Math 247: Test 1 (Wright, Spring 2020) **Name:** KEY

In-class test _____/70 points *Take home test* _____/30 points

1. (5 pts) Exercise and Happiness. Suppose you gather data from your classmates by asking how many hours they exercise per week and whether or not they are generally happy.

   List the variables and state whether each is categorical or numerical.

   Hours of exercise: numerical
   Happy (yes, no)?: categorical

   If you found that people who exercise more are generally happier, could you conclude that exercise CAUSES people to be happier? (Circle your answer.)

   YES (NO) — Note: This is not a controlled experiment. Observational studies do not establish cause-and-effect.

2. (5 pts) Write what each of the following symbols stands for: $\bar{x}$, x, n, s, $\sum$

   $\bar{x}$ = Sample mean
   x = data value (numerical)
   n = sample size (number of data values)
   s = Sample Standard deviation
   $\sum$ => sum (add values up)

3. (5 pts) Vitamin C and Allergies. A study done by the Mayo clinic examined whether breast-feeding mothers who were choosing to take large doses of Vitamin C had children who had a lower chance of having allergies. They found there was a 30% lower risk of children having allergies for the Vitamin C moms, as compared the children whose mothers were not choosing to take Vitamin C.

   What was the research question for the study? — Careful with verbs here "lowers" implies causation!
   Is there a link (an association) between mothers taking large doses of Vitamin C and allergies in children who are breast feeding.

   This study is (circle one) a randomized, controlled EXPERIMENT (an OBSERVATIONAL STUDY)

   Can we conclude from this study that taking large doses of Vitamin C actually caused the reduction in allergies? Briefly explain.

   No! Observational studies do not establish cause-and-effect due to possible confounders.

   Describe one potential confounder for this study. (* Note: A confounder must link the two variables to explain the effect seen.)

   Mothers who are choosing to take the Vitamin C may be taking other supplements and/or also may have a healthier lifestyle that impacts their kids' allergies (including not smoking?)

4. (9 pts) A researcher is interested in the effect of music on memory. She takes a random sample of 60 Cuesta College students then randomly assigns the students to one of three groups: those who will listen to quiet music, those who will listen to loud music, and those who will not listen to music. Each group works on a memorization task while listening to music (or not) and then takes a memory test.

What was the research question for the study?

Does music have an effect on memory formation and, if so, what type of music?

What is the population of interest?

Cuesta College students (college students in general)

What is the sample and sample size?

Sample: $n = 60$ Cuesta College students.

This study is (circle one) (a randomized, controlled EXPERIMENT) an OBSERVATIONAL STUDY

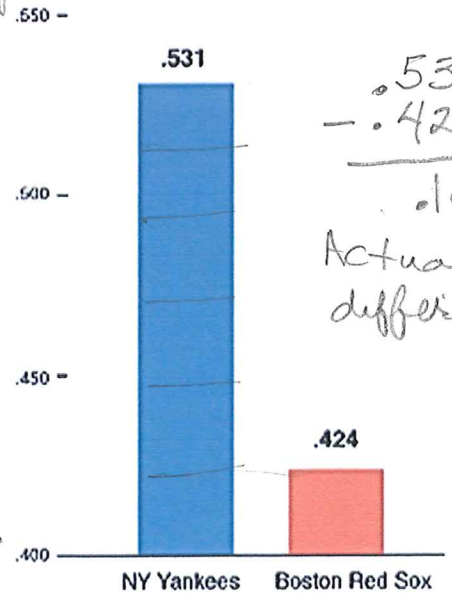
What is the treatment variable? Music

What is the response variable? Memory test results.

If the researcher found that students who listened to quiet music performed significantly better on the memory task, could she conclude that the music CAUSED the better memory results? Explain.

Yes, She conducted a controlled experiment and by randomly assigning students to groups, potential confounders are distributed through the groups, and not concentrated in one group.

5. (2 pts) The given graph shows the percentage of games won by the New York Yankees and the Boston Red Sox for one season. Explain why this graph is misleading.



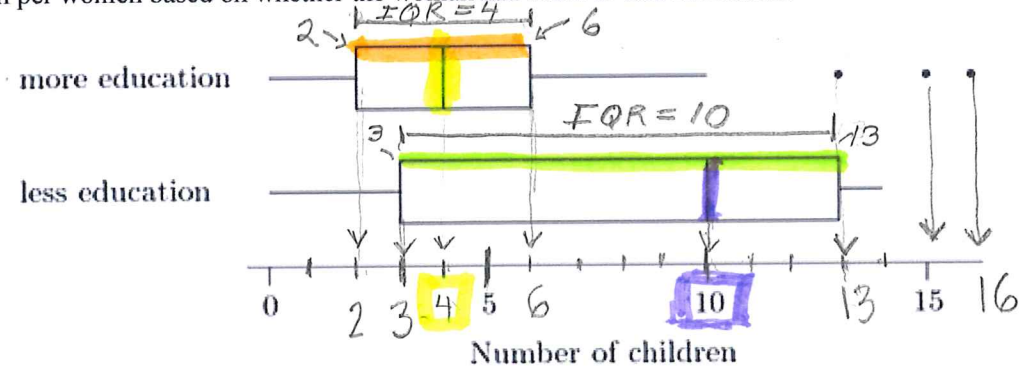
.531
−.424
.107
Actual difference

The graph is misleading because the y-axis does NOT begin at zero!

Not zero!

The difference in percent of games won looks HUGE (like 4x's!) but really the difference is only .107.

6. (10 pts) Women in rural, developing nations typically have many children. The Gates Foundation investigated the relationship between education of women and birthrate. The boxplots below show a graph of the number of children per women based on whether the woman had more or less education..



What does this graph suggest about the relationship between education of women and birthrate?

It suggests that women with more education typically have fewer children than those with less education.

Which group has outliers in the data and what are the approximate value of the outliers?

Women with more education. Three woman had unusually large numbers of children (13, 15, and 16 kids)

Excluding the outliers, which data set shows the most skewing?

Women with less education (skewed left)

Estimate the Median for both groups:

Median number of children for women with more education = 4 kids

Median number of children for women with less education = 10 kids

Estimate the IQR for both groups: $IQR = Q_3 - Q_1$

IQR for women with more education = $6 - 2 = 4$ kids

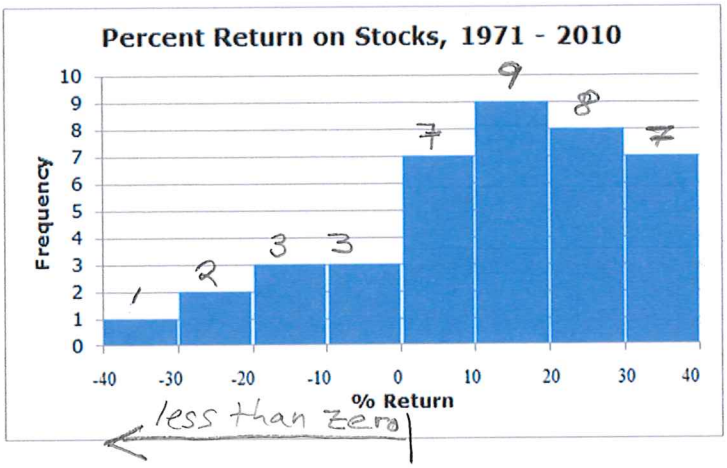
IQR for women with less education = $13 - 3 = 10$ kids

These being equal to the medians is totally a coincidence!

Excluding the outliers, which group's data showed the most variability?

less education group.

7. (10 pts) The percent return on stock for the NASDAQ from 1971 to 2010 is shown in the histogram.



Use the histogram to answer the following questions.

2 (a) How many stocks are were in this study?

40

2 (b) How many stocks had a percent return that was negative (less than 0)? 9

1 (c) What is the relative frequency (express as a fraction and a decimal) of the stocks that had a negative return?

9/40 = .225

2 (d) What is the shape of the distribution? left skewed

2 (e) Estimate the median value for the distribution: ≈ 15% (between 10 and 20%)

1 (f) Based on the shape we know what about the mean relative to the median?

(i) The mean is greater than the median

(ii) The mean is less than the median

(iii) The mean is exactly equal to the median

(iv) Can't tell.

8. (2 pts) Which of the histograms show the larger amount of variability in the data? Machine 2 (data is more spread out relative to the center)

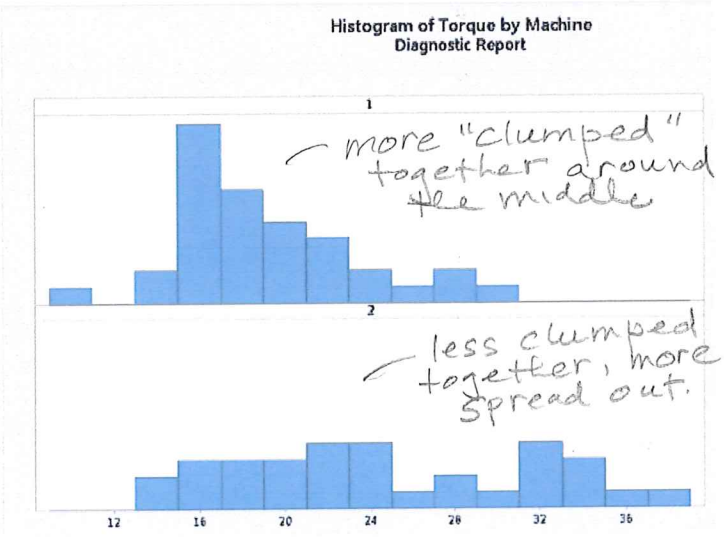
Just by looking at the histograms, compare the standard deviations of the two sets of data. (Circle the correct answer).

(a) SD for Machine 1 is greater than SD for Machine 2.

(b) SD for Machine 1 is less than SD for Machine 2.

(c) SD for Machine 1 is the same as SD for Machine 2

(d) Can't tell.



9. (12 pts) A random sample of 5 students were asked how many times they check their social media per day.

Their responses were 0, 9, 10, 16, 100

(a) Construct a dot plot of this data. (graphs will vary)

(a) Find the mean of the data and mark it with a triangle on the dotplot. Then find the median.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{135}{5} = 27$$

Med = 10

(b) Which is a more "typical value" for this data set, the mean or the median? MEAN (MEDIAN)

(c) How did the outlier affect the mean? Pulled it up, away from the center

(d) Because of this effect, we say that the mean is not "resistant"

4 (e) By hand, find the standard deviation of the data.

x	$x-\bar{x}$	$(x-\bar{x})^2$
0	−27	729
9	−18	324
10	−17	289
16	−11	121
100	73	5329
✓: $\Sigma(x-\bar{x}) \stackrel{?}{=} 0$ yes!		$\Sigma(x-\bar{x})^2 = 6792$

SD: $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

$= \sqrt{\frac{6792}{4}}$

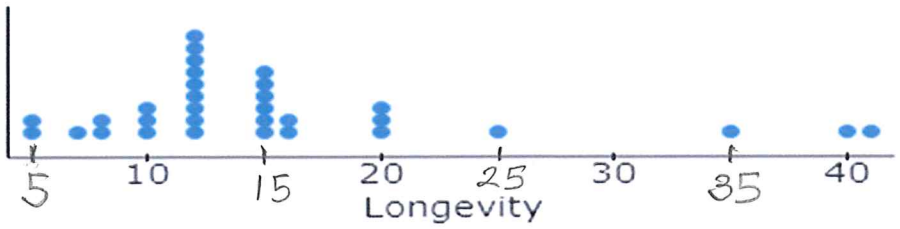
$= \sqrt{1698}$

$s = 41.2$ texts

(f) How did the outlier affect the standard deviation? Made it HUGE!

(g) Is the standard deviation resistant? No, the outlier drastically impacted it.

10. (10 pts) The lifespan (in years) for a number of different mammals in San Luis Obispo is graphed below, with the summary statistics shown below that. Use the graph and the summary stats to answer the questions.



Summary statistics:

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3
Longevity	32	15.4	77.09	8.78	1.55	12	36	5	41	11	16

(a) How many data values are there? $n = 32$

(b) How many mammals had a life span over 30 years? 3 mammals

(c) What proportion (relative frequency) of mammals had a lifespan over 30 years? $\frac{3}{32} = .094$

(d) Which would it be more appropriate to describe the center and variation of this data set: (circle one)

the mean and standard deviation

the median and IQR

Why? The data has outliers which would make the mean and SD not representative

Note: excluding the outliers, the data is pretty symmetric

(e) What is the five-number summary for this data set? 5, 11, 12, 16, 41 years

(f) Find the IQR.

$IQR = Q_3 - Q_1$
$= 16 - 11 = 5$ years

(g) Find the Lower Outlier and Upper Outlier Limits.

Lower $= Q_1 - 1.5\,IQR = 11 - 1.5(5) = 3.5$ years

Upper $= Q_3 + 1.5\,IQR = 16 + 1.5(5) = 23.5$ years

(h) Is the data value of 35 years an outlier? Explain how you can tell based on the Outlier Limits you found in (g).

Yes! Since 35 is over the upper limit of 23.5, we know this is an outlier.
(So is 25 years, surprisingly!)