

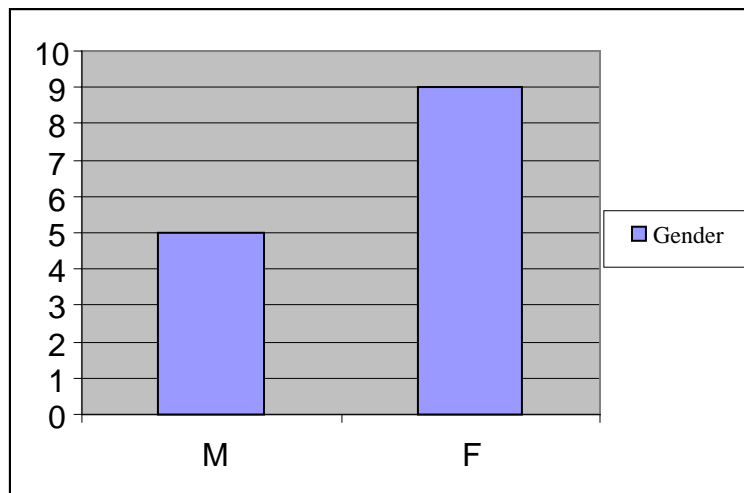
TECHNOLOGY SEMINAR - 07

Sheldon L. Epstein

Bachelor of Science – Electrical Engineering, Massachusetts Institute of Technology

In the last Technology Seminar (TS06), we ended our discussion of *Probability* (<http://en.wikipedia.org/wiki/Probability>) with an introduction to the study of health and illness *statistics*, which is called *Epidemiology* (<http://en.wikipedia.org/wiki/Epidemiology>). For this seminar, we will examine some important characteristics of *data* (<http://en.wikipedia.org/wiki/Data>) called *statistics* (<http://en.wikipedia.org/wiki/Statistic>). Briefly, a *statistic* is a result of applying a mathematical or statistical function to a set of data.

We begin by examining the easiest health and illness statistic to analyze, which is *gender* (<http://en.wikipedia.org/wiki/Gender>). There are only two possible choices; namely, Male and Female. Assume a classroom containing 14 students – 5 Male and 9 Female. Gender data may be displayed in a vertical *bar chart* (http://en.wikipedia.org/wiki/Bar_chart) of a *frequency distribution* (http://en.wikipedia.org/wiki/Frequency_distribution) as shown below.

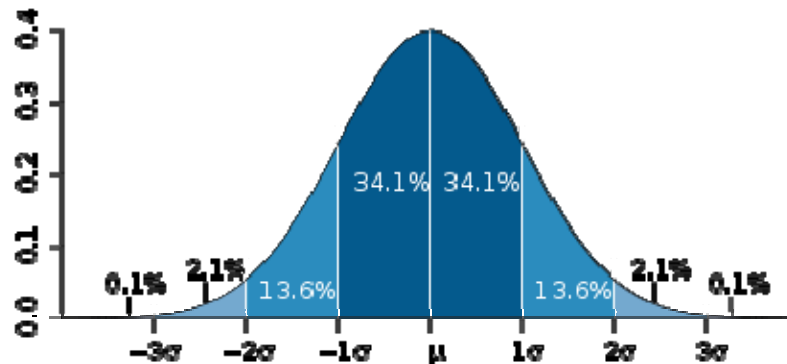


While analyses of absolute numbers are useful, they are not as useful as *probability distributions* (http://en.wikipedia.org/wiki/Probability_distribution). For example, suppose you wanted to compare gender distributions of all classrooms in a school or between schools. Then, you would find that expressing data as percentages, such as $5/14 = 0.357 = 35.7\%$ for Males and $9/14 = 0.643 = 64.3\%$ for Females, to be more useful.

Data sets such as Gender, where there are only 2, 3, 4, ... 10 mutually exclusive choices, generally are easily analyzed and not particularly useful or challenging. Instead, consider measurement data, such as *personal income* (http://en.wikipedia.org/wiki/Personal_income), *platelet count* (http://en.wikipedia.org/wiki/Platelet_count#High_and_low_counts), or weight. For large populations (e.g. 10,000 or more individuals), these measurement can yield hundreds or hundred-thousands of possible values. Therefore, *probability distributions* (http://en.wikipedia.org/wiki/Probability_distribution) are much more useful.

There are several important mathematical distribution functions – such as *binomial* and *Poisson*. For many classes of problems, measurement data have a *Normal distribution* (http://en.wikipedia.org/wiki/Normal_distribution). A *Normal distribution* is a specific mathematical function that depends solely on its *mean* (<http://en.wikipedia.org/wiki/Mean>) and

its *standard deviation* (http://en.wikipedia.org/wiki/Standard_deviation). In news media parlance, it is called the *bell-shaped curve* (http://en.wikipedia.org/wiki/Bell-shaped_curve). In technology, its name is the *Gaussian Curve* (http://en.wikipedia.org/wiki/Gaussian_curve). A Gaussian Curve drawn by http://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg is shown below:



where μ = *mean* and σ = *standard deviation*. Note that $34.1\% + 34.1\% = 68.2\%$ of the area within the curve is located within $\pm 1\sigma$ of the mean μ .

If your teacher ‘grades on the curve’, then students with scores within ± 1 standard deviation of the mean are given a grade of **C**. Scores between $+1\sigma$ and $+2\sigma$ earn a grade of **B**, while those with a higher score are awarded a grade of **A**.

Intelligence tests (http://en.wikipedia.org/wiki/Intelligence_tests), such as those that yield an *Intelligence Quotient (IQ)* score, are believed to generate IQ score data that are normally distributed. Whether this is true for the United States population as a whole is a hotly debated subject with significant political overtones. For example, see a controversial book that is described at http://en.wikipedia.org/wiki/The_Bell_Curve.

The next question to be considered when examining populations with a large number of individuals is whether it is necessary to conduct a *census* (<http://en.wikipedia.org/wiki/Census>) to measure or count each member or whether one could obtain an accurate estimate of a variable by *sampling* ([http://en.wikipedia.org/wiki/Sampling_\(statistics\)](http://en.wikipedia.org/wiki/Sampling_(statistics))) of a small percentage of members.

Article 1, Section 2 of the U.S. Constitution and the Fourteenth Amendment require an *Enumeration* (<http://en.wikipedia.org/wiki/Enumeration>) of our population every 10-years for purposes of apportionment of the House of Representatives and other important reasons. Republicans assert that *Enumeration* means a *census* in which every individual is counted. Democrats reply that *Enumeration* permits *sampling*, which they claim will be more accurate because minorities and aliens (non-citizens) have traditionally been under-counted. The method that is chosen will determine whether Illinois will lose a seat in the House of Representatives. See Supreme Court Justice Clarence Thomas’ dissenting opinion in *Utah et al v. Donald L. Evans, Sec. of Commerce et al* (<http://www.law.cornell.edu/supct/html/01-714.ZX1.html>). Also see Wikipedia entry at http://en.wikipedia.org/wiki/Utah_v._Evans.

This Technology Seminar note is at <http://www.k9ape.com/publicservice/PSM/TS07.pdf>. The INTERNET version contains active URL links for your convenience.