**FORMULAS:**

 OR 

(Formula Difference is in denominator – sq, roots then mult, vs. mult. then sq. root)

 

**STAT ESSENTIALS:** Be able to: 1) go through the steps leading to correlation and regression; 2) Conducting a correlation by hand and via SPSS; 3) Scatter Plot – building and interpreting; 4) Basics of Regression. **[NOTE: see Orange sheets for additional problems and resources.]**

**PROBLEMS:**

**1)** Identify steps that lead to regression:

1. Does the relationship make sense?
2. Draw a picture – here, a scatter plot
3. State your hypotheses
4. Calculate the correlation coefficient, r
5. Determine if r is statistically significant. If yes, regression. If no, fallback statistic is the y mean
6. Determine the line of best fit (least squares) equation

**2)** When should you not conduct regression? If the correlation is not statistically significant, no regression.

**3)** If a correlation is not statistically significant, what statistic represents your “best guess” estimate of a y variable for a given x variable? Mean of the Y variable.

**4)** What the … What type of relationship is this? **Spurrious** (one word will do here)



**5)** Located below are data collected for three paired variables. For which pairs would it appear appropriate to obtain a correlation coefficient? How did you figure this out?



If it makes sense to look at the relationship, draw a picture.





**6)** What is the value of “n” for the relationship between Exercise & Milk Consumption in the prior problem? 9

**7)** Assume that there is a statistically significant correlation (r = .89) between the length of new born female infants and their birth weight. Using length as the X variable, determine the coefficient of determination AND explain what the heck this means in this relationship. (the internet may help here)

r2 = (r\*r)\*100 = 79.21% The coefficient of determination in this relation represents the percent of variation in the y variable (weight) from its mean (y-bar) that may be explained by the linear relationship between length (X) and weight (Y).

**8)** To use a least squares equation to estimate values outside the range of data values is called exsanguination. Wait a minute, I know from countless CSI programs that that isn’t right. So what is it called? Extrapolation. You are assuming that the observed relationship extends beyond available data. That may not be the case. Therefore, regression generally should not be used to predict y values outside the range of the x variable.

**9)** Age & Exercise







Correlation Significance: from table,  = .05 [p-value .040]

d) x = 35: 10.499-.17997(35 years)=4.2 hours

 when x = 64: outside data range.

**10)** A medical researcher wishes to describe the relationship between the prescription cost of a brand name drug and its generic equivalent. The data (in dollars) are shown.





 A)



a) Correlation Significance: from table  = .01 [p = .006]

d) Estimate of y => -4.1475+.382(30) = **$7.31**

**11)** Given a line of best fit equation for the number of fires and acres burned is

Is the correlation between these two variables positive or negative? How do you know? Look at the sign of the slope. Positive as the slope is positive (+1.0297).

**12)** Hours Online The number of hours 12 students spent online during the weekend and the scores of each student who took a test the following Monday are shown below.











**Corr. Sig: from table  - .01 [p-value .001]**

**Estimates of Test Scores:**



**13)** Data on tornadoes and deaths are shown below.





A)



B)

Not a significant correlation at  = .05, therefore regression should not be done if that level was being used.

C) if done at  = .10, then could do regression. [ rest based on this alpha level]







D) If x = 1000: