Hello Stats classes. I hope you are all healthy + safe.

This is certainly a year you will never, ever forget. Especially you seniors. I truly feel bad for you guys.

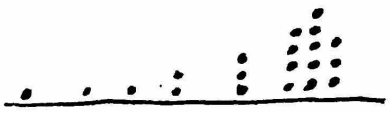
Attached is some exam review. Some of it requires a calculator, which you may not have. Don't worry about it. This is a good chunk of the review we would be doing for the final. So any work you do now will lighten your load later on.

Going forward, I have no idea what the rest of the year looks like at this point. In all honesty, I do hope we get back to school so we can end the year the right way. But until then, stay safe, practice good health habits to ensure others are safe, + do something nice for someone every day. They will appreciate it more than you can imagine.

-Mr. Cercato

Name _____ NUSTEP stats Exam Review

1. Use the dotplot to answer a+b



a) Describe the distribution

b) What is the relationship between the mean + median for this distribution?

2. Which of the following represents a quantitative variable?

a) College major b) Favorite movie genre c) # of credit hours

d) Relationship status e) Number of facebook friends f) Liking Coke or Pepsi

3. The variables above that are not quantitative are called what?

4, what is the main reason statisticians use random samples?

5a) what does a correlation coefficent do?

b) If this value is positive, what does it mean?

c) what if it's negative?

d) what is the range of possible values it can be?

e) Does this value have units?

6. Describe the difference between a Type I + Type II error.

7. If data is statistically significant, what do we do with the null hypothesis?

8. As sample size increases what happens to the mean of the sample? What about the standard deviation of the sample?

9a) What is the formula for finding a z-score?

b) What does the z-score value tell us?

c) If data is normally distributed with a mean of 81 + a standard deviation of 4, find the z-score for a value of 74 + interpret its meaning.

10) The Central Limit Theorem states: "For _____ samples with sufficiently _____ sample size, the distribution of sample statistics for a _____ or a _____ is _____ distributed and _____ at the value of the _____ parameter."

Name _____

AUSTEP Stats Review ~
Outliers

An outlier is

that causes a distribution to be _____ _____ or

_____ _____ .

Outliers have very little effect on the _____ value of a set of data. They do, however, have an effect on the _____ of the data. A large outlier will cause the mean to _____ + thus be _____ than the median. A small outlier will cause the mean to _____ + thus be _____ than the median.

To calculate the "fences", or boundaries, for outliers, use the _____ _____ rule:

1)

2)

Ex: A data set is made up of: 92, 88, 106, 169, 76, 72, 67, 10, 115, 73, 111, 59. Calculate the fences for the outliers + determine if any data is an outlier.

# NUSTEP Stats Boxplot Review

For each of the following problems:

a) Find the sample mean, median, mode, range, interquartile range, and standard deviation

b) Calculate the fences for possible outliers and list any outliers

c) Sketch a boxplot for the data

d) Give a 95% confidence interval for the population mean

1. The lengths, in inches, of 8 bones in the human body are 50.5, 27.5, 36.5, 23, 26, 49, 24, and 40

2. The data shows the population (in millions) of 12 randomly selected US states.

.621 .830 1.262 1.315 2.421 2.901 3.591 4.507 4.602 5.740 8.540 19.281

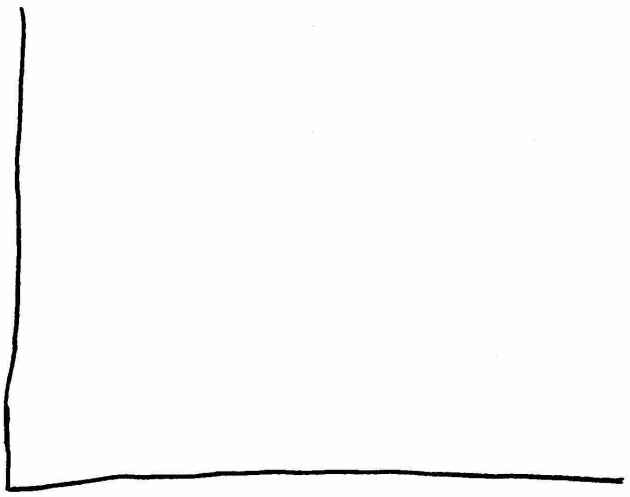
Name _____

AUSTEP Stats Review – Scatterplots

Ex: The table shows minutes studied for an exam + the exam score for 11 randomly selected students.

| Minutes studied (x) | 15 | 25 | 35 | 40 | 45 | 55 | 60 | 65 | 70 | 75 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Exam Score (y) | 50 | 57 | 62 | 68 | 75 | 75 | 83 | 78 | 81 | 86 | 91 |

a) Draw a scatterplot – be sure to label the axes



b) Find the equation of the least squares regression line for this data. Round your values to the nearest hundredth.

c) Add a regression line to your scatterplot.

d) What is the value of the correlation coefficient?

e) Using your least squares regression equation, predict the grade for someone who studied 80 minutes.

MUSTEP Stats Review ~ Calculator Use Name _____

## I. Finding a probability

Ex 1) The distribution of sample proportions of US adults with a college degree for random samples of size $n=500$ is $N(0.275, 0.45)$. What is the probability that a sample will have a proportion $\hat{p}$ that is more than 0.40? You must list the calculator commands you use. Round your answer to the nearest hundredth.

## II. Finding a "score"

Ex 2) Heights of 10 year old boys follow an approximate normal distribution with mean of $\mu = 55.5$ inches & standard deviation of $\sigma = 2.7$ inches. A parent says his 10 year old son is in the 90th percentile for height. To the nearest hundredth of an inch, how tall is this boy? Sketch a drawing of this distribution & label the height you obtain.

Name _____

NUSTEP Stats Review ~
z-scores & confidence intervals

1. Formula for finding a z-score?

2. What does a z-score tell us?

3. What is the purpose of a confidence interval?

Ex: The course project in a computer science class was to create 100 computer games of various levels of difficulty ranging from 1 (easy) to 20 (difficult). A random sample of 10 games produced the following difficulty ratings: 12, 5, 2, 4, 1, 4, 18, 10, 1, 16

a) Find the sample mean rating & standard deviation (round to nearest hundredth).

b) Find a z-score & tell what it means for each of the following ratings:

i) 17

ii) 3

c) Find a 95% confidence interval for this data.

d) Interpret this interval in context of the problem.

MUSTEP Stats Exam Review #s

1. Give 2 examples of categorical variables

2. Give 2 examples of quantitative variables

3. Records from the SPCA were collected to determine what breed of dog is most commonly adopted. What is this an example of?

4. A student conducting a survey asked if people prefer pizza or hamburgers. The student asked 60 people at a pizzeria. What is the bad part of his survey? How could it have been avoided?

5. The amount of variability in data is known as ...? The symbol for this (population) is ...?

6. What is the symbol for testing a hypothesis at the 5% significance level?

7. In order to estimate the average number of pairs of shoes high school students own, a random sample of 50 students was taken & a 90% confidence interval was found to be (3.7, 12.6). What does this mean in context of this problem?

8. Give 2 different bootstrap samples that could be drawn from a sample of 4, 7, 8, 9, 14.

9. When conducting a study, our conclusion deals with the _____ hypothesis we can either _____ it or not. When we reject it, we say the data is ....?

10. Explain what $N(47,3)$ means.

11. Two sets of data are given:

① 74, 77, 80, 83, 91 ② 15, 77, 80, 83, 91

a) Find the mean + median of each set of data (nearest tenth)

b) which set of data appears to have an outlier?

c) what value from (a) is most affected by the outlier?

---

Part II

12. A new generic drug is being used to treat high cholesterol. Insurance records show the average copayment by a patient for a month's supply is $24.75. A sample of 75 users of this drug found their average copayment to be $16.25.

a) Identify $24.75 and $16.25 as either a parameter or statistic + give the proper notation for each.

b) If we took 1000 random samples of size 75 from the population of all copayments for this drug + plotted these sample means on a dotplot, describe the shape of the plot + where its center would be.

13. Weights of 6 month old boys follow a normal distribution with $\mu = 17.3$ lbs and $\sigma = 3.2$ lbs. A parent says his 6 month old boy is in the 80th percentile for weight. How much does this boy weigh? Sketch a drawing of the distribution + label the height you obtain.

4) A sample of scores from Natural Disasters:

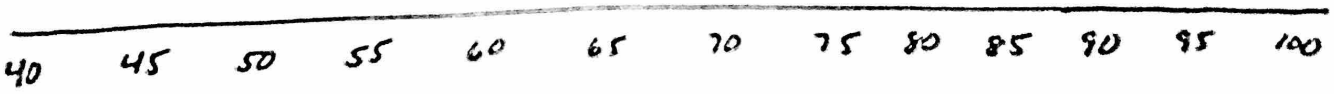
41, 64, 71, 78, 78, 81, 81, 81, 81, 86, 86, 94, 98

a) Give each of the following rounded to the nearest hundredth if necessary:

| | |
|---|---|
| mean | range |
| median | interquartile range |
| mode | standard deviation |

b) Find the upper + lower "fences" for outliers + identify any outliers.

c) Draw a boxplot




d) Give the 95% confidence interval for the population mean.

Two studies are described below. Use the descriptions to answer the questions that follow.

**Study 1:** Currently, the NFL will not televise home games in the local market if the game does not sell out 72 hours before kick-off. This is known as the Blackout Rule. However, there is talk of changing this rule. As part of the evidence, the league will find the average seating capacity of all 32 of their stadiums.

**Study 2:** It is said that the color of a room can increase your appetite, specifically that red rooms cause an increase in eating habits. To test this, 120 high school students are randomly assigned to one of two groups. The first group is placed in a red room and the second group is placed in a white room. They are kept there for a period of 3 hours without food. After 3 hours, a buffet is brought into each room and the amount of food consumed by each person is recorded.

(a) Which study is an experiment?

(b) Identify the explanatory and response variables in this experiment.

(c) Identify your variables from part (b) above as either quantitative or categorical.

(d) What type of study is the one that is not an experiment?

NUSTEP Stats Review #4

1. A sample of data is found to be normally distributed with a mean of 73 & a standard deviation of 6. This means that about 95% of the data falls between... ?

2. a) p-value = .08, $\alpha = 5\%$. which hypothesis is supported & why?

   b) p-value = .03, $\alpha = 5\%$. same questions ↑

3. If we don't use random sampling, what can we say about our results?

4. Does a correlation measure a cause/effect relationship or does it measure a linear relationship between 2 quantitative variables?

5. If we reject a true null hypothesis it is a _____ error. If we fail to reject a false null, it is a _____ error.

6. Using a sample size of 50, we find in a certain study that $\bar{x} = 127$ and $S_x = 14$. what happens to $\bar{x}$ and $S_x$ if we use a sample size of 200?

7. Test scores are normally distributed with a mean of 79 & a standard deviation of 5. Find the z-score for a test score of 71 & tell what this means.

... with a mean of .

8. What is the name of the theorem that says "when random samples are sufficiently large, the distribution of sample statistics is normal & centered at the population parameter"?

---

Part II

9. Approximately 15% of Americans are red heads (treat this as a known population parameter). A study on the relationship between hair color & profession found that in a random sample of 115 teachers, 24 of them were red heads. Test the hypothesis that the proportion of red headed teachers differs from the proportion of red headed Americans.

a) State the null & alternative hypotheses.

b) Calculate the test statistic & p-value

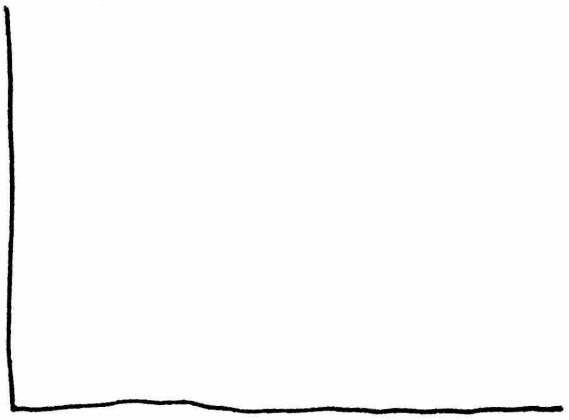
c) What do we conclude at a 1% level of significance?

10. The distribution of sample proportions of US adults with a high school diploma for random samples of size $n = 350$ is $N(0.418, 0.129)$. What is the probability that a sample will have a proportion $\hat{p}$ that is more than 0.56? You must list the calculator commands used to get full credit.

A teacher compiles the following data from 12 students regarding # of absences + their grade:

| absences (x) | 1 | 3 | 3 | 5 | 6 | 10 | 0 | 5 | 4 | 2 | 5 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| grade (y) | 94 | 88 | 84 | 83 | 78 | 74 | 93 | 86 | 83 | 89 | 80 | 77 |

a) Draw a scatterplot for the data (Be sure to label the axes)



b) What is the equation for the least squares regression line? (Nearest hundredth)

c) Add a regression line to your plot.

d) What is the correlation coefficient? (nearest hundredth)

12. Geologists calculated the depth of 47 rivers in the US. The found the average depth to be be 18.4 feet with a standard deviation of 8.3 feet. Test to see if this sample provides evidence that the average depth of all US rivers is greater than 15.5 feet. Show all details of your test.

AUSTEP STATS REVIEW #5

1. Identify each as a categorical or quantitive variable:
   Vehicle Type  Credit Score  Shoe Size  Favorite Sport

2. Records of student lunch purchases were used to determine if students prefer white or chocolate milk. Is this an experiment or obs. study? Why?

3. Fill in the blank w/ the missing symbol or meaning:
   sample mean _____  $S_x$ _____
   population mean _____  correlation coefficient _____
   $\sigma$ _____  $\alpha$ _____

4. A 95% confidence interval for a particular SAT test is (920, 1550). What does this mean?

5. Is 4,6,8,10,10 a good bootstrap from a sample of 4,6,8,9,10? Explain

6. If we obtain a result we are not expecting then the data is _____ _____ & we should _____ the null h

7. What would change the most due to the appearance of an outlier?
   mode  proportion  median  mean

8. How would we note data that is symmetric, bell shaped, has a mean of 80, & a standard deviation of 6?

---

Part II

9. Exam scores for a particular standardized test follow a normal distri with $\mu = 74.2$ and $\sigma = 7.8$. A student claims to have scored in the 90th percentile. What is his grade (to the nearest tenth). Sk a drawing of the distribution & label the score you obtain.

10. A sample of online purchases from Amazon.com has the following values (rounded to the nearest dollar):

34, 36, 38, 40, 40, 40, 41, 41, 45, 50, 50, 98

a) Give the sample mean, median, mode, range, IQR, + std. dev. Rnd. to the nearest hundredth if necessary.

b) Find the upper + lower fences for outliers + identify any.

c) Draw a boxplot. Indicate any outliers.

30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

d) Give a 95% confidence interval for the population mean.

11. Studies are finding stomach bacteria is essential for good health. A study of 212 subjects compared the number of unique bacterial genes in the stomachs of healthy patients + those with IBS. The study found that healthy subjects had significantly more unique bacterial genes.

a) Experiment or Obs. Study? Explain

b) List all variables in the description + identify as categorical or quant.

c) Which var's are explanatory + which are response?

d) State appropriate null + alt. hypotheses.