The central limit theorem

Chrysafis Vogiatzis

Lecture 10

Learning objectives

After these lectures, we will be able to:

- Explain why the normal distribution appears often in real life.
- Recall and use the central limit theorem.

Motivation: The normal distribution

Why is the normal distribution so ubiquitous? Why is it that in many instances we see normally distributed quantities around us?

Motivation: Testing a hypothesis

Consider the case of trying to predict the outcome of an election. A good way to do so would be to pick a sample n of potential voters, and ask them what they would be voting for. Say x say they are voting for Candidate 1: this could lead you to deduce that x/n is the proportion of the vote that Candidate 1 would get in the general election. But, what can you say for the distribution of the proportion? How probable is it that your prediction is off?

Introduction

These lecture notes are organized as follows. First, we motivate why the central limit theorem applies; later in the notes, we state the theorem in its entirety. We finish this lecture with an example.

Motivating the central limit theorem

Say we throw a "fair" die (uniform distribution of getting any of the six numbers) 100,000 times and we collect back the appearances of each number. We then plot our results (see Figure 1) and observe that, as expected, every number appears equally probably. Now, let's consider a game of Monopoly. In this board game, you throw 2 dies and the summation of the two numbers is the number of steps you are expected to take: note that this number goes from 2 (both

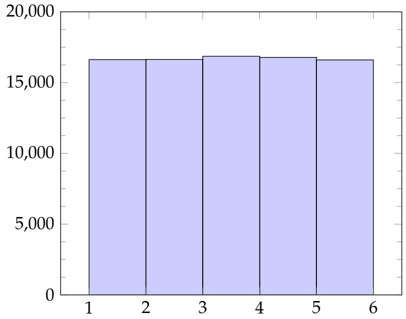


Figure 1: The number of occurrences for each number from 1 to 6 for one fair die.

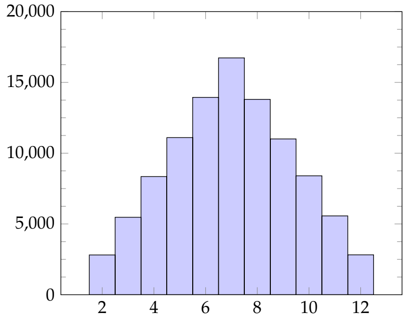


Figure 2: The number of occurrences for numbers from 2 to 12 for two fair dies.

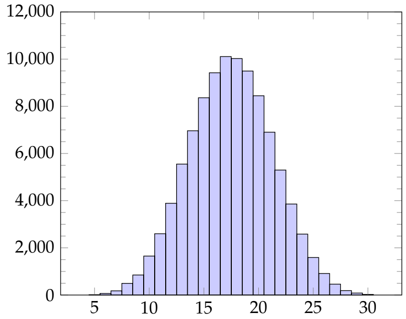


Figure 3: The number of occurrences for numbers from 5 to 30 for five fair dies.

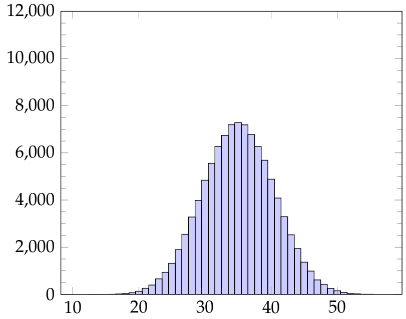


Figure 4: The number of occurrences for numbers from 10 to 60 for ten fair dies.

dies land on the side of 1) to 12 (both dies land on the side of 6). We already saw earlier how the probability of getting a 7 is higher than the rest, which is revealed also in Figure 2. There seems to be an interesting pattern emerging... Let's investigate this more!

We now proceed to show what happens when we toss 5 and 10 dies (see Figures 3 and 4). The pattern is even clearer now: it seems like **the summation of many random variables with the same distribution follows a normal distribution**! Let's see whether we can formally state what we observe.

Theorem 1 (The central limit theorem) *Let X_i, $i = 1, \dots, n$ be a series of independent, identically distributed random variables.* [1] *Also, define $Z = \sum_{i=1}^{n} X_i$ (i.e., as the summation of all random variables X_i) or define $Y = \sum_{i=1}^{n} X_i / n$ (i.e., as the average of all X_i).*

Then both Z and Y follow a normal distribution when n is large enough. [2]

What is the implication of this result? Say we are measuring some random variable that is an average of independent random variables coming from the same distribution; then this average is expected to be normally distributed! This is why the normal distribution appears

[1] Continuous or discrete, Bernoulli, binomial, geometric, Poisson, exponential, uniform, normal – any distribution. Note though that all random variables need to follow the same distribution.

[2] What constitutes "large enough"? We will investigate this later in the semester.

so often in real life. And this is pretty interesting since we live in a world full of data that does not seem to follow any "clean", nice distributions: yet, selecting samples from this data and analyzing them provides us with a nice normal distribution to work with.

Waiting for a bus

Assume that the time you have to wait for a bus every day is uniformly distributed between 0 and 4 minutes.

a) What is the probability you have to wait for more than 3 minutes for the bus today?

b) What is the probability you have to wait on average for more than 3 minutes for the bus during 5 days of waiting for the bus every day?

c) What is the probability you have to wait on average for more than 3 minutes for the bus during your stay in Urbana-Champaign?

a) The first one is pretty straightforward: uniform distribution, continuous between 0 and 4: hence, the probability is $\frac{1}{4} = 0.25$.

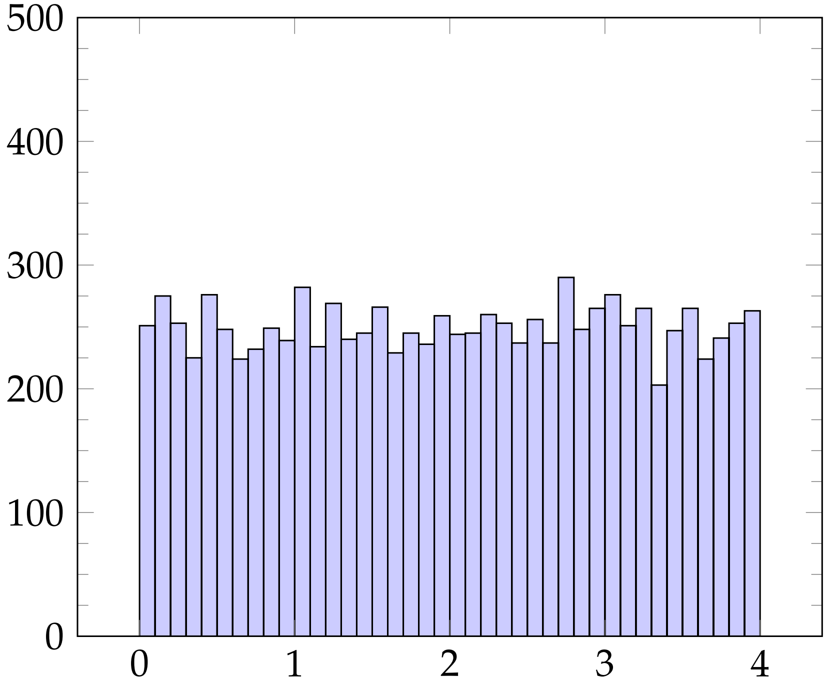
b) The second one is tougher. Is $n = 5$ (for 5 days of waiting for the bus every day) big enough for the central limit theorem to apply? And does it help to apply it?

c) The third one is similar to our previous reasoning–**but** keeping in mind that your stay in Urbana-Champaign is for 3-4 years, we may assume that n (number of days waiting for a bus) is pretty big. Does the central limit theorem help?

Based on the central limit theorem, the average time we wait for the bus is normally distributed. Let us visualize what this means (much like what we did for the dies earlier). We will generate 10000 random variables and present the results depending on the number of times each interval of numbers appears.

Finally, for the last question (where the central limit theorem would clearly hold as n is very big), we can immediately find the probability using a normal distribution. The details of the normal distribution will be discussed later in the semester.

Figure 5: The amount of time we spend waiting for one bus.



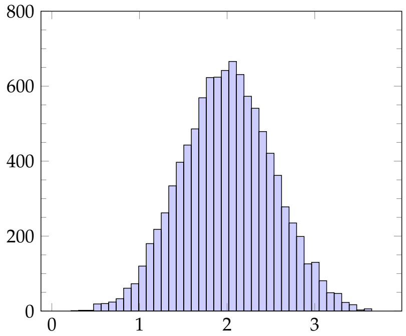


Figure 6: The average amount of time we spend waiting for a bus for a total of $n = 5$ days.

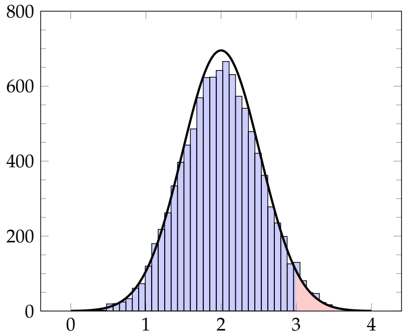


Figure 7: Visualizing the "probability" of waiting (on average) for more than 3 minutes in 5 days. It is shown in red.

The central limit theorem

Expectation and variance review

Let us recall a few important properties from calculating expectations and variances. Given a series of independent random variables $X_i, i = 1, \ldots, n$, we have that:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right].$$

$$Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var\left[X_i\right].$$

Moreover, recall that:

$$E\left[\alpha \cdot X\right] = \alpha \cdot E\left[X\right].$$

$$Var\left[\alpha \cdot X\right] = \alpha^2 \cdot Var\left[X\right].$$

Combining, we have that for quantities $\frac{\sum_{i=1}^{n} X_i}{n}$, we get:

$$E\left[\frac{\sum\limits_{i=1}^{n} X_i}{n}\right] = \frac{\sum\limits_{i=1}^{n} E\left[X_i\right]}{n}.$$

$$Var\left[\frac{\sum\limits_{i=1}^{n} X_i}{n}\right] = \frac{\sum\limits_{i=1}^{n} Var\left[X_i\right]}{n^2}.$$

Assuming that we have $\mu = E\left[X_1\right] = E\left[X_2\right] = \ldots = E\left[X_n\right]$ and $\sigma^2 = Var\left[X_1\right] = Var\left[X_2\right] = \ldots = Var\left[X_n\right]$, we may write that:

$$E\left[\sum_{i=1}^{n} X_i\right] = n \cdot \mu \qquad E\left[\frac{\sum\limits_{i=1}^{n} X_i}{n}\right] = \frac{n \cdot \mu}{n} = \mu.$$

$$Var\left[\sum_{i=1}^{n} X_i\right] = n \cdot \sigma^2 \qquad Var\left[\frac{\sum\limits_{i=1}^{n} X_i}{n}\right] = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The full theorem

We are now ready to state the central limit theorem in its entirety.

Theorem 2 (The central limit theorem) *Let X_i, $i = 1, \ldots, n$ be a series of independent, identically distributed random variables with expected value $E[X_i] = \mu$ and variance $Var[X_i] = \sigma^2$. Define $Z = \sum_{i=1}^{n} X_i$ (i.e., as the summation of all random variables X_i) and $Y = \sum_{i=1}^{n} X_i/n$ (i.e., as the average of all X_i).*

Then:

- *Z follows a normal distribution when n is large enough with parameters $\mu_Z = \sum_{i=1}^{n} E[X_i] = n \cdot \mu$ and $\sigma_Z^2 = \sum_{i=1}^{n} Var[X_i] = n \cdot \sigma^2$.*

$$\boxed{Z \sim \mathcal{N}\left(n \cdot \mu, n \cdot \sigma^2\right).}$$

- *Y follows a normal distribution when n is large enough with parameters $\mu_Y = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \mu$ and $\sigma_Z^2 = \frac{1}{n} \sum_{i=1}^{n} Var[X_i] = \frac{\sigma^2}{n}$.*

$$\boxed{Y \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).}$$

Random buses

The time you have to wait for a bus every day is uniformly distributed between 0 and 4 minutes. What is the probability you have to wait for more than or equal to 2.2 minutes for the bus on average in the next 300 days?

We may now fully use the central limit theorem.

- As $n = 300$ (big enough), we know that the average time you will wait for the bus in the next 300 days is normally distributed with mean μ and variance $\sigma^2/300$.
- The time you wait for a single bus is uniformly distributed with mean $\mu = \frac{0+4}{2} = 2$ minutes and variance $\sigma^2 = \frac{(4-0)^2}{12} = 4/3$ minutes^2.
- Combining, the average time T you wait for the bus follows $\mathcal{N}(2, \frac{4}{900})$.

Random buses (cont'd)

We are interested in $P(T > 2.2) = 1 - P(T \leq 2.2)$. First let us convert to the proper z value:

$$Z = \frac{X - \mu}{\sigma} = \frac{2.2 - 2}{2/30} = 3.$$

Looking at the z-table:

$$P(T \leq 2.2) = 0.9987 \implies P(T \geq 2.2) = 1 - 0.9987 = 0.0013.$$