CS 70 Discrete Mathematics and Probability Theory
Summer 2014 James Cook **Note 17**

# 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) an additive factor of 0.1 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 random variable 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]. We call such a family of random variables *independent, identically distributed*, or *i.i.d.* for short.

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 as $n$ grows: i.e, the probability that we are far from the mean $p$ will get smaller.

To see this, we need to compute $\mathrm{Var}(A_n)$. But $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is just a multiple of a sum of *independent* random variables.

**Theorem 17.1**: For any random variable $X$ and constant $c$, we have

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

And for *independent* random variables $X$ and $Y$, 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.

Before we prove this theorem, let us look more carefully at something we have been using implicitly for some time:

**Joint Distributions** Consider two random variables $X$ and $Y$ defined on the same probability space. By linearity of expectation, we know that $\mathrm{E}(X+Y) = \mathrm{E}(X)+\mathrm{E}(Y)$. Since $\mathrm{E}(X)$ can be calculated if we know the distribution of $X$ and $\mathrm{E}(Y)$ can be calculated if we know the distribution of $Y$, this means that $\mathrm{E}(X+Y)$ can be computed knowing only the two individual distributions. No information is needed about the *relationship* between $X$ and $Y$. This is not true if we need to compute, say, $\mathrm{E}((X+Y)^2)$, e.g. as when we computed the variance of a binomial r.v. This is because $\mathrm{E}((X+Y)^2) = \mathrm{E}(X^2)+2\mathrm{E}(XY)+\mathrm{E}(Y^2)$, and $\mathrm{E}(XY)$ depends on the relationship between $X$ and $Y$. How can we capture such a relationship?

Recall that the distribution of a single random variable $X$ is the collection of the probabilities of all events $X = a$, for all possible values of $a$ that $X$ can take on. When we have two random variables $X$ and $Y$, we can think of $(X,Y)$ as a "two-dimensional" random variable, in which case the events of interest are $X = a \wedge Y = b$ for all possible values of $(a,b)$ that $(X,Y)$ can take on. Thus, a natural generalization of the notion of distribution to multiple random variables is the following.

**Definition 17.1 (joint distribution)**: The *joint distribution* of two discrete random variables $X$ and $Y$ is the collection of values $\{(a,b,\Pr[X = a \wedge Y = b]) : (a,b) \in \mathscr{A} \times \mathscr{B}\}$, where $\mathscr{A}$ and $\mathscr{B}$ are the sets of all possible values taken by $X$ and $Y$ respectively.

This notion obviously generalizes to three or more random variables. Since we will write $\Pr[X = a \wedge Y = b]$ quite often, we will abbreviate it to $\Pr[X = a, Y = b]$.

Just like the distribution of a single random variable, the joint distribution is *normalized*, i.e.

$$\sum_{a\in\mathscr{A},b\in\mathscr{B}} \Pr[X = a, Y = b] = 1.$$

This follows from noticing that the events $X = a \wedge Y = b$, $a \in \mathscr{A}, b \in \mathscr{B}$, partition the sample space.

The joint distribution of two random variables fully describe their statistical relationships, and provides enough information to compute any probability or expectation involving the two random variables. For example,

$$\mathrm{E}(XY) = \sum_c c \times \Pr[XY = c] = \sum_a \sum_b ab \times \Pr[X = a, Y = b].$$

More generally, if $f$ is any function on $\mathbf{R} \times \mathbf{R}$,

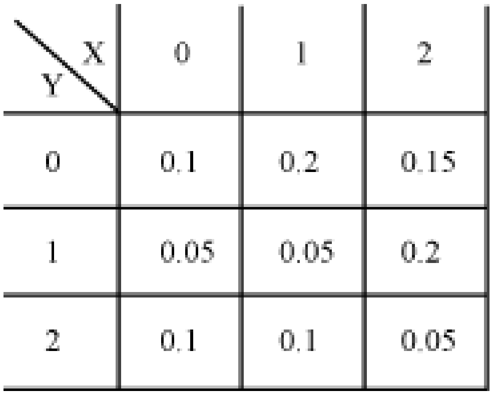
$$\mathrm{E}(f(X,Y)) = \sum_c c \times \Pr[f(X,Y) = c] = \sum_a \sum_b f(a,b) \times \Pr[X = a, Y = b].$$

Moreover, the individual distributions of $X$ and $Y$ can be recovered from the joint distribution as follows:

$$\begin{aligned} \Pr[X = a] &= \sum_{b\in\mathscr{B}} \Pr[X = a, Y = b] \qquad \forall a \in \mathscr{A}, && (1) \\ \Pr[Y = b] &= \sum_{a\in\mathscr{A}} \Pr[X = a, Y = b] \qquad \forall b \in \mathscr{B}. && (2) \end{aligned}$$

The first follows from the fact that the events $Y = b$, $b \in \mathscr{B}$, form a partition of the sample space $\Omega$, and so the events $X = a \wedge Y = b$, $b \in \mathscr{B}$ are disjoint and their union yields the event $X = a$. Similar logic applies to the second fact.

Pictorially, one can think of the joint distribution values as entries filling a table, with the columns indexed by the values that $X$ can take on and the rows indexed by the values $Y$ can take on (Figure 1). To get the



| Y \ X | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 0.1 | 0.2 | 0.15 |
| 1 | 0.05 | 0.05 | 0.2 |
| 2 | 0.1 | 0.1 | 0.05 |

Figure 1: A tabular representation of a joint distribution.

distribution of $X$, all one needs to do is to sum the entries in each of the columns. To get the distribution of $Y$, just sum the entries in each of the rows. This process is sometimes called *marginalization* and the individual distributions are sometimes called *marginal* distributions to differentiate them from the joint distribution.

**Independent Random Variables** Independence of random variables is defined in analogous fashion to independence for events:

**Definition 17.2 (independent r.vs)**: Random variables $X$ and $Y$ on the same probability space are said to be *independent* if the events $X = a$ and $Y = b$ are independent for all values $a, b$. Equivalently, the joint distribution of independent r.vs decomposes as

$$\Pr[X = a, Y = b] = \Pr[X = a]\Pr[Y = b] \quad \forall a, b.$$

Note that for independent r.vs, the joint distribution is fully specified by the marginal distributions.

Mutual independence of more than two r.vs is defined similarly. A very important example of independent r.vs is indicator r.vs for independent events. Thus, for example, if $\{X_i\}$ are indicator r.vs for the $i$th toss of a coin being Heads, then the $X_i$ are mutually independent r.vs.

We saw that the expectation of a sum of r.vs is the sum of the expectations of the individual r.vs. This is not true in general for variance. However, as the above theorem states, this is true if the random variables are independent. To see this, first we look at the expectation of a product of independent r.vs (which is a quantity that frequently shows up in variance calculations, as we have seen).

**Theorem 17.2**: For *independent* random variables $X$ and $Y$, we have $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

**Proof**: We have

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_a \sum_b ab \times \Pr[X = a, Y = b] \\
&= \sum_a \sum_b ab \times \Pr[X = a] \times \Pr[Y = b] \\
&= \left(\sum_a a \times \Pr[X = a]\right) \times \left(\sum_b b \times \Pr[Y = b]\right) \\
&= \mathrm{E}(X) \times \mathrm{E}(Y),
\end{aligned}$$

as required. In the second line here we made crucial use of independence. ■

For example, this theorem would have allowed us to conclude immediately in our random walk example at the beginning of Lecture Note 16 that $\mathrm{E}(X_i X_j) = \mathrm{E}(X_i)\mathrm{E}(X_j) = 0$, without the need for a calculation.

We now use the above theorem to conclude the nice property of the variance of independent random variables stated in the theorem above, namely that for independent random variables $X$ and $Y$, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$:

**Proof**: From the alternative formula for variance in Theorem 16.1, we have, using linearity of expectation extensively,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}((X+Y)^2) - \mathrm{E}(X+Y)^2 \\
&= \mathrm{E}(X^2) + \mathrm{E}(Y^2) + 2\mathrm{E}(XY) - (\mathrm{E}(X) + \mathrm{E}(Y))^2 \\
&= (\mathrm{E}(X^2) - \mathrm{E}(X)^2) + (\mathrm{E}(Y^2) - \mathrm{E}(Y)^2) + 2(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)).
\end{aligned}$$

Now *because X and Y are independent*, by Theorem 18.1 the final term in this expression is zero. Hence we get our result. ■

**Note:** The expression $\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ appearing in the above proof is called the *covariance* of $X$ and $Y$, and is a measure of the dependence between $X$ and $Y$. It is zero when $X$ and $Y$ are independent.

It is very important to remember that **neither** of these two results is true in general, without the assumption that $X$ and $Y$ are independent. As a simple example, note that even for a 0-1 r.v. $X$ with $\Pr[X = 1] = p$, $\mathrm{E}(X^2) = p$ is not equal to $\mathrm{E}(X)^2 = p^2$ (because of course $X$ and $X$ are not independent!).

Note also that the theorem does not quite say that variance is *linear* for independent random variables: it says only that variances sum. It is *not* true that $\mathrm{Var}(cX) = c\mathrm{Var}(X)$ for a constant $c$. It says that $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

The proof is left as a straightforward exercise.

We now return to our example of estimating the proportion of Democrats, where we were about to 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 $\sigma^2$ 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, $\varepsilon$ and $\delta$, both in the range $(0,1)$. The parameter $\varepsilon$ controls the *error* we are prepared to tolerate in our estimate, and $\delta$ 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 $\varepsilon = 0.1$ 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 $\delta$, we need to set

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

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 (3) becomes

$$n \geq \frac{p(1-p)}{\varepsilon^2 \delta}. \tag{4}$$

Since $p(1-p)$ is takes on its maximum value for $p = 1/2$, we can conclude that it is sufficient to choose $n$ such that:

$$n \geq \frac{1}{4\varepsilon^2 \delta}. \tag{5}$$

Plugging in $\varepsilon = 0.1$ and $\delta = 0.05$, we see that a sample size of $n = 500$ is sufficient. Notice that the size of the sample is independent of the total size of the population! This is how polls can accurately estimate quantities of interest for a population of several hundred million while sampling only a very small number of people.

## 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 $i$th 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?) This time, however, since we do not have any a priori bound on the mean $\mu$, it makes more sense to let $\varepsilon$ be the relative error. i.e. we wish to find an estimate that is within an additive error of $\varepsilon\mu$.

Using equation (3), but substituting $\varepsilon\mu$ in place of $\varepsilon$, it is enough for the sample size $n$ to satisfy

$$n \geq \frac{\sigma^2}{\mu^2} \times \frac{1}{\varepsilon^2 \delta}. \tag{6}$$

Here $\varepsilon$ and $\delta$ are the desired relative error and confidence respectively. Now of course we don't know the other two quantities, $\mu$ and $\sigma^2$, appearing in equation (6). In practice, we would use a lower bound on $\mu$ and an upper bound on $\sigma^2$ (just as we used a lower bound on $p$ in the Democrats problem). Plugging these bounds into equation (6) will ensure that our sample size is large enough.

For example, in the average wealth problem we could probably safely take $\mu$ 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 $\sigma^2$. Indeed, if there is at least one individual with wealth \$50 billion, then assuming a relatively small value of $\mu$ means that the variance must be at least about $\frac{(50\times 10^9)^2}{250\times 10^6} = 10^{13}$. (Check this.) 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 $\lambda$ — of course, this $\lambda$ 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 note). So $\frac{\sigma^2}{\mu^2} = \frac{1}{\lambda}$. Thus in this case a sample size of $n = \frac{1}{\lambda \varepsilon^2 \delta}$ suffices. (Again, in practice we would use a lower bound on $\lambda$.)

# 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 (LLN). 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 17.3**: **[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

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

**Proof**: Let $\mathrm{Var}(X_i) = \sigma^2$ be the common variance of the r.vs; we assume that $\sigma^2$ is finite[2]. With this (relatively mild) assumption, the LLN is an immediate consequence of Chebyshev's Inequality. For, as we have seen above, $\mathrm{E}(A_n) = \mu$ and $\mathrm{Var}(A_n) = \frac{\sigma^2}{n}$, so by Chebyshev we have

$$\Pr[|A_n - \mu| \geq \alpha] \leq \frac{\mathrm{Var}(A_n)}{\alpha^2} = \frac{\sigma^2}{n\alpha^2} \to 0 \qquad \text{as } n \to \infty.$$

This completes the proof. ■

Notice that the LLN says that the probability of *any* deviation $\alpha$ 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.

[2] If $\sigma^2$ is not finite, the LLN still holds but the proof is much trickier.