Notes for Recitation 22

Philosophy of Probability

Applying probability to real-world processes often involves a little bit of philosophy. Let's first consider this simple problem: What is the probability that

$$N = 2^{6972607} - 1$$

is a prime number? One might guess 1/10 or 1/100. Or one might get sophisticated and point out that the Prime Number Theorem implies that only about 1 in 5 million numbers in this range are prime. Or one can say that assigning a probability to this statement is nonsense because there is no randomness involved; the number is either prime or it isn't.

This question highlights the distinction between two philosophical approaches to probability. One school of thought says that probabilities can only be meaningfully applied to *repeatable processes* like rolling dice or flipping coins. In this view, the probability of an event represent the fraction of trials in which that event will occur. This view is sometimes called classical statistics, sampling theory, or the frequentist approach.

An alternate view is the Bayesian approach, in which a probability can be interpreted as a *degree of belief* in a proposition. A Bayesian would agree that the number above is either prime or composite, however would be perfectly willing to assign a probability to each possibility. The Bayesian approach is thus broader and willing to assign probabilities to any event, repeatable or not. One challenge with the Bayesian approach is coming up with reasonable *prior* probabilities for events that only occur once.

As an aside, it is not clear whether Bayes himself was Bayesian in this sense. However, a Bayesian would be willing to talk about the probability that Bayes was Bayesian while a sampling theorist would say that is nonsense because there is no repeatable process that generates Bayes' beliefs!

Getting back to prime numbers, there is a probabilistic primality test due to Rabin and Miller. If N is composite, there is at least a 3/4 chance that the test will discover this. (In the remaining 1/4 of the time, the test is inconclusive; it never produces a wrong answer.) Moreover, the test can be run again and again and the results are independent. So if N actually is composite, then the probability that $k = 100$ repetitions of the Rabin-Miller do not discover this is at most:

$$\left(\frac{1}{4}\right)^{100}$$

So 100 consecutive inconclusive answers would be extremely convincing evidence that N is prime! If you're comfortable using probability to describe your personal belief about primality after such an experiment, you might be a Bayesian. Otherwise, you might prefer more traditional views of probability.

The Bayesian/Frequentist divide is an interesting one philosophically, but relatively minor in practice; The mathematics of probability remains the same and either approach can lead one astray when modeling real-world processes if the model is based on unsound assumptions. This differences aren't relevant to the mathematics of probability that we teach in 6.042, but do come up in practical courses on statistics, estimation and decision theory.

A Sampling-Theory Approach to Polling

Consider a simple yes/no public opinion poll. In a classical view, every person in the population has a definite opinion and so we assume that there is some fraction p of the population would answer "yes" to the question and the remaining $1 - p$ fraction would answer "no". (Let's forget about the people who hang up on pollsters or launch into long stories about their dog — real pollsters have no such luxury!) Now, p is a fixed number, not a randomly-determined quantity. So trying to determine p by a random experiment is analogous to trying to determine whether N is prime or composite using a probabilistic primality test.

Probability slips into a poll since the pollster samples the opinions of a people selected uniformly and independently at random. The results are qualified by saying something like this:

> "One can say with 95% confidence that the maximum margin of sampling error is ± 4 percentage points."

This means that either the number, q, reported in the poll is within 0.04 of the actual fraction, p, or else an unlucky 1-in-20 event happened during the polling process; specifically, the pollster's random sample was not representative of the population at large. This is *not* the same thing as saying that there is a 95% chance that q is within 0.04 of p; it either is or it isn't, just as N is either prime or composite regardless of the Rabin-Miller test results.

Suppose that a coin that comes up heads with probability p is flipped n times. Then for all $\alpha < p$

$$\Pr\left\{\#\text{ heads} \le \alpha n\right\} \le \frac{1-\alpha}{1-\alpha/p} \cdot \frac{2^{nH(\alpha)}}{\sqrt{2\pi\alpha(1-\alpha)n}} \cdot p^{\alpha n}(1-p)^{(1-\alpha)n}$$

where:

$$H(\alpha) = \alpha \log_2 \frac{1}{\alpha} + (1-\alpha) \log_2 \frac{1}{1-\alpha}$$

1   Approximating the Cumulative Binomial Distribution Function

A coin that comes up heads with probability p is flipped n times. Find an upper bound on

$$\Pr\left\{\#\text{ heads} \ge \beta n\right\}$$

where $\beta > p$. Think about the number of tails and plug into the monster formula above.

Solution.

$$\Pr\left\{\#\text{ heads} \ge \beta n\right\} = \Pr\left\{\#\text{ tails} \le (1-\beta)n\right\}$$

Now tails comes up with probability $1-p$. So the answer is the same as above with α replaced by $1-\beta$ and p replaced by $1-p$:

$$\Pr\left\{\#\text{ heads} > \beta n\right\} \le \frac{\beta}{1-\frac{1-\beta}{1-p}} \cdot \frac{2^{nH(\beta)}}{\sqrt{2\pi\beta(1-\beta)n}} \cdot p^{\beta n}(1-p)^{(1-\beta)n}$$

Here we're using the fact that $H(1-\beta) = H(\beta)$. ∎

2   Gallup's Folly

A Gallup poll found that 45% of the adult population of the United States plan to vote Republican in the next election. Gallup polled 640 people and claims a margin of error of 3 percentage points.

Let's check Gallup's claim. Suppose that there are m adult Americans, of whom pm plan to vote Republican and $(1-p)m$ do not. Gallup polls n Americans selected uniformly and independently at random. Of these, qn plan to vote Republican and $(1-q)n$ do not. Gallup then estimates that the fraction of Americans who plan to vote Republican is q.

Note that the only randomization in this experiment is in who Gallup chooses to poll. So the sample space is all sequences of n adult Americans. The response of the i-th person polled is "yes" with probability p and "no" with probability $1-p$ since the person is selected uniformly at random. Furthermore, the n responses are mutually independent.

a. Give an upper bound on the probability that the poll's estimate will be 0.04 or more too low. Just write the expression; don't evaluate yet!

   Solution. We can regard each response as a coin flip that is heads with probability p. In these terms, qn is the total number of heads flipped. So we have:

   $$\Pr\{qn \le (p-0.04)n\}$$
   $$\le \frac{1-(p-0.04)}{1-(p-0.04)/p} \cdot \frac{2^{nH(p-0.04)}}{\sqrt{2\pi(p-0.04)(1-(p-0.04))n}} \cdot p^{(p-0.04)n}(1-p)^{(1-(p-0.04))n}$$

   ∎

b. Give an upper bound on the probability that the poll's estimate will be 0.04 or more too high. Again, just write the expression.

   Solution. Reasoning as before and using the answer to the preceding problem gives:

   $$\Pr\{qn > (p+0.04)n\}$$
   $$\le \frac{p+0.04}{1-\frac{1-(p+0.04)}{1-p}} \cdot \frac{2^{nH(p+0.04)}}{\sqrt{2\pi(p+0.04)(1-(p+0.04))n}} \cdot p^{(p+0.04)n}(1-p)^{(1-(p+0.04))n}$$

   ∎

c. The sum of these two answers is the probability that Gallup's poll will be off by 4 percentage points or more, one way or the other. Unfortunately, these expressions both depend on p— the unknown fraction of voters planning to vote Republican that Gallup is trying to estimate!

   However, the sum of these two expressions is maximized when $p = 0.5$. So evaluate the sum with $p = 0.5$ and $n = 640$ to upper bound the probability that Gallup's error is 0.04 or more. Pollsters usually try to ensure that there is a 95% chance that the actual percentage p lies within the poll's error range, which is $q \pm 0.04$ in this case. Is Gallup's poll properly designed?

Solution. The probability that the error is 0.04 or more is at most

$$\frac{.54}{1-.46/.5} \cdot \frac{2^{640\cdot H(.46)}}{\sqrt{2\pi\cdot.46\cdot.54\cdot640}} \cdot .5^{.46\cdot640}.5^{.54\cdot640}$$

$$+$$

$$\frac{.54}{1-.46/.5} \cdot \frac{2^{640\cdot H(.54)}}{\sqrt{2\pi\cdot.54\cdot.46\cdot640}} \cdot .5^{.54\cdot640}.5^{.46\cdot640}.$$

$$= \quad 2\cdot\frac{.54}{1-.46/.5} \cdot \frac{2^{640\cdot H(.54)}}{\sqrt{2\pi\cdot.54\cdot.46\cdot640}} \cdot .5^{.54\cdot640}.5^{.46\cdot640}$$

$$\leq \quad .427\cdot 2^{640\cdot H(.54)}(.5)^{640}$$

$$= \quad .427\cdot 2^{640\cdot H(.54)-640}$$

$$\leq \quad .427\cdot 2^{-3.008}$$

$$\leq \quad .054$$

This means that p will lie within the error range of a polled fraction with probability 0.946. So our estimates suggest Gallup's poll is not quite large enough to meet the claimed 0.95 probability. Since Gallup is a professional, we expect he's got the poll size right, by using a more accurate numerical estimation formula – or he may have considered it legitimate to round a very slightly larger margin of error down to 0.04. ∎

3   Noisy Channel

Suppose we are transmitting packets of data across a noisy channel. Each packet has probability .01 of being lost. Now suppose we are transmitting $10,000$ packets. What is the probability that at most 2% of the packets are lost?

Solution. Sending data over the noisy channel is analogous to flipping $10,000$ coins where the probability of heads is $p = 0.01$ (in this case, a coin coming up heads is equivalent to the packet being dropped). We want to know what the likelihood is of greater than $\alpha = .2$ of all the coins coming up heads. However, in this case, we have $a > p$, and cannot use the equation we developed earlier. However, we can ask ourselves the question in terms of number of tails, where the probability of tails is $p = 0.99$, and we want to know the probability of a at most $\alpha 0.98$ of them coming up tails.

Plugging this in to our equation, we find that the probability is approximately

$$\left(\frac{1 - .98}{1 - .98/.99}\right) \frac{2^{10000(.98log(\frac{.98}{.00}) + .02log(\frac{.01}{.02}))}}{\sqrt{2\pi(.98)(1 - .98)10000}} < 2^{-60}$$

■