Lecture Notes: Paradoxes in Probability and Statistics

Kris Sankaran

October 29, 2011

> I think it is much more interesting to live with uncertainty than to live with answers that might be wrong. – Richard Feynman

1 Introduction

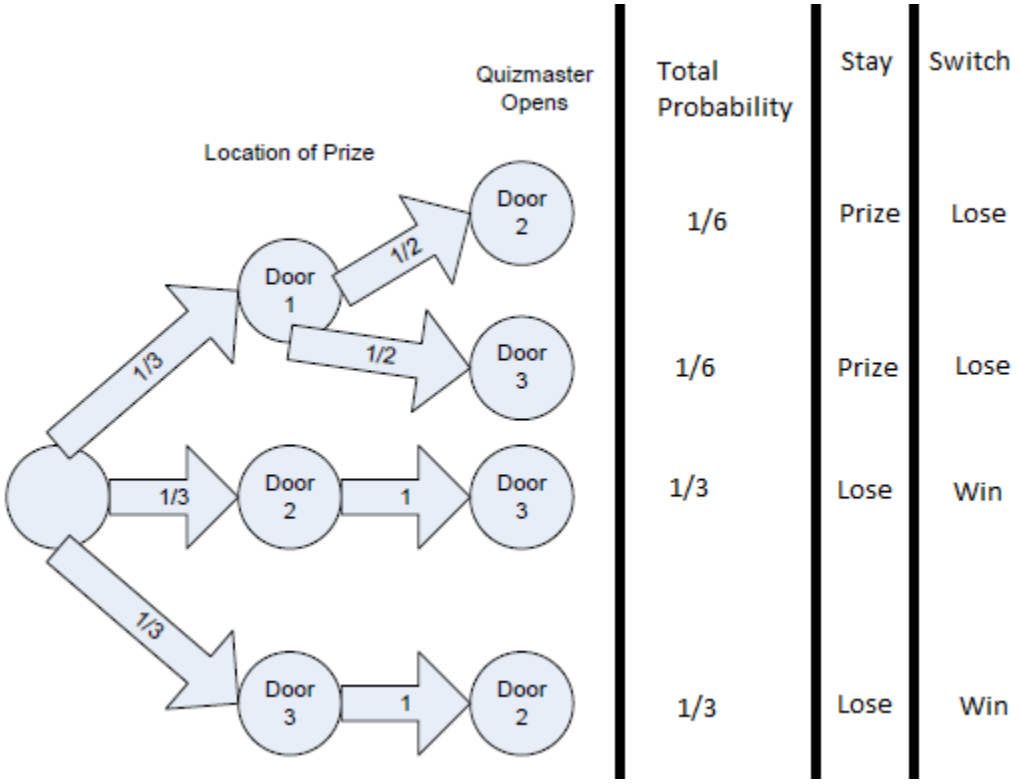
The purpose of these notes is to give a very brief and informal introduction to a few classic paradoxes. Indeed, paradoxes of probability and statistics abound at all levels of mathematical sophistication, and they begin to hint at the richness of these interrelated fields of study.

Ultimately, probability and statistics are about making decisions in the face of uncertainty. We often are asked to make predictions given whatever information we have about possibilities, chances, and past occurrences. Probability and statistics provide a sophisticated and mathematical framework for thinking about uncertainty and randomness. We'll find this way of thinking crucial in approaching the following counterintuitive problems about chance and information.

2 Monty Hall Paradox

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No.1, and the host who knows what's behind the doors, opens another door, say No.3, which has a goat. He then says to you, "Do you want to pick door No.2?" Is it to your advantage to switch your choice?

The following table shows the probability of every possible outcome if the player initially picks door No.1.



Therefore, the player should switch–doing so doubles the probability of winning from $\frac{1}{3}$ to $\frac{2}{3}$.

3 St. Petersburg Paradox

This problem was posed in 1713, and it is known as the St. Petersburg Paradox because the Bernoulli brothers, whose claims to fame include research on integral calculus, comets, and fluid flows, wrote about the question for the Academy of St. Petersburg. Here's the problem statement:

Consider the following gambling strategy. Bet \$1 on the outcome of a fair coin flip. If I'm right, I win a dollar and call it a day. If I loose, then I wager \$2 and flip it again, with the hope that, if I win this time, then I'll make up the loss from the previous bet and in fact make a \$1 extra. If I loose in the second flip too, then I make *another* bet, this time wagering $\$2^2$. The idea here is that, I'll make up the loss from the previous two tosses (\$1 + \$2), and in fact make an extra dollar. In general, if I loose the first $n-1$ bets, then I bet $\$2^{n-1}$ on the n^{th} bet, with the hope

that I would recuperate my losses from the first $n-1$ bets. To reiterate, we lose $\$2^{n-1}-1$ dollars from the first $n-1$ bets.

$$\begin{aligned}
\text{Losses from bets 1 to } n-1 &= \$1+\$2+\cdots+\$2^{n-2} \\
&= \sum_{i=0}^{n-2} 2^i \\
&= \frac{1-2^{n-1}}{1-2} \\
&= \$2^{n-1}-1
\end{aligned}$$

So for our n^{th} bet, we wager $\$2^{n-1}$ dollars. This seems like a fool-proof strategy for winning \$1, which should make us a little suspicious of this game already (otherwise, why isn't everyone playing this game?).

The key insight is that the expected amount of money that we need to lose before we win the game is infinite. This is not at all obvious at first glance, so we'll develop some probability tools to quantify this notion of an "expected value." For a rigorous definition and discussion of an "expected value" in probability, consult one of the recommended readings (or Wikipedia!). For our current puposes, the expected value corresponds to what we call the "average" in regular conversation.

In probability, a formula for computing the expected value of a function of a random variable (in non-math-speak this means a formula for computing the average value of a quantity that depends on some random number) $g(Z)$ where Z takes on values k with probabilities $\mathbf{P}(Z=k)=p_k$ is $\mathbf{E}(g(Z))=\sum_{k=0}^{\infty} g(k)\times p_k$. We let N be the random variable representing the first time that we win. We let $g(N)$ be the amount of money that we lose up to the time that we win. So, if $N=n$, the amount of money that we lose is $g(n)=\$(1+2+\cdots+2^{n-2})$, as computed above. Notice then that $\mathbf{P}(N=n)=\left(\frac{1}{2}\right)^n$, because we would have had to lose $n-1$ bets and win the n^{th} bet, which all occur with probability $\frac{1}{2}$. Using our formula from above, we have the following expected value for the amount of money we lose before we win

our last bet:

$$\begin{aligned}
\mathbf{E}\left(g\left(N\right)\right) &= \sum_{n=1}^{\infty}\left(1+2+\cdots+2^{n-2}\right)\left(\frac{1}{2}\right)^{n} \\
&= \sum_{n=1}^{\infty}\left(\frac{1-2^{n-1}}{1-2}\right)\left(\frac{1}{2}\right)^{n} \\
&= \sum_{n=1}^{\infty}\left(\frac{2^{n-1}-1}{2^{n}}\right) \\
&= \sum_{n=1}^{\infty}\left(\frac{1}{2}-\frac{1}{2^{n}}\right) \\
&= \infty
\end{aligned}$$

The gambler is going to have to lose quite a bit of money before winning! [Notes about computation: The second line follows from the formula for the sum of a finite geometric series. Between the fourth and fifth line, we used that $\sum_{n=1}^{\infty}\frac{1}{2}$ is an infinite sum while $\sum_{n=1}^{\infty}\frac{1}{2^n}$ is a finite sum, so the difference between the two is still infinite. The actual computation is not that important, and these formulas and facts can all be found on wikipedia: http://en.wikipedia.org/wiki/Geometric_series#Sum.]

So, we have rigorously established that, even though the game described in the St. Petersburg Paradox seems like a great game for the gambler, it could never actually be used to win money, because the expected amount of money lost before winning is infinite (in a precise, mathematical sense).

4 Simpson's Paradox

Up till now, I've made it seem like probabilisits and statisticians spend all their time thinking about fun puzzles (especially gambling ones). While there *are* lots of fun motivating paradoxes for important ideas in probability and statistics, there are also some paradoxes that emerge in the reality of working with data which can have an important impacts on decision-making in a variety of fields. Indeed, this wide scope of significance is one of the most appealing features of statistics.

Often, it's difficult to perform experiments to test particular hypotheses, so people analyze data that has been collected without any rigorous scientific protocal. For example, we might analyze health records instead of designing an experiment comparing different health practices. Without a control, conclusions drawn from this data can be difficult to interpret or assess. It is in this context that Simpson's paradox emerges.

Simpson's paradox occurs when there is a reversal in associations present between different groups when the groups are analyzed at different levels of resolution. This is much easier seen with data. Below, is a data set about legal policy that exhibits this phenomenon. It is based of off 4764 murder cases tried in Florida between 1973 and 1979:

Murderer	Death Sentence	Other	%
black	59	2448	2.4
white	72	2185	3.2

The proportion of black murders sent to death is about the same as the proportion of white murderers; in fact, the proportion for whites is slightly larger. Note what happens when we break this data down by the race of victims as well.

Victim	Murderer	Death	Other	%
black	black	11	2209	0.5
	white	0	111	0.0
white	black	48	239	16.7
	white	72	2074	3.4

The association reverses now: black murderers now face harsher sentences than their white counterparts. In particular, black murderers whose victims were white face much harsher sentences than any other group in the table, while no white murderers whose victims were black faced the death sentence. The reason the data suggested that the two groups faced the same percentage of death sentences in the original presentation was because 1) murderers seem to select victims of their same race and 2) the sentences of those murderers whose victims were white is much harsher than those whose victims are black.

5 Recommended Readings and Websites

These notes are only designed to give a brief and informal introduction to some of the interesting ideas of probability and statistics. Here are some readings and websites that you may enjoy and that would give you the opportunity to learn these concepts in more depth.

5.1 Probability

1. "50 Challenging Problems in Probability" by Frederick Mosteller
2. "Introduction to Probability Models" by Sheldon Ross
3. "An Introduction to Stochastic Processes" by Paul G. Hoel, Sidney C. Port, and Charles J. Stone
4. "Probability and Random Processes" by Geoffrey Grimmet and David Stirzaker
5. math.stackexchange.com (search for "probability")

5.2 Statistics

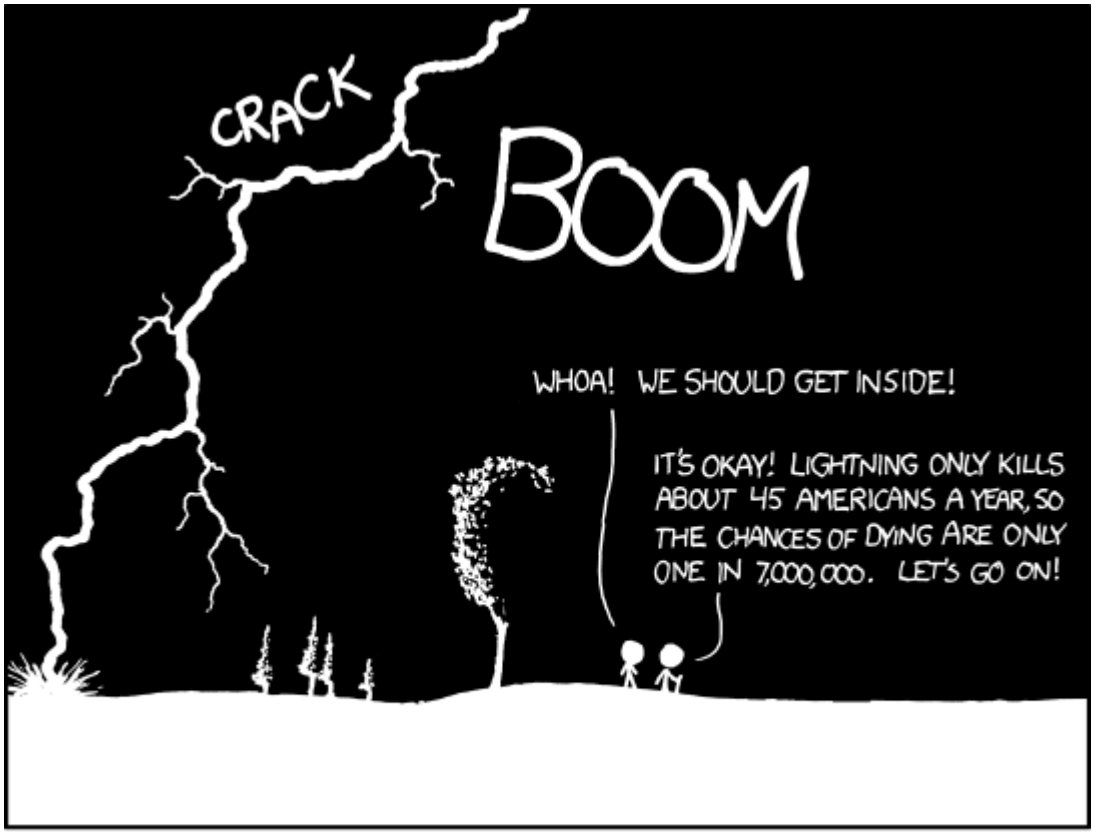
1. "All of Statistics" by Larry Wasserman
2. "Mathematical Statistics and Data Analysis" by John Rice
3. stats.stackexchange.com

5.3 Data and Data Visualization

1. "The Visual Display of Quantitative Information" by Edward Tufte
2. Flowingdata.com and "Visualize This" by Nathan Yau
3. Informationisbeautiful.net and "Information is Beautiful" by David McCandless
4. kaggle.com
5. datawithoutborders.cc

6 Conclusion

Happy exploring!



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.