Contents

Bounding a Binomial

Suppose X is the sum of n Bernoulli random variables, $X = X_1 + X_2 + \cdots + X_n$, where random variable $X_i \sim Ber(p)$ is 1 if success in trial i and 0 otherwise.

- E.g. number of heads in n coin flips, number of corrupted bits in message sent over network
- X is a Binomial random variable: $X \sim Bin(n, p)$

$$P(X = k = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \dots, n$$

(recall $\binom{n}{k}$ is the number of outcomes with exactly k successes and $n-k$ failures)

- Often n is large, e.g. $n = 12000$ bits in a 1500B packet. Often p is small, e.g. bit error rate $p = 10^{-6}$
- Then becomes hard to compute $Bin(n, p)$. Why?
- Hint: $\binom{100}{10} \approx 10^{13}$, $\binom{100}{20}$ exceeds double presicion range.

Extreme n and p arise commonly:

- number of errors in file writen to disk
- number of elements in a particular bucket in a large hash table
- number of server crashes in a day in a large data centre
- number of facebook login requests that go to a particular server

Let's apply Chernoff inequality

- $P(X \geq a) \leq e^{et} e^{\log E[e^{tX}]}$
- $X = \sum_{i=1}^{n} X_i$
- $E[e^{tX}] = E[e^{t \sum_{i=1}^{n} X_i}] = E[\prod_{i=1}^{n} e^{tX_i}]$

- Since the X_i are independent, $E[\prod_{i=1}^n e^{tX_i}] = \prod_{i=1}^n E[e^{tX_i}]$
- For a single Bernoulli random variable X_i with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$:

$$E(e^{tX_i}) = pe^t + (1-p)e^0 = pe^t + 1 - p = 1 + p(e^t - 1)$$

- So $E[e^{tX}] = \prod_{i=1}^n E[e^{tX_i}] \leq (e^{p(e^t+1)})^n - e^{np(e^t-1)}$
- $P(X \geq a) \leq e^{-ta} e^{\log E[e^{tX}] \leq e^{-ta+np(e^t-1)}}$
- Select $a = (1+\delta)np$

$$P(X \geq (1+\delta)np) \leq e^{-np(t(1+\delta)-e^t+1)}$$

- Try $t = log(1+\delta)$

$$P(X \geq (1+\delta)np) \leq e^{-np((1+\delta)\log(1+\delta)-(1+\delta)+1)} = e^{-np((1+\delta)\log(1+\delta)-\delta)}$$

- Note that $E[X] = np$ so we can rewrite this as

$$P(X \geq (1+\delta)\mu) \leq e^{-\mu((1+\delta)\log(1+\delta)-\delta)}$$

We just need the mean μ in order to calculate bound (no need for n or p).

Web Server Load

Requests to a web server

- Historically, server load averages 20 hits per second
- What is the probability that in 1 second we receive more than 50 hits
- Number of hits $X = \sum_i X_i$. Assume hits occur independently. Apply Chernoff bound for binomial RVs.
- $E[X] = 20 = np.(1+\delta)np = 50$ so $\delta = \frac{50}{np} - 1 = 2.5 - 1 = 1.5$

$$P(X \geq 30) \leq e^{-np((1+\delta)\log(1+\delta)-\delta)} = e^{-20(2.5\log(2.5)-1.5)} \approx 10^{-7}$$

- Will almost never exceed 50 hits (assuming independence assumption valid), so enough size for server to cope with this max load
- Why does this happen? When we add **independent** X_i in the 1's and 0's tend to cancel out providing n is large. Called **statistical multiplexing** which is very important for sizing data centres, networks, etc.

Sampling

Opinion poll

- Suppose we want to know what fraction of the population likes marmite. What do you do?
- Run a poll. Ask n people and report the fraction who like marmite.
- But how to choose n? And how accurate is this anyway?
- Suppose true fraction of population who likes marmite is p
- Suppose we ask n people chosen *uniformly at random* from the population (so we need to be careful about the way we choose people, e.g. what if we only ask Australians living in Ireland?)
- Let $X_i = 1$ is person i likes marmite and 0 otherwise. Let $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $X = nY$.
- We can use Chernoff to bound $P(X \geq (1+\delta)\mu)$ (and also $P(X \leq (1-\delta)\mu)$)
- How do we select n so that estimate is not more than 5% above the mean 95% of the time?
- That is, $P(Y \geq p + 0.05) = P(X \geq np + 0.05n) \leq 0.05$
- Now $P(X \geq np + 0.05n) = P(X \geq \mu + 0.05\frac{\mu}{p}) = P(X \geq (\frac{0.05}{p})\mu) \leq 0.05$
- Chernoff bound tells us:

$$P(X \geq (1 + \frac{0.05}{p})\mu) \leq e^{-\mu((1+\frac{0.05}{p})\log(1+\frac{0.05}{p})-\frac{0.05}{p})}$$

- We want $e^{-\mu((1+\frac{0.05}{p})\log(1+\frac{0.05}{p})-\frac{0.05}{p})} \geq 0.05$. So needs:

$$\mu = np \geq \frac{\log(0.05)}{(1+\frac{0.05}{p})\log(1+\frac{0.05}{p}) - \frac{0.05}{p}}$$

$$n \geq -\frac{\log(0.05)}{(p+0.05)\log(1+\frac{0.05}{p}) - 0.05}$$

- So we need $n \geq\approx 2436$ to ensure that 95% of the time $Y \leq p + 0.05$
- Computing a lower limit, we obtain a **confident interval**: $p - 0.05 \leq Y \leq p + 0.05$ more than 95% of the time.