Math 075 Final Review Sheet

"Categorical Analysis"

Things to Remember from this Unit:

- ✓ Classify data as quantitative or categorical
- ✓ How to convert a decimal proportion into a percentage.
- ✓ How to convert a percentage into a decimal proportion.
- ✓ How to find a decimal proportion and percentage from categorical data
- ✓ How to read and analyze categorical data from bar plots and pie charts
- ✓ How to use a percentage and a total to find an expected amount.

Classify the following as categorical or quantitative.

1. A person's opinion about raising or lowering taxes.
2. The number of cigarettes a person smokes each day.

Convert the following decimal proportions into percentages by multiplying by 100%.

3. 0.0487
4. 0.926
5. 0.0033

Convert the following percentages into a decimal proportion by dividing by 100 and removing the % sign.

6. 0.52%
7. 7.46%
8. 23.9%

9. A car dealership has a total of 177 vehicles. Of those vehicles, 58 are minivans. What is the decimal proportion of vehicles that are minivans? What percentage of the vehicles are minivans?
10. About 17% of the athletes in an athletic department have some kind of injury during the season. If the department has a total of 123 athletes this year, how many should we expect to have at least one injury during the season?

"Categorical Relationships"

Things to Remember from this Module:

- ✓ How to create two way tables from data OR percentages
- ✓ How to use two way tables to find regular, conditional, and joint percentages
- ✓ How to determine if categories are related or not from percentage information.

Use the two way table below to answer the following questions.

	Prefer Cats	Prefer Dogs	Do not like cats or dogs
Male	14	89	21
Female	56	47	13

11. What percent of the sample preferred dogs?

12. If a person was female, what percent preferred cats?

13. What proportion are both male and prefer dogs?

14. What percent of the people were either female or did not like cats or dogs?

15. What percentage of the males prefer cats?

16. 11.3% of males prefer cats and 48.3% of females prefer cats. Does this indicate that gender and liking cats are related or not? Explain why.

17. Does this indicate that a person's gender causes them to like cats?

18. Describe the process of making a two way table if you are only given the grand total and marginal and conditional percentages.

"Quantitative Data Analysis"

Things to Remember from this Module:

- ✓ How to calculate the Mean, Median, Mode, IQR, Range, Frequency (n) , Q1 and Q3
- ✓ How to write sentences to explain the Mean, Median, Mode, IQR, Range, Standard Deviation, Frequency (n) , Q1 and Q3
- ✓ Shapes of Histograms and Dot Plots
- ✓ Know the difference between categorical and quantitative variables
- ✓ Make a boxplot using data
- ✓ Read and interpret information from a box plot
- ✓ What is the correct measure of center and spread depending on the shape
- ✓ How to calculate two numbers that typical values are in between and whether something is an outlier or not
- ✓ Write an essay describing a data set by interpreting all of the sample statistics and graphs.

Practice questions

19. Use the following set of data: 7, 9, 9, 10, 11, 11, 12, 12, 12, 12, 15, 22
    a. Find the Median Q_2:
    b. Find the quartiles Q_1 and Q_3. Now find the Interquartile Range (IQR).

20. Create a boxplot from the data; don't forget to calculate if there are any outliers!!

21. Using the following data set: 1, 1, 2, 3, 3, 5, 6
    a. Find the Mean
    b. Find the Range
    c. Find the Standard Deviation

22. What are the four shapes discussed in chapter 2? For each shape, draw a histogram and boxplot with that shape.

23. If you are asked to describe a data set in an essay from graphs and sample statistics, what key things should you discuss?

24. If a data set is bell shaped, what is the best measure of center and spread? How do you find typical values? How do you identify unusual values?

25. If a data set is <u>not</u> bell shaped (skewed), what is the best measure of center and spread? How do you find typical values? How do you identify unusual values (outliers)?

"Linear Quantitative Relationships"

Things to Remember from this Module:

- ✓ How to calculate slope from the r and standard deviation
- ✓ Write a sentence to explain the slope as a rate of change with units
- ✓ How to find the y intercept from the means and slope
- ✓ Write the equation of a regression line from the slope and y-intercept
- ✓ Be able to judge the strength and direction of a linear relationship from a scatterplot and the correlation coefficient (r)
- ✓ Explanatory (X) and Response (Y), know how to determine which is which
- ✓ Write a sentence to explain r, r^2 and S_e
- ✓ No how to judge if a residual plot is evenly spread out
- ✓ No how to judge if a histogram of the residuals is bell shaped and centered at zero.
- ✓ Be able to identify possible confounding variables
- ✓ Do not extrapolate using your linear regression model
- ✓ Correlation is NOT Causation!!!!

Practice Questions:

26. We used StatCrunch to calculate the regression line for some data. The explanatory variable x represented number of years after 1990 and the response variable y represented the price of the home.

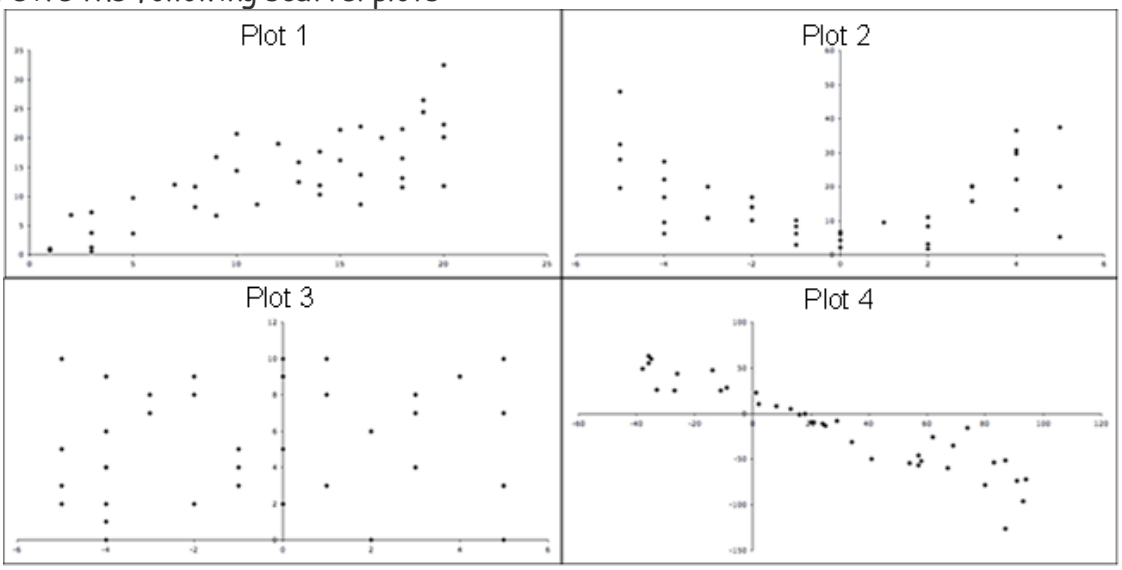
    Price of Home = 96000 + 11000 (Years since 1990)

    a. What is the slope? Write a sentence to explain the meaning of the slope

    b. What is the y-intercept? Write a sentence to explain the meaning of the y intercept
    c. Find the equation of the line representing Ann's house value

    d. Estimate Ann's house value in 2004 (year 14). Assume it is not an extrapolation.

27. Give the following scatterplots



    a. Write positive, negative, or no correlation next to each scatterplot

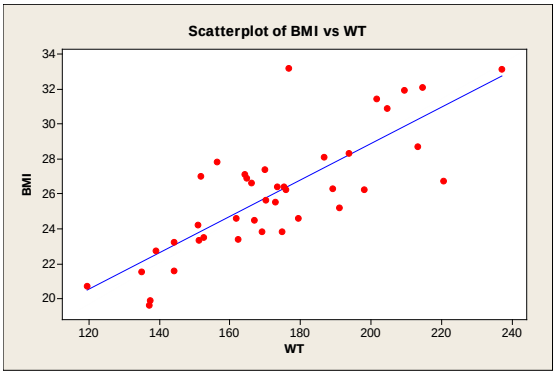
    b. Which graph has the strongest linear relationship
    c. Which graph has the strongest curved relationship

    d. The four correlation coefficients for the scatterplots shown are -0.1169, 0.7699, -0.9396, and 0.1632. Match the correlations to the plots.

28. Does a strong positive correlation PROVE that one variable causes the other variable to respond? Does no correlation always imply there is no relationship? Give an example of two variables that confirms your answers.

29. For an essay question on a relationship between two variables, what key things should you talk about?

30. We looked at 40 randomly selected men to analyze the relationship between the weight of a man and his BMI (Body Mass Index). A software found the following graphs and statistics.



Correlation Coefficient of WT and BMI = 0.800

The regression equation is Y = 8.02 + 0.104X

i. What does the scatterplot tell us about the relationship between the weight of a man and his Body Mass Index?

ii. A trainer said that if a man is heavy, it will cause him to have a large BMI. Does the data support this statement?

iii. What is the r value and what does it tell us?

iv. What is the r^2 value and what does it tell us? List other confounding variables thatnmight influence BMI besides weight?

v. Use the regression line to predict the BMI of a man that weighs 220 pounds? How accurate do you think this prediction is?

vi. Can we use the regression line to predict the BMI of a man that is 100 pounds? Why or why not?

31. Find the **least squares regression line**

Descriptive Statistics: Age, Distance

Variable	N	Mean	Median	TrMean	StDev
Age (X)	30	51.00	54.00	51.19	21.78
Distance (Y)	30	423.0	420.0	422.3	82.8

Variable	Minimum	Maximum	Q1	Q3
Age	18.00	82.00	27.75	71.25
Distance	280.0	590.0	360.0	467.5

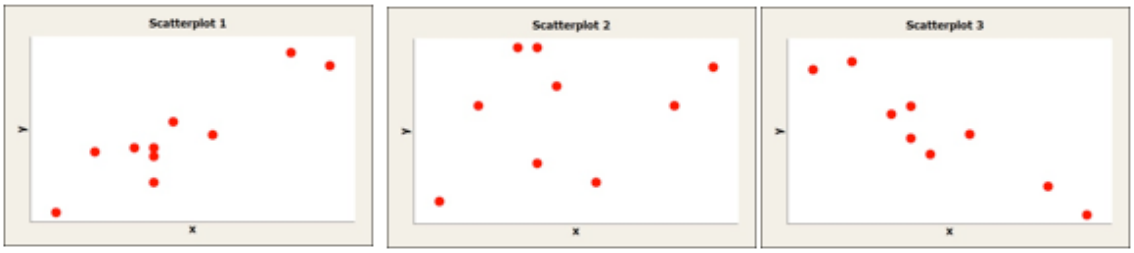
Correlations: Age, Distance

Pearson **correlation** of Age and Distance = -0.793

Regression Lines. Give the equations for slope and y intercept <u>of a regression line.</u>

$y = mx + b$ Where $m = \frac{r \bullet s_y}{s_x}$ and $b = \overline{y} - m \bullet \overline{x}$

Match each description of a set of measurements to a scatterplot.



32. x = average outdoor temperature and y = heating costs for a residence for 10 winter days
33. x = height (inches) and y = shoe size for 10 adults
34. x = height (inches) and y = score on an intelligence test for 10 teenagers

35. For the following data:
    a) Make a scatter plot. (x,y)

x	y
2	3
4	4
6	8
8	11
10	13
11	14

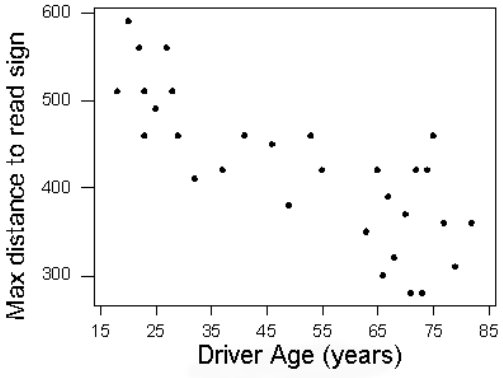
b) Describe the strength and direction of the scatterplot you just drew.

c) What does an outlier do to an r-squared value? Will the r-squared value increase or decrease?

Identify the explanatory and response variables in the following:

36. How the price of a package of meat is related to the weight of the package?
    a. Explanatory:
    b. Response:
37. Is the distance you walk related to the calories you burn?
    a. Explanatory:
    b. Response:

38. Use the following Scatterplot



a) The line we will use to describe this scatterplot is: **Y = 576 – 3X,** use the linear model to interpret the slope.

b) Interpret the y intercept.

"Curved Quantitative Relationships"

Things to Remember from this Module:

- ✓ Create exponential, logarithmic, and quadratic curves with technology, give the scope of the x – values, r-squared, standard deviation of the residual errors and use them to assess the fit of the curve to the data.
- ✓ Create a residual plot and the histogram of the residuals and be able to read them.
- ✓ Use exponential, quadratic, and logarithmic curves to make predictions in the scope of the x-values
- ✓ Use R-squared, and the Standard Deviation of the Residual Errors to determine which curve is the best fit.

Practice Problems

39. Write an essay on the following topic: How can we use curves to describe the relationship between two variables? How can we use R-squared and the Standard Deviation of the Residual Errors to assess which curve is the best fit?

40. Use the nonlinear data on the website www.matt-teachout.org .
Analyze the Mother's Age verses Birth Underweight data. The explanatory variable was the mothers age in years and the response variable was the birth weight in grams. Use Statcato find an exponential curve that fits the data. What is the R-squared? . (Remember the Standard Deviation of the Residuals is wrong in Statcato. The correct Standard Deviation of the Residuals for the exponential curve is 360.6 grams) Write a sentence explaining R-squared and two sentences explaining the Standard Deviation of the Residual Errors?

41. Use the nonlinear data on the website www.matt-teachout.org .
Analyze the Mother's Age verses Birth Underweight data. The explanatory variable was the mothers age in years and the response variable was the birth weight in grams. Use Statcato find a logarithmic curve that fits the data. What is the R-squared and the Standard Deviation of the Residuals? Write a sentence explaining R-squared and two sentences explaining the Standard Deviation of the Residual Errors?

42. Use the nonlinear data on the website www.matt-teachout.org . Analyze the Mother's Age verses Birth Underweight data. The explanatory variable was the mothers age in years and the response variable was the birth weight in grams. Use Statcato find a quadratic curve that fits the data. What is the R-squared and the Standard Deviation of the Residuals? Write a sentence explaining R-squared and two sentences explaining the Standard Deviation of the Residual Errors?

43. Which curve was the best fit for the Mother's Age vs birth underweight data? Explain why?

44. Use the exponential curve formula to predict the baby weight if the mom is 18 years old. How much error will be in that prediction?

45. Use the logarithmic curve formula to predict the baby weight if the mom is 22 years old. How much error will be in that prediction?

46. Use the quadratic curve formula to predict the baby weight if the mom is 30 years old. How much error should we expect in that prediction?

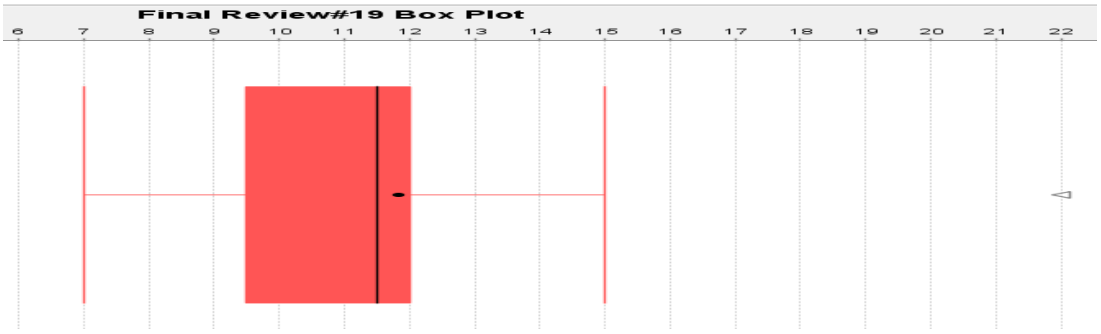
47. Go to the nonlinear data again. Find the quadratic curve that best fits the month in 2009 verses solar energy in kilowatt hours. What is the x and y coordinate of the vertex. Explain the meaning of both the x coordinate and the y coordinate.

Answer Key

1. Categorical
2. Quantitative
3. 4.87%
4. 92.6%
5. 0.33%
6. 0.0052
7. 0.0746
8. 0.239
9. 58/177 = 0.328 = 32.8%
10. 17% = 0.17, Amount = 0.17 x 123 = 20.91 (about 21 players with at least one injury)
11. 136/240 = 0.567 = 56.7%
12. 56/116 = 0.483 = 48.3%
13. 89/240 = 0.371 = 37.1%
14. 137/240 = 0.571 = 57.1%
15. 14/124 = 0.113 = 11.3%
16. It seems that gender is related to liking cats. If there was no relationship, then the percentages for men and women would be close. These percentages are significantly different indicating a relationship.
17. No. Relationships, Associations and Correlations do not imply causation. There are many confounding variables involved.
18. Use the marginal (regular) percentages to find the totals. Then use the conditional percentages to fill in the table. You will be multiplying the conditional percentages

19. Use the following set of data: 7, 9, 9, 10, 11, 11, 12, 12, 12, 12, 15, 22
    a. Find the Median Q_2: 11.5
    b. Find the quartiles Q_1 and Q_3. Now find the Interquartile Range (IQR).
       Q1=9.5 , Q2=11.5 , Q3=12 , IQR=2.5

20. Create a boxplot from the data; don't forget to calculate if there are any outliers!!



21. Using the following data set: 1, 1, 2, 3, 3, 5, 6
    a. Find the Mean 3
    b. Find the Range 5
    c. Find the Standard Deviation 1.9

22. Skewed Left (tail on left), Skewed Right (tail on right), Bell Shaped (Symmetric, Normal), Uniform (Rectangular shape)

23. Shape, Center (find best Average) , Spread (find how spread out typical numbers are and find two #s that typical numbers fall inbetween) , Outliers (should they be removed or not and if removed, what does that do to the shape?)

24. If a data set is bell shaped, the best measure of center is the mean. The average is also the mean. The best spread is the standard deviation. Typical values are 1 standard deviation from the mean. (mean – stand dev ≤ typical values ≤ mean + stand dev)
    Unusually high values are 2 or more standard deviations above the mean and unusually low values are 2 or more standard deviations below the mean.
    (unusually high ≥ mean + 2 stand dev)
    (unusually low ≤ mean – 2 stand dev)

25. If a data set is skewed, the best measure of center is the median. The average is also the median. The best spread is the interquartile range (IQR). Typical values are between the 1st and 3rd quartiles. (Q1 ≤ typical values ≤ Q3)
Unusually high values are 1.5 IQR above Q3 and unusually low values are 1.5 IQR below Q1.
(unusually high ≥ Q3 + 1.5 IQR)
(unusually low ≤ Q1 – 1.5 IQR)

26.

a. What is the slope? Write a sentence to explain the meaning of the slope m = 11000 , Housing prices are increasing $11000 per year
b. What is the y-intercept? Write a sentence to explain the meaning of the y intercept b = 96000 In 1990, (year zero) the price of a house was $96000
c. Find the equation of the line representing Ann's house value
y=11000x+96000

d. Estimate Ann's house value in 2004 (year 14). Assume it is not an extrapolation. $250000

27.

a. Plot 1: moderate positive correlation ; Plot 2: No correlation but curved relationship ; Plot 3: No correlation No relationship ; Plot 4: Strong Negative Correlation

b. Plot 4

c. Plot 2

d. Plot 1: 0.7699 ; Plot 2 & 3: 0.1632 or -0.1169 ; Plot 4: -0.9396

28. A strong Positive correlation indicates that the two variables have a positive linear relationship. That is to say as the explanatory variable increases, the response variable also increases. It does not imply that one variable causes the other because there may be lurking variables involved that may influence the explanatory and response variables. No correlation simply means no linear relationship. There can be a nonlinear relationship like a quadratic, logarithmic or exponential relationship.

29. You should talk about the scatterplot, r value, r squared, Standard Deviation of the Residual Errors, the type of relationship (correlation or nonlinear or none) and the strength and direction of the relationship (strong, moderate, weak, positive, negative). If there is correlation, you should also find the regression line of best fit and interpret the slope and y intercept. You should then use the residual plots to check the 6 regression criteria.

30.

i. There is a strong positive correlation between weight and BMI.

ii. It does not. Heavy men do tend to have larger BMI's, but that does not mean that one causes the other. We cannot make causation statements.

iii. The r value is 0.800 and this tells us that there is a strong positive correlation between weight and BMI

iv. R squared is 0.64 or 64%. So 64% of the variability in BMI can be accounted for by its linear relationship to the weight. There can be confounding variables such as height, amount of muscle, nutrition, etc.

v. BMI=30.9 This prediction is pretty accurate since the regression line fit the data very well.

vi. No. We should not use the regression line to predict the BMI of a 100 pound man since 100 is not in the scope of the data and would therefore be extrapolation and subject to a lot of error.

31.

m=-3.015 b=576.75

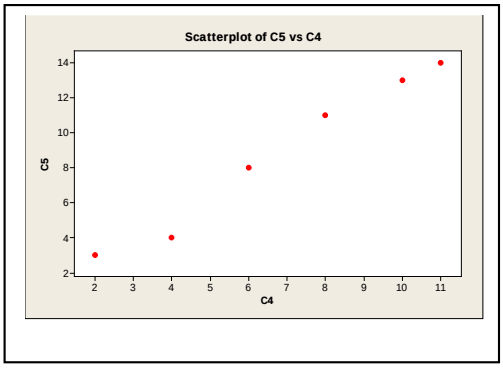
y= 576.75 - 3.015x

32. Plot 3
33. Plot 1
34. Plot 2

35. For the following data:

d) Make a scatter plot. (x,y)

x	y
2	3
4	4
6	8
8	11
10	13
11	14



a) Strong Positive relationship (correlation)

b) An outlier can have a dramatic effect on the shape of a scatterplot. Adding an outlier can sometimes decrease the correlation dramatically. It usually result in decrease in the strength of the linear (or curved relationship). The r-squared percentage will decrease if there is an influential outlier.

36.

a. Explanatory: Weight
b. Response: Price

37.

a. Explanatory: Distance
b. Response: Calories

38.

a) The slope means that as a person gets one year older, the distance to read the sign is decreasing 3 feet.

b) The y intercept of 576 feet would be the distance to read the sign at age zero. Does not make sense in the context of this problem. Notice (0,576) is not in the scope of the data.

39. Scatterplots often show curved patterns and we can find the curve the best describes the relationship. The curve with the highest R-squared value and the lowest Standard Deviation of the Residuals is the best fit.

40.

$y = 3424.43789\ (0.97724)^x$

R-squared = 0.193 = 19.3%

R-squared Sentence: 19.3% of the variability in birth weight can be explained by the exponential relationship with the mothers age.

Standard Deviation of Residuals Sentences: The points in the scatterplot are 360.6 grams from the exponential curve on average. If we use the exponential curve and the mothers age to predict an underweight baby's weight, our prediction could have an average error of 360.6 grams.

41.

Y = 4596.5933 - 810.3625 Ln (X)

R-squared = 0.181 = 18.1%

R-squared Sentence: 18.1% of the variability in birth weight can be explained by the logarithmic relationship with the mothers age.

Standard Deviation of Residuals = 356.8787

Standard Deviation of Residuals sentences: The points in the scatterplot are 356.9 grams from the log curve on average. If we use the log curve and the mothers age to predict an underweight baby's weight, our prediction could have an average error of 356.9 grams.

42.

Y = 2475.7179 + 4.14875 X + -0.91043 X^2

R-squared = 0.1926 = 19.3%

R-squared sentence: 19.3% of the variability in birth weight can be explained by the quadratic relationship with the mothers age.

Standard Deviation of the Residuals = 357.4413 grams

Standard Deviation of the Residuals sentences: The points in the scatterplot are 357.4 grams from the quadratic curve on average. If we use the quadratic curve and the mothers age to predict an underweight baby's weight, our prediction could have an average error of 357.4 grams.

43. None of the curves had a very strong relationship. The r-squared and standard deviations were almost the same. The Quadratic Curve and the Exponential curve had a lightly larger r-squared value than the logarithmic curve but were nearly equal (19.3%). Of those two curves, the quadratic had a slightly smaller standard deviation (357.4) than the exponential (360.6). So the quadratic would be the best fit.

44. 2262.7 grams (with 360.6 grams ave error)

45. 2091.7 grams (with 356.9 grams ave error)

46. 1780.8 grams (with 357.4 grams ave error)

47.

$y = 84.17045 + 425.60047x + (-33.22890)x^2$

The X coordinate of the vertex was 6.4 and y coordinate of the vertex was 1447. The maximum solar energy occurred at time 6.4 months (mid June). The maximum energy was 1447 kWh.