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) 10% (in relative terms) 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 26.1: *For any random variable X and constant c, we have*

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

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

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

[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.

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 is left as an exercise (Problem 1 on Homework 11). 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

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

In our original question, we had $\varepsilon = 0.1$ and $\delta = 0.05$. Notice that ε measures the *relative* error, i.e., the error as a *ratio* of the target value p. This seems like a more reasonable convention than the *absolute* error (based on $\Pr[|A_n - p| \geq \varepsilon]$). This is because a given absolute error (say, ± 0.1) might be quite small in the context of measuring a large value like $p = 0.5$, but very large when measuring a small value like $p = 0.05$. In contrast, the relative error treats all values of p equally.

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

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

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

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

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

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

Plugging in $\varepsilon = 0.1$ and $\delta = 0.05$, we see that a sample size of $n = 2000 \times \frac{1-p}{p}$ is sufficient.

At this point you should be worried. Why? Because our formula for the sample size contains p, and this is precisely the quantity we are trying to estimate! But we can get around this as follows. We just pick a value p' that we know for sure is less than p. (For example, in the Democrats problem we could certainly take $p' = \frac{1}{3}$.) Then we use p' in place of p in equation (2). Since p' is less than p, this will always lead us to take at least enough samples (why?). In the Democrats example, with $p' = \frac{1}{3}$, this means we would take a sample size of $n = 2000 \times 2 = 4000$.

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) \$20k (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 $\frac{(50\times 10^9)^2}{250\times 10^6} = 10^{13}$. (Check this.) However, this individual's contribution to the mean is only $\frac{50\times 10^9}{250\times 10^6} = 200$. 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.

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 26.2: **[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. For that, we need Chebyshev's inequality or some other quantitative tool.]

Actually 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.

To say this precisely, we need to define the "bell-shaped curve." This is the so-called Normal distribution, and it is the first (and only) non-discrete distribution we will meet in this course. For random variables that take on continuous real values, it no longer makes sense to talk about $\Pr[X = a]$. As an example, consider a r.v. X that has the uniform distribution on the continuous interval $[0,1]$. Then for any single point $0 \le a \le 1$, we have $\Pr[X = a] = 0$. However, clearly it is the case that, for example, $\Pr[\frac{1}{4} \le X \le \frac{3}{4}] = \frac{1}{2}$. So in place of point probabilities $\Pr[X = a]$, we need a different notion of "distribution" for continuous random variables.

Definition 26.1 (density function): For a real-valued r.v. X, a real-valued function $f(x)$ is called a (probability) density function for X if

$$\Pr[X \le a] = \int_{-\infty}^{a} f(x)dx.$$

Thus we can think of $f(x)$ as defining a curve, such that the area under the curve between points $x = a$ and $x = b$ is precisely $\Pr[a \le X \le b]$. Note that we must always have $\int_{-\infty}^{\infty} f(x)dx = 1$. (Why?) As an example, for the uniform distribution on $[0,1]$ the density would be

$$f(x) = \begin{cases} 0 & \text{for } x < 0; \\ 1 & \text{for } 0 \le x \le 1; \\ 0 & \text{for } x > 1. \end{cases}$$

[Check you agree with this. What would be the density for the uniform distribution on $[-1,1]$?]

Expectations of continuous r.v.'s are computed in an analagous way to those for discrete r.v.'s. Thus

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)dx.$$

And also

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2, \quad \text{where } \mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx.$$

[You should check that, for the uniform distribution on $[0,1]$, the expectation is $\frac{1}{2}$ and the variance is $\frac{1}{12}$.]

Now we are in a position to define the Normal distribution.

Definition 26.2 (Normal distribution): The Normal distribution with mean μ and variance σ^2 is the distribution with density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}.$$

[The reason for the constant factor $\frac{1}{\sigma\sqrt{2\pi}}$ is that this is the value of the integral of the exponential function $\mathrm{e}^{-(x-\mu)^2/\sigma^2}$. So we have to normalize by this value to get $\int_{-\infty}^{\infty} f(x)dx = 1$. If you like calculus, you might like to do the integrals to check that the expectation $\int x f(x)dx$ is indeed μ and that the variance is indeed σ^2.]

If you plot the above density function $f(x)$, you will see that it is a symmetrical bell-shaped curve centered around the mean μ. Its height at the mean is about 0.4, and its width is determined by the variance σ^2, as follows: 50% of the mass is contained in the interval of width 0.67σ either side of the mean, and 99.7% in the interval of width 3σ either side of the mean. (Note that, to get the correct scale, deviations are on the order of σ rather than σ^2.)

Now we are in a position to state the Central Limit Theorem. Because our treatment of continuous distributions has been rather sketchy, we shall be content with a rather imprecise statement. This can be made completely rigorous without too much extra effort.

Theorem 26.3: **[Central Limit Theorem]** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathrm{E}(X_i)$ and variance $\sigma^2 = \mathrm{Var}(X_i)$ (both assumed to be $< \infty$). Define $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then as $n \to \infty$, the distribution of A_n approaches the Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.*

Note that the variance is $\frac{\sigma^2}{n}$ (as we would expect) so the width of the bell-shaped curve decreases by a factor of $\sqrt{n}$ as n increases.

The Central Limit Theorem is actually a very striking fact. What it says is the following. If we take an average of n observations of absolutely any r.v. X, then the distribution of that average will be a bell-shaped curve centered at $\mu = \mathrm{E}(X)$. Thus all trace of the distribution of X disappears as n gets large: all distributions, no matter how complex,[2] look like the Normal distribution when they are averaged. The only effect of the original distribution is through the variance σ^2, which determines the width of the curve for a given value of n, and hence the rate at which the curve shrinks to a spike.

In class, we saw how the distribution of A_n behaves for increasing values of n, when the X_i have the geometric distribution with parameter $\frac{1}{6}$. In problem 5 of HW11, you will do the same thing for several different r.v.'s X_i, all of which have the same mean but very different distributions. You should observe the appearance of the bell-shaped curve in all cases (with a width determined by the variance of the particular X_i).

[2]We do need to assume that the mean and variance of X are finite.