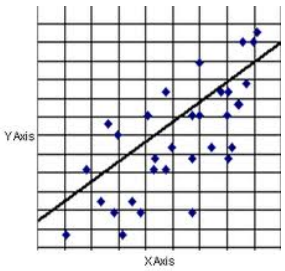
AP Statistics Summer Work Name _____
Correlation and Regression study



Correlation seeks to find a numerical measure of the strength of the relationship between two numerical variables. Linear regression attempts to model the relationship between these variables, once correlation has been established. Some variables have positive correlation, meaning as one increases the other does as well; for example height and shoe size – generally taller people have larger shoes. Some variables have negative correlation, meaning as one increases the other decreases; for example age of a car and its value – a car depreciates each year. Some variables may not be related. The graph on the right shows an example of what positive correlation looks like. Notice that the slope of the line is positive, but the points don't necessarily all lie on the line. The strength of the relationship is measured by proximity to the line that best fits the data. The strength is measured by the correlation coefficient r and the linear regression equation is of the form $\hat{y} = ax + b$.

Choose two numerical variables that you believe might be related.

_____ and _____

Examples. Use Google to search:

A type of food like candy/cereal/fast food: grams of sugar vs calories, fat grams vs grams of protein

Sports data: minutes played vs. points scored, years as a team vs championships won

Prices: # in package vs price per item or serving size vs price

Students: GPA vs ACT or SAT score, # of applicants vs # accepted

Cars: weight vs. horsepower of engine, age vs value

Jobs: years of college required vs avg starting salary

Music: # of weeks on top 40 chart vs highest rank achieved or # of copies sold

Any other pair of variables that sparks your interest……..

****Print your data and attach to this packet. Give the website source of your data:**

_____

What type of relationship / correlation do you think they have (positive / negative) and why?

Research / collect thirty pairs of data points on-line or in person. Tech savvy students can recreate this chart in Excel and print if you'd prefer.

	X =	**Y=**	XY	X^2	Y^2
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					

	X =	Y=	XY	X^2	Y^2
24					
25					
26					
27					
28					
29					
30					
Totals sums					

Remember that Σ means sum (total by adding). n is the # of points (so 30)
Formula for r

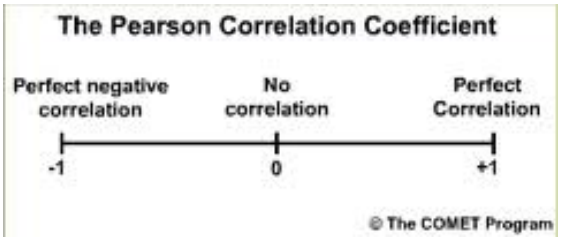
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Equation for correlation coefficient

Calculate r, the correlation coefficient. **Correlation scale**:

r=_____

The answer to r MUST be between -1 and 1 or you have an error in your calculations.



Between -1 and -.5 there is exists negative correlation, between -.5 and .5 there is no correlation, and between .5 and 1 there is positive correlation.

Type of correlation in your data _____

Meaning of correlation in terms of your two variables written in a sentence.

Use the following formulas for regression to find the slope and the intercept of the regression equation.

Slope = $\frac{(n\sum xy - \sum x \sum y)}{(n\sum x^2 - (\sum x)^2)}$ Intercept = $\frac{(\sum y - slope \sum x)}{(n)}$

Slope _____. Intercept _____

Equation in $\hat{y} = ax + b$ form: _____

Create a scatter plot of your data. You can use Excel or a similar computer program to create your graph and print it or create by hand on graph paper.

****Be sure to title and label your graph axes and graph your regression equation.**

The closer the r is to +1 or -1, the stronger the correlation and closer the points are to a straight line. When the correlation is strong, the accuracy of predictions based on the regression equation is high.

Accuracy of predictions_____ because _____

Use your equation to predict values of y for two different x values. Be sure to choose values inside range of original data set x values. Show your work.

1. For x = _____

2. For x = _____

We will be learning how to calculate r and the regression equation as well as graph on our graphing calculators during the first week of class so you can check your work. If you want to get a jump start check out entering data into lists, creating a scatterplot, graphing lines, and calculating r in the owner's manual that comes with your calculator or my calculator cheat sheet posted on AP Stats part of website.

In AP, it is necessary to write about your findings. Using a few sentences, address what you found out about your data and what you learned by doing this study?