Semester 1 Exam Study Guide Statistics Name: _____

The best way to prepare for a math final is to work out practice problems. Problems from old quizzes and tests will be the best problems to look at in preparation for the exam. You will need a graphing calculator for the exam. You will be permitted to use your Statistics FORMULAE on the exam. The material to be covered on the exam is outlined below.

Chapter 1: Exploring Data
Quiz 1 – Displaying Distributions with Graphs
Quiz 2 – Describing Distributions with Numbers (Mean, Std. Dev, 5-number summary, etc.)
Ch 1 Test

Chapter 2: Describing Location in a Distribution
Quiz 1 – Measures of Relative Standing (percentile, z-scores)
Quiz 2 – Normal Distributions
Ch 2 Test

Chapter 3: Examining Relationships
Quiz 1 – Scatterplots & Correlation
Quiz 2 – Regression & Residuals
Ch 3 Test

Answer Key to 1.1 Stats: Mr.Reddy

Chapter 1: Exploring Data

1.1 Displaying Distributions with Graphs (pp.4-36)

1. In statistics, what is meant by *individuals*? *Individuals* are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

2. In statistics, what is meant by a *variable*? A *variable* is any characteristic of an individual. A variable can take different values for different individuals.

3. What is meant by *exploratory data analysis*? *Exploratory data analysis* uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

4. What is the difference between a *categorical variable* and a *quantitative variable*? A *categorical variable* places an individual into one of several groups or categories. A *quantitative variable* takes numerical values for which arithmetic operations such as adding and averaging makes sense.

5. When is it useful to use a bar chart? Or a pie chart? *Bar charts* and *pie charts* are both used to display categorical data. A pie chart cannot be used unless we have information about all the categories to total 100%

6. What is meant by a *distribution*? How do you describe the overall pattern of a distribution? The *distribution* of a variable tells us what values the variable takes and how often it takes these values.
   To describe the overall pattern of a distribution:
   1) Give the center and spread
   2) See if the distribution has a simple shape that you can describe in a few words.

7. Define *range*: The *range* is the difference between the largest and smallest values of data distribution.

8. When is it better to use a *histogram* rather than a *dotplot*? When you have many data values.

9. What is meant by *frequency* in a histogram? The frequency = the number of counts in each class.

10. When setting a window for constructing a histogram on the TI-84:

    a. What is the significance of Xscl? The width of each class

    b. How do you choose the values of Xmin and Xmax? Xmin = smallest data value and Xmax = largest data value

    c. What is the significance of Ymax? Ymax must be larger than greatest frequency in any class.

11. Define *outlier*. An *outlier* is an individual observation that falls outside the overall pattern of the graph.

12. If a distribution is *symmetric*, what does its histogram look like? The right and left sides are approximately mirror images of each other.

13. If a distribution is *skewed right*, what does its histogram look like? The right side (larger values) extends much farther out than the left side.

14. If a distribution is *skewed left*, what does its histogram look like? The left side (smaller values) extends much farther out than the right side.

15. How is the *stemplot* of a distribution related to its histogram? A histogram is a shaded in stemplot – on the histogram we lose the individual data values but we keep the overall shape of the distribution.

16. When is it advantageous to split stems on a stemplot? When each stem has many leaves.

17. What is the purpose of a *back-to-back stemplot*? To compare the shapes of 2 distributions.

18. When is it useful to construct a *time plot*? For variables that are measured over time, such as the height of a growing child, seasonal variation, price of a stock.

1.2 Describing Distributions with Numbers (pp.37-63)

1. In statistics, what is the most common measurement of center? The arithmetic average, or *mean*.

2. Explain how to calculate the *mean*, $\bar{x}$. To find the *mean* of a set of observations, add their values and divide by the number of observations. $\bar{x} = \frac{1}{n}\sum x_i$

3. Explain how to calculate the *median*, *M*. The *median* M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median:
   1) arrange all the observations in order of size, from smallest to largest
   2) if the number of observations is odd, the median M is the center observation of the order list.
   3) if the number of observations is even, the median M is the mean of the 2 center observations
   in the ordered list.

4. Explain why the median is *resistant* to extreme observations, but the mean is *nonresistant*. The median is *resistant* because it is only based on the middle one or two observations of the ordered list. The mean is sensitive to the influence of a few extreme observations. Even if there are no outliers a skewed distribution will pull the mean toward the long tail.

5. In statistics, what is meant by *spread*? *Spread* is a way to measure the variability of the observations around the center. One of the most common ways to measure spread is to calculate the range of the data. The range is obviously sensitive to extreme measures.

6. Explain how to calculate Q_1 and Q_3. To calculate the quartiles:
   1) arrange the observations in increasing order and locate the median M in the list.
   2) Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
   3) Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

7. What is the *interquartile range?* The *IQR* is the distance between the first and third quartiles, $Q_3 - Q_1$.

8. How can we use IQR to determine outliers? An observations is an *outlier* it if is more than 1.5*IQR above the third quartile of below the first quartile.

9. What is the *five-number summary*? The *5 # summary* is Minimum, Q_1, Median, Q_3, Maximum.

10. How do we use the five-number summary to make a modified boxplot? A modified boxplot is a graph of the 5-number summary, with outliers plotted individually.
    - a central box spans the quartiles
    - a line in the box marks the median
    - observations more than 1.5*IQR outside the central box are plotted individually
    - lines extend from the box out to the smallest and largest observations that are not outliers.

11. What does *standard deviation* measure? How do we calculate it? The *standard deviation* is a measure of spread. It measures spread around the mean and should only be used when the mean is chosen as the measure of center.

$$s = \sqrt{\frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2}$$

12. What is the relationship between *variance* and *standard deviation*? The standard deviation s is the square root of the variance s^2.

13. When does *standard deviation* equal zero? The *standard deviation = 0* only when there is *no spread*. This happens only when all observations have the same value. Otherwise $s > 0$. As the observations become more spread out about their mean, s gets larger.

14. Is *standard deviation* resistant or nonresistant to extreme observations? Explain. s, like the mean, is not resistant. Strong skewness or a few outliers can make s very large.

Chapter 2: The Normal Distributions

2.1 Density Curves and the Normal Distributions (pp. 78-92)

1. What is a *density curve*? A *density curve* is a curve that
   1) is always on or above the x-axis
   2) has area exactly 1 underneath it.

2. What does the area under a *density curve* represent? The *area* under the curve and above any range of values is the proportion of all observations that fall in that range.

3. Where is the median of a *density curve* located? The *median* is the equal-areas point. Half the area under the curve is to the left, the other half of the area is to the right.

4. Where is the mean of a *density curve* located? The *mean* is the balance point of the density curve. The mean and median are the same for a symmetric density curve.

5. What is the difference between the *randInt* and *rand* commands on the TI-84? randint(a, b) returns a random integer between a and b.
   rand returns a random number between 0 and 1.
   rand (n) returns n random numbers between 0 and 1.
   randint(a, b, n) returns n random integers between a and b.

6. How would you describe the shape of a *normal curve*? Draw several examples. Symmetric, single-peaked, bell-shaped.

7. Where on the *normal curve* are *inflection points* located? The *inflection points* are at $\pm \sigma$. (These are the points where the curve changes concavity.)

8. Explain the *68-95-99.7 rule*. This rule states that for a normal curve, 68% of the data lies between ± 1 σ, 95% of the data lies between ± 2 σ, and 99.7% of the data lies between ± 3 σ.

9. What is a *percentile*? The *pth percentile* of a distribution is the value such that p% of the observations fall at or below it.

10. Is there a difference between the 80^{th} percentile and the top 80%? Explain. Yes. The 80^{th} percentile means 80% of the data values are equal or below. The top 80% means 80% of the values are equal or above.

11. Is there a difference between the 80^{th} percentile and the lower 80%? Explain. No, these are the same.

Percentiles are used when we are most interested in seeing where an individual observation stands relative to the other individuals in the distribution.

2.2 Standard Normal Calculations (pp. 93-111)

1. Explain how to *standardize* a variable. z-score = $\frac{x-\mu}{\sigma}$

2. What is the purpose of standardizing a variable? A standardized value tells us how many standard deviations the original observation falls away from the mean, and in which direction.

3. What is the *standard normal distribution*? A normal distribution with mean 0 and standard deviation 1.

4. What information does the *standard normal table* give? The table entry for each value z is the area under the curve to the left of z.

5. How do you use the standard normal table (Table A) to find the area under the standard normal curve to the left of a given *z-value*? Draw a sketch. What calculator syntax would you use to find this value?

   normalcdf *(-10, a)*

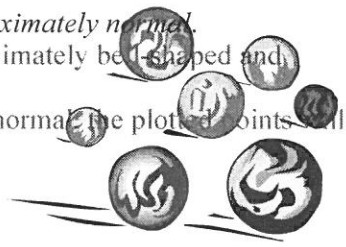
6. How do you use Table A to find the area under the standard normal curve to the right of a given *z-value*? Draw a sketch. What calculator syntax would you use to find this value?

   normalcdf *(a, 10)*

7. How do you use Table A to find the area under the standard normal curve between two given *z-values*? Draw a sketch. What calculator syntax would you use to find this value?

   normalcdf *(a, b)*

8. Describe two methods for assessing whether or not a distribution is *approximately normal*.
   1) construct a frequency histogram or stemplot. See if the graph is approximately bell-shaped and symmetric about the mean.
   2) Construct a normal probability plot. If the data distribution is close to normal, the plotted points will lie close to a straight line. (Nonnormal data will show a nonlinear trend)



Chapter 3 Notes: Examining Relationships
Answers

1. What is the difference between a **response variable** and an **explanatory variable**?
   The explanatory variable is your x variable. It is the variable that you are using to predict you y or response variable.

2. How are response and explanatory variables related to **dependent** and **independent** variable?
   Your prediction for the response variable depends on what your independent or explanatory variable is.

3. When is it appropriate to use a scatterplot to display data?
   When you have two groups of quantitative data.

4. Which variable always appears on the horizontal axis of a scatterplot?
   Explanatory variable

5. You can describe the overall pattern of a scatterplot by the ....
   Strength, Direction, and Form

6. Explain the difference between a **positive association** and a **negative association.**
   Two variables are positively associated when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.
   Two variables are negatively associated when above-average values of one tend to accompany below-average values of the other, and vice versa.

7. How can quantitative data which belongs to different categories be differentiated on a scatterplot?
   Use different colors for the points.

8. What does **correlation** measure?
   Correlation measures how well the data can be modeled by a linear relationship.

9. Explain why two variables must both be **quantitative** in order to find the correlation between them.
   Because the formula requires a standard deviation and mean for both variables, and categorical data does not have a measure of spread and center.

10. What is true about the relationship between two variables if the relationship is:
    a) Near 0? No linear relationship
    b) Near 1? The relationship is positive and could almost be perfectly be modeled by a line.
    c) Near -1? Negative
    d) Exactly 1? All the points would fall exactly on a line with positive slope
    e) Exactly -1? All the points would fall exactly on a line with negative slope

11. Is correlation resistant to extreme observations?
    No. Outliers can greatly change the value of *r*.

12. What does it mean if two variables have high correlation? It means that one the explanatory variable does a very good job of predicting what the response variable would be. The relationship follows a linear model very well and we can use a linear equation to make predictions about what the response variable could be based on a given explanatory.

13. What does it mean if two variables have weak correlation?
A linear model does not do a very good job of explain the relationship between the two variables.

14. What does it mean if two variables have no correlation?
A linear model would be useless for describing the relationship.

3.2 Least-Squares Regression (pp. 199-233)

1. In what way is a regression line a mathematical model?
It describes how a response variable y changes as an explanatory variable x changes. A regression line can be used to predict the value of y for a given value of x.
2. What is extrapolation and why is this dangerous?
Extrapolation is predicting a y value by extending the regression model to regions *outside* the range of the x-values of the data. It's dangerous because it introduces the questionable and untested assumption that the relationship between x and y does not change.

3. What is a least-squares regression line?
A *least-squares regression line* is a regression line that minimizes the sum of the squares of the residuals (the vertical distances of the data points from the line)

4. What is the formula for the equation of the least-squares regression line?

$$\hat{y} = a + bx, \text{ with slope } b = r\frac{s_y}{s_x} \text{ and intercept } a = \bar{y} - b\bar{x}$$

5. The least-squares regression line always passes through the point…
The grand mean $(\bar{x}, \bar{y})$

6. What is a residual?
A residual is the difference between an observed value of the response variable and the value predicted by the regression line. observed y – predicted y $= y - \hat{y}$

7. How can you calculate residuals on your calculator and use this to produce a residual plot?
To calculate residuals on your calculator: With data in Lists 1 and 2, enter y hat in highlighted L3 but replace the x in y hat with (L1) and press enter. Next with L4 highlighted, enter L2 minus L3 and press enter to see the residuals displayed in L4. To produce a residual plot: Create (turn on) a plot with L1 as the x variable and L4 as the y variable. Use ZOOM 9 to see the residual plot. NOTE: To compute the standard deviation of the residuals, calculate a 1-VAR STAT on the residual list.

8. If a least-squares regression line fits the data well, what characteristics should the residual plot exhibit?
Residual plots help us assess how well a regression line fits the data. If the points in a residual plot are randomly dispersed around the horizontal axis, then a linear regression model is appropriate for the data; otherwise a non-linear model is appropriate.

Chapter 3 Notes: Examining Relationships

9. How is the coefficient of determination defined?
   As r^2 – which is the square of the correlation coefficient, r. The coefficient of determination is equal to the percent of variation in one variable that is accounted for (predicted) by the other variable.

10. What is the formula for calculating the coefficient of determination?
11. Find r: $$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{[n(\Sigma x^2) - (\Sigma x)^2]\ [n(\Sigma y^2) - (\Sigma y)^2]}$$

    Then square the answer....DONE ON YOUR CALCULATOR!

12. if $r^2 = 0.95$, what can be concluded about the relationship between x and y?
    95% of the variation in y is explained by the linear model relating y to x. Remember, r^2 is a reported measure of how successful the regression model was in explaining the response variable.

3.3 Correlation and Regression Wisdom (pp. 233-251)

1. What are three limitations of correlation and regression?

1. The distinction between explanatory and response variables is important in regression; reversing x and y will yield a different least-squares regression line.
2). Correlation and regression lines describe only linear relationships.
3). Correlation and least-squares regression lines are not resistant.

2. Under what conditions does an outlier become an influential observation?

An observation is influential for a statistical calculation if removing it noticeably changes the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line. Influential points often have small residuals, because they pull the regression line toward themselves.

3. What is a lurking variable?

A hidden variable in a study, research, experiment, etc. that may affect the predictor variables given. In other words, a lurking variable is not among the explanatory or response variables in a study, but it may influence the variation in the response variable.

4. Why does association not imply causation?

A strong association between two variables is not enough evidence to draw conclusions about cause and effect.

Name _____KEY_____ Date _____ Block _____

Chapter 3: Describing Relationships

Reading Guide

Key Vocabulary:

- response variable
- explanatory variable
- independent variable
- dependent variable
- scatterplot
- positive association
- negative association
- linear
- correlation
- r-value
- regression line
- mathematical model
- least-squares regression line
- $\hat{y}$ "y-hat"
- SSM
- SSE
- r^2
- coefficient of determination
- residuals
- residual plot
- influential observation

Calculator Skills:

seq(X,X,min,max,scl)
2-Var Stats
sum
Diagnostic On
$\bar{x}, s_x, \bar{y}, s_y$
Clear All Lists
residual plot

3.1 Scatterplots and Correlation (pp.142-163)

1. What is the difference between a *response variable* and an *explanatory variable*? A response variable measures an outcome of a study. An explanatory variable may help explain or influence changes in a response variable.

2. How are response and explanatory variables related to *dependent* and *independent* variables? Explanatory variables are often referred to as independent variables and response variables are often referred to as dependent variables.

3. When is it appropriate to use a *scatterplot* to display data? When the goal is to display or compare the relationship of two quantitative variables.

4. Which variable always appears on the horizontal axis of a scatterplot? The explanatory variable

5. You can describe the overall pattern of a scatterplot by the direction, form, and strength.

6. Explain the difference between a *positive association* and a *negative association.* When the overall pattern moves from upper left to lower right, it's called a negative association. When the overall pattern moves from lower left to upper right, it's called a positive association.

7. What does *correlation* measure? Correlation measures the strength of the linear relationship between two quantitative variables.

8. What is true about the relationship between two variables if the *r-value* is:

   a. Near 0? - very weak to almost no linear relationship

   b. Near 1? - very strong to almost perfectly linear relationship

   c. Near -1? – very strong to almost perfectly linear relationship

   d. Exactly 1? – perfect linear relationship

   e. Exactly -1? – perfect linear relationship

9. Is *correlation* resistant to extreme observations? Explain. Correlation is not resistant to extreme observations, therefore outliers can greatly change the value of the correlation.

10. What does it mean if two variables have *high correlation*? If there is a high correlation, then the two variables closely agree and the scatter plot of the quantitative variables tend to follow a moderate to strong linear pattern.

11. What does it mean if two variables have *weak correlation*? If there is a weak correlation, then the two quantitative variables do not closely agree and the scatter plot of the variables tend to follow a scattered or non-linear pattern.

12. What does it mean if two variables have *no correlation*? If there is no correlation, then the strength of a linear relationship between two quantitative variables is non-existent.

13. Explain why two variables must both be *quantitative* in order to find the *correlation* between them. Scatterplots are the only choice for displaying the relationship between two quantitative variables. Quantitative variables are needed because correlation is measured numerically using the following interval: $-1 \leq r \leq 1$

14. Does a correlation close to -1 or 1 always guarantee a linear relationship? Explain. No. A scatterplot with a correlation that's close to -1 or 1 can have a curved form so it is necessary to always plot the given data. It is important to remember that you can calculate a correlation for any scatter plot; however r only measures straight-line or linear relationships.

3.2 Least-Squares Regression (pp.164-197)

1. What is a regression line? A straight line that describes how a response variable y changes as an explanatory variable x changes. Basically, it is a line used to model the data.

2. What is *extrapolation* and why is this dangerous? Extrapolation involves using a regression line to make predictions far outside the range of values of x that are used to obtain the line. Proceed with caution when making these predictions – often times they are not accurate.

3. What is a *least-squares regression line*? The least-squares regression line of y on x is a type of regression line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. (also called 'line of best fit') In other words, a least-squares regression line makes the errors in predicting y as small as possible by minimizing the sum of the squares of the residuals.

4. What is the formula for the equation of the *least-squares regression line*? Define each variable. Least-squares line: $\hat{y} = a + bx$; y-hat is the predicted value of y given a specific x; a is the y intercept – the predicted value of y when x is 0; b is the slope – the amount by which y is predicted to change when x increases by 1 unit.

5. The *least-squares regression line* always passes through the point **(x bar, y bar).**

6. What is a *residual*? A residual is the difference between an observed value of y and the value of y predicted by the regression line; i.e. **(residual = observed y – predicted y = y - $\hat{y}$)** In essence, a residual is an observable estimate of the unobservable statistical error.

7. What special property do the residuals have? The mean of the residuals will always equal zero. Why do they have this property? Because the sum of the residuals always equal zero.

8. What is a residual plot? A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. In other words, it is a scatter plot of the residuals against x (the explanatory variable) and turns the regression line horizontal.

9. How can you calculate *residuals* on your calculator and use this to produce a *residual plot*? To calculate residuals on your calculator: With data in Lists 1 and 2, enter y hat in highlighted L3 but replace the x in y hat with (L1) and press enter. Next with L4 highlighted, enter L2 minus L3 and press enter to see the residuals displayed in L4. To produce a residual plot: Create (turn on) a plot with L1 as the x variable and L4 as the y variable. Use ZOOM 9 to see the residual plot. NOTE: To compute the standard deviation of the residuals, calculate a 1-VAR STAT on the residual list.

10. If a *least-squares regression line* fits the data well, what characteristics should the *residual plot* exhibit? Residual plots help us assess how well a regression line fits the data. If the points in a residual plot are randomly dispersed around the horizontal axis, then a linear regression model is appropriate for the data; otherwise a non-linear model is appropriate.

11. What does the standard deviation of the residuals tell us? The standard deviation is roughly the average distance of the actual values form the least-squares regression line. It tells us the average distance of actual values from their expected values and thus measures prediction error.

12. How is the *coefficient of determination* defined? As r^2 – which is the square of the correlation coefficient, r. The coefficient of determination is equal to the percent of variation in one variable that is accounted for (predicted) by the other variable.

13. If $r^2 = 0.95$, what can be concluded about the relationship between x and y? 95% of the variation in y is explained by the linear model relating y to x. Remember, r^2 is a reported measure of how successful the regression model was in explaining the response variable.

14. What are three limitations of correlation and regression? 1. The distinction between explanatory and response variables is important in regression; reversing x and y will yield a different least-squares regression line. 2). Correlation and regression lines describe only linear relationships. 3). Correlation and least-squares regression lines are not resistant.

15. Under what conditions does an outlier become an *influential observation*? An observation is influential for a statistical calculation if removing it noticeably changes the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line. Influential points often have small residuals, because they pull the regression line toward themselves.

16. What is a *lurking variable*? A hidden variable in a study, research, experiment, etc. that may affect the predictor variables given. In other words, a lurking variable is not among the explanatory or response variables in a study, but it may influence the variation in the response variable.

17. Why does *association* not imply *causation*? A strong association between two variables is not enough evidence to draw conclusions about cause and effect.