TECHNOLOGY SEMINAR - 06

Sheldon L. Epstein
Bachelor of Science – Electrical Engineering, Massachusetts Institute of Technology

In the last Technology Seminar (TS05), we described how probability theory could be used in a *Coin Toss* game to study the question of whether a coin is *fair*. In that analysis, we proposed a *hypothesis* that the coin is *fair* and then discussed an experiment to obtain data that could be used to test the *hypothesis*.

You should be able to extend the *Coin Toss* game analysis to a *Die Toss* game, usually called *dice* (http://en.wikipedia.org/wiki/Dice) or *craps*, that is played with two 6-sided *dies* **A**,**B**. Each side of a die is embossed with one or more *pips* to indicate an *integer* drawn from a set [1…6]. In *craps*, an objective is to gamble on the outcome of a roll of two *dice* that is the sum of two *pips* on top faces of the *dies* and may be expressed as **A**+**B** = **B**+**A**, since order is immaterial.

There are 6 x 6 = 36 possible outcomes in a range [2…12]. Expanding on the *Coin Toss* example from last month's notes, an *Outcome Distribution Table* can be constructed as follows:

A+B	**OUTCOMES**	**P(A+B)**
2	1+1	1/36
3	1+2, 2+1	2/36
4	1+3, 2+2, 3+1	3/36
5	1+4,2+3, 3+2, 4+1	4/36
6	1+5, 2+4, 3+3, 4+2, 5+1	5/36
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1	6/36
8	2+6, 3+5, 4+3, 5+2, 6+2	5/36
9	3+6, 4+5, 5+4,3+6	4/36
10	4+6, 5+5, 6+4	3/36
11	5+6, 6+5	2/36
12	6+6	1/36

The *symmetry* of the *Outcome Distribution Table* is immediately obvious. Would *symmetry* also appear in *Outcome Distribution Tables* for **A**-**B**, **A**·**B**, **A**/**B** and $\mathbf{A}^{\mathbf{B}}$?

Analyses such as those for the *Coin Toss* and *Craps* games are very useful and can be expanded to other applications if care is taken. For example, consider how decimal digits appear in tables such as street addresses in telephone directories or prices in stock quotation tables where data have dimensions. What is the probability that the first or *most-significant digit* (MSD) is *1*?

You know that the MSD must be selected from the set [1…9] because a leading *0* is always suppressed (except for very rare occasions such as 0N356W style addresses). Would P(*1*) = 1/9? That might be a good first guess; but, it would be a wrong guess. In fact, P(*1*) ≈ 2/9 to 3/9. For an explanation of why that is so, please read descriptions of *Benford's Law* that appear at http://en.wikipedia.org/wiki/Benford_Law & http://mathworld.wolfram.com/BenfordsLaw.html.

Benford's Law was based on observation. It is one example of a situation where an experiment has already been conducted and we seek to construct a *hypothesis* that will link a *cause* to an *effect*?

One area in which we already have data on an effect; but, now seek cause(s), is health or illness of populations. The science is called *Epidemiology* (http://en.wikipedia.org/wiki/Epidemiology). Examples of recent *epidemiological* questions that have been posed include:

1. Does cigarette smoking cause lung cancer?
2. Does cellular telephone usage cause brain cancer?
3. Does drinking water containing *fluoride* reduce tooth decay?
4. Does drinking water containing *fluoride* cause cancer?
5. Does drinking soda pop contribute to *obesity* (http://en.wikipedia.org/wiki/Obesity) that leads to an increased *mortality rate* (http://en.wikipedia.org/wiki/Mortality_rate) or an increased *morbidity rate* (http://en.wikipedia.org/wiki/Morbidity_rate)?
6. Does continuous use of *aspirin* (http://en.wikipedia.org/wiki/Aspirin) reduce incidence of heart disease (http://en.wikipedia.org/wiki/Heart_disease)?

In each of these questions, an "experiment" has already been conducted because we already have population statistics. The next step is to construct a *hypothesis* and then evaluate existing data to test a hypothesis such as:

- *Cigarette smoking does cause lung cancer* (American Lung Association - http://www.lungusa.org/site/c.dvLUK9O0E/b.33484/k.438A/Quit_Smoking.htm), or
- *Cigarette smoking does NOT cause lung cancer* (The Tobacco Institute - http://www.lungusa.org/site/c.dvLUK9O0E/b.33484/k.438A/Quit_Smoking.htm).

Epidemiological questions are difficult to analyze because not everyone exposed to a cause experiences an effect and not everyone who experiences an effect has been exposed to a cause. Thus in the case of cigarette smoking, not every cigarette smoker contracts lung cancer and not every lung cancer patient is a cigarette smoker (for example, consider exposure to household *radon* gas (http://en.wikipedia.org/wiki/Radon).

Primary responsibility for collection of *epidemiological* data in the United States rests with the Dept. of Health & Human Services (HHS) Center for Disease Control and Prevention located in Atlanta (http://en.wikipedia.org/wiki/Center_for_Disease_Control_and_Prevention and see *Morbidity* and *Mortality* reports at http://www.cdc.gov/mmwr/).

Data collection is supplemented by State and municipal public health agencies. In Illinois, public health data are collected by the Illinois Dept. of Public Health (http://www.idph.state.il.us/health/statshome.htm). Health data are also collected by the Cook County Department of Public Health http://www.cookcountypublichealth.org/) and the Chicago Board of Health (search GOOGLE for long URL).

Data from these sources can be readily obtained for *Science Fair* (http://cssf.org/Student/stu-ijasguide.html) projects. Winners are eligible for college scholarship awards and prizes won by you are highly valued by college admissions officers.

This Technology Seminar note is at http://www.k9ape.com/publicservice/PSM/TS06.pdf. The INTERNET version contains active URL links for your convenience.