MATH 338

MIDTERM 1

THURSDAY, SEPTEMBER 29, 2016

Your name: _____

Your scores (to be filled in by Dr. Wynne):

Problem 1: ____/12

Problem 2: ____/12

Problem 3: ____/8

Problem 4: ____/11

Problem 5: ____/12

Total: ____/55

You have 75 minutes to complete this exam. This exam is closed book and closed notes with the exception of your formula sheet.

For full credit, show all work except for final numerical calculations (which can be done using a scientific calculator).

Problem 1. There are 50 words used in the book *Green Eggs and Ham*. The table shows their lengths:

Word Length (Letters)	1	2	3	4	5	8
Frequency	2	9	16	14	8	1

A) [4 pt] Find the five number summary of the lengths of the words in *Green Eggs and Ham*. (Hint: if you order all the numbers from smallest to largest, in what positions will Q1, Med, and Q3 be?)

Min: _____ 1

Q1: _____ 3

Med: _____ 3

Q3: _____ 4

Max: _____ 8

0.5 pts each for min and max, 1 pt each for Q1/Med/Q3

B) [1 pt] Find the mean of the lengths of the words in *Green Eggs and Ham*.

Mean = ((1)(2) + (2)(9) + (3)(16) + (4)(14) + (5)(8) + (8)(1))/50 = 3.44

Part C: Values are parameters since this is a population

C) [1 pt] The values you calculated in parts (A) and (B) are (circle one): statistics parameters

D) [3 pt] Sketch a relative frequency histogram of the word lengths. Label your axes.

0.5 pt shape of the histogram is roughly correct

1 pt bars are connected except for a space out to 8 OR separated as in binomial histogram from Lab 5

1 pt Y-Axis correctly labeled with Relative Frequency and axis limits are reasonable (up to 0.4 or so)

0.5 pt X-Axis correctly labeled with Word Length or equivalent and axis limits are reasonable

E) [1 pt] The distribution of word lengths is (circle the best answer):

symmetric skewed right skewed left

1 pt for skewed right, 0.5 pt for symmetric (looks roughly symmetric except for the outlier)

F) [2 pt] Is the eight-letter word an outlier? Justify your answer graphically and numerically.

1 pt showing gap in histogram, 1 pt showing (Q3 + 1.5*IQR) = 5.5 and 8 > 5.5

Problem 2. The "Four C's" used to evaluate a diamond summarize several variables including: its color, the number of blemishes on its surface, the total amount of light it reflects, and its weight.

A) [2 pt] Classify each variable above as categorical or quantitative (numerical).

Categorical: color

Quantitative: number of blemishes, light reflected, weight

0.5 pt per variable

B) [1 pt] The distribution of the number of blemishes is most likely (circle the best answer):

uniform normal binomial none of these

none of these – it's a discrete variable that's not in the binomial setting

C) [4 pt] Diamond cutters prefer to meet or exceed particular cutoff weights. Suppose that a diamond cutter plans to cut "1 carat" diamonds so that their weight is normally distributed with mean 1.033 carats and standard deviation 0.02 carats. What proportion of diamonds will weigh less than 1 carat?

z-score = (1 – 1.033)/0.02 = -1.65

P(X < 1 carat) = P(Z < -1.65) = 0.0495 – about 5% of diamonds will weigh less than 1 carat

3 pt using z-score framework and converting to z-score correctly

1 pt getting final answer correctly from table

D) [5 pt] Suppose that instead the diamond cutter wants only 0.1% (0.001) of 1 carat diamonds to fall below 1 carat. What must his <u>new</u> target mean weight be, if the standard deviation is the same? Round your answer to two decimal places.

New z-score for X = 1 is -3. So -3 = (1 – mu)/0.02, and mu = 1.06

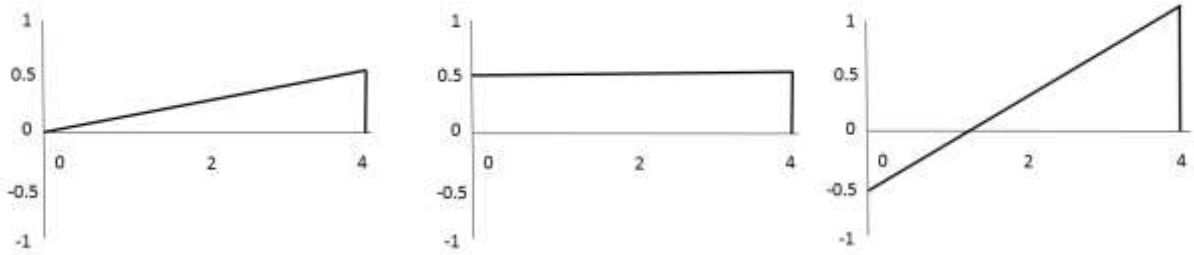
1 pt working in inverse normal distribution framework and reading z-score of -3 from table

3 pt using z-score framework and plugging in correctly for z, x, and sigma

1 pt solving correctly for mu to get final answer

Problem 3. Answer the miscellaneous questions about density curves and study design.

A) [1 pt] Which of the graphs below could depict the density curve of a real distribution (circle one)?



Left graph

B) [2 pt] For each of the two graphs you did not circle, explain why they cannot be density curves.

Center graph: area under the density curve is equal to 2

Right graph: density curve goes below the x-axis

1 pt each

C) [1 pt] True or False: A single observational study can prove a causal relationship between two variables? (circle True or False above)

False

A study wants to investigate the effect of a new drug on reducing high blood pressure. The study plans to record the blood pressure of subjects before they start the drug and after six months of taking the medication. One group will get the new drug and another group will get a drug currently available on the market.

D) [2 pt] Which of the following are potential confounding variables (circle at least one answer)?

Blood pressure Drug type Weight Age Favorite sports team

1 pt each Weight and Age, -1 pt for each other answer

E) [2 pt] Why do the researchers need a control group? Why does the control group get a currently available medication, instead of getting no treatment at all?

1 pt need control group to show that blood pressure reduction with the drug is greater than if no drug was given

1 pt need currently available medication to counter placebo effect OR for ethical reasons

Problem 4. The table shows the number of new Ebola cases in three different countries during the West African Ebola outbreak of 2014-15.

Year	Guinea	Liberia	Sierra Leone	Total
2014	2707	8018	9446	20171
2015	1097	2657	4656	8410
Total	3804	10675	14102	28581

An Ebola sufferer is selected at random from the 28,581 in the table.

A) [2 pt] What is the probability that this person is from Sierra Leone?

P(Sierra Leone) = 14102/28581 = 0.493

1 pt for correct setup, 1 pt for answer

B) [2 pt] What is the probability that this person is from Sierra Leone, given that this person contracted Ebola in 2015?

P(Sierra Leone | 2015) = 4656/8410 = 0.554

1 pt for correct setup, 1 pt for answer

C) [2 pt] Are the events "this person contracted Ebola in Sierra Leone" and "`this person contracted Ebola in 2015" independent? Show why or why not.

1 pt NO

1 pt P(Sierra Leone | 2015) is not equal to the unconditional probability, P(Sierra Leone)

Or P(2015 | Sierra Leone) is not equal to the unconditional probability, P(2015)

Or P(Sierra Leone and 2015) is not equal to P(Sierra Leone) times P(2015)

D) [5 pt] Suppose that the true incidence rate of Ebola in Sierra Leone was 0.2%. A test reads positive for Ebola in 99% of cases with Ebola, but also in 5% of cases without Ebola. A man from Sierra Leone tests positive for Ebola. Find the probability that he truly does have Ebola.

Using Bayes's Rule:

$$P(Ebola\,|test+) = \frac{P(test+|Ebola)\,P(Ebola)}{P(test+|Ebola)P(Ebola) + P(test+|no\ Ebola)P(no\ Ebola)}$$

$$P(Ebola\,|test+) = \frac{(0.99)(0.002)}{(0.99)(0.002) + (0.05)(0.998)} = 0.038$$

1 pt use Bayes's Rule and/or tree diagram, 3 pts plug in correctly, 1 pt final answer

Problem 5. Assume that, after today, we will have 34 more lecture and lab sessions this semester. Suppose that before each session, I generate a random number from a distribution described by a uniform density curve between 0 and 1. If the number is less than 0.1, I take attendance in that section, and otherwise I don't.

A) [1 pt] What is the probability that I take attendance at your next lecture period?

$P(X < 0.1) = 0.1$

B) [1 pt] The number of times I take attendance across the 34 class periods has which type of distribution (circle the best answer)?

uniform normal binomial none of these

Binomial – we are in the binomial setting here

C) [2 pt] Find the expected number of times that I take attendance over the 34 class periods.

1 pt $E(X) = np$

1 pt plug in $E(X) = (34)(0.1) = 3.4$ times

It can be shown that if X is a random variable described by a uniform distribution between 0 and 1, then $E(X) = 1/2$ and $Var(X) = 1/12$. Suppose that I generate 100 random numbers from this distribution.

D) [1 pt] The distribution of the sample mean of the 100 random numbers is (circle the best answer):

uniform normal binomial none of these

normal by Central Limit Theorem

E) [3 pt] Find the mean and standard deviation of the sample mean of the 100 random numbers.

1 pt use Central Limit Theorem to find mean and sd

1 pt mean of sample mean = mean of population = 1/2 = 0.5

1 pt sd of sample mean = sd of population/sqrt(n) = sqrt(1/12)/sqrt(100) = sqrt(1/1200) = 0.029

F) [4 pt] Use your answers to parts (D) and (E) to find the probability that I get a sample mean greater than 0.537.

3 pt work in cumulative normal framework: $P(\bar{x} > 0.537) = P(Z > z)$, $z = (0.537 - 0.5)/(0.029) = 1.28$

1 pt $P(\bar{x} > 0.537) = P(Z > 1.28) = 0.1003$ or about 10%

Extra Space. The tables below show a number of critical values *z* for the standard normal variable $Z \sim N(0, 1)$ and the corresponding cumulative proportions, corresponding to P(Z ≤ z).

z-score	Cumulative Proportion
-3.00	0.0013
-2.00	0.0228
-1.65	0.0495
-1.28	0.1003
-1.00	0.1587
-0.43	0.3336

z-score	Cumulative Proportion
0.43	0.6664
1.00	0.8413
1.28	0.8997
1.65	0.9505
2.00	0.9772
3.00	0.9987

The rest of this space to be used for extra work: