**Correlation Analysis**

**Purpose**

Correlation analysis tells about the linear relationship between two variables. A linear relationship is one that goes up or down when the other variable goes up or down. For instance, a positive linear relationship (or correlation) is one in which as one variable increases, so does the other. For example, as staff’s social/emotional skills increase, children’s assessment results for the social/emotional domain increase. An inverse relationship, in contrast, is one in which directions are opposing: as one variable increases, the other decreases, and vice versa. For example, as staff’s social/emotional skills increase, children’s externalizing behaviors decrease. Because it is a calculation of a linear relationship between things have numerical value, a correlation analysis must be used on numbers rather than qualitative data such as parent comments or other themes.

There are many common numeric correlations in the field of early childhood. One common correlation is between family income (a dollar amount) and assessment scores (numeric): as income goes up, so often too do assessment results such as kindergarten entry scores or other learning assessment results.

Correlation analysis helps make plain the important relationship between variables in a statistical way. Correlations can tell the direction of that relationship (positive or negative) as well as the strength of that relationship. More on the direction and strength of the relationship is described in the example below. Running a correlation can confirm whether something that staff perceive to be true can be backed by data. Having correlation data in hand can help programs understand a problem or pattern more fully and begin to identify ways to address it through changes to programs.

**How it Works, What it Means**

Understanding the relationship between two variables can be as simple as plotting X and Y data on grid paper or as complex as running a correlation statistic using statistical software. (Ironically, the statistical software works so quickly, that it has become a very easy process compared to pencil and paper.) Exel and similar programs can be used to run a correlation. However, it is helpful to describe the concept of the correlation by thinking about plotting data points on an X and Y axis and viewing scatterplots.

**Early Childhood Education Examples**

A program has data on children’s kindergarten entry profile (KEP) scores and family income in relation to Federal Poverty Guidelines. The KEP scores are 1, 2, or 3 for Not Yet, In Process, and Proficient, respectively. The data are loaded into Excel:



You suspect that there is a positive correlation between the child’s income and their KEP scores. In Excel, select the entire cell range:



Then select the scatterplot option from the **Insert** menu:



Selecting the upper-left option produces a chart like this. A trend line has also been added.



A trend line tells you what direction the relationship is between the two types of numerical data. The positive linear relationship between increasing KEP scores and increasing family income is clear. (If the relationship was negative, or inverse, the line would be “high to low” from left to right.)

If in this same database your program collects data on the number of social services programs each of the children’s families are accessing your program may be interested in knowing whether there is a relationship between family income and the number of programs being used. The data may look something like this:



The scatterplot for this variable relationship looks like this:



Trend line

This is an inverse correlation: as income goes up, the number of social services programs accessed goes down and as described above, a “negative” correlation.

When there is little or no relationship between the two variables, the slope of the trend line will be more level.

*The Correlation Coefficient*

Let’s say that seeing the relationship between the two variables of interest is useful, but the program would like to know the magnitude of that correlation. This requires running a statistic called a correlation coefficient. The correlation coefficient, or *r* value, lies between -1 and 1. If *r* is close to 0, the two variables are not related in a linear fashion and the trend line would be close to level. If the *r* value is close to -1 or 1, the relationship is strong and the trend line will have a slope up, or down. A negative value indicates an inverse relationship as is observed above in the family income and social services programs example.

To run a correlation coefficient, select **Formulas** on the toolbar, **Statistical formulas**, and **Pearson** (for Pearson Correlation Coefficient):



You will next need to enter the cell ranges of the two “arrays” – these are the ranges where the two types of data exist in the spreadsheet. Array 1 is for Income and Array 2 is for KEP:



The r value is shown in the window. It may also appear on the spreadsheet.

This result indicates that the relationship between family income and KEP is strong and positive (recall that a perfect position correlation is 1 and our result is .899).

**Remember**

* It is very important to visualize the data as well as run a correlation coefficient. Seeing the data helps you to understand the relationship between the variables. In some cases, a relationship is curvilinear or the “scatter” is very dispersed, which indicates a weak relationship.
* Correlation is not the same as a regression: regression requires a different set of statistical approaches. Correlations are often run prior to doing a regression. (Regression tells how much specific factors contribute to a given outcome.)
* Correlation is not causation. Be careful when communicating about variables that are highly correlated with one another.