Concentration Inequalities

Markov's inequality:

If X is a non-negative random variable, then for any $t > 0$

$$\Pr[X \geq t] \leq \frac{E[X]}{t}$$

Proof:

$$Z = \begin{cases} 0 & \text{if } X < t \\ t & \text{if } X \geq t \end{cases}$$

Clearly, with probality one,

$$Z \leq X$$

Take expectation of both

sides.

$$E[Z] \leq E[X]$$

The left-hand side is

$$E[Z] = 0 \cdot \Pr[X < t] + t \cdot \Pr[X \geq t]$$

$$= t \cdot \Pr[X \geq t]$$

Thus

$$t \cdot \Pr[X \geq t] \leq E[X]$$

Divide both sides by t. ∎

Chebyshev's inequality

If X is a random variable

If X is any random variable then for any $t > 0$,

$$\Pr[|X - E[X]| \geq t] \leq \frac{\mathrm{Var}(X)}{t^2}$$

Proof:

Let

$$Z = (X - E[X])^2$$

Then Z is non-negative.

Markov implies that

$$\Pr[Z \geq t^2] \leq \frac{E[Z]}{t^2} \qquad (*)$$

The left-hand side of $(*)$

$$\Pr[Z \geq t^2] = \Pr[(X - E[X])^2 \geq t^2]$$

$$= \Pr\left[|X - E[X]| \geq t\right]$$

The expectation on the right side of (*) is:

$$E[Z] = E\left[(X - E(X))^2\right]$$

$$= \mathrm{Var}(X)$$

- Very common case is that we have a random variable that is a sum or an average of many <u>independent</u> random variables.

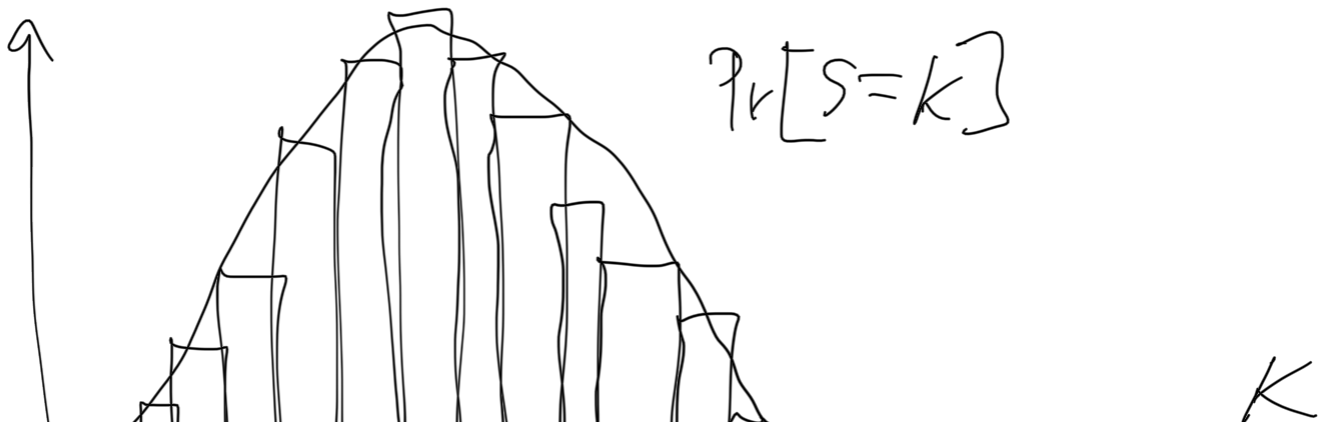
- The simplest example is

$X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli variables.

$$X_i = \begin{cases} 1 & \text{with probality } p \\ 0 & \text{with probality } 1-p. \end{cases}$$

$$S = \sum_{i=1}^{n} X_i$$

S has binomial distribution



Let's use Chebyshev's inequality for the binomial case

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E[Z] = p \qquad Var(Z) = \frac{p(1-p)}{n}$$

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \geq t\right] \leq \frac{p(1-p)}{nt}$$

Central limit theorem

If $X_1, X_2, \ldots$ are i.i.d.

with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$

Then for every $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left[\frac{\left(\sum_{i=1}^{n} x_i\right) - n\mu}{\sigma\sqrt{n}} \geq t\right] = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-x^2/2}\,dx$$

$$Z = \frac{\left(\sum_{i=1}^{n} x_i\right) - np}{\sigma\sqrt{n}}$$

Then Z converges in distribution to $N(0,1)$.

For $t \geq 1$,

$$\int_t^{\infty} e^{-x^2/2}\,dx \leq e^{-t^2/2}$$

So we expect

$$\Pr\left[\sum_{i=1}^{n} x_i - np \geq \right] \leq e^{-t^2/2}$$

$$\Pr\left[\frac{\left|\sum_{i=1}^{n} x_i - np\right|}{\sqrt{np(1-p)}} \geq t\right] \leq \cdots$$

$$\Pr\left[\left(\frac{1}{n}\sum x_i\right) - p \geq t\right]$$

$$= \Pr\left[\left(\frac{1}{n}\sum x_i - p\right)\sqrt{\frac{n}{p(1-p)}} \geq t\sqrt{\frac{n}{p(1-p)}}\right]$$

$$= \Pr\left[\frac{\sum_{i=1}^{n} x_i - np}{\sqrt{np(1-p)}} \geq \cdots\right]$$

$$\sim e^{-\frac{t^2 n}{2p(1-p)}}$$

This is exponentially better than Chebyshev's inequality.

We need a better bound.

Hoeffding's bound !