Always Look at the Data

By Tanya Hoskin, a statistician in the Mayo Clinic Department of Health Sciences Research who provides consultations through the Mayo Clinic CTSA BERD Resource.

It is one of the most obvious and yet easily forgotten steps in data analysis – "look at the data." Any consulting statistician will tell you that this is essential to an accurate and appropriate statistical analysis. Why? Well, it is never safe to assume that your data set is perfectly clean. Looking at the data can reveal obvious errors, particularly impossible or improbable values and logical inconsistencies. Furthermore, looking at the data gives you a clearer understanding of the variables and the values they are taking and can help you choose appropriate statistical analyses.

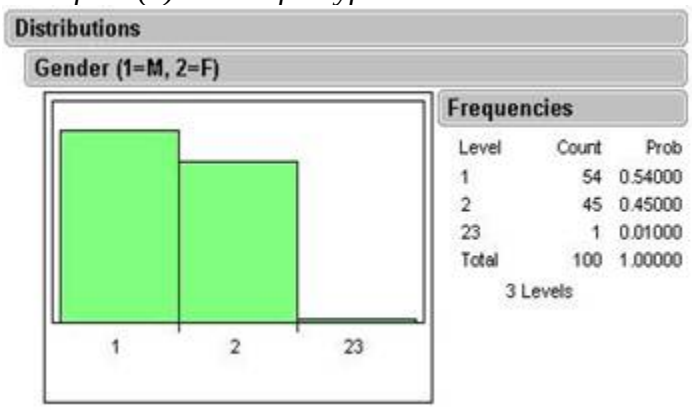
What I mean by "look at the data" is to look at numeric and graphical summaries to identify possible problems in the data and gain descriptive information. Here we will focus on data cleanup. I start by assuming you have a well-organized spreadsheet (see Data Basics). All computer output shown here was generated using JMP 5.0.1 statistical software.

Rule 1: Look at the levels of a categorical variable

Rule 1 deals with variables that classify subjects based on a category. Examples are gender, blood type and histologic stage. For categorical variables, there are usually a small number of possible values. Thus, it makes sense to look at the distribution of these variables (i.e., summarize the categories that occur in the data set) to make certain that each category is valid.

We consider data from a study of 100 patients:

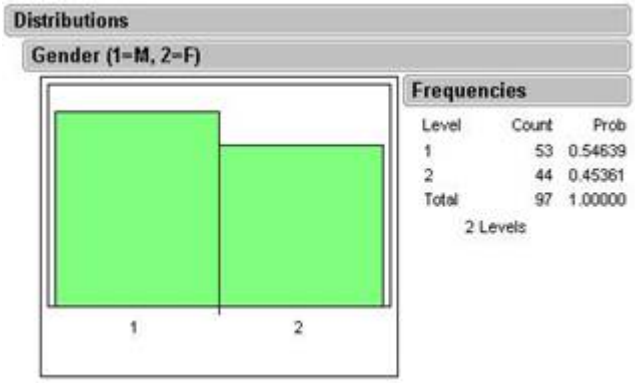
Example 1(a) – A simple typo



The bar chart on the left has one bar for each category or level of the variable. The fact that there are three bars rather than two tells us immediately that there is a problem since gender can only have two possible values. The frequency table on the right also shows us

that the levels appearing in the data set are 1, 2 and 23. We need to identify the subject with a 23 recorded, verify the correct gender, and edit the data set accordingly.

Example 1(b) – Missing observations



In this example, the gender variable has the correct number of levels and the correct numeric codes. It would be easy to mistakenly report that 55% of study participants were male and 45% were female. The point of this example is that you should always look at the total sample size included in the statistical procedure. Here the total sample size used was 97 subjects, but there were 100 subjects in the data set. Three patients do not have a value filled in for the gender variable (i.e., missing data). Many times missing data is not an error because it simply reflects reality. For a variable such as gender, however, we should have complete data for all patients. We need to identify the three patients and fill in their gender.

Rule 2 – Look at extreme observations of quantitative variables

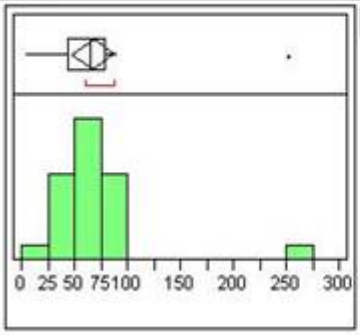
Rule 2 deals with variables that can be measured or counted on a numeric scale. Examples of this type of variable are age, systolic blood pressure, total cholesterol, height, weight, and the number of days from hospital admission to discharge. Unlike categorical variables, quantitative variables often have a large number of possible values. Because there are so many possible values, it does not make sense to look at each individual value. We can, however, look at the highest values and the lowest values recorded for a quantitative variable to identify any obvious problems.

We consider the distribution of the age variable in a study of 25 heart attack patients:

Example 2 – Multiple problems

Distributions

Age (years)



Quantiles

100.0%	maximum	253.00
99.5%		253.00
97.5%		253.00
90.0%		88.50
75.0%	quartile	79.25
50.0%	median	64.50
25.0%	quartile	44.00
10.0%		35.50
2.5%		4.00
0.5%		4.00
0.0%	minimum	4.00

Moments

Mean	67.916667
Std Dev	44.465831
Std Err Mean	9.0765498
upper 95% Mean	86.69294
lower 95% Mean	49.140393
N	24

Notice that this display looks different from the displays that we saw in Example 1. The figure on the left is called a histogram. Recall that a bar chart was used to display categorical variables in Example 1, with each bar representing a category. For a quantitative variable such as age or blood pressure, there are no inherent categories. A histogram displays quantitative data by grouping the numeric values into intervals. The bar height represents the number of observations falling into that interval. Also notice that the numeric summaries are different. In Example 1, because the variable was categorical, the software summarized the number (frequency) and proportion falling into each category. For a quantitative variable, the software summarizes the mean and standard deviation (under the heading "Moments") as well as the median, minimum, maximum, and other percentiles of the distribution (under the heading "Quantiles").

This histogram shows us that there is a subject with a recorded age above 250 years. Under the heading "Quantiles", we see that the maximum value for age is 253 years, clearly an impossible value. We would need to check this patient's record and correct the data set.

The histogram also shows that there is a patient with a recorded age in the interval from 0 years to 25 years. The quantile display shows us that the minimum age in the data set is 4 years. Although this is not an impossible value, it is an improbable value if this data set contains patients who had a heart attack. It is a good idea to check improbable values to make certain they were recorded correctly.

Finally, we look at the total number of observations included in the calculations. In the display above, in the last row under the heading "Moments", you see that N is 24. Thus, although there were 25 patients in the data set, the age variable was only present for 24. As with gender in Example 1(b), age is a variable for which we should have complete data.

Rule 3 – No one could write down all the rules

Rules 1 and 2 are the basics. Depending on your data set, there could be many additional data checks to perform. Some other data checks to consider:

- Check for duplicate observations. Data cleaning frequently reveals that some patients have multiple records in a data set. In some cases, there is a good reason. In other cases, it is a simple mistake. In either event, you need to be aware so you can handle it appropriately. Consult a statistician if you want to analyze multiple records per patient since this type of analysis requires special methodology.
- Check the order of dates. Date of recurrence should be after date of diagnosis. Date of death cannot be before date of diagnosis. The statements sound obvious, but these types of errors are often missed unless you pay close attention and check. Using software to calculate the time interval between two dates that should have a specific order is an easy way to perform this type of data check. For example, if you calculate the interval between date of death and date of diagnosis, negative values indicate a problem.
- Check logical relationships among variables. If one variable indicates that a patient did not have a CT scan and a second variable has the date of CT scan for that patient as 11/18/2001, the two variables are not consistent. If a data set has gender-specific variables, such as number of pregnancies, these variables should only be filled in for patients of the appropriate gender. The list of logic checks depends on the particular data set. Always take time to think about what logical relationships could be verified in your data.
- Use your medical and subject matter knowledge to identify more subtle inconsistencies.

Collecting your data in a consistent manner with a well-organized spreadsheet is the first step to clean data but not the last. Rare indeed would be a data set that did not suffer from at least a few typos or other mistakes. These rules don't help us find all errors but can help us find the obvious ones. Statisticians regularly check for these types of errors. If you are doing your own simple analyses, make sure you do not neglect this important step.

More information

The Mayo Clinic CTSA provides a biostatistical consulting service through its BERD Resource. More information can be found on the BERD home page.