MA 207 Exam 1 Review **Name:** Key **Date:** **Time:**

A gradebook for a summer statistics course includes columns for exam 1 score and course letter grade.

a. For each variable, state whether it is quantitative or categorical.

i. Exam 1 score quantitative - continuous (could in theory get 89.243)

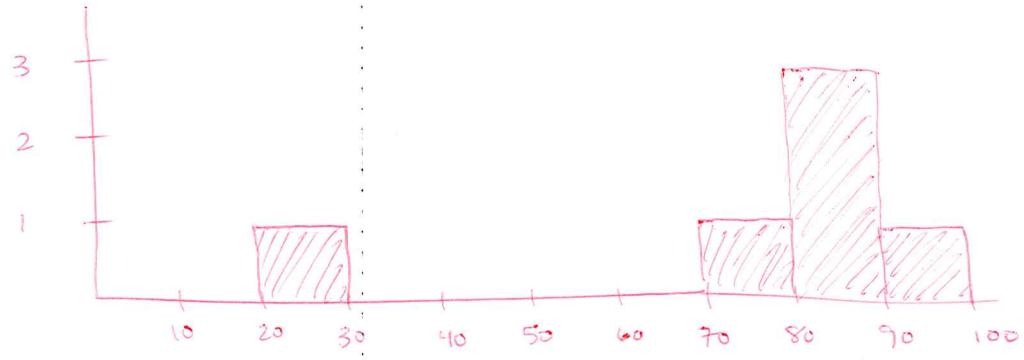
ii. Letter grade categorical

b. The exam 1 scores were: 75, 80, 22, 85, 92, 84. Make a stem-and-leaf plot for the exam 1 scores.

Ordered list: 22, 75, 80, 84, 85, 92

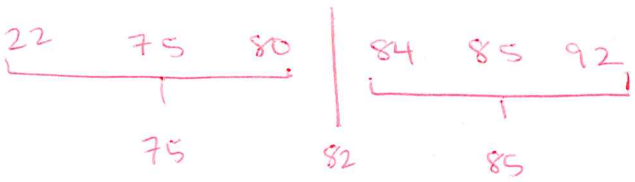
Stem	Leaf
2	2
3	
4	
5	
6	
7	5
8	0 4 5
9	2
10	

c. Make a histogram for the exam 1 scores using the intervals [0,10), [10,20), etc. on the horizontal axis.



d. Find the five number summary.

Min: 22
Q_1: 75
med: 82
Q_3: 85
max: 92



e. Find the range.

92 - 22 = 70

f. Find the interquartile range.

85 - 75 = 10

g. What score must a student be above to be a high outlier according to the IQR criterion?

$Q_3 + 1.5\,IQR = 85 + 1.5(10) = 100$

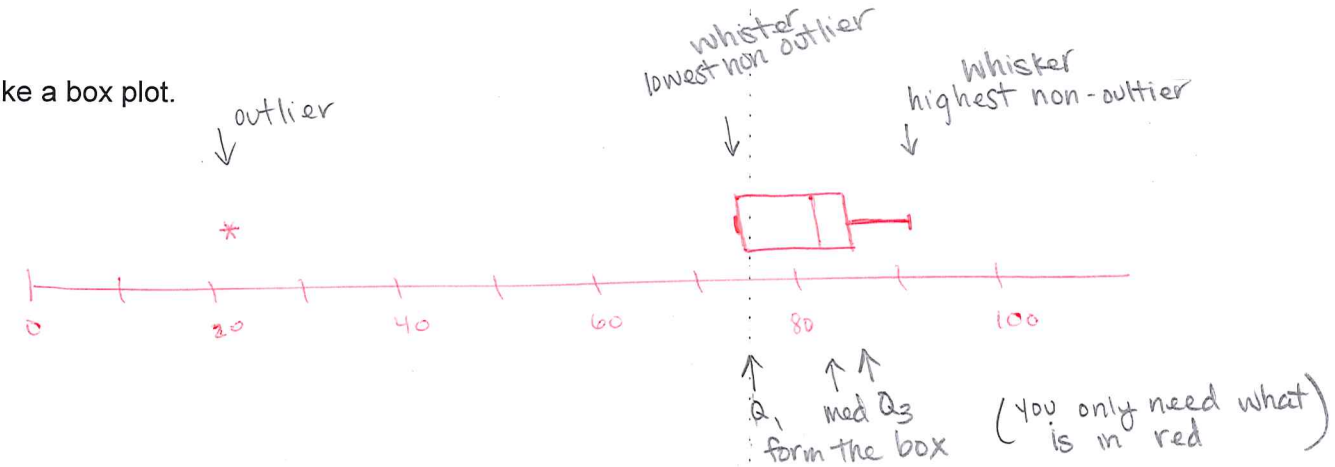
h. What score must a student be below to be a low outlier according to the IQR criterion?

$Q_1 - 1.5\,IQR = 75 - 1.5(10) = 60$

i. Are there any outliers according to the IQR criterion? If so, list them. If not, say there are none.

Yes; 22 is an outlier

j. Make a box plot.



k. The same exam 1 scores are listed again below. Calculate the mean (to one decimal place more than the observations) and the standard deviation (to one decimal place more than the mean) by setting up a chart as done in class.

Student #	Exam 1 score	$x-\bar{x}$	$(x-\bar{x})^2$
1	75	2	4
2	80	7	49
3	22	-51	2601
4	85	12	144
5	92	19	361
6	84	11	121
	$\Sigma x = 438$		3280

$\bar{x} = 73$

$\frac{\Sigma x}{n} = \frac{438}{6}$

$\frac{3280}{6-1} = 656$

$s = \sqrt{656} \approx 25.6$

l. For this part only, consider the data set *without* any outliers.

75
80
85
92
84

i. Find the mean of the data without outliers. Is it much different from the mean with outliers.?

83.2 yes; 10 pts higher

ii. Find the median of the data without outliers. Is it much different from the median with outliers.?

84 no; 2 pts higher

iii. In general, is the mean resistant to outliers?

No!

iv. In general, is the median resistant to outliers?

Yes!

m. In general, when is it more appropriate to use the mean as a measure of center, and when is it more appropriate to use the median as a measure of center?

If data is symmetric with no outliers, use mean otherwise

n. Describe the shape, center, and variability of the distribution and any striking deviations from the overall pattern.

① hard to have much shape with only 6 pieces of data. The low outlier skews the ~~chart~~ shape to the left. ② Median is 82. ③ IQR is 10

Hints (these hints will *not* be given on the exam): (1) For the shape, is the distribution symmetric, right skewed, or left skewed and how many modes does it have (e.g. unimodal, bimodal)? (2) For the center use *either* the median *or* the mean, whichever is appropriate for this particular data set. (3) For the variability, if the median was the appropriate measure of center, state the IQR to describe the variability, but if the mean was the appropriate measure of center, state the standard deviation to describe the variability. (4) If you used the IQR to measure variability, use the IQR criterion to find the outliers (you already did that above), but if you used the standard deviation to measure variability, then use the standard deviation criterion to find the outliers.]

④ one outlier according to the IQR criterion - 22.

2. In a study, seeds of a particular flower were germinated, and the heights x of the plants were measured after one month. The data was found to be approximately bell-shaped with mean 1.75 feet and standard deviation 0.25 feet.

a. What is the z-score for a plant that is 2 ft?

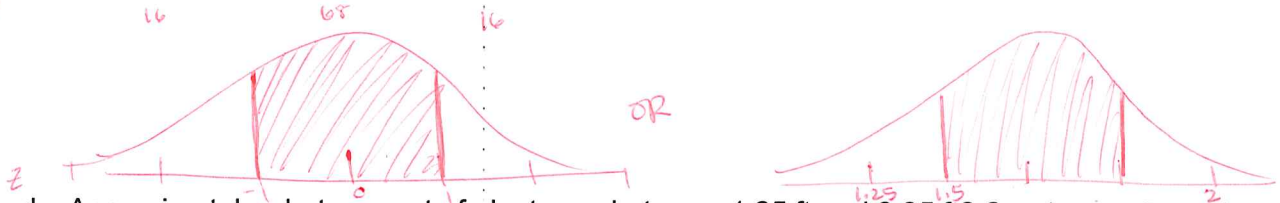
$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{(2 - 1.75)}{.25} = \boxed{1}$$

b. What is the z-score for a plant that is 1.5 ft?

$$z = \frac{(1.5 - 1.75)}{.25} = \boxed{-1}$$

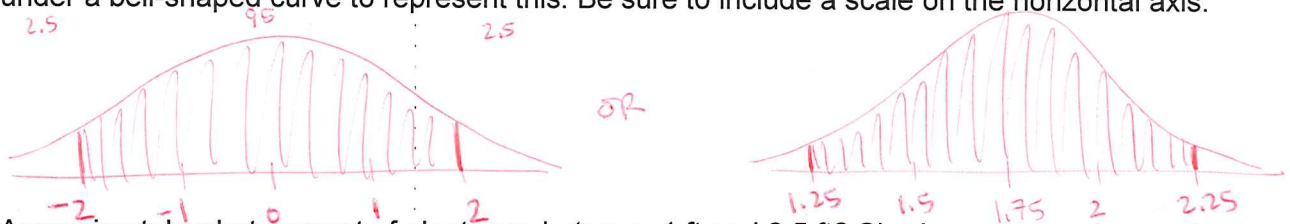
c. Approximately what percent of plants are between 1.5ft and 2 ft? Shade an appropriate region under a bell-shaped curve to represent this. Be sure to include a scale on the horizontal axis.

68%



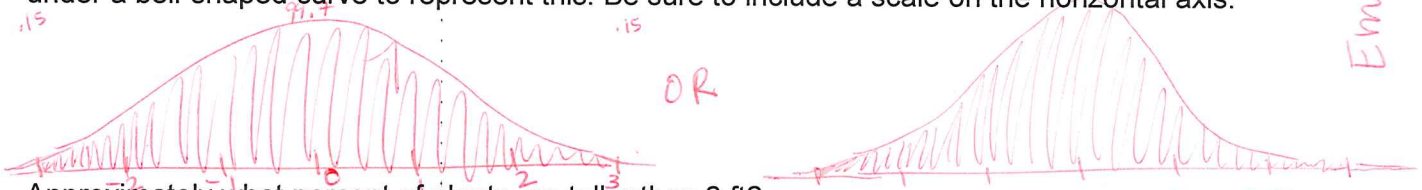
d. Approximately what percent of plants are between 1.25 ft and 2.25 ft? Shade an appropriate region under a bell-shaped curve to represent this. Be sure to include a scale on the horizontal axis.

95%



e. Approximately what percent of plants are between 1 ft and 2.5 ft? Shade an appropriate region under a bell-shaped curve to represent this. Be sure to include a scale on the horizontal axis.

99.7%



Empirical Rule

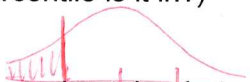
f. Approximately what percent of plants are taller than 2 ft?

$100 - 68 = 32$ $\quad \frac{32}{2} = \boxed{16\%}$

g. Approximately what percent of plants are taller than 1.25 ft?

Less than 1.25: $100 - 95 = 5$, $5/2 = 2.5$ / So taller than 1.25 is $\boxed{97.5\%}$

h. Approximately what percent of plants are shorter than 1.5 ft? (Equivalently, if a plant is 1.5 ft, what percentile is it in?)



$100 - 68 = 32$

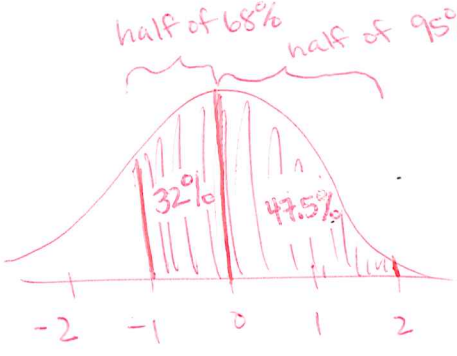
$32/2 = \boxed{16\%}$

i. Approximately what percent of plants are shorter than 2.25 ft?

$\boxed{97.5\%}$

j. Approximately what percent of plants are between 1.5 ft and 2.25 ft?

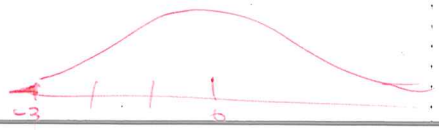
$\boxed{79.5\%}$



k. If a plant is 1 ft, what is its z-score?

$$z = \frac{1 - 1.75}{.25} = \boxed{-3}$$

l. If a plant is 1 ft, what percentile is it in?



$100 - 99.7 = .3$

$.3/2 = .15$

.15% lie below 1 ft, so $\boxed{0^{th}}$ percentile.

m. If a plant has a z-score of 1, what percentile is it in?

$z = 1$ 16% above $z=1$, 84% below $z=1$ **84th percentile**

n. If a plant has a z-score of 1, what is its height? (Remember to include units whenever appropriate.)

$1 = \frac{x - 1.75}{.25}$ $(1)(.25) = x - 1.75$ $2 = x$ **2 ft**

o. What is the z-score for a plant that has height $x = 1.6$ feet?

$z = \frac{1.6 - 1.75}{.25} = \frac{-.15}{.25} = -0.6$ there are no units for z-scores

p. If a plant has z-score $z = 1.6$, what is its height? (Include units.)

$1.6 = \frac{x - 1.75}{.25}$ $(1.6)(.25) = x - 1.75$ $.4 = x - 1.75$ $1.75 + .4 = x$ **2.15 ft**

q. What height must a plant be above to be a high outlier according to the standard deviation criterion?

$\bar{x} + 3s = 1.75 + 3(.25) =$ **2.5 ft**

r. What height must a plant be below to be a low outlier according to the standard deviation criterion?

$\bar{x} - 3s = 1.75 - 3(.25) =$ **1 ft**

3. Below is a contingency table showing survivals (S) and deaths (D) of patients after surgery in two hospitals (A and B). Give your answers are fractions (do not reduce the fractions).

	Hospital A	Hospital B	
Died	63	16	79
Survived	2037	784	2821
	2100	800	2900

(a) What proportion of subjects survived? What kind of proportion is this?

$2821/2900$ marginal

(b) What proportion of patients were at hospital A? What kind of proportion is this?

$2100/2900$ marginal

(c) What proportion of survivors were at hospital A? What kind of proportion is this?

$2037/2821$ conditional

(d) What proportion of hospital A patients survived? What kind of proportion is this?

$2037/2100$ conditional

(e) What proportion of patients were at hospital A and survived? What kind of proportion is this?

(intersection) $2037/2900$ joint

(f) What proportion of patients died at hospital B or survived at hospital A?

(union) $\frac{16 + 2037}{2900} = \frac{2053}{2900}$ has no name

(g) What proportion of patients were at hospital A or survived?

$\frac{63 + 2037 + 784}{2900} = \frac{2884}{2900}$ has no name

(h) True or false: Hospital A is better because more patients survived. Justify your answer.

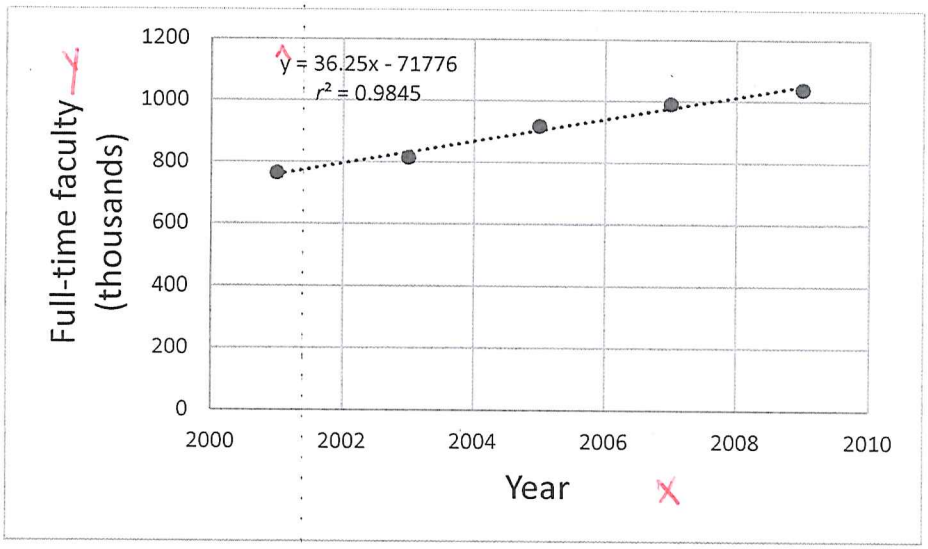
It is not the number of survivors that matters, but the proportion of survivors that is important.

$P(S|A) = 2037/2100 = .97$

$P(S|B) = 784/800 = .98$

(Also, the two hospitals have different populations

4. The scatter plot below shows number of full-time faculty members (in thousands) at four year colleges in certain years, together with the equation of the least squares regression line and the value of r^2.



(i) What is the explanatory variable? Year

(j) What is the response variable? full time faculty in thousands

(k) Describe the overall pattern (form, direction, strength) and any striking deviations from the overall pattern (regression outliers). [Hint: form is linear or nonlinear, direction is positive or negative, strength is strong or weak. There are no hints on the exam.]

linear, positive, strong, no outliers.

(l) In a complete sentence, explain what the value of r^2 means in the context of this problem. Be sure to include the actual value of r^2 in your explanation.

98.45% of the variation in the number of full-time faculty can be explained by the variation in years.

(m) Does the passage of time (an increase in the year) *cause* an increase in the number of faculty?

No. Association is not causation.

(n) What does the regression line predict for 2007? Give your answer to two decimal places and include units. Show your work using the equation of the regression line. Is this extrapolation: Yes or no?

$\hat{y} = 36.25(2007) - 71776 =$ 977.75 thousand full-time faculty.

OR 977,750

(o) Estimate the actual value in 2007 from the scatter plot. Include units.

1000 thousand

OR 1,000,000

(p) What is the residual for 2007? Include units.

actual - predicted

$1000 - 977.75 =$ 2.25 thousand full-time faculty.

(q) What is the *y*-intercept, including units? Explain what this means in a complete sentence in the context of this problem. Does this make sense in this context?

$\hat{y} = 36.25(0) - 71776 =$ −71,776

In the year 0 (two millenia ago), there were −71,776 thousand full-time faculty. This is not practical to have negative faculty – too far extrapolated

(r) What is the slope, including units? Explain what this means in a complete sentence in the context of this problem.

36.25 thousand full-time faculty per year $\left(\frac{\Delta y}{\Delta x}\right)$

If one more year passes, the number of full-time faculty goes up by 36.25 thousand.

(s) According to the regression line, during what year will there be 1,100,000 faculty members? Show your work using the equation of the regression line. Give your answer to the nearest year. Is this extrapolation: Yes or no? $\hat{y} = 1{,}100$ thousand

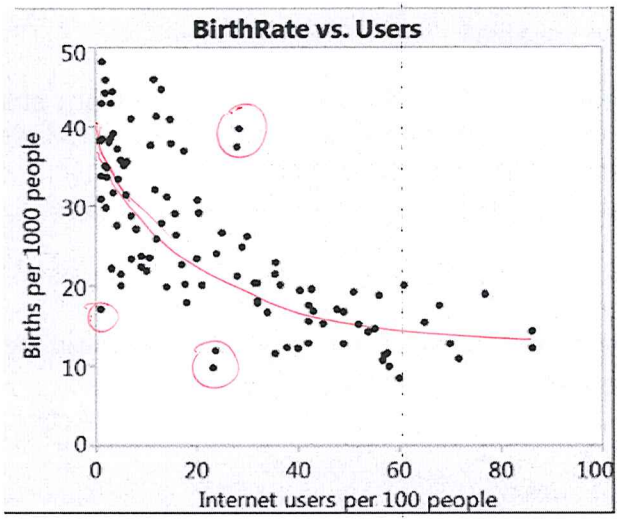
$36.25x - 71776 = 1100$

$36.25x = 72{,}876$

$x = \frac{72876}{36.25} \approx 2010$

X is beyond the highest x-valued observation

5. Consider the scatterplot below showing data available at www.worldbank.org.



a. What is the explanatory variable? Internet users per 100 people

b. What is the response variable? Births per 1000 people

c. Do changes in the explanatory variable *cause* changes in the response variable: Yes or no?

Association (or correlation), not causation

d. Describe the overall pattern and any striking deviations from the overall pattern.

Non linear, negative, weak, potential outliers circled

e. Does it make sense to compute a correlation coefficient for this association: Yes or no? Why or why not?

No! Correlation refers to a linear association.