Numeracy Nugget #4 - 'He Probably Doesn't Understand Probability'

George Rebane

The title of this nugget is a true statement, but to accept that requires that you first understand some things about probability. Let me attempt a short explanation.

Most of our knowledge is unreliable or probabilistic, meaning that most of what we know is known only to within a range of uncertainty. Examples – the time our school age youngster gets home in the afternoon (between 2:30 and 2:45PM), the odometer reading of our car (somewhere from 37,500 to 38,500 miles), the duration of World War 2 (oh, say about 5 to 6 years), the current US population (maybe 290 to 300 million), whether it will rain next week, (one will get you five that it will), … All of these little bits of knowledge and foreknowledge are technically called **random variables or RVs**. And after a little reflection, you will conclude that your entire store of knowledge is made up of nothing but random variables.

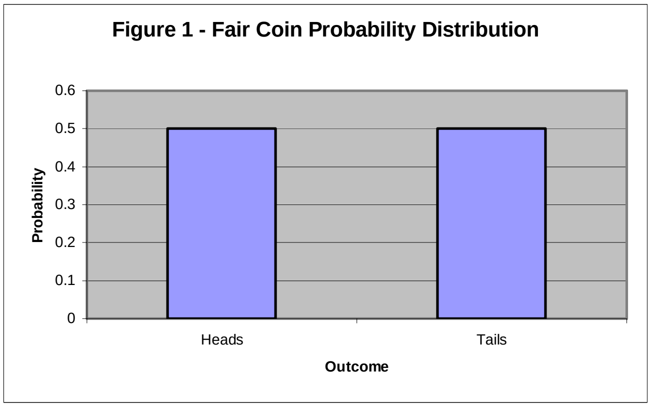


Figure 1 - Fair Coin Probability Distribution

More specifically, a RV may be a number, a word or statement, a 'logical' like TRUE or FALSE, or any proposition we can recall or communicate to another person. Well known RVs are the outcome of a coin flip and drawing a card from a 52-card deck. RVs come in two primary flavors – discrete (roll of a die) and continuous (tomorrow's high temperature). The important thing to keep in mind here is that **<u>the value of a RV is known only to within a probability distribution</u>**. Probability distributions therefore also come in two flavors – discrete and continuous – and, as we will discover, are themselves part of our overall beliefs. So, for the same RV, you may have a probability distribution that is different from mine.

Now what is a probability distribution? The simplest distribution is the discrete probability distribution that determines the outcome of a coin flip. For a fair coin it can be represented as a histogram (Figure 1) with equal 'weight' bars, each bar denoting the probability of heads or Pr(heads) being 50% or 0.5 along with the same weight for Pr(tails) = 0.5. Recalling NN3, here the 'certain event' is that the coin will land either heads or tails which requires that Pr(heads) + Pr(tails) = 1, the probability of certainty. (Also recall that 0 is the probability of an event that is either impossible or impossible to verify.) In general, the probabilities in a discrete distribution, that represent all of the possible outcomes, will sum to one. For example, it is easy to show (Figure 2) that the exhaustive 6x6 = 36 possible outcomes of rolling two dice is the set of numbers two through twelve having respective probabilities {1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36} which sum to one.

Before going on, we realize that now we know how to represent the distribution of an unfair coin, say, Pr(heads) = 0.6 and Pr(tails) = 0.4 by simply adjusting the heights (really the weight) of the histogram bars appropriately (Figure 3).

So where do RVs come from? Well, they are generated naturally through known or yet to be discovered 'laws' that govern all the processes that can happen in our universe. This is a bit deep, but nevertheless true. We can get a feel for the kinds of realworld processes that generate RVs by simulating one our PC. Your spreadsheet program has a little routine that outputs a (pseudo random) number anywhere from 0 to 1 with equal probability every time it runs. This says that any number from 0 to 1 can appear with equal likelihood, there are no favorites. We can then take this 0 to 1 interval and divide it up into bins or segments of appropriate lengths on the 0 to 1 interval by specifying where the bin boundaries are. Each bin represents an outcome for the RV that we wish to simulate. (Think of taking the histogram bars and laying them down end to end to make up a total length of 1.) Now taking a sequence of these random numbers and seeing which bins they fall into will exactly simulate the outcome of coin flips, or dice throws, or drawing of a card, or … . What we have simulated is the realworld process of actually flipping a coin, throwing dice, drawing cards, or any other process represented by how we set up our outcome bins on that 0 to 1 interval. The length of each bin is seen to represent the probability of that outcome.

Figure 3 - Biased Coin Probability Distribution

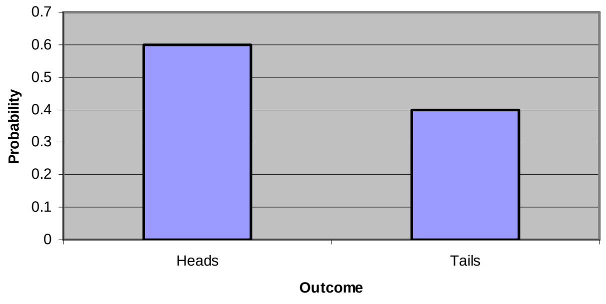
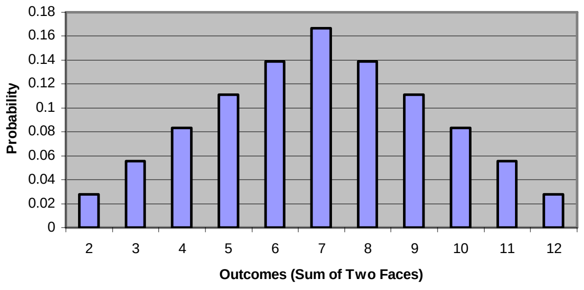


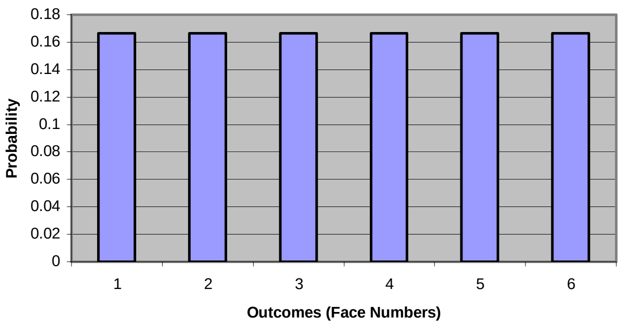
Figure 2 - Two Dice Probability Distribution



But how do we pick the lengths of the bins? Well, we now come full circle to confront our personal collection of beliefs (see NN3) since the bin lengths or probabilities that we set on our random number generator represent what we believe to be the *relative frequency* with which our simulator will spit out events represented by that bin. For example **if we believe** the theoretical arguments of how a single die behaves, we will set up six equal length bins labeled '1', '2', and so on. Each bin will be 1/6 or approximately 0.1667 long (Figure 4).

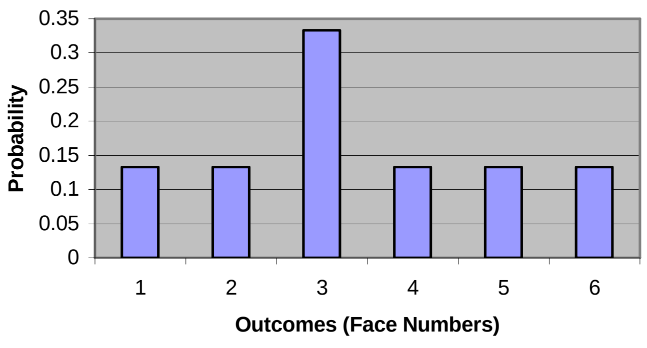
And when we ‘roll’ the simulated die a 1,000 times we will get each face approximately 166 times more or less, or the relative frequency of 1/6. As you might expect, the error in the fraction from the simulations will get smaller the more times we ‘roll the simulator’.

Figure 4 - Single Fair Die Probability Distribution



However, if we have reason to *believe* that the die is loaded so that, say, the face ‘3’ will come up with twice the normal probability or 2/6, then we would assign the outcome ‘3’ on our simulator a bin length of 2/6 = 1/3 and divide the remaining 2/3 of the random number generator among the five remaining faces giving each face a bin length of (2/3)/5 = 2/15 = 0.1333 (Figure 5). Of course it doesn’t matter in what order you place these labeled outcome bins on the 0 to 1 interval, only their relative lengths matter. **Probability is then the length of the bin(s) that you would select or accept (from some reliable source) that you *believe* would replicate your knowledge of how past outcomes of a RV have occurred or would best predict how the future outcome of a RV may come to pass.**

Figure 5 - Single Biased Die Probability Distribution



So now you should have a good working grasp of what a probability is in the sense of it representing random future outcomes, and how to set up a discrete probability distribution to simulate or model the occurrence of any set of discrete outcomes that represents your (here’s that word again) belief of what might happen. We will cover continuous random variables and their probability distributions in a future Nugget lest this one grow into a boulder.

Now we can go back to the beginning and verify that ‘He Probably Doesn’t Understand Probability’ is a true statement. Informally such a statement (or proposition) indicates that it’s most likely or that there is a high probability (say, greater than 0.8) that the stated claim is true. Then how big would you make the bin on the random number generator that represents the fraction of Americans who don’t understand probability in the sense we have just covered? Well, the US government says that it is upwards of 19 out of 20 people or a relative frequency of 19/20 = 0.95 or 95% of folks you could query at random out there in television land who have little or no working sense of random variables and probability. So you would make the ‘Doesn’t Understand Probability’ outcome bin about 0.95 long and the complement, or 0.05, be the bin length of ‘Understands Probability’. This distribution would replicate and/or predict how frequently you would run into both kinds of people. Therefore, the statement is true for a randomly picked ‘he’ (it works for a ‘she’ also).

Before we finish, let’s relate odds (or likelihood ratio LR) to probability. If your friend says ‘It’s 3 to 1 that the Raiders will win.’, he is really saying that he believes the probability Pr of the outcome can be computed from LR = 3/1 = Pr/(1 – Pr). So solving for Pr (remember high school algebra) gives Pr = LR/(1+LR) = (3/1)/(1+(3/1)) = 0.75. Now if you, as the clever probabilist, think that the probability of the Raiders winning is only 1/5 or 0.2, you would be willing to give odds of LR = 0.2/(1 - 0.2) = 0.25 = ¼ or ‘4 to 1’ that the Raiders will lose. In any event, you would be happy to take any bet your friend offers based on his quantified belief (that word again) of the future. In an upcoming Nugget we will examine how different beliefs (in terms of probability distributions) can lead to disagreements, disasters, dilemmas, and daring acts of apparent heroism.

Solution to NN3 (Monty Hall) Three Door Problem

The usual answer to the question ‘Should you switch the initially selected door?’ is that it doesn’t make any difference since there are two closed doors remaining so the chance (probability) is 50-50 that the prize is behind either door. This answer is wrong and points out how subtle and often difficult are problems that involve chance and risk - intuition does not always serve.

Consider the same problem but with a lot more doors, say, a thousand. The prize is behind only one door, unknown to you but known to Monty. You select a door which we, for easy reference, will call Door1. Monty now goes and opens all the other doors except Door253. Of course, the prize is not behind any of the open doors. Now Monty smiles and asks whether you want to keep Door1 or switch to Door253. At this point many people have an intuitional epiphany and instantly see that the correct answer is ‘Monty, I wanna switch, puhleeze let me switch!’ Why is that?

Well, the chance (probability) that you picked the right door – Door1 – in the first place is exactly 1/1000; not a good likelihood of winning. Since no one is allowed to move the prize once the game starts, this probability will not change no matter how many doors Monty now opens as long as he doesn’t open all the remaining doors (Door2 through Door1000). After you select Door1, the probability that the prize is behind <u>one of the other closed doors</u> is 999/1000, almost a certainty. And that probability will not change either no matter how many doors Monty opens as long as he doesn’t open all the remaining doors. If he opens just one door, the probability is still 999/1000 that the prize is behind one of the remaining 998 doors. If he opens another one, the probability is still 999/1000 that the prize is behind one of the remaining 997 closed doors. You now get the idea that if Monty continues opening all but one of the doors that you did not choose, the remaining closed door will hide the prize with a probability of 999/1000 So if Door253 remains closed, you know that its probability of hiding the prize is almost certainty or precisely 999/1000. You most definitely want to switch doors.

With this under our belt, we can see the solution to the three door problem on the TV show. The door that the contestant first selects has 1/3 probability of hiding the prize. The chance that the prize is behind one of the other two doors is 2/3. When Monty then eliminates one of those two doors by opening it, the remaining closed door now has all to itself the probability 2/3 of hiding the prize, and you would double your chances of winning from 1/3 to 2/3 if you switched. So the answer is, in this game you always switch when given the chance.

Anecdotally, when after much debate (some of it between later embarrassed mathematicians) this solution became known, the long-running ‘Let’s Make a Deal’ immediately lost its appeal and Monty was out of a job.

And now for the next Nugget Noodler (how does he come up with these inanities?) we have a little problem that builds on the subject of probabilities.

NN4 Problem: It is equally likely that a canvas sack contains either a red ball or a blue ball. Into this sack is dropped a red ball so that it now contains two balls. Later a random draw from the sack yields a red ball. What is the probability that the sack now contains one red ball?