MATH 338

MIDTERM 1

WEDNESDAY, FEBRUARY 28, 2018

Your name: ______________________

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

Problem 1:	____/11
Problem 2:	____/8.5
Problem 3:	____/15.5
Problem 4:	____/10
Problem 5:	____/10
Total:	____/55

You have 75 minutes to complete this exam.

You may refer to your (single-sided, prepared in advance) formula sheet. You may ask Dr. Wynne to clarify what a question is asking for. You may not ask other people for help or use any other resources.

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

1. The sex and body weight (in kg) of a sample of 144 cats was recorded. The table below shows some summary statistics. Use the table to answer the following questions:

	Female Cats	Male Cats
Count	47	97
Minimum Weight	2.00	2.00
Q1 (Weight)	2.15	2.50
Median Weight	2.30	2.90
Q3 (Weight)	2.50	3.20
Maximum Weight	3.00	3.90
Mean Weight	2.36	2.90
Std. Dev. of Weight	0.27	0.47

A) [1.5 pts] Name all variables in this study, and identify each variable as categorical or quantitative.

Sex is categorical and body weight is quantitative (numerical)

0.5 pts for naming the variables, 0.5 pts each for the type

B) [1.5 pts] Can we find the overall median body weight for the 144 cats from this table? If so, find it. If not, explain why not.

No, because we don't have enough information to order all of the 144 cats from lightest to heaviest

Or, we have no idea what weight corresponds to the average weight of the 72nd and 73rd heaviest cats

0.5 pts for no, 1 pt for any reasonable explanation involving ranking the body weights

C) [4 pts] In pounds (1 kg = 2.2 lbs), what is the overall mean body weight of the 144 cats?

2 pts the female cats weighed a total of 2.36*47 = 110.92 kg, and the male cats weighed a total of

2.90*97 = 281.30 kg, so the mean weight of the 144 cats was (110.92 + 281.30)/144 = 2.72 kg

2 pts using linear transformation rules, the mean body weight in lbs would be

2.72 * 2.2 = 5.99 (or 5.98 if rounded) lbs

D) [4 pts] If the distribution of male cat body weights were normal with mean and standard deviation given by the table, give three ranges of weights corresponding to, respectively, the lightest 95% of male cats, the heaviest 95%, and the middle 95%. Label which range corresponds to which group of cats.

1 pt by 68-95-99.7% rule, the middle 95% of male cats are between 2.90 – 2*0.47 and 2.90 + 2*0.47

kg, or between 1.96 and 3.84 kg

2 pts the lightest (heaviest) 95% of male cats correspond to a z-score of 1.65 (-1.65) according to

the table, so they have a cutoff weight of 1.65*(0.47) + 2.90 = 3.68 kg (or -1.65*0.47 + 2.90 = 2.12 kg)

1 pt the lightest 95% of male cats are lighter than 3.68 kg and heaviest 95% are heavier than 2.12 kg

2. A 2017 American Medical Association Report recorded the mortality rate (death rate) from various types of cancer in 3,047 counties and county-equivalents in the U.S.

A) [1 pt] What was a case in this study?

Each case in this study is a county (or a county-equivalent)

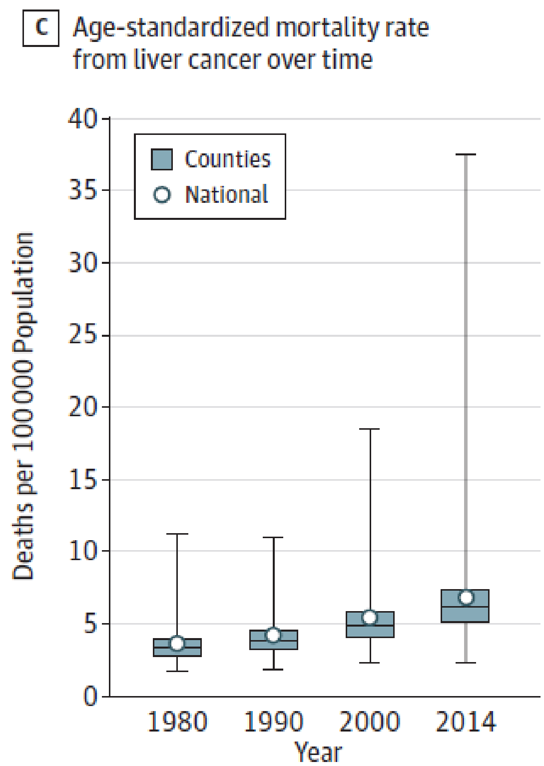
B) [2 pts] Should we treat the national mortality rate from liver cancer as a statistic or a parameter? Justify your answer.

A parameter, because we assume we are summarizing deaths for an entire population

Alternatively, points for explaining why we would have to treat this as coming from a sample

and therefore claiming it is a statistic

For parts C-E, refer to the figure to the right.



C) [1 pt] What plots are depicted in the figure (circle all correct answers)?

box plot   histogram   bar plot

1 pt for circling only box plot

D) [1 pt] In 1990, the distribution of liver cancer mortality rate among the counties was (circle the most likely correct answer):

skewed left   symmetric   skewed right

E) [3 pts] Which of these statements about the death rates from liver cancer are <u>definitely true</u> (circle the letter of each statement that can be proven from the figure)? 0.5 pts per circled/uncircled answer

a. Over time, the death rate from liver cancer in the average American county has increased

b. Over time, the variability among counties in death rates from liver cancer has increased

c. In 2000, less than 75% of counties had death rates from liver cancer below 5%

d. In 2014, at least half of all counties had death rates between 5 and 10 deaths per 100,000 population

e. According to our rule for numerically finding outliers, some counties in 2014 have outlier death rates

f. The distribution of liver cancer death rates among counties is unimodal

3. In lecture we discussed how to lose money at American roulette. It turns out that in French roulette there is only one green 0, and so 18 out of 37 slots are red and 18 out of 37 slots are black. If you bet $1 on black, you win $1 if black shows up and lose $1 if it doesn't.

A) [3 pts] Suppose you bet $1 on black for each of 8 independent spins of a French roulette wheel. Explain why the number of bets you win can be modeled as a binomial random variable.

Since the spins are independent, we check the B/N/S assumptions of BINS (1 pt each)

B – either we win or lose each bet; N – we commit to bet 8 times without knowing the result of any bet

S – each bet has probability 18/37 of winning

B) [3 pts] What is the probability that you win exactly 3 out of the 8 bets?

2 pts let X be the number of bets won, then X ~ B(8, 18/37) and we want to find P(X = 3)

1 pt use binomial distribution formula $P(X = 3) = \frac{8!}{3!(8-3)!}(\frac{18}{37})^3(1 - \frac{18}{37})^{(8-3)} = 0.23$

C) [5 pts] How much total money do you expect to win or lose over your 8 bets? Justify your answer.

Method 1: The mean of this binomial random variable is $8\left(\frac{18}{37}\right) = 3.89$, so we expect to win $1 on 3.89 bets and lose $1 on 8 – 3.89 = 4.11 bets, so we expect to lose a total of 22 cents over the 8 bets.

Method 2: On each bet, we have a 18/37 chance of winning $1 and a 19/37 chance of losing $1. Therefore, we expect to lose 1/37 of a dollar on a single bet. We can add the means of each bet to get the total amount we expect to win/lose. Since the bets are iid, we expect to lose 8/37 or 22 cents.

D) [5 pts] We design an experiment to test whether a book's "system" for roulette is better than betting haphazardly on red or black. Briefly describe these terms in the context of our experiment.

Experimental (Factor) Variable: the system used (either the book's system or haphazard betting)

1 pt for a description in context; 0.5 pts for just giving a definition

Response Variable: anything related to "how much money is won"

Control Group: the group of people that haphazardly bet on red or black (or just bet on black)

Or, the times/trips we bet our usual way, if we are the only subject

Random Assignment: each person in our study is randomly assigned to one of the two systems

(or, we randomly select whether to use the old system or the new system every time/trip)

Repetition: there are enough people playing each system (or we play enough times/trips) that chance variation in the results should be low compared to the influence of the system

4. A 2010 study investigated cognitive-behavioral therapy in a sample of 109 children with obsessive-compulsive disorder (OCD). After treatment, they recorded the children's scores on the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) and whether the clinician believed the OCD to be in remission.

A) [8 pts] 73 of the 80 children in remission had a Y-BOCS score of 14 or below, while 26 of the 29 children judged not to be remission had higher scores. Suppose the researchers decide to use "Y-BOCS score of 14 or below" as a test to diagnose whether a child is in remission. From this sample, estimate the sensitivity, specificity, positive predictive value, and negative predictive value of this test. Round all estimates to the nearest percent.

Sensitivity: 91%

Specificity: 90%

PPV: 96%

NPV: 79%

Space to show work:

1 pt: Sensitivity = 73/80 = 91.25% or about 91%

1 pt: Specificity = 26/29 = 89.7% or about 90%

1 pt out of 2 for sensitivity/specificity if you flipped them

For calculating PPV and NPV, I am showing a table below because it is the most space-efficient way to do this. However, this was not the only way to find PPV and NPV. 4 pts awarded for setting up a two-way table, a tree diagram, or two Bayes's rule equations and plugging in values correctly.

	Score $\leq$ 14	Score > 14	Total
In Remission	73	7	80
Not In Remission	3	26	29
Total	76	33	109

1 pt PPV = 73/76 = 96.1% or about 96%

1 pt NPV = 26/33 = 78.7% or about 79%

1 pt out of 2 for PPV/NPV if you flipped them

B) [2 pts] Suppose that the researchers instead decided to use "Y-BOCS score of 11 or below" to diagnose whether a child is in remission. How will this affect the sensitivity and specificity of the test? Circle the most likely outcome for both sensitivity and specificity. 1 pt each or 1 pt total for flipping them

Sensitivity will:	increase	decrease	stay the same
Specificity will:	increase	decrease	stay the same

If we decrease the cutoff score to 11, then likely, fewer than 73 of the 80 in remission will be under the cutoff, but more than 26 of the 29 not in remission will be above the cutoff.

5. Suppose that we plan to observe 25 independent realizations of the random variable $X \sim N(15, 5)$ and, independently, 100 independent realizations of the random variable $Y \sim N(12, 8)$.

A) [2 pts] Explain why we cannot, at this point, evaluate whether the statement $\bar{X} > \bar{Y}$ is true or false.

We are still in the planning stages and have not made any observations. Therefore, $\bar{X}$ and $\bar{Y}$ are random variables, and we do not know their values, so we cannot evaluate whether $\bar{X} > \bar{Y}$

Alternatively, $\bar{X} - \bar{Y} \sim N(3, 1.28)$ and there are some parts of that distribution both > 0 and < 0

B) [8 pts] For each of the ten statements below, circle the equality or inequality sign (<, =, or >) that correctly describes the relationship between the quantity on the left and the quantity on the right. There is exactly one correct answer per row.

μ_X	<	=	>	μ_Y
σ_x	<	=	>	σ_Y
$E(X+Y)$	<	=	>	$E(X-Y)$
$Var(X+Y)$	<	=	>	$Var(X-Y)$
$E(\bar{X})$	<	=	>	$E(\bar{Y})$
$Var(\bar{X})$	<	=	>	$Var(\bar{Y})$
$P(X < 10)$	<	=	>	$P(X \leq 10)$
$P(X < 10)$	<	=	>	$P(\bar{X} < 10)$
$P(\bar{X} < 10)$	<	=	>	$P(\bar{Y} < 10)$
$P(\bar{Y} < 10)$	<	=	>	$P(\bar{Y} > 14)$

0.5 pts for rows 1, 2, 3, and 7

1 pt for rows 4, 5, 6, 8, 9, and 10

Space to show work: Rows 1 and 2: there is no real work to show, this is just interpreting μ and σ

Rows 3 and 4 can be shown from the rules for linear combinations of independent random variables

Rows 5 and 6: By CLT $\bar{X} \sim N\left(15, \frac{5}{\sqrt{25}}\right)$ and $\bar{Y} \sim N\left(12, \frac{8}{\sqrt{100}}\right)$, or $\bar{X} \sim N(15, 1)$ and $\bar{Y} \sim N(12, 0.8)$

Row 7: Since X is a continuous random variable, P(X < x) and P(X $\leq$ x) are considered to be equal

Row 8: For the distribution of X, 10 corresponds to a z-score of -1. For the distribution of X-bar, 10 corresponds to a z-score of -5. Clearly there is more area to the left of -1 than to the left of -5

Row 9: For the distribution of Y-bar, 10 corresponds to a z-score of -2.5. Again, clearly there is more area to the left of -2.5 than to the left of -5

Row 10: for the distribution of Y-bar, 14 corresponds to a z-score of 2.5. There is equal area under the density curve describing Y-bar to the left of 10 and to the right of 14.

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 $P(Z \leq z)$.

z-score	Cumulative Proportion
-3.00	0.0013
-2.50	0.0062
-2.00	0.0228
-1.65	0.0495
-1.28	0.1003
-1.00	0.1587
-0.67	0.2514

z-score	Cumulative Proportion
0.67	0.7486
1.00	0.8413
1.28	0.8997
1.65	0.9505
2.00	0.9772
2.50	0.9938
3.00	0.9987

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