26 2025-03-28 | Week 10 | Lecture 25

Start reading 6.1, 6.2

Recall that we say that a sequence $X_1, X_2, \ldots$ of random variables is *independent and identically distributed (IID)* if

- The X_i's are independent rvs
- Every X_i has the same probability distribution as X_1

Theorem 80 (Central limit theorem)**.** *Suppose we have an IID sequence of random variables $X_1, X_2, \ldots$ with finite mean $\mathbb{E}[X_1] = \mu$ and finite variance $\mathrm{Var}(X_1) = \sigma^2 > 0$. Let*

$$S_n = X_1 + \ldots + X_n.$$

Then for any fixed $-\infty \le a \le b \le +\infty$, we have

$$\lim_{n\to\infty} \mathbb{P}\left[a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right] = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \tag{23}$$

The conclusion of the CLT can be stated in many equivalent ways. One useful way involves the sample average $\overline{X} := \frac{S_n}{n}$. The CLT says that

$$\overline{X} \approx \underbrace{\mu + \frac{\sigma}{\sqrt{n}} Z}_{\mathcal{N}(\mu, \frac{\sigma^2}{n})} \tag{24}$$

where $Z \sim \mathcal{N}(0,1)$.

Usually $n \ge 30$ is big enough for the approximation to be valid.

Example 81. A witch decides to make 50 batches of potions. Unfortunately, her hut is not very clean and bugs keep crawling into the brew. Suppose that the amount of bugs contaminating the ith batch is a random variable X_i with mean $\mu = 4g$ and standard deviation $\sigma = 1.5g$. Assuem the witch prepares her 50 batches independently. Let

$$\overline{X} = \frac{X_1 + \ldots + X_{50}}{50}$$

be the average amount of impurity.

Question: What is the probability that $3.5 \le \overline{X} \le 3.8$?
Solution: Here, $n = 50 > 30$ so we can apply the central limit theorem. By the central limit theorem, $\overline{X}$ has distribution approximately

$$\begin{aligned}
\overline{X} &\approx \mu + \frac{\sigma}{\sqrt{n}} Z \\
&= 4 + \frac{3}{2\sqrt{50}} Z
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ is a standard normal random variable. Therefore

$$\begin{aligned}
\mathbb{P}\left[3.5 \le \overline{X} \le 3.8\right] &\approx \mathbb{P}\left[3.5 \le 4 + \frac{3}{2\sqrt{50}} Z \le 3.8\right] \\
&= \mathbb{P}\left[-\frac{1}{2} \le \frac{3}{2\sqrt{50}} Z \le -\frac{1}{5}\right] \\
&= \mathbb{P}\left[-\frac{\sqrt{50}}{3} \le Z \le -\frac{2\sqrt{50}}{15}\right] \\
&= \mathbb{P}\left[-2.36 \le Z \le -.94\right] \\
&= \int_{-2.36}^{-.94} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&\approx .16
\end{aligned}$$

End of Example 81. □

Example 82 (Application of CLT)**.** The human genome consists of a sequence of nucleotides (A, T, C, G). For this problem we can think of the human genome as a sequence of nucleotides

`CAGGAAGCATTTATATAGAGGTAATAATTAAATTTAACTTTTTATTATTCTCTATGCCCAAGAGAATGTG...`

and the sequence consists of $n = 3.2 \times 10^9 = 3,200,000,000$ nucleotide letters.

The DNA mutation rate in humans is approximately $\mu = 2 \times 10^{-8} = 0.00000002$ mutations per nucleotide per generation. That means each time a person is born, each nucleotide either mutates (with proability μ) or doesn't mutate (with probability $1 - \mu$). Assume that nucleotides mutate *independently*, that is, whether one nucleotide undergoes mutates is indepdenent from any other nucleotides.

Let S_n be the number of new mutations in the genome of a randomly-selected newborn baby. We can write

$$S_n = X_1 + \ldots + X_n$$

where

$$X_i = \begin{cases} 1 & : \text{ mutation at site } i \\ 0 & : \text{ no mutation at site } i \end{cases}$$

Let M be the number of new mutations in the genome of a randomly selected newborn baby. Then $M \sim \text{Bin}(n, \mu)$ where the "number of trials" is $n = 3.2$ billion and the "success probability" is $\mu = 2 \times 10^{-8}$.

Question 1: On average, how many mutations does a newborn baby have?
Solution: We want

$$\begin{aligned} \mathbb{E}[S_n] &= \mathbb{E}[X_1 + \ldots + X_n] \\ &= \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] \\ &= n\mu \\ &= (3.2 \times 10^9)(2 \times 10^{-8}) \\ &= 64. \end{aligned} \tag{25}$$

Question 2: What is the probability distribution of S_n?
Solution: S_n is a sume of n "coin flips" each having success probability μ. Hence, S_n is a binomial random variable with number of trials n and success parameter μ.

Question 3: What percentage of people have between 56 and 72 mutations?
Solution: Use the central limit theorem. First observe that

$$\mathbb{E}[X_1] = 1 \cdot \mu + 0 \cdot (1 - \mu) = \mu \tag{26}$$

and

$$\mathbb{E}\left[X_1^2\right] = 1^2 \cdot \mu + 0^2 \cdot (1 - \mu) = \mu$$

Therefore

$$\begin{aligned} \sigma^2 &= \text{Var}(X_1) \\ &= \mathbb{E}\left[X_1^2\right] - (\mathbb{E}[X_1])^2 \\ &= \mu - \mu^2. \end{aligned}$$

Since μ is very small μ^2 is VERY VERY small, and hence we have $\sigma^2 \approx \mu$. Therefore

$$\sigma \approx \sqrt{\mu} = \sqrt{2} \times 10^{-4}. \tag{27}$$

Finally, we have

$$\sqrt{n} = 40000\sqrt{2}. \tag{28}$$

Therefore, using Eq. (23) from the Central limit theorem, we have

$$\mathbb{P}\left[a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

and plugging in the values from Eqs. (25), (27) and (28), we get

$$\mathbb{P}\left[a \leq \frac{S_n - 64}{8} \leq b\right] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

or equivalently

$$\mathbb{P}\left[64 + 8a \leq S_n \leq 64 + 8b\right] \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

Taking $a = -1$ and $b = 1$, we get

$$\mathbb{P}\left[56 \leq S_n \leq 72\right] \approx \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \approx .68.$$

Conclusion: about 68% of people have between 56 and 72 mutations.

What percentage of people have between 48 and 80 mutations?

End of Example 82. □