MTH 224, Spring 2024

Instructor: Bahman Angoshtari

Lecture 19

Section 5.6: Central limit theorem.

19.1. The central limit theorem (CLT)

The normal approximation of the binomial distribution is a special case of the central limit theorem (CLT). This result states that the sum of any sequence of i.i.d (independent, identically distributed) random variables is approximately normally distributed.

THEOREM 19.1. *Let $X_1, X_2 \ldots$ be i.i.d random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $Z_n = \frac{X_1+\cdots+X_n-n\mu}{\sqrt{n}\sigma}$. For any $t \in \mathbb{R}$,we have that $\lim_{n\to+\infty} \mathbb{P}(Z_n \leq t) = \phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx$. In particular,*

$$\lim_{n\to+\infty} \mathbb{P}(a \leq Z_n \leq b) = \phi(b) - \phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx.$$

EXAMPLE 19.2. We roll a fair die 12 times. Let S be the sum of the results, and Q be the product of the results. Estimate the probabilities $\mathbb{P}(S \geq 40)$ and $\mathbb{P}(Q \leq 100,000)$.

SOLUTION. Let X_i be the i^{th} result, and $S = X_1 + \cdots + X_{12}$. We have that

$$\mathbb{E}[X_i] = \frac{1}{6}(1+2+3+4+5+6) = \frac{1}{6} \cdot \frac{6(6+1)}{2} = \frac{7}{2},$$

and

$$\begin{aligned} \mathrm{Var}(X_i) &= \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 \\ &= \frac{1}{6}(1+2^2+3^2+4^2+5^2+6^2) - \frac{49}{4} \\ &= \frac{91}{6} - \frac{49}{4} = \frac{35}{12}. \end{aligned}$$

Therefore,we have $\mathbb{E}[S] = 12 \cdot \frac{7}{2} = 42$ and $\mathrm{Var}(S) = 12 \cdot \frac{35}{12} = 35$. By CLT, we can approximate S by a normal distribution with mean 42 and variance 35:

$$\mathbb{P}(S \geq 40) = \mathbb{P}\left(\frac{X_1 + \cdots + X_{12} - 42}{\sqrt{35}} \geq \frac{40-42}{\sqrt{35}}\right) \approx 1 - \phi\left(-\frac{2}{\sqrt{35}}\right) \approx 0.63.$$

Although $Q = \Pi_{i=1}^{12} X_i$ is not a sum, we do have that $\ln Q = \sum_{i=1}^{12} \ln X_i = \sum_{i=1}^{12} Y_i$, where $Y_i = \ln X_i$ are i.i.d r.v.s. Using a similar calculation as the one for finding $\mathbb{E}[X_i]$ and $\mathrm{Var}(X_i)$,we obtain that $\mathbb{E}[Y_i] \approx 1.1$ and $\mathrm{Var}(Y_i) \approx 0.3667$. So, $\mathbb{E}[\ln Q] \approx 12 \times 1.1 = 13.2$ and $\mathrm{Var}(\ln Q) \approx 12 \times 0.3667 = 4.39$. Finally, by CLT:

$$\mathbb{P}(Q \leq 100,000) = \mathbb{P}\left(Q \leq 10^5\right) = \mathbb{P}(\ln Q \leq 5 \cdot \ln 10) \approx \mathbb{P}\left(\frac{\ln Q - 13.2}{\sqrt{4.39}} \leq \frac{11.51 - 13.2}{\sqrt{4.39}}\right)$$
$$\approx \phi(-0.81) = 1 - \phi(0.81) \approx 0.21.$$

Example 19.3. We play the lottery every week. The probability to win is 1/20. How many weeks should we plan to play, so that the probability to win more than 10 times is at least 0.7?

Solution. Denote by W the number of weeks we play until we win for the 10^{th} time. Note that $W \sim \mathrm{NB}(10, 1/20)$, and recall that $W = X_1 + \cdots + X_{10}$, where $X_i \sim \mathrm{G}(1/20)$ are independent. We have $10 \cdot \mathbb{E}[X_i] = 10 \cdot \frac{1}{1/20} = 200$, and $10 \cdot \mathrm{Var}(X_i) = 10 \cdot \frac{\left(1-\frac{1}{20}\right)}{(1/20)^2} = 3800$.

Our goal is to find n such that $\mathbb{P}(W \leq n) \geq 0.7$. By CLT, we have

$$\mathbb{P}(W \leq n) = \mathbb{P}\left(\frac{W - 200}{\sqrt{3800}} \leq \frac{n - 200}{\sqrt{3800}}\right) \approx \phi\left(\frac{n - 200}{\sqrt{3800}}\right) \geq 0.7,$$

which implies that $\frac{n-200}{\sqrt{3800}} \geq \phi^{-1}(0.7) \approx 0.524$. It then follows that $n \geq 232.3$ (≈ 4.46 years).

Example 19.4. (Predicting the outcome of elections). In order to predict the outcome of a presidential election, a poll is taken. The predicted percentage of votes for candidate A is computed from the poll. How many people need to be included in the poll to validate the following claim:

"The predicted percentage is accurate within 1% with probability of at least ≥ 0.95" ?

Solution. Let $p =$ actual percentage, $n =$ poll size. Let $S_n =$ the number of people in favor of candidate A. Note that $S_n \sim \mathrm{Bin}(n, p)$ and that the predicted percentage is $\frac{S_n}{n}$.

We need to find n large enough so that $P\left(\left|\frac{S_n}{n} - p\right| \leq 0.01\right) \geq 0.95$. The rest of the solution will be discussed in the next lecture.