



"turning data into dollars"

Tom's Ten Data Tips – November 2009

Applied Probability Theory

Probability was only 'invented' fairly recently, and is a sub-branch of mathematics. Commonly, Fermat and Pascal (halfway 17th century) are considered the founding fathers of probability theory. Because chance events are all around us, they pervade all of our lives.

Applied probability is used in many disciplines like Operations Research, Game Theory, Engineering Telecommunications, Monte Carlo simulations, technical trading, modern Biology, and many, many other fields. By employing probability theory and statistics we cope with uncertainty and random perturbations all around us.

1. Our Brain Isn't Rigged (Yet?) For Calculating Proportions

It wasn't until the time of Galileo, shortly after Fermat and Pascal, that we began calculating expected outcomes from chance events using proportions. This began in the 2nd half of the 17th century. Mathematical problems are much harder to solve when you frame a question in terms of proportions as opposed to natural frequencies (see also tip# 6).

Because mankind has only been working with proportions for a few hundred years, some writers attribute human shortcomings to the way our brains are hardwired. There is also compelling evidence our neural processing of small numbers (up to three) works different from processing numbers four and larger (see e.g. The Math Gene). Probably not coincidentally, the way we *write* numbers 1-3 (variations of one to three lines) shows remarkable similarities across many, if not all, languages. Given these constraints, human tasks should be structured accordingly to avoid known processing bottlenecks. Taleb (2001): "... researchers of the brain believe that mathematical truths make little sense to our mind, particularly when it comes to the examination of random outcomes."

2. "Coincidences" Are Actually Quite Likely

What people consider a "remarkable coincidence" is a noteworthy event which has an a priori low probability of occurring. Let's take an example. I was recently walking in Paris, Rue du Faubourg (I live in

the Netherlands), and ran into a friend of my daughter. What a coincidence! It turns out that so-called coincidences are actually quite common, or rather: much more likely than we (generally) think. For any particular "coincidence" the probability is very low. However, there are many more *possible* coincidences than most people think. And since we forget or dismiss "ordinary" events as noise, the "coincidences" really stand out.

In a classroom with 23 kids, how large is the probability that two kids (or more) have their birthday on the same day? That turns out to be 50.73%, more than most people would estimate. When I talked to my daughter's friend, she told me that she was now working as a full-time fashion model. Of course therefore her a priori chance of being in Paris jumps up dramatically (but do we always discover these changes in prior odds??). Usually it's the overwhelmingly large number of *possible* noteworthy events that defies our imagination. Could this be a possible source for superstition?

3. Framing Is Everything

When you analyze a probability problem, the way you frame it is very important. History has taught us how major breakthroughs and advances have come from framing an existing problem in new ways. Let's take rolling dice, for example. When you throw *one* dice, the outcomes 1-6 have a uniform probability distribution. This simply means all six outcomes are equally likely. The outcome of throwing *two* dice (and summing) leads to 11 possible outcomes, a compound event.

Often compound events can be modeled as a *sequence* of two (or more) simple events. The same with throwing *two* dice: by framing this as a *sequence* of two simple events the problem is simple. Analyzing the compound event "sum of two dices" becomes much more tractable by framing it as a sequence of two simple events. Decomposition of difficult problems in a series of simpler challenges has led to significant advance in the theory of probability.

4. Statistics Is Probability Applied To Real World Events

One of the primary concerns of statistics is drawing inferences from data. These data points need to be collected in specific ways, to allow drawing valid conclusions about the "real world." What this typically means is that samples need to be unbiased and random so that every real world element has an equal chance of winding up in the data set.

If those conditions are met, advanced math can be used to draw conclusions that could be exceedingly difficult or costly to measure without the help of statistics.

Probability theory is axiomatic and therefore such problems have precise solutions. A conclusion is correct, or it's not. Statistics on the other hand, deals with real world phenomena, and therefore observations include measurement error and random variation. So solutions need to be "reasonable" so that a "normal" flow of events would likely lead to such outcomes.

5. Venn Diagrams Are Fun

Although I never realized it at the time they taught me this branch of math in high school, the algebra of events provides a universal language to derive mathematical laws that can be extremely powerful in probability theory. Commutative ($A + B = B + A$), Associative ($A + (B + C) = (A + B) + C$) and Distributive laws ($A(B + C) = AB + AC$) are the basic ingredients for almost everything you'll need on a day to day basis in applied probability.

Thinking which events can and cannot occur together, and the rules that govern their behavior is closely related to querying sets of data in a database. Once you know how to tie these together, you can quickly "guesstimate" the likelihood of particular outcomes while you are working through 'big' SQL queries. A full analytical solution can be hard and cumbersome, but some "handkerchief algebra" is worth its weight in gold to quickly test for the likeliness of a given outcome (e.g.: row count) and thus plausibility of a correct SQL statement.

6. Do Mental Arithmetic With Frequencies, *Not* Proportions

It's amazing how transforming a question from proportions to natural counts (frequencies) can make a very difficult problem easy to solve, all of a sudden. If a medical test is 90% reliable, and 1% of the population suffer this condition, how likely is it you are affected if you receive a "positive" test which suggests you're infected? Using Bayesian probability (proportions) you can solve this problem, but very few people can do this without paper and pencil or a calculator. Let's try using frequencies, now.

Within a population of 10,000, 100 people are affected and 9900 are healthy. When you test them all, 90 sick people get a positive result, but also 990 healthy people (8910 healthy people get a negative

result, and 10 sick people, too). So 90/990, or about 9% of “positive” results are with people who are actually sick. Easy. *If you use frequencies...*

7. Unlikely Occurrences Are *Under Estimated*

Human beings are not very “objective” observers. We are continuously inundated by impressions all around us, and to avoid “information flood” we imbue reality with meaning. Research shows that we sometimes infer patterns that aren’t really there. Our need for structure seduces us into making out regularity that isn’t really there. Just because something hasn’t happened (after all, we didn’t see it) does not mean it won’t happen. All market crashes of the last few decades are directly attributable to such phenomena. The fact that reckless (financial market) trading strategies have been “working” does not make them sound. Merely lucky. Russian roulette wins five times out of six, too. Nick Leeson & Jerome Kerviel come to mind. How many others are (still) lauded as “star” traders?

The general finding from psychological research is that we (severely) under estimate the likelihood of improbable events. Nick Leeson did not hedge against *both* a drop in stock prices *and* an earthquake in Kobe, exacerbating the former effect. That’s why Barings went down.

Our sense of assessing the magnitude of very small fractions (like 0.000001) badly needs cognitive help. For instance, by multiplying this fraction with the amount at risk, and then *visualize* the pile of dollar bills that are at stake. Is it a pile, a briefcase, trunk, or a truckload of bills?? Fooled by Randomness (Taleb, 2001) and Black Swan (Taleb, 2007) are two excellent books that shed more light on these topics.

8. Benford’s Law Was Discovered By Newcomb

Benford’s law (1938) is used as a method to detect fraud (in particular fraudulent bookkeeping). It pertains to any “natural” measure of magnitude, size, weight, length, etc. The larger the number of measurements taken, the closer the distribution of *the first (significant) digit* will follow this law. It turns out that in natural measurements, there are more outcomes *beginning* with a 1, than a 2, than a 3, etc. with the least number of entries beginning with a 9. These have the following probabilities: 1 – 30.1%, 2 – 17.6%, 3 – 12.5%, 4 – 9.7%, 5 – 7.9%, 6 – 6.7%, 7 – 5.8%, 8 – 5.1%, 9 – 4.6%.

A random generator does not follow Benford's law, and maybe very "smart" bookkeeping fraudsters will use this phenomenon to avoid being caught. J. Newcomb discovered the same law in 1881 (half a century earlier), but history has forgotten him, much like we think Watson (1736-1819) invented the steam engine (which Newcomen was the first to build in 1712, really).

9. Use Monte Carlo Simulations For Intractable Problems

The Monte Carlo algorithm was invented around 1946 by Stanislaw Ulam and John von Neumann while they were working on project Manhattan (making of the first atomic bomb).

The Monte Carlo algorithm is an efficient way to simulate chance processes, some of which may or may not even have an analytical solution. In the latter case, Monte Carlo simulations are sometimes the only way to get a grip on complicated problems. And *many* real life phenomena (often apparently "simple" ones, like queue management) are either exceedingly difficult, or at present unsolvable using mathematical analysis.

10. Applied Probability Is Qualitative As Much As Quantitative

By this I mean that *application* of probability theory is as much about framing the problem appropriately (see also tip# 3), assumptions, and considering the wider framework ("model") in which mathematics is applied, as it is a pure numbers game. Debacles like the (once huge) Long Term Capital Management hedge fund, devised and run by Nobel prize winners Scholes & Merton, have shown that getting it only a little bit wrong can cost serious money (at some point they controlled ~\$130 Bn). Most technical trading leverages arbitrage between the way financial markets are *supposed* to work, and the way they do. Apparently that spread can be rather profitable.

The current credit crisis is often blamed on risk models. But it was really a breakdown in corporate governance. Someone needs to decide which models to apply, and determine which model boundaries make it lose its validity (and therefore discontinue use). That's a governance call, that *can not* be grounded in numbers, at least not in numbers *alone*.