Concentration Inequalities in Probability Theory

Anuj Kumar Yadav

Suppose we have a biased coin such that the probability of occurring a head is some value p, i.e., $P(H) = p$, which is eventually not known to us. Now, to estimate p, we need to toss the coin a multiple number of times (n) and record the no. of the times the head occurs. The probability p can be estimated as $\hat{p}$ given by the fraction of no. of tosses in which heads occurs to the total no. of coin tosses which is eventually the sample mean $\tilde{S}_n$ where n is a very large number for estimate to be very accurate, as we have already seen in previous meetings.

Therefore, for any R.V, X which follows a certain probability distribution which is not known, it reveals out that knowing the distribution of $\tilde{S}_n$ can help us have a estimate on X, but it is not feasible to obtain the probability distribution of $\tilde{S}_n$ due to complexity as its computation involves convolution.
This is where certain Concentration lemmas help us to develop some bounds on the probability distribution of X without evaluating the sample mean, which are used to prove Law of large numbers.

Often, for a random variable X that we are interested in, we want to know - what is the probability that the value of the R.V, X is far or close to its mean or expected value, or saying how is it distributed around its expectation ?

Answer 1: Hidden in Concentration Lemmas and let's see the first generic answer which utilises expectation of the RV to establish a bound on the probability distribution:

I. Markov Inequality

For any **non-negative** Random variable $X \in \mathcal{X}$ associated with any probability distribution such that $X \geq 0$, then for all $a > 0$:

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{1}$$

Proof. Let X be a discrete random variable and A be a subset of the set $\mathcal{X}$, as X $\in \mathcal{X}$, defined as $A = \{x : X \geq a\}$, then expectation of X is given by:

$$E(X) = \sum_{x \in \mathcal{X}} x.P_X(x)$$

$$E(X) = \sum_{x \in A} x.P_X(x) + \sum_{x \notin A} x.P_X(x)$$

since, X is a non-negative random variable

$$\mathbb{E}(X) \geq \sum_{x \in A} x.P_X(x)$$

$$\mathbb{E}(X) \geq \sum_{x \in A} a.P_X(x)$$

$$\mathbb{E}(X) \geq a.\sum_{x \in A} P_X(x)$$

$$\sum_{x \in A} P_X(x) \leq \frac{E(X)}{a}$$

Therefore,

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

□

It can also be proofed considering that the X is a continuous random variable.

How can we relate it intuitively?

Example: Let $a = \frac{\mathbb{E}(X)}{\delta}$, where δ is a small quantity, then markov inequality implies that

$$P(X \geq \frac{\mathbb{E}(X)}{\delta}) \ \leq \ \delta$$

Therefore, the probability that a RV X is very large from its expected value is very small.

Real world Example: Suppose the avergae speed on a motor way is 60 Km/h. Then markov inequality implies that atmost $\frac{1}{2}$ of the vehicles drive at least 120 Km/hr.

Mathematical Example:A biased coin, when tossed give heads with a probability 0.1. It is tossed 200 times consecutively. Establish an upper bound on the probability that it gives heads at least 120 times ?

Answer: Let the random variable associated with a single toss of the coin be X which follows a Bernoulli distribution, Ber(0.1).
Since, the coin is tossed 200 times therefore, it corresponds to a binomial distribution with $n = 200$ and $p = 0.1$. Therefore, the mean of binomial R.V, Y is given by $E(Y) = np = 20$.
The markov inequality gives:

$$P(Y \geq 120) \leq \ 20/120$$

$$P(Y \geq 120) \leq \ 1/6$$

Therefore, Markov Inequality helps us establish this bound on the distribution of Y,

$$P(Y \geq 120) \leq \ 0.167; \ \ P(Y \geq 40) \leq \ 0.5; \ \ P(Y \geq 30) \leq \ 0.667; \ \ P(Y \geq 25) \leq \ 0.8 \tag{2}$$

Answer 2: The second answer to the same question is the Chebyshev Inequality which incorporates expectation as well as variance to establish bounds on distribution of X:

II. Chebyshev Inequality

Let X be any random variable associated with some probability distribution P_X, with mean μ and variance σ^2, then for all $c > 0$,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} \tag{3}$$

Proof. To prove the chebyshev inequality, we will use the result derived out of the markov Inequality. Let us define an another random variable Y such that $Y = X - \mu$,

$$P(|X - \mu| \geq c) = P(|Y| \geq c) = P(|Y|^2 \geq c^2)$$

using markov inequlity,

$$P(|Y|^2 \geq c^2) \leq \frac{E(|Y|^2)}{c^2}$$

$$P(|X - \mu|^2 \geq c^2) \leq \frac{E(|X - \mu|^2)}{c^2}$$

$$P(|X - \mu| \geq c) \leq \frac{E(|X - \mu|^2)}{c^2}$$

since, $E[(X - E(X))^2] = Var(X)$, we can have

$$P(|X - \mu| \geq c) \leq \frac{Var(X)}{c^2}$$

□

Intuitive Interpretation: let $c^2 = \frac{\sigma^2}{\delta}$, where δ can be considered a very small non negative number, then using the chebyshev Inequality, it can be inferred that:

$$P(|X-\mu| \geq \sqrt{\frac{\sigma^2}{\delta}}) \leq \delta$$

$$P(|X-\mu| \leq \sqrt{\frac{\sigma^2}{\delta}}) \geq 1-\delta$$

$$P(-\sqrt{\frac{\sigma^2}{\delta}} \leq X-\mu \leq \sqrt{\frac{\sigma^2}{\delta}}) \geq 1-\delta$$

$$P(\mu-\sqrt{\frac{\sigma^2}{\delta}} \leq X \leq \mu+\sqrt{\frac{\sigma^2}{\delta}}) \geq 1-\delta$$

The above equation can be seen as, for $\delta = 10^{-6}$, with probability/confidence $1-10^{-6}$, it can be inferred that, the X lies between $\mu \mp 10^3\sqrt{\sigma^2}$.

Mathematical Example:A biased coin, when tossed give heads with a probability 0.1. It is tossed 200 times consecutively. Establish an upper bound on the probability that it gives heads at least 120 times ?

Answer:Let Y be the binomial random variable with $n = 200$ and $p = 0.1$. We have mean = np = 20 and variance = npq = 18, on applying the chebyshev inequality, we have;

$$P(|Y-20| \geq 100) \leq \frac{18}{10^4}$$

$$P(Y \leq -80 \cup Y \geq 120) \leq \frac{18}{10^4}$$

$$P(Y \geq 120) \leq 0.0018$$

similarly, lets check the chebyshev bound on Y for different values of c.
c=20:

$$P(|Y-20| \geq 20) \leq \frac{18}{400}$$

$$P(Y \leq 0 \cup Y \geq 40) \leq \frac{18}{400}$$

$$P(Y \geq 40) \leq 0.045$$

for c=5;

$$P(Y \leq 15 \cup Y \geq 25) \leq 0.72$$

$$P(Y \geq 25) \leq 0.72$$

If we compare these bounds with those obtained using Markov inequality from Eq-10.2, it can be inferred that Chebyshev Inequality seems to be a quite stronger bound on the probability around the expected value as compared to the Markov Inequality.

III. Limitations:

However, in certain cases, it is possible that both Markov as well as chebyshev inequality fail to provide a tight bound on the probability especially for small random variables. There may be cases when they can turn out to be even non-informative. Therefore, this limits out their use and we need to have a tighter bound on the probability. This where, Chernoff bound comes into picture, which compensates the limits drwan by Markov and Chebyshev Inequality.
For example, lets consider our above example of tossing a biased coin 200 times, we had calculated that:
As per markov Inequality,

$$P(Y \geq 25) \leq \ 0.8$$

$$P(Y < 25) > \quad 1 - 0.8 = 0.2$$

Therefore, we have a bound on the CDF as

$$F_Y(24) \geq 0.2 \tag{4}$$

As per chebyshev Inequality, we have:

$$P(Y \geq 25) \leq 0.72$$

$$P(Y < 25) > 1 - 0.72 = 0.28$$

Therefore, we have a bound on the CDF as

$$F_Y(24) \geq 0.28 \tag{5}$$

Now, if we calculate the CDF of the same binomial distribution manually, we find that $P(Y \leq 24)$ i.e., $F_Y(24) = 0.855$.

Therefore, if we compare the actual value with the above bounds established with Markov and Chebyshev Inequalities, they seem to very loose, therefore, we need a more stronger bound on the probability distribution, which lays the scope for the Chernoff Bound.

Answer 3: The third answer to our same question is the Chernoff Inequality/bound which incorporates the knowledge of Moment Generating function(MGF) to establish bounds on the distribution of X:

The Moment Generating Function(MGF) associated with a Random variable X is defined as:

$$M_X(t) = \mathbb{E}[e^{t.X}] = \sum_x e^{tX}.P_X(x) \tag{6}$$

where t is some finite parameter.

The summation and integration are interchangeable depending on whether X is a Discrete or Continuous Random variable.

IV. Chernoff Bound

Let X be any random variable associated with some probability distribution P_X, and the MGF associated with X be $M_X(t)$, then for any $a > 0$:

$$P(X \geq a) \leq \inf_{t \geq 0} M_X(t).e^{-ta} \tag{7}$$

$$P(X \leq a) \leq \inf_{t \geq 0} M_X(-t).e^{ta} \tag{8}$$

Proof. We can simply use the Markov inequality to proof the chernoff bound. As stated, X is any random variable with MGF $M_X(t)$.

since, we know that for a monotonically increasing function g, for certain range of x we have:

$$Pr(X \geq a) = Pr(g(X) \geq g(a))$$

contrastingly, for a monotonically decreasing function f, for certain range of x we have:

$$Pr(X \geq a) = Pr(f(X) \leq f(a))$$

Using the above facts, for $t > 0$ and for any X, we can have;

$$P(X \geq a) = P(e^{tX} \geq e^{ta})$$

Since, e^{tX} is a one-one map and $e^{tX} > 0$, we can apply Markov Inequality on the above equation, and therefore,

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$$

$$P(X \geq a) \leq \mathbb{E}[e^{tX}].e^{-ta}$$

To, obtain the closest upper bound, the quantity can on the R.H.S can be minimized w.r.t the parameter t, given that $t > 0$ to preserve the monotonically increasing behaviour of e^{tX};
Therefore,

$$P(X \geq a) \leq \inf_{t \geq 0} \mathbb{E}[e^{tX}].e^{-ta} = \inf_{t \geq 0} M_X(t).e^{-ta} \tag{9}$$

similarly, we can prove the other equation for the Chernoff bound.
For $t > 0$,

$$P(X \geq a) = P(e^{-tX} \leq e^{-ta})$$

On taking complement both sides, we have

$$P(X \leq a) = P(e^{-tX} \geq e^{-ta})$$

since $e^{-tX} > 0$, we can apply the Markov inequality

$$P(X \leq a) = P(e^{-tX} \geq e^{-ta}) \leq \frac{\mathbb{E}[e^{-tX}]}{e^{-ta}}$$

$$P(X \leq a) \leq \mathbb{E}[e^{-tX}].e^{ta}$$

minimizing the quantity on R.H.S w.r.t the parameter t,

$$P(X \leq a) \leq \inf_{t \geq 0} \mathbb{E}[e^{-tX}].e^{ta} = \inf_{t \geq 0} M_X(-t)e^{at} \tag{10}$$

□

Intuitive Interpretation: To understand the above equations more intuitively, the equation-(10.7) can also be represented as following equations by substituting $X = Y - \mathbb{E}(Y)$ and $X = \mathbb{E}(Y) - Y$:

$$P(Y \geq \mathbb{E}(Y) + a) \leq \inf_{t \geq 0} M_{Y - \mathbb{E}(Y)}(t).e^{-ta} \tag{11}$$

$$P(Y \leq \mathbb{E}(Y) - a) \leq \inf_{t \geq 0} M_{\mathbb{E}(Y) - Y}(t).e^{-ta} \tag{12}$$

Now, if a is assumed to be a very large positive scalar (assume 10^6), it can be easily inferred from equations-(10.11) and (10.12) that the probability of having Y very far away from the expectation (either large or small) is very small as the term e^{-at} converges to a very small quantity for very large a.

Mathematical Example:
Now, lets apply the chernoff bound to the mathematical example, which we have been looking for so far, and see how good bound does chernoff inequality gives on the Random Variable.

A biased coin, when tossed give heads with a probability 0.1. It is tossed 200 times consecutively. Establish an upper bound on the probability that it gives heads at least 25 times ?

Answer: We will use the the Chernoff bound to establish the bound for the probability of $P(Y \geq 25)$:
we have:
$n = 200$

$p = 0.1$ and $q = 0.9$
using the equation-(10.7),

$$P(Y \geq 25) \leq \inf_{t \geq 0}(q + pe^t)^{200}.e^{-25t}$$

$$P(Y \geq 25) \leq \inf_{t \geq 0}(p(e^t - 1) + 1)^{200}.e^{-25t}$$

Note: *Here, we are using the taylor series series expansion of e^x to establish the bound on the above function, there could be numerous ways to solve for it and establish more tighter bound than established below.*

$$P(Y \geq 25) \leq \inf_{t \geq 0}(e^{p(e^t - 1)})^{200}.e^{-25t}$$

$$P(Y \geq 25) \leq \inf_{t \geq 0} e^{(20e^t - 25t - 20)}$$

$$P(Y \geq 25) \leq e^{\inf_{t \geq 0}(20e^t - 25t - 20)}$$

we can have the infimum at $t = 0.2231$,

$$P(Y \geq 25) \leq 0.5607$$

$$P(Y < 25) \geq 1 - 0.5607 = 0.44$$

Since, X is a discrete Random variable, we have:

$$F_Y(24) \geq 0.44 \tag{13}$$

Remark: On comparing the bound obtained on CDF obtained from Chernoff bound, i.e.,(10.13), with the bounds obtained from Markov and Chebyshev Inequality in (10.4) and (10.5) respectively, it can be concluded that in general, Chernoff bound establishes a more tighter bound, as compared to other two bounds, which makes Chernoff a stronger Concentration lemma.

A good reason behind this fact, is that more the Information we have about a Random variable, we can make more accurate conclusions on how is it distributed. Since, Markov Inequality just uses the first moment to establish the bound, Chebyshev Inequality uses the second moment to do so and the chernoff bound uses the complete Moment Generating function to establish the limits, therefore, in general Chernoff bound is the strongest and Markov inequality is the weakest concentration Inequality among the three.

REFERENCES

[1] Discussions from meetings on Concentration Lemmas, *Prof. Amitalok Budkuley, IIT Kharagpur*.
[2] Introduction to Probability, *Dimitri P. Bertsekas and John N. Tsitsiklis, MIT USA*.
[3] Lecture slides on Discrete Mathematics, *Prof. Kousha Etessami, University of Edinburg, UK*.
[4] Short Lecture on Estimates of Random variables, *Prof. Himanshu Tyagi, IISc Bangalore*.