Central Limit Theorem and the Law of Large Numbers
Class 6b, TF, AID-M
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Understand the statement of the law of large numbers.

2. Understand the statement of the central limit theorem.

3. Be able to use the central limit theorem to approximate probabilities of averages and sums of independent identically-distributed random variables.

2 Introduction

We all understand intuitively that the average of many measurements of the same unknown quantity tends to give a better estimate than a single measurement. Intuitively, this is because the random error of each measurement cancels out in the average. In these notes we will make this intuition precise in two ways: the law of large numbers (LoLN) and the central limit theorem (CLT).

Briefly, both the law of large numbers and central limit theorem are about many independent samples from same distribution. The LoLN tells us two things:

1. The average of many independent samples is (with high probability) close to the mean of the underlying distribution.

2. The density histogram of many independent samples is (with high probability) close to the graph of the density of the underlying distribution.

To be absolutely correct mathematically we need to make these statements more precise, but as stated they are a good way to think about the law of large numbers.

The central limit theorem says that the sum or average of many independent copies of a random variable is approximately a normal random variable. The CLT goes on to give precise values for the mean and standard deviation of the normal variable.

These are both remarkable facts. Perhaps just as remarkable is the fact that often in practice n does not have to be all that large.

2.1 There is more to experimentation than mathematics

The mathematics of the LoLN says that the average of a lot of independent samples from a random variable will almost certainly approach the mean of the variable. The mathematics cannot tell us if the tool or experiment is producing data worth averaging. For example, if the measuring device is defective or poorly calibrated then the average of many measurements will be a highly accurate estimate of the wrong thing! This is an example of systematic error or sampling bias, as opposed to the random error controlled by the law of large numbers.

3 The law of large numbers

Suppose X_1, X_2, …, X_n are independent random variables with the same underlying distribution. In this case, we say that the X_i are independent and identically-distributed, or i.i.d. In particular, the X_i all have the same mean μ and standard deviation σ.

Let $\overline{X}_n$ be the average of $X_1, \ldots, X_n$:

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that $\overline{X}_n$ is itself a random variable. The law of large numbers and central limit theorem tell us about the value and distribution of $\overline{X}_n$, respectively.

LoLN: As n grows, the probability that $\overline{X}_n$ is close to μ goes to 1.

CLT: As n grows, the distribution of $\overline{X}_n$ converges to the normal distribution $N(\mu, \sigma^2/n)$.

Before giving a more formal statement of the LoLN, let's unpack its meaning through a concrete example (we'll return to the CLT later on).

Example 1. Averages of Bernoulli random variables
Suppose each X_i is an independent flip of a fair coin, so $X_i \sim \text{Bernoulli}(0.5)$ and $\mu = 0.5$. Then $\overline{X}_n$ is the proportion of heads in n flips, and we expect that this proportion is close to 0.5 for large n. Randomness being what it is, this is not guaranteed; for example we could get 1000 heads in 1000 flips, though the probability of this occurring is very small.

So our intuition translates to: with high probability the sample average $\overline{X}_n$ is close to the mean 0.5 for large n. We'll demonstrate by doing some calculations in R. You can find the code used for 'class 6 prep' in the usual place on our website.

To start we'll look at the probability of being within 0.1 of the mean. We can express this probability as

$$P(|\overline{X}_n - 0.5| < 0.1) \quad \text{or equivalently} \quad P(0.4 \le \overline{X}_n \le 0.6)$$

The law of large numbers says that this probability goes to 1 as the number of flips n gets large. Our R code produces the following values for $P(0.4 \le \overline{X}_n \le 0.6)$.

$n = 10$:	`pbinom(6, 10, 0.5) - pbinom(3, 10, 0.5)`	= `0.65625`
$n = 50$:	`pbinom(30, 50, 0.5) - pbinom(19, 50, 0.5)`	= 0.8810795
$n = 100$:	`pbinom(60, 100, 0.5) - pbinom(39, 100, 0.5)`	= 0.9647998
$n = 500$:	`pbinom(300, 500, 0.5) - pbinom(199, 500, 0.5)`	= 0.9999941
$n = 1000$:	`pbinom(600, 1000, 0.5) - pbinom(399, 1000, 0.5)`	= 1

As predicted by the LoLN the probability goes to 1 as n grows.

We redo the computations to see the probability of being within 0.01 of the mean. Our R code produces the following values for $P(0.49 \le \overline{X}_n \le 0.51)$.

$n = 10$:	`pbinom(5, 10, 0.5) - pbinom(4, 10, 0.5)`	$= 0.2460937$
$n = 100$:	`pbinom(51, 100, 0.5) - pbinom(48, 100, 0.5)`	$= 0.2356466$
$n = 1000$:	`pbinom(510, 1000, 0.5) - pbinom(489, 1000, 0.5)`	$= 0.49334$
$n = 10000$:	`pbinom(5100, 10000, 0.5) - pbinom(4899, 10000, 0.5)`	$= 0.9555742$

Again we see the probability of being close to the mean going to 1 as n grows. Since 0.01 is smaller than 0.1 it takes larger values of n to raise the probability to near 1.

This convergence of the probability to 1 is the LoLN in action! Whenever you're confused, it will help you to keep this example in mind. So we see that the LoLN says that with high probability the average of a large number of independent trials from the same distribution will be very close to the underlying mean of the distribution. Now we're ready for the formal statement.

3.1 Formal statement of the law of large numbers

Theorem (Law of Large Numbers): Suppose X_1, X_2, …, X_n, … are i.i.d. random variables with mean μ. For each n, let $\overline{X}_n$ be the average of the first n variables. Then for any $a > 0$, we have

$$\lim_{n \to \infty} P(|\overline{X}_n - \mu| < a) \, = \, 1.$$

This says precisely that as n increases the probability of being within a of the mean goes to 1. Think of a as a small tolerance of error from the true mean μ.

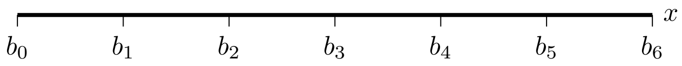
Looking back at Example 1, we see that for tosses of a fair coin: If we choose the number of tosses $n = 500$, then with probability $p = 0.99999$, the experimental frequency of heads $\overline{X}_n$ will be within $a = 0.1$ of 0.5. In words, this tells us that, on average, only 1 in 100,000 experiments will produce an experimental frequency outside this range. If we decrease the tolerance a and/or increase the probability p, then n will need to be larger.

4 Histograms

We can summarize multiple samples $x_1, \ldots, x_n$ of a random variable in a histogram. Here we want to carefully construct histograms so that they resemble the area under the pdf. We will then see how the LoLN applies to histograms.

The step-by-step instructions for constructing a density or frequency histogram are as follows.

1. Pick an interval of the real line and divide it into m intervals, with endpoints b_0, b_1, …, b_m. Usually these are equally sized, so let's assume this to start.



Six equally-sized bins

Each of the intervals is called a bin. For example, in the figure above the first bin is $[b_0, b_1]$ and the last bin is $[b_5, b_6]$. Each bin has a bin width, e.g. $b_1 - b_0$ is the first bin width. Usually the bins all have the same width, called the bin width of the histogram.

2. Place each x_i into the bin that contains its value. If x_i lies on the boundary of two bins, we'll put it in the left bin (this is the R default, though it can be changed).

3. To draw a frequency histogram: put a vertical bar above each bin. The height of the bar should equal the number of x_i in the bin.

4. To draw a density histogram: put a vertical bar above each bin. The area of the bar should equal the fraction of all data points that lie in the bin.

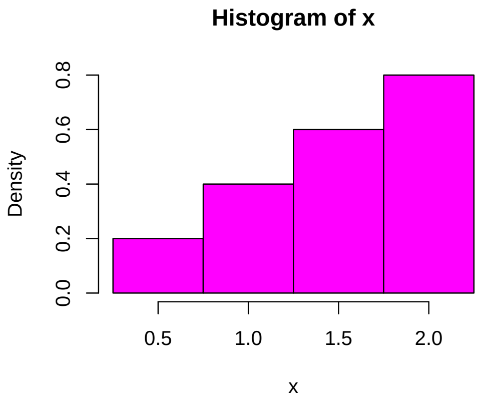
Notes:
1. When all the bins have the same width, the frequency histogram bars have area proportional to the count. So the density histogram results from simply by dividing the height of each bar by the total area of the frequency histogram. Ignoring the vertical scale, the two histograms look identical.

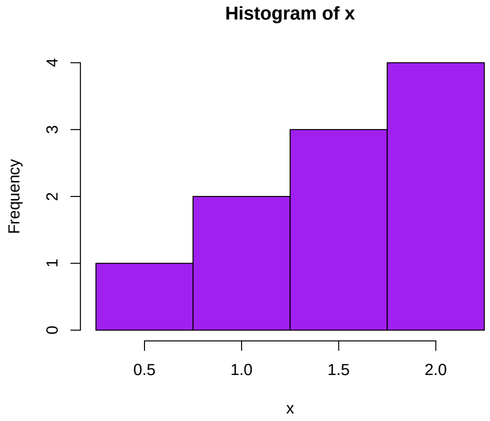
2. Caution: if the bin widths differ, the frequency and density histograms may look very different. There is an example of this below. Don't let anyone fool you by manipulating bin widths to produce a histogram that suits their mischievous purposes!

In 18.05, we'll stick with equally-sized bins. In general, we prefer the density histogram since its vertical scale is the same as that of the pdf.

Examples. Here are some examples of histograms, all with the data [0.5,1,1,1.5,1.5,1.5,2,2,2,2]. The R code that drew them is in the file 'class6-prep-b.r'. You can find it in the usual place on our website.

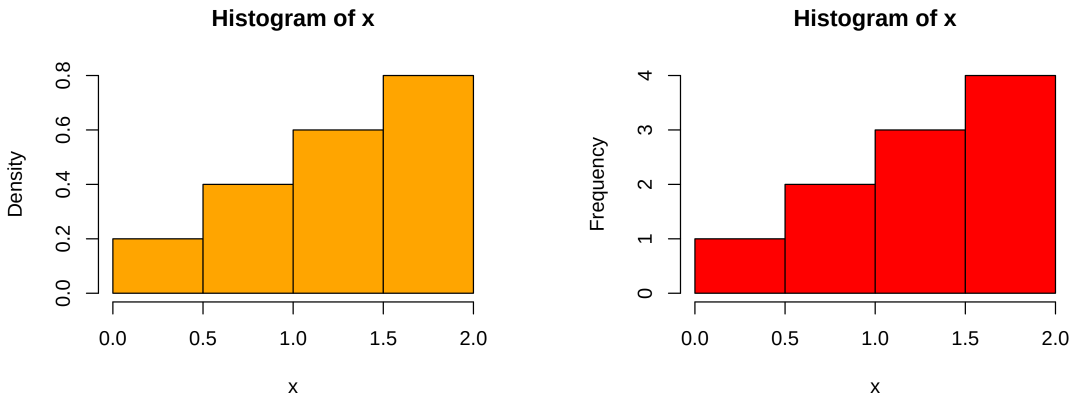
1. Here the density and frequency plots look the same but have different vertical scales.





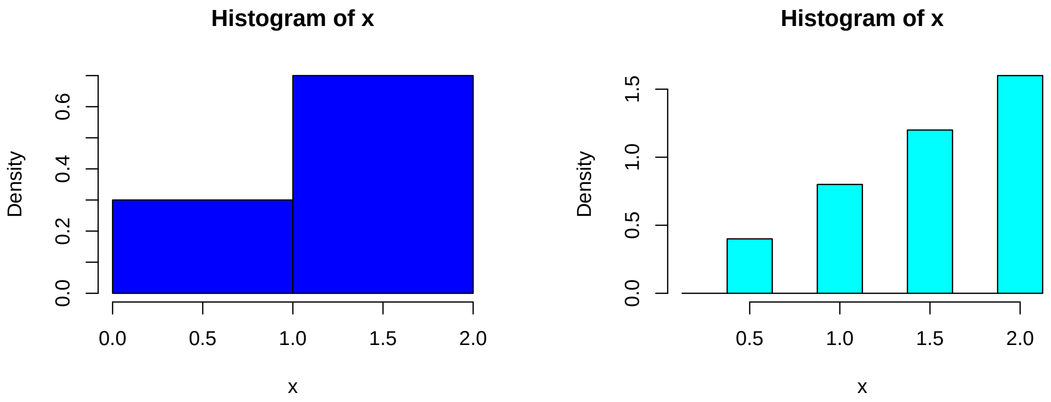
Bins centered at 0.5, 1, 1.5, 2, i.e. width 0.5, bounds at 0.25, 0.75, 1.25, 1.75, 2.25.

2. Note the values are all on the bin boundaries and are put into the left-hand bin. That is, the bins are right-closed, e.g the first bin is for values in the right-closed interval $(0, 0.5]$.



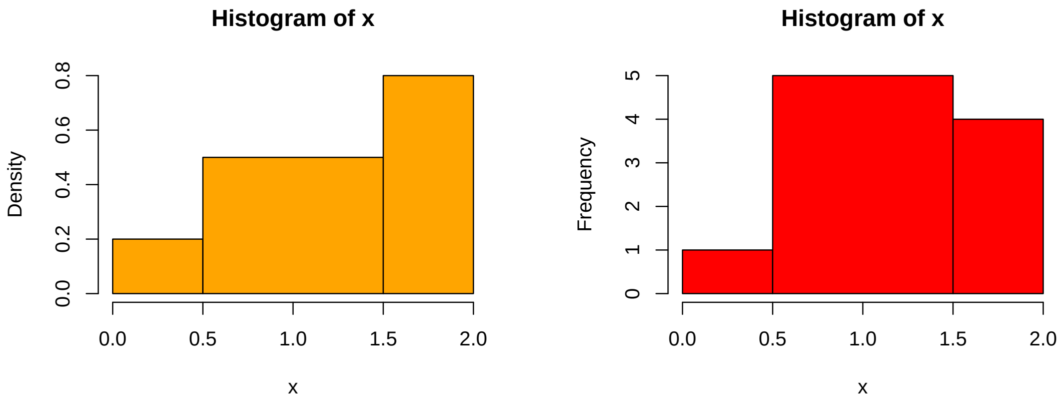
Bin bounds at 0, 0.5, 1, 1.5, 2.

3. Here we show density histograms based on different bin widths. Note that the scale keeps the total area equal to 1. The gaps are bins with zero counts.



Left: wide bins; Right: narrow bins.

4. Here we use unqual bin widths, so the density and frequency histograms look different



Don't be fooled! These are based on the same data.

The density histogram is the better choice with unequal bin widths. In fact, R will complain if you try to make a frequency histogram with unequal bin widths. Compare the frequency histogram with unequal bin widths with all the other histograms we drew for this data. It clearly looks different. What happened is that by combining the data in bins $(0.5, 1]$ and $(1, 1.5]$ into one bin $(0.5, 1.5)$ we effectively made the height of both smaller bins greater.

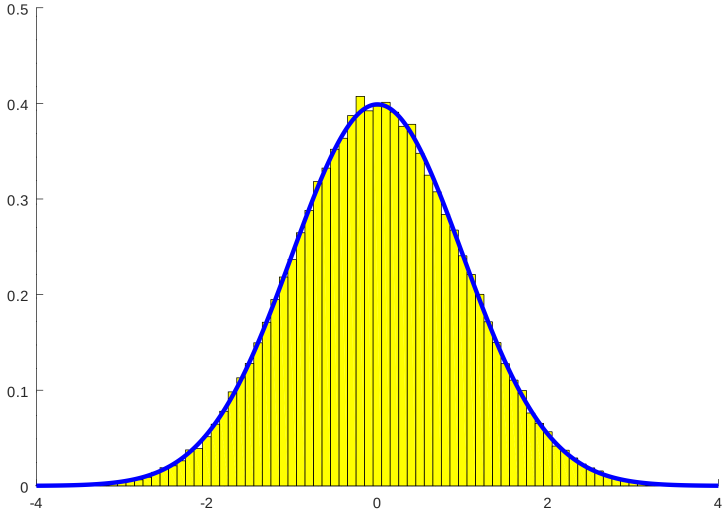
The reason the density histogram is nice is discussed in the next section.

4.1 The law of large numbers and histograms

The law of large number has an important consequence for density histograms.

LoLN for histograms: With high probability the density histogram of a large number of samples from a distribution is a good approximation of the graph of the underlying pdf $f(x)$ over the range of the histogram.

Let's illustrate this by generating a density histogram with bin width 0.1 from 100000 draws from a standard normal distribution. As you can see, the density histogram very closely tracks the graph of the standard normal pdf $\phi(z)$.



Density histogram of 10000 draws from a standard normal distribution, with $\phi(z)$ in blue.

5 The Central Limit Theorem

We now prepare for the statement of the CLT.

5.1 Standardization

Given a random variable X with mean μ and standard deviation σ, we define its standardization of X as the new random variable

$$Z = \frac{X - \mu}{\sigma}.$$

Note that Z has mean 0 and standard deviation 1. Note also that if X has a normal distribution, then the standardization of X is the standard normal distribution Z with mean 0 and variance 1. This explains the term 'standardization' and the notation of Z above.

5.2 Statement of the Central Limit Theorem

Suppose X_1, X_2, …, X_n, … are i.i.d. random variables each having mean μ and standard deviation σ. For each n, let S_n denote the sum and let $\overline{X}_n$ be the average of $X_1, \ldots, X_n$.

$$\begin{aligned} S_n &= X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \\ \overline{X}_n &= \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{S_n}{n}. \end{aligned}$$

The properties of mean and variance show

$$\begin{aligned} E[S_n] &= n\mu, & \text{Var}(S_n) &= n\sigma^2, & \sigma_{S_n} &= \sqrt{n}\,\sigma \\ E[\overline{X}_n] &= \mu, & \text{Var}(\overline{X}_n) &= \frac{\sigma^2}{n}, & \sigma_{\overline{X}_n} &= \frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Since they are multiples of each other, S_n and $\overline{X}_n$ have the same standardization

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

Central Limit Theorem: For large n,

$$\overline{X}_n \approx \text{N}(\mu, \sigma^2/n), \qquad S_n \approx \text{N}(n\mu, n\sigma^2), \qquad Z_n \approx \text{N}(0,1).$$

Notes: 1. In words: $\overline{X}_n$ is approximately a normal distribution with the same mean as X but a smaller variance.
2. S_n is approximately normal.
3. Standardized $\overline{X}_n$ and S_n are approximately standard normal.

The central limit theorem allows us to approximate a sum or average of i.i.d random variables by a normal random variable. This is extremely useful because it is usually easy to do computations with the normal distribution.

A precise statement of the CLT is that the cdf's of Z_n converge to $\Phi(z)$:

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z).$$

The proof of the Central Limit Theorem is more technical than we want to get in 18.05. It is accessible to anyone with a decent calculus background.

5.3 Standard Normal Probabilities

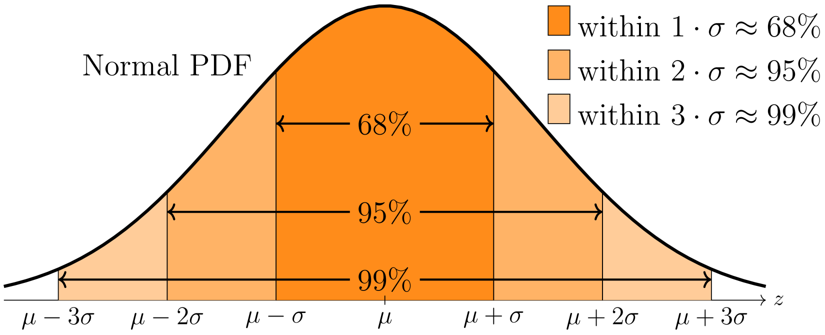
To apply the CLT, we will want to have some normal probabilities at our fingertips. The following probabilities appeared in Class 5. Let $Z \sim N(0, 1)$, a standard normal random variable. Then with rounding we have:

1. $P(|Z| < 1) \approx 0.68$
2. $P(|Z| < 2) \approx 0.95$; more precisely $P(|Z| < 1.96) \approx 0.95$.
3. $P(|Z| < 3) \approx 0.997$

These numbers are easily computed in R using `pnorm`. However, they are well worth remembering as rules of thumb. You should think of them as:

1. The probability that a normal random variable is within 1 standard deviation of its mean is 0.68.
2. The probability that a normal random variable is within 2 standard deviations of its mean is 0.95.
3. The probability that a normal random variable is within 3 standard deviations of its mean is 0.997.

This is shown graphically in the following figure.



Claim: From these numbers we can derive:

1. $P(Z < 1) \approx 0.84$
2. $P(Z < 2) \approx 0.977$
3. $P(Z < 3) \approx 0.999$

Proof: We know $P(|Z| < 1) = 0.68$. The remaining probability of 0.32 is in the two regions $Z > 1$ and $Z < -1$. These regions are referred to as the right-hand tail and the left-hand tail respectively. By symmetry each tail has area 0.16. Thus,

$$P(Z < 1) = P(|Z| < 1) + P(\text{left-hand tail}) = 0.84$$

The other two cases are handled similarly.

5.4 Applications of the CLT

Example 2. Flip a fair coin 100 times. Estimate the probability of more than 55 heads.

Solution: Let X_j be the result of the j^{th} flip, so $X_j = 1$ for heads and $X_j = 0$ for tails. The total number of heads is

$$S = X_1 + X_2 + \ldots + X_{100}.$$

We know $E[X_j] = 0.5$ and $\text{Var}(X_j) = 1/4$. Since $n = 100$, we have

$$E[S] = 50, \quad \text{Var}(S) = 25 \quad \text{and} \quad \sigma_S = 5.$$

The central limit theorem says that the standardization of S is approximately $\text{N}(0,1)$. The question asks for $P(S > 55)$. Standardizing and using the CLT we get

$$P(S > 55) = P\left(\frac{S-50}{5} > \frac{55-50}{5}\right) \approx P(Z > 1) = 0.16.$$

Here Z is a standard normal random variable and $P(Z > 1) = 1 - P(Z < 1) \approx 0.16$.

Example 3. Estimate the probability of more than 220 heads in 400 flips.

Solution: This is nearly identical to the previous example. Now $\mu_S = 200$ and $\sigma_S = 10$ and we want $P(S > 220)$. Standardizing and using the CLT we get:

$$P(S > 220) = P\left(\frac{S-\mu_S}{\sigma_S} > \frac{220-200}{10}\right) \approx P(Z > 2) = 0.025.$$

Again, $Z \sim \text{N}(0,1)$ and the rules of thumb show $P(Z > 2) = 0.025$.

Note: Even though $55/100 = 220/400$, the probability of more than 55 heads in 100 flips is larger than the probability of more than 220 heads in 400 flips. This is due to the LoLN and the larger value of n in the latter case.

Example 4. Estimate the probability of between 40 and 60 heads in 100 flips.

Solution: As in the first example, $E[S] = 50$, $\text{Var}(S) = 25$ and $\sigma_S = 5$. So

$$P(40 \le S \le 60) = P\left(\frac{40-50}{5} \le \frac{S-50}{5} \le \frac{60-50}{5}\right) \approx P\left(-2 \le Z \le 2\right)$$

We can compute the right-hand side using our rule of thumb. For a more accurate answer we use R:

$$\texttt{pnorm(2) - pnorm(-2)} = 0.954\ldots$$

Recall that in Section 3 we used the binomial distribution to compute an answer of 0.965.... So our approximate answer using CLT is off by about 1%.

Think: Would you expect the CLT method to give a better or worse approximation of $P(200 < S < 300)$ with $n = 500$?

We encourage you to check your answer using R.

Example 5. Polling. When taking a political poll the results are often reported as a number with a margin of error. For example 52% ± 3% favor candidate A. The rule of thumb is that if you poll n people then the margin of error is $\pm 1/\sqrt{n}$. We will now see exactly what this means and that it is an application of the central limit theorem.

Suppose there are 2 candidates A and B. Suppose further that the fraction of the population who prefer A is p_0. That is, if you ask a random person who they prefer then the probability they'll answer A is p_o

To run the poll a pollster selects n people at random and asks 'Do you support candidate A or candidate B?' Thus we can view the poll as a sequence of n independent Bernoulli(p_0) trials, X_1, X_2, ..., X_n, where X_i is 1 if the i$^{\text{th}}$ person prefers A and 0 if they prefer B. The fraction of people polled that prefer A is just the average $\overline{X}$.

We know that each $X_i \sim$ Bernoulli(p_0) so,

$$E[X_i] = p_0 \quad \text{ and } \sigma_{X_i} = \sqrt{p_0(1-p_0)}.$$

Therefore, the central limit theorem tells us that

$$\overline{X} \approx \text{N}(p_0, \sigma^2/n), \qquad \text{where } \sigma = \sqrt{p_0(1-p_0)}.$$

In a normal distribution 95% of the probability is within 2 standard deviations of the mean. This means that in 95% of polls of n people the sample mean $\overline{X}$ will be within $2\sigma/\sqrt{n}$ of the true mean p_0. The final step is to note that for any value of p_0 we have $\sigma \leq 1/2$. (It is an easy calculus exercise to see that 1/4 is the maximum value of $\sigma^2 = p_0(1-p_0)$.) This means that we can conservatively say that in 95% of polls of n people the sample mean $\overline{X}$ is within $1/\sqrt{n}$ of the true mean. The frequentist statistician then takes the interval $\overline{X} \pm 1/\sqrt{n}$ and calls it the 95% confidence interval for p_0.

A word of caution: it is tempting and common, **but wrong**, to think that there is a 95% probability the true fraction p_0 is in a particular confidence interval. This is subtle, but the error is the same one as thinking you have a disease if a test with a 95% true positive rate comes back positive. We will go into this in much more detail when we learn about confidence intervals.

5.5 Why use the CLT

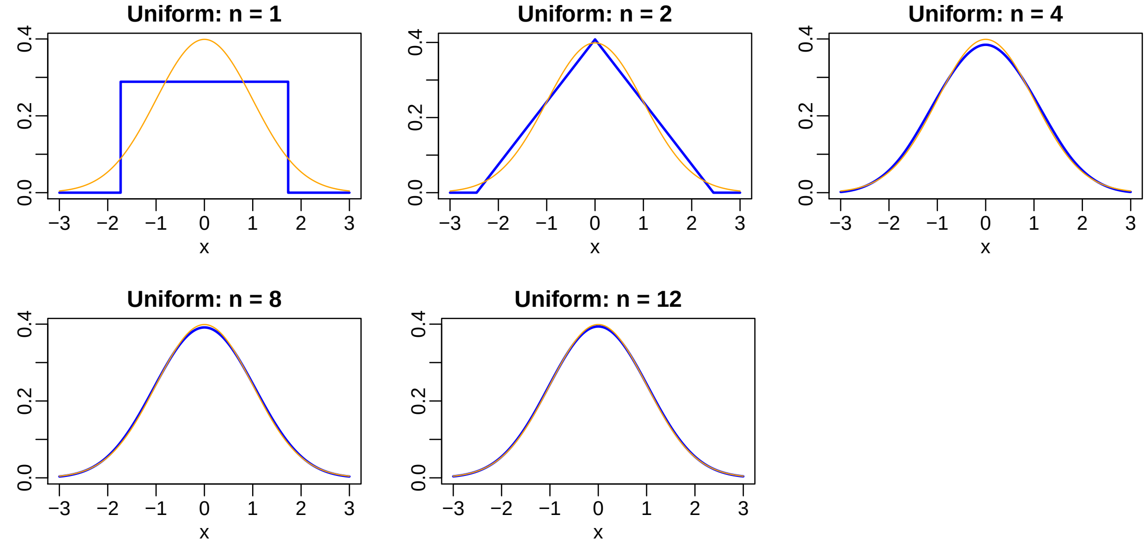
Since the probabilities in the above examples can be computed exactly using the binomial distribution, you may be wondering what is the point of finding an approximate answer using the CLT. In fact, we were only able to compute these probabilities exactly because the X_i were Bernoulli and so the sum S was binomial. In general, the distribution of the X_i may not be familiar, or may not even be known, so you will not be able to compute the probabilities for S exactly. It can also happen that the exact computation is possible in theory but too computationally intensive in practice, even for a computer. The power of the CLT is that it applies whenever X_i has a mean and a variance. Though the CLT applies to many distributions, we will see in the next section that some distributions require larger n for the approximation to be a good one.

5.6 How big does n have to be to apply the CLT?

Short answer: often, not that big.

The following sequences of pictures show the convergence of averages to a normal distribution.

First we show the standardized average of n i.i.d. **uniform** random variables with $n = 1, 2, 4, 8, 12$. The pdf of the average is in blue and the standard normal pdf is in red. By the time $n = 12$ the fit between the standardized average and the true normal looks very good.



Uniform: n = 1

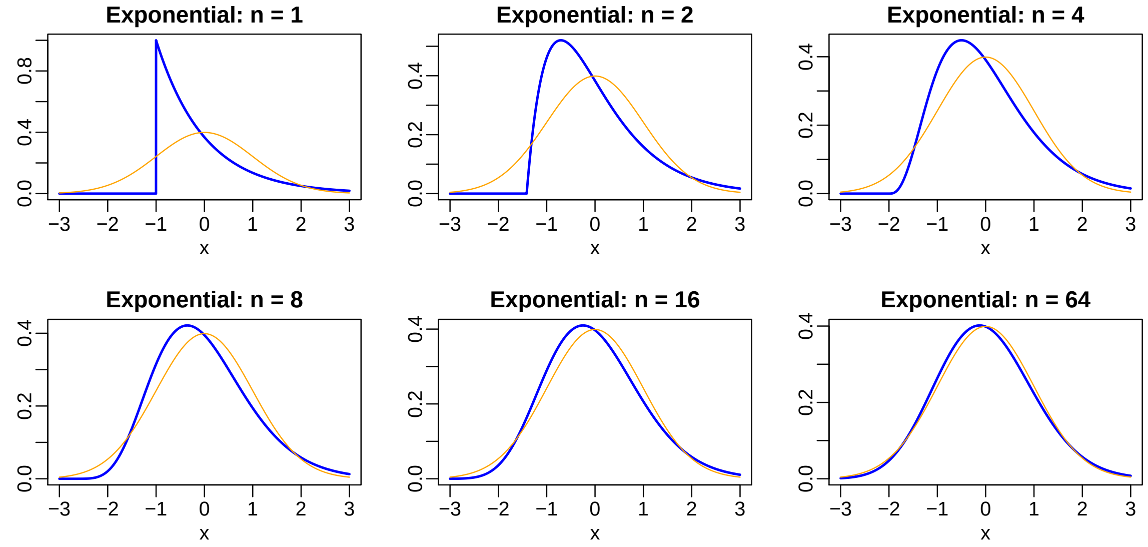
Uniform: n = 2

Uniform: n = 4

Uniform: n = 8

Uniform: n = 12

Next we show the standardized average of n i.i.d. **exponential** random variables with $n = 1, 2, 4, 8, 16, 64$. Notice that this asymmetric density takes more terms to converge to the normal density.



Exponential: n = 1

Exponential: n = 2

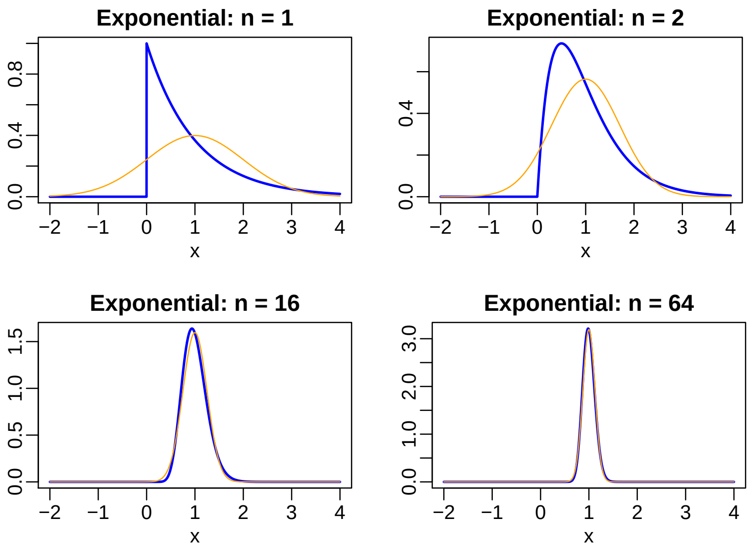
Exponential: n = 4

Exponential: n = 8

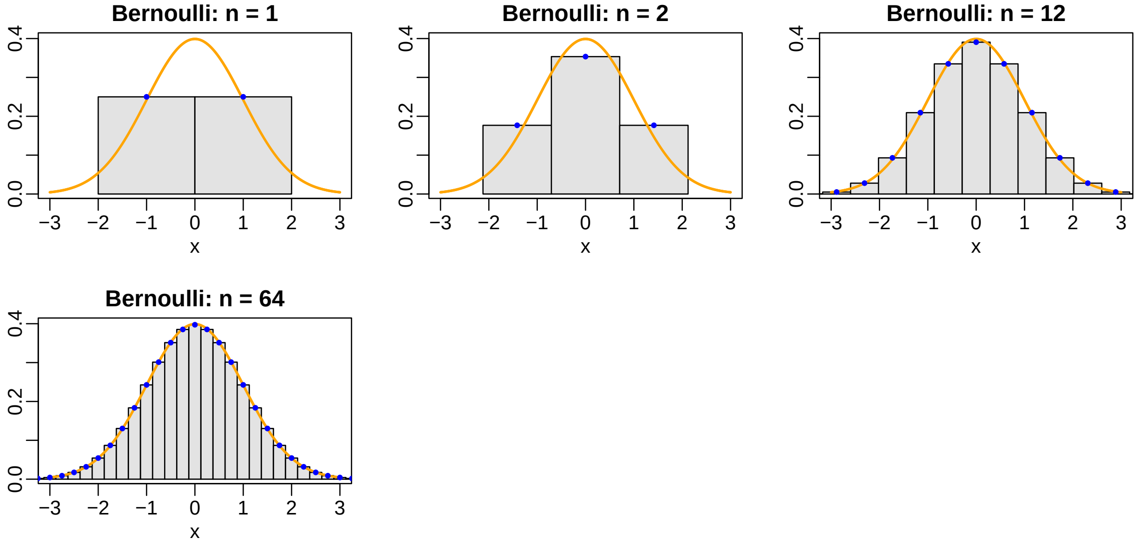
Exponential: n = 16

Exponential: n = 64

Next we show the (non-standardized) average of n exponential random variables with $n = 1, 2, 4, 16, 64$. Notice how this standard deviation shrinks as n grows, resulting in a spikier (more peaked) density.

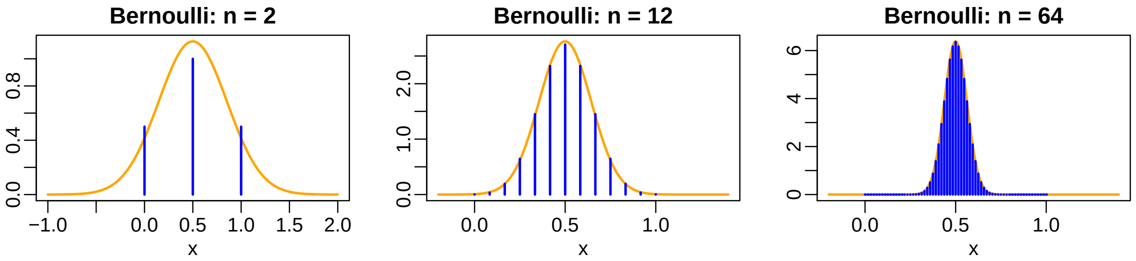


The central limit theorem works for discrete variables also. Here is the standardized average of n i.i.d. Bernoulli(0.5) random variables with $n = 1, 2, 12, 64$. Notice that as n grows, the average can take more values, which allows the discrete distribution to 'fill in' the normal density.



Note. In order to put the binomial (sum of Bernoulli) and normal distribution on the same axes, we had to convert the binomial probability mass function to a density. We did this by making it a bar graph with bars centered on each value and with bar width equal to the distance between values. Then the height of each bar is chosen so that the area equals the probability of the corresponding value.

Finally we show the (non-standardized) average of n Bernoulli(0.5) random variables, with $n = 4, 12, 64$. Notice how the standard deviation gets smaller resulting in a spikier (more peaked) density. (In these figures, rather than plotting colored bars, we made the bars white and only plotted a blue line at the center of each bar.



Note by Prof. Mayer: This script is the work of the original authors (Jeremy Orloff and Jonathan Bloom), which appear in the title and all praise for this work must be to them. The only modifications done to fit it to ”Theoretical fundamentals of AI and Data Science” (DIT, Faculty AI, Study course AIN-B) are leaving out certain topics and changing the class numbering.

The complete original script (and more: slides, exams, question sheets, and solutions) can be downloaded at MIT OpenCourseWare: Introduction To Probability And Statistics, 18.05, Spring 2014, Undergraduate

Prof. Mayer thanks Jeremy Orloff and Jonathan Bloom not only for making the course material open to the public, but also for giving him access to the original LaTeX sources to tailor the material better to the DITs needs. Without their help, this course would by far not have the quality of its current state.