# Master of Data Science Online Programme
# Course: Probability Theory
# SGA #5: Chebyshev's inequality and multivariate normal distribution

Student: Andrei Batyrov (Fall 2022)
arbatyrov@edu.hse.ru

Higher School of Economics (HSE)
National Research University
Faculty of Computer Science

March 22, 2023

## Contents

# Problem 1

A fair coin is tossed 400 times. Let $X$ be number of heads. Prove that

$$P(X > 240) \leq \frac{1}{32}. \tag{1}$$

(Hint: use Chebyshev's inequality and symmetry considerations.)

## Solution

*Proof.* It is clearly seen that $X$ is a random variable with binomial distribution for a fair coin, i.e. $X \sim Bin(n = 400, p = 0.5)$. In Problem 2 of SGA Week3 we have shown that the variance of a binomially distributed variable $X$ is

$$VarX = np(1-p), \tag{2}$$

where $p$ is the probability of "success" (head is obtained), and $n$ is the number of tosses. Similarly, it is easy to show that the expected value of $X$ is $\mathbb{E}X = np$. Indeed, since a binomially distributed variable is a sum (number of heads) of $n$ independent identically (Bernoulli) distributed random variables $Y_1, \ldots, Y_n$, then

$$\mathbb{E}X = \mathbb{E}(Y_1 + \cdots + Y_n) = \mathbb{E}Y_1 + \cdots + \mathbb{E}Y_n = n \cdot \mathbb{E}Y = np, \tag{3}$$

where $p$ is the probability of "success" (head is obtained), $n$ is the number of tosses, and $\mathbb{E}Y_1 = \cdots = \mathbb{E}Y_n = \mathbb{E}Y = 1 \times p + 0 \times (1-p) = p$ for the Bernoulli distribution.

Now let's consider the Chebyshev's inequality [2]:

$$P(|X - \mathbb{E}X| > \varepsilon) \leq \frac{VarX}{\varepsilon^2}, \tag{4}$$

where $\varepsilon$ is some real number. The inequality means that the probability of "the distance between an outcome of the random variable $X$ and its expected value $\mathbb{E}X$ being greater than $\varepsilon$" is no greater than the variance of $X$ divided by the same $\varepsilon$. By expanding the left-hand side of (4), we get

$$\begin{aligned} P(|X - \mathbb{E}X| > \varepsilon) &= P(-\varepsilon > X - \mathbb{E}X > \varepsilon) \\ &= P(\mathbb{E}X - \varepsilon > X > \mathbb{E}X + \varepsilon) \\ &= P(\mathbb{E}X - \varepsilon > X) + P(X > \mathbb{E}X + \varepsilon). \end{aligned} \tag{5}$$

Now let's first show that the binomial distribution is symmetric, if $p = 0.5$. As per [1], the *pmf* of the binomial distribution is

$$P(X = k) = \binom{n}{k} p^n (1-p)^{n-k}, \tag{6}$$

where $p$ is the probability of "success" (head is obtained), $n$ is the total number of tosses, $k$ is the number of "successful" tosses (number of heads). Since $p = 0.5$, $1 - p = 1 - 0.5 = 0.5 = p$, and (6) can be re-written as

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} p^k p^{n-k} = \binom{n}{k} p^n = \frac{n!}{k!(n-k)!} p^n. \tag{7}$$

To demand symmetry, the probability of the number of "successful" tosses $k$ which is $P(X = k)$ must be equal to the probability of the number "failed" tosses (number of tails) $n - k$ which is $P(X = n - k)$. Indeed, as per (7)

$$P(X = n-k) = \binom{n}{n-k} p^n = \frac{n!}{(n-k)!(n-(n-k))!} p^n = \frac{n!}{k!(n-k)!} p^n. \tag{8}$$

So, since $P(X = k) = P(X = n - k)$, the binomial distribution for tossing a fair coin ($p = 0.5$) is symmetric with the point of symmetry $\mathbb{E}X$ [3].

Since we have just shown that our distribution is symmetric, the probabilities $P(\mathbb{E}X - \varepsilon > X)$ and $P(X > \mathbb{E}X + \varepsilon)$ for symmetric points $\mathbb{E}X \pm \varepsilon$ are equal. Thus, the left-hand side of (4) is

$$P(|X - \mathbb{E}X| > \varepsilon) = 2P(X > \mathbb{E}X + \varepsilon) \Rightarrow P(X > \mathbb{E}X + \varepsilon) = \frac{1}{2} P(|X - \mathbb{E}X| > \varepsilon). \tag{9}$$

In our case $\mathbb{E}X+\varepsilon = 240$. Solving this equation for $\varepsilon$ yields $\varepsilon = 240-\mathbb{E}X = |\mathbb{E}X = np \text{ as per } (3)| = 240 - np = 240 - 400 \times 0.5 = 240 - 200 = 40$. Finally, as per (9) and (4), we get

$$\begin{aligned} P(X > 240) &= \frac{1}{2}P(|X - 200| > 40) \\ &\le \frac{1}{2} \cdot \frac{VarX}{\varepsilon^2} \\ &= |VarX = np(1-p) \text{ as per } (2)| \\ &= \frac{np(1-p)}{2\varepsilon^2} \\ &= \frac{400 \times 0.5 \times (1-0.5)}{2 \times 40^2} \\ &= \frac{100}{3200} \\ &= \frac{1}{32} \Rightarrow P(X > 240) \le \frac{1}{32}. \end{aligned} \tag{10}$$

■

## Discussion

Is the result $P(X > 240) \le 1/32$ expected or not? Since the binomial random variable is in fact the sum of $n$ independent identically (Bernoulli) distributed random variables, we can apply the Central Limit Theorem [5] to the average of Bernoulli random variables. However, we are interested not in the average but in the sum. To make sure, if we still can apply the CLT to the sum, let's compare the $Z$-scores of the average and the sum of Bernoulli random variables $Y_1, \dots, Y_n$.

$Z$-score of the average $\bar{Y} = (Y_1 + \cdots + Y_n)/n = X/n$ is

$$\begin{aligned} Z_{\bar{Y}} &= \frac{\bar{Y} - \mathbb{E}Y}{\sqrt{VarY}}\sqrt{n} \\ &= \frac{\bar{Y} - p}{\sqrt{p(1-p)}}\sqrt{n} \\ &= |p = 0.5 = 1 - p| \\ &= \frac{\bar{Y} - p}{p}\sqrt{n} \\ &= |\bar{Y} = X/n, \text{ since } X = Y_1 + \cdots + Y_n| \\ &= \frac{X/n - p}{p}\sqrt{n} \\ &= \frac{X - np}{p\sqrt{n}}. \end{aligned} \tag{11}$$

$Z$-score of the sum $X = Y_1 + \cdots + Y_n$ is

$$\begin{aligned} Z_X &= \frac{X - \mathbb{E}X}{\sqrt{VarX}}\sqrt{n} \\ &= |X = nY, \text{ since } X = Y_1 + \cdots + Y_n| \\ &= \frac{X - np}{\sqrt{Var(nY)}}\sqrt{n} \\ &= \frac{X - np}{\sqrt{n^2 VarY}}\sqrt{n} \\ &= \frac{X - np}{\sqrt{n^2 p(1-p)}}\sqrt{n} \\ &= |p = 0.5 = 1 - p| \\ &= \frac{X - np}{np}\sqrt{n} \\ &= \frac{X - np}{p\sqrt{n}} \Rightarrow Z_X = Z_{\bar{Y}}. \end{aligned} \tag{12}$$

We have obtained identical $Z$-scores, which means that the CLT holds not only for the average, but also for the sum of Bernoulli random variables (our case), and both $CDF_{Z_{\bar{Y}}}$ and $CDF_{Z_X}$ approach the $CDF_{\mathcal{N}(0,1)}$ as $n \to \infty$. So, in our case for $n = 400 \gg 1$ we can assume that our distribution is close to normal distribution with moments $\mu = \mathbb{E}X = 200$ and $\sigma = \sqrt{VarX} = \sqrt{100} = 10$, i.e. $X \sim \mathcal{N}(\mu = 200, \sigma^2 = 100)$. Now, as per [4], 99.7% of all values of $X$ are in the segment $[\mu - 3\sigma, \mu + 3\sigma] = [200 - 3 \times 10, 200 + 3 \times 10] = [170, 230]$, i.e. $P(170 \leq X \leq 230) \approx 0.997$. So it is no wonder that $P(X > 240)$ is bound by a rather small number $1/32 = 0.03125$ – we are in the right tail of the distribution.

## Answer

$P(X > 240) \leq 1/32$ indeed. This result means that for 400 tosses of a fair coin the probability of "$X$ being more than 40 heads away from its expected value of 200 heads" is less than or equal to $1/32 = 0.03125$. If we assume that for $n = 400 \gg 1$ we are close to normal distribution, this result is expected. Out of 400 tosses of a fair coin the chance to obtain more than 40 heads than expected (200) is rather low.

## References

[1] Ilya Schurov. *Binomial distribution*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756754.

[2] Ilya Schurov. *Chebyshev's inequality*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756849.

[3] Ilya Schurov. *Expected value of continuous random variable*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756824.

[4] Ilya Schurov. *Properties of normal distribution*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756858.

[5] Ilya Schurov. *Z-score and statement of central limit theorem*. Faculty of Computer Science, Higher School of Economics. URL: https://smartedu.hse.ru/mod/page/0/756856.