Education

Introduction to Research Statistical Analysis: An Overview of the Basics

Christian Vandever[1]

Author affiliations are listed at the end of this article.

Correspondence to:
Christian Vandever
HCA Healthcare Graduate Medical Education
2000 Health Park Drive
Brentwood TN, 37027
(Christian.Vandever@hca-healthcare.com)

Abstract

Description

This article covers many statistical ideas essential to research statistical analysis. Sample size is explained through the concepts of statistical significance level and power. Variable types and definitions are included to clarify necessities for how the analysis will be interpreted. Categorical and quantitative variable types are defined, as well as response and predictor variables. Statistical tests described include t-tests, ANOVA and chi-square tests. Multiple regression is also explored for both logistic and linear regression. Finally, the most common statistics produced by these methods are explored.

Keywords

statistical analysis; sample size; power; t-test; anova; chi-square; regression

Introduction

Statistical analysis is necessary for any research project seeking to make quantitative conclusions. The following is a primer for research-based statistical analysis. It is intended to be a high-level overview of appropriate statistical testing, while not diving too deep into any specific methodology. Some of the information is more applicable to retrospective projects, where analysis is performed on data that has already been collected, but most of it will be suitable to any type of research. This primer will help the reader understand research results in coordination with a statistician, not to perform the actual analysis. Analysis is commonly performed using statistical programming software such as R, SAS or SPSS. These allow for analysis to be replicated while minimizing the risk for an error. Resources are listed later for those working on analysis without a statistician.

Discussion

After coming up with a hypothesis for a study, including any variables to be used, one of the first steps is to think about the patient population to apply the question. Results are only relevant to the population that the underlying data represents. Since it is impractical to include everyone with a certain condition, a subset of the population of interest should be taken. This subset should be large enough to have power, which means there is enough data to deliver significant results and accurately reflect the study's population.

The first statistics of interest are related to significance level and power, alpha and beta. Alpha (α) is the significance level and probability of a type I error, the rejection of the null hypothesis when it is true. The null hypothesis is generally that there is no difference between the groups compared. A type I error is also known as a false positive. An example would be an analysis that finds one medication statistically better than another, when in reality there is no difference in efficacy between the two. Beta (β) is the probability of a type II error, the failure to reject the null hypothesis when it is actually false. A type II error is also known as a false negative. This occurs when the analysis finds there is no difference in two medications when in reality one works better than the other. Power is defined as 1-β and should be calculated prior to running any sort of statis-

© 2020 HCA Physician Services, Inc. d/b/a Emerald Medical Education

tical testing. Ideally, alpha should be as small as possible while power should be as large as possible. Power generally increases with a larger sample size, but so does cost and the effect of any bias in the study design. Additionally, as the sample size gets bigger, the chance for a statistically significant result goes up even though these results can be small differences that do not matter practically. Power calculators include the magnitude of the effect in order to combat the potential for exaggeration and only give significant results that have an actual impact. The calculators take inputs like the mean, effect size and desired power, and output the required minimum sample size for analysis. Effect size is calculated using statistical information on the variables of interest. If that information is not available, most tests have commonly used values for small, medium or large effect sizes.

When the desired patient population is decided, the next step is to define the variables previously chosen to be included. Variables come in different types that determine which statistical methods are appropriate and useful. One way variables can be split is into categorical and quantitative variables. (**Table 1**) Categorical variables place patients into groups, such as gender, race and smoking status. Quantitative variables measure or count some quantity of interest. Common quantitative variables in research include age and weight. An important note is that there can often be a choice for whether to treat a variable as quantitative or categorical. For example, in a study looking at body mass index (BMI), BMI could be defined as a quantitative variable or as a categorical variable, with each patient's BMI listed as a category (underweight, normal, overweight, and obese) rather than the discrete value. The decision whether a variable is quantitative or categorical will affect what conclusions can be made when interpreting results from statistical tests. Keep in mind that since quantitative variables are treated on a continuous scale it would be inappropriate to transform a variable like which medication was given into a quantitative variable with values 1, 2 and 3.

Both of these types of variables can also be split into response and predictor variables. (**Table 2**) Predictor variables are explanatory, or independent, variables that help explain changes in a response variable. Conversely, response variables are outcome, or dependent, variables whose changes can be partially explained by the predictor variables.

Choosing the correct statistical test depends on the types of variables defined and the question being answered. The appropriate test is determined by the variables being compared. Some common statistical tests include t-tests, ANOVA and chi-square tests.

T-tests compare whether there are differences in a quantitative variable between two values of a categorical variable. For example, a t-test could be useful to compare the length of stay for knee replacement surgery patients between those that took apixaban and those that took rivaroxaban. A t-test could examine whether there is a statistically significant difference in the length of stay between the two groups. The t-test will output a p-value, a number between zero and one, which represents the probability that the two groups could be as different as they are in the data, if they were actually the same. A value closer to zero suggests that the difference, in this case for length of stay, is more statistically significant than a number closer to one. Prior to collecting the data, set a significance level, the previously defined alpha. Alpha is typically set at 0.05, but is commonly reduced in order to limit the chance

Table 1. Categorical vs. Quantitative Variables

Categorical Variables	Quantitative Variables
Categorize patients into discrete groups	Continuous values that measure a variable
Patient categories are mutually exclusive	For time based studies, there would be a new variable for each measurement at each time
Examples: race, smoking status, demographic group	Examples: age, weight, heart rate, white blood cell count

Table 2. Response vs. Predictor Variables

Response Variables	Predictor Variables
Outcome variables	Explanatory variables
Should be the result of the predictor variables	Should help explain changes in the response variables
One variable per statistical test	Can be multiple variables that may have an impact on the response variable
Can be categorical or quantitative	Can be categorical or quantitative

of a type I error, or false positive. Going back to the example above, if alpha is set at 0.05 and the analysis gives a p-value of 0.039, then a statistically significant difference in length of stay is observed between apixaban and rivaroxaban patients. If the analysis gives a p-value of 0.91, then there was no statistical evidence of a difference in length of stay between the two medications. Other statistical summaries or methods examine how big of a difference that might be. These other summaries are known as post-hoc analysis since they are performed after the original test to provide additional context to the results.

Analysis of variance, or ANOVA, tests can observe mean differences in a quantitative variable between values of a categorical variable, typically with three or more values to distinguish from a t-test. ANOVA could add patients given dabigatran to the previous population and evaluate whether the length of stay was significantly different across the three medications. If the p-value is lower than the designated significance level then the hypothesis that length of stay was the same across the three medications is rejected. Summaries and post-hoc tests also could be performed to look at the differences between length of stay and which individual medications may have observed statistically significant differences in length of stay from the other medications.
A chi-square test examines the association between two categorical variables. An example would be to consider whether the rate of having a post-operative bleed is the same across patients provided with apixaban, rivaroxaban and dabigatran. A chi-square test can compute a p-value determining whether the bleeding rates were significantly different or not. Post-hoc tests could then give the bleeding rate for each medication, as well as a breakdown as to which specific medications may have a significantly different bleeding rate from each other.

A slightly more advanced way of examining a question can come through multiple regression. Regression allows more predictor variables to be analyzed and can act as a control when looking at associations between variables. Common control variables are age, sex and any comorbidities likely to affect the outcome variable that are not closely related to the other explanatory variables. Control variables can be especially important in reducing the effect of bias in a retrospective population. Since retrospective data was not built with the research question in mind, it is important to eliminate threats to the validity of the analysis. Testing that controls for confounding variables, such as regression, is often more valuable with retrospective data because it can ease these concerns. The two main types of regression are linear and logistic. Linear regression is used to predict differences in a quantitative, continuous response variable, such as length of stay. Logistic regression predicts differences in a dichotomous, categorical response variable, such as 90-day readmission. So whether the outcome variable is categorical or quantitative, regression can be appropriate. An example for each of these types could be found in two similar cases. For both examples define the predictor variables as age, gender and anticoagulant usage. In the first, use the predictor variables in a linear regression to evaluate their individual effects on length of stay, a quantitative variable. For the second, use the same predictor variables in a logistic regression to evaluate their individual effects on whether the patient had a 90-day readmission, a dichotomous categorical variable. Analysis can compute a p-value for each included predictor variable to determine whether they are significantly associated.
The statistical tests in this article generate an associated test statistic which determines

the probability the results could be acquired given that there is no association between the compared variables. These results often come with coefficients which can give the degree of the association and the degree to which one variable changes with another. Most tests, including all listed in this article, also have confidence intervals, which give a range for the correlation with a specified level of confidence. Even if these tests do not give statistically significant results, the results are still important. Not reporting statistically insignificant findings creates a bias in research. Ideas can be repeated enough times that eventually statistically significant results are reached, even though there is no true significance. In some cases with very large sample sizes, p-values will almost always be significant. In this case the effect size is critical as even the smallest, meaningless differences can be found to be statistically significant.

These variables and tests are just some things to keep in mind before, during and after the analysis process in order to make sure that the statistical reports are supporting the questions being answered. The patient population, types of variables and statistical tests are all important things to consider in the process of statistical analysis. Any results are only as useful as the process used to obtain them. This primer can be used as a reference to help ensure appropriate statistical analysis.

Glossary

Alpha (α): the significance level and probability of a type I error, the probability of a false positive
Analysis of variance/ANOVA: test observing mean differences in a quantitative variable between values of a categorical variable, typically with three or more values to distinguish from a t-test
Beta (β): the probability of a type II error, the probability of a false negative
Categorical variable: place patients into groups, such as gender, race or smoking status
Chi-square test: examines association between two categorical variables
Confidence interval: a range for the correlation with a specified level of confidence, 95% for example
Control variables: variables likely to affect the outcome variable that are not closely related to the other explanatory variables
Hypothesis: the idea being tested by statistical analysis
Linear regression: regression used to predict differences in a quantitative, continuous response variable, such as length of stay
Logistic regression: regression used to predict differences in a dichotomous, categorical response variable, such as 90-day readmission
Multiple regression: regression utilizing more than one predictor variable
Null hypothesis: the hypothesis that there are no significant differences for the variable(s) being tested
Patient population: the population the data is collected to represent
Post-hoc analysis: analysis performed after the original test to provide additional context to the results
Power: 1-beta, the probability of avoiding a type II error, avoiding a false negative
Predictor variable: explanatory, or independent, variables that help explain changes in a response variable
p-value: a value between zero and one, which represents the probability that the null hypothesis is true, usually compared against a significance level to judge statistical significance
Quantitative variable: variable measuring or counting some quantity of interest
Response variable: outcome, or dependent, variables whose changes can be partially explained by the predictor variables
Retrospective study: a study using previously existing data that was not originally collected for the purposes of the study
Sample size: the number of patients or observations used for the study
Significance level: alpha, the probability of a type I error, usually compared to a p-value to determine statistical significance
Statistical analysis: analysis of data using statistical testing to examine a research hypothesis
Statistical testing: testing used to examine the validity of a hypothesis using statistical calculations
Statistical significance: determine whether to reject the null hypothesis, whether the p-value is below the threshold of a predetermined significance level
T-test: test comparing whether there are differences in a quantitative variable between two values of a categorical variable

Conflicts of Interest

The author declares he has no conflicts of interest.

Christian Vandever is an employee of HCA Healthcare Graduate Medical Education, an organization affiliated with the journal's publisher.

This research was supported (in whole or in part) by HCA Healthcare and/or an HCA Healthcare affiliated entity. The views expressed in this publication represent those of the author(s) and do not necessarily represent the official views of HCA Healthcare or any of its affiliated entities.

Author Affiliation

1. HCA Healthcare Graduate Medical Education

Resources

1. Finding and Using Health Statistics. National Library of Medicine. Updated April 3, 2019. Accessed April 15,2020. https://www.nlm.nih.gov/nichsr/stats_tutorial/cover.html
2. Thomas E. An introduction to medical statistics for health care professionals: describing and presenting data. *Musculoskeletal Care*, 2004;2:218-228. https://doi.org/10.1002/msc.73
3. Thomas E. An introduction to medical statistics for health care professionals: Hypothesis tests and estimation. *Musculoskeletal Care*, 2005;3:102-108. https://doi.org/10.1002/msc.30
4. Thomas E. An introduction to medical statistics for health care professionals: basic statistical tests. *Musculoskeletal Care*, 2005;3:201-212. https://doi.org/10.1002/msc.11