The Journal of Foot & Ankle Surgery 56 (2017) 910–913

Contents lists available at ScienceDirect

The Journal of Foot & Ankle Surgery

journal homepage: www.jfas.org

Investigators' Corner

Assumptions of Statistical Tests: What Lies Beneath

Daniel C. Jupiter, PhD

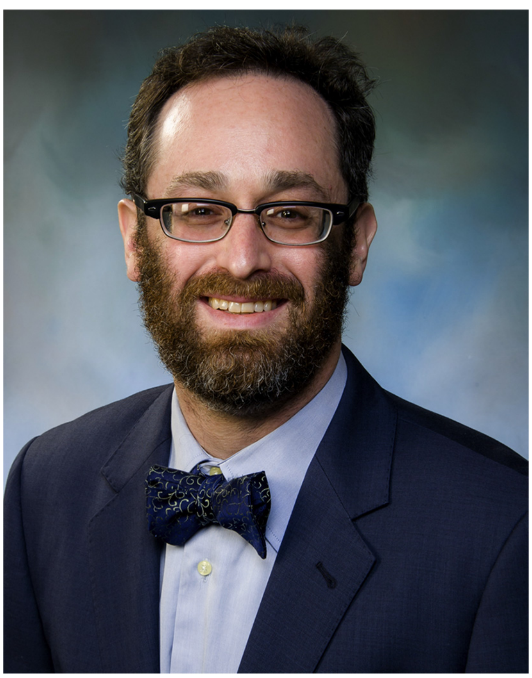
Assistant Professor, Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, Galveston, TX

ARTICLE INFO

Keywords:
homoscedasticity
independence
linearity
normality

ABSTRACT

We have discussed many statistical tests and tools in this series of commentaries, and while we have mentioned the underlying assumptions of the tests, we have not explored them in detail. We stop to look at some of the assumptions of the *t*-test and linear regression, justify and explain them, mention what can go wrong when the assumptions are not met, and suggest some solutions in this case.



While we have mentioned them in passing, and have delved into them slightly when considering non-parametric statistical tests (1), the assumptions underlying our statistical tests have most often been background players in our discussions. In this commentary, we will change focus, bringing assumptions to the foreground, and look closely at those of some of the more commonly used statistical tools. We will think about why the assumptions are important, why they are required, and what they mean. We will also discuss what happens when assumptions are not met, and mention some options for addressing these violations.

Conspicuous by its absence in this article is a review of tools to test assumptions; while this is clearly important, we prefer to leave this aspect of the discussion for later examination, as part of a broader, overarching discussion of regression models and model selection to be presented in a series of 5 articles. We will present here, then, an overview of the meanings and technicalities of the assumptions of tests. In a second commentary, we will discuss several notions of goodness of fit, and fidelity of models, given that our required model assumptions are met. In that commentary we will more thoroughly introduce residuals of various sorts, as part of metrics of fidelity. We will also use residuals in our diagnostic tools, to be introduced in a third commentary. Next in the series we will discuss predictive models, and the use of cross-validation for the verification of these models. This discussion profits from our earlier introduction of residuals. The exploration will be brought, in some sense, full circle, with a fifth commentary on how to make model selection decisions.

The Right Tool for the Job

Why worry about assumptions at all? Statistics is difficult enough as it is, as we have to think about which test to apply, how to interpret the results, and how to use the results to inform the clinical story we are trying to tell. If we already have more than 20 commentaries on use and interpretation of statistics in the Investigators' Corner series, why muddy the waters further with more technicality? Because statistics, like any tool, is agnostic to its use and users. One could, but shouldn't, use one's kitchen oven to make pottery. One could, but shouldn't, use the screw driver from the tool shed in performing a TAR. In either case, it is not the tool

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

Address correspondence to: Daniel C. Jupiter, PhD, 301 University Boulevard, 1.134G Ewing Hall, Galveston, TX 77555-1150.
E-mail address: dajupite@utmb.edu

1067-2516/$ - see front matter
http://dx.doi.org/10.1053/j.jfas.2017.05.022

that is to blame, but the user; in both cases the tool has been extended past where it should go, and the results obtained will be suboptimal, and perhaps disastrous. So too with statistics. Any time there are two groups of subjects, we can, but perhaps shouldn't always, compare variables between the groups using a Student's *t*-test. Any time we are looking at union rates vs. time since surgery, we can, but perhaps shouldn't always, examine trends using a linear regression. In this commentary we explore the limits of some of our tools, we see how far they can and should stretch, and we look at what happens when they break. We are looking to see what suitable inputs are needed to obtain meaningful outputs: what should be allowed in the oven if we want to bake beautiful breads. Rather than getting into technical weeds, or exploring esoteric tests, we will focus on several of the more commonly used statistical tests: *t*-test, ANOVA, and linear regression.

Assumptions of the Student's *t*-test and the ANOVA

While it is the most basic of the tests we will examine, many of the assumptions of the Student's *t*-test have echoes or are revisited in the other, more complex, tests which we use. The *t*-test is used to compare the mean values of a continuous variable in two groups. We see in this description of the test the origin of the requirements that the data has to meet in order for the test to produce meaningful results: we are examining continuous variables.

Normality

As such, since we usually prefer to deal with normally distributed continuous variables, we require it of our data; the data in each group should be normally distributed. This is actually a relatively subtle requirement, and not quite true as written. In unravelling the *t*-test, and looking at its mechanics, we learn that the statistic it computes, the *t*-statistic,

$$\frac{\mu(A)-\mu(B)}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}},$$

is distributed following a *t*-distribution. (We recall that $\mu(A)$ and $\mu(B)$ are the mean values in groups A and B, respectively; s_A^2 and s_B^2 are the within group variances in the two groups; and n_A and n_B are the numbers of subjects in each of the two groups.) In order that the above quantity follow the *t*-distribution, it must be the case that the data in each group satisfies a requirement closely related and very similar to the requirement that the data in each group be normally distributed.[1] We settle, then, on this last, the requirement that the data be normally distributed within each group separately.

Homoscedasticity

We notice in the formula above that the means in the two groups are compared, but the variances are combined, giving us a pooled estimate of the variance of the overall population. In fact, if we think about the null hypothesis of the *t*-test, it is that the two groups being compared are actually from the *same* population: that men and women are the same in terms of mean height, that time to union is identical with the two fixation types being compared. Given that this is the null hypothesis, our second assumption is completely natural: homoscedasticity, or that the variances in the two groups are the same. We are looking at two groups with the same variance and asking whether the means are the same. This assumption also makes sense from a mathematical point of view, in that if one group's variance is very large, and the other's small, one group will dominate. The problem is exacerbated if one group is much larger than the other. Of course, given that we are looking only at samples of our two groups, and there is thus sampling noise, we do not expect that the variances be exactly the same in the two groups, but we do require them to be relatively close.

Independence

While not explicit in the formulation of the *t*-test, the third assumption comes about from the explicit statement of the question which the test is hoping to answer: given two, distinct, groups, are the means of the continuous variable of interest the same? We note that the two groups are distinct. In other words, no individual appears in both groups, and the groups are independent, with no relationships between individuals in the two groups, except those that may arise by mere chance. This precludes studies where we look at individuals taking two distinct treatments; pre/post studies; studies of siblings or other closely related pairs of people; and any other similar designs.

The assumptions of the ANOVA are the same as those for the t-test, but of course involve three or more groups, rather than two. And, while the essence of the assumptions remains the same in considering two-way ANOVA, their statement becomes somewhat more complex.

Violations

We have listed three requirements: within group normality, equality of variance, and independence of observations. How seriously should each be taken, and why do they matter? For the first assumption, we can see the importance with a very simple example. We consider the mean incomes of people in two neighboring counties. If one county happens to contain a high powered CEO or two, the mean incomes in that county may well look very large, relative to those in the neighboring county. Remove those CEOs, however, and the incomes look essentially the same. The skew imparted by those data points makes the incomes in the one county non-normally distributed, and may confuse our inference about relative wealth, depending on the question we wish to answer. Normality matters. We have also mentioned that the formulas, the mathematics behind the test, rely on normality of data. But, having said this, the *t*-test is actually relatively robust to deviations from normality, and it is only data with great skew, or especially small data sets, that we worry about. In most cases, with moderate or large data sets, absent serious deviations from normality, we are not too concerned. The second assumption, homoscedasticity, while natural, is also somewhat odd from a scientific point of view: why would we expect that men's and women's heights have the same variance? What if we truly want to know whether the means are the same, and do not wish to be concerned with equality of variance? Again, the test is relatively robust to violations, but even better, statisticians have developed a variant of the *t*-test, the Welch-Satterthwaite *t*-test, that removes this assumption. While this new test is slightly less powerful than the most basic *t*-test, the strength of not having to concern oneself with the annoying homoscedasticity assumption is worth the loss of power—most statistical packages will default to, or at least provide, the Welch-Satterthwaite results. We are left with the third assumption, that the two groups be independent. While the previous two assumptions pertain to the content of the data, this assumption is about study design. We have discussed related issues in our earlier

[1] For those of a technical bent, the specific requirement is that the sampling distribution of the mean within each group is normally distributed. It is the fact that, due to the Central Limit Theorem, this holds for large enough sample sizes, irrespective of underlying distribution of data, that allows us to ignore the assumption of within group normality of the data, when groups are large.