

Hypothesis Development and Testing

In this module we will be examining Kids Count data and identifying how we can use the available indicators of status of children in the United States for statistical analysis. Kids Count is a project of the Annie E. Casey Foundation that tracks children on both a national and state-by-state basis measuring the educational, social, economic, and physical well-being of children. Data collected is available on-line and is used by a variety of individuals and organizations involved with projects such as welfare guidelines, health care initiatives, educational programs, and the development of a system of policy supports that can help parents become more successful both as workers and as parents.

Along with the data we have tools available that allow us to utilize both descriptive statistics and inferential statistics. Descriptive statistics are procedures for summarizing, organizing, graphing, and in general, describing quantitative information. Inferential statistics is used to make inferences about a population based on information about a sample drawn from that population.

1. The first step of this exercise is to identify a research question. The specification of a research topic should move from a general discussion of an area of interest to a more narrowly defined issue. Answer the following three questions:
 - a. What is the general issue or interest area on which you are writing?
 - b. What are the specific questions that you are asking or issues you intend to explore?
 - c. Why are these issues/questions important and sociologically interesting?
2. The next step of the exercise is to formulate a hypothesis. Hypotheses are statements of (or conjecture about) the relationships among the variables that a researcher intends to study. Hypotheses are generally testable statements of relations. In such cases, they are thought of as predictions, which, if confirmed, will support a theory.*
3. After writing your hypothesis, identify the dependent and independent variables that you will use to test your hypothesis. The dependent variable is defined as the presumed effect in a study; so called because it “depends” on another variable. It is the variable whose values are predicted by the independent variable, whether or not caused by it.* For example, in a study to see if there is a relationship between the teen violent crime arrest rates and the teen dropout rate the teen violent crime arrest rates might be the effect (dependent variable).
4. The independent variables are the presumed causes in a study. Independent variables are the variables that can be used to predict the

values of another variable.* In the above example, the independent variable might be the teen dropout rate.

5. When identifying dependent and independent variables the question of cause is preeminent. Cause is defined as an event, such as a change in one variable that produces another event, such as a change in a second variable. Be warned that there is no concept in statistics that is more troublesome than "cause." Researchers may disagree about what constitutes a cause and especially about how restrictive a set of conditions must be met before it is legitimate to talk of cause. Many social scientists would agree with the following – others would not: to attribute cause, for X to cause Y, three conditions are necessary (but not sufficient): (1) X must precede Y; (2) X and Y must covary; (3) no rival explanations account as well for the covariance of W and Y. Causal relations may be simple or multiple. In simple causation, whenever the first event (the cause/independent variable) happens, the second (the effect/dependent variable) always does too. Multiple causation is much more common in the social and behavioral sciences.* Multiple causes may be such that any one of several causes can produce the same effect (for example, teen violent crime arrest rates may be caused by teen dropout, children living in poverty, teen birth rate or some combination of the three). Multiple causes also may be such that no one of them will *necessarily* produce the effect, but several of them in combination make it more likely.

6. Examine your dependent variable as a trend over time. Choose the United States and one or two other states for your line graph. Consider choosing one of the worst of best-ranked states as a comparison. Choose your dependent variable as your indicator. Graph the data from 1990-1999.

Describe the overall national trend. Was there an increase, decrease, or did it stay the same. Put another way, nationally is your selected variable on the rise or is it decreasing?

7. Do the same for your independent variable(s).

8. To test the relationship between your dependent variable and the independent variable(s), open the excel file called "tool_us.xls". Make a scatter plot by using the pull down menu. Let x be the independent variable (plotted on the x axis) and y be the independent variable (plotted on the y axis). Cut and paste the scatter plot into a Word file and record the correlation coefficient. (An explanation of the correlation coefficient can be found below)

Are there any data points that seem to stand out --not part of the cluster of data points? These are called outliers. Click on an outlier to see which state is represented.

Was the hypothesis confirmed? Explain your answer.

9. Repeat step 8 for each of your independent variables.

Explanation of Correlation Coefficient

The correlation coefficient or Pearson's r is a measure of the degree of linear association existing between two variables. We want to pay close attention to both the direction and strength of the association. A positive correlation is indicated by the absence of a negative sign and means that variables are changing in the same direction. An increase or decrease in one variable corresponds to the same change in another variable. For example, we would expect that the more time a student studies for an exam (x) the higher the exam score (y). A negative relationship is indicated by a minus sign and means that as one variable increases there is a corresponding decrease in another variable. The strength of a relationship is indicated by the numeric value of the coefficient. Coefficients range from -1.0 to 1.0 . These values are examples of perfect correlations. In reality most values are found in between -1.0 and 1.0 . Correlations of $.30$ or less (either $+$ or $-$) are considered weak, $.31 - .70$ (either $+$ or $-$) are deemed moderate and $.71$ and above (either $+$ or $-$) considered strong. These are not absolute rules but should be used as a guide in interpretation. Note that the higher the correlation coefficient (either positive or negative), the more closely clustered the data points are in the shape of a diagonal line.

What might be a better measure of your dependent variable than the one(s) you tested?

* See: Vogt, Paul W. 1993. *Dictionary of Statistics and Methodology*. London: Sage Publications.