Concentration Inequalities

Barna Saha
Concentration Inequalities

• Markov Inequality:

\[ \text{Prob}(X \geq t) \leq \frac{E[X]}{t} \]

• Chebyshev Inequality:

\[ \text{Prob}(|X - E[X]| \geq t) \leq \frac{\text{Var}[X]}{t^2} \]
The Chernoff Bound

• Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in $\{0, 1\}$ with $E[X_i] = \rho_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$. Then the following holds for any $\delta > 0$

$$\Pr[X \geq (1 + \delta)\mu] < \left( \frac{e^{\delta}}{(1 + \delta)(1+\delta)} \right)^\mu$$

$$\Pr[X \leq (1 - \delta)\mu) < \left( \frac{e^{-\delta}}{(1 - \delta)(1-\delta)} \right)^\mu$$
The Chernoff Bound

• Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in $\{0, 1\}$ with $E[X_i] = p_i$.

Let $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$. Then the following holds for any $1 > \delta > 0$:

\[ Prob[X \geq (1 + \delta)\mu] \leq e^{\frac{-\mu \delta^2}{3}} \]

\[ Prob[X \leq (1 - \delta)\mu] \leq e^{\frac{-\mu \delta^2}{2}} \]
Coin Tossing Example

• Consider tossing n fair coins, that is each coin has equal probability \( \frac{1}{2} \) of returning a head or a tail. Obtain an upper bound on the probability of obtaining more than \( \frac{3}{4} \times n \) heads.

• Apply Markov, Chebyshev and the Chernoff.
Sampling: Chernoff Bound in Work

• **Estimating gene mutation:** We are interested in evaluating the probability that a particular gene mutation occurs in the population.

• **Popular Query:** We are interested in estimating the number of users searching for I-phone 8 release date.

• **Popular Item:** We are interested in the number of Amazon.com shoppers buying a particular beauty product in the last month.
Sampling: Chernoff Bound in Work

• **Estimating gene mutation:** We are interested in evaluating the probability that a particular gene mutation occurs in the population.

• Given a DNA sample, a lab test can determine if it carries the mutation. However, the test is expensive and we could only test it on a few such samples.
Sampling: Chernoff Bound in Work

• Popular Query: We are interested in estimating the number of users searching for iPhone 8 release date.

• We can examine the query log of every user to determine the total count of searches on iPhone 8 release date. However, that will require a huge amount of processing time.
Sampling: Chernoff Bound in Work

• **Popular Item:** We are interested in the number of Amazon.com shoppers buying a particular beauty product in the last month.

• We can examine the items purchased for every user in the last one month to find the number of users buying a particular beauty product. Again it will incur huge processing requirement.
Sampling: Chernoff Bound in Work

• Do the estimate on a small sample.

• Select the sample size so that estimate from the sample is reliable.
How large a sample shall we take?

- 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,

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