The Anatomy of Data

Joseph R. Dettori, PhD¹ and Daniel C. Norvell, PhD¹

“Data! Data! Data! I can’t make bricks without clay.”
—Sherlock Holmes, in Arthur Conan Doyle’s The Adventure of the Copper Beeches

Data is necessary to draw logical inferences in clinical research. Understanding the anatomy of data is essential if one is to obtain the correct interpretation of the data collected. The purposes of this article are to review the types of data encountered in clinical research, how they are used, and why it is important to know the differences among them.

**Types of Data**

Variables are usually labeled as qualitative (categorical) or quantitative. A qualitative variable, not to be confused with qualitative research,*¹ is one in which the variable categories are not described as numbers but instead by verbal groupings. On the other hand, quantitative variables are those in which the variables are measured in some numerical unit (Figure 1).

**Qualitative Data**

Qualitative or categorical data are counts of the number of participants or observations in each category. This data is often described with percentages or other ratios (eg, risks). Categorical data can fall into 2 classifications: nominal or ordinal. Nominal variables have 2 or more categories that have “names” (Latin nominalis, meaning pertaining to names), not numerical values. Additionally, they have no natural or intrinsic order to them. Examples of a nominal variable are sex (male or female), marital status (single, married, or divorced) and cause of spinal cord injury (motor vehicle accident, fall, or gunshot). Nominal variables can also include those that are noted to be either present or absent. When data variables have only 2 possible categories, they are often referred to as binary or dichotomous data. Examples of dichotomous presence/absence data include smoking status (yes or no), litigation pending (yes or no), and the presence of osteoporosis (yes or no).

Ordinal data is similar to nominal data except that, with ordinal data, there are categories that can be placed in distinct order or hierarchy (eg, category A is more or less severe than category B). This is illustrated with the American Spinal Injury Association (ASIA) classification where impairment is progressively greater starting with ASIA grade E (normal) and finishing with ASIA grade A (no sensory or motor function). One common mistake regarding ordinal data occurs when one assumes that the differences between adjacent categories are the same. In our example of the ASIA classification system, the difference between ASIA grades A and B may not be the same as between ASIA grades B and C, particularly with regard to how well a baseline ASIA grade predicts future function. Table 1 illustrates how categorical data is often displayed when comparing 2 interventions (early vs late surgery).

---

*¹Qualitative research is research where the data is not in the form of numbers. It is primarily exploratory and seeks to understand underlying reasons, opinions, and motivations of people. It provides insights into the problem or helps to develop ideas or hypotheses for potential quantitative research. Qualitative data collection methods vary using unstructured or semi-structured techniques such as focus groups (group discussions), individual interviews, and participation/observations. For an overview, see Grossoehme (2014).

---

¹ Spectrum Research, Inc, Tacoma, WA, USA

**Corresponding Author:**
Joseph R. Dettori, PhD, Spectrum Research, Inc, 705 South 9th Street, Suite 203, Tacoma, WA 98405, USA.
Email: joe@specri.com
There are 2 types of quantitative or numeric data: quantitative data and discrete data. Quantitative data can be divided into finer levels. When graphed, they form a distribution of values along a continuum. Distributions that form a bell-shaped curve are said to be approximately normally distributed, while all other distributions are nonnormally distributed. Approximate normal distributions should be accompanied by a mean and standard deviation (SD), while nonnormal distributions are better described by a median and interquartile range (IQR). Mean and median measures indicate where on the continuum the data tends to cluster. The SD or IQR describe the spread or variability of the data over the continuum, respectively. Examples of continuous data include a patient’s height and weight, range of spinal motion, or lumbar bone mineral density. Quantitative data is often summarized in a table such as Table 2 comparing 2 treatments (surgery vs conservative care) in patients with cervical myelopathy.

### A Relationship

Sometimes continuous data can be converted to categorical data to assist in the clinical interpretation. An example can be found in the low back pain outcomes score (LBOS). The LBOS is divided into 13 subscales ranging from current pain to leisure activity. The total score is based on a 75-point scale. Since a person could receive a score between 0 and 75 points, this score can be thought of as a continuous outcome. On the other hand, the authors of the score have also divided it into 4 potential outcome categories:

- Excellent: >65
- Good: 50-64
- Fair: 30-49
- Poor: 0-29

These arbitrary divisions create an ordinal categorical variable (ie, there is an order or hierarchy). Furthermore, it is not uncommon for authors to take categorical scores like these and divide them into “Good” and “Poor” outcomes (eg, <50 = Poor, ≥50 = Good) making them dichotomous or binary. Dichotomous outcomes are easier to work with in an analysis, since they allow for the calculation of risks and risk ratios. However, care must be taken to justify such arbitrary divisions. There is often a trade-off between clinical and statistical reasoning to form appropriate categories. Ideally, for most variables, cut points should be set a priori and not based primarily on a look at the data to determine what may seem most advantageous for the hypothesis to be tested.

### Why Does It Matter Whether a Variable Is Qualitative or Quantitative?

Understanding the data type assists in determining the type of statistical analysis used. As we have seen, descriptive statistical analyses assume that the variables have specific levels of measurement. Results from nominal and ordinal data are presented using frequency distributions. Results from discrete and continuous data are presented using measures of central tendency (mean, median, mode) and dispersion (SD, standard error, IQR). In hypothesis testing, the selection of the statistical test begins with determining the type of data used. Once that is done, other questions can easily lead a researcher to the right statistical test.

Choosing the best way to visually display data also begins with understanding data types. Nominal data can be displayed as a pie chart, column or bar chart, or stacked column or bar chart. Ordinal data is best shown as a column or bar chart while discrete and continuous data are best represented as a bar chart, histogram, line graph, or scatterplot.

### Summary

- A qualitative variable, also called categorical, is one in which the variable categories are not described as numbers but instead by verbal groupings. There are two classifications of categorical data: nominal and ordinal.
• Nominal variables have “names,” not numerical values.

• Ordinal data is similar to nominal data except that, with ordinal data, there are categories which can be placed in distinct order or hierarchy (e.g., category A is less severe than category B).

• Quantitative variables are those in which the variables are measured in some numerical unit.

• There are 2 types of quantitative data: discrete and continuous.

• Counts are discrete, measurements are continuous.

• Continuous data can be converted to categorical to assist in the clinical interpretation.

• Understanding the anatomy of data assists in determining the type of statistical analysis needed and the best way to visually display the data.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Analytic support for this work was provided by Spectrum Research, Inc. with funding from AOSpine.

References
1. Grossoehme DH. Overview of qualitative research. J Health Care Chaplain. 2014;20:109-122.
2. Greenough CG, Fraser RD. Assessment of outcome in patients with low-back pain. Spine (Phila Pa 1976). 1992;17:36-41.