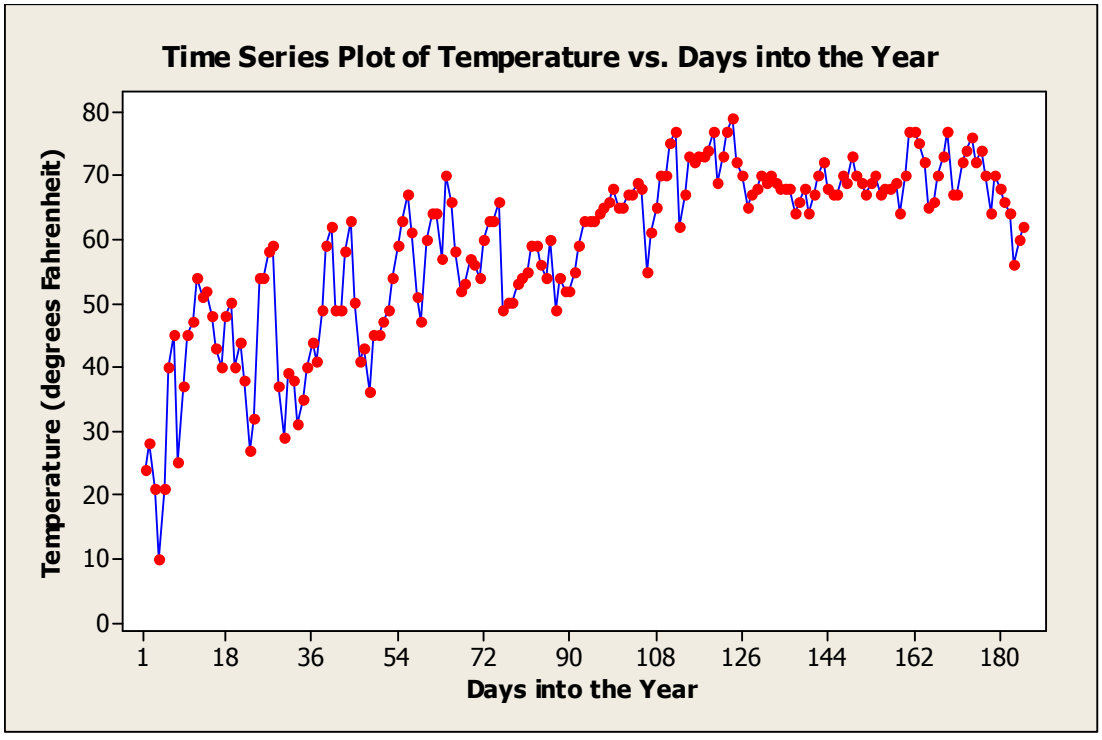
2006-1397: TEACHING THE IID ASSUMPTION IN ENGINEERING STATISTICS I

Mark Inlow, Rose-Hulman Institute of Technology

Teaching the IID Assumption in Engineering Statistics I

Many procedures taught in introductory statistics courses require that the data meet, at least approximately, the normality assumption. However, all require, in some form, the IID assumption, namely, that the observations comprising the sample are independently and identically distributed. Because of this, and the fact that statistical procedures are less robust to IID violations than normality violations and IID violations are much more difficult to handle, the IID assumption is the more crucial of the two.

In spite of this fact, we believe introductory statistics courses for engineers, and the corresponding texts, neither adequately stress the importance of the IID assumption nor provide adequate tools for assessing it. Our belief is based on observing students in upper level statistics courses unthinkingly apply IID analysis methods to data which is blatantly non-IID. We became aware of the extent of this problem when students in an advanced statistics course, after spending a week on time series analysis, blithely computed a confidence interval for the mean of the following nonstationary data using the IID formula.



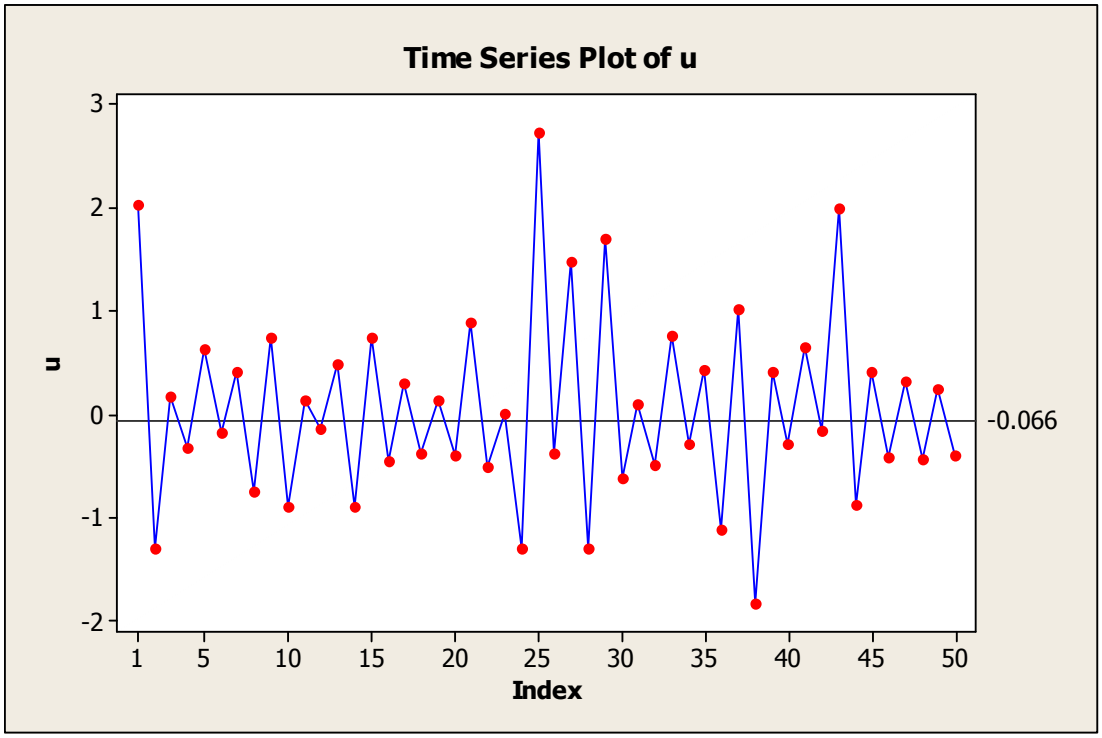
Although there is a trend in introductory statistics texts toward greater emphasis and assessment of the IID assumption, additional improvements are needed. One reason for this is that most of the effort at improving statistics pedagogy is directed toward the problem of teaching statistics at the precalculus level to non-engineering students. Such students are well-served by the emphasis on data acquired via simple random sampling whereas engineering students also need to know how to handle process data - data which is more likely to violate the IID assumption.

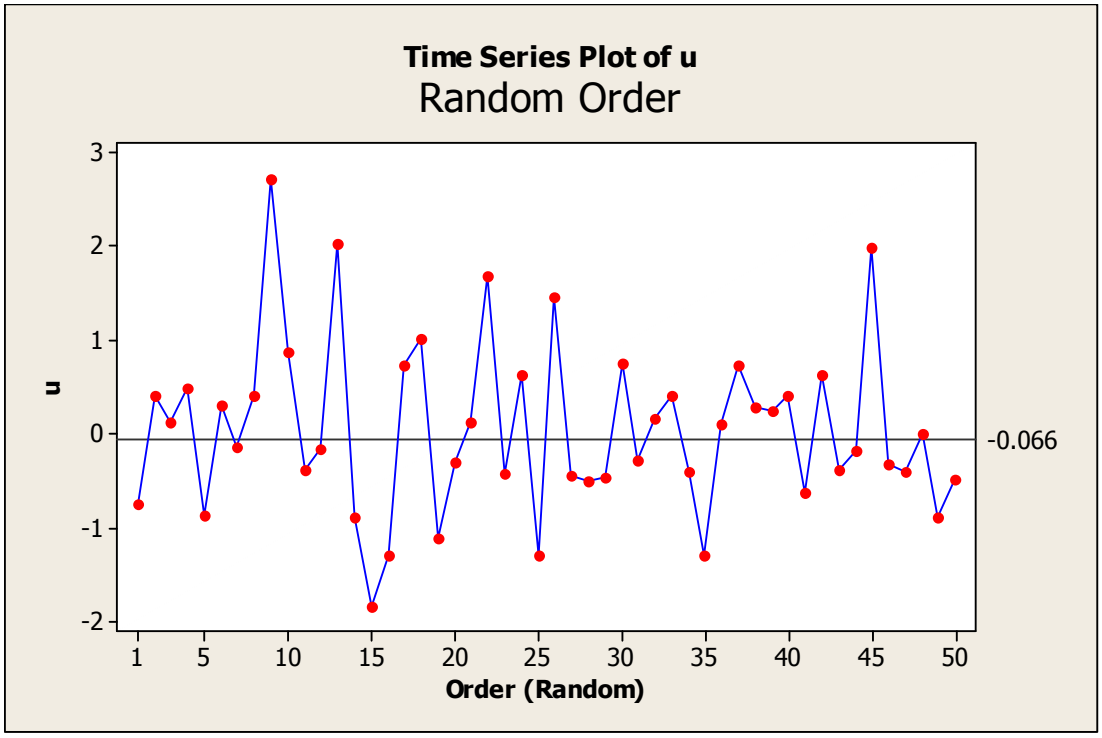
Following the example of one recent text, *Applied Statistics for Engineers and Scientists* by Petruccelli, Nandram, and Chen, our approach is to introduce students to the IID assumption as soon as possible. However, this is complicated by the fact that we begin the course with descriptive statistics instead of probability. Because of this we cannot define the IID assumption rigorously until later. Therefore we initially present the notion of IID data in an intuitive fashion. On the first day of class we have students measure their reaction times doing a simulation multiple times and then discuss how we should determine which students have the shortest reaction time. Students propose the usual candidates for summarizing an individual's reaction times such as the mean, median, trimmed mean, etc. The next class period we provide time series plots of the reaction times for randomly selected students and then resume the discussion. Since most individuals' reaction times decrease across the multiple trials, the students realize that their original proposals are not appropriate since there are trends in the data. Thus we set the stage for introducing methods for summarizing data and assessing the IID assumption.

Following the approach of David Moore and others of defining the distribution in the beginning of the course/text, we begin by defining the distribution of a data set to consist of the unique values which occur in the data and how often each value occurs. Since the distribution is typically much smaller than the data set, the question naturally arises as to what information is lost when the data is summarized by its distribution. The answer, of course, is that all information contained in the order of the data is lost. The ordering of the data will contain important information only if the observations are related (*they are not independent*) and/or the observations exhibit systematic trends (*they are not identically distributed*), that is, if the data violate the *IID* assumption. Since students intuitively understand what it means for the observations to be related and to possess trends, we initially define the IID assumption in terms of the absence of these characteristics. Further, by linking violation of the IID assumption with the idea of information in the ordering of the data, we arrive at a new visual method of assessing the IID assumption, namely, the random order time series plot. If the order of the data contains no relevant information, i.e., the data meet the IID assumption, then the visual character of the data set should be essentially unchanged if it is plotted in random order. Thus the IID assumption can be assessed by comparing the appearance of the data plotted in order, the times series plot, with the data plotted in random order, the random order plot. If the visual character of the two plots is similar, then the IID assumption is probably met. Of course, one can check the IID assumption using only the time series plot but comparison of this plot with random order plots facilitates interpretation by providing examples of IID data with the *same distribution*. These examples make it easier to determine if

1. apparent trends are systematic or simply random oscillations, and
2. if adjacent observations are related or autocorrelated since the random order plots will appear to be rougher or smoother than the original data if negative or positive autocorrelation is present.

An example of this is provided below. Although there aren't any trends in the time series plot, we note that it appears rougher than the random order plot. This is due to negative correlation - when the data is plotted in order, adjacent values have opposite sign.

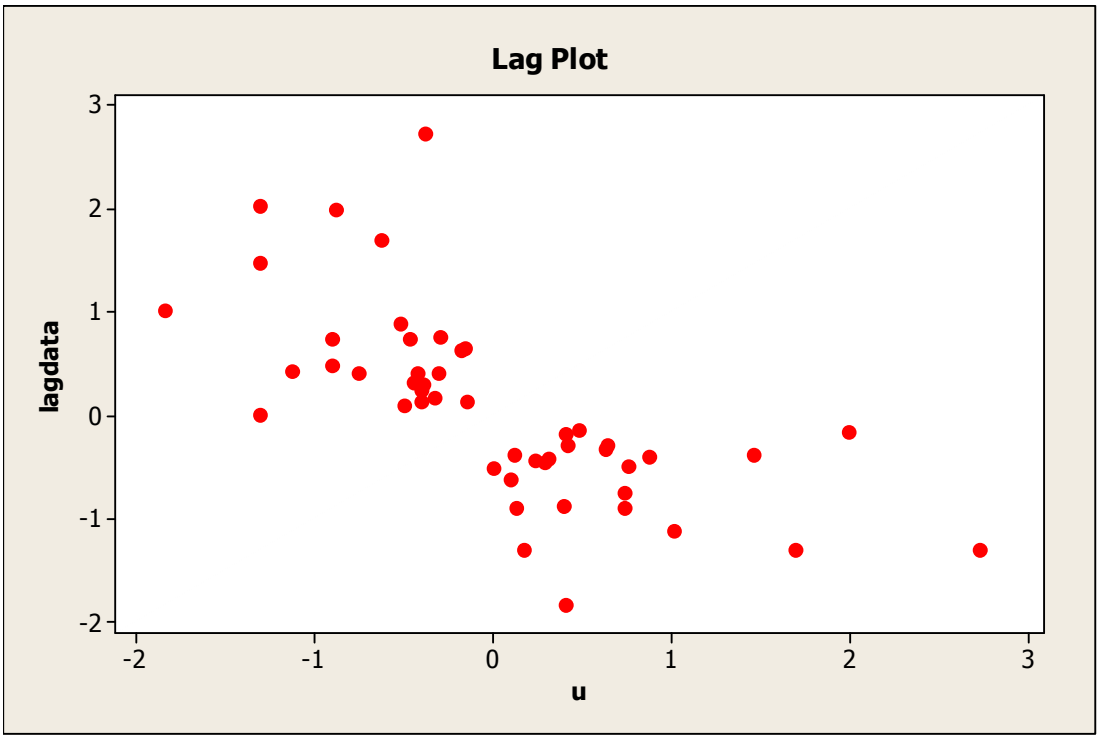

Time Series Plot of u
u
Index
-0.066


Time Series Plot of u
Random Order
u
Order (Random)
-0.066

Advantages of the times series/random order plot comparison approach are

1. **Conceptually simple**: Students intuitively understand randomizing the order of the data.
2. **Data unaltered**: Unlike other approaches for detecting trends which use smoothing and other data summarization techniques, the data values are not altered in any way.
3. **Many IID examples**: Since each random order plot provides an example of a time series plot of IID data, after assessing the IID-ness of a few data sets, students will have seen several examples of time series plots of IID data.

An additional benefit of the times series/random order plot comparison approach is that it convinces students students of the need for objective statistical methods very early in the course. Determining if data are IID can be tricky using time series plots and random order plots. In particular, it's difficult to determine if the data are dependent by comparing the smoothness of the two plots. After analyzing several data sets, some students begin to express frustration at the inherent subjectivity of using the time series/random order plot comparison approach to detect dependence in the data. Fortunately, by this time we have begun discussing descriptive statistics for bivariate data, in particular, the scatter plot and the sample correlation coefficient r. Thus, at this point, students easily grasp the fact that if the data lacks independence because adjacent observations are related, then a scatter plot of observation x(t-1) vs. observation x(t) (often called a lag plot) will exhibit a nonuniform distribution of points in the Cartesian plane. In particular, for the most common form of dependence, autocorrelation, the points will approximate a line. In this case, the correlation coefficient r provides an appropriate, objective measure of the strength of the relationship between adjacent observations. Below is the lag plot for the data of the preceding example. The corresponding correlation coefficient is -0.682.



Advantages of using the lag plot to assess lack of independence are

1. Scatter plot application: By showing that the scatter plot provides a useful solution to an existing problem (assessing independence) and providing a source of problems, the lag plot facilitates students' understanding of the scatter plot while simultaneously reinforcing their understanding of the IID assumption.
2. Correlation application: By showing that the correlation coefficient provides an objective measure of dependence (if the data is autocorrelated) and providing a source of problems, the lag plot facilitates students' understanding of the correlation coefficient including determining when it is and is not an appropriate measure of the strength of the relationship between two variables. Again, it does this while reinforcing the IID concept in students' minds.

An additional benefit of using the correlation coefficient of the lag plot to assess dependence is that it motivates the need for formal hypothesis testing procedures and provides a means of revisiting the IID assumption as soon as the p-value approach to hypothesis testing is presented. Although the sample correlation coefficient r provides an objective measure of dependence when the data is autocorrelated, there is still the issue of how large r must be for us to decide that its value is not an artifact of the data set but is due to autocorrelation in the data generating process. Thus we need a formal procedure for making this decision, i.e., the hypothesis test. Since most statistical packages provide p-values for testing that the "true" correlation coefficient is zero, this provides an example of hypothesis testing in a familiar scenario (IID assessment).

Of course, before discussing hypothesis testing, we cover probability theory. Here we return to the IID assumption and define it formally in the context of discussing the fact that a simple random sample is (at least approximately) a collection of IID random variables. At this point we reinforce the IID concept by showing how it enables us to easily derive the sampling distribution of the sample mean using the rules for describing the distributions of linear combinations of independent random variables. (We believe that, time permitting, introductory engineering statistics courses should, at the very least, teach the rules for linear combinations. Some courses/texts go further by teaching additional methods under the rubric of uncertainty analysis or error propagation.)

Finally, throughout the remainder of the course, we revisit the IID assumption since it appears in some form for the usual statistics procedures taught in an introductory engineering statistics course, i.e., one and two sample analysis, ANOVA, and regression. For example, if the random error assumptions of the regression model are met, then the residuals are normal and obey (approximately) the IID assumption. Due to their experience assessing the IID assumption via our methods, the students are well-prepared to interpret residual plots and determine if the residuals are autocorrelated.

Whenever someone advocates that a new method/approach be added to an introductory statistics course, some invariably reply that there's no room! In fact some advocate teaching fewer topics in introductory courses so that students better absorb the remaining material. We are sympathetic to these concerns.

We argue that the benefits of using the above methods and approach will outweigh the costs as follows:

1. The problem is real as demonstrated by motivated statistics students failing to grasp the IID concept in upper level courses.
2. We believe our approach helps solve the problem by
    i. Repeatedly revisiting the IID assumption throughout the course, and
    ii. providing graphical and formal procedures for testing the IID assumption analogous to the graphical and formal procedures (normal QQ plot and normality test) currently taught for assessing the normality assumption.
3. An additional benefit of our approach is that it informally introduces students to the concept of a sufficient statistic: If the data set is IID then the distribution captures all relevant information, i.e., the distribution is a sufficient statistic.
4. The overhead of our approach is very low since
    i. Students should actually collect data at some point in an introductory course. Using reaction time experiments, it is easy to get non-IID data.
    ii. The two methods we use, the random order plot and the lag plot, are simple versions of methods which should be in this course, the time series plot and the scatter plot, respectively. Further, we provide these methods as macros (Minitab) avoiding the overhead of teaching students the software details of constructing the plots.
    iii. Our approach simplifies residual analysis. The regression model model assumptions are met if the residuals are normally distributed and (essentially) IID with respect to the predictor variables. Our methods can be used to check that the residuals possess this latter property.
    iv. Because of i, ii, and iii, perhaps the only real overhead of our approach is the time it takes to discuss the IID assumption throughout the course. This should be a small addition to what instructors should be doing anyway, namely, conceptually connecting the various topics so that statistics is perceived as a unified whole organized around a few core concepts as opposed to a grab bag of loosely related techniques.

Since we only fully integrated the above methods and approach into our Engineering Statistics I course this academic year, we have yet to verify the benefit of these changes in advanced courses. However, given the fact that student performance on final exam questions involving the IID assumption has greatly improved, we believe students have already benefited from the changes. We are optimistic about their future performance.