CS 70 Discrete Mathematics for CS
Spring 2007 Luca Trevisan

Lecture 23

I.I.D. Random Variables

Estimating the bias of a coin

Question: We want to estimate the proportion p of Democrats in the US population, by taking a small random sample. How large does our sample have to be to guarantee that our estimate will be within (say) 4% of the true value with probability at least 0.95?

This is perhaps the most basic statistical estimation problem, and shows up everywhere. We will develop a simple solution that uses only Chebyshev's inequality. More refined methods can be used to get sharper results.

Let's denote the size of our sample by n (to be determined), and the number of Democrats in it by the random variable S_n. (The subscript n just reminds us that the r.v. depends on the size of the sample.) Then our estimate will be the value $A_n = \frac{1}{n}S_n$.

Now as has often been the case, we will find it helpful to write $S_n = X_1 + X_2 + \cdots + X_n$, where

$$X_i = \begin{cases} 1 & \text{if person } i \text{ in sample is a Democrat;} \\ 0 & \text{otherwise.} \end{cases}$$

Note that each X_i can be viewed as a coin toss, with Heads probability p (though of course we do not know the value of p!). And the coin tosses are independent.[1]

What is the expectation of our estimate?

$$\mathrm{E}(A_n) = \mathrm{E}(\tfrac{1}{n}S_n) = \tfrac{1}{n}\mathrm{E}(X_1 + X_2 + \cdots + X_n) = \tfrac{1}{n} \times (np) = p.$$

So for any value of n, our estimate will always have the correct expectation p. [Such a r.v. is often called an *unbiased estimator* of p.] Now presumably, as we increase our sample size n, our estimate should get more and more accurate. This will show up in the fact that the *variance* decreases with n: i.e., as n increases, the probability that we are far from the mean p will get smaller.

To see this, we need to compute $\mathrm{Var}(A_n)$. And since $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, we need to figure out how to compute the variance of a *sum* of random variables.

Theorem 23.1: *For any random variable X and constant c, we have*

$$\mathrm{Var}(cX) = c^2\mathrm{Var}(X).$$

[1] We are assuming here that the sampling is done "with replacement"; i.e., we select each person in the sample from the entire population, including those we have already picked. So there is a small chance that we will pick the same person twice.

And for independent *random variables X, Y, we have*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof: From the definition of variance, we have

$$\mathrm{Var}(cX) = \mathrm{E}((cX - \mathrm{E}(cX))^2) = \mathrm{E}((cX - c\mathrm{E}(X))^2) = \mathrm{E}(c^2(X - \mathrm{E}(X))^2) = c^2\mathrm{Var}(X).$$

The proof of the second claim was an earlier exercise. Note that the second claim does **not** in general hold unless X and Y are independent. □

Using Theorem 23.1, we can now compute $\mathrm{Var}(A_n)$:

$$\mathrm{Var}(A_n) = \mathrm{Var}(\tfrac{1}{n}\sum_{i=1}^{n} X_i) = (\tfrac{1}{n})^2\mathrm{Var}(\sum_{i=1}^{n} X_i) = (\tfrac{1}{n})^2 \sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

where we have written σ^2 for the variance of each of the X_i. So we see that the variance of A_n decreases linearly with n. This fact ensures that, as we take larger and larger sample sizes n, the probability that we deviate much from the expectation p gets smaller and smaller.

Let's now use Chebyshev's inequality to figure out how large n has to be to ensure a specified accuracy in our estimate of the proportion of Democrats p. A natural way to measure this is for us to specify two parameters, ε and δ, both in the range $(0,1)$. The parameter ε controls the *error* we are prepared to tolerate in our estimate, and δ controls the *confidence* we want to have in our estimate. A more precise version of our original question is then the following:

Question: For the Democrat-estimation problem above, how large does the sample size n have to be in order to ensure that

$$\Pr[|A_n - p| \geq \varepsilon] \leq \delta \ ?$$

In our original question, we had $\varepsilon = 0.04$ and $\delta = 0.05$.

Let's apply Chebyshev's inequality to answer our more precise question above. Since we know $\mathrm{Var}(A_n)$, this will be quite simple. From Chebyshev's inequality, we have

$$\Pr[|A_n - p| \geq \varepsilon] \leq \frac{\mathrm{Var}(A_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

To make this less than the desired value δ, we need to set

$$n \geq \sigma^2 \times \frac{1}{\varepsilon^2\delta}. \tag{1}$$

Now recall that $\sigma^2 = \mathrm{Var}(X_i)$ is the variance of a single sample X_i. So, since X_i is a 0/1-valued r.v., we have $\sigma^2 = p(1-p)$, and inequality (1) becomes

$$n \geq (p - p^2) \times \frac{1}{\varepsilon^2\delta}. \tag{2}$$

Plugging in $\varepsilon = 0.04$ and $\delta = 0.05$, we see that a sample size of $n = 12,500 \cdot (p - p^2)$ is sufficient.

Unfortunately we do not know what p is, but we know that $p - p^2 \leq 1/4$ for every value of p, and so $n = 3125$ is sufficient.

Estimating a general expectation

What if we wanted to estimate something a little more complex than the proportion of Democrats in the population, such as the average wealth of people in the US? Then we could use exactly the same scheme as above, except that now the r.v. X_i is the wealth of the ith person in our sample. Clearly $\mathrm{E}(X_i) = \mu$, the average wealth (which is what we are trying to estimate). And our estimate will again be $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, for a suitably chosen sample size n. Once again the X_i are i.i.d. random variables, so we again have $\mathrm{E}(A_n) = \mu$ and $\mathrm{Var}(A_n) = \frac{\sigma^2}{n}$, where $\sigma^2 = \mathrm{Var}(X_i)$ is the variance of the X_i. (Recall that the only facts we used about the X_i was that they were independent and had the same distribution — actually the same expectation and variance would be enough: why?)

From equation (1), it is enough for the sample size n to satisfy

$$n \geq \frac{\sigma^2}{\mu^2} \times \frac{1}{\varepsilon^2 \delta}. \tag{3}$$

Here ε and δ are the desired error and confidence respectively, as before. Now of course we don't know the other two quantities, μ and σ^2, appearing in equation (3). In practice, we would use a lower bound on μ and an upper bound on σ^2 (just as we used a lower bound on p in the Democrats problem). Plugging these bounds into equation (3) will ensure that our sample size is large enough.

For example, in the average wealth problem we could probably safely take μ to be at least (say) \$50k (probably more). However, the existence of people such as Bill Gates means that we would need to take a very high value for the variance σ^2. Indeed, if there is at least one individual with wealth \$50 billion, then assuming a relatively small value of μ means that the variance must be at least about $\frac{(50\times 10^9)^2}{300\times 10^6} \approx 8 \cdot 10^{12}$. (Check this.) However, this individual's contribution to the mean is only $\frac{50\times 10^9}{300\times 10^6} \approx 170$. There is really no way around this problem with simple uniform sampling: the uneven distribution of wealth means that the variance is inherently very large, and we will need a huge number of samples before we are likely to find anybody who is immensely wealthy. But if we don't include such people in our sample, then our estimate will be way too low.

(The very high variance in wealth distribution also means that the average wealth is not very representative of "typical" wealth. For example, in 1999, the average household wealth in the US was approximately \$ 214,000, but the median wealth was approximately \$ 60,000.)

As a further example, suppose we are trying to estimate the average rate of emission from a radioactive source, and we are willing to assume that the emissions follow a Poisson distribution with some unknown parameter λ — of course, this λ is precisely the expectation we are trying to estimate. Now in this case we have $\mu = \lambda$ and also $\sigma^2 = \lambda$ (see the previous lecture). So $\frac{\sigma^2}{\mu^2} = 1$, whatever the value of λ. Thus in this case a sample size of just $n = \frac{1}{\varepsilon^2\delta}$ suffices.

The Law of Large Numbers

The estimation method we used in the previous two sections is based on a principle that we accept as part of everyday life: namely, the Law of Large Numbers. This asserts that, if we observe some random variable many times, and take the average of the observations, then this average will converge to a *single value*, which is of course the expectation of the random variable. In other words, averaging tends to smooth out any large fluctuations, and the more averaging we do the better the smoothing.

Theorem 23.2: **[Law of Large Numbers]** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathrm{E}(X_i)$. Define $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\alpha > 0$, we have*

$$\Pr[|A_n - \mu| \geq \alpha] \to 0 \qquad \text{as } n \to \infty.$$

We will not prove this theorem here. Notice that it says that the probability of *any* deviation α from the mean, however small, tends to zero as the number of observations n in our average tends to infinity. Thus by taking n large enough, we can make the probability of any given deviation as small as we like. Note, however, that the Law of Large Numbers does not say anything about *how large n* has to be to achieve a certain accuracy.

If our random variables have finite variance (which is always the case if the random variables only take a finite number of possible values) then we can give a simple proof of the law of large numbers using Chebyshev's inequality. The proof for the general case is considerably more difficult.

Theorem 23.3: **[Law of Large Numbers – Finite Variance Case]** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathrm{E}(X_i)$ and variance $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Define $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\alpha > 0$, we have*

$$\Pr[|A_n - \mu| \geq \alpha] \to 0 \qquad \text{as } n \to \infty.$$

Proof: For every n, we have $\mathrm{E}(A_n) = \mu$ and $\mathrm{Var}(A_n) = \frac{\sigma^2}{n}$, so

$$\Pr[|A_n - \mu| \geq \alpha =] \leq \frac{\mathrm{Var}(A_n)}{\alpha^2} = \frac{\sigma^2}{\alpha^2} \cdot \frac{1}{n} \to 0$$

□

When we have finite variance, however, we can say something much stronger than the Law of Large Numbers: namely, the distribution of the sample average A_n, for large enough n, looks like a *bell-shaped curve* centered about the mean μ. The width of this curve decreases with n, so it approaches a sharp spike at μ. This fact is known as the Central Limit Theorem.