Forensic Geography Lab: Descriptive Statistics and Correlation

The purpose of this lab is to introduce students to a handful of basic statistics, several of which are critical to predicting crime at an introductory level.

Student Learning Outcomes

1. Student will calculate and interpret a battery of descriptive statistics.
2. Student will calculate and interpret correlation coefficients.

In order to be able to understand the geography of crime, it is necessary to be able to identify the social and economic conditions that appear to be causal factors in the pattern of crime. These causal factors are generally well known, and include things like income, ethnicity, age profile, residential stability, family structure among others. Criminologists and geographers have identified other elements in the landscape, like liquor stores and payday loan stores as contributing to a rise in crime rates.

If you would like to measure the effect of variables that you think are contributing to crime, you must understand how to calculate and interpret several key predictive statistics; chief among these statistics is regression. A good regression model allows you to figure out what factors increase or lower crime rates. Regression also allows you to identify places that fit the expected pattern, those places that have higher crime rates than expected (based on the model), as well as those places that have lower crime rates than one would expect based on the model.

However, before you move on to regression, it is necessary to understand correlation, and before that simple descriptive statistics. Why? Because you it is difficult successfully identify likely causal variables to include in the model without these introductory stats. It's also often necessary to prepare your data so that it will work properly in a regression model. These intro stats will help you know if you need to fix up your data so it will work better in the model.

This isn't a "stats" class, so we won't go into all the gory details, or answer all the myriad "what ifs" that go with learning how to do regression. If we did that, we'd run out of time during the semester! Instead, the goal is to show you the power of ~~the dark side~~ regression modeling so that you might find it worth your time to learn how to do it on your own, or inspire you to take more stats, or the Spatial Analysis Course.

Simple Descriptive Statistics

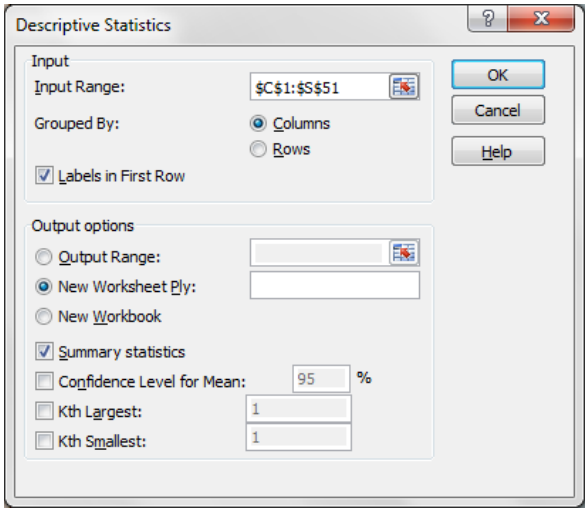
Descriptive statistics describe a group of numbers. Duh? You're probably familiar with "average" or "mean", and perhaps with median. Both are measures of central tendency; or try to identify what's average, normal or "in the middle". You may be less familiar with standard deviation, skewness, kurtosis and the coefficient of variation. Each of which is useful as you try to understand your dataset,

and how individual variables (like poverty, education levels, imprisonment rates) might interact with data on crime rates.

There are a lot of ways to calculate a big pile of stats (SPSS, ArcMap, GeoDa, etc.) but since we've been working with Excel in this course, we'll use it to calculate a page of descriptive statistics for this lab.

Here's how:

1. Open the file State_Crime_Data.xlsx It will be on the Y://courses/courses_sgraves/ drive in the Forensic folder or available via link in Moodle.
2. Find and select (click on) the tab at the bottom of the file called "Homework" to activate it.
3. Examine the data and note that it is contains a column of property crime rates labeled PCR_100K (property crimes per 100,000 persons).
4. Note that there are a number of other columns as well, each of which might logically have a relationship with the property crime rates in each state. For a fuller description of what each column heading means click on the "Labels-Key" tab.
5. To calculate a pile of descriptive statistics for this data set, activate the DATA tab in Excel from the top tool ribbon.
6. Find the Data Analysis icon among the icons on the tool ribbon. It's frequently in the top right corner, by itself.
    a. If the Data Analysis tool icon is missing, then click on the File tab. Select "options" in the left options column, click "Add Ins", click "Go…" (near the bottom) and make sure "Excel Add-ins" is active in the drop down menu to the left of the go button, check "Analysis ToolPak" and click OK. Follow instructions (if there are any) Go back to your Data tab.
7. Click on the Data Analysis tool. From the pop-up window, select "Descriptive Statistics" and click OK.
8. Click once in the "Input Range" box, to make sure your cursor is there the tool is ready to accept your input.
9. Click in Cell C1 (PCR_100K) and drag down and to the right until all the data and headers are highlighted.
10. Input range should appear as it does in the image to the right.
11. Make sure you check "Labels in First Row"
12. And select "Summary statistics" as it is in the image to the right.
13. Click OK.
14. Examine your results and answer the several question in Moodle.



Kurtosis

Kurtosis is a measure of the peakedness of observations in a histogram (or bell curve) around the mean. If your observations are crowded closely around the mean, and the tails of the curve are short, then the distribution is called "leptokurtic". If your observations are scattered widely around the mean and the curve is flat, the curve is described as "~~flattykurtic~~" "platykurtic".

Excel uses a negative-positive scale to express the kurtosis of a distribution. Progressively larger positive number indicate a more peaked or leptokurtic distribution, negative numbers a platykurtic distribution.

Skewness

Skewness is a measure of the asymmetry of a distribution of numbers. This is of some concern to geographers who want to do spatial analysis because badly skewed distributions can make predictive modeling and calculating correlation coefficients hazardous…and difficult to evaluate results accurately.

When data is positively skewed (a positive value in Excel) then there are more observations above, or "to the right" of the mean. If one or two observations are very far above average …these are called "outliers" (e.g., Bill Gates' income should he move into my neighborhood) the distribution would be very badly positively skewed (to the right).

Data can be negatively skewed as well when there are outliers well below the mean.

If there are outliers above and below the mean in equal proportion, it tends to move the skewness back toward zero, or neutral, but those can still cause some problems…and should appear in the kurtosis score.

Coefficient of Variation

A statistic related to Kurtosis and Skewness is the coefficient of variation. It is very simply the ratio between the mean and the standard deviation. It helps you understand how much variability is the within the dataset. You can compare the variability of data among your variables. You can perhaps pick out potentially troublesome variables with this simple statistic.

Excel doesn't automatically calculate it for you so you'll have to do it yourself. Here's how:

1. Find the standard deviation and divide it by the mean. You might want to multiply it by 100.
2. If your standard deviation is in cell B7 and the mean is cell B3, then the formula would be =B7/B3 or =(B7/B3)*100 if you hate fractional numbers.

Correlation

Correlation statistics help us understand possible relationships between variables. First we can see the "direction" of the relationship. So you might ask, "When one variable gets larger…does the other get larger? Or perhaps smaller?" or "When poverty increases, does neighborhood crime also increase?" If you find two variable both grow together, it's called a "positive correlation". Sometimes, as one variable gets bigger, the other gets smaller. That's called an "inverse" or "negative" correlation.

Correlation measures also help us measure the strength of relationships. You can answer "How *much* does one go up when the other goes up?" or "How quickly does crime rise when poverty increases?" or "Is the relationship between poverty and crime strong, or is it weak?"

To find the strength and direction of relationships between two sets of data, you can calculation one of several *correlation coefficients*. The correlation coefficients will be expressed as numbers between zero and one. A perfect correlation of (which are rare) indicates that for every unit of change in variable, there is a perfect, corresponding change in the second variable. If for every *increase* of unit of change in one variable the second one also increases by a proportionate amount, then the correlation coefficient will be 1.0 (positive). If on the other hand, for each unit of growth or upward movement in the first dataset, you find a corresponding, but opposite (negative) unit of movement in the second dataset, then you'll get a correlation coefficient of -1.0.

You may be interested to see in the dataset provide for this lab whether or not there are strong relationships between the property crime rate by state, and the various other columns of data in the dataset. Is there a relationship, for example, between the number of cops per person and the property crime rate? One would guess yes, but how to measure that relationship? You can do it in Excel. Here's one way:

1. Open your data and activate the "Homework" data tab.
2. Click on Data (top tab) and from the tool ribbon select Data Analysis tool.
3. From the Data Analysis options, select "Correlation"
4. The input range is the entire data set from column C to the right…click in C1 and click and drag down and to the right to highlight all your data.
5. Check "Labels in First Row" and click OK (see image)
6. Examine your results and answer a few questions about those results.