CHAPTER **42**

BIOSTATISTICS 101

Stephan J. LaPointe, DPM, PhD

INTRODUCTION

The assumption for the sake of this manuscript is that the reader has an interest in and understands the importance statistics are currently playing in our podiatric medical practice. This will only increase in the future. There is already more attention being paid to quality and value of clinical research. In addition efforts are already underway to measure our outcomes and prove that the care we provide is financially responsible and clinically meaningful. This new understanding will significantly impact, change and improve how we practice our profession. All parties involved, especially our patients, will benefit. An entire manuscript could be dedicated to this premise alone.

This manuscript is based primarily and almost entirely on a text Primer of Biostatistics, Fifth Edition, by Stanton A. Glantz, Ph.D. The book is well written and intended to be read by medical professionals and researchers. This manuscript for the sake of brevity hits on the salient points without the examples and much of the prose found in the text. The intent is to provide the reader with quick reference when reviewing journal articles or about to embark in study of their own. Certainly a more thorough explanation of the principles of statistics can be found in the Primer of Biostatistics. My recommendation would be to either purchase the fifth edition or wait for the seventh edition. The latter is likely to be released by the time you read this manuscript. This does not cover all the topics in the text, but provides an introduction to some of the statistical methods contained therein that should start someone on their journey to understanding statistics.

PART 1: CHARACTERIZING DATA

Normal Distribution

Part I will focus on ordinal data (data with ranking such as numbers) that comply with the normal distribution since this includes most of the studies and data we will be exposed to. What is the normal distribution? The normal or Gaussian distribution is what we learned in grade school when teachers graded our papers as the bell curve. The normal distribution occurs whenever the measurement is the result of small independent random factors. An individual measurement in the normal distribution is more likely to fall near the mean (or average) than far away from it and equally likely to be above or below the mean. If the data are not likely to follow these criteria then the normal distribution does not apply. This could occur if the students got the answers to the test and most scored at or near 100 since the data would be skewed and the scores cannot be above 100.

Mean and Standard Deviation

Whenever the measurements meet the criteria of the normal distribution two parameters can fully describe the data. They are the mean and the standard deviation. The mean is simply the average of the data.

$$\mu = \frac{\Sigma X}{N}$$

Where μ is the mean, ΣX is the arithmetic sum of the individual measurements and N is the total number of measurements.

The second parameter, the standard deviation, describes how the data varies about the mean. To define the variability of the data we calculate how far away from the mean each measurement is by subtracting the mean from each value. We square this value since, if a measurement is below the mean, it results in a negative value. This is called the population variance.

Population variance = sum of (measurement of one member – mean)^2 / number of members

The equivalent mathematical expression is:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Since the units of the population variance are squared (i.e. cm^2 in the case of length measurements) we take the square root of the population variance to arrive at the standard deviation. Mathematically this is expressed as:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

So now we have two parameters the mean and the standard deviation that define the data assuming the data follow the normal distribution.

Other characteristics of the normal distribution are that roughly 68 percent of the measurements fall within one standard deviation and about 95 percent fall within two standard deviations.

Sample Mean and Standard Deviation

Now, in most studies it is impossible to measure all the subjects so we sample some of the population. If we measure only a subset of the population with n measurements of a possible total N we call this the sample mean denoted as $\bar{X}$ and defined mathematically as:

$$\bar{X} = \frac{\Sigma X}{n}$$

The sample standard deviation is denoted as s (or SD) in the following equation:

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

The difference between the sample mean and standard deviation and the population mean and standard deviation is that we are substituting the entire population mean μ with the sample mean $\bar{X}$ and in calculating the sample standard deviation we are dividing by n -1 instead of n. The reason for the latter is that the sample will never show as much variability as the entire population. By dividing by a value less than n (i.e. n -1) we are increasing the standard deviation so we do not underestimate it.

Standard Error of the Mean

$\bar{X}$ and s are derived from a sample of n of the entire population N. How close are the sample mean and the standard deviation of the sample to actual mean and standard deviation to the underlying population? We introduce an additional parameter to help with this called the standard error of the mean. This statistic measures how close the sample mean is to the underlying entire population mean. We take say 50 samples of size n of N and calculate (here's where it gets tricky) the mean of the 50 sample means ($\overline{X_{\bar{X}}}$) and the standard deviation of the sample means ($s_{\bar{X}}$). The mean of the sample means $\overline{X_{\bar{X}}}$ will be equal to the mean μ of the entire population. However, $s_{\bar{X}}$ the standard deviation of the sample means does not equal the standard deviation of the underlying population. It is a measure of the variability of the means not the underlying population. The means themselves cancel out much of the underlying variability of the population so this value will always be smaller than the standard deviation of the population. This value is called the standard error of the mean. The true standard error of the mean ($\sigma_{\bar{X}}$) is defined as:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size.

The best estimate of the standard error of the mean, $s_{\bar{X}}$, from one single sample is:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Central Limit Theorem

One interesting characteristic of the distribution of the sample means is that it will follow the normal distribution (or bell curve) even if the underlying population data does not. The standard error of the mean will be larger for larger variability of the underlying population as noted by the standard deviation. This value will be smaller for a larger sample size as shown in the last equation above. This and the characteristics of the sample means and standard error of the means has led to the central limit theorem.

- *The distribution of sample means will be approximately normal regardless of the distribution of the values in the original population from which the samples were drawn.*
- *The mean value of the collection of all possible sample means will equal the mean of the original population.*
- *The standard deviation of the collection of all possible means of samples of a given size, called the standard error of the mean, depends on both the standard deviation of the original population and the size of the sample.*

One note is to be careful when authors report their data they often substitute the standard error of the mean in place of the standard deviation when describing their data since the former will be smaller. This, however, fools the reader to believing there is less variability of the population data and is a serious error.

PART II: TESTING FOR THE DIFFERENCE BETWEEN GROUPS

Tests of significance are how we determine whether a treatment has an effect on the underlying population. We first assume that there is no significant difference due to the treatment or treatments and call it the null hypothesis. The tests used to determine whether to accept or reject the null hypothesis are collectively called analysis of variance. How do we do this?

We estimate the variance in two ways. First we calculate the average variance for all the groups and call it the within groups variance. We have eliminated the treatment effects.

$$s^2_{wit} = \frac{s^2_{treatment}}{number\ of\ treatmnets}$$

The within groups variance s^2_{wit} is based on the mean for that sample and therefore is independent of and will not be affected by any of the treatments.

Next we calculate the estimate of the underlying population variance. Remember that we decided to test the null hypothesis that there is no difference between any treatment groups so this implies that all the samples are from the same underlying population. So the standard deviation of the sample means of the different groups will approximate the standard error of the means. From the equation we used to calculate the standard error of the means, $s_{\bar{X}}$:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

We square the equation multiply by n to solve for s^2 (now noted as s^2_{bet}) which will be the estimate of the population variation based on the sample means or the between group variance.

$$s^2_{bet} = n\, s^2_{\bar{X}}$$

$s_{\bar{X}}$ is the standard deviation of the same sample means i.e. the standard error the means.

The F test statistic

If the groups are from the same underlying population and the treatment has no effect then both estimates of variance from within and between the groups should be about the same value. We create the F test statistic.

$$F = \frac{s^2_{bet}}{s^2_{wit}}$$

If there is no difference between the groups this value should be close to 1. However, if F is a large number then there is more variance between the groups than within the groups and therefore we can reject the null hypothesis. There is indeed a difference between the groups and therefore the samples were not drawn from the same population, (ie, one of the groups is different due to the treatment effect).

The P value

The value of the F test statistic will depend on which individuals in the underlying population are tested. If we were to repeat this experiment 100 times when there is no difference between the groups most of the values of F will be close to 1 but some will be greater than 1. We want to calculate the value of F for which there is only a 5% chance that we reject the null hypothesis when indeed the null hypothesis is true. This value of F is determined to be a "big F" and we report that the P <0.05 (for 5%). It is possible, with sheer bad luck, to reject the null hypothesis 5% of the time at that level of F

The critical value of F must not be based on only 100 experiments but all possible experiments. There are typically an infinite number of experiments that can be performed due to the large size of the underlying population. So mathematicians have created tables for critical values of F that correspond to p <0.05 and p < 0.01.

To create these tables there are four underlying assumptions:

- *All the samples are independent of each other.*
- *Each sample must be randomly selected from the underlying population.*
- *The populations from which the samples are drawn must have a normal distribution*
- *The variances of each population must be equal, even when the means are different, i.e, when the treatment has an effect.*

Degrees of Freedom and F tables

The value of F depends on the size of the sample and number of samples that are under question. So does the F value at which P < 0.05. So the equations used to develop the F tables are dependent on two parameters known as the degree-of-freedom parameters denoted as ϑ. The numerator or between groups degrees of freedom from the F statistic is

$$\vartheta_n = m - 1$$

The within groups or denominator degrees of freedom:

$$\vartheta_d = m(n-1)$$

M is the number of samples and n is the sample size.

The degrees of freedom are simply the way the number of samples and the sample size enter into the mathematical formulas used to calculate the statistical tables of critical values of F.

The t tests – special case of two groups

The analysis of variance may include more than two groups and only states whether there is a difference between any of the groups, but not which groups are significantly different from each other. The t test is the most often used statistic in the medical literature and is often inappropriately used to compare multiple groups which we discuss further. When the t test is used to test multiple groups it increases the likelihood of erroneously rejecting the null hypothesis when it is indeed true. This effectively increases the chances of reporting that some intervention or treatment had an effect when it did not.

$$t = \frac{difference\ in\ samples\ means}{standard\ error\ of\ difference\ of\ sample\ means}$$

When this ratio is small there is no difference. When the ratio or t is big we will reject the null hypothesis and assert that the treatment had an effect. This is the same logic used in the analysis of variance. In both cases we are comparing the differences in magnitude of the means with the variability that would be expected within the samples.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}}}$$

But since we assume that the two samples were drawn from the same population the variances S_1^2 and S_2^2 are both estimates of the population variance. Therefore, we average the two and call it the pooled-variance estimate:

$$S^2 = {}^1/_2\ (S_1^2 + S_2^2)$$

The t-test statistic based on the pooled-variance estimate is therefore:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2}{n} + \frac{S^2}{n}}}$$

As with the F statistic there are a range of values for t dependent on which members of the population are tested. As with F, if the t value is big (either positive or negative), the null hypothesis will be rejected and the treatment will have had an effect. In the case of the t test it has two tails since we are subtracting the means and can have both negative and positive values. When there is no difference the t statistic approaches zero. Again the t values for P <0.05 and P <0.01 have been tabulated. As with the F statistic the sample size enters into the mathematical tables as degrees of freedom $\nu = 2(n-1)$ where n is the sample size. As the n increases it becomes easier to detect smaller differences between the groups. As it turns out the t test is a special case of analysis of variance where $F = t^2$ and there are two groups. Suffice it to say that this can be proven mathematically.

Post – hoc analysis

When there are more than two groups, we first perform the analysis of variance using the F statistic. Then if the data shows a significant difference we conduct post-hoc analyses also called multiple-comparison procedures to determine which groups defer significantly from each other. As previously asserted many authors simply perform multiple t tests between all the groups. Recall that we assume t statistic at the p <0.05 assumes that we are willing to accept a 5 percent chance (or one in 20) that we report a difference when one does not exist. If we repeat this process comparing group A with group B and then group B with C and then group A and C and so on we increase the chance that we will state that there is a difference when one does not exist. So for every comparison we make we add an additional 5 percent chance of erroneously reporting a difference when one does not exist. So for example with five groups of data there would be 10 possible pair wise comparisons or a 50% chance of stating that there is difference when there is not.

The Bonferroni t test

The Bonferroni t test is a simple arithmetic modification of the t test that takes into account the increase likelihood of reporting a difference when making more than one comparison. Instead of taking the value of t at p <0.05 you divide that by the number of comparisons to be made. In the case of three comparisons take the 0.05/3 for the 1.6 percent chance of reporting a difference when it does not exist. This works well until there are about 10 comparisons at which this becomes overly conservative.

Holm t test

Another more powerful post hoc analysis based on the Bonferroni concept is the Holm t test. It is a step down procedure. Here the divide p value sequentially for the given number of tests in the case of three pair wise comparisons 0.05/3 = 0.0167; 0.05 /2 = 0.025; 0.05. Then based on the degrees of freedom, the t statistic for each value of p is determined and compared to the calculated values starting for the largest difference (smallest p value and largest t value). If the null hypothesis is rejected (there is a significant difference) then the next comparison is made otherwise all following stepwise comparisons are assumed to be consistent with the null hypothesis (not significant).

Student Newman Keuls test (SNK or Newman Keuls test)

There is another mathematical model and statistical post hoc analysis, the Student Newman Keuls test with a corresponding table for its q statistic. This model gives a more realistic estimate of the true likelihood of making the error of stating there is a difference when there is not. The first step is still to perform the analysis of variance. Then calculate the q statistic as follows:

$$q = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{s^2_{wit}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Here $\bar{X}_A$ and $\bar{X}_B$ are the means being compared, s^2_{wit} is the variance within the treatment groups as calculated for the analysis of variance and n_A and n_B are the sample sizes. The degrees of freedom ϑ_d is the denominator degrees of freedom again from the analysis of variance. The table is listed by degrees of freedom for each value of α_t the total accepted risk for falsely rejecting the null hypothesis usually $p < 0.05$ or 0.01. The means are listed in order from smallest to largest. The first step is to compare the largest to the smallest mean. Then compare the largest with the second smallest until the largest is compared to the second largest. Then repeat the same for the second largest until all comparisons are made. In these tables there is a p value needed to determine the q value. The p is the number of means in the comparison. For example if there are four means then the value of p for the q statistic is the number of means in the comparison will be four when comparing the fourth to the first. Comparing the third to the first is p of 3 and so on. If no significant difference exists between two means then assume none exists within the remaining means.

The Tukey test

The SNK was derived from the Tukey Test and is computed identically to the former with one exception. In the SNK test we used different P's representing the number of means included within the comparison being made to arrive at a different q value. So for the SNK we have different critical values for q for each comparison. With the Tukey test the P value is set to m, the number of means or groups in the study. So there is only one P value. (Do not confuse this P with the P [probability] value for acceptable error). Between the Tukey and the SNK the Tukey will be more conservative and less likely to reject the null hypothesis. The SNK proponents feel that since analysis of variance is done first it controls for overall error rate.

Which post-hoc test to perform?

Unadjusted t tests are also known as Fisher's protected Least Significant Difference Test yield too many errors and the Bonferonni is at the opposite end and too conservative. Dr. Glantz prefers the SNK over the Tukey but relates that it is somewhat liberal in detecting differences. Ultimately he recommends the Holm test as the best compromise since it is less conservative than Tukey or Bonferonni, while at the same time controlling for overall risk of false-positive tests at the pairwise level. The important thing is to understand the difference between the tests and how they are performed.

PART III: POWER

What does "not statistically significant" mean

When researchers make statistical comparisons of data and find a statically significant difference between treatment groups they declare as such. However, they often make the wrong assumption that if there was no difference found that there was no statistically significant difference. Not necessarily. They just failed to prove that there was a difference, but the treatment may still have an effect. As a critical reader we need to ask whether the test performed was robust enough to detect a difference if it existed. This relates to the power of the study.

Type I and Type II errors

There are two types of errors. Type I errors or false positives occur when we state there is a significant difference and there is not. Type II errors or false negatives occur when we state that there is no difference when there is one. The commonly accepted value for a big statistic where we are comfortable declaring that there is a difference is 0.05 and this is denoted by $\alpha = 0.05$. This is our minimum p value. Type I errors occur when one rejects the null hypothesis and it is true. The risk of having a Type I error is expressed by α in most cases 0.05 or 5 percent.

The probability of a Type II error where we accept the null hypothesis and it is false or a false negative is denoted by the Greek symbol β. Here there is a treatment effect and we fail to find it. The power of the test is the probability of finding a difference when it exists or true positive and is relative to the Type II error probability by $1 - \beta$ (Table 1).

What affects our risk of type II errors of stating that there is no difference when one exists? Or otherwise stated what variables affect power and are there things we can do to improve it? The size of the treatment effect plays a role. The larger the effect of the treatment the easier it is to detect. So one thing we need to do is determine how small an effect that is worth detecting. This can be subjective.

Table 1.

TYPE 1 AND TYPE 2 ERRORS

	ACTUAL SITUATION	
Conclude from observations	**Treatment has an effect**	**Treatment has not effect**
Treatment has an effect	True positive Correct conclusion 1 - ß	False positive Type I error α
Treatment has no effect	False negative Type II error ß	True negative Correct conclusion 1 - α

Type I and II errors are related. The harder or more stringent we are in making the test to prove there is an effect or make ß smaller, the more likely we are of missing a true effect or make ß bigger or the power (1 – ß) smaller. There is only one way to improve on both simultaneously and that is to increase the sample size or number of subjects.

The variables that play an important role in the power of a test are the underlying variability in the population, the size of the treatment effect, what probability you will accept for a significant result, and number of subjects.

The size of the Type I Error α

Requiring that there be stronger evidence before reporting a significant difference we are effectively changing the α value to a smaller value. This has the effect of pushing out the point at which we will state there is a significant difference further out of the bell curve distribution. The power of the test lies in the points that are outside of that previous α value. Increasing α we are making more of the points lie within the bell curve and decrease the power.

The size of the treatment effect

The larger the effect of the treatment the further apart the two bell curves that represent the two groups will be. Since the power can be represented by the data that lies outside the overlapping of the two curves the power again is increased with larger treatment effect.

The population variability

The more variability within each group the wider the bell curve for each group. For a given treatment this increase in the width of the bell curve will again increase the amount each of the distributions will overlap and decrease the power.

It is common to combine the size of the treatment effect and the population variability into a ratio called the noncentrality parameter and using this dimensionless ratio to help characterize the treatment effect and variance as a ratio:

$$\phi\ (noncentrality\ parameter) = \frac{\delta\ (treatment\ effect)}{\sigma\ (\ population\ standard\ deviation)}$$

Bigger sample size equals more power

Investigators have no control over the size of the treatment effect or population variability and the probability to determine there is a significant difference has historically been α = 0.05. There is one additional variable that investigators will be able to control and that is the sample size. This will increase power for two reasons.

Increasing the sample size increases the number of degrees of freedom which in turn decreases the value of the t statistic that will result in a significant difference. Secondly as the sample size increases this will result in higher t values overall when comparing to the distribution for no effect.

Although t tests were used to explain power, the principles apply to all statistics including analysis of variance. The exact method for calculating power is dependent on the test and its mathematical model.

Summary of power

The power of a test tells us the likelihood that the hypothesis of no treatment effect will be rejected when the treatment has an effect.

The more stringent we are with our requirement for detecting a difference, the less power of the experiment.

The smaller the size of the treatment effect when compared to the population variation the lower the power.

Larger sample sizes result in more power.

The exact procedure to measure power is dependent on the test in question and its underlying mathematical model. We typically aim for power of approximately 80% for an acceptable level. If an author reports no statistically significate difference, he or she should provide a power analysis. Otherwise we need to be careful, if not skeptical, of the ability of the experimental design to find a difference.

PART IV: EACH SUBJECT RECEIVES THE TREATMENT(S)

When each group in study contains different subjects the largest source of variability is the variability between each subject. If the study can be designed where each subject receives the treatment or treatments then we can significantly reduce the population variability. Each subject will in effect serve as the control for the treatment. This helps to isolate the effect of treatment from the variability among subjects.

Paired t test: one subject before and after one treatment

In experiments when the subject can be observed before and after a treatment, we can eliminate the variability between subjects by measuring the change the treatment introduces within each subject. Remember that for t

$$t = \frac{parameter\ estimate - true\ value\ of\ population\ parameter}{stadard\ error\ of\ the\ parameter\ estimate}$$

The parameter we are going to address is the change in each individual due to the treatment or δ. We let d equal each individual change and $\bar{d}$ is the mean change. So the standard error of the difference is

$$s_d = \sqrt{\frac{\Sigma(d-\bar{d})^2}{n-1}}$$

So the standard error of the difference is

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

And then

$$t = \frac{\bar{d} - \delta}{s_{\bar{d}}}$$

Since we want to test the hypothesis that there is no change with treatment $\delta = 0$ for no change with the treatment.

This value of t is then compared to the critical value of t where the degrees of freedom is defined by the equation: v = n – 1.

The t test like all the tests we have discussed require a normally distributed population and in this case we are referring to the changes due to the treatment effect have to be normally distributed.

To summarize for the paired t test

Compute the change in response the occurs for each individual d.

Compute the mean change $\bar{d}$ and the standard error of the mean changes .

Compute t by dividing the former by the latter = $\frac{\bar{d}}{s_{\bar{d}}}$.

Compare this t with the critical value for v = n – 1 degrees of freedom where n is the number of subjects.

Repeated measures analysis of variance

When each subject undergoes more than one treatment we have a parallel situation to analysis variance. Except again each subject behaves as a control and we have eliminated a significant source of variability which is the between subject variability. In this model we will be left with within subjects variations. Not all subjects will respond the same to the treatment. So within subject variation will be divided into two sources the variation due to the treatment and the residual variation (since each subject is likely to have some variability in response to a particular treatment).

There is a detailed mathematical treatise on the development of the of repeated measures analysis of variation in Chapter 9 of the text Primer of Biostatistics, Fifth Edition. For our purposes we will show the logic and skip the more rigorous mathematical development of the equations.

The total sum of the squares equals the sum of the sum of the squares for within the treatment groups and between the treatment groups.

$$SS_{tot} = SS_{bet\ subjects} + SS_{wit\ subjects}$$

We can further divide the within group sum of the squares into the effect of the treatment and residual variation of the subject to the treatment.

$$SS_{wit\ subjects} = SS_{treat} + SS_{res}$$

So we can solve for SS_{res}

$$SS_{res} = SS_{wit\ subjects} - SS_{treat}$$

The same can be done for the degrees of freedom

$$DF_{res} = DF_{wit\ subjects} - DF_{treat} = n(m-1) - (m-1) = (n-1)(m-1)$$

Now statisticians throw us a curve ball and call the SS/DF the mean square and label it the MS. This is not a mean in the mathematical sense of the word but is really an estimate of the variance. Hence:

$$MS_{treat} = \frac{SS_{treat}}{DF_{treat}} \quad and \quad MS_{res} = \frac{SS_{res}}{DF_{res}}$$

So here

$$F = \frac{MS_{treat}}{MS_{res}}$$

And we use the DF_{treat} numerator degrees of freedom and the DF_{res} for the F statistic tables. This is analogous to the F statistic for the analysis of variance. For analysis of variance we used the variance within each group in the numerator and the variance between the groups using the means for the groups. Here we have eliminated the between subjects variability since each subject serves as its own control. The numerator represents the variance due to the treatment and denominator is the residual variance. If this is from the same population i.e. there is no treatment affect the F will be small or closer to one. As the treatment introduces more variance the F will be large just as in the analysis of variance.