

Descriptive Data Analysis

This 13th article of the Basics of Research series is first in a short series on statistical analysis. These articles will discuss creating your statistical analysis plan, levels of measurement, descriptive statistics, probability theory, inferential statistics, and general considerations for interpretation of the results of a statistical analysis.

Statistical Analysis Plan

The most important part of any research project is the planning process. This statement is as true for data analysis as for any of the other steps in the research process. The development of your statistical analysis plan should not be delayed until after you have your data in hand. Rather, the investigator should select the statistics to describe the sample and to analyze the data for each research question or hypothesis before initiating the study. Most grant applications will require this information, but these decisions should be made regardless of application for funding.

Investigators should plan to first describe their sample. They should identify the important demographic characteristics of the sample, such as sex, age, and race. These variables will be the same for most studies. Other sample characteristics, such as diagnosis, weight, height, Glasgow Coma Scale, and so forth, also may be important to provide. Descriptive statistics will describe these variables.

Next, investigators should plan the analyses for each research question/hypothesis. A table may be useful for this activity. In the first column should be the research question/hypothesis; in the second, all relevant variables (and timing information if needed); and in the third, the statistical test to be used. This process helps investigators ensure that they are collecting all needed data, at the right time, to answer their question. After all study data have been collected is not the time to discover that an important piece of data has been missed.

Investigators uncomfortable with statistical analysis should consult a statistician early in the planning phase. A statistician will help them determine what statistical analyses are most appropriate for answering the research questions/hypotheses, taking into consideration the types of data to be collected. Although statisticians may seem intimidating, investigators should consider them an important member of the research team and avail themselves of their expertise.

To help diminish the stress of a statistical consultation, investigators should prepare a list of questions before the meeting. In creating the list of questions, they need to start with the research question.¹ If the investigators have an idea of what statistics to use, then the questions for the statistician are related to whether the proposed analyses are appropriate and what other statistics should be considered.

If the investigators have no idea of what statistics to use, the first question should be what statistics are appropriate for the research questions being addressed.

Investigators should take advantage of the meeting with the statistician to find out why the analysis is appropriate and to increase their knowledge of statistics. They need to be able to defend their choice of statistics at presentations and within publications.

Several advantages result from having a plan for data analysis before starting the study. The most obvious is that the investigators are not left wondering what to do with all of the data they now have in their computer. A plan speeds the process of data analysis. If a computer program will be used, the commands for the analysis can even be written before data collection is complete. In this case, as soon as all of the data are entered, the investigators run the predetermined programs, and the analysis is ready for interpretation.

The second advantage of planning the statistical analysis before the study is an increase in scientific integrity. The investigators who have a plan ahead of time are less likely to bend the analysis to suit their purpose. A plan also prevents the process of repeating analyses until something is found that is statistically significant. A post hoc (after the fact) approach to statistical analysis is inappropriate and increases the chance of making a type I statistical error (see a future issue of this series on hypothesis testing for a discussion of type I errors).² If post hoc analyses are used, a technique such as Bonferroni adjustment is needed to decrease the chance of a type I error.²

Before beginning a study, the investigators should identify the computer, the data entry method,³ and the data analysis software they will use for the study. They also should spend time during the early phases of the project becoming familiar with the software to be used. Data analysis will proceed more smoothly if the investigators do not need to stop and ask for technical assistance.

Statistical Analysis

Overview

Statistical analysis can be a complex process. However, the statistics required for studies most commonly done in critical care transport research are fairly straightforward. Statistics are

generally descriptive (describing what is) or inferential (determining the likelihood of a real difference being present in the population). To select the most appropriate statistics, investigators need to know which type of question they are asking and the level of measurement being used for the variables. This article presents information on levels of measurement and descriptive statistics, leaving probability theory and inferential statistics for a future issue.

Level of Measurement

Level of measurement refers to the amount of information contained within the data element and to some extent the degree of detail present. Data elements are measured at the nominal (categorical), ordinal, interval, or ratio (continuous) level. The term *nominal level (or categorical) data* refers to data that can only be put into groups. For example, the demographic data elements of race and religion are measured at the nominal level. This means that the values consist of categories such as white, African American, Native American, Asian, and other. With nominal level data, no category is better than another, and the difference between categories cannot be determined. For example, neither Catholic nor Protestant is better than the other, and whether Catholic is closer to Protestant or closer to Jewish is not known. The only interpretation possible is that two subjects are or are not the same on this variable.

A specific subset of nominal level data is dichotomous data. Dichotomous data are nominal but have only two possible categories. A common example is mortality. The values for mortality are either live or die. Other dichotomous variables are things that can be measured as yes or no, on or off, or present or absent. Dichotomous variables possess characteristics beyond those of other nominal level data, but such a discussion is beyond the scope of this article.

Ordinal level data are one step up from nominal data. As the name implies, ordinal data have an inherent order. Data values such as *never*, *sometimes*, *often*, and *always* have order. An individual would interpret *sometimes* as being more frequently than *never* and *always* as more frequently than *often*. However, the difference in magnitude is not known with ordinal level data. It cannot be said of ordinal level data that *always* is twice as frequent as *often* or that the distance between *never* and *sometimes* is the same as the distance between *sometimes* and *often*. The only interpretation available is that of which is greater or which is lesser.

At the interval level of measurement, distances between data elements can be determined. Temperature is the most common variable measured at the interval level. With temperature, the difference between 40 and 50 degrees is the same as the difference between 50 and 60 degrees. However, interval level data have no true zero. Consequently, multiplication is not allowed with interval level data. This means that you cannot say that 10 degrees is twice as hot as 5 degrees. Because there is no true zero, a reference point does not exist. Zero degrees Fahrenheit or Celsius are arbitrary numbers that do not relate to the amount of temperature present. In contrast, temperature measured in Kelvin has a true zero⁴ (absolute zero), where

the absence of all energy serves as the reference point. Unlike with Fahrenheit or Celsius, at 20 degrees Kelvin, molecules are moving twice as fast as at 10 degrees Kelvin.

Data elements that have the characteristic of a true zero are measured at the ratio (continuous) level of measurement. Common examples are height, weight, and heart rate. Ratio level data are measured at the highest level of measurement and contain the greatest amount of information. Ratio level data can be transformed by addition, subtraction, multiplication, or division without altering their relative values. Ratio level data can be analyzed with the widest range of statistical methods. Ratio level data are often required for use with the most powerful statistics.²

Data elements can always be reduced in their level of measurement but can never be increased. For example, if income is measured at the ratio level (in exact dollar amounts), the data elements can be reduced to ordinal level data by the creation of categories (Table 1). However, data measured at the ordinal level of measurement cannot be increased to the ratio level of measurement. If we ask a subject what category her income fits into, we can never determine from the raw data her exact income level. Consequently, data should always be collected at the highest level of measurement possible and converted at the time of data analysis if a lower level of measurement is desired. This recommendation does not hold when there is strong reason to believe that a subject will not be truthful if she is required to provide exact data or when the data cannot be expected to be accurately measured at the higher level of measurement.

Descriptive Statistics

Descriptive statistics are numbers that summarize the data with the purpose of describing what occurred in the sample. In contrast, inferential statistics are numbers that allow the investigator to determine whether there are differences between two or more samples and whether these differences are likely to be present in the population of interest. Descriptive statistics also can be used to compare samples from one study with another. Descriptive statistics also help researchers detect sample characteristics that may influence their conclusions. For example, if a sample of air medical personnel included 400 women and only 20 men, the investigator would need to be careful about generalizing the findings to male air transport personnel.

Frequency distributions are often the first analyses to be done on a data set. Frequency distributions are a valuable

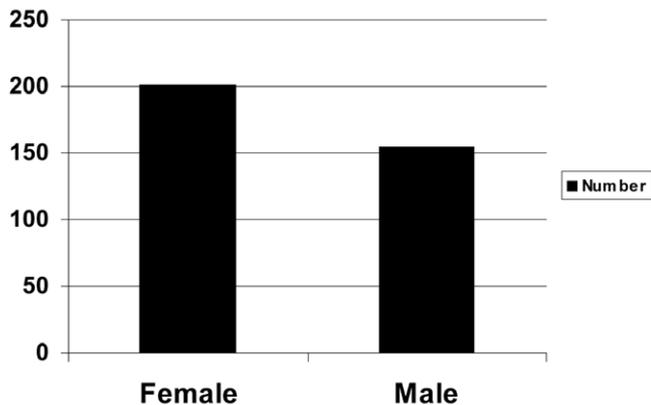
Table 1. Yearly Income

Ratio Level Data	Ordinal Level Data
4,590	< 10,000
11,230	10,000-20,000
25,600	20,001-30,000
33,775	30,001-40,000

Data from the ratio level column can be reclassified as a value from the ordinal level column but not vice versa.

Table 2. Frequency Distribution

Gender	Number (%)
Male	154 (43.4)
Female	201 (56.6)

Figure 1. Frequency Distribution (Histogram)**Table 3. Bivariate Frequency Distribution**

Gender	Professional Role		
	Paramedic	Nurse	Physician
Male	74	62	12
Female	53	142	5

Table 4. Mode for Interval or Ratio Level Data

Age	Frequency
18	5
23	1
31	1
40	3
42	4
43	2
45	1
46	2
54	1
65	1

Mode = 18

method for describing nominal or ordinal level data (discrete data). Because discrete data only characterize the quantity within categories, a frequency distribution adequately describes nominal or ordinal level data. Frequency distributions also can help detect data entry errors.

A frequency distribution consists of a description of the number of subjects selecting each possible option and may include the percentage of the sample that this number repre-

sents. For example, a frequency distribution for gender would describe how many men and how many women were in the sample. A frequency distribution can be shown using numeric values or using graphical techniques (Table 2 and Figure 1). Frequency distributions are often univariate (one variable only) but may be bivariate (including two variables). A bivariate frequency distribution is often presented as a table with the name and values of one variable across the top and the name and values of the second variable down the left side. Table 3 is an example of a bivariate frequency distribution. Multivariate frequency distributions describing more than two variables at one time are possible but become more complex and are beyond the scope of this article.

Measures of central tendency are statistics that describe where the middle of the sample lies. The lowest level measure of central tendency is the mode. The mode is the value most frequently occurring within the dataset. A mode can be used with all levels of measurement and is the primary measure of central tendency available for nominal level data. When examining the number of patients with trauma, cardiac, or other medical problems, the category having the most patients represents the mode. If evaluating the ages of a sample, the age represented by the most subjects is the mode.

Although the mode provides helpful information for nominal or ordinal level data with only a few categories, the mode may be of little value with interval or ratio level data. In Table 4, age 18 is the most common age; however, the sample is generally much older than that. Overall the sample consists mostly of individuals in their 40s, although not many subjects have the exact same age in that small range.

The next measure of central tendency is the median, the value that is in the exact middle of the sample. The median is the point at which half of the subjects lie above this value and half of the subjects lie below it. For example, the ages of the nine subjects in sample 1 of Table 5 are arranged in numeric order. The fifth age, 45, is in the exact middle of the sample; consequently, age 45 is the median. The median is a better measure of central tendency than mode because it is not influenced by an accidental grouping of values away from the true center of the data. However, the median cannot be determined for nominal level data because no order is present within the data.

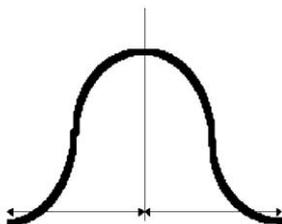
The mean (or average) is the most common measure of central tendency. The mean is calculated by adding up the value for all subjects and dividing by the total number of subjects (n). For sample 1 in Table 5, the mean is 38.3.

The mean is more sensitive to outliers and more influenced by the distribution of the values than is the median. In Table 5 two data sets are demonstrated. Both have the same median but have very different means. Consequently, different pieces of information are available with the two measures, and one or both may be relevant to the research at hand.

The use of means to describe a dataset should be limited to interval and ratio level data. Nominal level data do not have a true numeric value, so it is not possible to compute a mean. Although ordinal data might be represented using numeric values, the conceptual intervals between the values may not be the same; therefore, the mean would be difficult to inter-

Table 5. Median and Mean

	Sample 1	Sample 2
	18	34
	20	36
	21	38
	36	42
	45	45
	46	55
	52	60
	53	58
	54	62
Median	45	45
Mean	38.3	47

Figure 2. Normal Distribution (Bell Curve)**Table 6. Range and Standard Deviation Example**

	Sample 1	Sample 2
	19	33
	31	42
	42	48
	44	51
	55	53
	61	54
	62	56
	69	60
	76	62
Range	57	29
Mean	51	51
Standard Deviation	17.4	9.0

pret. For example, for never (0), occasionally (1), and always (2), the conceptual difference between never and occasionally may not be the same as the difference between occasionally and always. Thus, the numbers 0, 1, and 2 do not accurately represent the conceptual distance between values, and the mean would be skewed accordingly.

Measures of central tendency provide information on where the majority of data lie. However, these measures do not inform the reader regarding the distribution of data across possible values or their variability from one subject to the next. One method for describing a collection of values is called distribution. A normal distribution is typically described as

being bell shaped, with a middle that is exactly in the center of the distribution. In addition, the tails (sides) of the distribution are symmetric, having the exact same shape (Figure 2). The presence of a normal distribution of data in the population is a common assumption for inferential statistics.

One measure of variability is the range. Range is the difference between the greatest value and the smallest value. For the example in Table 6, the range in sample 1 is 57 and in sample 2 is 29. The range informs the reader that one set of data is more spread out (more variance) than the other.

The range, like the mean of a sample, is very sensitive to outliers or measurements that are greatly different than the rest of the sample. In Table 6, the range is wide, mostly because of the influence of only two values, 19 and 76. The rest of the values are clustered between 36 and 66, so a range of 56 could be misleading.

A measure of variability that minimizes the effects of outliers is standard deviation. A standard deviation is a mathematical calculation of the variance of all the measurements in a sample. The standard deviation can be viewed as the average distance from the mean that each of the values lies. The mathematical equations for calculating the standard deviation can be found in Burns and Grove² and are easily performed by basic statistical software programs or standard spreadsheets.

Pearson's R is another common descriptive statistic. This statistic describes the relationship between two variables. Although Pearson's R can be used descriptively, it is more commonly used as an inferential statistic. This topic will be covered in a later issue of the series.

Conclusion

In conclusion, I would like to stress that the best research studies are initiated with a statistical plan already created. This plan may or may not have been developed with the assistance of a statistician. The first step of data analysis is usually to describe the sample and then subgroups within the sample. Frequency distribution, mean, median, mode, range, and standard deviation are the most commonly used statistics for accomplishing this task.

In the next issue in the series, the basics of probability theory will be discussed. This information will be used as a background to the discussion of inferential statistics.

References

1. Thompson CB, Panacek EA. Clinical research and critical care transport: How to get started. *Air Med J* 2006;25:107-11.
2. Burns N, Grove SK. *The practice of nursing research: Conduct, critique, and utilization*. 5th ed. St. Louis: Elsevier Saunders; 2005.
3. Thompson CB, Panacek EA. Data management. *Air Med J* 2008;27:156-8.
4. U.S. Metric Association. Metric system temperature (Kelvin and degree Celsius). Available at www.lamar.colostate.edu/~hillger/temps.htm. Accessed November 30, 2008.

Cheryl Bagley Thompson, PhD, RN, is an associate professor and assistant dean of informatics and learning technologies at the University of Nebraska Medical Center College of Nursing in Omaha. She can be reached at cbthompson@unmc.edu.

1067-991X/\$36.00
Copyright 2009 by Air Medical Journal Associates
doi:10.1016/j.amj.2008.12.001