CORRELATION

- Correlation coefficient: statistical index of the degree to which two variables are associated, or related.

- We can determine whether one variable is related to another by seeing whether scores on the two variables *covary*---whether they vary together.

EXAMPLE OF CORRELATION

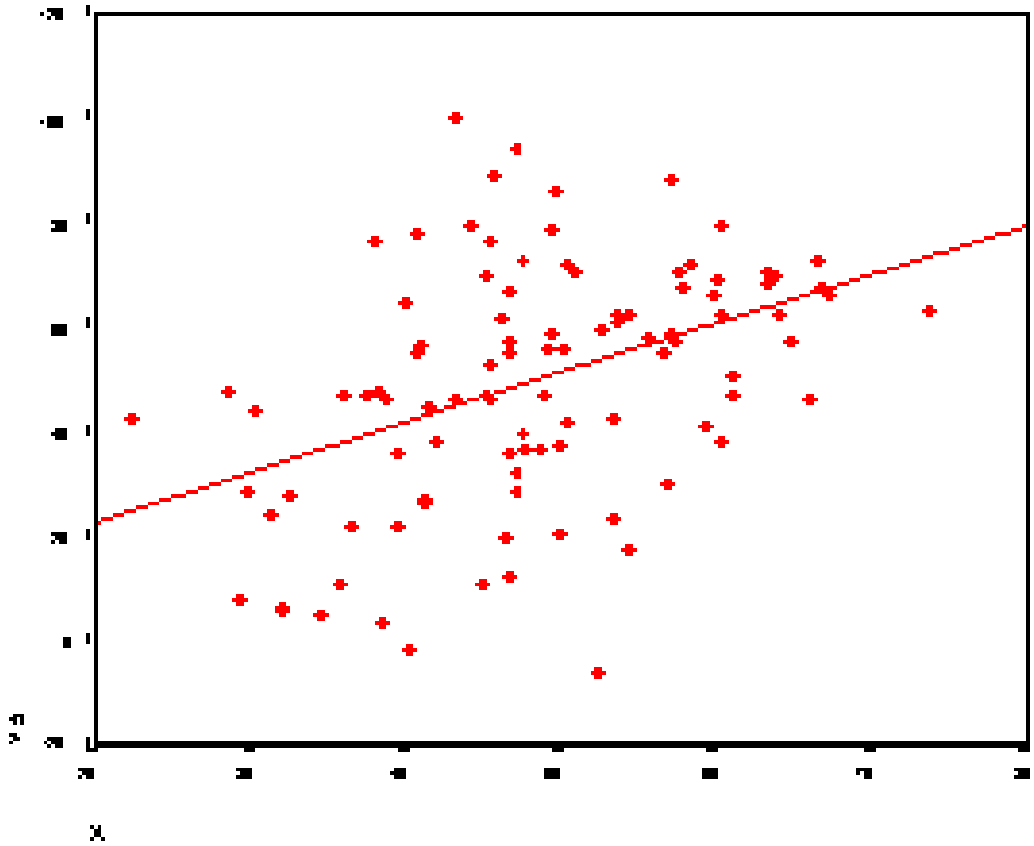
Is there an association between:

- Advertisement and sales?
- Speed and Distance
- Preparation Hours for exam and Marks obtained
- Children's IQ and Parents' IQ
- Number of Employees and Productivity?

SCATTERPLOT

- The relationship between any two variables can be portrayed graphically on an x- and y- axis.
- Each subject i_1 has $(x_{1,} y_{1)}$. When score s for an entire sample are plotted, the result is called **scatter plot**.

- Scatterplot



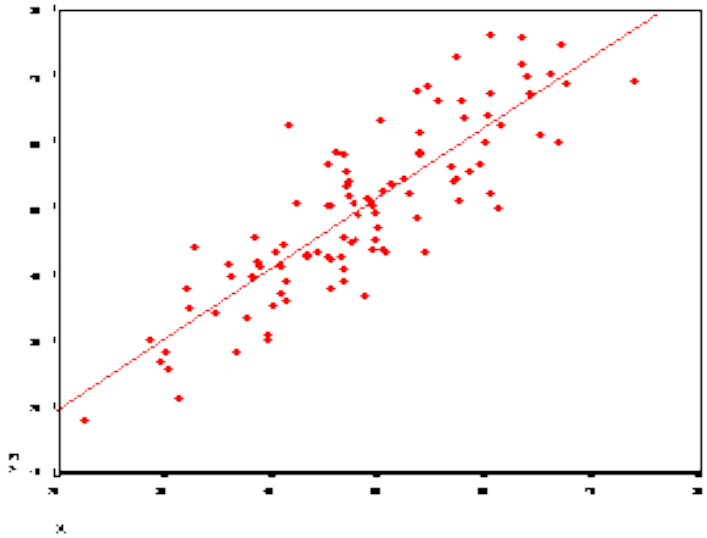
DIRECTION OF THE RELATIONSHIP

Variables can be positively or negatively correlated.

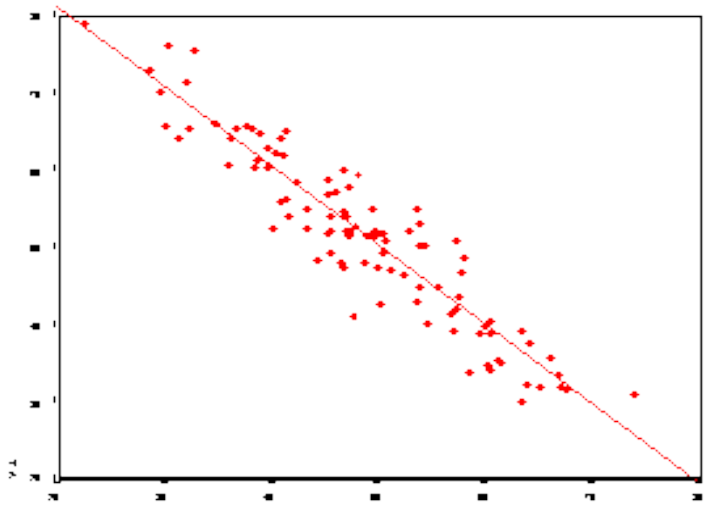
- Positive correlation: A value of one variable increase, value of other variable increase.

- Negative correlation: A value of one variable increase, value of other variable decrease.

r = .85



r = -.94

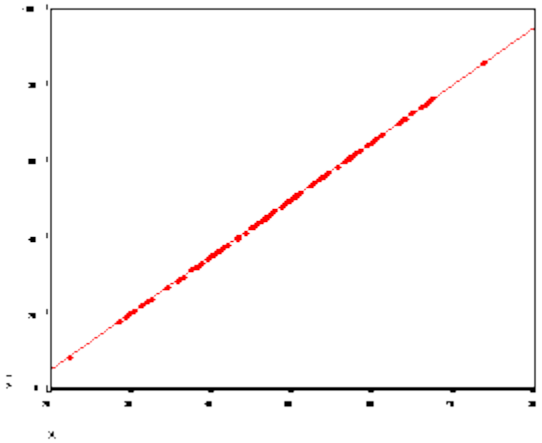


STRENGTH OF THE RELATIONSHIP

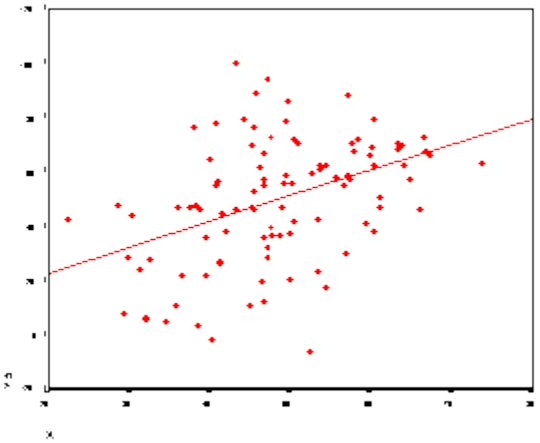
The magnitude of correlation:

- Indicated by its numerical value
- ignoring the sign
- expresses the strength of the linear relationship between the variables.

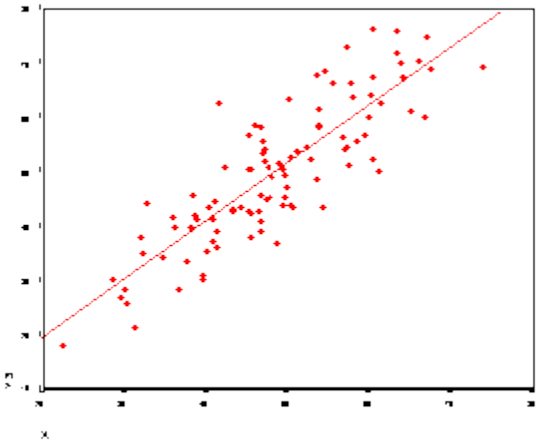
r =1.00



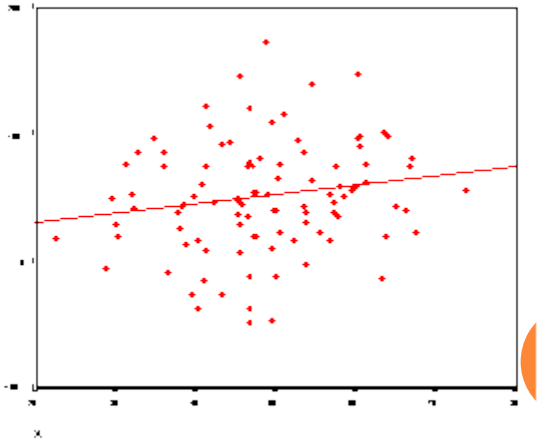
r = .42



r =.85



r =.17



PEARSON'S CORRELATION COEFFICIENT

There are many kinds of correlation coefficients but the most commonly used measure of correlation is the Pearson's correlation coefficient. (r)

- The Pearson r range between -1 to +1.
- Sign indicate the direction.
- The numerical value indicates the strength.
- Perfect correlation : -1 or 1
- No correlation: 0
- A correlation of zero indicates the value are not linearly related.
- However, it is possible they are related in **curvilinear** fashion.

STANDARDIZED RELATIONSHIP

- The Pearson r can be thought of as a standardized measure of the association between two variables.
- That is, a correlation between two variables equal to .64 is the same strength of relationship as the correlation of .64 for two entirely different variables.
- The metric by which we gauge associations is a standard metric.
- Also, it turns out that correlation can be thought of as a relationship between two variables that have first been standardized or converted to z scores.

$$r = \frac{\sum Z_x Z_y}{N - 1}$$

CORRELATION REPRESENTS A LINEAR RELATIONSHIP

- Correlation involves a linear relationship.
- "Linear" refers to the fact that, when we graph our two variables, and there is a correlation, we get a line of points.
- Correlation tells you how much two variables are *linearly related*, not necessarily how much they are related in general.
- There are some cases that two variables may have a strong, or even perfect, relationship, yet the relationship is not at all linear. In these cases, the correlation coefficient might be zero.

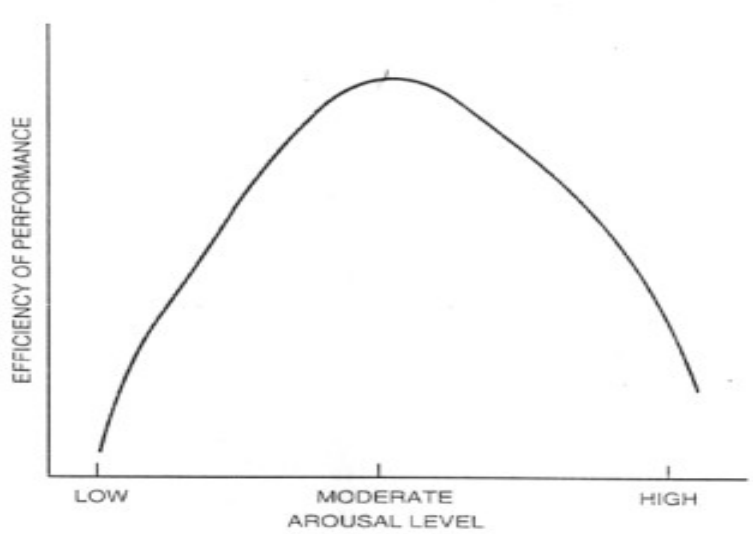
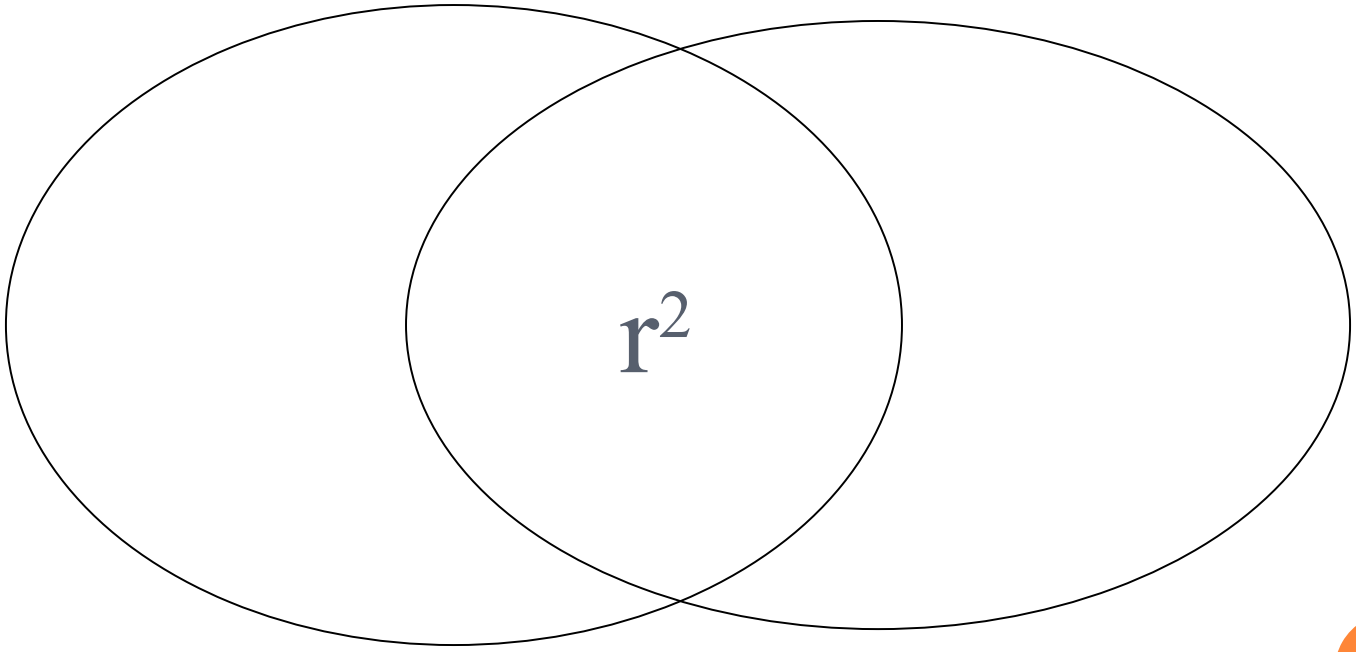


Figure 4-5 Hypothesized relationship between performance efficiency and level of arousal, illustrating a curvilinear relationship.

COEFFICIENT OF DETERMINATION R^2

- The percentage of shared variance is represented by the square of the correlation coefficient, r^2 .
- Variance indicates the amount of variability in a set of data.
- If the two variables are correlated, that means that we can account for some of the variance in one variable by the other variable.

COEFFICIENT OF DETERMINATION R^2



STATISTICAL SIGNIFICANCE OF R

- A correlation coefficient calculated on a sample is statistically significant if it has a very probability of being zero in the population.
- In other words, to test r for significance, we test the null hypothesis that, in the population the correlation is zero by computing a t statistic.
- Ho: r = 0
- H_A: r = 0

SOME CONSIDERATION IN INTERPRETING CORRELATION

1. Correlation represents a linear relations.

- Correlation tells you how much two variables are linearly related, not necessarily how much they are related in general.
- There are some cases that two variables may have a strong perfect relationship but not linear. For example, there can be a curvilinear relationship.

SOME CONSIDERATION IN INTERPRETING CORRELATION

2. Restricted range (Slide: Truncated)

- Correlation can be deceiving if the full information about each of the variable is not available. A correlation between two variable is smaller if the range of one or both variables is truncated.
- Because the full variation of one variables is not available, there is not enough information to see the two variables covary together.

SOME CONSIDERATION IN INTERPRETING CORRELATION

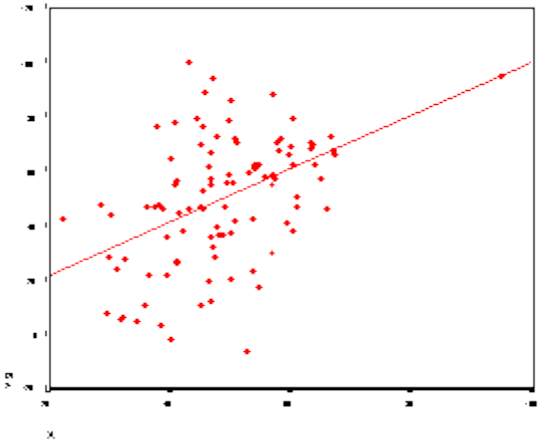
3. Outliers

- Outliers are scores that are so obviously deviant from the remainder of the data.

- On-line outliers ---- artificially inflate the correlation coefficient.
- Off-line outliers --- artificially deflate the correlation coefficient

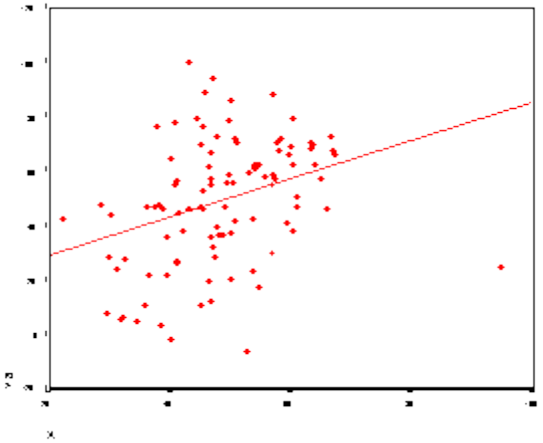
ON-LINE OUTLIER

- An outlier which falls near where the regression line would normally fall would necessarily increase the size of the correlation coefficient, as seen below.
- r = .457



OFF-LINE OUTLIERS

- An outlier that falls some distance away from the original regression line would decrease the size of the correlation coefficient, as seen below:
- r = .336



CORRELATION AND CAUSATION_

- Two things that go together may not necessarily mean that there is a causation.
- One variable can be strongly related to another, yet not cause it. Correlation does not imply causality.

- When there is a correlation between X and Y.
- Does X cause Y or Y cause X, or both?
- Or is there a third variable Z causing both X and Y , and therefore, X and Y are correlated?