Probability Models Do Not Generate Your Data

"How are these data distributed?" is the wrong place to start your analysis.

Donald J. Wheeler

The number of major hurricanes in the Atlantic since 1940 (used in the February Column) are shown as a histogram in Figure 1. Some analysts would begin their treatment of these data with a consideration of whether they might be considered to be distributed according to a Poisson distribution.

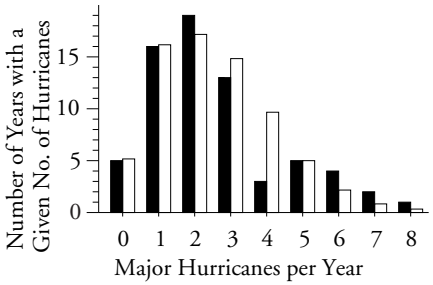

Figure 1. Major Atlantic Hurricanes 1940-2007 (Black)
Expected Numbers According to Poisson Model (White)

The 68 data in Figure 1 have an average of 2.60. Using this value as the mean value for a Poisson distribution we can carry out any one of several different tests collectively known as "goodness-of-fit" tests. Skipping over the details, the result of such a test would be that there is "no detectable lack-of-fit between the data and a Poisson distribution with a mean of 2.60." Based on this, many analysts would proceed to use techniques that are appropriate for collections of Poisson observations. For example they might transform the data in some manner, or they might compute probability limits to use in analyzing these data. Such actions would be wrong on several levels.

First of all, since the analysis above began with a goodness-of-fit test, we need to understand just what a goodness-of-fit test does and does not do. The key is in the language used. A statement that "There is no detectable lack-of-fit" is not the same as saying, "These data were generated using a particular probability model." The double negative is not the same as the positive. But all too often in practice the double negative is used as if it were a positive statement. In fact, with enough data, you will *always* reject any particular probability model. This is why all goodness-of-fit tests are actually *lack-of-fit* tests, and using the correct terminology will help to prevent the error described here.

The only time that we can ever make a positive statement that the data were generated using a particular probability model is when we are generating artificial data sets. While this may be appropriate when working out the details for some new type of analysis, it has no place in the analysis of real-world data.

But don't we need to know if the data are normally distributed or if they satisfy certain other conditions before we can use our analysis techniques? No, this notion is based on confusion between the purpose of data analysis and the way we teach statistical inference. When we analyze data we want to know if a change has occurred, or if two things are different, because we are going to take some action based on the outcome of the analysis. As long as we can be reasonably sure about our decision, the particular alpha-level, or confidence level is no longer important. As long as our procedure is reasonably conservative, so that the potential signals are not readily confused with the probable noise, we have a technique that we can use.

In statistical inference we rarely apply distributional assumptions to the original data, X. Instead, we work with some transformation of the data, Y. Our statistic Y will generally have a known distribution which will characterize some aspect of the distribution of X. Using the known distribution of Y, we compute critical values, compare these with the observed value for Y, and decide what this tells us about the distribution of X. This mathematical approach may seem very exact and precise, but in practice it is always approximate.

Thus, a common point of confusion for students of statistics is to think that we have to know the distribution of X before we can perform any analysis. Thankfully this is not true. In fact, this point of confusion was a major obstacle to the development of analysis techniques in the Nineteenth Century.

Finally, the overwhelming obstacle to using the question of how the data are distributed as the starting point for an analysis is the fact that the data may not be homogeneous. In January we discovered that the data in Figure 1 came from two different systems. For 33 years the number of major hurricanes averaged about 1.7 per year, with an upper bound of 4 per year. For the other 35 years the multi-decadal tropical oscillation resulted in an average of about 3.5 major hurricanes per year, with an upper bound of 9 or 10 per year. So which of these two different systems does the Poisson model in Figure 1 represent? Neither one.

Your data are the result of a process or system, and like all things in this world, these processes and systems are subject to change. For this reason alone, the primary question of data analysis is, and always has been, "Are these data reasonably homogeneous or do they contain evidence of a lack of homogeneity?" And the primary tool for examining any set of data for homogeneity is the process behavior chart. This is why any data analysis should always begin by using the context of the data to place the data on a process behavior chart in some rational manner. To do otherwise may well result in patent nonsense.

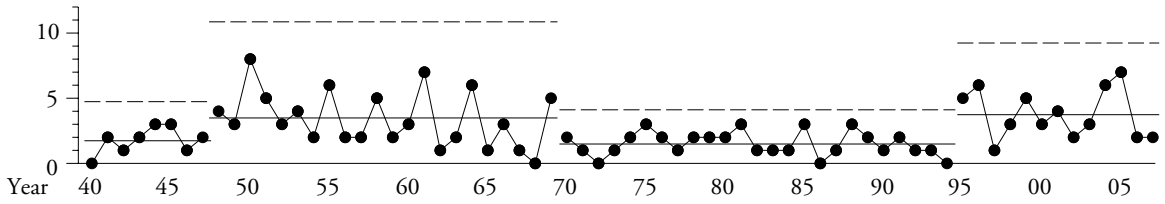

Figure 2. *X* Chart for the Number of Major Hurricanes in the Atlantic 1940-2007