Central limit theorem simply explained

NDH802 - Spring 2021

The research question

Suppose we want to investigate the cheating behaviors among the Swedish students in a school year. We assume 2 outcomes, to cheat or not to cheat. Each student takes ten exams. The probability of cheating is 10% (similarly applicable for all exams, all students). To make it fancier, we write it in statistical language. Let X denote the cheated exams, $X \sim Bin(10, 0.1)$.

Kindly note that these assumptions are purely hypothetical.

Data simulation

Normally, we don't know the so-called *true* distribution, but for the sake of learning the CLT, let's pretend we do to simulate some data. In the Rmd file you'll find the data simulation process. It is, however, not the focus of this illustration.

Now we have a so-called **population** data that includes 360000 rows like these. Each row represents a student with his/her ID and number of exams he/she cheated. For example, student #1 cheated on 1/10 exam(s), student #2 cheated on 0/10 exam, student #3 cheated on 1/10 exam(s), and so on.

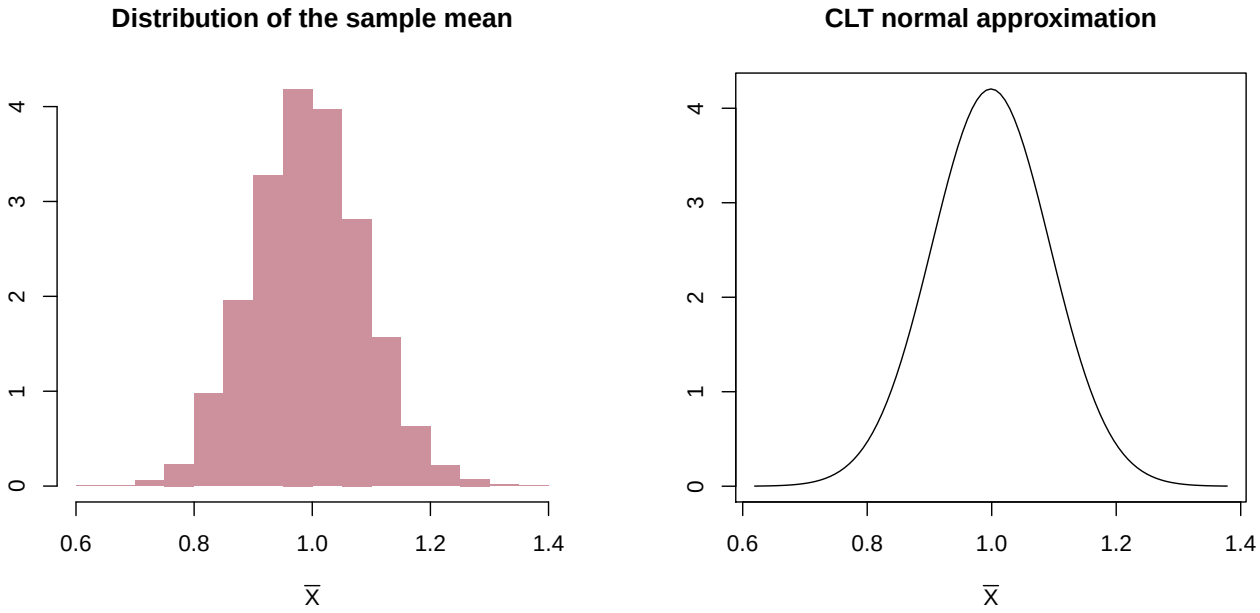
```
##    studentID cheat
## 1          1     1
## 2          2     0
## 3          3     1
## 4          4     0
## 5          5     1
## 6          6     1
## 7          7     0
## 8          8     1
## 9          9     2
## 10        10     1
```

Once we have created the data, we pretend (again) it has nothing to do with us. Two activities we often do when we learn statistics are to imagine and to pretend we know or don't know things, depending on the questions.

Back to our research question. If we had unlimited time and resources, we would investigate each and every student and record his/her cheating behavior. Unfortunately, we don't. That's when sampling comes to our rescue.

Sampling

Still, let's imagine we have a huge budget. We draw a sample of 100 students to investigate, and record the sample mean of the cheated exams. We call it $\bar{X}_i$. We repeat it, say 10000 times, i.e, we finally have 10000 values of $\bar{X}_i$. We put them together and call it (vector) $\bar{X}$. The distribution of $\bar{X}$ is plotted in the histogram below (left hand side).



Now, let's stop imagining. We usually work with tight budget. We are therefore grateful to the statistician who laid out one of the most powerful foundations of statistics. **The Central Limit Theorem (CLT) suggests that when we have a huge number of observations, the sample mean follows a normal distribution**. Formally, when n is large, $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$. This curve is plotted above (right hand side). (*Small note:* I guess the trickiest part here is that previously when you learn about $\bar{X}_n$, it's a number, when we discuss the CLT, $\bar{X}_n$ is a random variable.)

Let's compute and compare some probabilities, one with the imaginary data we collected, one with the normal approximation. For example, we would like to know the probability that a random student cheats on more than 1 exam, or $P(X > 1)$.

```
cut_point = 1

#If we had data, we simply count the number observations that satisfy the condition and divide it by th
data_prob = sum(sample_mean > cut_point)/length(sample_mean)

#This we practised serveral times when we learn about the normal distribution
clt_prob = pnorm(cut_point, mean = mu, sd = sqrt(sigma_squared_clt), lower.tail = F)

#Print out the results
data.frame(data_prob, clt_prob)
```

```
##   data_prob  clt_prob
## 1    0.4652 0.4948149
```

Pretty close right? We are computing this with sample size of 100. When the sample size gets larger, the results would get closer.

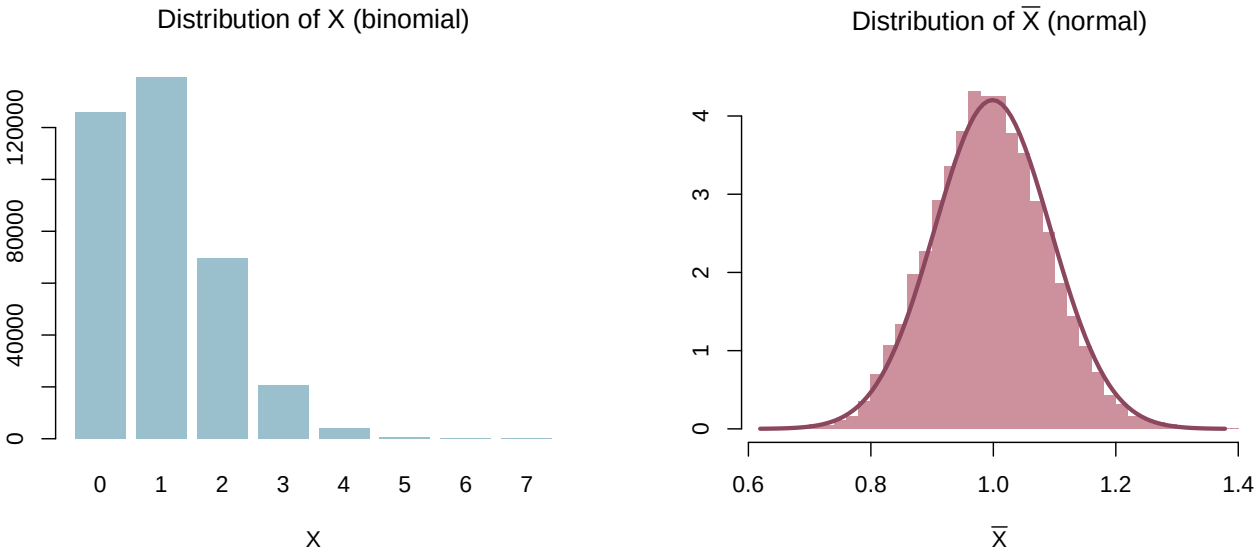
Last time we pose an interesting question, **how large is large and how close is close?** The answer is beautifully captured in the variance of the normal approximation $\frac{\sigma^2}{n}$. Technically, when n tends to ∞, the variance tends to 0. In real life, it means when n gets larger and larger, the variance get smaller and smaller, i.e., we are more and more certain about the μ.

If you'd like, you can try modifying the sample size in the Rmd file to get a better hang of "converging".

Key takeaways

(i) Distribution of the random variable X is **different from** the distribution of the sample mean $\bar{X}$.
(ii) When the number of observations is large, the sampling distribution of the sample mean $\bar{X}$ converges (approximates) to **the normal distribution** $N(\mu, \frac{\sigma^2}{n})$, **regardless** of the distribution of X.

For example, the blue bars illustrate the number of cheated exams following binomial distribution, while the pink bars represent the **average** number of cheated exams (the **sample mean**), following normal distribution when the sample size is large.



That's pretty much it. We hope you had fun learning about the CLT and understand its importance. If you have general questions about the CLT, feel free to contact Emelie/Huong. If you have questions about the Rmd file, please contact Huong.