Chebyshev's Inequality

We have seen that, intuitively, the variance (or, more correctly the standard deviation) is a measure of "spread", or deviation from the mean. Our next goal is to make this intuition quantitatively precise. What we can show is the following:

Theorem 14.1: **[Chebyshev's Inequality]** For a random variable X with expectation $\mathrm{E}(X) = \mu$, and for any $\alpha > 0$,

$$\Pr[|X-\mu| \geq \alpha] \leq \frac{\mathrm{Var}(X)}{\alpha^2}.$$

Before proving Chebyshev's inequality, let's pause to consider what it says. It tells us that the probability of any given deviation, α, from the mean, either above it or below it (note the absolute value sign), is at most $\frac{\mathrm{Var}(X)}{\alpha^2}$. As expected, this deviation probability will be small if the variance is small. An immediate corollary of Chebyshev's inequality is the following:

Corollary 14.2: For a random variable X with expectation $\mathrm{E}(X) = \mu$, and standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$,

$$\Pr[|X-\mu| \geq \beta\sigma] \leq \frac{1}{\beta^2}.$$

Proof: Plug $\alpha = \beta\sigma$ into Chebyshev's inequality. □

So, for example, we see that the probability of deviating from the mean by more than (say) two standard deviations on either side is at most $\frac{1}{4}$. In this sense, the standard deviation is a good working definition of the "width" or "spread" of a distribution.

We should now go back and prove Chebyshev's inequality. The proof will make use of the following simpler bound, which applies only to *non-negative* random variables (i.e., r.v.'s which take only values ≥ 0).

Theorem 14.3: **[Markov's Inequality]** For a *non-negative* random variable X with expectation $\mathrm{E}(X) = \mu$, and any $\alpha > 0$,

$$\Pr[X \geq \alpha] \leq \frac{\mathrm{E}(X)}{\alpha}.$$

Proof: From the definition of expectation, we have

$$\begin{aligned}
\mathrm{E}(X) &= \textstyle\sum_a a \times \Pr[X = a] \\
&\geq \textstyle\sum_{a\geq\alpha} a \times \Pr[X = a] \\
&\geq \alpha \textstyle\sum_{a\geq\alpha} \Pr[X = a] \\
&= \alpha \Pr[X \geq \alpha].
\end{aligned}$$

The crucial step here is the second line, where we have used the fact that X takes on only non-negative values. (Why is this step not valid otherwise?) □

There is an intuitive way of understanding Markov's inequality through an analogy of a seesaw. Imagine that the distribution of a non-negative random variable X is resting on a fulcrum, $\mu = \mathrm{E}(X)$. We are trying to find an upper bound on the percentage of the distribution which lies beyond $k\mu$, i.e. $\Pr[X \geq k\mu]$. In other words, we seek to add as much weight m_2 as possible on the seesaw at $k\mu$ while minimizing the effect it has on the seesaw's balance. This weight will represent the upper bound we are searching for. To minimize the weight's effect, we must imagine that the weight of the distribution which lies beyond $k\mu$ is concentrated at exactly $k\mu$. However, to keep things stable and maximize the weight at $k\mu$, we must add another weight m_1 as far left to the fulcrum as we can so that m_2 is as large as it can be. The farthest we can go to the left is 0, since X is non-negative. Moreover, the two weights m_1 and m_2 must add up to 1, since they represent the area under the distribution curve:

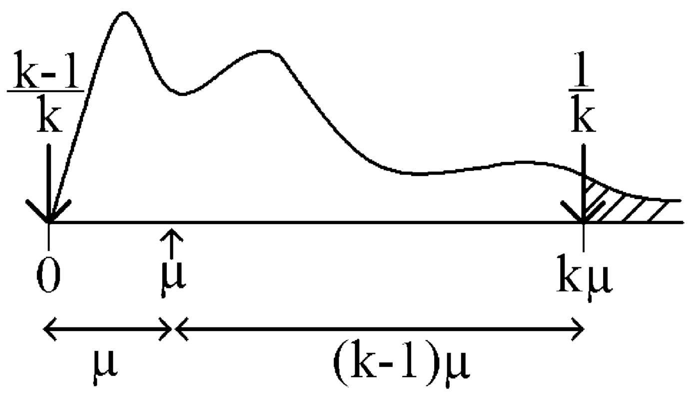

Figure 1: Markov's inequality interpreted as balancing a seesaw.

Since the lever arms are in the ratio $k-1$ to 1, a unit weight at $k\mu$ balances $k-1$ units of weight at 0. So the weights should be $\frac{k-1}{k}$ at 0 and $\frac{1}{k}$ at $k\mu$, which is exactly Markov's bound.

Now we can prove Chebyshev's inequality quite easily.

Proof of Theorem 16.2 Define the r.v. $Y = (X-\mu)^2$. Note that $\mathrm{E}(Y) = \mathrm{E}((X-\mu)^2) = \mathrm{Var}(X)$. Also, notice that the probability we are interested in, $\Pr[|X-\mu| \geq \alpha]$, is exactly the same as $\Pr[Y \geq \alpha^2]$. (Why?) Moreover, Y is obviously non-negative, so we can apply Markov's inequality to it to get

$$\Pr[Y \geq \alpha^2] \leq \frac{\mathrm{E}(Y)}{\alpha^2} = \frac{\mathrm{Var}(X)}{\alpha^2}.$$

This completes the proof. □

Examples

Here are some examples of applications of Chebyshev's inequality (you should check the algebra in them):

1. **Coin tosses.** Let X be the number of Heads in n tosses of a fair coin. The probability that X deviates from $\mu = \frac{n}{2}$ by more than $\sqrt{n}$ is at most $\frac{1}{4}$. The probability that it deviates by more than $5\sqrt{n}$ is at most $\frac{1}{100}$.

2. **Fixed points.** Let X be the number of fixed points in a random permutation of n items; recall that $\mathrm{E}(X) = \mathrm{Var}(X) = 1$. Thus the probability that more than (say) 10 students get their own homeworks after shuffling is at most $\frac{1}{100}$, however large n is.

In some special cases, including the coin tossing example above, it is possible to get much tighter bounds on the probability of deviations from the mean. However, for general random variables Chebyshev's inequality is essentially the only tool. Its power derives from the fact that it can be applied to *any* random variable.

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) and 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 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] 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 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 $\mathrm{Var}(A_n)$. But $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is just a multiple of a sum of *independent* random variables.

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

where we have written σ^2 for the variance of each of the X_i. So we see that <u>the variance of A_n decreases linearly with n</u>. 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

[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.

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

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

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 δ, we need to set

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

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

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

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:

$$n \geq \frac{1}{4\varepsilon^2\delta}. \tag{3}$$

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

$$\Pr[|A_n - \mu| \geq \varepsilon\mu] \leq \delta.$$

Using equation (1), but substituting $\varepsilon\mu$ in place of ε, it is enough for the sample size n to satisfy

$$n \geq \frac{\sigma^2}{\mu^2} \times \frac{1}{\varepsilon^2\delta}. \tag{4}$$

Here ε and δ are the desired relative error and confidence respectively. Now of course we don't know the other two quantities, μ and σ^2, appearing in equation (4). 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 (4) 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.) 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 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 14.4: **[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.v.'s; we assume that σ^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. $\square$

Notice that the LLN 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.

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