The Journal of Foot & Ankle Surgery 57 (2018) 427–431

Contents lists available at ScienceDirect

The Journal of Foot & Ankle Surgery

journal homepage: www.jfas.org

Investigators' Corner

The Doctor Is In! Diagnostic Analysis

Daniel C. Jupiter, PhD

Associate Professor, Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, Galveston, TX

ARTICLE INFO

Keywords:
boxplot
diagnostic test
influential point
Q-Q plot
residual

ABSTRACT

To make meaningful inferences based on our regression models, we must ensure that we have met the necessary assumptions of these tests. In this commentary, we review these assumptions and those for the *t*-test and analysis of variance, and introduce a variety of methods, formal and informal, numeric and visual, for assessing conformity with the assumptions.

The first thing that I like to do as an analyst, when I receive a data set from a collaborator, is to poke around a bit, informally. I look to see whether the data have been coded correctly; is, for example, the body mass index truly a numeric variable or have words errantly wandered into that column of the spreadsheet? Is the differentiation between an unrecorded ankle-brachial index and incompressible veins clear? Are there some strange values in the age variable, such as a negative age? One might argue that this process should be called "data cleaning"; however, a more liberal view sees this as a first diagnostic pass over the data. We are looking to diagnose whether the data are even suitable for analytical purposes.

Once I am assured that the data are reasonable, in good order, and that no egregious errors are present, I will perform the descriptive analysis of the data that I have doubtless planned with my collaborator. These are the basic descriptions of means, medians, minima, and maxima for continuous variables, and counts within categories for discrete variables. More exploration, more cleaning … but, again, diagnostic in nature. If I notice that one of my several ankle-brachial index categories has but a few patients, and I realize this will lead to a difficult to interpret analysis or cause mathematical issues in the multivariate tests, I will make some decisions about recoding variables. This is diagnosis, purchasing insurance, ensuring that I am not walking into a minefield; perhaps a bit of prophylaxis, as well as diagnostics.

Why do I mention these seemingly pedestrian explorations of data? This is supposed to be a technical article about diagnostic tests. First, as I have just suggested, I believe that these steps are, indeed, part and parcel of the diagnostic analysis of data. Second, I hope to reinforce that the act of performing an analysis does not consist of merely pressing a button and having a machine spit out a ticker tape of statistical answers. Rather, the analytical process is one in which the investigator and analyst are actively involved the whole way through, ensuring that we are not committing an error of garbage in, garbage out.

Truly, the analysis is a journey of watchfulness from beginning to end—from these initial examinations of the data to the final assessments of whether our results have any real world meaning. Along the way, we perform the specific diagnostic tests that are the subject of this commentary: diagnostics designed, among other things, to assess whether the assumptions of our statistical tests are met.

The Importance of Assumptions

The present commentary is the third in a series of 5 specifically targeted at thinking about regression models and model selection. We started the series by discussing the assumptions of statistical tests (1), outlining the different assumptions for different families of tests, and thinking about the problems that arise when we violate those assumptions. Our second article put this in context by looking at the goals for the models we might wish to build and what they are used for: description, identification of risk factors, or prediction (2). Given some of the tools that these 2 articles introduced, we are now in a position to look back at our assumptions and consider further tools that will allow us to determine whether they are met. In the next, our last 2, commentaries, we will first look at predictive modeling and some aspects thereof. Finally, we will conclude with an overarching discussion of how statisticians make model choices: given the goals of the model one is building, how does one decide which variables are appropriate to be included.

To recapitulate quickly, we noted in our first commentary on assumptions that the Student *t*-test and analysis of variance (ANOVA) relied on the following assumptions: normality of the data within each

Financial Disclosure: None reported.
Conflict of Interest: None reported.

Address correspondence to: Daniel C. Jupiter, PhD, Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, 301 University Boulevard, 1.134G Ewing Hall, Galveston, TX 77555-1150.
E-mail address: dajupite@utmb.edu (D.C. Jupiter).

1067-2516/$ - see front matter
https://doi.org/10.1053/j.jfas.2018.01.001

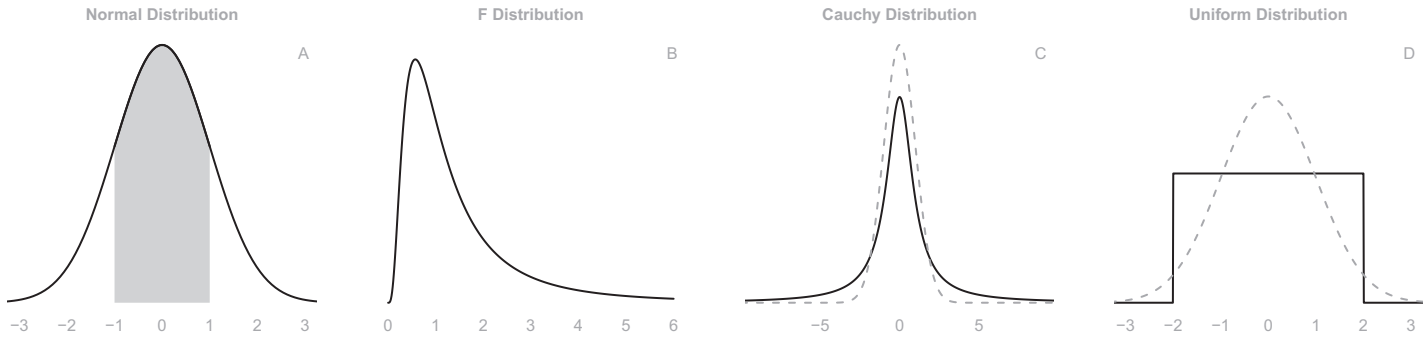

Fig. 1. Probability distributions. *(A)* A normal distribution, with the data within 1 standard deviation of the mean highlighted in gray. *(B)* The F distribution, an example of a skewed distribution. *(C)* The Cauchy distribution, a leptokurtotic distribution. The normal is drawn on the same axes, as a dashed line. Note the "fat tails" of the Cauchy distribution. *(D)* The uniform distribution, a platykurtotic distribution. The normal is drawn on the same axes, as a dashed line.

of the study groups, a common variance across all the study groups, and independence of observations. Linear regression relies on the errors in the model being normally distributed, having common variance at each value of the independent variables, and independence of observations. We further require that the linear regression equation is linear in the coefficients, although not necessarily linear in the variables, and that the relationship being modeled is indeed linear. Whatever the purpose for one using a model (2), meeting the assumptions is essential. It does not, after all, matter for what purpose you are constructing a building; if the foundations are not stable, the building will not stand.

We also revisit 2 technical points that we have made a few times. First, it is not the populations underlying the groups studied using the *t*-test or the samples themselves that we are studying, in particular, and that were drawn from those populations, that we require to be normal. Rather, it is a more complicated object, the sampling distribution of the mean from the underlying populations that we require to be normal. Not having this sampling distribution to hand—and, absent the entire underlying population, we never will—we are content to state that our assumption is that the samples we are examining are normally distributed. Second, we do not have the errors in the linear regression: the errors are the deviations of the observed data points from the mean values predicted by the true model. This true model is posited to exist in the universe; however, we will never see it, because it is a theoretical object. Instead, we have the model estimated from our sample of data and the deviations of the observed data from that approximate model. These deviations are the residuals, and we assess their normality, rather than that of the errors. A third wrinkle: usually one tests the assumptions of a tool before using it. We can only examine residuals once we have them, which occurs after we have built the model and used the linear regression, whose very assumptions we would like to check. Truly, then, diagnostic testing is part of our analytical journey from beginning to end.

Assumptions of the *t*-Test and ANOVA

Normality

To discuss the most important of our assumptions, normality, we first briefly consider the normal distribution. Although we all know what the normal distribution looks like—the standard bell curve to which we appeal when our college grades are not what we desire—we are perhaps less familiar with the characteristics of that distribution that make it useful for diagnostic testing. Of note, the bell curve is symmetric, the mean is the same as the median, and roughly 68% of the distribution is within 1 standard deviation of the mean (Fig. 1A). These are descriptive, not prescriptive, in that a distribution might have these characteristics and still not be normal, but they are a useful starting point. For example, if a distribution is asymmetric, or *skewed*, with many more values in the right (or left) tail than in the left (or right), it cannot be normal (Fig. 1B). Alternately, if a distribution has significantly more (or less) than 68% of the data within 1 standard deviation of the mean, it cannot be normal. We call such distributions *leptokurtotic* or *platykurtotic* (Fig. 1C,D).

Plots for Assessing Normality

These descriptions inspire 2 visual tools for assessment of normality: the boxplot and the normal Q-Q plot. The boxplot will be familiar to most readers, and examples are provided in Fig. 2. As the legend for Fig. 2A indicates, the plot displays the maximum, minimum, and median of the data, along with the first and third quantile. This gives an impression of the overall distribution and symmetry of the data. The plot also contains a notion of the spread of the data in the adjacent points, which roughly encode the standard deviation of the data. If our data are normal, we would expect the plot to be symmetric, and we would expect to see few points outside the adjacent points.

The normal Q-Q plot is a plot of the observed data against the corresponding theoretical quantiles of a normal distribution. That is, we plot the observed data on 1 axis, and, for each point in the data set, we find and plot on the other axis that value in the normal distribution representing the same percentile as our original data point[1]. Although this definition is involved, suffice to say that if the data are normal, the points in the Q-Q plot will be on the 45° diagonal line through the origin of the graph. Any severe deviation from this indicates a lack of normality, and indeed with a trained eye, patterns of skew or kurtosis can be derived by looking at such pictures. We give examples in Fig. 3.

These are our first diagnostic tests. If we are running a *t*-test, we separately plot the data in the 2 groups, in boxplots and Q-Q plots. We look for the symmetric boxplots as described that are not too "fat." We look for our data to lie roughly on the 45° line in the Q-Q plot. If we see serious violations of these requirements, we are led to believe that our data are not normal.

[1]We recall that for a given observation in a data set, the percentile corresponding to that observation is the percentage of data in the data set with a smaller value than the given observation.