**Basics of Statistics and Key Tools / Programs for Labs**

***\*\* For tools, see separate doc, especially for Python and R.***

**Basics of Statistics:**

1. **Mean – sum of all the numbers divided by the total number in the sample.**
2. **Variability –** refers to how "spread out" a group of scores is. See below:

These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

****

1. **Variance:**

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the variance is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 1. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is simply the previous column squared.



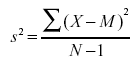
The formula for the variance is (this is for a sample):

http://onlinestatbook.com/2/summarizing_distributions/graphics/pop_var.gif

If the variance in a sample is used to estimate the variance in a

population, then the previous formula underestimates the variance and

the following formula should be used:



where s2 is the estimate of the variance and M is the sample mean. Note that M is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample, this formula is most often used.

1. Standard Deviation (Sx): It is just the square root of the variance.
2. **Correlation r:**

We are going to compute the correlation between the variables X and Y

shown in Table 1 Below (NOT Table. 1 above). We begin by computing the mean for X and subtracting

this mean from all values of X. The new variable is called "x." The variable

"y" is computed similarly. The variables x and y are said to be *deviation*

*scores* because each score is a deviation from the mean. Notice that the

means of x and y are both 0. Next we create a new column by multiplying

x and y.

Before proceeding with the calculations, let's consider why the sum of the

xy column reveals the relationship between X and Y. If there were no

relationship between X and Y, then positive values of x would be just as

likely to be paired with negative values of y as with positive values. This

would make negative values of xy as likely as positive values and the sum

would be small. On the other hand, consider Table 1 in which high values

of X are associated with high values of Y and low values of X are associated

with low values of Y. You can see that positive values of x are associated

with positive values of y and negative values of x are associated with

negative values of y. In all cases, the product of x and y is positive, resulting

in a high total for the xy column. Finally, if there were a negative relationship

then positive values of x would be associated with negative values of y and

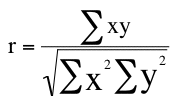
negative values of x would be associated with positive values of y. This

would lead to negative values for xy.

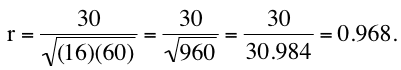
Table 1 - A. Calculation of r.



Pearson's r is designed so that the correlation between height and weight is the same whether height is measured in inches or in feet. To achieve this property, Pearson's correlation is computed by dividing the sum of the xy column (Σxy) by the square root of the product of the sum of the x2 column (Σx2) and the sum of the y2 column (Σy2). The resulting formula is:



and therefore



An alternative computational formula that avoids the step of computing deviation scores is:

