Chapter 5. Multiple Random Variables

5.7: Limit Theorems

Slides (Google Drive) Alex Tsun Video (YouTube)

This is definitely one of the most important sections in the entire text! The Central Limit Theorem is used everywhere in statistics (hypothesis testing), and it also has its applications in computing probabilities. We'll see three results here, each getting more powerful and surprising.

If $X_1, \ldots, X_n$ are iid random variables with mean μ and variance σ^2, then we define the sample mean to be $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. We'll see the following results:

- The expectation of the sample mean $\mathbb{E}\left[\overline{X}_n\right]$ is exactly the true mean μ, and the variance $\mathsf{Var}\left(\overline{X}_n\right) = \sigma^2/n$ goes to 0 as you get more samples.
- (Law of Large Numbers) As $n \to \infty$, the sample mean $\overline{X}_n$ converges (in probability) to the true mean μ. That is, as you get more samples, you will be able to get an excellent estimate of μ.
- (Central Limit Theorem) In fact, $\overline{X_n}$ follows a *Normal distribution* as $n \to \infty$ (in practice n as low as 30 is good enough for this to be true). When we talk about the distribution of $\overline{X_n}$, this means: if we take n samples and take the sample mean, another n samples and take the sample mean, and so on, how will these sample means look in a histogram? This is crazy - regardless of what the distribution of X_i's were (discrete, continuous), their average will be approximately Normal! We'll see pictures and describe this more soon!

5.7.1 The Sample Mean

Before we start, we will define the sample mean of n random variables, and compute its mean and variance.

Definition 5.7.1: The Sample Mean + Properties

Let $X_1, X_2, \ldots, X_n$ be a sequence of iid (independent and identically distributed) random variables with mean μ and variance σ^2. The **sample mean** is:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

Further:

$$\mathbb{E}\left[\bar{X}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

Also, since the X_i's are independent:

$$\mathsf{Var}\left(\bar{X}_n\right) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathsf{Var}\left(X_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Again, none of this is "mind-blowing" to prove: we just used linearity of expectation and properties of variance to show this.

What is this saying? Basically, if you wanted to estimate the mean height of the U.S. population by sampling n people uniformly at random:

- In expectation, your sample average will be "on point" at $\mathbb{E}\left[\bar{X}_n\right] = \mu$. This even includes the case $n = 1$: if you just sample one person, on average, you will be correct. However, the variance is high.

- The variance of your estimate (the sample mean) for the true mean goes down (σ^2/n) as your sample size n gets larger. This makes sense right? If you have more samples, you have more confidence in your estimate because you are more "sure" (less variance).

In fact, as $n \to \infty$, the variance of the sample mean approaches 0. A distribution with mean μ and variance 0 is essentially the degenerate random variable that takes on μ with probability 1. We'll actually see that the Law of Large Numbers argues exactly that!

5.7.2 The Law of Large Numbers (LLN)

Using the fact that the variance is approaching 0 as $n \to \infty$, we can argue that, by averaging more and more samples ($n \to \infty$), we get a really good estimate of the true mean μ since the variance of the sample mean is $\sigma^2/n \to 0$ (as we showed earlier). Here is the formal mathematical statement:

Theorem 5.7.1: The Law of Large Numbers

Weak Law of Large Numbers (WLLN): Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with mean μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ **converges *in probability*** to μ. That is for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| > \epsilon\right) = 0$$

Strong Law of Large Numbers (SLLN): Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with mean μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ **converges *almost surely*** to μ. That is:

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

The SLLN implies the WLLN, but not vice versa. The difference is subtle and is basically swapping the limit and probability operations.

The proof the WLLN will be given in 6.1 when we prove Chebyshev's inequality, but the proof of the SLLN is out of the scope of this class and much harder to prove.

5.7.3 The Central Limit Theorem (CLT)

Theorem 5.7.2: The Central Limit Theorem (CLT)

Let $X_1, \dots X_n$ be a sequence of independent and identically distributed random variables with mean μ and (finite) variance σ^2. We've seen that the sample mean $\bar{X}_n$ has mean μ and variance $\frac{\sigma^2}{n}$. Then as $n \to \infty$, the following equivalent statements hold:

1. $\bar{X}_n \to \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
2. $\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \to \mathcal{N}(0, 1)$
3. $\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$. This is not "technically" correct, but is useful for applications.
4. $\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \to \mathcal{N}(0, 1)$

The mean or variance are not a surprise (we computed these at the beginning of these notes for any sample mean); the importance of the CLT is, regardless of the distribution of X_i's, the sample mean approaches a Normal distribution as $n \to \infty$.

We will prove the central limit theorem in 5.11 using MGFs, but take a second to appreciate this crazy result! The LLN say that as $n \to \infty$, the sample mean of iid variables $\overline{X}_n$ converges to μ. The CLT says that, as $n \to \infty$, the sample mean actually converges to a Normal distribution! For ***any*** original distribution of the X_i's (discrete or continuous), the average/sum will become approximately normally distributed.

If you're still having trouble with figuring out what "the distribution of the sample mean" means, that's completely normal (double pun!). Let's consider $n = 2$, so we just take the average of $X_1 + X_2$, which is $\frac{X_1+X_2}{2}$. The distribution of $X_1 + X_2$ means: if we repeatedly sample X_1, X_2 and add them, what might the density look like? For example, if $X_1, X_2 \sim \text{Unif}(0, 1)$ (continuous), we showed the density of $X_1 + X_2$ looked like a triangle. We figured out how to compute the PMF/PDF of the sum using convolution in 5.5, and the average is just dividing this by 2: $\frac{X_1+X_2}{2}$, which you can find the PMF/PDF by transforming RVs in 4.4. On the next page, you'll see exactly the CLT applied to these Uniform distributions. With $n = 1$, it looks (and is) Uniform. When $n = 2$, you get the triangular shape. And as n gets larger, it starts looking more and more like a Normal!

You'll see some examples below of how we start with some arbitrary distributions and how the density function of their mean becomes shaped like a Gaussian (you know how to compute the pdf of the mean now using convolution in 5.5 and transforming RV's in 4.4)!

On the next two pages, we'll see some visual "proof" of this surprising result!

Let's see the CLT applied to the (discrete) Uniform distribution.

- The first ($n = 1$) of the four graphs below shows a **discrete** $\frac{1}{29} \cdot \text{Unif}(0, 29)$ PMF in the dots (and a blue line with the curve of the normal distribution with the same mean and variance). That is, $\mathbb{P}(X = k) = \frac{1}{30}$ for each value in the range $\left\{0, \frac{1}{29}, \frac{2}{29}, \ldots, \frac{28}{29}, 1\right\}$.
- The second graph ($n = 2$) has the average of two of these distributions, again with a blue line with the curve of the normal distribution with the same mean and variance. Remember we expected this triangular distribution when summing either discrete or continuous Uniforms. (e.g., when summing two fair 6-sided die rolls, you're most likely to get a 7, and the probability goes down linearly as you approach 2 or 12. See the example in 5.5 if you forgot how we got this!
- The third ($n = 3$) and fourth ($n = 4$) have the average of 3 and 4 identically distributed random variables respectively, each of the distribution shown in the distribution in the first graph. We can see that as we average more, the sum approaches a normal distribution.

Again, if you don't believe me, you can compute the PMF yourself using convolution: first add two Unif$(0, 1)$, then convolve it with a third, and a fourth!

Despite this being a discrete random variable, when we take an average of many, there become increasingly many values we can get between 0 and 1. The average of these iid discrete rv's approaches a continuous Normal random variable even after just averaging 4 of them!

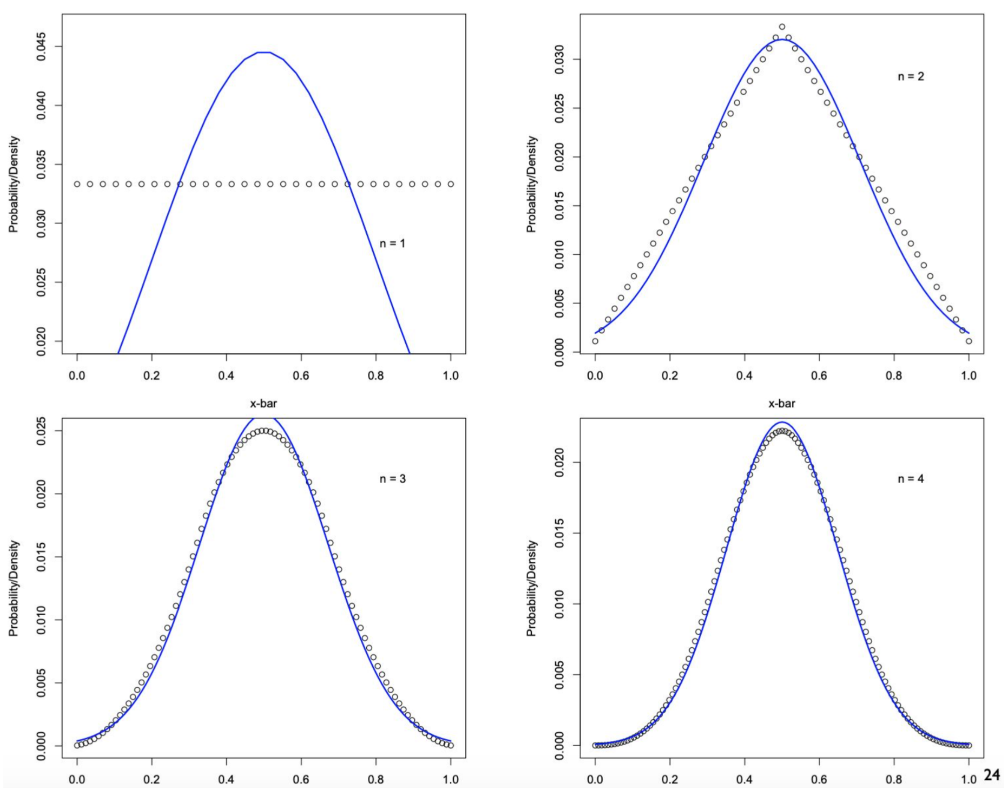
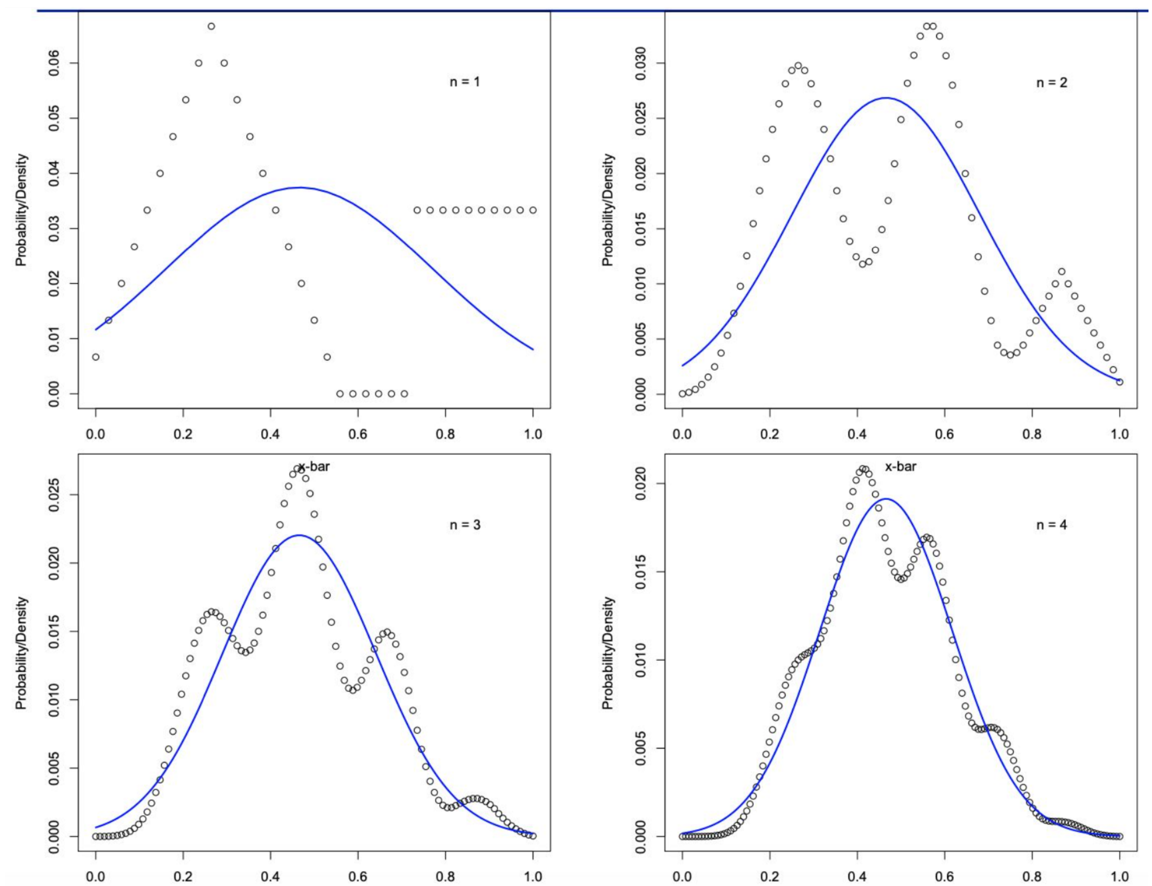


Image Credit: Larry Ruzzo (a previous University of Washington CSE 312 instructor).

You might still be skeptical, because the Uniform distribution is "nice" and already looked pretty "Normal" even with $n = 2$ samples. We now illustrate the same idea with a strange distribution shown in the first ($n = 1$) of the four graphs below, illustrated with the dots (instead of a "nice" uniform distribution). Even this crazy distribution nearly looks Normal after just averaging 4 of them. This is the power of the CLT!



What we are getting at here is that, regardless of the distribution, as we have more independent and identically distributed random variables, the average follows a Normal distribution (with the same mean and variance as the sample mean).

Now let's see how we can apply the CLT to problems! There were four different equivalent forms (just scaling/shifting) stated, but I find it easier to just look at the problem and decide what's best. Seeing examples is the best way to understand!

Example(s)

Let's consider the example of flipping a fair coin 40 times independently. What's the probability of getting between 15 to 25 heads? First compute this exactly and then give an approximation using the CLT.

Solution Define X to be the number of heads in the 40 flips. Then we have $X \sim \text{Bin}(n = 40, p = \frac{1}{2})$, so we just sum the Binomial PMF:

$$\mathbb{P}\left(15 \leq X \leq 25\right) = \sum_{k=15}^{25} \binom{40}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{40-k} \approx 0.9193.$$

Now, let's use the CLT. Since X can be thought of as the sum of 40 iid $\text{Ber}(\frac{1}{2})$ RVs, we can apply the CLT. We have $\mathbb{E}\left[X\right] = np = 40(\frac{1}{2}) = 20$ and $\mathsf{Var}\left(X\right) = np(1-p) = 40(\frac{1}{2})(1-\frac{1}{2}) = 10$. So we can use the approximation $X \approx \mathcal{N}(\mu = 20, \sigma^2 = 10)$.

This gives us the following good but not great approximation:

$$\begin{aligned}
\mathbb{P}\left(15 \leq X \leq 25\right) &\approx \mathbb{P}\left(15 \leq \mathcal{N}(20, 10) \leq 25\right) \\
&= \mathbb{P}\left(\frac{15-20}{\sqrt{10}} \leq Z \leq \frac{25-20}{\sqrt{10}}\right) && \text{[standardize]} \\
&\approx \mathbb{P}\left(-1.58 \leq Z \leq 1.58\right) \\
&= \Phi(1.58) - \Phi(-1.58) \\
&= 0.8862
\end{aligned}$$

$\square$

We'll see how to improve our approximation below!

5.7.4 The Continuity Correction

Notice that in the prior example in computing $\mathbb{P}\left(15 \leq X \leq 25\right)$, we sum over $25 - 15 + 1 = 11$ terms of the PMF. However, our integral $\mathbb{P}\left(15 \leq \mathcal{N}(20, 10) \leq 25\right)$ has width $25 - 15 = 10$. We'll always be off-by-one since the number of integers in $[a, b]$ is $(b - a) + 1$ (for integers $a \leq b$) and not $b - a$ (e.g., the number of integers between $[12, 15]$ is $(15 - 12) + 1 = 4 : \{12, 13, 14, 15\}$).

The continuity correction says we should add 0.5 in each direction. That is, we should ask for $\mathbb{P}\left(a - 0.5 \leq X \leq b + 0.5\right)$ instead so the width is $b - a + 1$ instead. Notice that if we do the final calculation, to approximate $\mathbb{P}\left(15 \leq X \leq 25\right)$ using the central limit theorem, now with the continuity correction, we get the following:

Example(s)

Use the continuity correction to get a better estimate than we did earlier for the coin problem.

Solution We'll apply the exact same steps, except changing the bounds from 15 and 25 to 14.5 and 25.5.

$$\begin{aligned}
\mathbb{P}(15 \leq X \leq 25) &\approx \mathbb{P}(14.5 \leq \mathcal{N}(20, 10) \leq 25.5) && \text{[apply continuity correction]} \\
&= \mathbb{P}\left(\frac{14.5 - 20}{\sqrt{10}} \leq Z \leq \frac{25.5 - 20}{\sqrt{10}}\right) \\
&\approx \mathbb{P}(-1.74 \leq Z \leq 1.74) \\
&= \Phi(1.74) - \Phi(-1.74) \\
&\approx 0.9182
\end{aligned}$$

Notice that this is much closer to the exact answer from the first part of the prior example (0.9193) than approximating with the central limit theorem without the continuity correction! □

Definition 5.7.2: The Continuity Correction

When approximating an integer-valued (**discrete**) random variable X with a continuous one Y (such as in the CLT), if asked to find a $\mathbb{P}(a \leq X \leq b)$ for integers $a \leq b$, you should compute $\mathbb{P}(a - 0.5 \leq Y \leq b + 0.5)$ so that the width of the interval being integrated is the same as the number of terms summed over $(b - a + 1)$. This is called the **continuity correction**.

Note: If you are applying the CLT to sums/averages of *continuous* RVs instead, you should **not** apply the continuity correction.

See the additional exercises below to get more practice with the CLT!

5.7.5 Exercises

1. Each day, the number of customers who come to the CSE 312 probability gift shop is approximately Poi(11). Approximate the probability that, after the quarter ends ($9 \times 7 = 63$ days), that we had over 700 customers.

   Solution: The total number of customers that come is $X = X_1 + \cdots + X_{63}$, where each $X_i \sim \text{Poi}(11)$ has $\mathbb{E}[X_i] = \mathsf{Var}(X_i) = \lambda = 11$ from the chart. By the CLT, $X \approx \mathcal{N}(\mu = 63 \cdot 11, \sigma^2 = 63 \cdot 11)$ (sum of the means and sum of the variances). Hence,

$$\begin{aligned}
\mathbb{P}(X \geq 700) &\approx \mathbb{P}(X \geq 699.5) && \text{[continuity correction]} \\
&\approx \mathbb{P}(\mathcal{N}(693, 693) \geq 699.5) && \text{[CLT]} \\
&= \mathbb{P}\left(Z \geq \frac{699.5 - 693}{\sqrt{693}}\right) && \text{[standardize]} \\
&= 1 - \Phi(0.2469) \\
&= 1 - 0.598 \\
&= 0.402
\end{aligned}$$

   Note that you could compute this exactly as well since you know the sum of iid Poissons is Poisson. In fact, $X \sim \text{Poi}(693)$ (the average rate in 63 days is 693 per 63 days), and you could do a sum which would be very annoying.

2. Suppose I have a flashlight which requires one battery to operate, and I have 18 identical batteries. I want to go camping for a week ($24 \times 7 = 168$) hours. If the lifetime of a single battery is Exp(0.1),

what's the probability my flashlight can operate for the entirety of my trip?

Solution: The total lifetime of the battery is $X = X_1 + \cdots + X_{18}$ where each $X_i \sim Exp(0.1)$ has $\mathbb{E}[X_i] = \frac{1}{0.1} = 10$ and $\mathsf{Var}(X_i) = \frac{1}{0.1^2} = 100$. Hence, $\mathbb{E}[X] = 180$ and $\mathsf{Var}(X) = 1800$ by linearity of expectation and since variance adds for independent rvs. In fact, $X \sim \text{Gamma}(r = 18, \lambda = 0.1)$, but we don't have a closed-form for its CDF. By the CLT, $X \approx \mathcal{N}(\mu = 180, \sigma^2 = 1800)$, so

$$
\begin{aligned}
\mathbb{P}(X \geq 168) &\approx \mathbb{P}(\mathcal{N}(180, 1800) \geq 168) && \text{[CLT]} \\
&= \mathbb{P}\left(Z \geq \frac{168 - 180}{\sqrt{1800}}\right) && \text{[standardize]} \\
&= 1 - \Phi(-0.28284) \\
&= \Phi(0.28284) && \text{[symmetry of Normal]} \\
&= 0.611
\end{aligned}
$$

Note that we don't use the continuity correction here because the RV's we are summing are already continuous RVs.