Statistics

Editors:

Seth A. Strope, MPH

Authors:

Christian J. Nelson, PhD

Last Updated:

Monday, March 1, 2021

1. Introduction

A basic knowledge of statistics is needed for urologists to make informed decisions about treatments, read the medical literature, and conduct their own research and quality improvement studies. This section of the AUA core curriculum provides an introduction to statistics. Recommendations for further study are provided at the end of the chapter.

2. The Importance of Sample Size: Confidence Intervals and Power

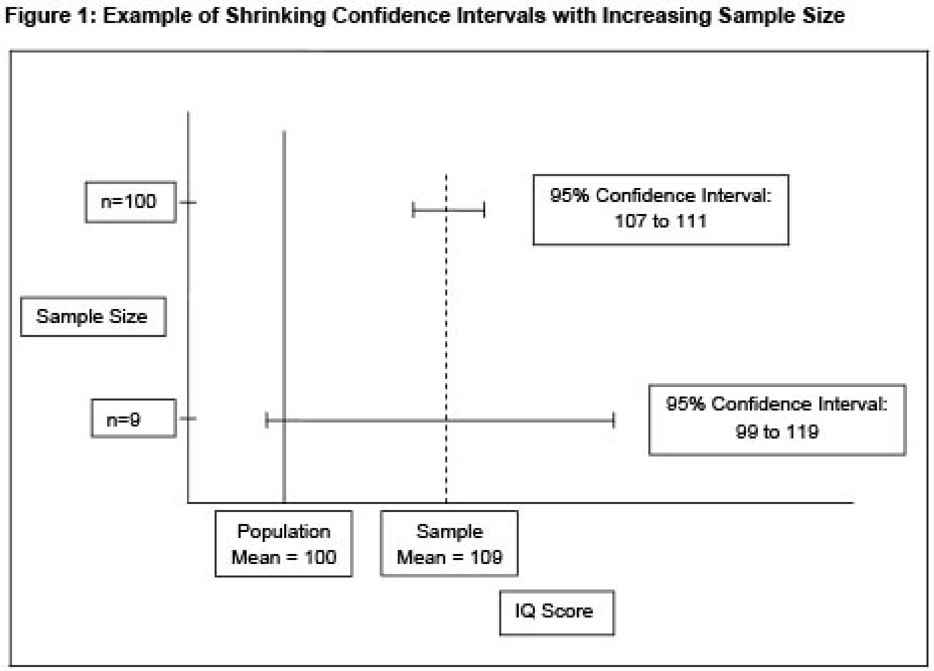

Figure 1: Example of Shrinking Confidence Intervals with Increasing Sample Size

The basic concept is actually very simple: **as the sample size increases**, **the sample becomes a better estimate of the population**. As the sample becomes a better estimate of the population, the sample statistics (numeric information from the sample) become a better estimate of the population values. This basic concept translates into how precise we believe the statistics are for a specific sample size. **The larger the sample size**, **the more precise the statistics**. Statisticians discuss this concept as "sampling error," and as our statistics become more precise, there is less "sampling error." Thus, as the sample size increases, there is a reduction in sampling error.[1]

Through our understanding of probability and the properties of the normal distribution, statisticians have developed a way to measure sampling error. **The term used for the measurement of sampling error is** "**standard error**." Standard error basically represents what its name implies - the standard amount of error that one would expect in a sample given the sample size.[1]

Statisticians have translated the concept of standard error into the development of "confidence intervals." Most often, these confidence intervals are reported as **95% confidence intervals** (although you may at times see a 99% confidence interval). This range represents the best estimate (our confidence) where 95% of the sample means would fall around the population value for a specific distribution if we selected multiple samples.[2] A sample mean is considered a "point-estimate," and represents our best guess at the population mean given the characteristics and size of the sample. If a second sample was taken with the same characteristics and sample size as the first sample, and a sample mean was calculated, this would most likely produce a sample mean with a different value than what was calculated from the first sample (a product of sample error). This second sample mean would be considered another "point-estimate." If this process was repeated for multiple samples with the same characteristics and sample size, these point-estimates could be put together to form a range of possible values of sample means. Using the concepts of probability, we can calculate the interval that contains 95% of these sample means. This interval is the 95% confidence interval. The length of this interval represents how much error there is in our sample mean. Since, as stated above, smaller samples produce more sampling error than larger samples, **smaller samples will have wider confidence intervals** as compared to larger samples. As the sample size increases, the width of the confidence intervals will decrease.

The concept of standard error and confidence intervals becomes important when we are trying to determine if there are statistically significant differences. For example, "average intelligence" is considered a score of 100 on an IQ test. A researcher is interested in determining if first year college students are smarter than "average." He takes a sample of nine first year college students, and the mean IQ for these nine students is 109. However, since it was a small sample size (n=9), there is a wide confidence interval that ranges from 99 to 119 (this is usually stated in publications as: 95% CI: 99-119). Since the confidence interval contains an IQ of 100, we would conclude that there is no statistical difference and state that first year college students are not smarter than average (see **Figure 1**). However, if the researcher had originally collected a larger sample of 100 first year college students, which had produced a mean IQ of 109, the width of the 95% confidence interval would shrink to 107-111 (95% CI: 107 to 111). Note that this confidence interval is not very wide, and indicates relatively low sampling error with the larger sample size. Since the confidence interval does not contain the IQ score of 100, we would conclude statistical significance and state that it is likely the first year college students are significantly more intelligent than average (see **Figure 1**). Since a 95% confidence interval was used, then would translate into a p value of $p < 0.05$. This is derived since there is only a 5% chance that scores outside the 95% confidence interval belong to that specific distribution. The 5% is converted to proportions and represents the p level of $p < 0.05$. If a 99% confidence interval was used, than the p level would be $p < 0.01$.This example illustrated the problem with small sample size: the confidence intervals may be too wide to “detect” or “see” differences that are actually present (this is a **Type II error**).[2]

This translates into the concept of “statistical power.” Statistical power is the chance of finding statistical significance when a true difference is present. Since we often assume our interventions will make a difference, this translates into “the chance of finding statistical significance.” Often researchers will ask, “Is the study powered correctly?” or “Does the study have the power to detect such small changes?” Piecing the logic together on sample size, sampling error, and confidence intervals, the “power” of a study will be determined primarily by the sample size. Larger studies have the “power” to detect small differences or associations since the width of the confidence intervals will be narrow. [3] However, sample size is also directly related to resources and the cost of running a study. Oftentimes, large sample sizes are an expensive luxury. Thus, researchers try to plan the number of subjects they need for a study prior to starting the study. If the researcher has a general idea of the effect of the study (the differences study will produce), the researcher can then plan on how narrow his confidence interval will need to be to determine statistical differences. The spread of the confidence interval will then help the researcher determine the appropriate sample size. **An appropriately powered study will have an 80% chance of finding statistical significance if the differences the researcher predicts are produced by the study**.[3]

3. Selecting the Correct Statistical Test

When deciding on a statistical test it is first important to understand the difference between a dependent variable and an independent variable, and how these variables can be measured. The “dependent” variable or “DV” is the variable that is being measured for change. The dependent variable is thought to be influenced by the treatment or by other variables in the study, and is sometimes call the outcome variable. The “independent” variable is either the variable manipulated by the researcher (in experimental designs, this is group assignment) or the variable or variables thought to influence the dependent variable or outcome variable. For example, consider a study investigating a new drug to treat erectile dysfunction (ED). If the study was comparing the new drug versus a placebo, and investigating if this new drug improved erectile function, the independent variable would be group assignment (new treatment vs. placebo) and the dependent or outcome variable would be the measure of erectile function. In a different example, if a study was investigating the association that age and diabetes have on urinary flow rate, age and diabetes would be the independent variables and the urinary flow rates would be the dependent or outcome variable.

The researcher then needs to develop an “operational” definition for all of the variables in the study. An operational definition is how the researcher chooses to define or measure the variables. For example, erectile function can be defined in a number of different ways. One measure of erectile function is the International Index of Erectile Function (IIEF), which has an Erectile Function Domain (EFD). Erectile function can be defined on a continuum with the EFD, where lower scores indicate poorer erectile function and higher scores indicate better function. Importantly, if erectile function is operationalized in this way it would be measured as a **continuous variable** (EFD scores range from 6-30). The researcher may also choose to measure a variable as a binary variable, as either a condition exists or it does not. One way to measure erectile function as a binary (dichotomous) variable is to define a subject as having ED or not having ED. The EFD has a cutoff score which indicates the presence of erectile dysfunction (EFD < 26) versus no erectile dysfunction (EFD ≥26). The researcher could use this cutoff method and this would produce a binary variable (ED: yes/no).

3.1 How is the Dependent Variable Measured: Continuous or Binary?

The way the dependent variable is measured will help decide the type of statistical tests that are used to analyze the data. Variables that are measured as continuous variables produce descriptive statistics such as means, standard deviations, and variances. If the dependent variable is measured as a continuous variable, the ability to calculate means and standard deviations allow the researcher to use a certain group of statistical tests. These groups of tests are: t-tests, Analysis of Variance (ANOVA), correlation coefficients, and linear multiple regression (Newton, 2013).

Variables that are measured as binary variables produce descriptive statistics such as frequency, or the number of subjects in each category, which are often reported as percent (30% have ED, 70% do not have ED). Since binary outcomes are reported in percentages, it is not possible to calculate statistics such as means, standard deviations, or variance. As a result, the researcher cannot use the same statistical procedures that are used when the dependent variable is measured as a continuous variable. The types of statistical procedures which are appropriate when the dependent variable is measured as a binary variable are: chi-square, Fisher’s exact test, point-biserial correlation coefficients, and logistic regression. [2]

3.2 What Is The Study Design?

Once you have determined how the dependent variable is measured, the next consideration is the design of the study. This discussion will discuss two basic types of study design: Group Designs and Correlational Designs.

3.2.1 Group Designs

Group designs are experiments in which the researcher is comparing two or more groups, and assessing if the groups are different in terms of the dependent variable. Randomized controlled trials (RCTs) fall into this category. Since these designs use random assignment and control for important elements of the study, these are considered “**experimental designs**,” which means the researcher can conclude cause and effect from these studies. “**Quasi-experimental designs**” are also group designs. These types of designs compare two or more groups, but do not use random assignment or control for important elements of the study. Retrospective cohort studies also can be structured as a group design comparing two or more groups. These types of studies cannot conclude cause and effect because of the lack of randomization and proper control. For the purpose of this discussion, “ **repeated-measures designs**” will be considered group designs. A repeated-measures design follows one group over time and assesses the group two or more times (repeating the measure). A simple example is a pre-post design where one group is assessed prior to an intervention and the same group is assessed following the intervention. You might hear these designs described as “using subjects as their own control.” Cause and effect cannot be concluded from repeated-measures designs because there is no separate control group. [4]

3.2.2 Correlational Designs

Correlational designs do not compare groups. Correlational designs are interested in the relationship between two or more variables within a group. An example of a simple correlational design is finding the association between age and erectile function. Note that there are no group differences in this design, and the focus is on the relationship between variables within a group. Correlational designs can explore the relationship between multiple independent variables and one dependent variable (how age and diabetes are related to erectile function).[4]

4. Statistical Procedures When the Dependent Variable is Measured as a Continuous Variable

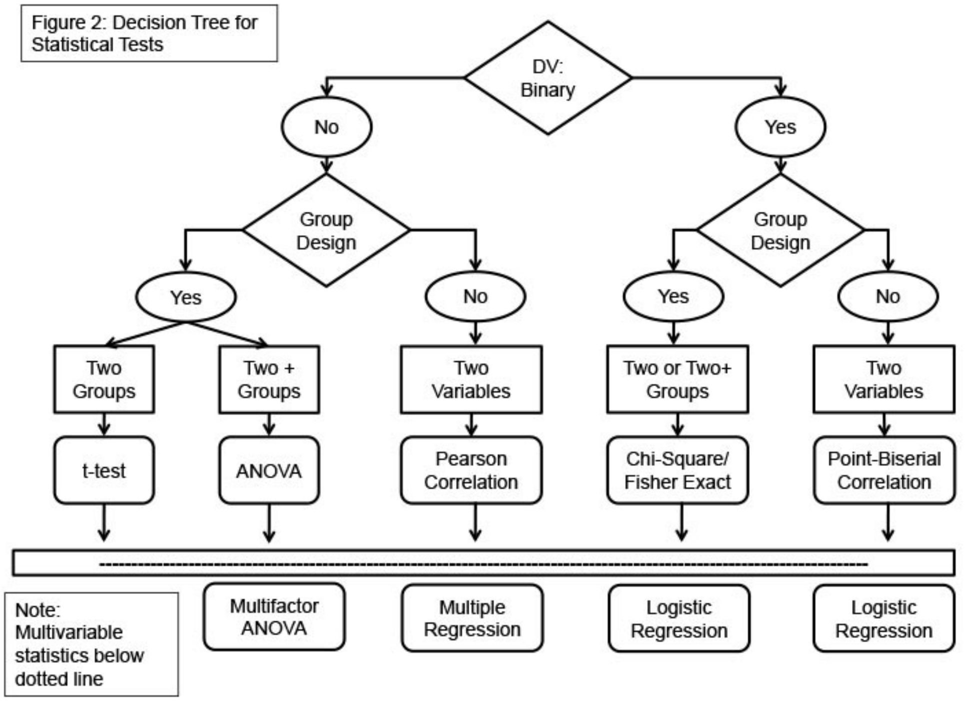

Figure 2: Decision Tree for Statistical Tests

When deciding on a statistical test, the first question to ask is: how is the dependent variable measured? Once this question is answered, the next important consideration is the research design of the study. Below are the types of statistical procedures used for specific research designs when the dependent variable is measured as a continuous variable. See **Figure 2** for a decision tree when deciding on a statistical test.

4.1 Group Designs

If the dependent variable is assessed as a continuous variable, and the design is a group design, the first statistical tests to consider are t-tests or Analysis of Variance (ANOVA). The decision between a t-test and ANOVA will usually depend on how many groups or repeated-measures the researcher is comparing. If the researcher is comparing two groups or two repeated-measures, the first test to consider is a t-test. If there are more than two groups or repeated-measures, an ANOVA may be the appropriate test. These tests are considered parametric test, which make certain assumptions about the probability of the distribution. One important assumption is that the distribution is normally distributed. If the underlying assumptions are not valid (e.g., the data is not normally distributed), then non-parametric tests would be more appropriate. The non-parametric tests which could be used in place of t-tests are the Mann-Whitney U test and Wilcoxon signed-rank test. The Kruskal-Wallis one-way analysis of variance is a non-parametric test which could be used instead of an ANOVA[5].

4.1.1 T-tests

There are two t-tests that will be covered in this discussion. The most common t-test is the independent measures t-test, which compares means from two independent or different groups. For example, if a researcher was investigating a new drug to treat ED, and compared the means of erectile function between the two groups (a group of men who received the new treatment versus a different group of men who received placebo), then an independent measures t-test would be the correct statistical test. This is appropriate as long as erectile function (the dependent variable) was measured as a continuous variable. One important assumption of a t-test is that the dependent variable, erectile function in this example, is normally distributed. If this assumption is violated, then the appropriate non-parametric test would be the Mann-Whitney U test.

The second t-test is the related samples t-test (also known as repeated measures t-test, paired samples t-test, or matched samples t-test). This t-test would be used for a repeated-measures research design. A common example is a pre-post design where a researcher is examining changes before and after an intervention. To illustrate, if a researcher was testing the efficacy of a new drug to treat ED in one group, and measured erectile function before the start of the drug treatment (pre-intervention), and then assessed the men after six weeks of treatment (post-intervention, with the same measure), a related-samples t-test would be the appropriate t-test. This test also assumes that the dependent variable (erectile function) is normally distributed. If normality cannot be assumed, then the appropriate test non-parametric test would be the Wilcoxon signed-rank test.

4.1.2 ANOVA

ANOVAs apply to the same type of research designs that are appropriate for t-tests, but are used when a study is comparing three or more groups or repeated-measures. The independent-measures ANOVA (the “standard” ANOVA and usually just called ANOVA), is used when there are three different or independent groups. For example, consider a study where a researcher is investigating a new treatment for LUTS. If the researcher compared three groups (a group of men who received a novel drug versus a group of men who received the standard agent versus a different group of men who received placebo), an independent-measures ANOVA would be the appropriate statistical test. As with t-tests, we assume the dependent variable is normally distributed. If normality cannot be assumed, the appropriate non-parametric test would be that Kruskal-Wallis one-way analysis of variance.

The second type of ANOVA applies to a repeated-measures research design, which has three or more repeated measures. A common example of this is a pre-post design where a researcher is examining changes before and after an intervention, with a third follow-up after the intervention to determine if the benefits of the intervention were sustained over time. Consider the example of researchers testing the efficacy of a new treatment for nocturnal enuresis. The investigators assessed the sample of men before they were given the new treatment (pre-intervention), and assessed them again six weeks following the start of treatment (post-intervention, with the same measure). They were then assessed again six months later. In this case, the researcher would have three repeated measures on the same sample, and a repeated-measures ANOVA would be the appropriate test.

ANOVAs can also be “multifactorial.” The language associated with ANOVAs labels independent variables as “factors.” Thus, a multifactorial ANOVA is an ANOVA, which incorporates more than one independent variable. Consider the example above from the independent-measures ANOVA where the researchers were comparing three groups (new treatment, standard treatment, and placebo) and determining if there were differences between the three groups on LUTS measured as a continuous variable (IPSS score). The group assignment would be considered the independent variable. It is one independent variable with three levels or groups. The researcher might also be concerned how diabetes impacts outcome in this study. The researcher could add the presence of diabetes as a second “factor” or independent variable. This could be assessed as diabetes versus no diabetes, and would have two levels. The researcher would then have two factors or independent variables (treatment group assignment with three levels and presence of diabetes with two levels), and they would be determining the impact of these independent variables on the dependent variable. The appropriate statistical test would be a multifactorial ANOVA and the levels of the independent variables would be used to describe the test. Thus, this would be described as a 3 X 2 multifactorial ANOVA. This is also an example of a “multivariable model” where there are two or more independent variables and one dependent variable.

This multifactorial ANOVA would produce two main effects. The first main effect would be the determination of statistical significance in IPSS score change between the treatment groups, and the second main effect would be determination of statistical significance in IPSS score change based on diabetes (yes/no). The benefit of a multifactorial ANOVA is that it also determines if there is an “interaction” between the main factors. An interaction is where the effect of one factor depends on the levels of second factor. In this example, there would be an interaction if the benefit or effect of the new medication on LUTS is different based on the presence of diabetes.

4.2 Correlational Designs

4.2.1 Correlation

When assessing the relationship between two variables, the appropriate test is usually the Pearson correlation coefficient. This is the standard correlation that is reported most often. For example, if a researcher is interested in the relationship or association between age and erectile function, the correct statistical test would be the Pearson correlation coefficient ("r"). The value of a correlation coefficient ranges from zero (which indicates no relationship) to one (which indicates a perfect linear relationship). The negative or positive sign indicates the direction of the relationship. A good guide to interpret the strength of the correlation is: small, r=0.10; medium, r=0.30; and large r=0.50.[3] Another way to interpret the importance of the relationship between variables is to calculate r squared. When a researcher squares the correlation coefficient it produces the coefficient of determination. When multiplying the coefficient of determination by 100, this represents the percent of variance one variable explains of a second variable. [6]

In addition to the Pearson correlations coefficient, another common correlation is the Spearman's rank correlation coefficient. This can be used when the variables are measured in "rank" order (e.g., 1 st, 2nd, 3rd), and is also a non-parametric correlation coefficient that can be used when underlying parametric assumptions for the Pearson correlation coefficient (e.g., normal distribution) are not met.[5]

4.2.2 Multiple Linear Regression

Many times a researcher is interested in the association of multiple variables related to one outcome variable. This would be another example of a "multivariable model" where more than one independent variable is associated with one dependent variable. A simple example of this is if a researcher is interested in age and number of vascular co-morbidities, and how they both are related to erectile function. We know age is related to erectile function and we also know the presence and number of vascular co-morbidities are related to erectile function. The difficult part is that age and vascular co-morbidities are also related to each other, and we need to control for this relationship in some way. Age will have an impact on erectile function; however the relationship may not be related to age alone. Since age is related to vascular co-morbidities, the only reason age may impact erectile function is because of an increase in vascular co-morbidities as a person ages. Statistical procedures can mathematically "control" for the impact each independent variable has on each other. In doing so, the procedure would remove the association between vascular co-morbidities and age (and inverse). The procedure "purifies" each variable. The association between age and erectile function is then assessed without the impact of vascular co-morbidities. Simultaneously, the relationship between vascular co-morbidities and erectile function is assessed without the impact of age. From a mathematical perspective, this can be very complicated. However, from a conceptual perspective, this is easier to understand. Regression would allow us to view the independent contribution of age and the independent contribution of vascular co-morbidities on erectile function.

The researcher can enter a number of independent or predictor variables in the regression model predicting the dependent variable. The general rule of thumb is that there should be at least 15 subjects per predictor variable. Ratios below this level can lead to spurious results. All regression-based models work in the same way in that they simultaneously control for the relationships among the independent/predictor variables. **The difference between multiple linear regression and logistic regression is how the dependent variable is measured**. If the dependent/outcome variable is continuous, the researcher should use linear multiple regression. If the dependent/outcome variable is binary, the researcher should use logistic regression.

Of note, both the multifactorial ANOVA and multiple regressions are multivariable models. The ANOVA statistical tests are actually very specific examples of a multiple regression, and many of the analyses performed using ANOVA can also be conducted with multiple regression. The results reported from an ANOVA analysis are generally more intuitive and this, in addition to common practice, suggest the ANOVA models are often times the first choice for analyzing data from group designs. However, there may be times when you are reading literature and analyses that you thought should have been run with ANOVA, but were actually run with multiple regression.

5. Statistical Procedures When the Dependent Variable is Measured as a Binary Variable

5.1 Group Designs

5.1.1 Chi-Square And Fisher's Exact Test

These group designs are similar to the group designs described above, except that the dependent variable is a binary variable. For example, if erectile function is measured as ED versus no ED, and the design of the study is exploring the difference between a group of men who received a new medication for ED versus a group of men who received a placebo, the appropriate statistical test for this data is **Chi-square**. This test will indicate if there are significant differences in the percentage in each group (percent of men with ED in the treatment group vs. the percent of men with ED in the placebo). **Fisher's exact test** should be used when the number of subjects in any subgroup is below five.[7] The data in this example will produce four subgroups: men with ED in the placebo group, men without ED in the placebo group, men with ED in the treatment group, and men without ED in the treatment group. If the sample size for any of these subgroups (often times called "cells") is below five, then a Fisher's exact test should be used instead of the Chi-square test.

When discussing group designs in which the dependent variable was continuous, the number of groups or repeated measures was an important consideration when deciding between a t-test and an ANOVA. However, when the dependent variable is binary, Chi-square or Fisher's exact test continue to be the appropriate test with increasing groups or repeated measures. For example, if the researcher was comparing three groups (a new medication group vs. a standard treatment group vs. a placebo group), and the dependent variable of erectile function was measured as binary (ED vs. No ED), the appropriate tests to consider would still be Chi-Square or Fisher's exact test. This is also true for repeated-measures designs with two or more repeated measures, as the first tests to consider are Chi-square and Fisher's exact test.

5.1.2 Logistic Regression

When considering group designs, researchers may also want to add one or more independent variables in addition to the independent variable of group assignment. As in a previous example where a researcher was testing a new drug for ED and comparing a new drug group versus placebo group, the researcher may also want to determine if diabetes also impacts the dependent variable. In this case, there would be the independent variable of treatment group (new treatment group versus placebo group), and an additional independent variable of diabetes (yes/no). If the dependent variable was binary (ED vs. no ED), logistic regression would be the appropriate test. Logistic regression is similar to multiple regression discussed above, except that in logistic regression the outcome variable is binary. When a multi-factorial ANOVA was discussed above, we considered a similar design where the dependent variable was measured as a continuous variable. In the multi-factorial ANOVA, the interaction between group assignment and diabetes was tested. There is also a way to test for an "interaction" between group assignment and diabetes in regression models, and this interaction could be tested with logistic regression.

5.2 Correlational Designs

5.2.1 Point-Biserial Correlation

When a researcher is interested in the association between two variables, the general construct is a correlation coefficient. When the independent variable is continuous and the dependent variable is binary, this type of correlation is called a point-biserial correlation. For example, if a researcher was interested in the relationship or association between age and erectile function, where erectile function was measured as ED versus no ED, the correct statistical test would be the point-biserial correlation. This is similar to a Pearson correlation and is interpreted in the same way. The only functional difference is that from a technical perspective, the correlation should be labeled as a point-biserial correlation.

5.2.1 Logistic Regression

As stated above in the discussion of linear multiple regression, many times a researcher is interested in the association of multiple independent variables related to one dependent variable. Using the example above from linear multiple regression, a researcher may be interested in age and number of vascular co-morbidities and how they are both related to erectile function. If erectile function (the dependent variable) was measured as a binary variable (no ED versus ED), the appropriate statistical test is logistic regression. The regression analysis would control for the relationship between age and vascular co-morbidities, and would provide the researcher with the independent association between each independent/predictor variable and the dependent variable.

Presentations

References

1 Sarndal, C., Swensson, B., and Wretman, J., Model Assisted Survey Sampling. 1st ed. 1992, New York, NY: Springer.

2 † Newton, R.R. and K.E. Rudestam, Your statistical consultant : answers to your data analysis questions. 2nd ed. 2013, Thousand Oaks: SAGE Publications. xxii, 356 p.

This is one of the more “user friendly” statistical references. This book goes through statistical procedures and also discusses the short-comings of the most frequently used statistical tests.

3 Ellis, P.D., The essential guide to effect sizes : statistical power, meta-analysis, and the interpretation of research results. 2010, Cambridge ; New York: Cambridge University Press. xvii, 173 p.

4 Rosenthal, R. and R.L. Rosnow, Essentials of behavioral research : methods and data analysis. 3rd ed. 2008, Boston: McGraw-Hill. xxi, 842 p.

5 Corder, G.W. and D.I. Foreman, Nonparametric statistics for non-statisticians : a step-by-step approach. 2009, Hoboken, N.J.: Wiley. xiii, 247 p.

6 Steel, R.G.D., J.H. Torrie, and D.A. Dickey, Principles and procedures of statistics : a biometrical approach. 3rd ed. McGraw-Hill series in probability and statistics. 1997, New York: McGraw-Hill. xx, 666 p.

7 Agresti, A., A Survey of Exact Inference for Contingency Tables. Statistical Science, 1992. 7(1): p. 131-153.