CLARK UNIVERSITY

Exploring data
Math 218, Mathematical Statistics

D Joyce, Spring 2016

Survey of the chapter on summarizing and exploring data.

Types of values that experimental variables can have. Categorial versus numerical values. Categorical values don't have intrinsic numbers associated to them. For instance, a Bernoulli (also called binary) value is one of two values, such as Heads or Tails, Success or Failure, candidate A or candidate B. For ease of use, these categories are often associated with numbers, Success with 1 and Failure with 0, but the numbers 1 and 0 aren't intrinsic to the categories.

Categorical data are either nominal or ordinal. Nominal categories have names but nothing else. Ordinal categories can be ranked from lowest to highest, but the ranks aren't intended to be equally spaced in any way.

Numerical values are either discrete or continuous, meaning the values are either integers or real numbers. This is a minor distinction as actual measurements are usually discrete (think dollars and cents). Treatment of discrete and continuous values is pretty much the same.

Numerical values can be either interval or ratio. If the difference between two values is meaningful, but the ratios are not, then you're looking at an interval scale. When the ratios are also important, then a ratio scale as well. Other kinds of values can come up besides these two standard types, for instance, two- and three-dimensional values.

Look at exercises p. 150, 1 and 4 to illustrate types of values.

Probability theoretical variables versus experimental data. Many of the same terms are used for both, and they're closely related, but they're not the same.

A *random sample* in probability is an array of random variables

$$\mathbf{X} = (X_1, X_1, \ldots, X_n)$$

that are independent and have the same distribution. The distribution comes from a family of distributions with one or more parameters $\theta_1, \theta_2, \ldots, \theta_k$ (where k is usually 1 or 2). For instance, a Bernoulli distribution has one parameter $\theta_1 = p$, the probability of success. For another example, a normal distribution has two parameters $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

After the data is/are collected, we have n data values $x_1, x_2, \ldots, x_n$, also called a random sample, but these are outcome values, specific numbers.

The sample mean in probability is the random variable

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The data also has a sample mean, which is a specific number,

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Thus, $\overline{x}$ is an outcome for the random variable $\overline{X}$. This sample mean $\overline{x}$ can be used to estimate the parameter p for a Bernoulli distribution or the parameter μ for a normal distribution. A lot of this course is about finding good estimators for parameters, and $\overline{x}$ as an estimator of μ is the most important example.

There is also a sample median $\tilde{X}$ and its outcome value $\tilde{x}$, and quartiles and percentiles etc. More generally, there are order statistics. If the outcome values are listed in order, you get the order statistics. The smallest value is X_{min}, also denoted $X_{(1)}$, the largest is X_{max}, also denoted $X_{(n)}$, and the intermediate values denoted

$$X_{\text{min}} = X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} = X_{\text{max}}.$$

Their outcome values are denoted with lowercase x's.

Besides the central statistics of mean and median, there are also statistics of spread, or dispersion. The primary ones are standard deviation and variance which we'll study in detail later in the course.

Graphical and visual techniques. Histograms and Pareto charts, stem & leaf plots, box & whiskers plots (in their many variations) are all used to help display the data to help us get a feel for the distribution of data.

A good place to find out about this kind of thing is Edward Tufte's book *The Visual Display of Quantitative Information*. It goes beyond statistical displays and considers displays of general quantitative information. If you find it interesting, you may want to read his other books.

You may have seen logarithmic and/or exponential graph paper. Plotting data using these scales can convert graphs of logs, powers, or exponents into straight lines that are visually easy to recognize. They're also used in statistics when the data is one of these forms.

A very useful scale or plot for statistics is the normal one. If data is normally distributed, then when it is plotted on a normal plot, then the data will appear to be linear, but nonnormal data will not appear linear.

Bivariate data. When only one variable X is sampled in the experiment, we get the sample $X_1, X_2, \ldots, X_n$, but when two variables X and Y are sampled, we get a pair of samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ that are somehow related. Many experiments are designed to see what the connection is between two such values, and many of the statistical tests we'll study in later chapters are about this kind of bivariate data.

Simpson's paradox. If we have time, we'll examine the example in the text.

Covariance and correlation coefficients. Correlation and causation.

Straight line regression. We'll look at this in detail in a later chapter. We'll only have time here to descibe the basic concept.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`