Applications of Chernoff Bound

Barna Saha

Estimating Sample Size

- Let p be the unknown probability that a gene mutates.
- Entire dataset size=N
- Sample size=n
- In the sample $\hat{n}$ of them have been mutated
- Estimated probability of mutation

$$\hat{p} = \frac{\hat{n}}{n}$$

Is this a reliable estimate?

When is $\hat{p} = \frac{\hat{n}}{n}$ a reliable estimate?

- Must satisfy

$$Prob(|\hat{p} - p| > \delta) \leq \gamma$$

- Or,

Confidence parameter

$$Prob(\hat{p} \in [p - \delta, p + \delta]) \geq (1 - \gamma)$$

Error tolerance

Estimating Sample Size

- Define indicator random variables X_i which is 1 if the i-th sampled element has the desired property (mutation/i-phone 8 query..) and 0 otherwise.

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

$$E[X] = E[\sum_{i=1}^{n} X_i] = nE[X_i] = nProb(X_i = 1) = np$$

Estimating Sample Size

- We have

$$
\begin{aligned}
&Prob(|\hat{p} - p| > \delta) \\
&= Prob(|n\hat{p} - np| > n\delta) \\
&= Prob(|X - E[X]| > n\delta) \\
&= Prob(|X - E[X]| > E[X]\frac{\delta}{p}) \\
&\leq 2e^{\frac{\frac{-n\delta^2}{p^2}}{3}} = 2e^{\frac{-n\frac{\delta^2}{p}}{3}} \leq 2e^{\frac{-n\delta^2}{3}}
\end{aligned}
$$

Estimating Sample Size

$$Prob(|\hat{p} - p| > \delta) \leq 2e^{\frac{-n\delta^2}{3}}$$

-
- We want $2e^{\frac{-n\delta^2}{3}} \leq \gamma$

$$\frac{2}{e^{\frac{n\delta^2}{3}}} \leq \gamma$$

$$e^{\frac{n\delta^2}{3}} \geq \frac{2}{\gamma}$$

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

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

Estimating Sample Size

$$Prob(|\hat{p} - p| > \delta) \leq 2e^{\frac{-n\delta^2}{3}}$$

-
- We want $2e^{\frac{-n\delta^2}{3}} \leq \gamma$

$$\frac{2}{e^{\frac{n\delta^2}{3}}} \leq \gamma$$

$$e^{\frac{n\delta^2}{3}} \geq \frac{2}{\gamma}$$

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

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

$$\gamma = 0.1, \delta = 0.2, n \geq 225$$

$$\gamma = 0.01, \delta = 0.2, n \geq 397$$

$$\gamma = 0.001, \delta = 0.2, n \geq 570$$

$$\gamma = 0.0001, \delta = 0.2, n \geq 742$$

Estimating Sample Size

$$Prob(|\hat{p} - p| > \delta) \leq 2e^{\frac{-n\delta^2}{3}}$$

-
- We want $2e^{\frac{-n\delta^2}{3}} \leq \gamma$

$$\gamma = 0.1, \delta = 0.2, n \geq 225$$
$$\gamma = 0.1, \delta = 0.02, n \geq 22468$$
$$\gamma = 0.1, \delta = 0.002, n \geq 2246799$$

$$\frac{2}{e^{\frac{n\delta^2}{3}}} \leq \gamma$$

$$e^{\frac{n\delta^2}{3}} \geq \frac{2}{\gamma}$$

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

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

Repeating Reservoir Sampling

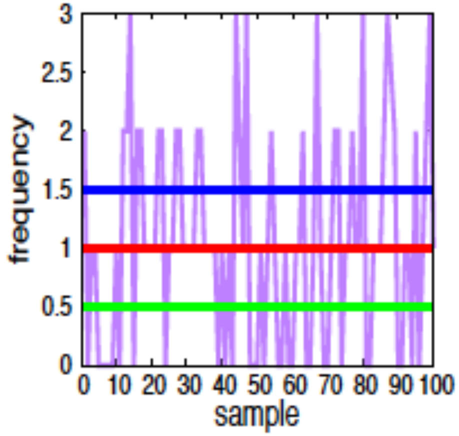

Figure 1: $m = 100$

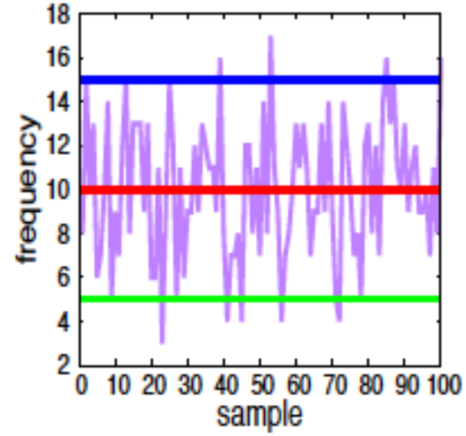

Figure 2: $m = 1000$

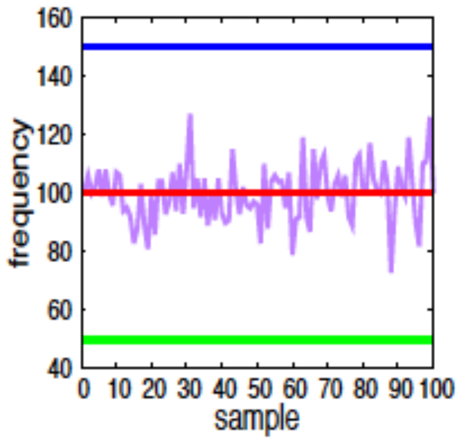

Figure 3: $m = 10000$

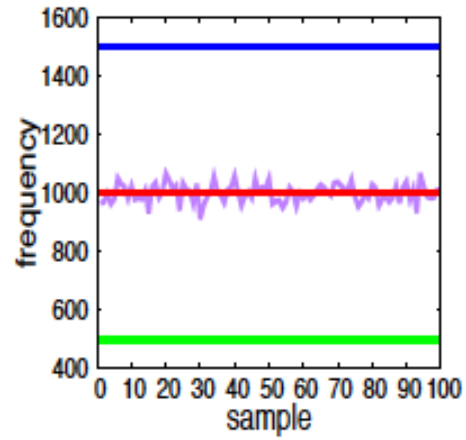

Figure 4: $m = 100000$

Repeating Reservoir Sampling

- Number of items=100
- According to the proof of the reservoir sampling each item has 1/100 chance of being stored in the reservoir
- Consider the 1st item and define an indicator random variable X_i which is 1 if the the 1st item is stored at the end of the algorithm on its i-th run.
- We run the algorithm "m" times.

Repeating Reservoir Sampling

- Define $X = \sum_{i=1}^{m} X_i$

$$E[X] = \sum_{i=1}^{m} E[X_i] = \frac{m}{100}$$

- We want the frequency of all items to be within $\frac{m}{100} \pm \frac{m}{200}$ with high probability.

Repeating Reservoir Sampling

- We want the frequency of all items to be within $\frac{m}{100} \pm \frac{m}{200}$ with high probability.

- For the 1st item

$$Prob(|X - E[X]| \geq 0.5 * E[X]) \leq 2e^{-\frac{m}{1200}}$$

Repeating Reservoir Sampling

- We want the frequency of all items to be within $\frac{m}{100} \pm \frac{m}{200}$ with high probability.

- For the 1st item

$$Prob(|X - E[X]| \geq 0.5 * E[X]) \leq 2e^{-\frac{m}{1200}}$$

- For the 2nd item

$$Prob(|X - E[X]| \geq 0.5 * E[X]) \leq 2e^{-\frac{m}{1200}}$$

- For the 100th item

$$Prob(|X - E[X]| \geq 0.5 * E[X]) \leq 2e^{-\frac{m}{1200}}$$

Repeating Reservoir Sampling

- We want the frequency of all items to be within $\frac{m}{100} \pm \frac{m}{200}$ with high probability.

- Prob there exists at least one item out of 100 such that its frequency is not in the range is

$$\leq 100 * 2e^{-\frac{m}{1200}}$$

Chernoff+Union Bound

- Random Load Balancing
  - Suppose a content delivery network like YouTube receives a million content requests per minute. Each request needs to be served from one of the 1000 servers. How should one distribute the load so that no server is overloaded.

Chernoff+Union Bound

- Random Load Balancing
  - Suppose a content delivery network like YouTube receives a million content requests per minute. Each request needs to be served from one of the 1000 servers. How should one distribute the load so that no server is overloaded.
  - Assign each request to a random server.

Random Load Balancing

- Let there be “n” requests and “k” servers
- Consider server “i”
- Define an indicator random variable X_j which will be 1 if request j is assigned to server i and 0 otherwise.
- Load on machine i: $X = \sum_{j=1}^{n} X_j$
- $E[X] = \sum_{j=1}^{n} E[X_j] = \frac{n}{k}$ [Apply the Chernoff Bound]

Random Load Balancing

- Applying the Chernoff Bound we get

$$Prob\left(X \geq \frac{n}{k} + 3\sqrt{\frac{n\ln k}{k}}\right)$$

$$= Prob\left(X \geq \frac{n}{k}(1 + 3\sqrt{\frac{\ln k}{n/k}})\right)$$

$$\leq e^{-\frac{\frac{n}{k}\frac{9\ln k}{n/k}}{3}} = \frac{1}{k^3}$$

Random Load Balancing

- Apply Union Bound

- Prob(there exists at least one server which is overloaded) $\leq \frac{1}{k^2}$