The Four Central Concepts of Probability and Statistics
Carl Wagner – June 17, 2010

It is the hallmark of a civilized person to be able to burst into tears upon reading a column of statistics.

Bertrand Russell

1. *Probabilities are relative frequencies, either (i) observed or (ii) predicted.*

(i). Observed relative frequencies: Suppose that I toss a die twenty four times, and the numbers 5,1,4,4,2,3,6,1,2,5,3,1,2,5,5,1,2,6,1,2,5,4,5, and 1 come up. I record these outcomes, both as *frequencies* and as *relative frequencies*, in the following table:

Outcome	Observed Frequency	Observed Relative Frequency
1	6	6/24
2	5	5/24
3	2	2/24
4	3	3/24
5	6	6/24
6	2	2/24
Column sums:	24	1

Note that the sum of all the observed frequencies is equal to the total number of observations, and the sum of all the observed relative frequencies is equal to 1. This is always the case. Relative frequencies always take a value between 0 and 1 inclusive. The term *empirical probability* means exactly the same thing as *observed relative frequency.*

(ii). Predicted relative frequencies: Suppose that I am about to toss a die twenty four times, and want to predict the relative frequencies of the outcomes 1,2,3,4,5,and 6. I might examine the die, notice how symmetrical it is, how it doesn't seem to be weighted on a particular side, etc.. On the basis of this examination, I decide to predict that the six possible outcomes will occur equally often, i.e., that the relative frequency of each will be 4/24. These *predicted relative frequencies* are also called *theoretical probabilities.* They are also said to constitute a *probability model.* In this case (and many others) where we estimate that the relative frequencies of all the possible outcomes are identical, we are using what is called the *uniform model.* There are many, many other probability models (e.g., the *binomial model, the geometric model, the hypergeometric mode, the Poisson model, the normal model, etc.),* and long experience has identified in which situations each model is appropriate (i.e., yields good estimates of relative frequencies). You will encounter some of these models, particularly the binomial and the normal model, in any good course in probability and statistics.

2. *Statistical theory gives us tools to evaluate probability models (also called "hypotheses") in the light of observed relative frequencies.*

Naturally, it is rare that a probability model gives an exactly correct prediction, i.e, we don't expect that, once we make our observations, the observed relative frequencies will exactly match the predicted relative frequencies. But, if our model is a reasonable one, we expect the observed relative frequencies to be "reasonably close" to the predicted relative frequencies. The elaboration of the meaning of "reasonably close" is developed in the part of statistics called *hypothesis testing.* (In the die tossing case above, we would use something called the *chi-squared test* to check the reasonableness of the uniform model in the light of the observed relative frequencies.)The basic idea is that when our observations deviate too greatly from what our model predicts, we reject the model.

3. *In hypothesis testing, the proper way to measure deviation from what is predicted by a model is* ***not*** *in terms of percentages, but in terms of an important unit of measurement called the "standard deviation."*

Here is a simple example: Suppose I am planning to flip a coin 100 times and you are planning to flip another coin 400 times. We both think that our coins have an equal chance of coming up "heads" or "tails," and so I estimate the frequency of heads in 100 flips to be 50 (i.e., I estimate the relative frequency to be 50/100 = ½) , and you estimate the frequency of heads in 400 flips to be 200 (i.e., you estimate the relative frequency to be 200/400 = ½) .Now I actually flip my coin, and get 55 heads. And you flip your coin and get 220 heads. We each observed a greater number of heads than we predicted. In each case we observed 10% more heads than we predicted ((55 – 50)/50 = 1/10 = 10%, and (220 – 200)/200 = 1/10 =10%). So it looks like our models performed equally on the prediction task (whether equally badly or equally well is yet to be determined, and depends on how stringently we decide to evaluate models). *But in fact (and this is one of the most important insights of statistical theory) my model has performed considerably better than yours!* It turns out that 55 is just one "standard deviation" above 50, but 220 is two "standard deviations" above 200. In terms of standard deviations, your observation is much further away from what your model predicted than my observation is from what my model predicted. So your observations cast much more doubt on the reasonableness of your model. The business of determining the standard deviation for a probability model is somewhat complicated, but you will learn how to do it in any good probability and statistics course. Just keep your eye peeled for its appearance.

4. *Statistical theory also gives us the tools for estimating features of a population, based on observation of a random sample of that population. The accuracy of such estimates depends only on the size of the sample,* ***not*** *on the percentage of the population sampled.*

This is another case where thinking in terms of percentages, which is so reasonable in many areas of life, is **completely inappropriate** in statistics. This strikes most people as very strange (the fancy mathematician's term is "counter-intuitive), but it is one of the pillars of statistical wisdom. Here is an example: Suppose I take a random sample of 400 Tennessee voters to see which presidential candidate they support, and I want to do the same thing in California, and have my estimates have an equal margin of error. Common sense would suggest that since California has a much larger population than Tennessee that I would have to poll a larger number of California voters to get the same margin of error. But common sense is nonsense in this case! Polling 400 Californians will give me the same margin of error as polling 400 Tennesseans. In fact, the margin of error of each poll is roughly 5 percentage points. Most statistics courses make a really big deal about how to compute the margin of error under various degrees of stringency, but for practical estimation here is what you should remember: If you sample n people, (1) take the square root of n, (2) take the reciprocal of that square root, and (3) express the result as a percentage. (That gives the margin of error with what is known as “95% confidence.”). In our example, we take the square root of 400, getting 20. Then we take the reciprocal of 20, getting 1/20. Then we convert 1/20 = 0.05 to a percentage, getting 5%. This is how the margin of error that you see reported in newspaper articles is derived. In particular, suppose that 40% of 400 people polled in Tennessee favored Obama. The newspapers would report that the percentage favoring Obama was 40%, with a margin of error of 5%. This is sometimes written $40\% \pm 5\%$ (with $\pm$ read as “plus or minus”). Another way to put it is that based on the poll, the set of numbers between 35% and 45%, inclusive, usually denoted [35%, 45%] is a “95% confidence interval” for the percentage of all Tennessee voters that favor Obama. This means that if we use this method for constructing confidence intervals in many different polls, then about 95% of the time the true population percentage will lie in that interval.