Learning Curve

A Student’s Guide to the Classification and Operationalization of Variables in the Conceptualization and Design of a Clinical Study: Part 2

Chittaranjan Andrade

ABSTRACT

Students without prior research experience may not know how to conceptualize and design a study. This is the second of a two-part article that explains how an understanding of the classification and operationalization of variables is the key to the process. Variables need to be operationalized; that is, defined in a way that permits their accurate measurement. They may be operationalized as categorical or continuous variables. Categorical variables are expressed as category frequencies in the sample as a whole, while continuous variables are expressed as absolute numbers for each subject in the sample. Continuous variables should not be converted into categorical variables; there are many reasons for this, the most important being that precision and statistical power are lost. However, in certain circumstances, such as when variables cannot be accurately measured, when there is an administrative or public health need, or when the data are not normally distributed, it may be justifiable to do so. Confounding variables are those that increase (or decrease) the apparent effect of an independent variable on the dependent variable, thereby producing spurious (or suppressing true) relationships. These and other concepts are explained with the help of clinically relevant examples.

Keywords: Categorical variable, continuous variable, confounding variable, measurement of variables

Introduction

The first article in this two-part series explained what independent and dependent variables are, how an understanding of these is important in framing hypotheses, and what operationalization of a variable entails. This article is the second part; it discusses categorical and continuous variables and explains the importance of identifying and studying confounding variables.

Categorical and Continuous Variables

Categorical variables are also known as discrete or qualitative variables. These are variables that are operationalized as categories. The value of the variable is expressed as a category count (also known as category frequency) for the sample, or as a cell count (or cell frequency) when the data are presented in a table. Sex is an example of a categorical variable; it is operationalized into male and female categories and is expressed as the number (count, frequency) of males and the number (count, frequency) of females in the sample. Religion (categories: Hindu, Muslim, Christian, Other), place of residence (categories: rural, semi-urban, urban), occupation (categories: unskilled, semi-skilled, skilled), family history of mental illness (categories: present, absent), HIV test result (categories: positive, negative), two-year survival (categories: alive, dead), and response to a question (categories: yes, no) are other examples of categorical variables.

Continuous variables are also known as quantitative variables. These are variables that are operationalized as continuous and explains the importance of identifying and studying confounding variables.

Confounding variables are those that increase (or decrease) the apparent effect of an independent variable on the dependent variable, thereby producing spurious (or suppressing true) relationships.

These and other concepts are explained with the help of clinically relevant examples.

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-Commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
number for each unit in the sample. Age is an example of a continuous variable; it is represented by a number for each subject in the sample. Other examples of continuous variables include height, number of previous depressive episodes, total score on a depression rating scale, red blood cell count, and duration of survival after treatment for cancer. Whereas the value of a categorical variable is expressed as category frequencies in the sample, the value of a continuous variable is expressed as mean, median, mode, range, standard deviation, and/or interquartile range for the sample.

## Expressing Continuous Variables as Categorical Variables

Continuous variables can be converted into categorical variables. Thus, instead of expressing age as a value for each subject (and as mean and standard deviation for the sample), we can describe age as a category size, as shown in Table 1. There are many reasons why this is not good statistical practice:\(^2,3:\)

1. **We lose precision.** Imagine that we classify the variable “Examination result” into pass and fail categories with fail defined as a score of \(<40\) marks and pass as a score of \(\geq 40\) marks. With such a classification, we have no idea whether those who passed barely managed to pass or passed with flying colors. This is exactly what happens when, in a drug trial, we classify patients as responders and nonresponders; we have no idea whether those who responded improved partially and so continued to exhibit residual symptoms, or recovered completely. And, if a male subject in our study does not tell us his exact age but hedges, saying “I’m 20 to 29 years old,” and we are unhappy with his statement, why would we record age as presented in Table 1? Loss of precision impairs our ability to see and understand finer details in the data.

2. **We may classify subjects in a way that defies common sense.** With reference to Table 1, subjects who are 20 and 29 years old, who are nine years apart in age, are classified in the same group (20 to 29 years) whereas subjects who are 29 and 30 years old, who are just one year apart, are classified in different groups (20 to 29 years and 30 to 39 years).

3. **The boundaries of the categories are arbitrary.** There is no mathematical reason why we should prefer categories that increase in units of ten (e.g., 20 to 29, 30 to 39) as opposed, say, to categories that increase in units of eight (e.g., 20 to 27, 28 to 35).

4. **Statistical significance is harder to achieve in tests applied to categorical data.** So, if continuous data are converted into categorical data, the analyses may be contaminated by type 2 (false negative) errors.

The above notwithstanding, there are certain situations in which continuous variables may justifiably be converted into categorical variables:\(^2,3:\)

1. **There is an administrative or public health need.** As an example of an administrative need, age may be categorized into pediatric, adult, and geriatric groups for hospital services. As an example of a public health need, blood pressure and low-density lipoprotein cholesterol values may be split into different categories, using different cut-off values, for category-specific treatment guidelines. Similarly, classifying patients into responders and nonresponders satisfies a public health need; it helps people understand how many patients can be expected to improve to at least the cut-off point (e.g., 30% improvement) that was used to define treatment response.

2. **The variable cannot be accurately measured.** This happens, for example, when rural or illiterate patients are unable to state their exact age but are able to say that they are in their twenties, thirties, or forties.

3. **A variable shows a nonlinear association with the dependent variable.** As an example, in the well-known Yerkes–Dodson curve, low and high stress are both associated with poorer performance or achievement, whereas moderate stress is associated with higher performance or achievement. There may, likewise, be nonlinear associations between alcohol intake and ischemic heart disease events.

4. **The data are skewed, that is, there are some subjects (outliers) with extreme values.** In such situations, the mean is not an appropriate measure of central tendency; rather, the median is appropriate. The data, then, are either ranked and studied using non-parametric tests or categorized and further studied. As an example, data on physical exercise variables are usually skewed: most people do little exercise, some people exercise moderately, and a few people exercise vigorously. For statistical analysis, such data may be categorized into tertiles, quartiles, or quintiles.

5. **The data are presented in a histogram.** When histograms are necessary to explain the data, the only way to do so is to present continuous data in class intervals (or categories) along the X-axis, with frequency count displayed on the Y-axis.

6. **When risks need to be calculated.** In logistic regression analyses, continuous data may be converted into categories (e.g., quintiles) so that odds ratios (e.g., for highest vs. lowest quintiles) can be calculated. Such a strategy can be used, for example, to examine the influence of baseline low-density lipoprotein cholesterol level on the five-year risk of an ischemic heart disease event. As a simpler example, in a randomized, placebo-controlled trial, the relative risk of response or remission can only be determined if a cut-off is applied to continuous data (obtained using a rating scale) to classify patients into response or remission categories.

## Confounding Variables

A discussion on variables is incomplete without a section on confounding variables. Consider the following example. We study data on mortality associated
with helmet use in a thousand two-wheeler traffic accident cases (Table 2). Here, wearing a helmet is the (categorical) independent variable, and occurrence of death is the (categorical) dependent variable. It would indeed seem, from the statistically significant finding in Table 2, that wearing a helmet protects the rider from serious head injuries and death. Can we conclude that the data prove a cause and effect relationship and hence that wearing a helmet should be mandated for two-wheeler riders? To the layperson’s eye, it would seem so.

This was an observational study, not a randomized controlled trial. So, we must consider another possibility. What if personality factors are responsible for reckless riding (resulting in more serious and potentially fatal accidents) as well as for a disregard for safety measures such as helmet use? If such is the case, then recklessness, rather than not wearing a helmet, would partly or wholly explain the mortality risk. So, personality may be a confounding variable that influences the association between wearing a helmet and the risk of death in a traffic accident. Expressed otherwise, people who are careful by nature may ride carefully and be less likely to suffer an accident. People who are careful are also more likely to obey laws and wear helmets. So, carefulness, as a personality trait, is what saves lives; wearing a helmet is merely a marker for carefulness and hence a lower risk of accidents.

Similarly, we may find that overcrowding of wards is associated with a higher risk of postoperative infection. It may not be the crowding of beds in the wards that increases the risk; rather, when wards are crowded, the number of visitors proportionately increases, and the risk of germs being brought into the ward also proportionately increases. Thus, visitor density, and not bed density, may explain the relationship between overcrowding of wards and postoperative infection. Bed density is just a marker of (increased) risk of infection.

In an example cited in the first article in this series, age is the confounding variable that explains the association between the number of teeth and body weight in preschool children. In an example relevant to perinatal psychiatry, the increased risk of autism spectrum disorder (ASD) associated with antidepressant (AD) use during pregnancy may not be because of AD exposure; it may be because of genetic factors or behavioral changes associated with depression. The use of AD to treat depression during pregnancy is therefore merely a marker for the increased risk of ASD. Thus, the genetic factors and behavioral changes are confounding variables that partly or wholly explain the association between AD exposure and the risk of ASD.

Readers may note that confounding variables may also mask relationships between independent and dependent variables. For example, a study may find that stress has no significant effect on performance. However, had motivation been examined as a confounding variable, the study might have found that stress increased performance in persons with high motivation and decreased performance in persons with low motivation, resulting in a net absence of effect in the sample as a whole. A more extensive discussion on confounding variables is available elsewhere.

From this discussion, it should be clear that once the independent variable has been defined, confounding variables comprise all the other variables that can either increase or decrease the value of the dependent variable. In good research, therefore, all variables which influence the dependent variable should be measured and studied, and not just the independent variable(s) of interest. It would be disastrous to complete a study and then discover that an important confound had not been studied.

### Table 2

| Riders wearing helmet | Survived the Accident | Died in the Accident |
|-----------------------|-----------------------|----------------------|
|                       | 500                   | 50                   |
| Riders not wearing helmet | 300                 | 150                  |

Note: Chi-square = 90.91; df = 1; P < 0.001.

### Learning Curve

It is important to identify and study all important dependent and independent variables related to the study’s subject. This requires careful thought at the time of preparation of the research protocol, itself. As an example, in a study on sociodemographic and clinical predictors of AD response in patients with major depressive disorder, after data collection is complete, it is too late to remember that adherence to AD treatment should also have been studied.

Studying a large number of variables improves the understanding of the subject of the study as well as allows the examination of the influence of confounding variables. Thus, if a researcher wishes to examine the effects of diet on ischemic heart disease, it is not sufficient to collect information only about dietary habits and the occurrence of myocardial infarction in a large cohort of subjects. A far better design would be to include:

1. The following independent variables: age, sex, dietary habits, exercise patterns, smoking, alcohol intake, family history of ischemic heart disease, medical history of diabetes and hypertension, and so on.

2. The following dependent variables: occurrence(s) of angina during follow-up, occurrence(s) of myocardial infarction during follow-up, need for angioplasty or other surgical intervention, and occurrence of cardiovascular death.

It is important to study the same variable using different instruments. This is because not all instruments are equal in sensitivity, specificity, reliability, validity, and other characteristics. Furthermore, different instruments may measure different aspects of the same variable, or different concepts of the same variable when the variable is abstract (e.g., personality, depression, and psychosis). Thus, as explained in the first part of this article, when studying the influence of medication on depression, it is a good idea to use several different methods for the assessment of the disorder, and not just one method, and several methods for the assessment of the same dependent variable, and not just one.
Declaration of Conflicting Interests
The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author received no financial support for the research, authorship, and/or publication of this article.

References
1. Andrade C. A student’s guide to the classification and operationalization of variables in the conceptualization and design of a clinical study: Part 1. Indian J Psychol Med 2021; (in press).
2. Streiner DL. Breaking up is hard to do: The heartbreak of dichotomizing continuous data. Can J Psychiatry 2002; 47(3): 262–266.
3. Andrade C. Age as a variable: Continuous or categorical? Indian J Psychiatry 2017; 59(4): 524–525.
4. Andrade C. Antidepressant use in pregnancy and risk of autism spectrum disorders: A critical examination of the evidence. J Clin Psychiatry 2013; 74(9): 940–941.
5. Andrade C. Genes as unmeasured and unknown confounds in studies of neurodevelopmental outcomes after antidepressant prescription during pregnancy. J Clin Psychiatry 2020; 81(3): 20f13463.
6. Johnston R, Jones K, and Manley D. Confounding and collinearity in regression analysis: A cautionary tale and an alternative procedure, illustrated by studies of British voting behavior. Qual Quant 2018; 52(4): 1957–1976.
7. Rhodes AE, Lin E, and Streiner DL. Confronting the confounders: The meaning, detection, and treatment of confounders in research. Can J Psychiatry 1999; 44(2): 175–179.
8. Mamdani M, Sykora K, Li P, et al. Reader’s guide to critical appraisal of cohort studies: Part 2. Assessing potential for confounding. BMJ 2005; 330(7497): 960–962.
9. Andrade C. Prenatal depression and infant health: The importance of inadequately measured, unmeasured, and unknown confounds. Indian J Psychol Med 2018; 40(4): 395–397.