Chebyshev's inequality: Polling Application

In the last lecture, we covered Chebyshev's inequality:

Theorem 10.1: **[Chebyshev's Inequality]** *For a random variable X with expectation $\mathbb{E}[X] = \mu$, and for any $a > 0$,*

$$\mathbf{P}(|X-\mu| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

Let us now continue the polling application. Suppose we want to figure out the preference of a population. For example, we want to know the fraction p of the number of democrats in California. How many people should we ask before we get a reliable answer? Let X_i model the answer of a person, where $X_i = 1$ if the person is a democrat and 0 otherwise. Then $X_i = 1$ with probability p.

We want to estimate the parameter p in the model from observing the outcome of the random experiment, which is a standard problem in statistics. An estimator takes the data collected from polling and outputs an estimate $\hat{p}$ of the parameter p. The data in this problem is the n responses $X_1, \ldots, X_n$ of the people. We estimate $\hat{p}$ from the n observations, so $\hat{p}$ will be a function of the data. A reasonable estimate here is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

the fraction of the people polled who say they are democrats. How reliable is this estimate? p is not a random variable, but $\hat{p}$ is a random variable because it is a function of the X_is which are random variables (our data). We start by computing the expectation of $\hat{p}$:

$$\mathbb{E}[\hat{p}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{np}{n} = p$$

On the average, our estimator will give p, which is a desirable property. When $\mathbb{E}[\hat{p}] = p$, we say our estimator is unbiased.

How many samples do we need to collect until we get a reliable estimate for p? We first need to define what we mean by reliable estimator. A good estimator $\hat{p}$ will be close to p with high probability. For example, we would like an interval of width 0.1 around $\hat{p}$ to contain p with 95% probability. More specifically, we want the interval around $\hat{p}$ to satisfy $\mathbf{P}(|\hat{p}-p| > 0.1) \leq 0.05$. If this property holds, then $[\hat{p}-0.1, \hat{p}+0.1]$ is said to be the 95% confidence interval.

We connect this probability and the variance by using Chebyshev's inequality:

$$\mathbf{P}(|\hat{p}-p| > 0.1) \leq \frac{\text{Var}(\hat{p})}{0.1^2}.$$

We would like the right hand side to be ≤ 0.05. The variance of $\hat{p}$ can be written as

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

where we have used the variance of a binomial distribution which we previously computed ($\sum_i X_i \sim Bin(n,p)$). Therefore, we need

$$\frac{\frac{p(1-p)}{n}}{0.1^2} \leq 0.05,$$

or equivalently:

$$n \geq \frac{p(1-p)}{0.05 \times (0.1)^2}. \tag{1}$$

If we knew p, then (1) would tell us how many people we need to poll. But the problem is that we don't know p. Observe though that the maximum value of $p(1-p)$ is 0.25 (achieved at $p = 0.5$). So if we are a bit conservative and choose n such that

$$n \geq \frac{0.25}{0.05 \times (0.1)^2} = 500,$$

then the inequality (1) will be satisfied regardless of the true value of p.

Independent Random Variables and the Law of Large Numbers

The implicit assumption in our question "How many people should we ask before we get a reliable answer?" is that as we poll more and more people our answer becomes more reliable. Why would that be true? Will the reliability keep improving as we poll more people? This point is clear when we look at the variance of $\hat{p}$: the variance decreases when n increases. As the amount of data increases, the variance shrinks by $\frac{1}{n}$. The distribution of $\hat{p}$ is shrinking closer and closer around the mean. Reliability increases because the variance is decreasing.

We seek to understand this point better: Why does the variance decrease?

Suppose we have two random variables X and Y, and we want to compute $\mathbb{E}[X+Y]$. By linearity of expectation, we know $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This says that the relation between the two random variables does not affect the expectation of their sum. Remember the problem of randomly handing out homeworks to students. Suppose X indicates when the first student gets his homework, and Y indicates when the second student gets his homework. Then whether or not the first student gets his homework back affects the outcome for the second student. Therefore there is obviously a dependence between X and Y. Nevertheless it is still true that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

On the other hand, the variance of $X+Y$ is affected by the dependence between the variables. In general,

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

We consider the following example to illustrate this point.

Suppose X takes values 1 and -1 each with probability 1/2, and that Y has the same distribution. Given the distributions, we can compute $\mathbb{E}[X+Y]$ which is 0 because each expectation is 0. However, the individual distributions of the random variables do not f specify the relationship between the random variables. We will compare 3 cases:

1) Suppose we have the relationship $X = 1 \implies Y = 1$, and $X = -1 \implies Y = -1$. Then the above table summarizes the possible events, where each entry in the table is the probability of the intersection of events, i.e $\mathbf{P}(X = a, Y = b)$ for $a, b \in \{-1, 1\}$:

Table 1:

	X=1	X=-1
Y=1	0.5	0
Y=-1	0	0.5

If $Z = X+Y$, then $Z = 2$ with probability $\frac{1}{2}$, and $Z = -2$ with probability $\frac{1}{2}$. Therefore, the variance of Z can be computed to be $\text{Var}(Z) = 4$. In this case, the two variables reinforce each other so we get a distribution with a bigger variance.

2) Suppose we have the relationship $X = 1 \implies Y = -1$, and $X = -1 \implies Y = 1$. Then the probabilities of the possible events becomes:

Table 2:

	X=1	X=-1
Y=1	0	0.5
Y=-1	0.5	0

When compared to the first case, the distribution of X or Y *individually* is the same, but the relationship between them changed. The pmf of Z in this case is : $\mathbf{P}(Z = 0) = 1$. Therefore $\text{Var}(Z) = 0$. Notice that here the two variables cancel each other out.

Thus far we have considered two cases, one with a positive correlation between X and Y and one with a negative correlation. We will now define the notion of two random variables being independent. We know what it means for two events to be independent. How do we define the notion of two random variables being independent? A random variable is described by a collection of events. The events connected to a random variable X are all the events $\{X = a\}$ for all the possible values a that X can take on We want to define a relationship between X and Y which is a generalization of the independence of events. Any event associated with Y should be independent of any event in X.

Definition 10.1 (Independent Random Variables): X and Y are independent if the events $\{X = a\}$ and $\{Y = b\}$ are independent for all a, b.

3) Let us now go back to our binary example and consider the case when X and Y are independent, We can compute the 2 by 2 table using the fact that X and Y are independent since we can write the probabilities of the intersections as products: $\mathbf{P}(X = +1, Y = -1) = \mathbf{P}(X = +1)\mathbf{P}(Y = -1)$

What is the distribution of Z in this case? $Z = -2$ with probability 0.25 (when X and Y are -1), $Z = +2$ with probability 0.25, and $Z = 0$ with probability 0.5. This case falls in between the cases 1 and 2: $\text{Var}(X+Y) = 0.5 \times 4 = 2 = \text{Var}(X) + \text{Var}(Y)$ (only the points -2 and 2 contribute to the variance).

Table 3:

	X=1	X=-1
Y=1	0.25	0.25
Y=-1	0.25	0.25

Observe for this example that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. This turns out to be true whenever the random variables X and Y are independent. We will prove this fact at the end of the lecture.

The polling example was a special case of this result applied to the binomial distribution. In that example we have $\mathrm{Var}(\sum_i X_i) = np(1-p)$, where $p(1-p)$ is the variance of each of the X_i. So in the polling example, the variance of the sum is the sum of the variances. This is key to the fact that we get increasing reliability as n increases. If the X_is are all strongly positively correlated (for example they survey people that are related to each other), the results are not independent in this case. All the probability will be around $\sum_i X_i = n$ and $\sum_i X_i = 0$. So the variance of $\sum X_i$ will be of the order of n^2. We are interested in $\frac{1}{n^2}\mathrm{Var}(\sum_i X_i)$, which will not go to zero with n.

Now assuming that the variance of a sum of independent variables is equal to the sum of the variances, we will state one of the most important theorems in probability which is the law of large numbers:

Theorem 10.2: **[Law of Large Numbers]** *Let $X_1, X_2, \ldots, X_n$ be random variables each having the same distribution with the common expectation $\mu = \mathbb{E}[X_i]$. Suppose every pair of the random variables is independent. Define $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $a > 0$, we have*

$$\mathbf{P}[|A_n - \mu| \geq a] \to 0 \qquad \text{as } n \to \infty.$$

Proof: Let $\mathrm{Var}(X_i) = \sigma^2$ be the common variance of the r.v.'s; we assume that σ^2 is finite[1]. With this (relatively mild) assumption, the law of large numbers is an immediate consequence of Chebyshev's Inequality. For, as we have seen above, $\mathbb{E}[A_n] = \mu$ and $\mathrm{Var}(A_n) = \frac{\sigma^2}{n}$, so by Chebyshev we have

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

This completes the proof. □

Now let us go back and prove useful facts about the expectation and variance of independent random variables:

1- If X and Y are independent, then $E(XY) = E(X)E(Y)$.

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

Proof:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{\omega\in\Omega}\mathbf{P}(\omega)X(\omega)Y(\omega)\\ &= \sum_{a,b}\mathbf{P}(X=a,Y=b)ab \qquad \text{(Why?)}\\ &= \sum_{a,b}\mathbf{P}(X=a)\mathbf{P}(Y=b)ab\\ &= \sum_{a}\mathbf{P}(X=a)a\cdot\sum_{b}\mathbf{P}(X=b)b\\ &= \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

□

2- If X and Y are independent, then $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$.

Proof: Let μ_X and μ_Y be the means of X and Y respectively.

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y-\mu_X-\mu_Y)^2]\\ &= \mathbb{E}[(X-\mu_X+Y-\mu_Y)^2]\\ &= \mathbb{E}[(X-\mu_X)^2]+\mathbb{E}[(Y-\mu_Y)^2]-2\mathbb{E}[(X-\mu_X)(Y-\mu+y)]\\ &= \mathbb{E}[(X-\mu_X)^2]+\mathbb{E}[(Y-\mu_Y)^2]-2\mathbb{E}[X-\mu_X]\mathbb{E}[Y-\mu_Y]\\ &= \mathrm{Var}(X)+\mathrm{Var}(Y) \text{ where the cross term is 0.}\end{aligned}$$

□

The variance of a sum has a cross term, but when the variables are independent then the cross term is zero. The above calculation can be generalized to show that

$$\mathrm{Var}(\sum_i X_i)=\sum_i \mathrm{Var}(X_i)$$

whenever every pair of the random variables are independent.