10. (Continuation from # 5) **Numerical variables** take on numeric values, while **categorical variables** place each subject into a category. For example, political party affiliation as Republican, Democrat, or Independent is a categorical variable. And gas mileage of a car would be a numeric variable. Think about the structure of numeric and categorical variables. Name two weaknesses and two strengths of each categorical and numeric variables.

11. **Data** is the actual values of the variable. The singular of data is **datum**. Give some made-up example data for variables in a study looking at the eating habits of dogs. Give some made-up example data for variables in a study looking at the sleeping habits of adolescents. How are these alike? How are they different?

12. **Chance models** generate data from random processes. Flipping a coin is such a model and can be to **simulate** data. For an example of such a chance model, consider the game show *Let's Make A Deal* , where a classic dilemma presents itself: Suppose there are 3 doors. Behind one of the doors is the Big Deal, worth thousands of dollars. Behind the other two are trivial prizes worth nothing. You initially choose a door, and then an incorrect door is revealed. The host then asks if you want to switch doors or not - should you? What is the best strategy? You can also conduct a simulation of this scenario by performing it and repeating it many times.

13. The College Bookstore is conducting a survey to see if the students prefer online textbooks or physical textbooks. The questions are collecting information on gender, classification (freshmen, sophomore, etc.), as well as their preference on textbooks. Sixty students participated in the survey. Consider the ideas of population, sample, parameter, statistic, variable, categorical, numeric, and data. Highlight as much information as you can about this survey. What are maybe some ways you might want to analyze this data?

14. On *The Price is Right*, there is a game called Lucky Seven. In this game, the contestant gets $7 in the form of dollar bills. The first number in the price of a car is revealed. The contestant must then guess the next number in the price of the car. If the guess is correct, the game moves on to the next number. If the guess is wrong, for each digit the contestant is off costs them one of their dollars. To win the car, they must have $1 left after the 5-digit price is revealed to buy the car. What strategy is the best strategy to take to win the car?

15. True/False: *A sample should have the same characteristics as the population it is representing.* Is it strictly true/false? Do you agree with it? Is there a problem with this? What are your thoughts?

16. A **pie chart** displays qualitative data. Typically, data in a pie chart shows category ownership and adds up to 100 percent. Do you think it is ever possible for a pie chart to have percentages that add up to over 100? Why or why not?

17. **Probability** is a mathematical tool used to study randomness. It is usually expressed in the form of a decimal or a fraction. Probability deals with the chance or likelihood of an event occurring. When flipping a coin, the outcome if the coin is fair is expected to be 50% heads. That is, there is a 0.5 probability of a heads occurring, and a 0.5 probability of a tails occurring. If you flip 10 coins and all turn out to be tails, what do you think the probability of a heads is on the next flip? Why?

18. A **histogram** displays percents or counts of quantitative data. It is represented as vertical bars that touch each other, with the bottom of the graph being a horizontal axis showing the quantitative variable. Why do you think histogram bars must touch?

19. **Qualitative data** places data into categories using categorical variables. **Quantitative data** are always numbers and can be discrete (can be counted, whole numbers) or continuous (on a continuum, like height or weight, can be a decimal). Imagine you see a large amount of cars in a parking lot. Create a categorical variable, a discrete quantitative variable, and a continuous quantitative variable to describe these cars.

20. (Continuation from #16) If the data in a set adds up to over 100 percent, can you devise a way to display the data? Explain your method.

21. Suppose you want to conduct a survey at Sterling College to see what brand of soft drink is the student favorite. You want to make sure your **sampling method** is as good as possible. What makes a good sample? How might you sample students in a "fair" way? What does fair even mean in this case? Discuss.

22. Sometimes sampling is done, for practical reasons, without replacement. This means once a subject is chosen, it is not eligible to be selected again. When is this fair? When does it make sense? Consider population size - if the population is a class of 20 and you select 1 person, that is a 1/20, or 0.05 chance of being selected. If you then choose a second person without replacement, there is now a 1 in 19 chance, or 0.0526 chance of being selected. Is this change significant? Could it be? Now, consider the population is size 20,000,000 and you select 1 person, then a second person without replacement. How different is it now? What do you feel about population size when it comes to sampling without replacement?

23. Janice is conducting a survey in the cafeteria. She is asking every student that walks by how they feel about the college's latest actions. Does this type of survey seem "fair"? Are there any problems with it? Explain.

24. Are you familiar with rock-paper-scissors? Is it a fair game? Defend your assertion.

25. A **simple random sample**, also called SRS, is conducted when a group of n individuals are equally likely to be chosen as any other group of n individuals. Indeed, the key is randomization. How could this randomization be achieved?

26. (Continuation) Three other variations on the SRS are the **stratified sample** and the **cluster sample** and the **systematic sample**. In a stratified sample, the population is divided into groups and the members of the group are randomly chosen proportionally. In a cluster sample, the population is divided into groups and then some entire groups are randomly surveyed. In a systematic sample, a random starting point is chosen and then every n^{th} subject is chosen. In a phone survey where the survey leaders have a phone book (or spreadsheet with phone numbers), which of these methods would it be best to use? Why?

27. (Continuation from # 24) You and a friend, who has played the game much more than you, conduct a **simulation** to see how fair rock-paper-scissors is. In 12 different rounds, you won 4 games and your friend won 8 games. What is your **proportion** of wins? If you conducted 120 different rounds, how do you feel the proportion would change? What if it were 120,000 rounds?

28. (Continuation) Winning and losing is an example of a **binary variable**, a categorical variable with only two outcomes. In the previous problem, you only won 4 games - does this mean you were unlucky or is your friend better at the game? In other words, you can write these assertions as such: *Experienced rock-paper-scissors players have a greater than 50% chance of winning games in the long run.* And: *Experienced rock-paper-scissors players equal chance (50%) of winning or losing games in the long run.* The first assertion is called the **alternative hypothesis**, declaring that there is an effect

that can be sensed. The second assertion is called the **null hypothesis**, declaring that the results happen by random chance alone. Can you devise a way to test these hypotheses? If not, it's okay - just think about it real hard and explain your thoughts about how one might go about evaluating to see if the null or alternative hypothesis is true.

29. Consider the following data about Sterling College: school colors, classification of students (freshmen, sophomore, etc.), average monthly temperature in Fahrenheit, and annual tuition paid by students. For each of these data sets, differentiate it from the other 3 data sets in as much detail as possible.

30. A survey is conducted online with anyone choosing to respond as the sample. What can you say about the sampling method here? Is it a random sample? Why or why not? Can the results be trusted? Is it black and white or gray? Explain.

31. (Continuation from # 29) Of the school colors and the classification of students from # 29, how did you find they differed? We say that data that are qualitative or categorical are called **nominal data**. It possesses no natural ordering. On the other hand, **ordinal data** have a natural ordering. Which of school colors or student classification is nominal and which is ordinal? Can you come up with another variable for data collection for Sterling College that would also be nominal? ordinal?

32. (Continuation from # 28) We can write the null and alternative hypotheses using symbols to make it easier to keep up with. We often write the null hypothesis H_0 and the alternative hypothesis H_a. Also, the sample size is known as **n**. So to represent the null and alternative hypotheses using symbols, let p be the parameter proportion of wins of

an experienced rock-paper-scissors player. Then:

$$H_a : p > \frac{1}{2}$$

$$H_0 : p = \frac{1}{2}$$

This is called a **one-sided** hypothesis because the alternative hypothesis involves an inequality. Imagine a **two-sided** hypothesis: What would that mean and look like?

33. A survey is conducted among 6 random people about who they are going to vote for in the next Presidential election. What can you say about the sampling here? Do you feel the results can be trusted? If not, why not - it is random! Explain.

34. A doctor's office allows patients to complete a survey after receiving service. Out of 120 patients one week, 13 chose to complete the survey, and only 16% were happy with their service. The doctors are really upset. Should they be? What can you say about the sample here? Are there any issues? Explain.

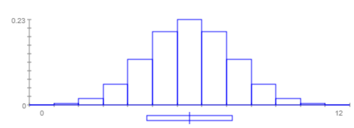
35. (Continuation from # 31) Of the average monthly temperature of Fahrenheit and annual tuition paid by students, there is a key difference. What difference did you find? Data that is numeric and has defined intervals (such as degrees in Fahrenheit) are called **interval data**. Calculations can be done on numeric data, but ratios cannot be done - does it make sense to say 20 degrees is -1 times colder than -20 degrees since $\frac{-20}{20} = -1$? Explain. How could this be fixed? Does the annual tuition paid by students share this problem? If yes, how can it be fixed? If not, what is different?

36. A scientist is conducting an **experiment**, a controlled study to investigate the relationship between two variables. In this experiment, the scientist is measuring the amount of

direct sunlight plants receive daily and measuring their growth on a weekly basis. What do you feel the point of this experiment would be? Can you identify the variables? What key difference is there between the variables?

37. In the election of 2020, a poll conducted at a Trump rally found that 97% of likely voters would be voting for Trump. What can you say about this?

38. (Continuation from #32) Attempting to factor experience and skill into paper-rock-scissors chances seems really difficult - how would you even begin making a model for that? So, to try and model the paper-rock-scissors situation, it makes more sense to begin with the null hypothesis, the chance-alone assertion. Assuming the null hypothesis is true, there should be a 50% chance to win each game. If we conduct a simulation with chance of success 50% and sample size n = 12 and we conduct many samples (such as 1000) and then graph the results in a histogram, we get the following:



We call this the **null distribution**, since it is the values obtained when the null hypothesis is true. Notice the peak of the histogram occurs around 6 games - can you explain this? Will this distribution look the same every time this simulation is run? Explain.

39. (Continuation) The real question is whether our observed statistic differed significantly from that of the null distribution. Recall that 8 out of 12 games, your experienced friend won. That is a $2/3 \approx 0.67$ proportion of successes for your friend. Under the null hypothesis, this proportion should be 0.5 (why?). Where does 8 lie in the null distribution in the previous question? We need a way to determine "how extreme" the 0.67 proportion really is. We define the **p-value** as the **probability** (essentially the chance) of obtaining your observed statistic or one more extreme in the direction of the alternative hypothesis, assuming the null hypothesis is true. Why is it necessary to include "or one more extreme" when computing a p-value?

40. (Continuation) Think about the size of a p-value. A large p-value means it is fairly likely for the observed statistic (or one more extreme) to occur purely by chance, again, because the null hypothesis is assumed to be true. A small p-value means it is unlikely to obtain the result assuming the null hypothesis is true. If you are attempting to accept the alternative hypothesis, are you looking for a small p-value or a high p-value? What threshold of a p-value do you feel is low (or high, if you chose high) enough? Defend your reasoning.

41. In the early-to-mid 1900s and even as late as the 1970s, black males in the South were used as subjects in an experiment about the effects of syphilis, even after penicillin became a widely-available treatment decades earlier. Think about this experiment for a moment. Discuss.

42. A researcher finds through a study that 99.98% of people in car crashes were clothed! Therefore, the researcher concludes that clothes cause car crashes. Discuss.

43. (Continuation from # 35) With annual tuition paid by students at Sterling College, all meaningful calculations can be done on the dollar amounts, including ratios. This means

it is called **ratio data**. What value is there in having a true zero? (By the way, what do I mean when I say "true zero"?) Does it accomplish anything mathematically? How are ratio data more flexible for calculations than interval data?

44. In 1971, the Stanford prison experiment was conducted by psychology professor Phillip Zimbardo using college students. In the study, volunteers were assigned either to be guards or prisoners, randomly, and placed in a mock prison where they had to live out their roles for a week. The students assigned as guards quickly began subjecting the prisoners to psychological torture and harassed the prisoners, and the prisoners largely accepted the assault. This went on for days and escalated until eventually, the study was called off after the abuse began to get physical and obviously off-limits. What are your thoughts on the Stanford prison experiment?

45. A political poll takes an opponent's sentence "Only a horrible human being would say I want to kill babies" and asks respondents if they could support the opponent who said "I want to kill babies." Surprisingly (or not) enough, 98% of people said they could not support the opponent. Discuss.

46. **Explanatory variables** cause a change in a second variable and are also sometimes called the independent variable. **Response variables**, also called dependent variables, are changed by a corresponding change in the explanatory variable. Can you think of a scenario where one variable might cause a change in another? Explain.

47. **Institutional Review Boards (IRBs)** serve research institutions and colleges/universities in conducting studies in an ethical way. In this process, they ensure that risk is minimized to participants, and participants must be given **informed consent**, among other requirements. Informed consent means that the risks and role of the study must clearly be explained to subjects of the study in advance. Do you feel informed consent is im-

portant? Why or why not?

48. Nominal, ordinal, interval, and ratio data are also called **scales of measurement** since they define the regions on which data can be measured. They also define what statistics are possible to calculate for a given data set. For data on a nominal scale, what kind of summary statistic do you think can be done? (Note: We haven't actually discussed any yet - but if you were to design such a statistic, what would it do?)

49. In the following situation, describe any unethical behavior, if any, and how it could impact the reliability of the resulting data: *A researcher is collecting data in the community. She takes a stroll around a particular neighborhood to survey participants because the residents match her race and have middle-upper-class homes, which feel safer to her.*

50. Is it ever appropriate not to have informed consent ahead of time? Consider human behavior and how human behavior changes when presented with information. If there ever is such a situation, how would one go about conducting a study or experiment? Does it still maintain an ethical foundation? Explain.

51. Are chance models reality? Discuss.

52. In the following situation, describe any unethical behavior, if any, and how it could impact the reliability of the resulting data: *A survey leader is conducting a survey in a particular neighborhood. She visits 40 homes to ask her questions. Since she visited*

during the middle of the day, 32 of the homes appeared to have no one home. She recorded only the 8 homes that responded and hoped to add to it with a new neighborhood the next day.

53. A **frequency** is the number of times a value of the data occurs. If we wanted to divide by the total set of all outcomes, we'd arrive at a **relative frequency** instead. Frequency is used to count data values. For what scale of measurement do you feel frequency would be most useful in describing? Why?

54. For each frequency shown in the table below, calculate the relative frequency associated with each data value.

Data Value	Frequency	Relative Frequency
2	3	
3	5	
4	3	
5	6	
6	2	
7	1	

55. (Continuation) The **cumulative relative frequency** is the accumulation of the previous relative frequencies. For example, if the first relative frequency is 0.05 and the second is 0.20, the first two cumulative relative frequencies would be 0.05 and 0.25, since they are added. Calculate the cumulative relative frequencies of each data value in the table above.

56. Is the misuse of statistics a real problem? Many people distrust statistics because of how they can be manipulated. What are some ways you can guard against the misuse of statistics?

57. We have discussed p-values to assess the strength of evidence against the null hypothesis. Another way to find evidence against the null hypothesis is to standardize the value of the observed statistic. To do this, we need two new statistics - the **mean** and the **standard deviation**. For now, think of the mean as the "average" or "middle" of the null distribution. Think of the standard deviation as the "spread" or the average distance of each datum from the mean in the null distribution. We won't actually calculate these now. The formula to standardize the observed statistic is:

$$\text{Standardized statistic } = z = \frac{\text{observed statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}$$

An important idea is that *observed statistics that have z values of 2 or 3 or more can be considered "extreme" and unlikely in the null distribution.* These observations would fall in the **tail** of the distribution. In a hypothetical scenario, suppose the mean is 5 and the standard deviation is 1. Would an observed value of 6 be considered extreme or not? If not, what would?

58. (Continuation) If we were trying to find evidence against the null hypothesis, we can use either standardized statistics or p-value. Consider a statistic with a high standardized z statistic, such as over 2. Would the p-value for the observed statistic be low or high? How do you know?

59. In the following situation, describe any unethical behavior, if any, and how it could impact the reliability of the resulting data: *A study is done about a large national pizza brand. The study is paid for by the brand itself and used for sales purposes. In the study, the pizza and a competitor's pizza are served (in the brand-name boxes) to participants, and the competitor's pizzas are colder and staler. To report their results, they compiled that 44% enjoyed their pizza more, 35% supported the competitor's pizza, and the rest had no preference. In their national ad campaign, the company reported "Most people prefer our pizza over Competitor's Pizza."*

60. Consider the **z standardized statistic** (also often called a "z-score"). In words, what do you think it measures, exactly? It may help to look back at the formula for inspiration and to try a few sample cases.

61. (Continuation from #40) Let's think back to the rock-paper-scissors simulation run by you and your experienced friend. In the 12 trials, your experienced friend won 8 games, and you won 4. This gave your friend a **sample proportion** $\hat{p} = 8/12 \approx 0.67$. (The "hat" on the p identifies it as an estimate of a parameter.) Recall that the null hypothesis is that an experienced player would win a game 50% of the time. This serves as the mean under the null distribution. Let's assume the standard deviation of the null distribution is 3 (we'll spare ourselves that calculation for now). What would the z-score for the observed $\hat{p}$ be?

62. (Continuation) How extreme is the z-score you found in the last problem? If the observed statistic were instead 10, would it be more extreme or less extreme than 8? In general, what can you infer about the observed statistic as the distance from the mean increases?

63. (Continuation) If we increased n from 12 to 120, do you think that would increase, decrease, or have no effect on the strength of the evidence against the null hypothesis? What if $n = 120,000$ instead? In general, can you say anything about the sample size? Note we aren't mathematically proving anything here - this is just your intuition.

64. (Continuation) Given the z-score obtained in #61, do you think there is enough evidence to **reject the null hypothesis**? Or should we **fail to reject the null hypothesis** due to lack of evidence against it? This is called a **test of significance**. We will define more of the numbers soon. Do you think increasing the sample size should make it easier or more difficult to reject the null hypothesis? Note: If we reject the null hypothesis, we are NOT directly saying the alternative hypothesis is true. Rather, we are saying that,

given the evidence we have, we are making a judgment that the alternative hypothesis is the most plausible result. Likewise, if we fail to reject the null hypothesis, we are not saying the null hypothesis is true, but rather that we just lack enough evidence to repudiate it entirely.

65. In a class of eight students, all 8 picked "1" when given the choice of "1" or "2". Is this result surprising? Would it be more surprising if a class of 80 had all chosen "1"? What can you say, generally, about this?

66. In a two-sided test of significance, the alternative hypothesis goes in two directions, not just "greater than" or "less than". The extreme values must go in both directions, not just one. A p-value for a two-sided test, then, must find out how frequently the observed statistic or a more extreme one occurred in one tail of the distribution and adding that to the corresponding probability in the opposite tail of the null distribution. Think about this for a moment. Then, consider this: Will a two-sided test provide more evidence against the null hypothesis than a one-sided test or less? Why do you think this? Because of your answer, which test do you think is used more often in practice?

67. (Continuation) Can you think of a situation where you might want to do a two-sided test of significance rather than a one-sided test?

68. If you knew NOTHING about the content, would you rather have fewer questions on a True/False test or more questions? What significance does this have for sample size and tests of significance?

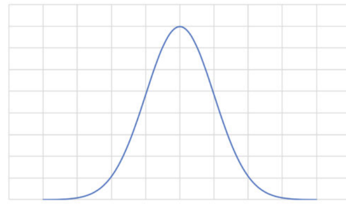
69. Imagine you have lost \$400 in the past 20 rounds of blackjack at the casino. Should you go all-in on your next round since you're obviously due for a good round? Explain.

70. What is the difference between $\hat{p}$ and p? What about the difference between these and p-value?

71. The null distribution of sample proportions that have occurred so far have all had similar characteristics:

- They often follow a bell-shaped curve.
- They are centered at the null hypothesis value for p.
- Their standard deviation (or more generally, their **variability**) is influenced by sample size.

Many of these situations, which occur very often in simulations and real-life data sets, is called **normally distributed** or a **normal distribution**. It is called normal, in part, because of how often it truly does come up in statistics.



There are an infinite number of normal distributions, but they all look similar. The mean of the normal distribution is called the **location parameter**, and the standard deviation is called the **shape parameter**. What do you think each of these parameters will do to the distribution curve?

72. (Continuation) In the early 1900s, in a day and age where computers weren't available, theories and rules were created concerning the Normal Distributions that made calculations easier. One of the largest and most important results, even today, is the famous **Central Limit Theorem**. Very generally speaking, it says that if our sample size, n, is large enough, the distribution of sample proportions will be Normally Distributed, centered at p, and with a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$. How large is large enough? Unfortunately, there is not a magical answer when it becomes true, but larger is always better. Suppose we want to roll a six-sided die and count the proportion that lands on "1". The die is rolled 400 times and comes up with 72 ones. Calculate the z-score associated with this result. Is this an extreme result?

73. a **census** is a survey of the entire population. What are some practical considerations concerning a census? When might a sample be the preferred option instead?

74. The **sample mean** is typically denoted $\bar{x}$, while the sample proportion, as previously

noted, is $\hat{p}$. While the population proportion is sometimes represented as p, it is also represented as Π, the capital Greek letter Pi. Meanwhile the **sample standard deviation** is represented as s, while the **population standard deviation** is represented as σ, the lowercase Greek letter, sigma. Sometimes the expression "That's Greek to me!" to exclaim that some idea is unknown to the speaker; therefore, why do you think Greek letters might be used for population parameters?

75. With categorical variables, can we make inferences about which of the following: the population proportion vs the population mean? With quantitative variables, is this the same case? What is different?

76. Consider what makes a SRS random. To help with this consideration, realize that the most common form of non-SRS sampling is **convenience sampling**, or simply collecting data from whatever sources can be conveniently obtained. A common example of convenience sampling is just asking your classmates your survey. What do you think makes a sampling process truly random? And is it possible to be truly random? What is "good enough"?

77. Read this sentence aloud: "Tall people with suits are suspicious of penguins with tutus." How many seconds do you think it took you? Suppose we have the following claim: "On average, it takes someone 4 seconds to read the given sentence." We are interested in finding out if the actual time taken differs from this. Will this be a one-sided or two-sided hypothesis?

78. A statistic is said to be **unbiased** if the statistic consistently estimates the population parameter. When using simple random sampling, the average of the sample means from different samples equals the population mean, and the average of the sample proportions from different samples equals the population proportion. That's great for simple random

sampling! This allows us to generalize a sample to the larger population and safely assume that the observed sample statistic will be "in the ball park" of the population parameter. What do you think "in the ball park" even means in this case? And how "big" can this ball park be? To answer this, consider what other statistics may be important to factor in. Do you have any ideas?

79. Recall the Central Limit Theorem that tells us if our sample size, n, is large enough, the distribution of sample proportions will be Normally Distributed, centered at p, and with a standard deviation of $\sqrt{\frac{p(1-p)}{n}}$. It also applies to the distribution of sample means as follows: The mean is equal to μ (μ is the Greek letter "mu" and represents the population mean.), and the standard deviation equals $\frac{\sigma}{\sqrt{n}}$. However, we usually don't know σ. Instead, we can estimate it using s, giving $\frac{s}{\sqrt{n}}$, a statistic called the **standard error** of sample means. However, s is NOT unbiased! Do you think this will "mess up" using the Central Limit Theorem?

80. Recall that when the number of samples is large, the standard deviation of the sampling distribution of sample proportions is found using the formula $\sqrt{\frac{p(1-p)}{n}}$. If the population is "large enough" compared to the sample size, this formula actually works for the standard deviation of the sample proportions estimate as well! However, in practice, we rarely know what p (or Π, if using Greek notation) is, so how might we estimate this standard deviation?

81. (Continuation from #78) Simple random sampling produces unbiased estimates for the mean and proportion, but what if we don't have simple random sampling? Suppose we have convenience sampling and ask the first 30 people to come through the cafeteria - do you feel this will have the same unbiased property? Why or why not? In the absence of SRS, do you think we can still safely generalize to the population?

82. A p-value of 0.05 is often considered the statistically significant threshold; that is, a result lower than 0.05 p-value is considered ”extreme” and unlikely to occur by chance alone. This number, 0.05, is largely arbitrary and was chosen by a prominent statistician in the early 1900s. Do you feel this 5% number is a fair number? Are there instances where a p-value of 0.05 may not be appropriate?

83. (Continuation from #80) How ”large” is large enough for the sample proportion standard deviation formula to apply? Convention says that 20 times more than the sample is a large enough population. This means the sample must be less than or equal to 5% of the population. Why do you think this is?

84. (Continuation of #79) Using the Central Limit Theorem, we can find our z-statistic of sample means, using the information stated in #79. Recall that to standardize a statistic, we can use the following formula:

$$\text{Standardized statistic } = z = \frac{\text{observed statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}$$

This yields us:

$$z = \frac{x - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

where x is the observed statistic, and μ_0 is the hypothesized value of the population mean under the null distribution. But as stated, we rarely know what σ is, so if we estimate it using s, we get:

$$z = \frac{x - \mu_0}{\frac{s}{\sqrt{n}}}$$

However, this is wrong! This is NOT a z-statistic! In fact, the usage of s makes it not a Normally distributed variable anymore. In fact, this is a new distribution, called a **t-distribution**. Like the Normal, this is a mound-shaped distribution, but has more variability. Likewise, the statistical significance test to estimate a single mean is called the **one-sample t-test**. For the sample proportions test, a test statistic of greater than 2 or less than -2 was considered ”extreme”; do you think this guideline also applies for the one-sample t-test? Why do you think that?

85. When decision-making is important to consider, the 0.05 p-value threshold can be used in a test of significance to reject the null hypothesis or fail to reject the null hypothesis. This is called α, or the **significance level**. A certain hospital has had 8 of the last 10 patients die after having heart transplants. It would be important to know if this was purely chance or if something worse was happening. The probability of one such patient dying after a heart transplant is 0.456, but as you increase it to 8, the probability shrinks to 0.00001. Should you reject the null hypothesis? (What IS the null hypothesis?) Do you believe a 5% threshold is a good threshold for this scenario?

86. (Continuation from # 77) Let $H_a : \mu \neq 4$ and $H_0 : \mu = 4$. In a sample of 20 students, they gave the following data for the number of seconds it took to read the require sentence: $\{2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 8, 10, 10, 12\}$ Develop a graphical display for this data that makes sense. Why did you choose the way you did? What can you notice about the distribution of data? Does it have the same characteristics at first blush as the Normal distribution or not?

87. (Continuation from #85) It seems silly that the α-level has to be even set at all; after all, why accept the outcome occurring by chance 5% of the time? Why not set it to 1%? That would make your results more accurate, right? Even still, why not set it to 0% and don't leave anything to chance - this would be ideal, right? What is flawed about this line of thinking?

88. (Continuation from #86) To calculate the mean, we must add up the data points and divide by n, the sample size. What is the mean of the data in #86? Does this number accurately portray the "middle" or "average" of the data? Explain this - after all, we have been using the mean all along to describe the middle or average. What does this tell us?

89. (Continuation from #86) A distribution of data is **skewed** if it is not symmetric and the bulk of the values fall towards one side or the other. If plotted, the tail of a skewed distribution goes far in one direction. A **right-skewed** distribution has a long right tail, while a **left-skewed** distribution has a long left tail. What type of skew does the data from #86 have?

90. Here's another way to think of significance level: *The* α*-level controls the probability that you mistakenly reject a null hypothesis that is actually true.* Is there ever any danger in making this mistake? Explain. See if you can come up with a real-life situation where this might prove harmful or even deadly.

91. (Continuation from #88) Can you devise a way, other than the mean, to find a central number that best represents the data set {2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 8, 10, 10, 12} ? Explain how to find your number. Now, see if you can find a second such number. Are these numbers the same? Are they a better center than the mean?

92. You can actually make two types of mistakes when you perform a test of significance. The first occurs if the null hypothesis is actually true but the researcher rejects the null hypothesis. This is called a **Type I error**, or a "false alarm". In this situation, a researcher believes he/she has found an effect when there really is none to be found. The second type of mistake occurs if the null hypothesis is actually false but the researcher fails to reject the null hypothesis. This is called a **Type II error**, or a "missed opportunity". In this case, the researcher fails to detect a real effect or difference that is present. Imagine a court case: A jury is deciding the fate of a defendant who is accused of a heinous crime, and a human judgment must be made! In this context, interpret Type I and Type II errors. How could these errors be dangerous? Do you feel one error is more dangerous than the other?

93. Can we ever truly know if a Type I or Type II error occurs in a statistical situation? Explain.

94. (Continuation of #92) Now we pose an interesting dilemma: Back to the court case, our judicial system is set up to minimize the likelihood of making a Type I error in a criminal case (what outcome would this be?). In practice, this means "innocent until proven otherwise" or in other words, the evidence must prove guilt "beyond a reasonable doubt". Think about this - what effect does this practice have on making a Type II error? What does this generally imply about the relationship between Type I and Type II errors?

95. (Continuation from #91) The **median** is the middle number in a data set. If there are an odd number of data values, the median is the $[\frac{n+1}{2}]$th observation, where n is the sample size. If there is an even number, the median is found by adding up the $[\frac{n}{2}]$th and $[\frac{n}{2}+1]$th observations and dividing by 2. What is the median of the data set in #91?

96. We have a data set $\{3, 4, 5, 6, 7, 8{,}9\}$. Calculate the mean and median of this data set. What do you notice? Now, add 100 to the data set. Without calculating new values, which do you think will change more - the mean or the median? Why?

97. (Continuation) The number 100 in the previous question is said to be an **outlier**, since it is very far away from the rest of the data. How many such outliers would need to be added to substantially change the median? What do you think this means?

98. A statistic is said to be **resistant** or robust if its value does not change considerably when outliers are included in a data set. Do you think the mean or median are resistant? Why do you think that?

99. Another rarely-used measure of center (besides the mean and median) is the **mode**. This is the data value that occurs the most often. Would this ever be useful? In a skewed distribution, which measure of center do you feel is most appropriate? Why?

100. The probability of rejecting a false null hypothesis is called the **power** of a test. In other words, it is the ability of a test to find a real effect if there is one. It is $1 -$ probability of Type II error. Ganzfeld studies involve testing for psychic abilities among two participants. In one of these studies, each participant is placed in separate rooms and one is shown an image. The other participant is shown one of four possible images (one that matches the first participant's), and they must "psychically" choose the correct image. Statistician Jessica Utts in 2010 reported out of 2,124 sessions, there were a total of 709 correct answers. If the subjects had NO psychic ability, what would the probability be of guessing the target image correctly? What would the null and alternative hypotheses be if we wanted to test whether the data provide strong evidence of psychic ability? Calculate an appropriate test statistic, and decide whether to reject or fail to reject the null hypothesis. Explain what a Type I and Type II error would be here, and explain power in this situation. What does this study tell you overall about psychic abilities, if anything?

101. Two studies are done to gage the support for making Election Day a national holiday. In the first study, they found 60% plus or minus 5% were in favor. In the second study they found that 58% plus or minus 3% were in favor. Which study is more convincing? What do you think the point of the **margin of error** (the plus or minus numbers) happens to be?

102. If I were to say that I was "95% confident" about something, what would that mean?

103. The **Empirical Rule** says that approximately 68% of the data falls within ± 1 standard deviation in a Normally-distributed model; meanwhile, 95% of the data falls within ± 2 standard deviations; and almost all (about 99.7%) of the data falls within ± 3 standard deviations. Of particular interest is the idea that 2 standard deviations capture 95% of the spread around the mean. How much of the data lie beyond +2 standard deviations, approximately?

104. A **confidence interval** is a range, centered at the sample estimate of the parameter, that attempts to capture the true parameter's location a certain percentage of the time. Most often, a 95% confidence interval is chosen. If we are attempting to capture the population mean inside a confidence interval, obviously our sample mean will serve as the center for my confidence interval - this is my "best guess estimate" of where the population mean resides. What other statistic do you think we will need to calculate the confidence interval? Why?

105. A rough estimate method for calculating a 95% confidence interval of the population proportion is called the **2SD method**. Thus, the 95% 2SD confidence interval for Π is given by:

$$\hat{p} \pm 2 \times \sqrt{\hat{p}(1-\hat{p})/n}$$

Think about the formula. Interpret each part of it. What statistic do you think $\sqrt{\hat{p}(1-\hat{p})/n}$ is? Why?

106. Time for a thought experiment! There are 10 studies done all finding the average price of a current-year Prius. Do you think the car-to-car standard deviation from the mean will be higher or the study-to-study sampling standard deviation from the mean will be

higher? Why do you think this? Do you think any other factor is important?

107. In general, the formula for a confidence interval is as follows:

$$\text{sample statistic} \pm \text{multiplier} \cdot \text{ SD of statistic}$$

For approximately Normal data (such as sample proportions with large enough sample size), a 90% confidence interval has a multipler of approximately 1.645; for a 95% confidence interval, the multiplier is approximately 1.96; for a 99% confidence interval, the multiplier is approximately 2.576. With this in mind, explain where the 2SD method comes from.

108. (Continuation) Observe that the multiplier increases as the confidence level increases. What does this mean about the size of the confidence interval? Think about this: If we wanted to more confident in our result, what happens to the size of the confidence interval? What about vice-versa? What does this mean?

109. If we go back to the 2SD method and consider it for a quantiative variable now instead of categorical, we must use the mean and SD of the mean (ie, the standard error of the mean). (NOTE: Notice an important point here - the formula for confidence interval uses the SD of the statistic, NOT the SD of the sample!) The standard error for $\bar{x}$ is $s/\sqrt{n}$. Therefore, the "quick and dirty" 2SD method for a 95% confidence interval about the mean gives us

$$\bar{x} \pm 2 \times \frac{s}{\sqrt{n}}$$

What is a limitation of this method?

110. In a general confidence interval for the population mean, what distribution will the

multiplier come from? Recall that for the population proportion, the Normal distribution was used. Will it work here? Why do you think this?

111. Another thought experiment: Would any sample size work for a confidence interval for the mean? Consider a sample size of 2. Do you think you will have enough data to validly find a confidence range? Why? What about a sample size of 5? 8? 10? When is it big enough to be okay? What do you think? Is there anything else to consider here?

112. Yet another thought experiment: How could you cut the margin of error in half? Discuss.

113. (Continuation) Is there a statistic, besides the confidence level, that can affect the confidence interval? It might help to look at the formula.

114. In general, a confidence interval about the mean is difficult to calculate by hand - finding out the multiplier actually involves some really nasty math. In practice, tables of values can be used, or software can be used as well. If you wanted to use a table to calculate a multiplier value (often called a "critical value"), can you think of any disadvantages? (I know, you've probably never seen such a table - the point is to think about inherent limitations of a table over software.)

115. In November 2013, a Gallup poll asked 1,039 U.S. adults how much they planned to personally spend on Christmas gifts. The average was $704. Suppose the standard

deviation was \$150. Use the 2SD method to approximate and interpret at 95% confidence interval for μ.

116. (Continuation) If the SD were \$300 instead, what would change? Don't calculate, but compare how the confidence interval should change.

117. (Continuation) This poll had 562 males and 477 females. If we looked at each subgroup separately and found the confidence intervals for each (assuming the same SD as the overall sample), how would you predict the confidence intervals would change? Why? Actually calculate them to see if you were correct. What are your thoughts?

118. (Continuation) In general, hypothesize about how confidence intervals change as the sample size increases. Why does this happen?

119. Identify THREE different ways to decrease the confidence interval for a population mean.

120. **Nonrandom errors**, also called systematic errors, are errors that are not due to random error from the sampling method. Instead, nonrandom errors can be caused by bias, discrimination, untrustworthy survey takers, and other characteristics such as sex, race, disposition, etc. can all be causes of nonrandom errors. Is there any easy way to eliminate nonrandom errors? Is this truly a statistical question at all?

121. We will return to a topic we have touched a couple of times already - explanatory versus response variables and possible association between them. Consider smoking cigarettes and lung cancer - is there an **association** between these variables? In other words, does information about one variable's status (smoking - yes/no) give you any predictive power over the other variable (probability of developing lung cancer)? What would be the explanatory variable? Response variable? Does this *necessarily* mean that one causes the other? Explain.

122. Why do you believe there is a natural temptation to jump to causal explanations? As a statistical-minded person, what is a better approach?

123. How might one untangle the relationship between association and causation? Any ideas?

124. A study was conducted and found that ownership of a Lamborghini was associated with ownership of a yacht. The researchers were then curious to if yacht ownership caused Lamborghini ownership - do you think there is an association here? Could it be reversed? Is causation a good explanation here? More to the point, is there a better explanation?

125. Imagine you are planning a study about pet ownership, and you want to see whether owning a pet has an association on health (good, okay, poor). What would be the explanatory variable? Response variable? Is there just one way to answer this question?

126. A **confounding variable** is a variable that is related to both the explanatory and response variables in such a way that it cannot be separated from the effects of the explanatory variable. A study conducted to see if there was a relationship between if people participated in recreational hunting or not and the amount of farmland in an area. They found an association between the proportion of hunters in an area and the amount of farmland. They then concluded that farmers like to hunt often. Is this a good conclusion? Why or why not? Can you find any confounding variable to explain this?

127. An **observational study** is a study done where the researcher merely watches and records data, but does not intervene. Recall what an experiment is. How does an observational study differ from an experiment? Please note, this is a major distinction in science. Also, how does your answer to the previous question relate to confounding variables and effects?

128. In an experiment, you assign one or more **treatments** to a group in the study. You may also assign a **control** group to monitor the effects the treatment may have on the treatment group. What sort of treatment do you think the control group gets? Why?

129. Suppose you have 10 males and 6 females in a small experiment. If you feel that gender may affect the result, how could you assign the groups to balance the study with respect to the possible confounding variable as much as possible?

130. If you are conducting a large experiment with several hundred participants in each group, you can't possibly try to balance each group for every possible confounding variable - so what to do, what to do! That's my question for you: Is there a strategy you feel you can employ that will neutralize the effects of any confounding variables?

131. A **quasi-experiment** is an experiment where the explanatory variable is manipulated non-randomly. For example, a study about the grade increase based on two different versions of Calculus taught at a school would be difficult to randomly assign students to a class - many students would probably be upset. Can you think of a way that researchers in this study could still glean useful results? Realize that the students signing up for Class A and the students signing up for Class B may not be directly comparable.

132. In a **double-blind** study, neither the subjects nor the researchers know what treatment group each subject is in. When could this be useful? What effect might this have on bias?

133. (Continuation from #130) Randomization is the best way to deal with any confounding effects. Therefore, a **randomized experiment** uses randomization (through software, for instance) to make random assignments into each group. What effect might this have on our ability to draw cause-and-effect conclusions?

134. It appears that experiments are much more useful than observational studies! When might observational studies actually be useful? Are there any cases where it is preferred over experiments? Explain.

135. We can find confidence intervals using more rigorous math also. You'll still need a TI-84 calculator or other statistical software. To do this for the confidence interval of the mean, we need to know that the **degrees of freedom** is $n-1$. For now, just know that the degrees of freedom is a statistic that helps control the shape of the distribution. We say that a **random variable** X is distributed as a t-statistic with 19 degrees of freedom as follows:

$$X \sim t_{19}$$

Besides the degrees of freedom, what other mathematical formulas will you need to figure out the confidence interval? (It's not clear yet what role the degrees of freedom plays, but it will soon become apparent.)

136. A certain study asked participants one of two questions: 1) Are you having a good day? 2) Are you having a bad day? The respondents' results were recorded below, along with what question they were asked:

	Good Day?	Bad Day?
Yes	67	43
No	34	55

What are the variables in this study? How many individuals were surveyed?

137. (Continuation) Are the group sizes equal - in other words, were the same number of people asked each question? In this case, it may be useful to look at the relative frequencies instead of the frequencies themselves. If you do this, what does the data look like now? Do you feel the way a question was asked matters? Or is there too little evidence to know?

138. (Continuation) Suppose I wish to know if the respondents' responses were significantly different from one another based on the question they received - in other words, might "Are you having a good day?" versus "Are you having a bad day?" affect the perception and therefore response of participants. Let Π_1 be the population proportion for "good day" responses and Π_2 be the population proportion for the "bad day" responses. How might you go about setting up a null hypothesis and alternative hypothesis here?

139. (Continuation from #135) To compute a confidence interval, you also need to know the standard deviation of the sample. This can be found on most calculators. If you are curious about the actual calculation, here is the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

The Σ symbol means to add up all of the results from the $(x_i - \bar{x})^2$ part. From examining the formula, can you try and explain what the standard deviation seeks to measure? We have discussed this before in a less-rigorous way - let's see if it makes some sense now.

140. (Continuation from #136,#138) We can compute the standardized statistic (which is a z-statistic) as follows:

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\hat{p_1}$ is the sample proportion of group 1, $\hat{p_2}$ is the sample proportion of group 2, n_1 is the sample size of group 1, n_2 is the sample size of group 2, and $\hat{p}$ is called a pooled proportion - it is the combined proportion for both groups put together. Calculate this z-statistic for the data in #136.

141. (Continuation) The margin of error for a confidence interval for two proportions involves a multiplier (it's the same multiplier we discussed earlier, the 1.645, 1.96, or 2.576 numbers depending on the confidence level) and the standard error of the population proportion. It is given by this formula:

$$\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

where $\hat{p_1}, \hat{p_2}, n_1,$ and n_2 are the same as the previous question. With this information, calculate the margin of error for the data in #136.

142. (Continuation) In #140, we discussed a **pooled** estimate. This is only valid if we assume the two groups are homogeneous - meaning there is little difference in variation

between them. How might you modify the denominator in #140 to accommodate a difference in the variation between groups?

143. (Continuation from #139) We continue to seek finding a confidence interval mathematically. We have discussed degrees of freedom and standard deviation of the sample. We also need to know the multiplier of the margin of error. The math behind this part is very complicated, but a calculator can solve it for us, such as a TI-84. If we are wanting a 95% confidence interval centered at the sample mean, what percentage of the distribution will lie beyond that 95% on the right side? (It's NOT 5%!)

144. In #135, we introduced an idea called a **random variable**. This necessitates more explanation. First, a **sample space** is the collection of all possible outcomes of an event. For example, rolling a die has the same space 1, 2, 3, 4, 5, and 6. A **random variable** is a mathematical function that takes an event in the sample space as the domain and sends it to a real number between 0 and 1. This number is the probability of that event occurring. We can write $P(X = 1) = 1/6$ to indicate that the probability of rolling a 1 on a standard six-sided die is 1/6. Our random variable is X. Fill out the entire probability model for rolling a six-sided die - in addition to the one answer I've given you already, you should have 5 more.

145. (Continuation) The **expected value** of a random variable is $\Sigma x P(x)$, or the sum of each value of the random variable multiplied by its probability and then summed with every other product. In the case of rolling a six-sided die, what is the expected value of a die roll?

146. (Continuation from #143) Statisticians, by convention, often measure a test statistic using a "less than or equal to" approach, meaning the percentage of the distribution that lies to the left of the top edge of our confidence interval is what the computer or calculator

is looking for. In #143, you should've found that 2.5% of the distribution lies to the right of our confidence interval. This means 0.975 of the distribution lies less than or equal to the top edge of our confidence interval. This number is what the TI-84 is looking for when telling you what the multiplier has to be. You also need the degrees of freedom. Now, suppose our data set is {8.6,9.4,7.9,6.8,8.3,7.3,9.2,8.7,11.4,10.3,5.4,5.5,6.9} We want to find a confidence interval for the mean. What distribution will we use? What is n? What is your degrees of freedom? On your TI-84, compute the multiplier for the confidence interval by using the invT function with arguments of (0.975,14). What multiplier did you get?

147. Suppose there is a lottery where you can select any 3-digit number, including leading zeros (so 005 is acceptable). If the ticket costs $1 to play, and you win $500 for your number being picked, what is the expected value? You can also read this as "What is the average expected gain or loss on each ticket?"

148. A probability distribution is given below. Find the expected value.

x	P(x)
0	P(x=0)=4/50
1	P(x=1)=8/50
2	P(x=2)= 16/50
3	P(x=3) = 14/50
4	P(x=4) = 6/50
5	(x=5) = 2/50

149. Think of expected value as guaranteed to happen over a long number of trials or events. This is called the **Law of Large Numbers**. How might this explain how lotteries and

casinos are able to be profitable? If the cost of a lottery ticket were below the expected value of the ticket, what can you conclude?

150. (Continuation from #146) To compute the confidence interval, we also need the standard error of the mean. Compute this for the data set in #146 using the formula:

$$\frac{s}{\sqrt{n}}$$

151. (Continuation) Now we can finally create our confidence interval! First, calculate the mean of the data in #146 to serve as our parameter estimate. Then, complete the confidence interval using the following formula discussed previously:

$$\text{sample statistic} \pm \text{multiplier} \cdot \text{ SD of statistic}$$

What is your 95% confidence interval for the mean?

152. (Continuation) Summarize all of the steps necessary to calculate a confidence interval step-by-step. Instead of Tinv, which uses the t-distribution, when might you use the Ninv operation to get the multiplier for a Normal distribution?

153. Calculate a 95% confidence interval of the mean for the following data {6, 8, 9, 11, 14, 15, 16, 17, 19}

154. Calculate a 95% confidence interval of the population proportion for the following data

$\{ 0.56, 0.55, 0.57, 0.58, 0.56, 0.56, 0.55, 0.59\}$ Assume the population is 20 times or more bigger than the sample size.

155. If a 90% confidence interval were desired for the mean, how might one adjust the procedure of calculating the multiplier for the confidence interval?

156. When should you use the median rather than the mean?

157. Many times we want to compare two means. Can you devise a way to do this? You don't have to come up with a test or anything, but is there a way we could combine the two data sets into one data set, in a way?

158. A study was conducted in 1999, published in *Pediatrics*, that studied whether children who were breastfed during infancy differed cognitively at age four from their peers who weren't breastfed. The study involved 323 white children from similar backgrounds and socioeconomic status. What do you think a null hypothesis and alternative hypothesis here would look like? Hint: We will be looking at the two means.

159. The **range** is the distance between the highest and lowest number in a data set. Is the range useful? Why or why not?

160. If we are looking for a statistically significant *difference* between two means, what does that mean? Think about what "difference" means mathematically.

161. (Continuation from #158) Is the study in #158 a controlled experiment or an observational study? Can cause-and-effect be determined, or just association? Could you redesign the study to change this? Would this even be ethical? Explain.

162. In comparing two means, μ_1 and μ_2, we will be looking to see if $\mu_1 - \mu_2 \neq 0$. In other words, we want to see if the difference in means is significantly different from zero. Why? What does this tell us?

163. The standardized t-statistic for a difference in means is:

$$t = \frac{\bar{x_1} - \bar{x_2} - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Interpret this statistic.

164. The **upper quartile** is the median of the upper half of the data, while the **lower quartile** is the median of the lower half of the data. In other words, the upper quartile is the 75th **percentile**, meaning 75% of the data distribution lies to the left of the upper quartile. Likewise, the lower quartile is the 25th percentile. Subtracting the upper and lower quartile gives the **inter-quartile range**, or IQR. Think about what this statistic is - can you figure out what information it tells us? Is there another statistic that it could be compared to? Is it any different?

165. The degrees of freedom for two population means is somewhat complicated. Here is the formula:

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{(\frac{1}{n_1-1})(\frac{s_2^2}{n_1})^2 + (\frac{1}{n_2-1})(\frac{s_2^2}{n_2})^2}$$

Luckily, computers and calculators can calculate this easily! Notice the **variances** (the square of the standard deviations) are not pooled, as we did previously. Any idea why this may be?

166. A professor is testing whether a face-to-face class has better final exam scores than an online class. The tables below show the class scores:

Online Class:

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Face-to-Face Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Just by looking at it and getting a feel for the data, do you think the face-to-face class has better final exam scores?

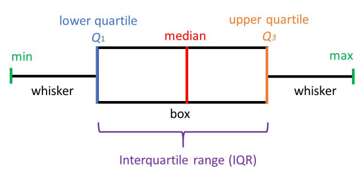
167. (Continuation) Follow-Up from the previous question: What is the null and alternative hypotheses? Will the test be one-tailed or two-tailed?

168. Like the median, the IQR is resistant to outliers. The standard deviation is not resistant; in fact, it s extremely sensitive to extreme values. Do you have any ideas when the IQR

should be used?

169. (Continuation from #166) On a TI84, conduct a **two-sample t-test**. To do this, input the data into List1 and List2, go to TESTS, press 2SampTTest, and hit Enter. L1 is the first list, and L2 is the second list. Make sure to change the μ_1 and μ_2 appropriately. Make sure it is NOT Pooled (never do this!). What result do you get? What do you conclude?

170. A **boxplot**, or box-and-whisker plot, is a visual display of the range, IQR, and median. The box displays the middle 50% of the data (the IQR) with a line inside the box indicating the median while the whiskers extend outward to the maximum and minimum values to display the range. An example is shown below:



Given the data set {70,70,73,78,81,83,87,90,95,100,100}, create a boxplot of the data. Given the data set {45,66,75,78,79,80,85,91,96}, create a boxplot of the data. What is similar about the distributions? What is different? What is an advantage of a boxplot?

171. Instead of comparing two means, imagine you have two sets of data that are **matched data**, or paired data. This means that both data sets are related to one another; usually, this means that one data set is a "before" measurement on a group of individuals, and the other data set is the "after" measurement on that same group of individuals. To perform a test of significance with such data, you will find the difference of the "After" data minus the "Before" data, and then perform a one-sample t-test on the resulting differenced data set. With that in mind, consider a study done to investigate the effectiveness of hypnotism in reducing pain. Subjects recorded their pain level before and after hypnotism. The results are shown below. Are the sensory measurements, on average, lower after hypnotism? Use a 5% significance level. On your TI84, you can input the differenced list int L1, go to TESTS, then T-Test, enter 0 for μ_0, and 1 for frequency. You should also choose "$< \mu_0$" since this is a one-sided test that is looking for a less-than answer. What is your result?

Subject:	A	B	C	D	E	F	G	H
Before:	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After:	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

172. A nutritionist wants to compare two groups of customers who have been trying two different diets for the past 3 months and wants to see if there is any meaningful difference in their weight loss between the two diets. The results are shown below. Conduct an appropriate test of significance at the 5% level and give your result.

Diet	Person A	Person B	Person C	Person D	Person E	Person F	Person G	Person H
Diet 1	3	5	17	11	19	15	0	3
Diet 2	4	6	14	16	11	0	7	9

173. A football coach wants to test whether his strength conditioning class increased his players' maximum lift on the bench press. The results of the four players in the class are shown below. Perform an appropriate test of significance and give your result.

Weight	Player 1	Player 2	Player 3	Player 4
Before	205	241	338	368
After	295	252	330	360

174. A tutor wants to know if his tutoring has significantly improved his students' grades. Their math class average before his tutoring is shown below, along with their grade one semester after tutoring. Perform an appropriate test of signficance and give your result.

Tutoring	Student 1	Student 2	Student 3	Student 4
Before	77	82	62	70
After	80	81	75	74

175. Think for a moment about all of the tests of significance, confidence intervals, and statistics we have discussed so far. Now is a good moment to stop and take a deep breath! Try and summarize, from memory, as many of these as possible.

176. With our discussion of proportions, we have been secretly using a distribution called the **binomial distribution**. In this distribution, there is a fixed number of trials, and only two independent outcomes, "success" or "failure". When doing proportions, we were using p as the probability of success; we will continue that notation here. Let the random variable $X =$ the number of successes obtained in the n independent trials. Then, the mean $\mu = np$ and the standard deviation is $\sqrt{np(1-p)}$. Suppose you are flipping a coin. What is p? If you flipped the coin 10 times, what is the mean? The standard deviation?

177. Could you do a two-sample test instead of the paired test for paired data? Why or why not? What is the problem?

178. What is the relationship between association and paired data?

179. On the TI84, you can use binompdf to find $P(X = \text{value})$ and binomcdf to find $P(X \leq \text{value})$. Suppose there is a 30% chance of a student not completing their Bachelor's degree (moment of silence please!). In a group of 20 randomly selected students, what is the probability that exactly half of them will not complete their Bachelor's degree? What is the probability that half or less of them will not complete their Bachelor's degree?

180. If 25% of Sterling students regularly volunteer in the community, if we select 15 students at random, what is the probability that you could find a group of more than 5 of them willing to volunteer? (Hint: Probabilities have to add to 1! We want $P(X > 5)$, but the TI84 will give us $P(X \leq 5)$. How can we fix this?)

181. Does paired design always have to involve the same subjects? Can you come up with a situation where it wouldn't?

182. Similar to a binomial distribution is the **geometric distribution**. In this distribution, trials (of "successes" and "failures") are repeated until the first success, then stopped. The random variable X = the number of trials until a success, the mean $\mu = \frac{1}{p}$, and the standard deviation is $\sqrt{\frac{(1-p)}{p^2}}$. Let's a bakery has a 0.02 probability of making a defective cookie and having to expel it from the conveyor. What is the probability that a defective cookie is found on or before the 20th cookie? On the TI84, you can use geometpdf for $P(X = \text{value})$ questions and geometcdf for $P(X \leq \text{value})$. The parameters will be your probability and then number of the trial you are on.

183. Does your bowl size affect how much you eat? Food psychologist Brian Wanisnk (at mindlesseating.org) ran such an experiment. In this experiment, two sessions were held. During the first session, participants were assigned to receive either a small bowl or a large bowl and allowed to take as many M&Ms as they wanted. At a following session, the

bowl sizes were switched. What would be your null hypothesis? Alternative hypothesis? Was this an observational study or an experiment? How does that affect the conclusions?

184. The literacy rate for a nation measures the proportion of age 15 and over who can read and write. Let's say the literacy rate for students in Mexico is 22% (I'm sure it's actually much higher than this! This is just for illustrative purposes.). Let X = the number of Mexican students you must ask until you find one that says he/she is literate. What is the probability you find a literate person on your 5th student asked? 6th? 7th?

185. In a paired/matched design, what is the population mean of the difference under the null hypothesis?

186. (Continuation of #183) The data that shows how many M&Ms participants took from each bowl is shown below.

Small Bowl	33	24	35	24	40	33	88	36	65	38	28	50	26	34	51	25	26
Large Bowl	41	92	61	19	21	35	42	50	11	104	97	36	43	62	33	62	32

If you wanted to see if there was a statistically significant difference between the bowl sizes, what would be your null hypothesis and alternative hypothesis? Use a TI84 to find a result. Give your feedback.

187. True/False: The shape of the null distribution of the sample mean difference is approximately normal for pairs of 20 or more.

188. True/False: The null distribution of the sample mean difference is centered at the

hypothesized value of the population mean difference.

189. Does the variability in the sample differences affected the variability of the null distribution of the sample mean difference?

190. (Continuation of #186) Use a TI84 to find a 95% confidence interval for the data in #186. Does this interval contain 0? What does that imply?

191. (Continuation) Consider the previous question and the relationship between a confidence interval and a significance test. Can you find any link between the two? Explain.

192. (Continuation) What does the result of the confidence interval and significance test mean in practical terms for us - does bowl size really not matter? Or could there be other factors here? What, if so? Could the experiment be repeated differently somehow?

193. (Continuation) Suppose you did repeat the experiment again with different tweaks this time in the hopes of obtaining a significant result. Comment on this. Is this ethical? What are your thoughts?

194. Consider the t-distribution that we have used extensively over the past 100 or so questions. In many of these cases, the sample size can be really small, or the distribution can be skewed, and it still works relatively well - the t distribution is relatively robust and resistant to outliers. However, as the sample size increases, it becomes increasingly close to the Normal distribution. In fact, many statisticians say that beyond and n=20 or n=30, you can just use the Normal distribution instead in all cases. What are your thoughts on this? Is approximation really the same as giving exact results? In practical matters, does it matter? When might it matter?

195. Another probability distribution that is sometimes interesting is the **Poisson distribution**. It gives the probability of a number of events occurring in a fixed interval of time or space if these events happen with a known average rate and independently of the time since the last event. Let X = the number of occurrences in the interval of interest. Let's say your Voicemail receives 6 calls every day between 8 AM and 10 AM. What is the probability that you receive one call in the next 15 minutes? Assume it is 8 AM. To do this, find out the mean, which is the frequency per unit of time, ie, the number off calls you receive in a 15-minute time period. Take the "average" on a TI84 calculator and calculate using the poissonpdf function. How many calls can you expect?

196. (Continuation) What if we want to find the probability that you get more than one call in 15 minutes? To do this on the TI84, you'll use the poissoncdf function, and you'll have to say 1 - that result, since we are reversing the probability. (Huh? Try to explain that. Why do we have to say 1 - result? This is called the **complement probability rule**. It works because probability must add to 1. Explain this in a way that makes sense.)

197. Earlier, it was said that the proportions we have worked with "secretly" used the Binomial distribution - this is true, but only partly. When the binomial distribution has many trials, it can closely approximate the Normal distribution. The Normal distribution, as previously shown, is a smooth, continuous mound-shaped distribution. The Binomial distribution is a **discrete distribution**, as has been the geometric and Poisson distributions previously discussed. This means it takes on only a finite number of

values instead of values on a continuum. Considering these facts, can you sketch an idea of what you think a binomial distribution with 5 trials may look like? 10 trials? 20 trials? 40 trials? Remember, as the trials increase, it should look more like a smooth, mound-shaped Normal distribution.

198. Why do we use z-statistics for proportions but t-statistics for means?

199. A customer service center receives about ten emails every half-hour. What is the probability that the customer service center receives more than 15 emails in the next half-hour? Use a TI84 calculator to help you.

200. According to a recent Pew Internet Project poll, girls between the ages of 14 and 17 send an average of 187 SMS messages or other social media messages each day. Let X = the number of such messages a girl aged 14 to 17 sends per day. What is the probability that such a girl sends exactly 175 messages per day? What is the probability she sends less than or equal to 150 messages per day?

201. A study at Virginia Tech studied whether vehicles come to a complete stop at an intersection with four-way stop signs. One of the variables look at was the position of the car - that is, whether it was alone, leading a group of other cars, or following in a group of cars. The question they were seeking an answer to is if there is any type of association between the arrival position of the vehicle and whether or not it comes to a complete stop for at least an instant. What could your null and alternative hypotheses be here? What would ”no association” mean in this context?

202. The Normal distribution that we have already looked at is called a **continuous distribution** since its values fall on a continuum and is a smooth, connected curve. It turns out that the area underneath the curve (and above the x-axis) for the bell-shaped Normal distribution is 1. This can also be thought of as 100%. So, the Normal distribution's area underneath the curve ranges from 0 to 1 - hmm...is there another concept we have talked about a good bit that also must always be between 0 and 1? Do you think there's a possible connection here?

203. On your paper, draw a coordinate axis (x and y axes). Now, draw a heavy dark line starting at 0 and moving along the positive x-axis for as long as you like. Now here's the challenge - begin drawing upward at the point you stopped at to begin forming two sides of a rectangle, but there's a catch - your rectangle must have an area of exactly 1. Can you do it? If you figure out a way to do it, repeat the process several more times - can you figure out a general approach?

204. (Continuation of #201) Can you devise a different way to state the null and alternative hypotheses? Specifically, what is true about the probability that each group of arrival positions will stop? Can you write this out using symbols?

205. (Continuation of #203) Draw one final set of coordinate axes out, but this time, make a large space between 0 and 1. Repeat the process from #203, but only extend the bottom side of your rectangle out to 1 - how high must your second side go to ensure a rectangle of area 1? How do you know?

206. (Continuation) The **uniform distribution** is a probability distribution where all events are equally likely to occur. All of the rectangles you drew were examples of the uniform distribution, and the **standard uniform distribution** has a length and height of 1. Think about the area of the rectangle, and think about how much of it is shaded to the

left of some vertical line that crosses through your rectangle. If your rectangle starts at the origin (0,0) and goes to some point (b,0), can you figure out a way to find the x-value that will produce a vertical line cutting the rectangle in half? More generally, what if the uniform distribution started at (a,0) and went to (b,0), what would the formula be then? Is x-value of any interest to us? What does it tell us about the distribution?

207. Suppose three candidates are in a closely-followed election. Candidate A has 35% of the vote. Candidate B has 32% of the vote. Candidate C has 33% of the vote. Do you think, statistically, there is a significant difference between these candidates? How far apart would one of the candidates need to be for you to be relatively certain of a difference?

208. So far, we have discussed the Normal distribution, the uniform distribution, the binomial distribution, the geometric distribution, and the Poisson distribution. Compare and contrast these and when to use each one.

209. Let's return to our talk on ethics some, as it's been a while. An investigation of a famous social psychologist, Diederik Stapel, has led to the retraction of his articles from many of the world's top journals for his field. He was a former professor at Tilburg University in the Netherlands. Recently, an investigation concluded that he is guilty of fraud on an epic scale. The claims include falsifying data in over 55 papers and 10 doctoral dissertations he supervised. We'll delve deeper into this story in subsequent questions. For now, I'm curious as to what ethical breeches you feel he made - and don't think about how he just hurt his own career – how could this have greatly harmed others?

210. (Continuation of #207) If we were to create some type of test of significance to test the differences between candidates, what might that look like? What would that procedure

look like? Note, I'm not asking for a fully hashed-out idea here; I just want a general conceptual plan.

211. (Continuation of #209) In an interview he did with the *New York Times*, he said that he did not deny his actions, and that they were driven by deceit, but he claims it is much more complicated. His field, social psychology, is one he greatly loved, but Stapel was frustrated by the "messiness of experimental data, which rarely led to clear conclusions". Hmm...what do you think about his assertion? Is there some truth in it? Discuss.

212. (Continuation) Stapel's interview continued where he said that he sought to find "sexy results that journals found attractive", and that "It was a quest for aesthetics, for beauty – instead of the truth". Is this danger a very real threat to all practicing statisticians, scientists, and researchers? What do you think about his words? (P.S. Don't look at him as pathetic or evil; his desires are very common and temptation befalls us all. Not excusing his behavior, but we'd be amiss if we didn't acknowledge some universal truths here.)

213. (Continuation of #210) Let's say one of your plans to test this statistical significance between 3 groups is to take differences between each pair and do a test on each pair to see if the difference is large enough to reject the null. Think about why this might be a poor idea - you have 3 different tests to do, and if the null hypothesis is true 5% of the time in each case, we are increasing our overall chance to reject the null in error. Can you explain this oddity with an example? For example, if skydiving has a 1% death rate, would you rather skydive once or 500 times? How does this relate to our situation? Why is it an important consideration?

214. (Continuation of #212) Clearly, Stapel was very wrong in what he did, but it is a very real ethical concern faced by scientific academia. Journals will rarely publish results that

are NOT statistically significant, and publishing in journals is often a requirement to work as a professor or researcher. Discuss your thoughts on this. Why is this an unfortunate practice? Should journals be more open and accommodating to non-significant results that aren't "sexy"?

215. The **chi-square statistic** is a standardized statistic for summarizing the relationship between two categorical variables which has a predictable null distribution, making it a popular test. The **chi-square distribution** is a non-negative, right-skewed distribution used to test for an association between two categorical variables. The general form of this statistic is:

$$\Sigma_{i=1}^{n} \frac{((\text{observed - expected})^2}{\text{expected}}$$

, where "observed" refers to the observed cell counts in a two-way table, and the "expected" refers to the expected cell counts, computed assuming the null hypothesis to be true. To find the expected cell counts, multiply the row total times the column total and divide by the table total. Researchers wanted to find out if views on marriage differed significantly between different generations. They asked a total of 2,622 Millennials, Gen X'ers, and Boomers. The result is shown in the table below. Can you figure out the chi-square test statistic for this table? What is your conclusion?

Group	Millennials	Gen X'ers	Boomers	Other	Total
Marriage obsolete? Yes	236	313	401	68	1018
Marriage obsolete? No	300	416	745	143	1604
Total	536	729	1146	211	2622

216. (Continuation of #214) Stapel's investigators found that he was guilty of several malicious practices, including:

- creating datasets that confirmed expectations
- altering data
- changing the measurement instrument without reporting it
- misrepresenting the number of experimental subjects

Comment briefly on each of these and how each pose an ethical threat to the practice of sound statistical research.

217. In February 2013, Quinnipiac University conducted a poll asking about a controversial policing tactic in New York City: *As you may know, there is a police practice known as 'stop and frisk', where police stop and question a person they suspect of wrongdoing and, if necessary, search that person. Do you approve or disapprove of this police practice?"* The results are shown in the table below, divided by age group. Conduct a Chi-Square test to see if there is a significance difference among age groups.

Age	18-34	35-44	55+	Total
Approve	68	128	173	369
Disapprove	164	204	165	533
Total	232	332	338	902

218. A large study finds that the following table shows the number of televisions owned by American families. A random sample of 600 families in the western U.S. resulted in the data in the second table. At the 1% significance level, does it appear that the western U.S. families are largely different than the American population as a whole when it comes to television ownership?

Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
Total	600

What is the degrees of freedom? On a TI84, go to STAT TESTS, choose Chi2 GOF, and run the test. What is your conclusion?

219. It is often said that statistics can be manipulated to say whatever one wants them to say. This ties into our ethical discussion about proper behavior and usage of statistics. It is easy to suffer from confirmation bias that allows you to prove only what you already thought was true. Ethics is a central cornerstone of a sound statistical foundation, though. Should the "truth" play center stage? What does the "truth" even mean? If multiple interpretations of our results could create opposing views on the implications, how should we approach that? What role do you feel ethics plays in statistics?

220. Suppose we want to compare more than two groups on a quantitative response. Why don't we simply conduct all the possible two-sample comparisons between the groups? What is flawed with that idea? (We've touched on this earlier.)

221. A null hypothesis for a particular research experiment is $H_0 : \mu_1 = \mu_2 = \mu_3$. What does this mean? What would the alternative hypothesis most likely be? Explain.

222. The **F-distribution** is the next distribution we will study. This distribution is a continuous, non-negative, and right-skewed statistic. Explain what all of these mean, and sketch an example of what the curve for this distribution's probability might look like.

223. (Continuation) The **F-statistic** is a ratio of variation between the groups to the variation within the groups. The statistical test using the F-statistic is usually called the **ANOVA** test, or Analysis of Variance. Based solely on the name alone and the explanation of the statistic, what do you think this test does?

224. One of the assumptions of the ANOVA test is homogeneity of the population variances (the square of the standard deviations). This means that, roughly, the variances of each group must be about the same. What type of graphical measure could be a good way to easily compare the spread of the different groups? Support your decision.

225. An example of an ANOVA test is shown below. For now, we will focus on the "MS" column. The **mean square for treatment**, which in this case is 40.02, is a measure of variation between the three groups. Meanwhile, the **mean square for error**, which is 3.16 in our example, is a measure of the within-group variation. Study the table for a moment, and look back at the definition of the F-statistic: How do you think the two mean squares relate to the F-statistic?

Source	df	SS	MS	F	p-value
Treatment	2	80.04	40.02	12.67	0.0000
Error	54	170.53	3.16		
Total	56	250.56			

226. (Continuation) The degrees of freedom (df) in the treatment and error are divided by the SS to get the mean squares. The SS are the **sum of squares for treatment** and the **sum of squares for error**. These statistics calculate the difference from each datum point to the mean (either the within-group mean or between-group mean), squares the difference, and then adds all of the results together. To get the mean squares, we divide the sum of squares by the degrees of freedom. How is this similar to the formula for standard deviation? How is it different?

227. (Continuation) Notice the p-value says 0.0000...can it really be 0? What's going on here? Discuss.

228. (Continuation) The ANOVA test is designed to test if there is ANY difference at all between the different groups.Therefore, if the result is rejecting the null hypothesis, and it finds that there is a significant difference in means, what does that tell you? Are you done? Or no? If not, what can you do now?

229. If the ANOVA test is used for comparing three or more means, why do we use the t-test for two means? Why not just use the ANOVA test all the time?

230. Can you devise a situation where we might want to compare 3 or more groups? Would ANOVA be appropriate here? Why or why not? Can you come up with a situation where ANOVA wouldn't be appropriate? What would be?

231. Let's compare and contrast ANOVA with the Chi-Square Goodness of Fit test we discussed earlier. The Chi-Square GOF test works when all groups are categorical in nature. The ANOVA test works when the response variable is a numerical variable, but the explanatory variable is categorical. Can you think of a situation where the ANOVA might apply?

232. Interpret the F-statistic: $\frac{MS_{\text{between}}}{MS_{\text{within}}}$ (If you haven't figured this out yet, the mean squares are a type of variance, or squared standard deviation).

233. Let's calculate an ANOVA by hand, at least once, just so we can see how it's done. Three different diet plans are being tested for mean weight loss. The entries in the table

for weight losses are shown below.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

First, calculate the standard deviations for each group. Then, use the following formula to find the Sum of Squares for between groups:

$$\Sigma[\frac{s_j^2}{n_j}] - \frac{(\Sigma s_j)^2}{n}$$

, where s_j is each group's standard deviation. Next, we'll need the Sum of Squares total, which can be found using the formula:

$$\Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

Next, to get the Sum of Squares for within groups, use the formula:

$$SS(\text{within}) = SS(\text{total}) - SS(\text{between})$$

The degrees of freedom for between groups is the number in each group - 1, so record that. The degrees of freedom for within groups is the total sample size of all groups put together minus the number of groups, so record that also. The Mean squares for between groups (Treatment) is found by dividing the SS between groups by its degrees of freedom, and likewise for the Mean squares for within groups (Error). Finally, to get the F-statistic, divide the Mean squares for treatment by the Mean squares for error. Report your result.

234. (Continuation) It's no surprise that, following the above calculations, that most ANOVAs are done by computer these days. On your TI84, list the values from Plan 1, Plan 2, and Plan 3 as lists in the calculator, then under STAT TESTS, select ANOVA and input your lists. Then hit Enter. Much easier, eh? Your result should match your result from the previous question.

235. A scientist is performing an experiment to see how different types of soil cover would affect slicing tomato production. Tomato plants were grown under three different conditions, and groups of plants had one of the following treatments listed in the table. Calculate the ANOVA test to see if the conditions differ significantly for at least one of

the growing conditions. What is your result? (Hint: Use the TI84 calculator!)

Bare	Ground Cover	Plastic	Straw	Compost
2625	5348	6583	7285	6277
2997	5682	8560	6897	7818
4915	5482	3830	9230	8677

236. Suppose a study conducted in California asked participants how much money they would donate to the college. Participants were either Freshmen, Sophomores, Juniors, or Seniors. If it was desired to see whether there was any difference in the mean amount donated by each group, would ANOVA be appropriate here? What if instead of asking how much money for donation, it simply asked ”yes” or a ”no” – what changes are needed now?

237. A drug company wants to conduct an experiment to test the effect of a new cholesterol medication. The company takes 15 patients and divides them into 3 groups of 5 each receiving either 0 mg, 50 mg, or 100mg of the medication. They want to see whether there is a significant difference in mean cholesterol levels between the groups after 30 days. The results are shown in the table below. Conduct an ANOVA test to decide if there is a significant difference.

0 mg	50 mg	100 mg
210	210	180
240	240	210
270	240	210
270	270	210
300	270	240

238. ANOVA tests simply tell you that there is or is not a difference. Where that difference lies, it doesn’t say. There are many, many tests to assess this situation, but they are largely beyond the scope of this class. We could do the t-test, though; is there a way we might could judiciously go about it to minimize Type I error chances?

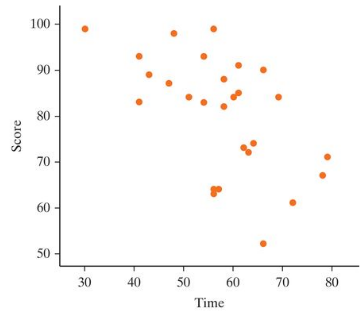
239. ANOVA also assumes that the individual group populations are Normally distributed. What type of graphical measure could be looked at ahead of time to ensure this assumption holds?

240. Summarize your findings on the ANOVA test. When is it most useful?

241. There is a strong relationship between the number of McDonald's restaurants in an area and the number of cases of lung cancer – does this mean Big Macs (or, heaven forbid, McNuggets) cause lung cancer?

242. As you might can see from the last question, we are returning to our topic of association. Let's recap for a bit. Do you feel there is an association between child age and height? Does this relationship last forever?

243. In the questions that follow, we will be addressing relationships between two quantitative variables. To help us visualize this, we are introducing a new graphical product called a **scatterplot**. In this graph, we put the explanatory variable along the x-axis and the response variable on the y-axis. Looking at the labels and graph, what do you suppose or guess this scatterplot is graphing?



244. We want a way to measure statistically the association revealed by a scatterplot. Let's notice that the association tells us the strength, direction, and form of the relationship between the data. What do you think each of these might mean?

245. When we try to measure association, we will also have to watch out for outliers. This could greatly change our results. If we do have outliers, should we throw them out? Why or why not? Discuss.

246. Much of the time, the association between two quantitative variables is linear in nature. If so, we have a particularly useful statistical tool under our belt - the **correlation coefficient**, or simply, the correlation. This is a statistic that measures the strength and direction of the *linear* association between two quantitative variables. It is denoted by the letter r. It ranges from -1 (very strong (perfect, actually) negative linear association) to +1 (very strong (perfect) positive linear association). What do you hypothesize a correlation of 0 means?

247. (Continuation from #245) Calling ALL unusual observations outliers isn't technically correct; some observations may follow the pattern of the scatterplot generally, but their

extreme nature (usually very far along the explanatory variable axis) makes them particularly influential - these **influential observations** can drastically change the correlation if removed from a data set. How to truly deal with influential points is beyond the scope of this class, but brainstorm for a moment; do you have any ideas? Is there a way we could shrink or lessen the impact of the point without removing it entirely?

248. So we know that correlation can range between -1 and 1. What would you say is a weak correlation? Moderate correlation? Strong correlation? I know you probably don't know, but I'm curious to your perceptions.

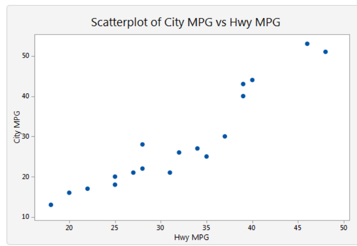
249. A positive correlation will produce a scatterplot where the data points generally move upward as the graph moves from left to right. Draw an example of non-perfect positive correlation in a scatterplot. Now, assign variable labels that make sense to the axes. Explain your reasoning.

250. (Continuation) Let's do the same with negative correlation. Can you find a real-life situation that has negative correlation?

251. We sometimes want to test whether the correlation is significantly different from 0. Why? We will revisit this idea in a bit.

252. Our attention turns to an interesting dilemma: We can measure the correlation, which shows us the strength and direction of the linear association, but what if we want to

actually *show* that linear association? Meaning, what if I wanted to draw a line on my scatterplot that got as close to as many points as possible? An example scatterplot it shown below. Strategize about how you might do this.



253. The **least squares regression line**, sometimes called the "best fit line" or simply the "regression line", is the line that gets as close to as many points as possible. So in the previous question, the least squares regression line is what you are attempting to find. However, the real math behind it is a little unwieldy. In a bit, we will give some intuition behind it, but let's go ahead and give the equation to the regression line:

$$\hat{y} = a + bx$$

Why do you think the y has a hat on it? (We've seen hats before!) Thinking back to a past Algebra class, do you think you know what a and b are? Explain.

254. Assertion: *The regression line must always pass through the point* $(\bar{x}, \bar{y})$. Do you think this is true? Can you support your answer intuitively or with evidence?

255. In the regression equation, b can also be written as $b = r(\frac{s_y}{s_x})$, where s is the standard deviations of x and y, respectively. Earlier, it was mentioned that we sometimes want

to test whether the correlation is significantly different from 0. So challenge: Explain why testing whether the slope of the regression line is zero or not is essentially the same thing.

256. Consider the data in the table below. Let's find a regression equation using the TI84 calculator. In the STAT list editor, enter the x data into L1 and the y data into list L2. On the STAT TESTS menu, select LinRegTTest, and go through the remaining prompts. Hit Enter. What is your regression equation? What is the correlation? Is the slope and correlation significantly different from 0?

x	2	3	3	4	5	6	6
y	7	9	13	8	11	16	10

257. (Continuation) You may have noticed a statistic in the output that we haven't mentioned, r^2. This is called the **coefficient of determination**, and is literally just the correlation squared. What it tells us, though, is important: It gives the proportion (or percent, if multiplied by 100) of the variation in the response variable that can be explained by variation in the explanatory variable. Suppose a study plots third exam grades as the explanatory variable and final exam grades as the response variable, and in the analysis, an $r^2 = 0.65$ was found. What does this tell us?

258. How might a regression line serve as a basis for making predictions?

259. (Continuation) When using a regression line to make predictions, we must be very careful about **extrapolation**, or predicting values for the response variable for values of the explanatory variable that we actually don't know! For instance, if we wanted to predict snowfall by the amount of melted water left in a rain gage, we might have the following data:

Water (x)	0.5	2.6	0.9	0.1
Snowfall (y)	4.6	12.8	8.3	0.9

If you wanted to predict that 5 inches in your rain gage meant 45 inches of snow, is that a fair prediction? Why or why not? When could extrapolation be particularly deceptive?

260. The actual math behind the least squares regression line and how it is found is beyond the scope of this class, but I did want to briefly discuss the idea of the **residual**, which is $y - \hat{y}$, or the difference between the observed y-value and the predicted y-value. Think about this: If the linear model actually captures the data well, what might we expect to see if looking at the residuals?

261. (Continuation) The least squares regression line is actually found by taking the residuals, squaring them, adding them up, and then minimizing that sum using Calculus. The purpose of squaring is to get rid of any negative signs that may be appearing due to the difference found in the residual formula. But here's a question: Why not just use absolute value? It gets rid of negative signs, too, and appears to be much easier to work with than squaring a bunch of numbers! So, why do you think we square the residuals?

262. Why do you think the name is called "least squares"?

263. Data are collected on the relationship on the number of hours per week practicing a musical instrument and scores on a math test. The regression line found is $\hat{y} = 72.5 + 2.8x$. For someone who practices their musical instrument 5 hours a week, what score would you predict? Do you feel this model is reasonable? When would it not be reasonable?

264. A researcher is exploring how alcohol consumption may affect driving performance. In the table below, the x values are the mL of alcohol consumed by each participant in the study. The y values are the score (from 0 to 100 in tenths) on a closed-course driving simulator after 30 minutes of alcohol consumption. Is it reasonable to model this using linear regression? If so, find the linear regression equation, the correlation, and the r^2. Predict the driving score of someone who drank 75 mL of alcohol. What about 3000 mL of alcohol? Is there any danger in either of these predictions?

x	23	47	102	156	165	173	188	202	203
y	98	99	93	91	88	84	86	80	77

265. Summarize what you know about regression, correlation, and scatterplots.

266. We have reached the end of the course! What are 3 topics you feel you know fairly well? What are 3 topics you feel you didn't learn well at all? What are 3 topics you'd like to know more about?

1 Appendix A: Glossary

- 2SD Method: This is a method of approximating a confidence interval by taking the parameter estimate statistic and the standard deviation of the statistic (standard error) and extending two standard deviations in each direction from the statistic.
- Alternative hypothesis: This is the explanation that a real effect occurs; it is typically the research conjecture.
- ANOVA: Analysis of Variance test; an overall test of multiple means that explores the variation between groups compared to the variation within groups; used for 3 or more groups to test a difference in at least one of the means.
- Association: Two variables are associated or related if the distribution of one variable has a relationship with values of the other variable.
- Biased: A sampling method is biased if statistics from different samples consistently overestimate or consistently underestimate the population parameter of interest.
- Binary variable: Categorical variable with only two outcomes.
- Binomial distribution: A discrete distribution that involves a fixed number of independent trials; the random variable X is defined as the number of successes in n trials.
- Boxplot: A visual display of the range, IQR, and median, summarizing the behavior of a quantitative variable
- Categorical variable: Variables that take on values that are names or labels.
- Census: Data gathered on the entire population; not done very often in practice due to cost and difficulty.
- Central Limit Theorem: A mathematical theory (a very important one) that predicts the behavior of the null distribution when certain validity conditions are met; under these certain conditions, the Normal approximation can be used for the sampling distribution.
- Chance model: A real or computerized process to generate data according to a well-understood set of conditions called a probability model; data generated in a chance model should be random and based on the underlying probability.
- Chi-square distribution: A non-negative, right-skewed distribution used for the chi-square tests for an association between two categorical variables (among other tests).
- Chi-square statistic: A standardized statistic for summarizing the relationship between two categorical variables which has a predictable null distribution (the chi-square distribution); we use this for the goodness-of-fit test.
- Cluster random sample: A method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters; every individual in the chosen clusters is included in the sample.

- Coefficient of determination: also called r^2; a statistic that measures the percentage of total variation in the response variable that is explained by the linear relationship with the explanatory variable; it is literally the square of the correlation.
- Complementary Probability Rule: The complement of an event A is 1 - probability of event A.
- Confidence interval: AN inference tool used to estimate the value of a parameter by forming a range of values around the parameter estimate based on the uncertainty due to the randomness in the sampling method.
- Confidence level: The probability or percent of the time that the true parameter will be captured by the confidence interval; it is usually 95% but not always.
- Confounding variable: A variable that is related to both the explanatory and response variable in such a way that its effects on the response variable cannot be separated from those of the explanatory variable.
- Convenience Sampling: A nonrandom sample of a population; typically, this takes the form of surveying or using individuals that are easy or convenient at the time instead of a random sample.
- Continuous distribution: A distribution where the variable is continuous, ie, measured on a continuum.
- Control: A placebo or do-nothing group in an experiment.
- Correlation (correlation coefficient): Statistic that measures the direction and strength of a linear relationship between two quantitative variables.
- Cumulative relative frequency: The sum of the relative frequencies for all values that are less than or equal to the given value.
- Data:a set of observations
- Datum: the singular of data
- Degrees of Freedom: The number of objects in a sample that a free to vary.
- Difference: A subtraction of two variables.
- Discrete Distribution: a distribution which can only take finite values of its variable.
- Double-blind: A study design where neither the subjects nor those evaluating the response know which treatment group each subject is in.
- Empirical Rule: A shorthand memory device for the Normal distribution; 68% of data falls with $\pm$ 1 standard deviation of the mean; 95% of the data falls within $\pm$ 2 standard deviations of the mean; almost all (99.7%) of the data falls within $\pm$ 3 standard deviations of the mean.
- Expected Value: Expected arithmetic average when an experiment is repeated many times; can be thought of as the mean.

- Experiment: A study in which researchers actively assign subjects to treatment groups.
- Explanatory variable: The variable that, if the alternative hypothesis is true, is explaining the changes in the response variable, sometimes know as the independent or predictor variable.
- Extrapolation: Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data.
- F-distribution: The distribution of the F-statistic; non-negative and skewed right; it is used for the ANOVA test.
- F-statistic: Ratio of variation between the groups to the variation within the groups; used for the ANOVA test.
- Fail to reject the null hypothesis:
- Frequency: the number of times a value of the data occurs.
- Geometric Distribution:A discrete distribution in which the trials occur until the first success; the random variable X is the number of trials until the first success.
- Histogram: A graph used with quantitative variables; this takes the form of a bar graph where the bars touch; the x-axis is the explanatory variable, and the y-axis is the response variable.
- Influential observations: may or may not be outliers; points of data that greatly influence the correlation and regression line; special care must be taken with these points.
- Institutional Review Board (IRB): A committee tasked with oversight of research programs that involve human subjects.
- Interquartile Range: The difference between the upper quartile and the lower quartile; the middle 50% of the data; a measure of spread of a quantitative variable.
- Interval data: Data that is numeric (so arithmetic operations make sense), but no true zero.
- Law of Large Numbers: As the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency probability approaches zero.
- Least squares regression line: The best fit line on a scatterplot, in the sense that the line gets as "close" as possible to all the data points.
- Left-skewed: Most of the observations tend to fall on the right side of the distribution with a left tail extending outwards.
- Location parameter: Another phrase for the mean; the mean controls where the distribution is centered.
- Lower Quartile: The value for which 25% of the data lie below.

- Margin of error: Half of the width of a confidence interval; the confidence interval multiplier times the standard error of the statistic.
- Matched (paired) data: Data where two very similar individuals (or, in some cases, the same individual in a "before" and "after" situation) are grouped instead of using totally independent groups; allows the reduction of variability in the response variable; calculated using the differences between values.
- Mean: A number that measures the center of a set of data; commonly called the "average"; it is the sum of all values in the population divided by the number of values in the population (population size).
- Mean square for error: Denominator of the F-statistic; measures within-group variation; it is similar to averaging the standard deviations (squared) across the groups being compared.
- Mean square for treatment: Numerator of the F-statistic; measures the variation among the group means
- Median: The middle data value when the data are sorted in order from smallest to largest.
- Mode: The value that occurs most frequently in a set of data.
- Nominal data: data that merely labels observations, but does not have order or number.
- Nonrandom errors: Reasons why the statistic may not be close to the parameter that are separate from random chance error. These can include bias or systematic error built-in to the process and will not be picked up by the statistical process.
- Normal distribution (normally distributed): A classic bell-shaped distribution that occurs frequently in statistics; it is a continuous, mound-shaped distribution centered at its mean μ with a spread determined by its standard deviation σ.
- Null distribution: The distribution that represents what could happen in the study assuming the null hypothesis is true.
- Null hypothesis: The explanation that an event occurs by random chance alone; this is the statement in a research study that "no effect" occurs or that nothing interesting happens; the null hypothesis is what most statistical tests are based on.
- Numerical variable: variables that take on values that are indicated by numbers.
- Observational studies:Studies in which researchers merely observe, but do not intervene, and measure variables of interest.
- One sample t-test: A statistical test on one sample of data that uses the Student's t distribution to test inferences about the population mean.
- One-sided hypothesis: An alternative hypothesis that is "greater than" or "less than", and as such, goes in only one direction.
- Ordinal data: Data that has a defined order, but is not numeric.

- Outlier: An observation that does not fit the overall pattern of the distribution; outliers can cause major problems with statistics.
- P-value: The probability of obtaining a value of the statistic at least as extreme as the observed statistic when the null hypothesis is true; in many studies, a p-value of 0.05 or less is considered significant.
- Parameter: The number that describes a property about a population as a whole; in practice, parameters are usually not known and must be estimated by statistics.
- Percentile: A number that divides ordered data into hundredths; the median is the 50th percentile; the lower and upper quartile are the 25th and 75th percentile, respectively.
- Pie chart: A circular display of categorical data.
- Poisson distribution: A discrete distribution that counts the number of times a certain event will occur in a specific interval
- Pooled: Data that is pooled means it is grouped together.
- Population: The entire collection of subjects we are interested in; we usually take a sample from the underlying population.
- Population mean: μ; the average of the entire population.
- Population proportion: p or Π; the proportion of the entire population.
- Population standard deviation: σ; the standard deviation of the entire population.
- Power: The probability of rejecting a false null hypothesis; in other words, it is the ability of a statistical test to find a real effect if there is one.
- Probability: A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur.
- Proportion: The number of successes divided by the total number in the sample.
- Qualitative data: Data that is not numeric in nature; arithmetic operations (adding, subtracting, etc.) don't make sense.
- Quantitative data: Data for which arithmetic operations (adding, subtracting, etc.) make sense.
- Quasi-experiment: Experiments that manipulate the explanatory variable in a nonrandom way.
- Random variable: a function from the sample space to the real number interval 0 to 1; converts list of a sample space into a probability
- Randomized experiment: A study where experimental units are randomly assigned to two or more treatment conditinos and the explanatory variable is actively imposed of the subjects.
- Range: The highest data point minus the lowest point.

- Ratio data: Data that is numeric and also has a true zero; all arithmetic operations can work on it, and is the highest scale of measurement of data.
- Reject the null hypothesis: Enough evidence has been presented to claim the null hypothesis is probably not true.
- Relative frequency: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes.
- Resistant (robust): A statistic is resistant (or robust) if its value does not change much when extreme observations are added or removed from a data set.
- Residual: The vertical distances between a point and the least squares regression line; found by $y - \hat{y}$.
- Response variable: The variable that, if the alternative hypothesis is true, is impacted by the explanatory variable; also called the dependent variable.
- Right-skewed: The bulk of the observations fall on the left side of the distribution with a tail to the right.
- Sample: The set of observed values; a sample is a subset taken from an underlying population
- Sample mean: $\bar{x}$, the average of the sample.
- Sample proportion: $\hat{p}$, the proportion of the sample.
- Sample space: The set of all possible outcomes of an experiment
- Sample standard deviation: s, the standard deviation of the sample.
- Sampling method: A way of surveying a sample for a given study.
- Scales of measurement: There are four scales of measurement; these dictate what types of statistics can be done on the data; they are (in order) nominal, ordinal, interval, and ratio.
- Scatterplot: A graphical summary of the relationship between two quantitative variables; each dot on the scatterplot shows the values for both variables.
- Shape parameter: Another word for the standard deviation; controls the shape of the distribution.
- Significance level: A value used as a criterion for deciding how small a p-value needs to be to provide convincing evidence against the null hypothesis; it is 1 - confidence level; also called α, or alpha. The most common is 0.05.
- Simple random sample: A sampling method that ensures that every sample of size n is equally likely to be the sample selected from the population; in other words, every individual in the population has equal chances of being selected as every other individual.
- Simulation: A simulation repeats a trial many times over, usually on a computer, to create an empirical distribution that mimics theoretical results.

- Skewed: The bulk of the values in the distribution fall on one side; the distribution is not symmetric.
- Standard deviation: A measure of the spread of the distribution; the average distance each point is from the mean; the sum of squared deviations from the mean divided by the degrees of freedom.
- Standard error: An estimate of the standard deviation of a statistic that is based on the data; usually, the statistic this is based on is the parameter estimate sample statistic.
- Standard uniform distribution: A uniform distribution on (0,1).
- Statistic: A number computed from a sample; it is often used to estimate a parameter of the underlying population
- Statistical significance: Unlikely to occur just by random chance; the most common threshold of statistical significance is a p-value of 0.05.
- Stratified random sample:A method for selecting a random sample used to ensure that subgroups of the populatino are represented adequately; divide the population into groups (Strata); use simple random sampling to identify a proportionate number of individuals from each stratum.
- Systematic random sample: A kethod for selecting a random sample, where the starting point is randomly chosen, and then every nth individual is selected.
- t-statistic (t-distribution): The standardized statistic (and underlying distribution) for the sample mean using the standard deviation of $\bar{x}$.
- Tail: The outer edge of a distribution; each distribution has a left and right tail.
- Test of significance: A procedure for measuring the strength of evidence against a null hypothesis about the parameter of interest.
- Treatment: The assigned conditions in an experiment.
- Two sample t-test: The t-test done for two samples; refer to Appendix B for more info.
- Two-sided hypothesis: in a two-sided hypothesis, p-values are considered that are at least as extreme as our observed result (the statistic estimating the parameter) in either direction; this usually takes the form of a "not equal to" alternative hypothesis.
- Type I error: The mistake made when one rejects the null hypothesis when it is actually true (a false alarm).
- Type II error: The mistake made when one fails to reject a null hypothesis that is actually false (missed opportunity).
- Unbiased: On average, across many random samples, the sampling method produces statistics whose average is the value of the population parameter; in other words, the expected value of the statistic is the population parameter.

- Uniform distribution: A continuous random variable that has equally likely outcomes over the domain (a,b).
- Upper Quartile: The value of which 75% of the data lie below.
- Variable: A characteristic of interest for each individual in a population.
- Variability: The spread of a distribution; typically measured by the standard deviation or the variance (the square of the standard deviation)
- z-statistic: the linear transformation of the form $z = \frac{x-\mu}{\sigma}$; used for Normally-distributed data tests.

2 Appendix B: Summary of Common Inferential Methods

2.1 Hypothesis Test for One-Sample Proportion

This hypothesis test involves estimating a parameter by using the sample proportion $\hat{p}$ and an interval around it.

This test uses the z-statistic. The formula is

$$z = \frac{\hat{p} - \Pi}{\frac{\hat{p}(1-\hat{p})}{\sqrt{n}}}$$

where $\hat{p}$ is the sample proportion, Π is the hypothesized value of the population proportion under the null hypothesis, and the denominator is the standard error of the population proportion.

The p-value can be found using a TI84 Calculator (see Appendix C) or using software. A common threshold for rejecting the null hypothesis is 0.05, but it can theoretically be any value between 0 and 1.

Practice:

1. The CEO of a large electric utility claims that 80 percent of his 1,000,000 customers are very satisfied with the service they receive. To test this claim, the local newspaper surveyed 100 customers, using simple random sampling. Among the sampled customers, 73 percent say they are very satisfied. Based on these findings, can we reject the CEO's hypothesis that 80% of the customers are very satisfied? Use a 0.05 level of significance.

2. In the previous example, the CEO claimed that 80 percent of the customers were satisfied. Perhaps a better question would be if at least 80 percent of the company's customers are very satisfied. Again, 100 customers are surveyed using simple random sampling. The result: 73 percent are very satisfied. Based on these results, should we accept or reject the CEO's hypothesis? Assume a significance level of 0.05.

3. Newborn babies are more likely to be boys than girls. A random sample found 13,173 boys were born among 25,468 newborn children. The sample proportion of boys was 0.5172. Is this sample evidence that the birth of boys is mor ecommon than the birth of girls in the entire population? Use a significance level of 0.05.

4. A coin is flipped 40 times. In those 40 times, 34 of them came up Heads. Is there strong enough evidence that the coin is biased? Use a significance level of 0.05.

2.2 Hypothesis Test for Single Population Mean

This hypothesis test involves estimating a parameter by using the sample mean $\bar{x}$ and an interval around it.

This test uses the t-statistic. The formula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where $\bar{x}$ is the observed sample mean, μ_0 is the hypothesized population mean under the null hypothesis, s is the sample standard deviation, and n is the sample size.

The p-value can be calculated using a TI84 Calculator (see Appendix C) or using software. A common threshold for rejecting the null hypothesis is 0.05, but it can theoretically be any value between 0 and 1.

Practice:

1. A college football coach records the mean weight that his players can bench press as 275 pounds. Three of his players thought the mean weight was more than than that amount. They asked 12 teammates to lift and test the results. Their results are: 205, 215, 225, 252, 265, 275, 313, 316, 338, 341, 345, and 385. Conduct a hypothesis test using a 5% level of significance.

2. Students in a class believe that the mean score on the first test is a 65. The instructor thinks the mean is higher than that. He samples ten students and obtains scores 65, 65, 70, 67, 66, 63, 63, 68, 72, and 71. He uses a 5% significance level. What is the result of such a significance test?

3. The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass. 1.11, 1.07, 1.11, 1.07, 1.12, 1.08, 0.98, 0.98, 1.02, 0.95, 0.95. Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use significance level of 0.05.

4. Your company wants to improve sales. Past sales data indicate that the average sale was \$100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicate an average of \$130, with a standard deviation of \$15. Did the training work? Test your hypothesis at 5% α level.

2.3 Hypothesis Test for Matched or Paired Samples

This hypothesis test involves comparing two population means when you have two samples in which observations from one sample can be paired or matched to the ot her sample.

This test uses the t-statistic. To calculate this statistic, we will use a slightly different formula than the one introduced in the questions, just to show you different notations. First, calculate the difference d_i of each of the i observations, then calculate the mean of the d_i, which we will call $\bar{d}$. Next, calculate the standard deviation of the differences, which we will call s_d. Now, the t-statistic will be given by the formula:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

where n is the sample size.

The p-value can be calculated using a TI84 Calculator (see Appendix C) or using software. A common threshold for rejecting the null hypothesis is 0.05, but it can theoretically be any value between 0 and 1.

Practice:

1. For a given exam, 20 students scored the following on a pre-test given before instruction: {18,21,16,22,19,24,17,21,23,18,14,16,16,19,18,20,12,22,15,17} and received the following results on the test: {22,25,17,24,16,29,20,23,19,20,15,15,18,26,18,24,18,25,19,16} (Note: The data is paired, so the first score in the first set is the same person as the first score in the second set, and so forth.) At the 5% significance level, what do you conclude?

2. Suppose 14 individuals are taking a medicine to help improve their diabetes. Their A1C levels before the medicine are {5.6, 6.4, 6.6, 6.8, 7.4, 7.6, 8.3, 8.5, 8.7, 9.2, 10.3, 10.4, 10.4, 10.5} and after taking the medicine for 30 days, their A1C levels are {4.5, 5.3, 5.3, 5.5, 6.7, 7.5, 7.7, 7.8, 8.2, 8.5, 8.8, 9.1, 9.3, 9.5 } (Note: The data is paired, so the first score in the first set is the same person as the first score in the second set, and so forth.) At the 5% significance level, what do you conclude?

3. A teacher allowed students to retake their test again for a better score. She wanted to know if there was a significant gain in score. The scores are shown below. Use a 5% significance level. What do you conclude?

Subject	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

4. A swim coach measured the swim times for the team before the semester and after the semester. Results are shown in the table below. Use a 5% significance level to decide if the team improved or not.

Before	After
14.3	11.3
15.5	14.7
15.6	15.5
14.7	14.8
15. 7	14.1
17.3	15.3
14.1	14.0
15.0	13.0

2.4 Hypothesis Testing for Two Proportions

This hypothesis test involves comparing two proportions. Here, we will use a pooled sample proportion to compute the standard error of the sampling distribution:

$$p_{\text{pooled}} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

where p_1 is the sample proportion from population 1, p_2 is the sample proportion from population 2, n_1 is the sample size of sample 1, and n_2 is the sample size of sample 2.

The standard error, using the pooled estimate, is given by:

$$\sqrt{p \cdot (1-p) \cdot [\frac{1}{n_1} + \frac{1}{n_2}]}$$

This test uses the z-statistic. The formula:

$$z = \frac{p_1 - p_2}{\text{SE}}$$

where SE is the standard error formula above. Note the hypothesized proportion difference under the null hypothesis is 0, so it is excluded from the numerator of the z-statistic.

The p-value can be calculated using a TI84 Calculator (see Appendix C) or using software. A common threshold for rejecting the null hypothesis is 0.05, but it can theoretically be any value between 0 and 1.

Practice:

1. A certain drug company is developing a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a SRS of 100 women and 200 men from a population of 100,000 volunteers. At the end of the study, 38% of women caught a cold and 51% of men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

2. Suppose a sample of 200 New York voters found 88 who voted for the Republican presidential candidate, while a sample of 300 California voters found 143 who voted for the same candidate. Test the claim that there is no difference between the two states in the proportions who favored the Republican candidate.

3. Is there a significant difference between the proportion of Republicans and the proportion of Democrats that identify "Immigration" as a major problem? In a particular survey, 198 of 300 Republicans surveyed said "yes" to the question, while 250 Democrats surveyed said "yes" to the question. Test the claim that there is no difference between the two party views.

4. In a recent survey, 63% of men viewed COVID as a major problem facing society, while 78% of women said the same. The survey was conducted among 1,035 males and 975 females. Is there evidence to support that the proportion of males that think COVID is a major issue is different from the proportion of females who think COVID is a major issue?

2.5 Hypothesis Testing for Two Means

This hypothesis test involves comparing two means to see if they significantly differ from one another.

For this test, we will let d be the hypothesized value of the null hypothesis difference if it were true. It is usually 0.

The standard error formula is:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The degrees of freedom formula is really ugly and not elaborated upon here.

This test uses the t-statistic. The formula:

$$t = \frac{(\bar{x_1} - \bar{x_2} - d)}{\text{SE}}$$

where $\bar{x_1}$ is the mean of the first group, $\bar{x_2}$ is the mean of the second group, and d and the SE are defined above.

The p-value can be calculated using a TI84 Calculator (see Appendix C) or using software. A common threshold for rejecting the null hypothesis is 0.05, but it can theoretically be any value between 0 and 1.

Practice:

1. Within a school district, students are randomly assigned to one of two Math teachers - Mrs. Smith and Mrs. Jones. After the assignment, Mrs. Smith had 30 students, and Mrs. Jones had 25 students. At the end of the year, each class took the same standardized test. Mrs. Smith's students had an average score of 78, with a standard deviation of 10; Mrs. Jones's students had an average score of 85, with a standard deviation of 15. Are their scores significantly different? Use a 5% level of significance.

2. A company has developed a new battery. The engineer in charge claims that the new battery will operate continuously for at least 7 minutes longer than the old battery. To test the claim, the company selects a SRS of 100 new batteries and 100 old batteries. The old batteries ran continuously for 190 minutes with a standard deviation of 20 minutes; the new batteries ran with a mean time of 200 minutes with a standard deviation of 40 minutes. Test the engineer's claim; use 0.05 level of significance.

3. At one KFC restaurant, a SRS of 30 orders of potato wedges contained an average of 280 calories with a standard deviation of 16 calories. At a second KFC restaurant down the road, a SRS of 25 orders of potato wedges contained an average of 304 calories, with a standard deviation of 22 calories. Is there significant evidence of a difference in the number of calories in each store's potato wedges? Use 5% level of significance.

4. Under normal conditions, is the average body temperature the same for men and women? Medical researchers interested in this question collected data from a large number of men and women, and random samples from that data are presented below. Is there sufficient evidence to indicate that mean body temperatures differ for men and

women?

Men	Women
96.9	97.8
97.4	98.0
97.5	98.2
97.8	98.2
97.8	98.2
97.9	98.6
98.0	98.8
98.6	99.2
98.8	99.4

2.6 Confidence Interval for Population Proportion

A confidence interval for proportions is computed when we have a categorical characteristic being measured - for example, political party or gender. We find the proportion of individuals as $\hat{p}$ in our sample and use it to estimate the population parameter Π.

The general form for a confidence interval is:

$$\text{observed statistic} \pm \text{multiplier} \cdot \text{standard error of the observed statistic}$$

As already mentioned, $\hat{p}$ will be our observed statistic. The formula for standard error of the proportion is:

$$\sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

The multiplier is trickier to figure out, but depends on the confidence level and the z-distribution. You can use a calculator or software to figure out the multiplier (most will give it to you if you just specify the confidence level), or you can use the table below for some commonly-use confidence levels:

Confidence Level	Multiplier
90%	1.645
95%	1.96
99%	2.576

The confidence interval will form a range around the observed statistic that captures the population parameter C% of the time, where C is the confidence level.

Practice:

1. Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. 500 randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people sampled, 421 responded yes. Use a 95% confidence level to computer a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

2. Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250, 98 reported owning a tablet. Use a 95% confidence level to compute a confidence interval estimate for the true proportion of people who own tablets.

3. A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation. Compute a 95% confidence interval for this data.

4. A financial officer for a company wants to estimate the percent of accounts receivable that are more than 30 days overdue. He surveys 500 accounts and finds that 300 are more than 30 days overdue. Compute a 95% confidence interval for the true percent of accounts receivable that are more than 30 days overdue.

2.7 Confidence Interval for Mean

This procedure finds a confidence interval around the sample mean, $\bar{x}$, in order to estimate the population mean, μ.

The general form for a confidence interval is:

$$\text{observed statistic} \pm \text{multiplier} \cdot \text{standard error of the observed statistic}$$

As already mentioned, $\bar{x}$ will be our observed statistic. The formula for standard error of the mean is:

$$\frac{s}{\sqrt{n}}$$

The multiplier is trickier to figure out. Using your TI84 or software to figure it out, you just need to input the confidence level you are looking for and the degrees of freedom, which is found by:

$$df = n - 1$$

Practice:

1. Suppose a SRS of 150 students is drawn from a population of 3000 students. Among sampled students, the average IQ was 115 with a standard deviation of 10. What is the 95% confidence interval for the students IQ scores?
2. A new therapeutic for a disease is being tested. Among a SRS of 255 adults, the average improvement over two weeks while taking the medicine on a patient-reported scale was 2.6 with a standard deviation of 0.6. What is the 95% confidence interval for the therapeutic score?
3. A group of 10 foot surgery patients had a mean weight of 240 pounds with a standard deviation of 25 pounds. Find a confidence interval for a sample mean for the true mean weight of all foot surgery patients. Use 95% confidence level.
4. In a survey of 355 college students, they self-reported their GPA as a sample mean of 3.1 on a 4.0 scale with a standard deviation of 0.3. What is the 95% confidence interval for this?

2.8 Chi-Square Goodness of Fit Test

The Chi-square Goodness of Fit test is done is applied when you have a categorical variable from a single population to determine whether sample data are consistent with a hypothesized distribution.

For this, we will use the expected frequency count. To calculate this, use the following formula:

$$E_i = np_i$$

where E_i is the expected frequency count of the i'th level of the categorical varaible, p_i is the hypothesized proportion of observations in level i, and n is the sample size.

The chi-square test statistic is calculated as:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency count for the i'th level of the categorical level, and E_i is defined above.

The p-value can be found by using a TI84 calculator or software, and is found using the formula for degrees of freedom:

$$df = k - 1$$

where k is the number of levels of the categorical variable.

Practice:

1. A toy company prints baseball cards. The company claims that 30% of the cards are rookies, 60% are non-All Star veterans, and 10% are veteran All-Stars. Suppose a SRS of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with the company's claim? Use 5% significance level.

2. Suppose we expect that M&Ms to have equal numbers of each of the six colors. A SRS of 600 M&M candies is made with the following distribution: 212 are blue, 147 are orange, 1003 are gree, 50 are red, 46 are yellow, 42 are brown. Is it likely that the M&Ms are distributed evenly? Test at the 5% significance level.

3. A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of 6's	Number of Rolls
0	48
1	35
2	15
3	3

   Note that the usual probability with 3 dice of rolling a 6 is as follows: zero sixes, 0.58 ; one six, 0.345; two sixes, 0.07; three sixes, 0.005. Are the dice fair? To answer this, perform a Chi-square goodness of fit test with 5% significance level.

4. Suppose a town of 2,500 residents wants to see if their demographic makeup has changed in the past 5 years. During the last census, the results were as follows:

Ethnicity	Distribution
White	0.743
Black	0.216
American-Indian	0.012
Hispanic	0.012
Asian	0.008
Others	0.009

Five years later, the observed count of White citizens was 1732, of Black citizens was 538, of American-Indian citizens was 32, of Hispanic citizens was 42, of Asians was 133, and of Others was 23. Is there significant evidence that the makeup of the town has changed? Use 5% level.

2.9 One-Way ANOVA

A one-way ANOVA is used when the explanatory variable is categorical and response variable is quantitative, and there is at least 3 levels of the explanatory variable. We wish to see if there is any statistically significant difference among all of the means of each group.

The ANOVA also assumes independence of each individual, normality of the underlying populations, equality of variance among the groups, and that the design is a randomized experiment.

With ANOVA for three groups, the null hypothesis is $H_o : \mu_1 = \mu 2 = \mu 3$, and the alternative is that at least one of the means differs from the others.

Calculation of the ANOVA is typically done by software or TI84 (and higher) calculators. However, it can be done by hand when the number in the sample sizes is relatively small.

First, you'll need to calculate the grand mean, which is the mean, $\bar{x}$, of all observations. Second, calculate the mean of each group, $\bar{x_j}$, as we will need these.

Next, calculate the between-groups sum of squares:

$$\Sigma n_j(\bar{x_j} - \bar{x})^2$$

which is each group mean subtracted by the grand mean, squared, and multiplied by the number in that group.

The next step is to calculate the within-groups sum of squares:

$$\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} (x_{ij} - \bar{x_j})^2$$

which is each individual in group j subtracted by the mean in group j, squared, and added up for every group into one single sum.

The degrees of freedom for between-groups is the number in each group - 1. The degrees of freedom for within-groups is the total sample size minus the number of groups.

Divide each Sum of Squares by their respective degrees of freedom to get the respective Mean Squares, the Mean Squares for Treatment is the between-groups statistic, and the Mean Squares for Error is the within-groups statistic.

The F-statistic for the ANOVA test is:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

Got all that? Again, a TI84 calculator or software will do this for you also, much quicker.

Practice:

1. You survey people about their social media use and assign them groups of "low", "medium", and "high". You want to see if there is a difference in hours of sleep per night. A small number of the people surveyed appears in the table below. Conduct a one-way ANOVA and decide if there is a significant difference in the group means. Use 5% significance.

Low	Medium	High
6.5	6	7
8.5	11	9.3
7.1	6	3
7.5	8	5
11.6	9	7
8.3	8.6	8. 9

2. You are curious if there is a difference in mean price of a 20oz bottle of soda based on the brand. You collect data from 4 different stores on Coke, Pepsi, and Sprite. The data is shown below. Conduct a one-way ANOVA and determine if there is a significant difference in the group means. Use 5% significance.

Coke	Pepsi	Sprite
1.99	1.99	1.99
2.19	2.09	1.79
1.99	2.09	1.89
1.95	1.99	2.05

3. A farmer is experimenting with types of fertilizer to see if there is an increase in crop yield. The following results were obtained over four fields. Conduct a one-way ANOVA and determine if there is a significant difference in group means. Use 5% significance.

Treatment 1	Treatment 2	Treatment 3
40.7	63.2	45.5
38.9	31.1	40.6
44.5	46.6	35.7
48.9	39.7	42.5

4. A student is studying whether college students of different classifications weigh differently. So far, she only has 4 students from each classification. Would her results so far be significant? Use a one-way ANOVA to find out. Use 5% significance.

Freshmen	Sophomore	Junior	Senior
165	177	213	155
141	255	189	156
304	185	175	202
173	170	177	215

2.10 Linear Regression

When we have two quantitative variables, we often wish to see if there is an association, specifically a linear association, between those variables. This leads us to the regression equation:

$$\hat{y} = a + bx$$

where $\hat{y}$ is the estimate of y, and x is the value of our explanatory variable at each data point. The parameter a serves as the y-intercept of the regression equation and can be found by the following formula:

$$a = \bar{y} - b\bar{x}$$

and b can be found by the formula:

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

On the TI84 and software, you can get the correlation and r^2, as well as the entire equation of the regression line easily. Many calculators that are just $8-$10 will even do this.

Practice:

1. A student wants to find out if there is any relationship between the distance from school and the cost of school supplies this semester. The results are summarized in the table below. Find the r, r^2, and regression line for this data.

Distance from School	Cost of Supplies
0.3	245
14.6	233
3.5	185
26.7	202
0.8	245
11.7	288
1.2	212
10.5	255

2. A student wants to predict the cost of a textbook based on the number of pages in the textbook. The first few books he chose are below. Summarize the results of a regression, and predict how much a textbook would cost if it has 1015 pages.

Number of Pages	Cost of Textbook
212	65
756	133
516	185
1367	262
188	145
415	178
522	166
888	199

3. Do you think the weight of a car is related to its fuel efficiency? The weight in pounds of several cars is shown below, along with their fuel efficiency in mpg. Conduct a regression analysis of this, and predict the fuel efficiency for a truck weighing 4320 pounds.

Weight	Fuel Efficiency
1765	39.6
2755	24.6
2223	33.5
4115	17.9
1512	42.7
2756	34.6
4576	15.4
3002	32.2

4. The age in years and the height in inches is given for several school children. Use this and a linear regression to predict the height of a 10 year old child.

Age (in years)	Height (in inches)
7.3	46
4.5	37
5.8	39
6.6	41
7.8	49
8.4	53
11.8	58
9.5	55

3 Appendix C: TI84 Statistical Tutorial

There are many, many uses of the TI84 for statistics, but I'll go over a few of them here. If in doubt, a search engine is always your friend!

- STAT $->$ EDIT to place all your data in Lists
- STAT $->$ CALC $->$ 1-VAR STATS L1 to calculate the descriptive statistics - mean, standard deviation (sample), standard deviation (population), sample size, lower quartile, median, and upper quartile.
- To get to most of the distributions, hit 2nd VARS (which should be DISTR). There are quite a few of these commands talked about in the questions, but a few will be discussed as follows. They all work basically the same.
- invnorm(percentage,mean,sd) - gives a z-statistic if you know the percentile, mean, and sd
- normalcdf(lower,upper,mean,sd) - finds the probability on an interval using the Normal distribution. To enter ∞, use 10 to the 99th power.
- tcdf(lower,upper,df) - finds the probability on an interval using the t-distribution. Again, use 10 to the 99th power for ∞.
- invT(left tail area, df) - gives the t-statistic if you know the left tail area and df.
- binompdf(#trials, prob of success, successes desired-optional) - uses the binomial distribution to give a probability of a specific outcome
- binomialcdf(# trials, prob of success, #successes desired) - same as previous, but returns the probability of all outcomes equal to or less than the desired probability.
- If you go to STAT $->$ TESTS, you'll arrive at all of the tests that we do.
- T-Test - this is the one-sample t-test
- 2-SampleTTest - this is the two-sample t-test
- 1-PropZTest - single proportion z-test
- 2-PropZTest - two proportion z-test
- Chi-Square GOF Test - the Goodness of Fit Chi-Square Test
- TInterval - the confidence interval for a t-statistic
- 2-SampTInt - the confidence interval for a two-sample t-statistic
- 1-PropZInt - the confidence interval for the one-proportion z-statistic
- LingRegTTest - computes the linear regression on data, and also does a t-test on the value of the slope and correlation to see if they are significantly different from zero.
- ANOVA - computes a one-way ANOVA test for comparing the means of 2 to 20 populations. Shows an F-statistic and gives the sum of squares and mean squares.

4 Appendix D: Culminating Project

As your final act of mastery in this course, you are to work collaboratively on an authentic statistics project. This project requires you to develop, execute, and analyze a survey, observational study, or experiment. You may do this venture solo, with 1 partner, or with 2 partners. You will entirely share grades, so make sure if you have a partner that you fully trust their work and work ethic!

Purpose: This will give you real experience working with statistics in a way that is meaningful. You must form a research hypothesis (or multiple hypotheses), design the study/survey/experiment, conduct the study/survey/experiment, collect the data, describe the data, perform analyses, and draw valid conclusions.

Topic: You are entirely free to choose your own topic! Choose something that is interesting to you, BUT beware that degree of difficulty will factor into the grade. Therefore, if you choose a question that is easily answerable without much effort, your grade will suffer. Likewise, if you choose a very challenging topic that is beyond the scope of this class, but you deliver solid work and do a good analysis, your grade will be slightly more lenient. The American Statistical Association has some ideas if you want to explore at https://www.amstat.org/asa/education/Statistics-Students.aspx

Proposal: You must develop a topic proposal submitted on Canvas a minimum of 14 days before the due date. (The due date will be communicated on Canvas and multiple times in class.) The proposal itself is not for a grade, but it MUST be turned in in order to get my approval. Your proposal must clearly explain your hypotheses, your explanatory and response variables, the test and/or intervals you will use to analyze the results, and how you will collect the data so the results for the inference will be satisfied. If you intend to use human data, you must also make sure your study will be safe and ethical (anonymous, able to quit at any time, informed consent, etc.) NO personal information that is sensitive to your subjects is permissible (so no grades/GPA).

Paper: You must write a minimum 5-8 page paper, 12 size font, Times New Roman, double-spaced. Graphs and tables are included in the length, but should take up no more than half of the total length. The key is communication and organization. Make sure all components of the paper are focused on answering the question(s) of interest and that correct statistical vocabulary is used. The paper should include:

- Title Page - include the title, the author(s), and the college. This should be on its own page (does NOT count towards the 5-8 pages).

- Abstract: The Abstract should summarize the entire paper on its own page (does NOT count towards the 5-8 pages) and be 100-250 words. The hypothesis and/or questions should be addressed, and the major results found.

- Introduction: The introduction should discuss what question you are trying to answer and why you chose this topic. Also, define the parameters of interest and set up the rest of the paper.

- Methods: The Methods section should detail your data collection - be specific! Explain how your survey/study/experiment was designed and how it was executed/conducted. Make sure to include detailed information about the sample and population of interest. In this section, you will also show the analysis and tests you plan to conduct and what your stated hypotheses are.

- Results: The Results section should give your raw data, graphs, summary statistics, and any other relevant statistics. In particular, you need to detail the results of your analysis on the main questions of the study/survey/experiment and how significant any results are. Please provide graphs, tables, and displays as appropriate to make it easier to understand the flow of information. Make sure the graphs are well-labeled, easy to compare, and help answer the question of interest.

- Discussion: In this section, you will have already listed your results, but now is the time for you to discuss the implications of those results. Is your result important? What questions are left on the table, unanswered? Could further research be done on this topic -if so, what? What possible errors (Type I, Type II, nonrandom, etc.), limitations, or interpretations are necessary to acknowledge? Any other critical reflections you have goes here.

Presentation: The group will also be responsible for a 7-15 minute oral presentation to the class. All group members do not need to speak in the presentation, but if questions are asked, they may be directed at anyone in the group. The presentation should summarize the results of the paper and clearly communicate the statistical importance of the survey/study/experiment. Some ideas that should definitely be discussed, both in the paper, and the presentation, should be the Research Question, the Parameters, the hypotheses, any assumptions and conditions, variables, statistical tests used, and the overall data collection methodology, as well as the ethics and safety concerns. As part of your presentation, you may opt to have a poster or PowerPoint to help you in presenting, but this is not required.

Grade: You will receive one test grade for the paper and a second test grade for the presentation, so in total, this amounts to two test grades. A rubric will be given out closer to time that details a bit more of what is expected to get maximum points.

5 Appendix E: Class Discourse Grading Sheet

Directions: This sheet will show the Class Discourse grade for each students. During each week of class, students can earn up to 4 points in Class Discourse. 1 point is earned each time a student presents work on the board, and 1 point is earned each time a student substantially contributes to class discussion. The professor will have this sheet with him/her each day and make marks as appropriate. Each checked box signifies 1 point gained. Class Discourse point opportunities end at Week 13.

Student	Wk 1	Wk 2	Wk 3	Wk 4	Wk 5	Wk 6	Wk 7	Wk 8	Wk 9	Wk 10	Wk 11	Wk 12	Wk 13
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐
	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐	☐☐ ☐☐

The total Class Discourse points possible are 50, with up to 2 bonus points available, for a total of 52.

6 Appendix F: Student Self-Evaluation Of Course

Directions: Answer each of the following questions at the end of the course, and turn in when instructed to do so. Turning a completed form in is worth 10 bonus points to your final grade.

Your name:	
Course:	
Semester:	

1. Did I work as hard as I could have? 1a. Explain your answer to 1. (optional)	Yes ☐ No ☐
2. Did I set and maintain high standards for myself? 2a. Explain your answer to 2. (optional)	Yes ☐ No ☐
3. Did I spend enough time to do quality work? 3a. Explain your answer to 3. (optional)	Yes ☐ No ☐
4. Did I regulate my procrastination, distractions and temptations in order to complete me work? 4a. Explain your answer to 4. (optional)	Yes ☐ No ☐
5. Did I make good use of available resources? 5a. Explain your answer to 5. (optional)	Yes ☐ No ☐
6. Did I ask questions if I needed help? 6a. Explain your answer to 6. (optional)	Yes ☐ No ☐
7. Did I review my work for possible errors? 7a. Explain your answer to 7. (optional)	Yes ☐ No ☐
8. Did I examine best practices for similar work? 8a. Explain your answer to 8. (optional)	Yes ☐ No ☐
9. Is my work something I am proud of and would show to a larger audience? 9a. Explain your answer to 9. (optional)	Yes ☐ No ☐
10. Sign your name, indicating your acknowledgment and agreement to all statements above:	

Questions taken from the article Growth Mindset: Personal Accountability and Reflection by Jackie Gerstein, Ed.D.

7 Appendix G: Student Evaluation of Course

Directions: Answer each of the following questions at the end of the course, and turn in when instructed to do so. Turning a completed form in is worth 10 bonus points to your final grade.

Most of the questions below will ask how well you agree with the statement given. On this scale, "1" is "Strongly Disagree", "2" is "Disagree", "3" is "Agree", and "4" is "Strongly Agree".

One question (number 12), instead of agreement, asks you for an overall rating. This question, while highly desired that you answer, IS optional and will not detract from your bonus point opportunity.

Finally, the final three questions, 13-15, are open-ended and expect short responses. If you need more room than what's given, you can write on the back of the sheet or a blank sheet of paper and turn that in as well.

Ultimately, please be aware that this evaluation is ONLY meant to help me become a better educator. It will not be used for any other purpose.

Your Name:	
Course:	
Semester:	

1. The professor explained concepts clearly.	1	2	3	4
2. The professor clearly articulated the standards of good performance for the course.	1	2	3	4
3. The professor provided guidance for understanding the course exercises.	1	2	3	4
4. The professor increased my understanding of course material.	1	2	3	4
5. The professor helped me achieve my goals.	1	2	3	4
6. The professor was helpful when I had difficulty performing well in the class.	1	2	3	4
7. The professor was helpful to me individually in email, office hours, etc.	1	2	3	4
8. If I mentioned any difficulty at all, the professor was helpful to me in answering my questions.	1	2	3	4
9. The professor provided clear constructive feedback.	1	2	3	4
10. The professor encouraged student questions and participation	1	2	3	4
11. The professor was effective in creating an environment conducive to learning.	1	2	3	4
12. Overall, considering both the limitations and possibilities of the subject matter and the course itself, how would you rate the overall effectiveness of this professor?	1	2	3	4

13. Please identify what you consider to be the strengths of this course.	
14. Please identify area(s) where you think the course could be improved.	
15. What advice would you give to another student who is considering taking this course?	

Questions taken from the UC Berkeley Center for Teaching & Learning "Course Evaluations Question Bank"