Concentration Inequalities and the Laws of Large Numbers

Suppose we have a biased coin, but we don't know what the bias is. To estimate the bias, we toss the coin n times and count how many Heads we observe. Then our estimate of the bias is given by $\hat{p} = \frac{1}{n}S_n$, where S_n is the number of Heads in the n tosses. Is this a good estimate? Let p denote the true bias of the coin, which is unknown to us. Since $\mathbb{E}[S_n] = np$, we see that the estimator $\hat{p}$ has the correct expected value: $\mathbb{E}[\hat{p}] = \frac{1}{n}\mathbb{E}[S_n] = p$. This means when n is sufficiently large, we can expect $\hat{p}$ to be very close to p; this is a manifestation of the *Law of Large Numbers*, which we shall see at the end of this note.

How large should n be to guarantee that our estimate $\hat{p}$ is within an error ε of the true bias p, i.e., $|\hat{p}-p| \leq \varepsilon$? The answer is that we can never guarantee with absolute certainty that $|\hat{p}-p| \leq \varepsilon$. This is because S_n is a random variable that can take any integer values between 0 and n, and thus $\hat{p} = \frac{1}{n}S_n$ is also a random variable that can take any values between 0 and 1. So regardless of the value of the true bias $p \in (0,1)$, it is possible that in our experiment of n coin tosses we observe n Heads (this happens with probability p^n), in which case $S_n = n$ and $\hat{p} = 1$. Similarly, it is also possible that all n coin tosses come up Tails (this happens with probability $(1-p)^n$), in which case $S_n = 0$ and $\hat{p} = 0$.

So instead of requiring that $|\hat{p}-p| \leq \varepsilon$ with absolute certainty, we relax our requirement and only demand that $|\hat{p}-p| \leq \varepsilon$ with *confidence* $1-\delta$, namely, $\mathbb{P}[|\hat{p}-p| \leq \varepsilon] \geq 1-\delta$. This means there is a small probability δ that we make an error of more than ε, but with high probability (at least $1-\delta$) our estimate is very close to p. Now we can state our result: to guarantee that $\mathbb{P}[|\hat{p}-p| \leq \varepsilon] \geq 1-\delta$, it suffices to toss the coin n times where

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

To prove such a result, we use a mathematical tool called *Chebyshev's Inequality*, which provides a quantitative, probabilistic bound on how far away a random variable can be from its expected value.

1 Markov's Inequality

Before discussing Chebyshev's inequality, we first prove the following simpler bound, which applies only to *nonnegative* random variables (i.e., r.v.'s which take only values ≥ 0).

Markov's inequality is intuitively similar to the notion that not everyone can score better than average. More precisely, at most half the people can score at least twice the average; when scores are non-negative.

Theorem 17.1 (**Markov's Inequality**). *For a* nonnegative *random variable X (i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$) with finite mean,*

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c},$$

for any positive constant c.

Proof. Let $\mathscr{A}$ denote the range of X and consider any constant $c \in \mathscr{A}$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{a \in \mathscr{A}} a \times \mathbb{P}[X = a] \\
&\geq \sum_{a \geq c} a \times \mathbb{P}[X = a] \\
&\geq \sum_{a \geq c} c \times \mathbb{P}[X = a] \\
&= c \sum_{a \geq c} \mathbb{P}[X = a] \\
&= c\,\mathbb{P}[X \geq c],
\end{aligned}$$

where the first line is just the definition of expectation, while the second line follows from the fact that X is a nonnegative random variable (i.e., all $a \in \mathscr{A}$ satisfies $a \geq 0$). Rearranging the last inequality gives us the desired result. □

Note in the proof above, we get $E[X] \geq c\mathbb{P}[X \geq c]$. Taking $c = 2E[x]$, the statement is that at most $1/2$ the people can be at least twice the average. The proof itself is analagous to this observation.

A slicker proof can be provided using the indicator function $I\{\cdot\}$, defined as

$$I\{\text{statement}\} = \begin{cases} 1, & \text{if statement is true,} \\ 0, & \text{if statement is false.} \end{cases}$$

Alternative proof of Theorem 17.1. Since X is a nonnegative random variable and $c > 0$, we have, for all $\omega \in \Omega$,

$$X(\omega) \geq c\,I\{X(\omega) \geq c\}, \tag{1}$$

since the right hand side is 0 if $X(\omega) < c$ and is c if $X(\omega) \geq c$. Multiplying both sides by $\mathbb{P}[\omega]$ and summing over $\omega \in \Omega$ gives

$$\mathbb{E}[X] \geq c\,\mathbb{E}[I\{X \geq c\}] = c\,\mathbb{P}[X \geq c],$$

where the first inequality follows from (1) and the fact that $\mathbb{P}[\omega] \geq 0$ for all $\omega \in \Omega$, while the last equality follows from the fact that $I\{X \geq c\}$ is an indicator random variable. □

If we have a random variable Y that is not necessarily nonnegative, then the same line of argument adopted above can be applied to prove the following result for the absolute value $|Y|$ of Y:

Theorem 17.2 (Generalized Markov's Inequality). *Let Y be an arbitrary random variable with finite mean. Then, for any positive constants c and r,*

$$\mathbb{P}[|Y| \geq c] \leq \frac{\mathbb{E}[|Y|^r]}{c^r}.$$

Proof. For $c > 0$ and $r > 0$, we have

$$|Y|^r \geq |Y|^r I\{|Y| \geq c\} \geq c^r I\{|Y| \geq c\}.$$

(Note that the last inequality would not hold if r were negative.) Taking expectations of both sides gives

$$\mathbb{E}[|Y|^r] \geq c^r\,\mathbb{E}[I\{|Y| \geq c\}] = c^r\,\mathbb{P}[|Y| \geq c],$$

and dividing by c^r leads to the desired result. □

There is an intuitive (leveraging your physical intuition) way to understand Markov's inequality through an analogy of balancing a seesaw, illustrated in Figure 1. Imagine that the probability distribution of a nonnegative random variable X is resting on a fulcrum at $\mu = \mathbb{E}[X]$. We are trying to find an upper bound on the percentage of the distribution which lies beyond $k\mu$, i.e., $\mathbb{P}[X \geq k\mu]$. In other words, we seek to add as much mass m_2 as possible on the seesaw at $k\mu$ while minimizing the effect it has on the seesaw's balance. This mass will represent the upper bound we are searching for. To minimize the mass's effect, we must imagine that the mass of the distribution which lies beyond $k\mu$ is concentrated at exactly $k\mu$. However, to keep things stable and maximize the mass at $k\mu$, we must add another mass 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 assumed to be nonnegative. Moreover, the two masss m_1 and m_2 must add up to 1, since they represent the area under the distribution curve. Since the lever arms are in the ratio $k-1$ to 1, a unit mass at $k\mu$ balances $k-1$ units of mass at 0. So the masses should be $\frac{k-1}{k}$ at 0 and $\frac{1}{k}$ at $k\mu$, which is exactly Markov's bound with $\alpha = k\mu$.

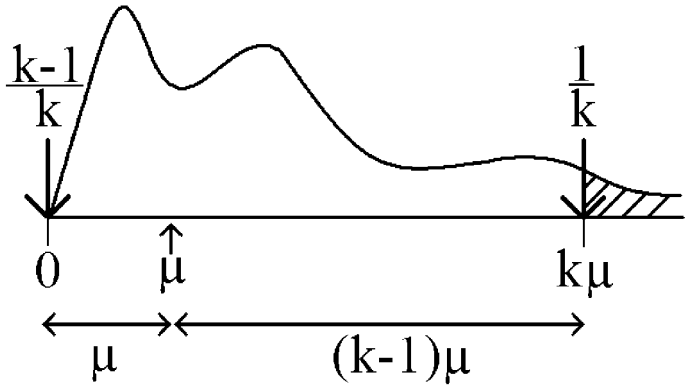

Figure 1: Markov's inequality interpreted as balancing a seesaw.

Example: Coin Tosses

Consider tossing a fair coin n times and let X denote the number of heads observed. What is the probability of observing more than $\frac{3}{4}n$ heads? Let's apply Markov inequality to obtain an upper bound on $\mathbb{P}[X \geq \frac{3}{4}n]$. Since $X \sim \text{Binomial}(n, \frac{1}{2})$, we have $\mathbb{E}[X] = \frac{1}{2}n$, so

$$\mathbb{P}[X \geq \tfrac{3}{4}n] \leq \frac{\mathbb{E}[X]}{\frac{3}{4}n} = \frac{2}{3},$$

which does not depend on n. Is this a good upper bound? The exact answer for $\mathbb{P}[X \geq \frac{3}{4}n]$ is given by

$$\mathbb{P}[X \geq \tfrac{3}{4}n] = \sum_{k=\lceil \frac{3}{4}n \rceil}^{n} \binom{n}{k} \frac{1}{2^n},$$

which is $\approx 5.5 \times 10^{-2}$ for $n = 10$ and $\approx 2.8 \times 10^{-7}$ for $n = 100$. As these numerical values show, $\mathbb{P}[X \geq \frac{3}{4}n]$ is a decreasing function of n, and so Markov's inequality does not provide a very good upper bound for this example. In the next section, we will see how to obtain an improved bound.

2 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. We can now make this intuition quantitatively precise:

Theorem 17.3 (Chebyshev's Inequality). *For a random variable X with finite expectation $\mathbb{E}[X] = \mu$,*

$$\mathbb{P}[|X - \mu| \geq c] \leq \frac{\text{Var}(X)}{c^2}, \tag{2}$$

and for any positive constant c.

Proof. Define $Y = (X - \mu)^2$ and note that $\mathbb{E}[Y] = \mathbb{E}\left[(X-\mu)^2\right] = \text{Var}(X)$. Also, notice that the event that we are interested in, $|X - \mu| \geq c$, is exactly the same as the event $Y = (X-\mu)^2 \geq c^2$. Therefore, $\mathbb{P}[|X - \mu| \geq c] = \mathbb{P}[Y \geq c^2]$. Moreover, Y is obviously nonnegative, so we can apply Markov's inequality in Theorem 17.1 to get

$$\mathbb{P}[|X - \mu| \geq c] = \mathbb{P}[Y \geq c^2] \leq \frac{\mathbb{E}[Y]}{c^2} = \frac{\text{Var}(X)}{c^2}.$$

This completes the proof. □

Here is a simpler proof using Generalized Markov's inequality from Theorem 17.2:

Alternative proof of Theorem 17.3. Define $Y = X - \mu$. Then, since $|Y|^2 = |X - \mu|^2 = (X-\mu)^2$, we have $\mathbb{E}\left[|Y|^2\right] = \mathbb{E}\left[(X-\mu)^2\right] = \text{Var}(X)$. Hence, (2) follows from applying Theorem 17.2 to $Y = X - \mu$ with $r = 2$. □

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

Corollary 17.1. *For any random variable X with finite expectation $\mathbb{E}[X] = \mu$ and finite standard deviation $\sigma = \sqrt{\text{Var}(X)}$,*

$$\mathbb{P}[\,|X - \mu| \geq k\sigma] \leq \frac{1}{k^2},$$

for any constant $k > 0$.

Proof. Plug $c = k\sigma$ 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.

In some special cases, it is possible to get tighter bounds on the probability of deviations from the mean. However, for general random variables that only have means and variances, Chebyshev's inequality is essentially the only tool. Its power derives from the fact that it can be applied to *any* random variable with a mean and a variance.

Example: Coin Tosses Revisited

Let's revisit the coin toss example considered above and apply Chebyshev's inequality to obtain an upper bound on $\mathbb{P}[X \geq \frac{3}{4}n]$. Recalling $\mathbb{E}[X] = \frac{n}{2}$, we obtain

$$\mathbb{P}[X \geq \tfrac{3}{4}n] = \mathbb{P}[X - \tfrac{n}{2} \geq \tfrac{n}{4}] \leq \mathbb{P}[|X - \tfrac{n}{2}| \geq \tfrac{n}{4}] \leq \frac{\mathrm{Var}(X)}{\left(\frac{n}{4}\right)^2}.$$

Since $X \sim \text{Binomial}(n, \frac{1}{2})$, we have $\mathrm{Var}(X) = n\frac{1}{2}(1-\frac{1}{2}) = \frac{n}{4}$, so we obtain

$$\mathbb{P}[X \geq \tfrac{3}{4}n] \leq \frac{\mathrm{Var}(X)}{\left(\frac{n}{4}\right)^2} = \frac{4}{n},$$

which is much better than the constant bound of $\frac{2}{3}$ given by Markov's inequality.

3 Applications

In this section, we discuss applications of Chebyshev's inequality to estimation problems.

3.1 Estimating the Bias of a Coin

Let us go back to our motivating example of estimating the bias of a coin. Recall that we have a coin of unknown bias p, and our estimate of p is $\hat{p} = \frac{1}{n}S_n$ where S_n is the number of Heads in n coin tosses.

As usual, we will find it helpful to write $S_n = X_1 + \cdots + X_n$, where $X_i = 1$ if the i-th coin toss comes up Heads and $X_i = 0$ otherwise, and the random variables $X_1, \ldots, X_n$ are independent and identically distributed. Then $\mathbb{E}[X_i] = \mathbb{P}[X_i = 1] = p$, so by linearity of expectation,

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\tfrac{1}{n}S_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p.$$

What about the variance of $\hat{p}$? Note that since the X_i's are independent, the variance of $S_n = \sum_{i=1}^{n} X_i$ is equal to the sum of the variances:

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

where we have written σ^2 for the variance of each of the X_i.

So we see that the variance of $\hat{p}$ 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 true bias p. As we discussed in the beginning of this note, 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.

Applying Chebyshev's inequality, we have

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

To make the right hand side less than the desired value δ, we need to set

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

Now recall that $\sigma^2 = \text{Var}(X_i)$ is the variance of a single sample X_i. So, since X_i is an indicator random variable with $\mathbb{P}[X_i = 1] = p$, we have $\sigma^2 = p(1-p)$, and inequality (3) becomes $n \geq \frac{p(1-p)}{\varepsilon^2 \delta}$. Since $p(1-p)$ is maximized[1] when $p = 1/2$, we can conclude that it is sufficient to choose n such that:

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

as we claimed earlier.

For example, plugging in $\varepsilon = 0.1$ and $\delta = 0.05$, we see that a sample size of $n = 500$ is sufficient to get an estimate $\hat{p}$ that is accurate to within an error of 0.1 with probability at least 95%.

As a concrete example, consider the problem of estimating the proportion p of Democrats in the US population, by taking a small random sample. We can model this as the problem of estimating the bias of a coin above, where each coin toss corresponds to a person that we select randomly from the entire population. And the coin tosses are independent.[2] Our calculation above shows that to get an estimate $\hat{p}$ that is accurate to within an error of 0.1 with probability at least 95%, it suffices to sample $n = 500$ people. In particular, 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.

3.2 Estimating a General Expectation

What if we wanted to estimate something a little more complex than the bias of a coin? For example, suppose we want to estimate the average wealth of people in the US. We can model this as the problem of estimating the expected value of an unknown probability distribution. Then we can use exactly the same scheme as before, except that now we sample the random variables $X_1, X_2, \ldots, X_n$ independently from our unknown distribution. Clearly $\mathbb{E}[X_i] = \mu$, the expected value that we are trying to estimate. Our estimate of μ will be $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$, for a suitably chosen sample size n.

Following the same calculation as before, we have $\mathbb{E}[\hat{\mu}] = \mu$ and $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$, where $\sigma^2 = \text{Var}(X_i)$ 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 $\hat{\mu}$ that is within an additive error of $\varepsilon\mu$:

$$\mathbb{P}[|\hat{\mu} - \mu| \geq \varepsilon\mu] \leq \delta.$$

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

$$n \geq \frac{\sigma^2}{\mu^2} \times \frac{1}{\varepsilon^2 \delta}. \tag{5}$$

Here ε and $1-\delta$ are the desired relative error and confidence level respectively. Now of course we don't know the other two quantities, μ and σ^2, appearing in equation (5). In practice, we would use a lower bound

[1] Use calculus to prove this.

[2] 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.

on μ and an upper bound on σ^2 (just as we used an upper bound on $p(1-p)$ in the coin tossing problem). Plugging these bounds into equation (5) 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) \$20,000 (probably more). However, the existence of very wealthy 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 in a population of size 325 million, then assuming a relatively small value of μ means that the variance must be at least about $\frac{(50\times 10^9)^2}{325\times 10^6} \approx 7.7\times 10^{12}$. 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.

4 The Law of Large Numbers

The estimation method we used in the previous 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 17.4 (Law of Large Numbers). *Let $X_1, X_2, \ldots$, be a sequence of i.i.d. (independent and identically distributed) random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ for all i. Then, their partial sums $S_n = X_1 + X_2 + \cdots + X_n$ satisfy*

$$\mathbb{P}\left[\left|\tfrac{1}{n}S_n - \mu\right| < \varepsilon\right] \to 1 \qquad \text{as } n \to \infty,$$

for every $\varepsilon > 0$, however small.

Proof. Let $\mathrm{Var}(X_i) = \sigma^2$ be the common variance of the r.v.'s; we assume that σ^2 is finite.[3] With this (relatively mild) assumption, the LLN is an immediate consequence of Theorem 17.3. Since $X_1, X_2, \ldots$ are i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, we have $\mathbb{E}\left[\frac{1}{n}S_n\right] = \mu$ and $\mathrm{Var}(\frac{1}{n}S_n) = \frac{\sigma^2}{n}$, so by Chebyshev's inequality we have

$$\mathbb{P}\left[\left|\tfrac{1}{n}S_n - \mu\right| \geq \varepsilon\right] \leq \frac{\mathrm{Var}(\frac{1}{n}S_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \text{as } n \to \infty.$$

Hence, $\mathbb{P}\left[\left|\frac{1}{n}S_n - \mu\right| < \varepsilon\right] = 1 - \mathbb{P}\left[\left|\frac{1}{n}S_n - \mu\right| \geq \varepsilon\right] \to 1$ as $n \to \infty$. □

Notice that the LLN says that the probability of *any* deviation ε 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.

[3] If σ^2 is not finite, the LLN still holds but the proof requires more advanced mathematical tools.