CMPT 210: Probability and Computing

Lecture 23

Sharan Vaswani

November 28, 2024

Recap

- **Tail inequalities** bound the probability that the r.v. takes a value much different from its mean.
- **Markov's Theorem**: If X is a non-negative random variable, then for all $x > 0$, $\Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x}$.
- **Chebyshev's Theorem**: For a r.v. X and all $x > 0$, $\Pr[|X - \mathbb{E}[X]| \geq x] \leq \frac{\mathrm{Var}[X]}{x^2}$.

Voter Poll

Q: Suppose there is an election between two candidates Donald Trump and Kamala Harris, and we are hired by Harris' election campaign to estimate her chances of winning the election. In particular, we want to estimate p, the fraction of voters favoring Harris before the election. We conduct a voter poll – selecting (typically calling) people uniformly at random (with replacement so that we can choose a person twice) and try to estimate p. What is the number of people we should poll to estimate p such that our estimation is reasonably accurate?

Define X_i to be the indicator r.v. equal to 1 iff person i that we called favors Harris.

Assumption (1): The X_i r.v's are mutually independent since the people we poll are chosen randomly and we assume that their opinions do not affect each other.

Assumption (2): The people we call are identically distributed i.e. $X_i = 1$ with probability p.

Suppose we poll n people and define $S_n := \sum_{i=1}^{n} X_i$ as the r.v. equal to the total number of people (amongst the ones we polled) that prefer Harris. $\frac{S_n}{n}$ is the *statistical estimate* of p.

Q: What is the distribution of S_n?

Voter Poll

Goal: Given some desired (ϵ, δ), we want to find for what n is our estimate for p accurate up to an error $\epsilon > 0$ and with probability $1 - \delta$ (for $\delta \in (0, 1)$). Formally, we want to find an n such that:

$$\Pr\left[\left|\frac{S_n}{n} - p\right| < \epsilon\right] \geq 1 - \delta\,.$$

Since $S_n \sim \text{Bin}(n, p)$, $\mathbb{E}[S_n] = np$ and hence, $\mathbb{E}\left[\frac{S_n}{n}\right] = p$, meaning that our estimate is *unbiased* – in expectation, the estimate is equal to p. Hence, the above statement is equivalent to,

$$\Pr\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| < \epsilon\right] \geq 1 - \delta$$

Hence, we can use Chebyshev's Theorem for the r.v. $\frac{S_n}{n}$ with $x = \epsilon$ to bound the LHS

$$\Pr\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| < \epsilon\right] = 1 - \Pr\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| \geq \epsilon\right] \geq 1 - \frac{\text{Var}[S_n/n]}{\epsilon^2}.$$

In order to achieve our goal, it is sufficient to find an n such that,

$$1 - \frac{\text{Var}[S_n/n]}{\epsilon^2} \geq 1 - \delta \implies \frac{\text{Var}[S_n/n]}{\epsilon^2} \leq \delta$$

Voter Poll

Let us calculate the $\text{Var}[S_n/n]$.

$$\begin{aligned}\text{Var}[S_n/n] &= \frac{1}{n^2}\text{Var}[S_n] && \text{(Using the property of variance)}\\ &= \frac{1}{n^2}\, n\, p\,(1-p) = \frac{p\,(1-p)}{n} && \text{(Using the variance of the Binomial distribution)}\end{aligned}$$

Hence, we want to find n s.t.

$$\frac{p\,(1-p)}{n\epsilon^2} \le \delta \implies n \ge \frac{p(1-p)}{\epsilon^2\,\delta}$$

But we do not know p! If $n \ge \max_p \frac{p(1-p)}{\epsilon^2\,\delta}$, then for any p, $n \ge \frac{p(1-p)}{\epsilon^2\,\delta}$. So the problem is to compute $\max_p \frac{p(1-p)}{\epsilon^2\,\delta}$. This is a concave function and is maximized at $p = 1/2$.

Hence, $n \ge \frac{1}{4\epsilon^2\delta}$ is sufficient to ensure that $\Pr\left[\left|\frac{S_n}{n} - p\right| < \epsilon\right] \ge 1-\delta$ meaning that we have estimated p upto an error ϵ and this bound is true with high probability equal to $1-\delta$.

For example, if $\epsilon = 0.01$ and $\delta = 0.01$ meaning that we want the bound to hold with probability 0.99, then, we require $n \ge 250000$.

Pairwise Independent Sampling

Claim: Let $G_1, G_2, \ldots, G_n$ be pairwise independent random variables with the same mean μ and standard deviation σ. Define $S_n := \sum_{i=1}^{n} G_i$, then,

$$\Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{1}{n}\left(\frac{\sigma}{\epsilon}\right)^2.$$

Proof: Let us compute $\mathbb{E}[S_n/n]$ and $\mathrm{Var}[S_n/n]$.

$$\mathbb{E}[S_n] = \mathbb{E}\left[\sum_{i=1}^{n} G_i\right] = \sum_{i=1}^{n} \mathbb{E}[G_i] = n\mu \implies \mathbb{E}[S_n/n] = \frac{1}{n}\mathbb{E}[S_n] = \mu$$

(Using linearity of expectation)

$$\mathrm{Var}[S_n] = \mathrm{Var}\left[\sum_{i=1}^{n} G_i\right] = \sum_{i=1}^{n} \mathrm{Var}[G_i] = n\sigma^2$$

(Using linearity of variance for pairwise independent r.v's)

$$\implies \mathrm{Var}[S_n/n] = \frac{1}{n^2}\mathrm{Var}[S_n] = \frac{\sigma^2}{n}$$

Pairwise Independent Sampling

Using Chebyshev's Theorem,

$$\Pr\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| \geq \epsilon\right] = \Pr\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\mathrm{Var}[S_n/n]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Hence, for arbitrary pairwise independent r.v's, if n increases, the probability of deviation from the mean μ decreases.

Weak Law of Large Numbers: Let $G_1, G_2, \ldots, G_n$ be pairwise independent variables with the same mean μ and (finite) standard deviation σ. Define $X_n := \frac{\sum_{i=1}^{n} G_i}{n}$, then for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr[|X_n - \mu| \leq \epsilon] = 1.$$

Proof: Follows from the theorem on pairwise independent sampling since $\lim_{n\to\infty} \Pr[|X_n - \mu| \leq \epsilon] = \lim_{n\to\infty} \left[1 - \frac{\sigma^2}{n\epsilon^2}\right] = 1$.

Questions?