IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 28, Issue 5. Ser. 4 (May. 2026), PP 58-73
www.iosrjournals.org

Guide To Choosing Appropriate Statistics For Research Analysis

Dr. Emmanuel I. Ogbenjuwa
Director, Center For Entrepreneurship Studies, Otukpo-Nigeria

Dr. Raymond O. Obinozie
Accounting Unit,
Dekalb County Juvenile Court,
Georgia-USA

Abstract
So many researchers and graduate students get enmeshed with the data analysis section of their dissertations and often resort to rule of thumbs in deploying incoherent statistics to analyze data. The results are misleadingg outcomes, resulting in Type 1 and/or II errors. Hypothesis testing yield false narratives and the study conclusions, totally unreliable. In this study, we offer basic guides to choice of statistics that best answers your research questions and relevantly test your hypotheses. The integrity of a study is the ability to inform the research community, with high degree of reliability, the outcome that is objective and dependable. Statistics is the life wire of research and a poor deployment can mess up all the efforts of the researcher. We highlight the importance of statistics with particular reference to social science research, discussion on data analysis with examples on data cleaning and management, testing for instrument validity and reliability statistically, determining the sample size with the help of statistics. Assumptions highlighted are outliers, normality, missing data, multi-collinearity, and homogeneity of variance/regression. The study closed with visits to some main effects parametric statistics such as Analysis of variance (ANOVA), Analysis of co-variance (ANCOVA), Factorial ANOVA, Correlation, Pearson correlation and Regression analysis all with practical examples. We are hopeful the study will add to knowledge in the area of data analysis. We are however mindful that ours is not exhaustive discuss on this important phenomenon.

Keywords: *Assumptions; Multi-collinearity; Type 1 and 11 errors; Statistical significance*

Date of Submission: 21-05-2026 Date of Acceptance: 31-05-2026

I. Introduction

This is a general guideline for choosing a statistical analysis and should not be construed as hard and fast rules. Usually your data could be analyzed in multiple ways, each of which could yield legitimate answers. The number of dependent variables (sometimes referred to as outcome variables), the nature of your independent variables (sometimes referred to as predictors) significantly determine the choice of statistics. You also want to consider the nature of your dependent variable, namely whether it is an interval variable, ordinal or categorical variable, and whether it is normally distributed.

Importance of Statistics

Robust statistics are indispensable to social science research. This is largely because Social science demographic data do not follow the bell-shaped curve but are frequently heavy-tailed, skewed, or multimodal. The application of weak statistics to non-normal distributions heavily increases the risk of false positives. Robust statistic general increases the internal validity of designs and lead to reliable outcomes.

Traditional models (like ordinary least squares regression) are highly sensitive to extreme values. Robust techniques—such as trimmed means or robust regression—accurately represent the majority of the data without letting a few extreme responses skew the entire outcome. Strong statistics equally ameliorate for the weakness of the non-experimental designs and help transforms categorical values to interval or ratio levels of measurement. Statistical methods play a massive role in the "credibility movement" within the social and behavioral sciences. They produce more trustworthy parameter estimates across replication studies, even when samples are contaminated by measurement errors or extreme participant responses.

Data Analysis

Statistics helps researchers organize data in a meaningful way. According to Frankfort-Nachmias and Nachmias (2008) descriptive statistics is used to summarize and organize data, while inferential statistics serves

the purpose of allowing researchers to make inferences about the observation. The result of statistics will allow the researcher to generalize findings and ameliorate weakness inherent in non-experimental design. Frankfort-Nachmias and Nachmias (2008) held that the validity of quasi-experimental design can be ensured by using statistical data analysis techniques to achieve the necessary control.

Validity and Reliability of the Measuring Instrument

The intrinsic value of an article is measured by the utility derived from it by readers. According to Bell, Distefano, and Morgan (2010), the essential features of a scientific inquiry are transparency and replication. Research is worthless if it does not extend or add to knowledge or contribute to building the body of knowledge. To achieve this purpose, researchers must ensure the study receives the confidence of the research community in respect to the reliability of the instrument and the validity of the design and the result. Bleijenbergh, Korzilius, and Verschuren (2011) argued that practice-oriented research is necessary to develop independent criteria for evaluating the quality of practice. Smith (2012) held that group designs do more to minimize internal threats to the validity than single-case designs. According to Smith, researchers and scholars presently utilize criteria developed for theory-oriented research in assessing the quality of practice-oriented research.

The Likert scale, for instance has been used by social scientists with a long history of reliability. Levy, Richardson, Lounsbury, Stewart, Gibson, and Drost (2011) employed the 5-point Likert scale to measure personality traits along with Guttmann scale with reported reliability index of 95% factor loading.

In order to reduce measuring error and in effect enhance the ability of the research findings, researchers ensure that strategies be enshrined to enhance the validity of the design. According to Frankfort–Nachmias and Nachmias (2008), validity is ensuring that researchers indeed measure what they think they are measuring. It is possible to set out to measure a particular construct but end up measuring another, thereby inputting error into the design. A researcher may set out to measure attitude but will unknowingly measure intelligence or ability.

Frankfort–Nachmias and Nachmias (2008) argued that the validity of measurement does influence the validity of the conclusion drawn about the hypothesis. To ensure content validity, the researcher subjects the questionnaire to peer review by independent scholars. According to Frankfort–Nachmias and Nachmias (2008), specialists in the field of study can be consulted to evaluate the questionnaire to ensure that it captures the construct it intends to measure.

Reliability of the instrument, on the other hand, concerns with the consistency of the measuring instrument to bring out same result at each use. Reliability of a measuring instrument is measured by its consistency in producing the same results under similar circumstances. Frankfort-Nachmias and Nachmias (2008) traced the application of Likert scale in social science to the development of a 24-item scale by Wayne Kirchner to measure attitudes toward employment of senior citizens. The SPSS has features to assess the reliability of the scale through the internal consistency coefficient. According to Green and Salkind (2011), internal consistency coefficient (ICC) is conducted through the reliability analysis program, a test which evaluates the rate of consistencies of the instrument. This represents an estimate of scale's reliability. The application of factor analysis procedure allows for the reduction of large numbers of overlapping variables. Green and Salkind (2011) argued that factor analysis "can yield factors that represent different dimensions of a broad conceptual system" (p. 313).

To conduct the factor on the measuring scale, the SPSS tool is helpful. It can be used to evaluate the measuring instrument for content, predictive, and construct validity in order to reduce threats to validity. A reliability test ensures consistent responses of the instrument across constructs, while validity checks ensure that meaningful and useful inferences can be drawn from using the instrument.

Sampling Strategy

Sampling strategy is the steps taken in selecting participants for the study to ensure that selected samples represent the population of interest. Frankfort–Nachmias and Nachmias (2008) argued that it is impossible, impractical, or extremely expensive to collect data from all the units in the population; hence, partial or selected units are chosen to observe, study, and make a generalization of the result to the entire population. For the generalization to be valid, the selection process must be free from bias and subjectivity. That means that every unit within the population must stand equal chance of being selected. If the researcher selects predetermined samples, this only confirms that he/she observes only what he/she wants to observe. It then means that he/she comes to study with a preconceived idea, and in desperation to confirm it, selects the samples which bear the characteristics he/she desires to observe. Bernardi (2011) used a deductive quota-sampling strategy adapted to satisfy the contending demands of variety of family formations, experiences, or characteristics that allow for comparability.

According to Frankfort-Nachmias and Nachmias (2008), stratified sampling ensures that different groups within the population, which exhibit common characteristics with the population, are fairly represented in each stratum.

Frankfort-Nachmias and Nachmias (2008) held that there is a need to have the idea of the standard deviation of the entire population when calculating the sample size from a population.

Sample size = $n = \frac{S^2}{SE}$

Where n = Sample size

S^2 = Variance of the (variable) sample

SE = Standard error of the mean.

Frankfort-Nachmias and Nachmias (2008) denounced fallacies associated with determining sample size, such as, "increase in samples will increase the precision of the sample result, or that the sample size must be a percentage of the population". Another misconception jettisoned by Frankfort-Nachmias and Nachmias (2008) is that the sample size must be 2000. Frankfort-Nachmias and Nachmias (2008) held that in calculating the sample size, the researcher must take cognizance of the standard error. The standard error is a statistical tool that indicates the closeness of the sample result with the parameter. Sample or standard error is calculated as:

$SE = \frac{S}{vn}$

Another statistic used to determine the sample size is the confidence interval. A confidence interval measures the degree or the chance that the postulation is correct. A 95% interval means that there is 5% chance that estimates will be wrong. According to Frankfort-Nachmias and Nachmias (2008), confidence interval and standard errors are commonly employed in surveys and opinion polls as checks against forecast errors. Aczel and Sounderpandian (2002) argued that the true population variances within the different strata are not usually known, and have to be estimated. Aczel and Sounderpandian (2002) held that the population variance could be estimated as:

$S^2_i = \frac{\Sigma(x - x)^2}{ni\text{-}1}$

Researchers employs the G Power statistical tool to calculate the sample size at 90% confidence interval. The G power is used to calculate the sample size when the effect size is given with known alpha beta of the critical value (Faul, Erdfelder, Lang, & Buchner, 2007). *F* test for Linear Multiple regression fixed model, for 2 numbers of predictors, a priori sample computation is: At the alpha of $(\propto errpro) = .05$, power $(1\text{-}\beta pro) = .90$, medium effect size of .15, yield the following: Critical *F* (3.1038387), the sample size is 540. Sherperis (RSCH 8250: Power and effect size) argued that prior to conducting research, it is incumbent on the researcher to conduct power analysis to aid in the determination of the sample size. The use of G power statistical tool reduces the chances of making type II errors. The calculated alpha level and the power measure the likely occurrence of error, and estimate the statistically significant difference in the sampling process.

In order to reduce the chances of making type 11 errors, the sample size should be large enough to represent the population of interest. According to Tailor (2012), a fairly easy way to calculate the sample size is to add the levels of the predictive and outcome variables and multiply them by 90 as a basic power requirement $(1\text{-}\beta pro)$.

Frankfort-Nachmias and Nachmias (2008) held that non response introduces bias to study, and offered techniques to compensate for such bias, such as substituted values or data cleaning to remove redundant entries.

Data Management Issues

Before you begin to clean your dataset to get ready to conduct your statistical analyses, you need to go over some important data management issues which will aid in the organization of data and the planning process for data analysis.

The first thing is to create a codebook for the dataset. The codebook contains the following: variable names, variable labels, value labels, and a list of any changes made to the dataset such as creating new variables or fixing raw variables. It is very helpful and can prevent you from making a mistake when going to conduct your analysis. You can easily create a codebook using the table function in Microsoft Word or using Excel.

The next thing is to create an analysis plan. Before tackling your data analysis, is strongly suggested you create a detailed analysis plan that you can follow when you are conducting your analyses. Your analysis plan should include your plans for data cleaning, your variable modifications and hypotheses research testing, as well as which graphs, tables and figures you will use to display your data.

This should be a step-by-step list of what you need to do to your data for your particular research project. This can be done as an outlying form in Microsoft Word.

And the last thing is your data security. As a researcher, you have a responsibility to protect your research data. You should always make sure your participant's personal information cannot be discovered by others. Always keep your files password protected and keep your participants' personal information such as names, contact information separate from the research data.

Assumptions of Statistics

Outliers

One of the first things to do when cleaning data in preparation for conducting inferential statistics is to look for outliers in the variables. Outliers are scores in your variables that are extreme in value, either greatly higher or lower than all the other scores for that variable. Most statisticians state that outliers are any values which have standardized scores in excess of the absolute value of 3.29, which is either positive or negative 3.29 for that variable. So, in other words, a score more than three standard deviations from the mean. Outliers can lead to both type 1 and type 2 error, thereby making your solution unreliable.

An easy way to search for outliers is to create standardized scores, z-scores for all of your variables. You can easily do this under the 'descriptives' tabs in SPSS. After you've created your standardized scores, run frequencies on your new standardized scores. Any variables with values in excess of the absolute value of 3.29 have outliers.

You can delete your outliers from your variables, but this reduces your sample size and it's not recommended. You can transform your variables. So, you could multiply the variable by its logarithm, square root or inverse. However, this makes the variable more difficult to interpret. What I recommend is that you modify your outlier so it's not as extreme. Sometimes this is called (Windsorizing). You can make your outliers for a variable one unit larger or smaller than the next most extreme score. What you do is you find the value whose standardized score is closest to the absolute value of 3.29 without going over that and then add one to that value. This value will now be used to replace your outliers for that variable.

Normality of Variables

Normality refers to a distribution of scores where the mean equals the median equals the mode. It is a bell-shaped curve. Most inferential statistics have an assumption of normality of variables. When your variable is normally distributed, it means it is not skewed or kurtotic. A lot of times, outliers will lead to non-normality, so always deal with outliers first and then check for non-normality.

There are many ways to check for normality of variables. Visually, you can run histograms of your variables and you can get histograms under the frequencies tab in SPSS and then look to see if your variables are normally distributed. If they are, that bell-shaped curve. You can also check the values of skewness and kurtosis for each of your variables. Then take your value of your skewness and your kurtosis and you can get these, this information under your 'descriptives' tab in SPSS and divide the skewness and divide the kurtosis by its standard error. And if that resulting value is greater than the absolute value of 2, then you have violated this assumption.

If after dealing with outliers and you still have non-normality, you can transform your non-normal variables to make them more normal. You could take the logarithm, square root or inverse of that variable and each of these can easily be done using the compute function in SPSS. And these usually try to – these do make the interpretation of your variables a little bit more difficult, but they can reduce the non-normality of your dependent variable.

However, if you don't want to modify your non-normal variables, some of the parametric statistics such as independent t-test and one-way repeated measures, ANOVA, there is a non-parametric equivalent that you can use to- with your non-normal variable.

Missing Data

The next assumption to discuss is the missing data. If you have less than 5% missing data in a variable, you usually don't need to worry about your missing data having any impact on your results. However, if you have more than 5% missing data in a variable, you should think about replacing that missing data with an estimated value. Missing data will lower your sample size and make your results less generalizable. To find which variables have more than 5% missing data, you can run frequencies on each variable and in the frequency table it will tell you how much missing data you have for each variable.

There are many ways to deal with missing data. First you can delete the missing data. Either you delete the variables that have too much missing data or you delete participants in your dataset that have too many missing variables. You can either use listwise deletion, which is a conservative approach which deletes participants that have any missing data plan, any variable. Or you can use pairwise deletion which deletes only participants when they have missing data for a particular analysis. And that is the preferred method of deletion if you are going to delete your missing data. You can also treat the missing data as another category or level in your independent variable and run analyses to see if those with missing data in that variable are significantly different than others on whatever your outcome variable is. Lastly, you could estimate your missing data with one of the many methods that are out there. The first one is prior knowledge. Here you just insert a value for your missing data based on your prior knowledge of that variable for your population of interest. Basically, it's a well-educated guess. Sometimes, you might know the norm of a particular variable for your particular age-group and you would use that value to replace the missing data for your participants. Another way that you can estimate missing data is

insert the mean, and this is the most popular method of replacing missing data. You can insert what we call the grand mean for that variable, which is the mean of all the participants in your dataset for that particular variable or you can insert a group mean for that variable. So, basically, you split your dataset into relevant groups for your research project and then create means for all relevant groups and then replace the missing data point for each participant based on which group they're in. And it's usually recommended that you use a group mean whenever possible over a grand mean because that's more of an accurate replacement value for your missing data. There are also many other advanced methods to replace missing data such as conducting regression analyses and expectation maximization and these are much more complicated way to replace missing data.

Multicollinearity

Multicollinearity is when you have variables that are too highly correlated with each other, greater than the absolute value of .8. Multicollinearity can increase the error in analysis and weaken that analysis. In some cases, if the multicollinearity is very high, you won't even be able to find the solution for your statistical analysis.

To find multicollinearity, run bivariate correlations, either Pearson or point biserial, depending on the scale of measurement of your variables, on all pairs of variables to look at their correlations. And if any of those correlations are greater than point – the absolute value of .8, you know you have multicollinearity. To fix multicollinearity, you can delete any of the variable pairs or just one of those variables that are too highly correlated from your analysis, and this is a conservative approach. Or you can combine the variable pairs into one variable and use that new variable in your analysis. So, for example, if you wanted to look at two variables such as physical abuse and psychological abuse and use those as predictors in a multiple regression but yet they are too highly correlated with each other, you could then create a combined variable where you add physical abuse and psychologic abuse and create an average score for that and then use that average score as an independent variable in that multiple regression.

Homogeneity of Variance

Homogeneity of variance refers to the expectation that the variance in one level of your independent variable should be the same as the variance at all other levels of your independent variable on that dependent variable of interest. If you violate this, you have heterogeneity of variance or heterogeneity. To find out if you have homogeneity of variance, you should conduct a Levene's test. This option is under the Options tab for most of your analyses in SPSS. And if you have a significant Levene's test, then you have violated this assumption.

For many of analyses, if you have a large enough sample size, your analysis is robust to this assumption, meaning a slight to moderate violation of this assumption will not greatly impact your results, so you don't have to worry about it if you have violated this assumption. However, if you are worried about the impact of heterogeneity of variance, you can transform your variables. So, taking the square root, the logarithm of the inverse or you could select a more stringent alpha level – for example, .01 instead of .05 and then use that for your analysis.

Homogeneity of Regression

The last assumption I want to go over is homogeneity of regression. This assumption assumes that the slope or steepness of the regression between your dependent variable and the covariate is equal for each level or group of your independent variable. In other words, the relationship between your dependent variable and the covariate should be the same for each level or group of your independent variable. Violation of this assumption signifies that you have a significant interaction between your covariate and the independent variable on that dependent variable. If you have heterogeneity of regression, you should not be using that particular covariate in your analysis. You can test for homogeneity of regression by using the general linear model analysis in SPSS. Click on Model and then Custom and insert all the variables and their interactions into the model. If any of the interactions that include the covariate are significant, then you have violated this assumption.

Unfortunately, you can't just modify your variables to fix a violation of this assumption. You can only delete the covariate from the analysis. You can, however, make your covariate into a categorical variable and use it as another independent variable in your analysis. This is known as a blocking design. This will give you some information as to the impact of the covariate on the dependent variable, but you can't use the covariate as a traditional covariate in the analysis.

ANOVA

Analysis of variance, or ANOVA, is a hypothesis-testing procedure that is used to evaluate mean differences between two or more treatments, or populations. If a difference between the means is statistically reliable, or significant, the difference is expected with a certain probability to reappear if the study is replicated. A nonsignificant difference implies that you cannot rule out the possibility that mean differences that do exist in your sample data occurred by chance. One way repeated-measures ANOVA

A repeated measures anova allows the experimenter to administer different treatments to the same participants at different times. According to Morrow (RSCH 8250 Repeated measures Anova-Conceptual) a repeated measures anova enable repeated measurement of each participants, producing more than one dependent variable scores for each participants.

A repeated measures anova, also known as a within subjects anova, is an analysis that we use when we have each participant in our dataset measured more than once on a the same dependent variable. In other words, we have two or more outcome dependent variable scores for each participant. Each participant is measured repeatedly on the same dependent variable. Your dependent variable must be continuous in nature. A repeated measures anova is just an extension of a dependent t-test/paired samples t-test. With a dependent t-test, you are limited to only two scores on a dependent variable per participant. That is, your participant can only be measured twice on the same dependent variable. However, with the repeated measures anova, you can unlimited numbers of measurements on the same dependent variable per participant in your dataset.

Repeated measures anova can be one-way, which is where you have only one independent variable, that is you have repeated measurements across only one independent variable. And repeated measures anovas can be factorial, that is where you have two or more independent variables so you have repeated measurements across two or more independent variables

Repeated measures anova can also answer the question, "Are there reliable mean differences between your different levels of your independent variable." A repeated measures anova can tell us if the score on the dependent variable changes for your participants across the different levels of the independent variable.

Research scenario #1. Professor Okopi is interested in knowing if amount of coke drinking changes in his sample of first-year college students. He measures weekly coke drinking, a continuous variable, at the beginning of the first semester in a sample of 600 first-year students. He also measures weekly coke drinking at the end of the first semester and at the end of the second semester in the same sample of 600 first-year students. Professor Okopi wants to see if weekly coke drinking changes across the three timepoints in his sample of first-year students.

In this scenario, the independent variable is time and there are three levels or groups: Beginning of first semester, end of first semester, and end of the second semester. All participants are measured at each timepoint. The dependent variable, which is measured three times for each participant is weekly coke drinking. For this scenario, the null hypothesis would be that coke drinking does not change over the course of the three timepoints. The alternative or research hypothesis would be that coke drinking changes over the course of the three timepoints.

Research scenario #2. Dr. Chris wants to test the impact of caffeine on memory in his sample of 200 college students. He gives his group of college students 100 milligrams of Dr. Chris and then 30 minutes later gives them a memory test. The next day, Dr. Chris gives the same group of college students 200 milligrams of caffeine and then 30 minutes later gives them the same memory test. On the third day, Dr. Chris gives the students 300 milligrams of caffeine and then 30 minutes later gives them the same memory test. His goal is to see if memory changes across the levels of the independent variable. In this scenario, the independent variable is amount of caffeine and there are three levels or groups: 100 milligrams, 200 milligrams, and 300 milligrams. And all participants get all three levels of the independent variable. The dependent variable for this scenario is score on a memory test. For this research scenario, the null hypothesis is that memory doesn't change depending on the amount of caffeine taken. The alterative or research hypothesis is that memory changes depending on the amount of caffeine taken. Both of these scenarios that I have mentioned can easily be answered using repeated measures anova.

There are four advantages to using a repeated measures anova versus a between groups anova where there are different participants in each level of the independent variable.

- ✓ First, it minimizes individual differences because the same group of participants are being used for each level of the independent variable. In other words, it reduces error in the anova equation.
- ✓ Second, repeated measures anova is more economical because you need fewer participants compared to between subjects anova to achieve adequate power.
- ✓ Third, it is useful to study data that changes across time, something that a between subjects anova cannot do. And, lastly for advantages, repeated measures anova is more sensitive to detecting change due to smaller standard error values. In other words, it is a more powerful statistic compared to a between subjects anova.

There are two disadvantages to using a repeated measures anova compared to a between subjects anova.

- ➢ First, one disadvantage could be carryover effects. Carryover effects are when a participant's response in the second treatment or second level of an independent variable is altered by the lingering aftereffects of the first treatment, the first level of an independent variable.

For example, remember the research scenario I discussed before with the different caffeine levels and memory? A possible carryover effect that could happen in that scenario is that participants could possibly still

be impacted by the earlier amount of caffeine, the 100 milligrams, when they are being exposed to the second level of the independent variable, the 200 milligrams.

- Second, another disadvantage could be progressive error. Progressive error is when your participants' performance changes consistently over time, repeated measurements, due to fatigue or practice and not just due to your manipulation. For example, your participants could become very tired due to your repeatedly measuring them on some outcome variable due to the nature of the experiment that they are in. because of this, their performance or their score on the dependent variable is degraded due to fatigue and not whatever you are manipulating in the study.

There are a couple of things you can do as a researcher to deal with these disadvantages. First, you can counterbalance the levels of your independent variable. Counterbalancing refers to presenting the levels of your independent variable in all possible orders to different participants. For example, in the scenario with the three levels of caffeine – 100, 200 and 300 milligrams of caffeine – you can counterbalance the levels and have some of your participants be exposed first to 100 milligrams, then 200 milligrams, then 300 milligrams of caffeine; others be exposed to 200, then 300, then 100 milligrams of caffeine, etc., using all possible orders of the three levels of the independent variable. Counterbalancing can help control for carryover effects. Second, you could increase the time in between measuring your dependent variable. In other words, increase the amount of time between the levels of your independent variable. For example, in the scenario with the caffeine and memory, you can make sure there is enough time in between when they get the first dosage of caffeine and the next dosage of caffeine so that there are minimal aftereffects or fatigue due to participating. Increasing the time in between levels of your independent variable can reduce both carryover effects as well as progressive error.

Assumptions in Repeated Measures Anova

- Assumption number 1, outliers. Outliers can greatly impact your repeated measures anova equation. They can either incorrectly increase type 1 error or decrease type 2 error, the significance of your repeated measures anova. Outliers can also reduce the generalizability of your results. You should always deals with outliers in your dependent variable prior to conducting your repeated measures anova. Refer to the data cleaning assumptions for more information on outliers.
- Assumption number 2, normality of the dependent variable. The scores for your dependent variable at each measurement in your repeated measures anova must be normally distributed. That is, having no skewness or kurtosis. Refer to the data cleaning assumptions for more information on normality of variables.
- Assumption number 3 is sphericity. Another assumption is that you should have sphericity, which is equal variances in your levels of your independent variable. In other words, you need to have similar dispersion of scores on the dependent variable for each level of your independent variable. You can conduct a (Mockley's) test to see if you have violated this assumption. Refer to the data cleaning assumptions for more information on sphericity.
- Lastly, assumption number 4 is missing data. Repeated measures anova is sensitive to missing data in the variables. If you have missing data, you run the risk of lowering the statistical power of your test due to lowering sample size. You should decide on what to do about your missing data prior to running your repeated measures anova. Refer to the data cleaning assumptions module for more information on missing data. If you violate any of these assumptions, you can either address the assumptions and attempt to modify your data so you no longer violate the assumptions, or you can use a nonparametric equivalent to a oneway repeated measures anova, which is a (Friedman) anova. This statistical test is less powerful than a oneway repeated measures anova. There is no nonparametric equivalent for a factorial repeated measures anova, so my best advice would be to make sure you address these assumptions and modify your data so you're able to conduct a parametric analysis such as the repeated measures anova.

Factorial ANOVA

A one way between-subjects ANOVA has three assumptions to be met. According to Morrow (2012) one of the assumptions states that the group score must be independent of each other. That means, the scores in each group must not be related to each other. Secondly, the population must exhibit normality. A population with normal distribution has the scores that are distributed evenly below and above the mean. The third assumption of ANOVA is that the population must have homogeneity of variance.

Levene's test of equality of error of variance

According to Field (23012) Leven's test is to evaluate if there is significant difference between group variance. In the Leven's test conducted here, a non-significance result is produced so the assumption of equal variance is met. If the assumption of equal variance is not met then a further step to transform the data would be taken to equalize the variance.

Analysis of Covariance (ANCOVA)

The analysis of Covariance extends the principles of analysis of variance further by including the possible effects of unmeasured variables which may impede the understanding of the experimental manipulation which could affect the outcome variable. ANCOVA enables the experimenter control for the effect of confounding variables and be able to measure the degree to which the independent variable predicts the dependent variable with confidence.

Homogeneity of regression is assumed in ANCOVA. According to Morrow (RSCH 8250 Analysis of Covariance-Conceptual), the slope of the regression between the covariate and the dependent variable is assumed to be equal for each level of the independent variable. Another assumption is the normality of the distribution, which means that the scores of the covariate and the dependent variable must be normally distributed. There must also be the homogeneity of variance between the levels of the independent variables.

Analysis of covariance, ANCOVA, is an extension of analysis of variance where you now have an additional variable, a covariate added to the equation. The purpose of this covariate or covariates because you can have more than one, is to statistically control for the effect of that variable, the covariate in the analysis of variance equation. A covariate is a variable that you believe is correlated with your dependent variable but is not correlated to your independent variable. Covariates should be continuous, however you can have categorical covariates if they are dummy-coded. In other words, you are controlling for the amount of variance the covariate brings to the equation. You are removing that variance and then testing to see if there are group differences on your outcome variable.

With analysis of covariance, you are adjusting the dependent, the outcome, to be what the dependent variable score would be if the participants score on the same on the covariate.

Assumptions in Analysis of Covariance

There are many assumptions that first must be addressed before you can perform your analysis of covariance.

- The first assumption is outliers. Outliers can greatly impact your analysis of covariance equation. It can actually cause heterogeneity of regression. You should always deal with outliers in your dependent variable prior to conducting your analysis of covariance. Refer to the data cleaning assumptions for more information on outliers.
- The next assumption is homogeneity of regression. This assumption refers to the fact that it is assumed that the slope or steepness of the regression between your dependent variable and the covariate is equal for each level or group of the independent variable. In other words, the relationship between the dependent variable and the covariate should be the same for each level of your independent variable. This is a very serious assumption for analysis of covariance. If this assumption is violated, thereby giving you heterogeneity of regression, you should not be including the covariate in the analysis of variance equation. Refer to the data cleaning assumptions for more information on homogeneity of regression.
- The next assumption is normality of the dependent variable. The scores for your dependent variable and your analysis of covariance must be normally distributed. That is, having no skewedness or ketosis. Refer to the data cleaning assumptions for more information on normality of variables.
- Another assumption is homogeneity of variance. In most cases, this assumption refers to equal variances between your levels of your independent variable. For analysis of covariance, it refers to equal covariances. To satisfy this assumption, you must have equal variances for your covariate scores across all levels or groups of your independent variable. In other words, your covariances must be equal across your levels of the independent variable. Refer to the data cleaning assumptions for more information on homogeneity – on homogeneity of variance.
- Another assumption is multicollinearity. For this assumption, you must make sure your covariates, if you have more than one, are not too highly correlated with each other, that is, greater than the absolute value of .8. Having two highly correlated covariates can weaken the statistical power of your analysis. Refer to the data cleaning assumptions for more information on multicollinearity.
- And, lastly, an assumption that we must address before conducting an analysis of covariance is missing data. Analysis of covariance is sensitive to missing data and the variables. If you have missing data, you run the risk of lowering the statistical power of your test due to lowering sample size. You should decide on what to do about your missing data prior to running your analysis of covariance. Refer to the data cleaning assumptions for more information on missing data.

Correlation

Correlation is a statistical technique that describes a relationship between two or more naturally occurring variables. And a bivariate correlation is a correlation of just two variables.

Characteristics of a Relationship

There are three characteristics of a relationship. The first is direction. And direction can be either positive or negative. A positive correlation is, both variables change in the same direction. As one goes up, the other goes up. As one goes down, the other goes down. A negative correlation is when both variables move in opposite directions. When one variable goes up, the other variable goes down. The other characteristic of a relationship is the form of the relationship. Correlations can either be linear or nonlinear. For example, correlations can be curvilinear. And lastly, the degree of the relationship. A perfect correlation is plus or minus 1.00. When you have no correlation, that value will be zero. So your correlations can range from negative 1.00 to positive 1.00.

Pearson Correlation

A Pearson correlation is a type of correlation that measures the degree and the direction of the linear relationship between two variables which we designate as X and Y. And these variables must be continuous in nature. Some alternative names for the Pearson correlation is the Pearson R and the Pearson product moment correlation.

Sample Research Questions

Some examples of questions that can be answered using a Pearson correlation are:

Question one: Is there a relationship between alcohol use and missing class? In this case, these are both dependent variables. The first one I'll designate as X, and that's alcohol use. And the second variable I'll designate as Y, number of classes missed. The second question is, Is there a relationship between spirituality and attitudes towards abstinence? And again, I'll designate the first variable as X, spirituality, and the second variable as Y, attitudes towards abstinence. And both of these questions can be addressed using a Pearson correlation.

Demonstration

The basic formula for a Pearson correlation is as follows: your correlation is equal to the degree to which X and Y, your two variables, vary together divided by the degree to which X and Y vary separately. Or another way of stating that is, your correlation is equal to the covariability of X and Y divided by the variability of X and Y separately. A more specific formula that you can use is as follows: your correlation is equal to SP divided by your sum of squares for X, your sum of squares for Y, and you take the square root of that. And SP is known as the sum of products of the deviations, where you take the sum of X and Y together minus the sum of X multiplied by sum of Y divided by N, your sample size. And your degrees of freedom for a Pearson correlation is going to be equal to N minus 2, your sample size minus 2. A more detailed formula is as follows: the computational formula, again, SP is divided by the square root of your sum of squares for your first group--for X-and your sum of squares for your next variable, which is Y.

Correlation simply describes a relationship between two variables. It does not explain why the two variables are related. A correlation cannot be interpreted as proof of a cause-and-effect relationship between the two variables. So just because you find a significant relationship between two variables, you cannot state that one causes the other. You don't have that information.

Effect Size

The effect size for a Pearson correlation is known as the coefficient of determination, and to calculate that, all you have to do is square the value of your correlation. This is the amount of variability and one variable that can be determined from the relationship with the other variable.

Demonstration

Now let me give you an example using the formula for the Pearson correlation. My research question is as follows: Is there a relationship between the amount someone drinks and the amount someone smokes? My X variable is alcohol use, and my Y variable is number of cigarettes smoked. My null hypothesis is that there is no population correlation, and my alternative or research hypothesis is that there is a real correlation. So let's calculate this Pearson R. So I have my X values. Again, that's alcohol. 1, 2, 4, and 5. I have my Y values, and that's number of cigarettes, and that is 3, 6, 4, and 7. Now let's find the cross products. All I'm doing is multiplying the value in X to the value in Y. That's 3, 12, 16, and 35. I find that the sum of X values is equal to 12. The sum of the Y values is equal to 20, and the sum of the cross product, XY, is equal to 66. And my sum of squares for X is equal to 10, and my sum of squares for Y is also equal to 10. Now, I have 4 participants, so my degrees of freedom is N minus 2, or 4 minus 2, which equals 2. So for this example, with 2 degrees of freedom, I'm going to choose an alpha level of 0.05 and two-tailed. To find the critical values that I need to surpass in order to achieve significance, I can go to my Pearson correlation table, which is located in the appendices of your statistics textbook, and look up the critical values that I need to surpass for degrees of freedom 2 alpha level of 0.05 and two-tailed, and I find that the critical value that I need to surpass is equal to plus or minus 0.95. So now let's plug

in this information. Again, your correlation is equal to the sum of the products divided by the square root of the sum of squares X and your sum of squares Y. And to calculate your sum of the products is equal to the sum of XY minus the sum of X times the sum of Y divided by N. So here, SP is equal to 66 minus 12... times 20 divided by 4. So my SP is equal to 6. So my correlation is equal to 6 divided by the square root of 10 times 10, or my correlation is equal to... positive 0.6. So how do I interpret this? I would put R, and in parentheses, my degrees of freedom-- in this case, is equal to 2-- equals to the value of R, which is 0.60, and comma NS. It is nonsignificant, because my correlation of 0.60 did not surpass the critical value of 0.95. So here is a nonsignificant relationship between my two variables, and I put a comma two-tailed. And I will interpret this as, there is not... a significant... relationship between alcohol... and smoking. I can also calculate the effect size for this statistic, and the effect size is R squared, which is equal to 0.60 squared, which is equal to 0.36. So my R squared for this analysis is 0.36.

Now let's go over an example using SPSS. Open up your SPSS program. Now find the data set that you want to use to analyze your data. Click on File; click on Open; click on Data. Now search for the data set that you're going to use. Once you have found the data set, click on the data set, then click on Open. And make sure your data set window appears on your screen. Now, to conduct a Pearson correlation, you have to click on Analyze, Correlate, and Bivariate. And now the bivariate correlation dialog box will appear on your screen. Now you must choose the two variables that you want to conduct a Pearson correlation for. I want to look at the relationship between depression and stress, so I'm going to scroll down in my box on the left, and I'm going to find my variable depression, click on it, and click on the right arrow key to move that to the dialog box on the right. And then I'm going to click on the variable stress and then click on the right arrow key to move that to the dialog box on the right. If you look down here where it says "Correlation Coefficients," you have three choices. The one for Pearson should be checked, and it already is. And then for test of significance, you have two-tailed or one-tailed, and I'm going to choose a two-tailed test of significance. And then you want to make sure this box that says "Flag Significant Correlations" is checked, and it is. Now you have to just click Okay. And you'll see that SPSS is going to give you a correlation table that contains the correlation, the Pearson correlation, between depression and stress. Now, as you see looking at this correlation table, you'll notice a couple of things. First is the values in the diagonal are 1, because that's the correlation between each variable and itself. The values in the off diagonals is the correlation, and one thing you'll see about this correlation table is, it is symmetric. The values that appear below the diagonal are exactly the same as the values that appear above the diagonal. They're duplicates. So you only have to look at either above or below the diagonal to get your information. So as you see here, the relationship between stress and depression, the value for the Pearson correlation is 0.644. It gives you your significance two-tailed of 0.001, and SPSS also indicates with two asterisks that the correlation is significant at the 0.01 level. And your N or your sample size for this analysis is 144. So how do I interpret this? I would use R. My degrees of freedom is 142, because it is N minus 2 equals 0.64, the value of my correlation, comma P less than 0.01, or you can use the exact significance: P equals 0.000 and then comma two-tailed. So what I would be saying is that... there is a significant... positive... relationship... between depression and stress. I can also calculate the effect size for this analysis, and that is just R squared. So in this case, it's 0.64 squared, which is equal to 0.41. Using correlation, I can't say anything about the causality, so I can't say that depression causes stress or that stress causes depression. I can only say that there is a significant relationship between these two variables. Now let's learn a little bit about regression.

Regression

Regression is a statistical technique for finding the best-fitting straight line, the regression line, for a set of data. The goal is to derive an equation that is used to predict the value of one variable when given the value of the other. Definition of Multiple Regression. Multiple regression is a statistical technique that allows one to assess the relationship between one dependent variable, or criterion, and several independent variables, or predictors. The dependent variable must be continuous and the independent variables should also be continuous. Though occasionally you can see discrete or dichotomous variables that are dummy coded. Multiple regression is an extension of bivariate regression, in which several independent variables, instead of just one independent variable, are combined to predict a value on a dependent variable for each participant in your dataset. Multiple regression is a very flexible statistic. It can be used with experimental, observational, as well as survey research.

Demonstration

Two sample questions that can be addressed using regression are as follows: Question one, can we predict the number of classes missed if we know how much students drink? And in this case, my variable is alcohol use, and we want to use that to see if we can predict the number of classes missed.

Question two, can we predict attitudes towards abstinence if we know someone's level of spirituality? So we want to see if we can predict someone's attitudes towards abstinence using the variable spirituality.

The equation you would use to conduct a regression is as follows: Y equals B times X plus A, where Y is your predicted score, B is your slope of the line, and that can be found by taking your sum of the products and dividing it by the sum of squares of X. And X is the known variable of the variable X, so the variable that you're using to predict this unknown quantity of the variable Y, and A is your Y intercept, the place where the line crosses-- your regression line crosses the Y axis when X is 0. Here's a picture of a regression line. Your Y values are placed on the vertical axis, and your X values are placed on the horizontal axis. And you see the Yintercept is where Y intersects a line where X is 0, and you can see the slope of the line. The regression line shows the relationship between these two variables.

Questions Answered by Multiple Regression

What are some questions that can be answered by multiple regression? First, multiple regression can answer the question, how strong is the relationship between the dependent variable and the independent variables? Multiple regression will tell you how strong that association is between your set of predictors or your independent variables, and your criterion, your dependent variable. It will tell you how strong the relationship is between that combination of predictors and your criterion. Second, multiple regression can answer the question, what is the importance of each of the independent variables to the dependent variable? Multiple regression will tell you which predictor or independent variable has the strongest relationship with your criterion, or dependent variable. You'll be able to tell which predictor has the most overlap with your criterion.

Demonstration

A scenario that can be addressed using multiple regression. Here is the scenario. Dr Agada is interested in studying what attitudinal factors predict job performance. She gives a survey to 500 workers at a local cable company. The survey contains questions that measure three different attitudinal variables: attitude towards the company, attitude towards coworkers, and attitude towards management. All of these are measured on a one to ten scale, with ten being very positive attitudes scale of measurement. Dr Agada also gathers information on each worker's job performance, which is rated on a scale of one to five, with five being above average performance. Dr Agada is interested in seeing how this combination of three attitudinal variables relate to job performance. She hypothesizes that more positive attitudes will be related to higher job performance. In this scenario, the independent variables or predictors are the three attitudinal variables -- attitude towards the company, attitude towards coworkers, and attitude towards management, and all three of these variables are continuous. The dependent variable, or criterion, is job performance, and this variable is also continuous. So how would I do my hypothesis for this scenario? The null hypothesis is that there is no relationship between the set of three attitudinal variables in job performance. The alternative, or research hypothesis, is that there is a positive relationship between the three attitudinal variables and job performance. This research scenario can easily be answered using multiple regression.

Assumptions in Multiple Regression

There are many assumptions that first must be addressed before you can perform your multiple regression:

- ✓ The first assumption is outliers. Outliers can greatly impact your multiple regression equation. They can affect the precision of the estimation of the regression weights. You should always deal with outliers in your variables, both the independent variables and dependent variable, prior to conducting your multiple regression. Refer to the data cleaning assumptions for more information on outliers.
- ✓ The second assumption is ratio of cases to predictors. Multiple regression can be sensitive to sample size. If your sample is too small, you will not get an accurate prediction equation. To be able to accurately test for your multiple correlation, and your individual regression coefficients, you need to have a sample size of at least N greater than or equal to 104 plus M, where M is the number of predictors in your multiple regression. For example, if you have five predictors or independent variables that you want to use in your multiple regression equation, you need at least 109 participants -- 104 plus five in your sample, in order for you to have an acceptable ratio of cases to predictors.
- ✓ The third assumption is multicollinearity. Multiple regression is sensitive to multicollinearity, which is when you have at least two of your predictors, your IVs, in your equation, that are too highly correlated with each other. Multicollinearity can make your regression equation unreliable and it can give you large standard errors in your equation. Refer to the data cleaning module for more information on multicollinearity.
- ✓ Assumption four is normality of variables. There is no assumption that your variables need to be normally distributed, so having no skewness or kurtosis in order to conduct a multiple regression. However, your prediction equation is enhanced if all of your variables are normally distributed. Refer to the data cleaning assumptions for more information on normality of variables.

Types of Multiple Regression

There are three different types of multiple regression:

- ✓ The first type is standard, or also known as simultaneous or direct multiple regression. In this type of multiple regression, all of your predictors or independent variables are entered into the equation at the same time. Each predictor is assigned only its unique variance that it contributes to the equation. Variance refers to the amount of overlap the predictor has with that criterion. No predictor is assigned the overlapping variance, and the overlapping variance refers to the overlap that is shared among the predictors. The overlapping variance still is part of the adjusted R-squared, it just isn't assigned to an individual predictor. Standard multiple regression is the most widely used type of multiple regression. It needs the fewest number of participants compared to the other types of multiple regression, though it does still need at least 104 plus M.
- ✓ The second type of multiple regression is sequential, or hierarchical, multiple regression. In this type of multiple regression, the researcher enters the predictors into the equation in an order specified by the researcher. The order is dependent upon prior theory, and/or research. Researchers can enter each predictor individually, at each step, or they can enter sets of predictors at each step in the regression equation. Overlapping variance is assigned to the predictors in the order of entry into the regression equation. This type of multiple regression is used when you want to include covariance into your regression equation, and/or when you want to add interactions into your regression equation. The N solution in a sequential multiple regression will be exactly the same as your solution if you conducted a standard multiple regression.
- ✓ The third type of multiple regression is statistical or stepwise multiple regression. In this type of multiple regression, the order of entry for the independent variables is dependent on statistical criteria. The software package in this case, SPSS (if you choose to use it), will decide which predictor is put into the equation at each step based on statistical criteria that the researcher decides on. Each predictor is given their unique variance and the overlapping variance at its point of entry into the regression equation. You can use the forward method where the independent variable that has the largest bivariate correlation with the dependent variable is entered into the regression equation first, or you can use the backward method where the independent variable with the smallest bivariate correlation with the dependent variable is deleted from the equation first. Based on whether or not all of your predictors end up in the final N solution of your multiple regression, your N solution using a statistical multiple regression may be different from the N solution if you would have used a standard multiple regression.

Logistic Regression

Logistic regression is a statistical technique that allows a researcher to predict a categorical outcome, the criterion or dependent variable from a set of independent variables also known as predictors that can be continuous or categorical. Logistic regression is similar to multiple regression in that it allows multiple independent variables to predict or relate to one dependent variable. The difference with logistic regression, the dependent variable is categorical.

Logistic regression is a very flexible technique. It has few assumptions that must be addressed in order to conduct it. Logistic regression can be binomial having only two levels or groups in the dependent variable, or multinomial, more than two levels or groups in the dependent variable.

Questions Answered by Logistic Regression

A logistic regression can answer many questions. Logistic regression can answer the question, "How strong is the relationship between the dependent variable and the independent variables." Logistic regression will tell you how strong the association is between your set of predictors, your IV's and your criterion, your DV. It will tell u how strong the relationship is between the combination of predictors and your criterion.

Logistic regression can answer the question what is the importance of each of the independent variables to classifying the outcome or dependent variable. Logistic regression will tell you which predictor, which independent variable has the strongest relationship with your criterion. You'll be able to tell which predictor does the best job at correctly classifying your outcome.

Let's use the binomial logistic regression for this demonstration:

Sample Research Scenario

Professor Okibe wants to see which variables are the best predictors of successfully earning a PhD within four years, which is coded as yes or no. The variables he is interested in using as predictors or independent variables are the following: Significant other support, Peer support, Financial support from the university and Mentor support. All of these variables are measured on a 1 to 5 scale, they are all continuous. His goal is to see if these variable accurately predict whether or not someone will successfully earn their PhD within four years and which variable has the strongest relationship with the outcome.

In this scenario, there are four independent variables: Significant other support, Peer support, Financial support from the university and Mentor support. All of these variables are continuous. Higher values of these variables indicate higher levels of the variable. For example, a value of 5 means the highest level of support. The dependent variable is successfully earning a PhD within four years, which is coded as yes/no. This variable is categorical. More specifically, it is dichotomous, meaning it has only two levels.

For this scenario, the null hypotheses would be that, this set of predictors or independent variables do not significantly predict the outcome of successfully earning a PhD within four years. The alternative or research hypothesis would be that this set of predictors or independent variables do significantly predict the outcome of successfully earning a PhD within four years.

Interpreting the Logistic Regression

I want to go over some common terms used to interpret logistic regression.

- The first term is "omnibus chi-square test." This is the overall test of your logistic regression. It tells you if your set of predictors, the independent variables predicts your outcome variable, your dependent variable. If the omnibus chi-square test is significant, this signifies you have a good model fit, that your set of independent variables is related to your dependent variable.
- The next term is the Wald chi-square test. This is the significance test for the individual independent variables in your logistic regression. A significant Wald test tells u that your independent variable is a significant predictor of the dependent variable.
- The next term is "Odds ratio" or E superscript B. The odds ratio signifies the increase or decrease if the odds radio is less than 1 in odds of being in one outcome category when the value of the predictor increases by one unit. These show the researcher how strong the relationship is between each individual independent variable and the dependent or outcome variable. If there is no relationship between the independent variable and the dependent variable, your odds ratio will be one. If there is no relationship between the independent variable and the dependent variable, your odds ratio will be one. If there is a negative relationship between the independent variable and the dependent variable, your odds ratio will be less than one. If there is a positive relationship between the independent variable and the dependent variable, your odds ratio will be greater than one. Odds ratios are also known as relative risk.
- Lastly, the Cox & Snell R-square, this is the overall effect size for your logistic regression. It tells you the amount of overlap between your set of predictors and your dependent variable.

Assumptions in Logistic Regression

There are relatively few assumptions that must be met in order to conduct a logistic regression. It is relatively free of restrictions:

- The first assumption is outliers. Just like with multiple regression, outliers can greatly impact your logistic regression equation. You should always deal with outliers prior to conducting a logistic regression. Refer to the data cleaning assumptions for more information on outliers.
- Assumption number two is ratio of cases to independent variables. Logistic regression is sensitive to small sample sizes. If you have too few cases or participants, then you run the possibility of producing inaccurate parameter estimates and high standard errors for your logistic regression equation. Many researchers recommend having at least 10 participants per independent variable for your logistic regression.
- Assumption number three is multicollinearity. In order to conduct a logistic regression, you must make sure you do not have multicollinearity among your independent variables. In other words, you don't want to have too high of correlations, for example, greater than the absolute value of .8, between any of the independent variables. If you do have multicollinearity, you may have larger standard errors and may not even be able to find a solution for your logistic regression. You should delete or combine any variable pairs that are too highly correlated before performing your logistic regression. Refer to the data cleaning assumptions for more information on multicollinearity.

Types of Logistic Regression

Just like with multiple regression, there are three different types of logistic regression.

- Standard or direct logistic regression is a type of logistic regression where all of your predictors, your independent variables, are entered into the regression equation at the same time.
- Sequential or hierarchical logistic regression is a type of logistic regression where the researcher enters the predictors into the regression equation in an order specified by the researcher. The order is dependent upon prior theory and/or research. Researchers can enter each predictor individually in each step or enter sets of predictors at each step in the regression equation.

This type of logistic regression is used when you want to include covariates in your regression equation and/or when you want to add interactions into your regression equation. The end solution in a sequential logistic regression will be the exact same solution if you conducted a standard logistic regression.

- Lastly, statistical or step-wise logistic regression is a type of logistic regression in which the order of entry for the independent variables is dependent on statistical criteria. The software package - in this case, SPSS – will decide which predictors put in the equation at each step based on statistical criteria that the researcher decides on. Based on whether or not all of your predictors end up in the final end solution of your logistic regression, your end solution- end solution using a statistical logistic regression may be different from the end solution if you would have used a standard logistic regression.

Statistical Power

The decision a jury has to reach (guilty vs. not guilt) is similar to the decision a researcher makes when assessing a relationship. There are four interrelated components that influence the conclusions you might reach from a statistical test in a research project. The logic of statistical inference with respect to these components is often difficult to understand and explain. This paper attempts to clarify the four components and describe their interrelationships.

The four components are:

- **sample size**, or the number of units (e.g., people) accessible to the study
- **effect size**, or the salience of the treatment relative to the noise in measurement
- **alpha level**, or significance level), or the odds that the observed result is due to chance
- **power**, or the odds that you will observe a treatment effect when it occurs

Given values for any three of these components, it is possible to compute the value of the fourth. For instance, you might want to determine what a reasonable sample size would be for a study. If you could make reasonable estimates of the effect size, alpha level and power, it would be simple to compute (or, more likely, look up in a table) the sample size. Some of these components will be more manipulable than others depending on the circumstances of the project.

For example, if the project is an evaluation of an educational program or counseling program with a specific number of available consumers, the sample size is set or predetermined. Or, if the drug dosage in a program has to be small due to its potential negative side effects, the effect size may consequently be small. The goal is to achieve a balance of the four components that allows the maximum level of power to detect an effect if one exists, given programmatic, logistical or financial constraints on the other components.

Power at its most basic form is the likelihood of finding statistically significant differences when statistically significant differences actually do exist. In other words, power is the likelihood of recognizing when a true difference exists. As you may recall from other research courses, Type II error is the potential for failing to reject the null hypothesis when a true difference exists. Therefore, power is directly related to a Type II error. The more power that we have in any research study, the less chance there is that we would fail to reject the null hypothesis when the null hypothesis should in fact be rejected. As you might already be guessing, when the likelihood of making a Type I error increases, the likelihood of making a Type II error is decreased. Therefore, there is an inverse relationship between Type I and Type II error, so while large sample sizes increase the likelihood of finding statistically significant differences, small samples sizes increase the likelihood of a Type II error. As you can see in this slide, statistical power is expressed as one minus beta. Therefore, Type II error is expressed as beta.

While the definition of power is fairly straightforward, the power of a study is dependent on several factors. A, sample size, B, effect size, and C, alpha level.

The larger the sample size we have, the larger the statistical power. It's important to recognize that a research study may result in a statistically significant finding -- not because the results are meaningful, but simply because the sample size is so large that the statistical test picks up on very minor deviations or differences among the groups. Conducting a power analysis prior to starting a research project helps the researcher to determine an adequate sample size that may result in appropriate statistical significant without excess or increased chance for a Type I or a Type II error.

Power and Alpha Level

The alpha level also plays a significant role in understanding power. As you may recall, the alpha level is the chance of error that researchers are willing to take in determining statistical significance. By setting the alpha level at .05, researchers indicate that they are willing to accept a five percent chance of error in their statistical analysis. If researchers set the level of significance at .10, the likelihood of finding a statistically significant difference increases. For example, in an F test for an ANOVA, setting the level of significance at .10 would increase the chance that the F observed would be larger than the F critical value.

Power and Research

Power is usually considered to be adequate at .80. This means that the researcher is accepting an 80 percent chance of finding a statistically significant difference when it actually does exist. However, it's important to note that the researcher is also accepting a 20 percent chance of a Type II error.

Hopefully, you can see that power is a critical element in any research study. However, power is an underreported aspect in most research. While the basic principles of power analysis are fairly straightforward, the calculation of power can be somewhat complex. Unfortunately, statistical packages that are primarily used for research such as SPSS or PASW do not compute power; therefore, computer programs have been developed that will help you to calculate power. G*Power is a free software program that is widely used to conduct power analysis. You'll do well to download and acquaint self with it.

Researchers rely heavily on statistical analysis software such as SPSS or PASW. Each of these statistical software packages reports limited types of effect size measures. Thus, many researchers who do in fact include effect size measures in their manuscripts for professional journals only report Eta-squared or Cohen's D, because this are what are provided by the statistical software packages.

Eta Squared

Eta-squared refers to the strength of association between the independent variables and the dependent variable. Eta-squared provides us with the amount of variance that is accounted for in the dependent variable by the independent variable. If the strength of association is weak than we know that the independent variable has less meaning or relevance to the dependent variable. Conversely, if the strength of association is strong, it is clear that the independent variable has a high degree of relevance to the dependent variable.

So I'm guessing you're now asking the question, what do I do if I want to report effect size? Well, if you're looking at association and you're interested in using statistical analysis software, then simply consider reporting partial Eta-squared. You should recognize, however, that Eta-squared and partial Eta-squared are upwardly biased estimates, especially when sample sizes are small for a study. Remember that you do have the option of reporting other effect size estimates and that there are more thorough calculations of effect size available to you.

Effect Size for Correlation

If you're interested in effect size for correlation than the coefficient of determination, or R-squared, is the appropriate calculation. R-squared is the proportion of variance that is explained when examining the association between variables. This calculation ranges from zero to one. It's important for your to note that R-squared is a positive number and it does not indicate the direction of the relationship. To put this in practical terms, if R-squared has a value of .21, this means that 21 percent of the variance of either variable is shared with the other variable.

Effect Size for Regression

If you're interested in tackling effect size for a regression analysis, then Cohen's F-squared is the appropriate effect size measure. When examining this particular table, you can see that the calculation for F-squared is R-squared divided by one minus R-squared. R-squared is simply the squared multiple correlation. With multiple regression, we are attempting to determine how well future outcomes will likely be predicted by the regression model. When interpreting effect sizes for Cohen's F-squared, you should consider .02 a small effect size, .15 a medium effect size, and .35 a large effect siz

Effect Size for Chi Square

Another measure of effect size that you may need to consider is the one for Chi square analysis. The phi coefficient is the standard effect size calculation for a Chi square. Phi is related to the point-biserial correlation and it estimates the extent of the relationship between two variables. Phi is computed by finding the square root of the Chi square statistic and dividing by the sample size.

Effect Size for T-test and ANOVA

When calculating effect size for T-tests and ANOVA, some of the most often used calculations are Omega Squared, Cohen's D, and Cohen's F.

II. Conclusion

The choice of scientific research method and design is never done hazardly but there are undergirding factors to consider. Every researcher brings with him personal ontologies, world views or philosophical assumptions which enable him to study natures. No researcher can pretend or at best may be unconscious of the powers of his world view in shaping his perceptions about phenomenon. A poorly structured paper mixes up the

choice of methodologies, the types of questions asked without a recourse to the historical alignments of the research components. This paper clearly shows the path to these assumptions and the relationship between the philosophy of the researcher, the types of research questions advances, the assumed relationship between the variables, the research instruments to be deployed, the design which aligns with the world views and the eventual conclusion and generalization of findings.

The study clearly elucidates emerging strategies in designs such as quantitative, qualitative and mixed method designs which are products of the world views of the researcher. Each of these designs were properly discussed along with the various approaches, sampling strategies, the concept of validity, credibility and reliability of instruments and how these qualities can be statistically ensured.

Finally, the study presents the rubrics for the choice of appropriate statistics to test for our main effects, data management and organization. Social science researchers use robust statistics to ameliorate for the internal weakness in design owing to the absence of true experiment. Every statistic come with general and specific assumptions which must be met before deployment. A poor choice of statistics leads to spurious, unreliable and misleading outcome. It is hope that this study will aid a lot of researchers approaching their work with absolute confidence and dexterity.

Reference

[1]. Aczel, A. D; & Sounderpandian, J. (2002). Complete Business Statistics, (5th Ed.). New Delhi, India: Tata Mcgraw-Hill Publishing Company Limited, India.

[2]. Atkinson, A. C., Riani, M., & Riani, M. (2000). Robust Diagnostic Regression Analysis (Vol. 2). New York: Springer.

[3]. Atkinson, A. C., Riani, M., & Cerioli, A. (2004). Exploring Multivariate Data With The Forward Search (Vol. 1). New York: Springer.

[4]. Bell, A. B., Distefano, C., & Morgan, G. B. (2010). A Premier On Disseminating Applied Quantitative Research. Journal Of Early Intervention,32 (5), 370-383. Doi. 10.1177/1053815110389462

[5]. Bleijenbergh, I., Korzilius, H., & Verschuren, P. (2011). Methodological Criteria For Internal Validity And Utility Of Practice-Oriented Research. Journal Of Qual-Quant, 2011 (45), 145- 156. Doi.10.1007/S11135 – 010-9361-5.

[6]. Bruin, J. 2006.Newtest: Command To Compute New Test. UCLA: Statistical Consulting Group. Https://Stats.Oarc.Ucla.Edu/Stata/Ado/Analysis/.

[7]. Crandell, J. L., Voils, C. I., Chang, Y., & Sandelowski, M. (2011). Bayesian Data Augmentation Methods For The Synthesis Of Qualitative And Quantitative Research Findings. Journal Of Qual-Quant, 2011 (45), 653 – 669. Doi:10.1007/S11135-010-9375-Z

[8]. Frankfort-Nachmias, C. & Nachmias, D. (2008). Research Methods In The Social Sciences, (7th Ed.). NY: Worth Publishers

[9]. Hampel F.R., Ronchetti E.M., Rousseeuw P.J., And Stahel W.A. (1986). Robust Statistics: The Approach Based On Influence Functions, 502 Pp. New York: Wiley

[10]. Hampel, F. R. (1968). Contributions To The Theory Of Robust Estimation. University Of California, Berkeley.

[11]. Hampel F.R. (1974). The Influence Curve And Its Role In Robust Estimation. Journal Of The American Statistical Association 69, 383–393. Retrieved From Https://Www.Baeldung.Com/Cs/Robust-Estimators-In-Robust-Statistics

[12]. Hand-Out On "Comparing Quantitative And Qualitative Research" Retrieved Http://Www.Experiment-Resources.Com/Quantitative-And-Qualitative-Research.Html

[13]. Horner, R. H., Swaminathan, H., Sugai, G. & Smolkowski, K. (2012). Considerations For The Systematic Analysis And Use Of Single-Case Research. Journal Of Education And Treatment Of Children, 35(2), 269-290. Doi:10.1353/Etc.2012.0011.

[14]. Huber P.J. (1964). Robust Estimation Of A Location Parameter. The Annals Of Mathematical Statistics 35, 73–101. [Introduces M-Estimators. Started Much Work On Robustness.] Huber P.J. (1981). Robust Statistics, 308 Pp. New York: Wiley.

[15]. Leeper, J. D. (2010). Introduction To SAS. UCLA: Statistical Consulting Group. From Https://Stats.Oarc.Ucla.Edu/Sas/Modules/Introduction-To-The-Features-Of-Sas/ (Accessed August 22, 2021).

[16]. Lecué, G., & Lerasle, M. (2020). Robust Machine Learning By Median-Of-Means: Theory And Practice. Retrieved From Https://Www.Sciencedirect.Com/Topics/Engineering/Robust-Estimator#:~:Text=Robust%20estimators%20are%20preferred%20when,Obtain%20a%20narrow%20tolerance%20band.

[17]. Liu, R. Y. (1990). On A Notion Of Data Depth Based On Random Simplices. The Annals Of Statistics, 405-414. Retrieved From Https://Mathworld.Wolfram.Com/Robustestimation.Html#:~:Text=An%20estimation%20technique%20which%20is,Used%20to%20optimize%20the%20algorithm.

[18]. Morrow, J. A. (2011): RSCH 8250: Walden University Lecture On Data Cleaning And Dealing With Assumptions. Retrieved From Https://Class.Waldenu.Edu/Webapps/Portal/Frameset.Jsp?Tab_Tab_Group_Id=_2_1&URl=%2Fwebapps%2Fblackboard%2Fexecute%2Flauncher%3Ftype%3dcourse%26id%3D_1983490_1%26url%3D

[19]. Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., & Murphy, S. A. (2012). Experimental Design And Primary Data Analysis Methods For Comparing Adaptive Interventions. Journal Of Psychological Methods Advance Online Publication Dol:10.1037/A0029372.

[20]. Prasad, A., Suggala, A. S., Balakrishnan, S., & Ravikumar, P. (2020). Robust Estimation Via Robust Gradient Estimation. Journal Of The Royal Statistical Society Series B: Statistical Methodology, 82(3), 601-627.

[21]. Rousseeuw, P. J., & Leroy, A. M. (2005). Robust Regression And Outlier Detection. John Wiley & Sons. Retrieved From Https://Projecteuclid.Org/Journals/Annals-Of-Statistics/Volume-50/Issue- Retrieved From Https://Www.Forbes.Com/Sites/Ellevate/2021/06/08/The-Importance-Of-Using-Robust-Analysis-To-Understand-Change/?Sh=1c2d6b943dab

[22]. Rousseeuw, P. J., & Hubert, M. (1999). Regression Depth. Journal Of The American Statistical Association, 94(446), 388-402. Retrieved From Https://Www.Adelaide.Edu.Au/Aiml/Our-Research/Machine-Learning/Robust-Statistics.

[23]. Tukey J.W. (1960). A Survey Of Sampling From Contaminated Distributions. Contributions To Probability And Statistics: Essays In Honor Of Harald Hotelling (Ed. I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, And H.B. Mann), Pp. 448–485. Stanford: Stanford University Press.

[24]. Tukey, J. W. (1975). Mathematics And The Picturing Of Data. In Proceedings Of The International Congress Of Mathematicians, Vancouver, 1975 (Vol. 2, Pp. 523-531). Retrieved From Https://Statisticsbyjim.Com/Basics/Robust-Statistics/