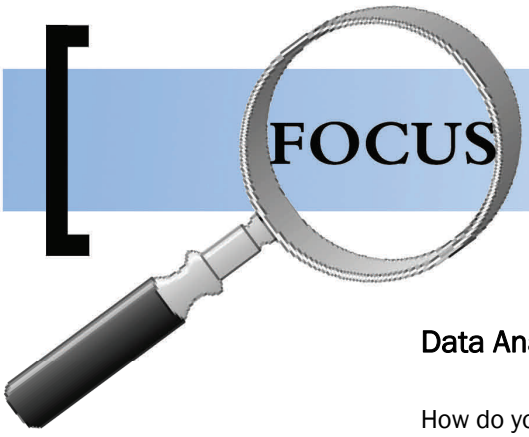
VOLUME 3, ISSUE 5

FOCUS on Field Epidemiology



Data Analysis Basics: Variables and Distribution

CONTRIBUTORS

Author:

Kim Brunette, MPH

Amy Nelson, PhD, MPH

FOCUS Workgroup*

Reviewers:

FOCUS Workgroup*

Production Editors:

Tara P. Rybka, MPH

Lorraine Alexander, DrPH

Rachel A. Wilfert, MD, MPH

Editor in chief:

Pia D.M. MacDonald, PhD, MPH

*** All members of the FOCUS Workgroup are named on the last page of this issue.**

UNC SCHOOL OF PUBLIC HEALTH

NORTH CAROLINA CENTER FOR PUBLIC HEALTH PREPAREDNESS

The North Carolina Center for Public Health Preparedness is funded by Grant/Cooperative Agreement Number U90/CCU424255 from the Centers for Disease Control and Prevention. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the views of the CDC.

How do you know whether a chemical spill in a factory caused illness in the workers? How do you know what food caused an outbreak of salmonella in your community?

In a field investigation, you often want to know whether a particular exposure (e.g., a chemical spill) is associated with any possible illness, or which of many possible exposures is associated with a particular illness (e.g., what was the potential cause of an outbreak of salmonella). You start the process of answering these questions by choosing a study design, developing a questionnaire, and gathering data in the field. All of these were discussed in previous issues of FOCUS. Once these steps are completed and you have collected your data, what comes next?

Unlike the depiction of epidemiologists in some television shows, after gathering data you don't simply have a brilliant flash of insight and solve the outbreak; you actually have to sit down and analyze that data! It is not the most glamorous part of the epidemiologist's job, but when the data lead to the source of an outbreak, the analysis is definitely rewarding. This issue of FOCUS will take you through the basic steps of descriptive data analysis, including types of variables, basic coding principles and simple univariate data analysis.

Types of Variables

Before delving into analysis, let's take a moment to discuss variables. This may seem a trivial topic to those with analysis experience, but variables are not a trivial matter. Much like people, variables come in many different sizes and shapes. Most field epidemiology, however, relies on garden-variety continuous and categorical variables.

Continuous variables are always numeric and theoretically can be any number, positive or negative (in reality, this depends upon the variable). Examples of continuous variables are age in years, weight, blood pressure readings, indoor and outdoor temperature, concentrations of pollutants in the air or water, and other measurements.

Categorical variables contain information that can be sorted into categories, rather like sorting information into bins. Every piece of information belongs in one—and only one—bin. There are several types of categorical variables: ordinal, nominal, and dichotomous or binary.

An **ordinal** variable is any categorical variable with some intrinsic order or numeric value. For example, we might categorize information on the educational status of a group of people into a variable called EDUCATION. One person may not have graduated from high school, another might have graduated from high school but received no further education, a third could have some college education or have received some other post-secondary training, and another might have graduated from college. The education levels of all members of the group will fit neatly into these categories, and the categories have an intrinsic order. A college graduate has more education than a high school graduate, and a high school graduate

has more education than someone who did not graduate from high school. Thus as the categories go from 1 to 5, the level of education increases. Other examples of ordinal variables are:

- agreement (for example, strongly disagree, disagree, neutral, agree, strongly agree)
- rating (for example, excellent, good, fair, poor)
- frequency (for example, always, often, sometimes, never)
- or any other scale (for example, "On a scale of 1 to 5, how much do you like peanuts?")

A **nominal** variable is a categorical variable *without* any intrinsic order. For example, say we have a variable called RESIDE that characterizes the part of the United States in which a person lives—the Northeast, the South, the Midwest, the Southwest, or the Northwest. The categories of this variable have no numeric value or order. Residence in the Northwest has no quantitative value compared to the Northeast. Other examples of nominal variables include sex (male, female), nationality (American, Mexican, French), race/ethnicity (African American, Hispanic, White, Asian American), or favorite pet (dog, cat, fish, snake).

A **dichotomous**, or **binary** variable is a categorical variable that has only 2 levels or categories. Many dichotomous variables represent the answer to a yes or no question. For example, "Did you attend the church picnic on May 24?" or "Did you eat potato salad at the picnic?" A variable does not have to be a yes/no variable to be dichotomous–it just has to have only 2 categories, such as sex (male/female).

Coding

Once you have gathered your questionnaire or other data, you may choose to code the data for entry into a database. Coding is the process of translating the information gathered from questionnaires or other investigations into something that can be analyzed, usually using a computer program. Coding involves assigning a value to the information given in a questionnaire, and often that value is given a label. In addition, coding can make the data more consistent. For example, if you have the question "Sex?" you might end up with the answers "Male", "Female", or "M", "F", etc. Coding will avoid such inconsistencies.

A common coding system (code and label) for dichotomous variables is the following:

0 = No 1 = Yes,

where the number 1 is the value assigned, and Yes is the label or meaning of that value.

Some like to use a system of ones and twos, where

1 = No 2 = Yes.

This brings out an important point in coding. When you assign a value to a piece of information, you must also make it clear what that value means. In the first example given above, 1 = Yes, but in the second example, 1 = No. Either way is fine, as long as it is clear how the data are coded. You can make it clear by creating a **data dictionary** as a separate file to accompany the dataset.

Similarly, we might code the dichotomous variable for sex:

0 = Female 1 = Male

Dichotomous variables can also be **dummy** variables. A "dummy" variable is any variable that is coded to have 2 levels, like the yes/no variables and male/female variables above. They can also be used to represent or stand in for more complicated variables. This is especially useful when you have many values that are more meaningful when analyzed in terms of a yes/no response.

For example, you may have collected data on the number of cigarettes smoked per week, with 75 different responses ranging from no cigarettes at all to 3 packs a week, but you can recode these data as a dummy variable: 1 = Smokes (at all), 0 = Non-smoker. You could also do this for education (1 = Any post-high school education, 0 = No post-high school education), food consumption (1 = Ate item at all during time period, 0 = Did not eat the item), and many other variables. This type of coding is useful in the later stages of analysis.

Many analysis software packages allow you to attach a label to variable values. Then the computer automatically labels the 0's as male and the 1's as female, which makes your life much easier when you are looking at the output, as shown in the example below:

Without label:	Variable SEX	Frequency	Percent
	0	21	60%
	1	14	40%
With label:	Variable SEX	Frequency	Percent
	Male	21	60%
	Female	14	40%

The coding process is similar with other categorical variables. For the variable EDUCATION mentioned above, we might code as follows:

0 = Did not graduate from high school
1 = High school graduate
2 = Some college or post-high school education
3 = College graduate.

Note that for this ordinal categorical variable, we need to be consistent with the numbering, because the value of the code assigned has significance. The higher the code, the more educated the respondent is. We could have also coded this variable in reverse order, so that 0 = College graduate, and 3 = Did not graduate from high school. In this case, the higher the code, the *l*ess educated the respondent is. Either way is fine, as long as we remember the coding when interpreting the analysis.

The following is an example of bad coding:

0 = Some college or post-high school education
1 = High school graduate
2 = College graduate
3 = Did not graduate from high school

The data we are trying to code has an inherent order, but the coding in this example does not follow that order. This is *not* appropriate coding for an ordinal categorical variable.

For a nominal categorical variable, however, the order makes no difference. Although we code each category with a number, the number does not represent a numerical value. For example, using the variable RESIDE mentioned above,

1 = Northeast
2 = South
3 = Northwest
4 = Midwest
5 = Southwest.

It doesn't matter what order we use for these categories. Midwest can be coded as 4, 2 or 5, because there is not an ordered value associated with each response.

Coding continuous variables is straightforward. If someone gives his or her age as 37 years, you enter it into the database as 37. But what if you decided that you would rather use age *categories* instead of the data you collected in years?

Creating categories from a continuous variable is common, and can easily be done using analysis software. With a software package, you can break down a continuous variable such as age into categories by creating an ordinal categorical variable, such as the following:

AGECAT
1 = 0–9 years old
2 = 10–19 years old
3 = 20–39 years old
4 = 40–59 years old
5 = 60 years or older.

You may also need to code responses from fill-in-the-blank and open-ended questions. With an open-ended question such as "Why did you choose not to see a doctor about this illness?", respondents will all answer somewhat differently. Also, you may give response choices for a particular question but offer an "other (specify)" option as well, where respondents can write whatever response they choose. These types of open-ended questions can be a lot of work to analyze. One way to analyze the information is to group together responses with similar themes. For the question above, responses of "didn't feel sick enough to see a doctor," "symptoms stopped," and "the illness didn't last very long," could all be grouped together as "the illness was not severe."

You will also need to code "don't know" responses. Typically, "don't know" is coded as 9.

Coding Tip:

Though you do not code until the data is gathered, you should think about *how* you are going to code while designing your questionnaire, before you gather any data. This will help you to collect the data in a format you can use.

Data Cleaning

One of the first steps in analyzing data is to look at the data and "clean" it of any obvious errors due to incorrect data entry.

If there are outliers (really high or really low numbers), are those numbers correct? An age value of 110 years could be an error for someone who was really 10 or 11 (or 101!).

Was a value entered that doesn't exist for the variable? For example, if 1 = male and 0 = female, and the number "2" was entered, there is clearly an error.

If there are missing values, did the person not give an answer, or was it accidentally not entered into the database?

Some software programs allow the user to set defined limits when entering data. That prevents a person from entering a 2 when only 1, 0, or missing are acceptable values. Limits can also be set for continuous and nominal variables, for example allowing only up to 3 digits for age, or limiting the words that can be entered. You can also assign field types for most kinds of variables based on what type of data the field should contain (e.g., formatting dates as mm/dd/yyyy or specifying numeric values or text). Also, some study protocols allow missing data to be entered from other sources. For example, if a person did not answer the question about age, that information could be available from a medical record being used in the study.

As a check on data entry, many data entry systems allow you to enter all of the data twice, and then compare the two entries to see if there are discrepancies. This process is called "double-entry."

Univariate data analysis, which we discuss next, is also a useful way to check the quality of the data, including checking for outliers.

Univariate Data Analysis

Univariate data analysis is important for many reasons. Looking at each variable alone, we learn a lot about the information collected. Also, univariate analysis is a good method of checking the quality of the data. Inconsistencies or unexpected results should always be investigated, using the original data as the reference point.

Simple frequencies can tell you whether most cases are young, or whether most cases are female, or whether many study participants share some other characteristic in which you are interested. First, examine the **univariate** (one variable) distribution. Graphs and tables can give you a good idea of what your variables look like, and can also help you find some of the data cleaning problems noted above.

Continuous Variables

An initial look at continuous variables can give you several important pieces of information:

- Do all subjects have data, or are values missing?
- Are most values clumped together, or is there a lot of variation? (Figures 1a and 1b)
- Are there outliers?
- Do the minimum and maximum values make sense, or could there be mistakes in the coding?

We can also conduct a univariate analysis of continuous variables to gain valuable information. Commonly used statistics include these:

Mean – average of all values of this variable in the dataset (Figure 2).

Median – the middle of the distribution, the number where half of the values are above and half are below (Figure 2).

Mode – the value that occurs most (Figure 2).

Range of values – from minimum value to maximum value (Figure 2).

Standard deviation – a measure of how variable the data are. In Figure 1a, the standard deviation is 20.4; in Figure 1b, it is 7.6. A large standard deviation compared to the values of the variable indicates that the data are widely distributed. Standard deviations can be calculated most easily with a software program, although they can be calculated by hand.

Figure 1a. Example of widely distributed age values

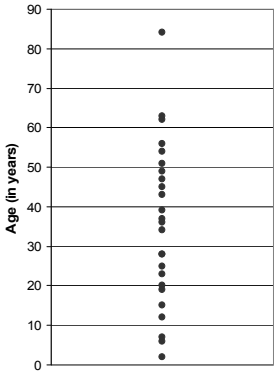
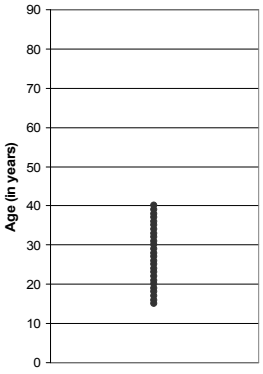


Figure 1b. Example of narrowly distributed age values



Distribution – shows whether most values occur low in the range, high in the range, or grouped in the middle (Figure 3).

Percentiles – the percent of the distribution that is equal to or below a certain value (Figure 3). In Figure 3a, the 25th percentile occurs at 4 years because 25% of the total number of respondents are 4 years or younger.

Figure 2. Statistics describing a continuous variable distribution

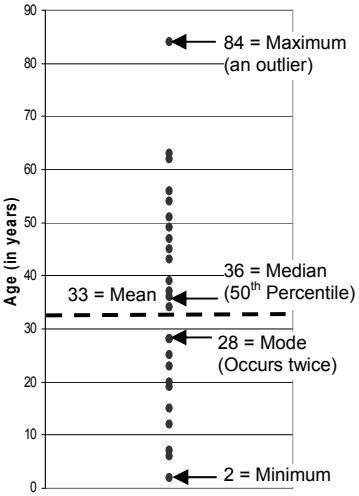


Figure 3a. "Bell-shaped" distribution for variable AGE

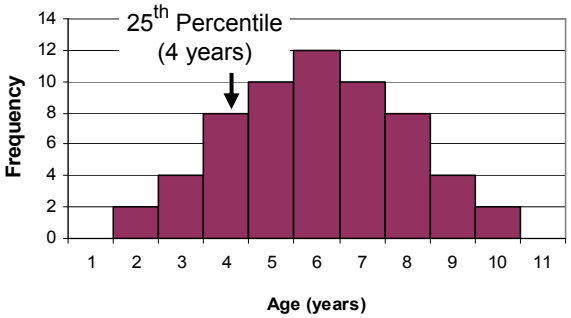
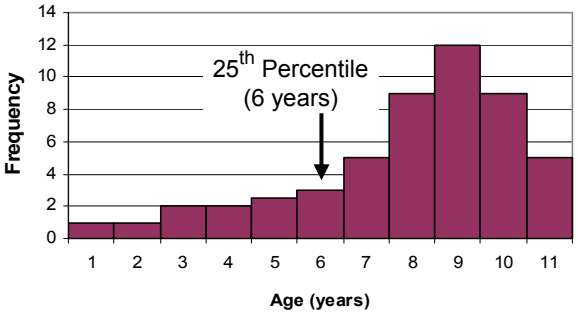


Figure 3b. Skewed distribution for variable AGE



Categorical Data

The distribution of categorical variables should also be examined before undertaking more in-depth analyses.

For example, let's look at the variable RESIDE, shown in Figure 4a. We see that more people live in the South than elsewhere, while fewer people live in the Southwest than elsewhere. The geographic area of residence may be important in some studies, because lifestyles and cultural influences tend to differ in different parts of the country. One way to put these results in context is to compare this

distribution to the "expected" distribution of residence. If we selected our study participants randomly from the US population, we would expect to have the same distribution of people in each region, assuming that they all have equal population sizes, as shown in Figure 4b.

Figure 4a: Number of people answering sample questionnaire who reside in 5 regions of the United States

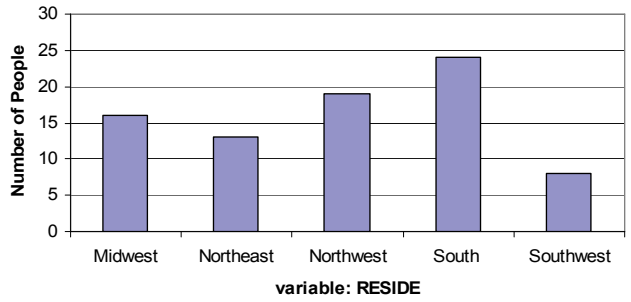
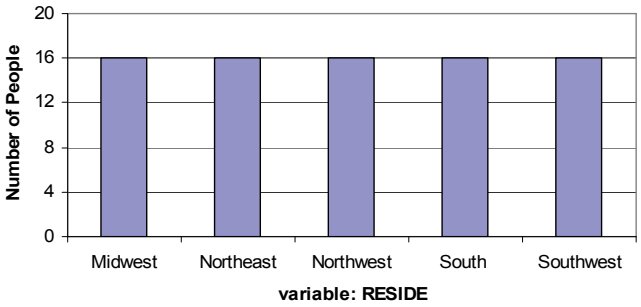


Figure 4b: Expected number of people residing in 5 regions of the United States



Another way we can take a look at this categorical data is by listing the data categories in tables. For example, Table 1 gives us the same information as Figure 4a, but in a different format. Many tables display the frequency (number) as well as the percent of the total individuals in the table. When graphs or tables are used to report and describe data, they should include all relevant information in the titles, headers, and labels, such as the study population, dates of the study, and number of participants.

In the example above, an epidemiologist would want to compare the observed number of people in each region (Figure 4a) to the expected number of people in each region (Figure 4b). Comparing the observed data with the expected data is a way for epidemiologists to see if there is something out of the ordinary associated with the observed data. We can go through the process with our ordinal categorical variable EDUCATION. Figure 5a shows the observed distribution of education levels in a given population of adults. Although the graph is labeled with education levels, those labels represent numbers used in coding (0 for less than high school education, 1 for high school graduates, and so on).

Table 1: Number of people answering sample questionnaire who reside in 5 regions of the United States

	Frequency	Percent
Midwest	16	20%
Northeast	13	16%
Northwest	19	24%
South	24	30%
Southwest	8	10%
Total	80	100%

This graph contains useful descriptive information about the study population. We can also compare this to the expected distribution of education among our study participants. Information from the US Census Bureau on the educational attainment of the US population aged 20 years or older is shown in Figure 5b. (1) This is the expected distribution of education level for the US population. By looking at the graphs and comparing the categories, we see that our study population appears to be more educated that we would have expected.

Figure 5a. Observed data on level of education from a hypothetical questionnaire

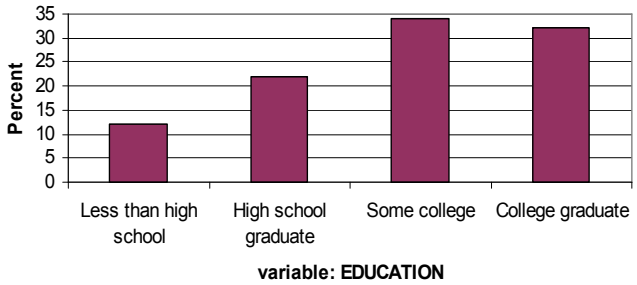
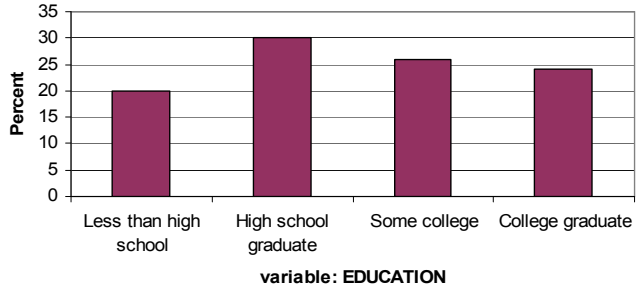


Figure 5b. Data on the education level of the US population aged 20 years or older, from the US Census Bureau



Are the observed data really that different from the expected data? This is the type of question an epidemiologist might explore further. One way to compare observed categorical data to expected categorical data is by using a statistical test such as chi-square. The next two issues of FOCUS will discuss chi-squares and other types of more extensive data analysis.

CONTACT US:

The North Carolina Center for Public Health Preparedness
The University of North Carolina at Chapel Hill
Campus Box 8165
Chapel Hill, NC 27599-8165

Phone: 919-843-5561
Fax: 919-843-5563
Email: nccphp@unc.edu

REFERENCES:

1. US Census Bureau. Educational Attainment in the United States: 2003---Detailed Tables for Current Population Report, P20-550 (All Races). Available at: http://www.census.gov/population/www/socdemo/education/cps2003.html. Accessed December 11, 2006.

FOCUS Workgroup:

- Lorraine Alexander, DrPH
- Meredith Anderson, MPH
- David Bergmire-Sweat, MPH
- Kim Brunette, MPH
- Anjum Hajat, MPH
- Pia D.M. MacDonald, PhD, MPH
- Gloria C. Mejia, DDS, MPH
- Amy Nelson, PhD, MPH
- Tara P. Rybka, MPH
- Rachel A. Wilfert, MD, MPH

If you would like to receive electronic copies of FOCUS on Field Epidemiology, please fill out the form below:

- NAME: _____
- DEGREE (S): _____
- AFFILIATION: _____
- E-MAIL ADDRESS: _____
- May we e-mail any of your colleagues? If so, please include their e-mail addresses here:

  _____

  _____

  _____

Please fax to: (919) 919-843-5563

or mail to: North Carolina Center for Public Health Preparedness
The University of North Carolina at Chapel Hill
Campus Box 8165
Chapel Hill, NC 27599-8165

Or go online: http://www.sph.unc.edu/nccphp/focus/

UPCOMING TOPICS!

- Data Analysis: Simple Statistical Tests
- Advanced Data Analysis: Methods to Control for Confounding
- Collecting Specimens in Outbreak Investigations

We are on the web!
http://www.sph.unc.edu/nccphp