Written by Ryan O'Donnell

1.3 The proof

We will not prove Chernoff Bound 2. However the proof is not much more than an elaboration on our proof of Chernoff Bound 1. Again, the idea is to let $\lambda > 0$ be a small "scale", pretty close to ϵ, actually. You then consider the random variable $(1+\lambda)^X$, or relatedly, $e^{\lambda X}$. Finally, you bound, e.g.,

$$\mathbf{Pr}[X \geq (1+\epsilon)\mu] = \mathbf{Pr}[e^{\lambda X} \geq e^{(1+\epsilon)\mu}] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{(1+\epsilon)\mu}},$$

using Markov's Inequality, and you can compute $\mathbf{E}[e^{\lambda X}]$ as

$$\mathbf{E}[e^{\lambda X}] = \mathbf{E}[e^{\lambda X_1 + \cdots + \lambda X_n}] = \mathbf{E}[e^{\lambda X_1} \cdots e^{\lambda X_n}] = \mathbf{E}[e^{\lambda X_1}] \cdots \mathbf{E}[e^{\lambda X_n}],$$

using the fact that $X_1, \ldots, X_n$ are independent.

2 Sampling

The #1 use of the Chernoff Bound is probably in Sampling/Polling. Suppose you want to know what fraction of the population approves of the current president. What do you do?

Well, you do a poll. Roughly speaking, you call up n random people and ask them if they approve of the president. Then you take this empirical fraction of people and claim that's a good estimate of the true fraction of the entire population that approves of the president. But is it a good estimate? And how big should n be?

Actually, there are *two* sources of error in this process. There's the probability that you obtain a "good" estimate. And there's the extent to which your estimate is "good". Take a close look at all those poll results that come out these days and you'll find that they use phrases like,

"This poll is accurate to within $\pm 2\%$*, 19 times out of* 20*."*

What this means is that they did a poll, they published an estimate of the true fraction of people supporting the president, and they make the following claim about their estimate: There is a 1/20 chance that their estimate is just completely way off. Otherwise, i.e., with probability $19/20 = 95\%$, their estimate is within $\pm 2\%$ of the truth.

This whole 95% thing is called the "confidence" of the estimate, and its presence is inevitable. There's just no way you can legitimately say, "My polling estimate is 100% guaranteed to be within $\pm 2\%$ of the truth." Because if you sample n people at random, you know, there's a chance they all happen to live in Massachusetts, say (albeit an unlikely, much-less-than-5% chance), in which case your approval rating estimate for a Democratic president is going to much higher than the overall country-wide truth.

To borrow a phrase from Learning Theory, these polling numbers are "Probably Approximately Correct" — i.e., probably (with chance at least 95% over the choice of people), the empirical average is approximately (within $\pm 2\%$, say) correct (vis-a-vis the true fraction of the population).

2.1 Analysis

How do pollsters, and how can we, make such statements?

Let the true fraction of the population that approves of the president be p, a number in the range $0 \leq p \leq 1$. This is the "correct answer" that we are trying to elicit.

Suppose we ask n uniformly randomly chosen people for their opinion, and let each person be chosen *independently*. We are choosing people "*with* replacement". (I.e., it's possible, albeit a very slim chance, that we may ask the same person more than once.) Let X_i be the indicator random variable that the ith person we ask approves of the president. Here is the key observation:

Fact: $X_i \sim \text{Bernoulli}(p)$, and $X_1, \ldots, X_n$ are independent.

Let $X = X_1 + \cdots + X_n$, and let $\overline{X} = X/n$. The empirical fraction $\overline{X}$ is the estimate we will publish, our guess at p.

Question: How large does n have to be so that we get good "accuracy" with high "confidence"? More precisely, suppose our pollster boss wants our estimate to have accuracy θ and confidence $1 - \delta$, meaning

$$\mathbf{Pr}\left[|\overline{X} - p| \leq \theta\right] \geq 1 - \delta.$$

How large do we have to make n?

Answer: Let's start by using the Two-sided Chernoff Bound on X. Since $X \sim \text{Binomial}(n, p)$, we have $\mathbf{E}[X] = np$. So for any $\epsilon \geq 0$, we have

$$\begin{aligned} \mathbf{Pr}[|X - pn| \geq \epsilon pn] &\leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon} \cdot pn\right) \\ \Leftrightarrow \mathbf{Pr}[|\overline{X} - p| \geq \epsilon p] &\leq 2\exp\left(-\frac{\epsilon^2}{2+\epsilon} \cdot pn\right). \end{aligned}$$

Here the two events inside the $\mathbf{Pr}[\cdot]$ are the same event; we just divided by n.

We want accuracy θ; i.e., we want $\overline{X}$ to be within θ of p with high probability. (In our original example, $\theta = 2\% = .02$.) We need to get $\theta = \epsilon p$, so we should take $\epsilon = \theta/p$.[1] Doing so, we get

$$\mathbf{Pr}[|\overline{X} - p| \geq \theta] \leq 2\exp\left(-\frac{\theta^2/p^2}{2+\theta/p} \cdot pn\right) = 2\exp\left(-\frac{\theta^2}{2p+\theta} \cdot n\right).$$

Okay, what about getting confidence $1 - \delta$? Let's look at that bound on the right. Having that n inside the $\exp(-\cdot)$ is great — it tells us the bigger n is, the less chance that our estimate is off by more than θ. As for the $\frac{\theta^2}{2p+\theta}$, well, the bigger that term is, the better. The bigger p is, the smaller that factor is, but the biggest p could be is 1. I.e.,

$$\frac{\theta^2}{2p+\theta} \geq \frac{\theta^2}{2+\theta},$$

and therefore we have

$$\mathbf{Pr}[|\overline{X} - p| \geq \theta] \leq 2\exp\left(-\frac{\theta^2}{2+\theta} \cdot n\right).$$

[1] Worried about $p = 0$? In that case, $\overline{X}$ will correctly be 0 100% of the time!

So if we want confidence $1-\delta$ in the estimate (think, e.g., $\delta = 1/20$), we would like the right-hand side in the above to be at most δ.

$$\begin{aligned} \delta \geq 2\exp\left(-\frac{\theta^2}{2+\theta}\cdot n\right) \quad &\Leftrightarrow \quad \exp\left(\frac{\theta^2}{2+\theta}n\right) \geq \frac{2}{\delta} \\ &\Leftrightarrow \quad \frac{\theta^2}{2+\theta}n \geq \ln\frac{2}{\delta} \\ &\Leftrightarrow \quad n \geq \frac{2+\theta}{\theta^2}\ln\frac{2}{\delta}. \end{aligned}$$

We have thus proved the following very important theorem. (NB: As is traditional, we've called the accuracy "ϵ" in the below, rather than "θ".)

Sampling Theorem: Suppose we use independent, uniformly random samples to estimate p, the fraction of a population with some property. If the number of samples n we use satisfies

$$n \geq \frac{2+\epsilon}{\epsilon^2}\ln\frac{2}{\delta},$$

then we can assert that our estimate $\overline{X}$ satisfies

$$\overline{X} \in [p-\epsilon, p+\epsilon] \text{ with probability at least } 1-\delta.$$

Some comments:

- That range $[p-\epsilon, p+\epsilon]$ is sometimes called the *confidence interval*.
- Due to the slightly complicated statement of the bound, sometimes people will just write the slightly worse bounds

  $$n \geq \frac{3}{\epsilon^2}\ln\frac{2}{\delta},$$

  or even

  $$n \geq O\left(\frac{1}{\epsilon^2}\ln\frac{2}{\delta}\right).$$

- One beauty of the Sampling Theorem is that the number of samples n you need *does not depend on the size of the total population*. In other words, it doesn't matter how big the country is, the number of samples you need to get a certain accuracy and a certain confidence only depends on that accuracy and confidence.
- In the example we talked about earlier we were interested in accuracy $\epsilon = 2\%$ and confidence 95%, meaning $\delta = 1/20$. So the Sampling Theorem tells us we need at least

  $$n \geq \frac{2+.02}{(.02)^2}\ln\frac{2}{1/20} = 5050\ln 40 \approx 18600.$$

  Not so bad: you only need to call 18600 or so folks! Er, well, actually, you need to get 18600 folks to respond. And you need to make sure that the events "person responds" and "person approves of the president" are independent. (Hmm... maybe being a pollster is not as easy as it sounds...)
- As you can see from the form of the bound in the Sampling Theorem, the really costly thing is getting high accuracy: $1/\epsilon^2$ is a fairly high price to have to pay in the number of samples. On the other hand, getting really high confidence is really cheap: because of the ln, it hardly costs anything to get δ really tiny.