19.4 Estimation by Random Sampling

Democratic politicians were astonished in 2010 when their early polls of sample voters showed Republican Scott Brown was favored by a majority of voters and so would win the special election to fill the Senate seat that the late Democrat Teddy Kennedy had occupied for over 40 years. Based on their poll results, they mounted an intense, but ultimately unsuccessful, effort to save the seat for their party.

19.4.1 A Voter Poll

Suppose at some time before the election that p was the fraction of voters favoring Scott Brown. We want to estimate this unknown fraction p. Suppose we have some random process for selecting voters from registration lists that selects each voter with equal probability. We can define an indicator variable, K, by the rule that $K = 1$ if the random voter most prefers Brown, and $K = 0$ otherwise.

Now to estimate p, we take a large number, n, of random choices of voters[3] and count the fraction who favor Brown. That is, we define variables $K_1, K_2, \ldots$, where K_i is interpreted to be the indicator variable for the event that the ith chosen voter prefers Brown. Since our choices are made independently, the K_i's are independent. So formally, we model our estimation process by assuming we have mutually independent indicator variables $K_1, K_2, \ldots$, each with the same probability, p, of being equal to 1. Now let S_n be their sum, that is,

$$S_n ::= \sum_{i=1}^{n} K_i. \tag{19.16}$$

The variable S_n/n describes the fraction of sampled voters who favor Scott Brown. Most people intuitively, and correctly, expect this sample fraction to give a useful approximation to the unknown fraction, p.

So we will use the sample value, S_n/n, as our *statistical estimate* of p. We know that S_n has a binomial distribution with parameters n and p; we can choose n, but p is unknown.

How Large a Sample?

Suppose we want our estimate to be within 0.04 of the fraction, p, at least 95% of the time. This means we want

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \leq 0.04\right] \geq 0.95\,. \tag{19.17}$$

So we'd better determine the number, n, of times we must poll voters so that inequality (19.17) will hold. Chebyshev's Theorem offers a simple way to determine such a n.

S_n is binomially distributed. Equation (19.15), combined with the fact that $p(1-p)$ is maximized when $p = 1 - p$, that is, when $p = 1/2$ (check for yourself!),

[3] We're choosing a random voter n times *with replacement*. We don't remove a chosen voter from the set of voters eligible to be chosen later; so we might choose the same voter more than once! We would get a slightly better estimate if we required n *different* people to be chosen, but doing so complicates both the selection process and its analysis for little gain.

gives

$$\text{Var}[S_n] = n(p(1-p)) \leq n \cdot \frac{1}{4} = \frac{n}{4}. \tag{19.18}$$

Next, we bound the variance of S_n/n:

$$\begin{aligned}
\text{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \text{Var}[S_n] && \text{(Square Multiple Rule for Variance (19.9))} \\
&\leq \left(\frac{1}{n}\right)^2 \frac{n}{4} && \text{(by (19.18))} \\
&= \frac{1}{4n} && (19.19)
\end{aligned}$$

Using Chebyshev's bound and (19.19) we have:

$$\Pr\left[\left|\frac{S_n}{n} - p\right| \geq 0.04\right] \leq \frac{\text{Var}[S_n/n]}{(0.04)^2} \leq \frac{1}{4n(0.04)^2} = \frac{156.25}{n} \tag{19.20}$$

To make our our estimate with 95% confidence, we want the righthand side of (19.20) to be at most 1/20. So we choose n so that

$$\frac{156.25}{n} \leq \frac{1}{20},$$

that is,

$$n \geq 3,125.$$

Section 19.6.2 describes how to get tighter estimates of the tails of binomial distributions that lead to a bound on n that is about four times smaller than the one above. But working through this example using only the variance illustrates an approach to estimation that is applicable to arbitrary random variables, not just binomial variables.

19.4.2 Matching Birthdays

There are important cases where the relevant distributions are not binomial because the mutual independence properties of the voter preference example do not hold. In these cases, estimation methods based on Chebyshev's Theorem may be the best approach. Birthday Matching is an example. We already saw in Section 16.4 that in a class of 95 students, it is virtually certain that at least one pair of students will have the same birthday, which suggests that several pairs of students are likely to have the same birthday. How many matched birthdays should we expect?

As before, suppose there are n students and d days in the year, and let M be the number of pairs of students with matching birthdays. Now it will be easy to

calculate the expected number of pairs of students with matching birthdays. Then we can take the same approach as we did in estimating voter preferences to get an estimate of the probability of getting a number of pairs close to the expected number.

Unlike the situation with voter preferences, having matching birthdays for different pairs of students are not mutually independent events. Knowing Alice's birthday matches Bob's tells us nothing about who Carol matches, and knowing Alice has the same birthday as Carol tells us nothing about who Bob matches. But if Alice matches Bob and Alice matches Carol, it's certain that Bob and Carol match as well! The events that various pairs of students have matching birthdays are not mutually independent, and indeed not even three-way independent. The best we can say is that they are *pairwise* independent. This will allow us to apply the same reasoning to Birthday Matching as we did for voter preference. Namely, let $B_1, B_2, \ldots, B_n$ be the birthdays of n independently chosen people, and let $E_{i,j}$ be the indicator variable for the event that the ith and jth people chosen have the same birthdays, that is, the event $[B_i = B_j]$. So in our probability model, the B_i's are mutually independent variables, and the $E_{i,j}$'s are pairwise independent. Also, the expectations of $E_{i,j}$ for $i \neq j$ equals the probability that $B_i = B_j$, namely, $1/d$.

Now, M, the number of matching pairs of birthdays among the n choices, is simply the sum of the $E_{i,j}$'s:

$$M ::= \sum_{1 \leq i < j \leq n} E_{i,j}. \tag{19.21}$$

So by linearity of expectation

$$\mathrm{Ex}[M] = \mathrm{Ex}\left[\sum_{1 \leq i < j \leq n} E_{i,j}\right] = \sum_{1 \leq i < j \leq n} \mathrm{Ex}[E_{i,j}] = \binom{n}{2} \cdot \frac{1}{d}.$$

Similarly,

$$\begin{aligned}
\mathrm{Var}[M] &= \mathrm{Var}\left[\sum_{1 \leq i < j \leq n} E_{i,j}\right] \\
&= \sum_{1 \leq i < j \leq n} \mathrm{Var}[E_{i,j}] && \text{(Theorem 19.3.8)} \\
&= \binom{n}{2} \cdot \frac{1}{d}\left(1 - \frac{1}{d}\right). && \text{(Corollary 19.3.2)}
\end{aligned}$$

In particular, for a class of $n = 95$ students with $d = 365$ possible birthdays, we have $\text{Ex}[M] \approx 12.23$ and $\text{Var}[M] \approx 12.23(1-1/365) < 12.2$. So by Chebyshev's Theorem

$$\Pr[|M - \text{Ex}[M]| \geq x] < \frac{12.2}{x^2}.$$

Letting $x = 7$, we conclude that there is a better than 75% chance that in a class of 95 students, the number of pairs of students with the same birthday will be within 7 of 12.23, that is, between 6 and 19.

19.4.3 Pairwise Independent Sampling

The reasoning we used above to analyze voter polling and matching birthdays is very similar. We summarize it in slightly more general form with a basic result called the Pairwise Independent Sampling Theorem. In particular, we do not need to restrict ourselves to sums of zero-one valued variables, or to variables with the same distribution. For simplicity, we state the Theorem for pairwise independent variables with possibly different distributions but with the same mean and variance.

Theorem 19.4.1 (Pairwise Independent Sampling)**.** *Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and deviation, σ. Define*

$$S_n ::= \sum_{i=1}^{n} G_i. \tag{19.22}$$

Then

$$\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] \leq \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.$$

Proof. We observe first that the expectation of S_n/n is μ:

$$\begin{aligned}
\text{Ex}\left[\frac{S_n}{n}\right] &= \text{Ex}\left[\frac{\sum_{i=1}^{n} G_i}{n}\right] && \text{(def of } S_n\text{)} \\
&= \frac{\sum_{i=1}^{n} \text{Ex}[G_i]}{n} && \text{(linearity of expectation)} \\
&= \frac{\sum_{i=1}^{n} \mu}{n} \\
&= \frac{n\mu}{n} = \mu.
\end{aligned}$$

The second important property of S_n/n is that its variance is the variance of G_i

divided by n:

$$\begin{aligned}
\operatorname{Var}\left[\frac{S_n}{n}\right] &= \left(\frac{1}{n}\right)^2 \operatorname{Var}[S_n] && \text{(Square Multiple Rule for Variance (19.9))}\\
&= \frac{1}{n^2}\operatorname{Var}\left[\sum_{i=1}^{n} G_i\right] && \text{(def of } S_n\text{)}\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}[G_i] && \text{(pairwise independent additivity)}\\
&= \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. && (19.23)
\end{aligned}$$

This is enough to apply Chebyshev's Theorem and conclude:

$$\begin{aligned}
\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq x\right] &\leq \frac{\operatorname{Var}[S_n/n]}{x^2}. && \text{(Chebyshev's bound)}\\
&= \frac{\sigma^2/n}{x^2} && \text{(by (19.23))}\\
&= \frac{1}{n}\left(\frac{\sigma}{x}\right)^2.
\end{aligned}$$

■

The Pairwise Independent Sampling Theorem provides a quantitative general statement about how the average of independent samples of a random variable approaches the mean. In particular, it proves what is known as the Law of Large Numbers[4]: by choosing a large enough sample size, we can get arbitrarily accurate estimates of the mean with confidence arbitrarily close to 100%.

Corollary 19.4.2. *[Weak Law of Large Numbers] Let $G_1, \ldots, G_n$ be pairwise independent variables with the same mean, μ, and the same finite deviation, and let*

$$S_n ::= \frac{\sum_{i=1}^{n} G_i}{n}.$$

Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr[|S_n - \mu| \leq \epsilon] = 1.$$

[4]This is the *Weak* Law of Large Numbers. As you might suppose, there is also a Strong Law, but it's outside the scope of 6.042.

19.5 Confidence versus Probability

So Chebyshev's Bound implies that sampling 3,125 voters will yield a fraction that, 95% of the time, is within 0.04 of the actual fraction of the voting population who prefer Brown.

Notice that the actual size of the voting population was never considered because *it did not matter*. People who have not studied probability theory often insist that the population size should influence the sample size. But our analysis shows that polling a little over 3000 people people is always sufficient, regardless of whether there are ten thousand, or a million, or a billion voters. You should think about an intuitive explanation that might persuade someone who thinks population size matters.

Now suppose a pollster actually takes a sample of 3,125 random voters to estimate the fraction of voters who prefer Brown, and the pollster finds that 1250 of them prefer Brown. It's tempting, **but sloppy**, to say that this means:

False Claim. *With probability 0.95, the fraction, p, of voters who prefer Brown is* $1250/3125 \pm 0.04$. *Since* $1250/3125 - 0.04 > 1/3$*, there is a 95% chance that more than a third of the voters prefer Brown to all other candidates.*

What's objectionable about this statement is that it talks about the probability or "chance" that a real world fact is true, namely that the actual fraction, p, of voters favoring Brown is more than 1/3. But p is what it is, and it simply makes no sense to talk about the probability that it is something else. For example, suppose p is actually 0.3; then it's nonsense to ask about the probability that it is within 0.04 of 1250/3125. It simply isn't.

This example of voter preference is typical: we want to estimate a fixed, unknown real-world quantity. But *being unknown does not make this quantity a random variable*, so it makes no sense to talk about the probability that it has some property.

A more careful summary of what we have accomplished goes this way:

> We have described a probabilistic procedure for estimating the value of the actual fraction, p. The probability that *our estimation procedure* will yield a value within 0.04 of p is 0.95.

This is a bit of a mouthful, so special phrasing closer to the sloppy language is commonly used. The pollster would describe his conclusion by saying that

> At the 95% *confidence level*, the fraction of voters who prefer Brown is $1250/3125 \pm 0.04$.

So confidence levels refer to the results of estimation procedures for real-world quantities. The phrase "confidence level" should be heard as a reminder that some statistical procedure was used to obtain an estimate, and in judging the credibility of the estimate, it may be important to learn just what this procedure was.

6.042J / 18.062J Mathematics for Computer Science
Spring 2015