Units!

Math 247: Test 1 (Wright-Spring 19) **Name:** KEY

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

1. (2 pts)Suppose you gather data from your classmates by asking their eye color and age.

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

   Eye color: Categorical
   Age: numerical

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

   $\bar{x}$ = Sample mean
   x = data values
   n = Sample size
   s = sample standard deviation
   $\sum$ = Sigma $\Rightarrow$ sum of

3. (9 pts) The September 2011 issue of the "Berkeley Wellness Letter" said that coffee reduces the chance of prostate cancer. A study of 48,0000 male health care professionals showed that those consuming the most coffee (six or more cups per day) had a 60% reduced risk of developing advanced prostate cancer.

   What was the research question for the study?
   Is there a link between coffee drinking and prostate cancer?

   What was the sample for the study?
   48,000 male health care professionals (HUGE Sample!)

   What is the (implied) population?
   Male health care professionals
   or even all men.

   This study is (circle one) A RANDOMIZED EXPERIMENT (AN OBSERVATIONAL STUDY)

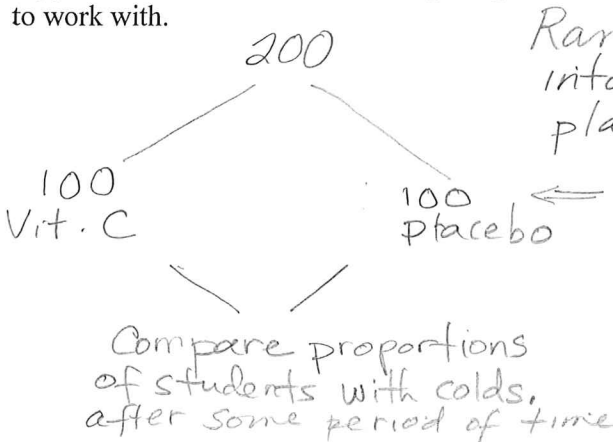
   What are the variables in this study? Coffee drinking and prostate cancer status

   Which of these variables is the Treatment (Factor)? Coffee drinking (number of cups)

   Which of these variable is the Outcome (Response)? Prostate cancer status

   Was the reporter correct in saying coffee REDUCES* the chance of prostate cancer (*which implies drinking more coffee CAUSED the reduction in prostate cancer)? Briefly explain your answer:
   No. This is an observational study so we can't conclude cause-and-effect.

4. (4 pts) (a) Briefly describe the design of a controlled experiment to determine whether the use of vitamin C supplements reduces the chance of getting a cold for college students. Assume you have 200 college students to work with.



Randomly assign the 200 students into a Vitamin C group and a placebo group.

Double blind: Neither the subjects nor the researchers should know who is taking what.

(b) If you found that the students who took vitamin got fewer colds in your experiment, would it be correct to state that vitamin C CAUSED the students to get fewer colds? Assume you've done a perfect experiment! Briefly explain your answer.

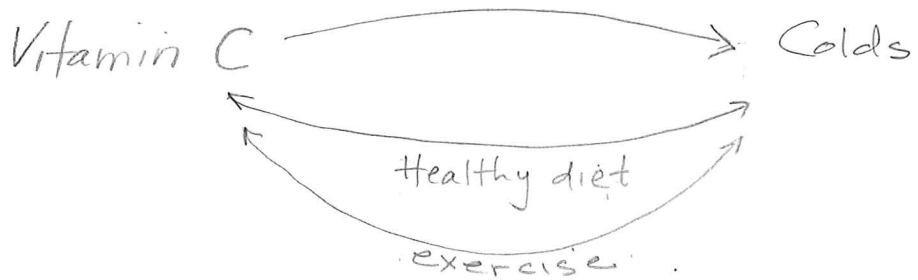
Yes, since this is a controlled experiment, we can conclude cause-and-effect. Any confounders (like healthier lifestyles) would be distributed evenly between the groups by random assignment.

5. (3 pts) Suppose instead of designing an experiment about vitamin C and colds, you find 100 students who don't take vitamin C and 100 students who do take vitamin C are compare whether or not they get a cold over a 6-week period. You find that those who do take vitamin C get fewer colds.

1 Would it be correct to state that your study shows that vitamin C CAUSES people to get fewer colds? Why or why not?

No, it's an observational study so we can't conclude cause-and-effect.

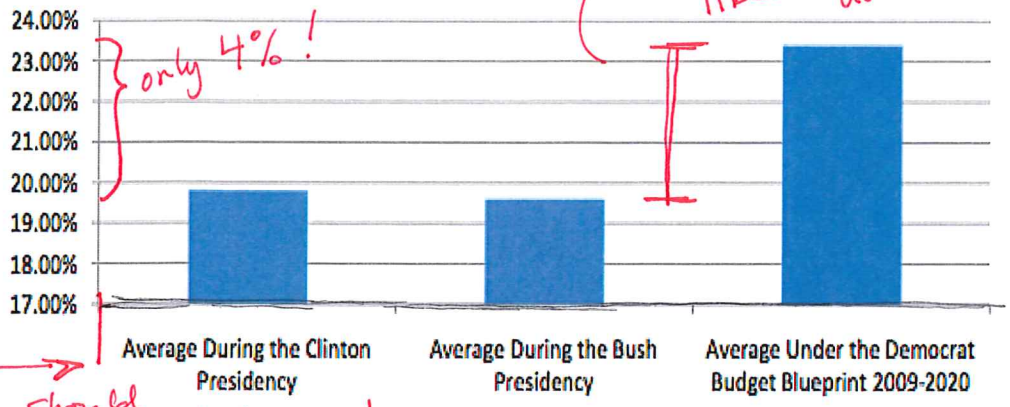
2 Describe one potential confounder in this situation. Describe how the confounder ties the Treatment variable to the Response variable.



If people are already choosing to take Vitamin C, they may also have a generally healthier lifestyle, (no smoking, good diet, more exercise) so these factors would connect the Treatment (taking Vit C) to the Response (incidence of colds) and affect the results.

6. (3 pts) The given graph shows Federal spending as a share of the economy.

   Explain why this graph is deceptive. (Write NOTHING about politics, strictly make an observation about how the graph is set up!)

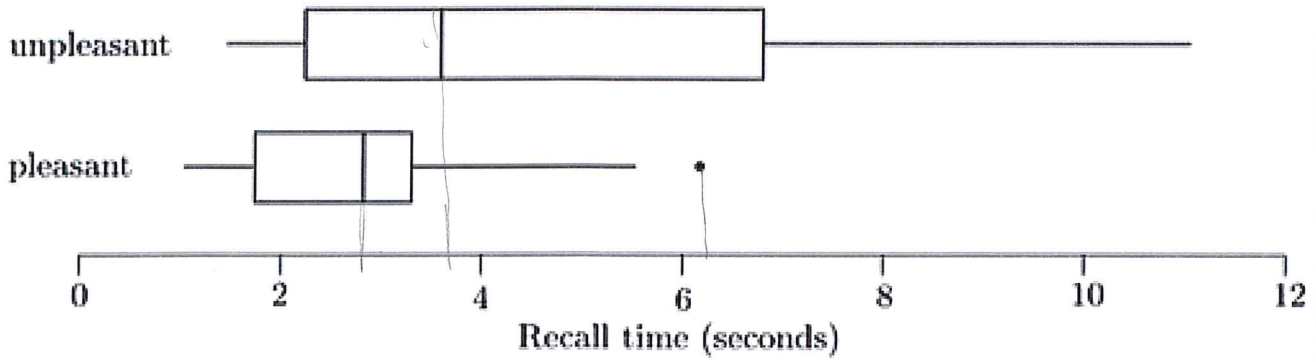


The y-axis doesn't begin at ZERO so the differences in bar height are exaggerated.

should begin at zero!

7. (5 pts) **Memory recall times** In a study of memory recall times, a series of words was shown to a subject on a computer screen. For each word, the subject was instructed to recall either a pleasant or an unpleasant memory associated with that word. (Example: word = "ocean"; round 1, recall a pleasant memory; round 2, recall an unpleasant memory).

   When the subject was able to recall a memory, they pressed a bar on the computer keyboard. The boxplots below show the recall times (in seconds) for twenty pleasant memories and for twenty unpleasant memories.



Estimate the median for both groups:

Median time for unpleasant memory = 3.7 s (answers will vary)

Median time for pleasant memory = 2.9 s (answers will vary)

Based on these graphs, did subjects typically have an easier or harder time recalling an unpleasant memory?

Harder, since it took longer to recall it.

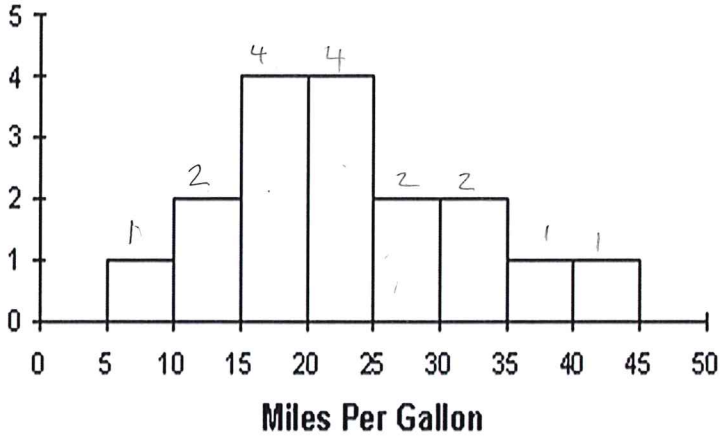
Which set of recall times (type of memory) showed the most variability?

Unpleasant memories showed the most variability

Which data set has an outlier and what is the approximate value of the outlier?

Pleasant memories; 6.2 seconds (answers will vary)

8. (6 pts) The distribution of gas mileage (mpg) for the top selling cars in 2015 are shown in the graph.



Use the histogram to answer the following questions.

(a) How many cars were in this study?

17 cars

(b) How many cars had a gas mileage under 15 mpg? 3

(c) What is the relative frequency (express as a fraction and a decimal) of the cars that had a gas mileage under 15 mpg? 3/17 = .176

(d) What is the shape of the distribution? Slightly right skewed

(e) Estimate the median value for the distribution: 22.5 mpg

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

Right Skew

(i) The mean is greater than the median

(ii) The mean is less than the median

Answer should be consistent with (d)

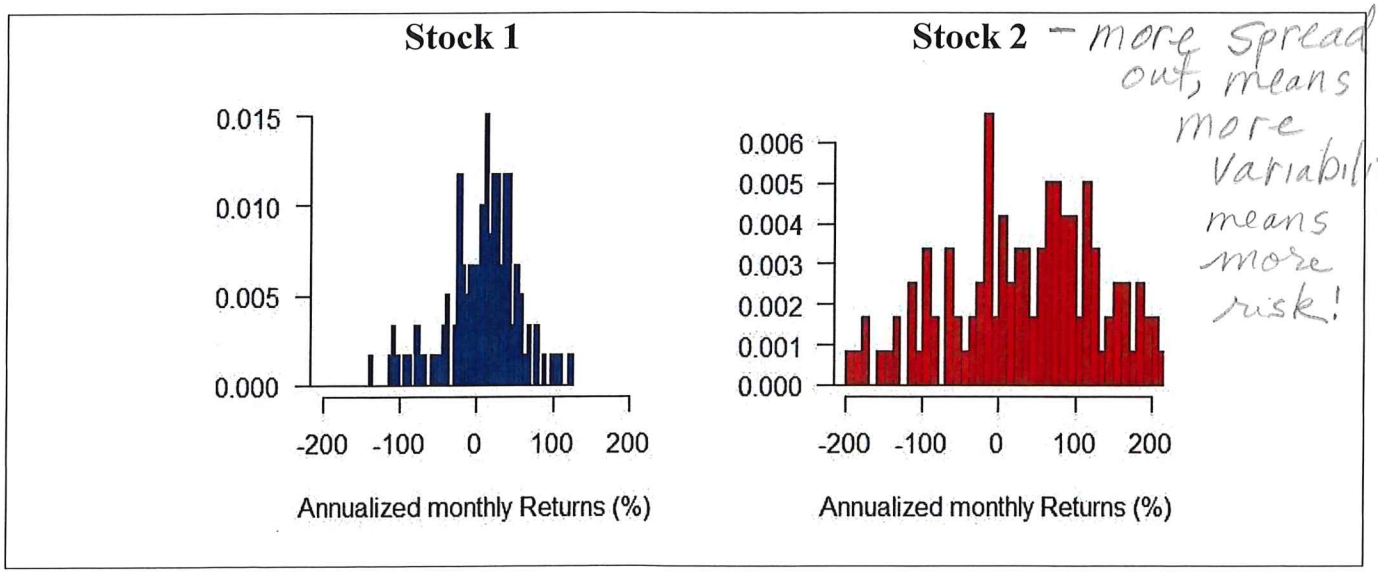
(iii) The mean is exactly equal to the median

(iv) Can't tell.

only in the case of perfect symmetry!

9. (2 pts) The two histograms below show the monthly returns (interest) for stocks for the S and for GM over a 10 year period. Variability is a measure of risk in stock investment (more variability means more risk). Which was the riskier stock?

(a) Stock 1 is riskier (b) Stock 2 is riskier (c) They have equal risk (d) Can't tell

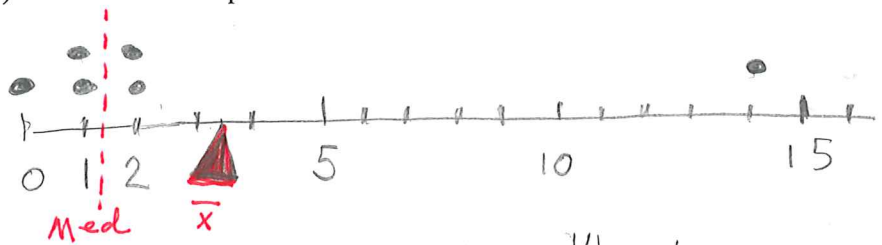


more spread out, means more variability, means more risk!

10. (18 pts) A random sample of 6 students were asked how many pets they have.
Their responses were 2, 0, 1, 14, 1, 2

0, 1, 1, 2, 2, 14

1 (a) Construct a dot plot of this data. $n = 6$



1 (b) What data value seems to be an outlier? 14 pets

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

$$\bar{x} = \frac{\Sigma x}{n} = \frac{20}{6} = 3.333 \text{ pets}$$

$$\text{Med} = 1.5 \text{ pets}$$

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

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

1 (f) Because of this effect, we say that the mean is not resistant

8 (g) By hand, find the standard deviation of the data.

x	$x - \bar{x}$	$(x-\bar{x})^2$
0	−3.333	11.109
1	−2.333	5.443
1	−2.333	5.443
2	−1.333	1.777
2	−1.333	1.777
14	10.667	113.785 ← big impact wow!

outlier →

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

$$= \sqrt{\frac{139.334}{6-1}}$$

$s = 5.279$ pets

$\Sigma x - \bar{x} \approx 0$ (not exact due to rounding!)

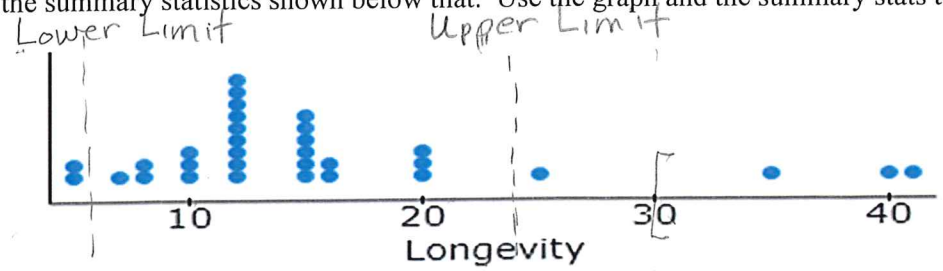
$\Sigma (x-\bar{x})^2 = 139.334$

1 (h) How did the outlier affect the standard deviation? Made it much larger!

1 (i) Is the standard deviation resistant? No — the outlier had a big effect on the value of the S.D.!

whoops! typo

11. (13 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

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

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

1 (c) What proportion (relative frequency) of mammals had a lifespan over 30 years? $\frac{3}{32} = .094 = 9.4\%$

2 (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 is skewed to the right! Also there appears to be outliers

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

1 (f) Find the IQR. $IQR = Q_3 - Q_1$
$= 16 - 11$
$= 5$ years

4 (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

1 (h) Is the data value of 25 years an outlier? Explain how you can tell based on the Outlier Limits you found in (g).

Yes, 25 is a potential outlier since it is beyond the fence (i.e. it's greater than 23.5)