Today:
Normal approximation to the probability histogram
Continuity correction
The Central Limit Theorem

Tomorrow:
Sampling Surveys (Chapter 19)

Reminder: Requests for regrades need to be in my hand no later than the end of lecture tomorrow.

Office hours today: 1-2 pm, 393 Evans

Probability histograms:

- One bin per possible value
- Height of bin is the probability that value occurs
- Theoretical only - these are actual probabilities, not observed proportions.

Probability histograms:

- One bin per possible value
- Height of bin is the probability that value occurs
- Theoretical only - these are actual probabilities, not observed proportions.

Probability histograms:

- One bin per possible value
- Height of bin is the probability that value occurs
- Theoretical only - these are actual probabilities, not observed proportions.

Probability histograms:

- One bin per possible value
- Height of bin is the probability that value occurs
- Theoretical only - these are actual probabilities, not observed proportions.

Probability histograms:

- One bin per possible value
- Height of bin is the probability that value occurs
- Theoretical only - these are actual probabilities, not observed proportions.

Yesterday:

- Drew a probability histogram for 4 tosses of a fair coin, counting the number of heads.
- Drew empirical histograms where we repeated the process many times.
- The empirical histogram converged to the probability histogram as the process was repeated more and more often.

Yesterday:

- Drew a probability histogram for 4 tosses of a fair coin, counting the number of heads.
- Drew empirical histograms where we repeated the process many times.
- The empirical histogram converged to the probability histogram as the process was repeated more and more often.

Yesterday:

- Drew a probability histogram for 4 tosses of a fair coin, counting the number of heads.
- Drew empirical histograms where we repeated the process many times.
- The empirical histogram converged to the probability histogram as the process was repeated more and more often.

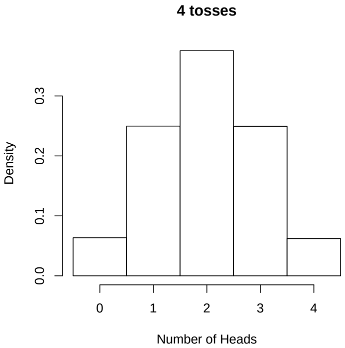
Question: What happens to the probability histogram when we increase the number of tosses?

Consider tossing a fair coin, counting the number of heads.

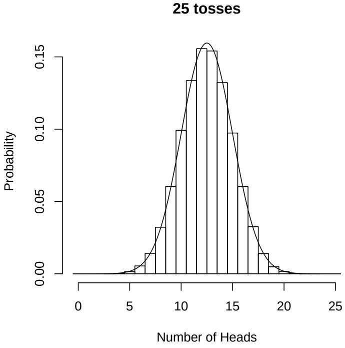
The box is [0 1]

Mean of box: $\frac{0+1}{2} = 0.5$

SD of box: $(1-0)\sqrt{\frac{1}{2} \times \frac{1}{2}} = 0.5$

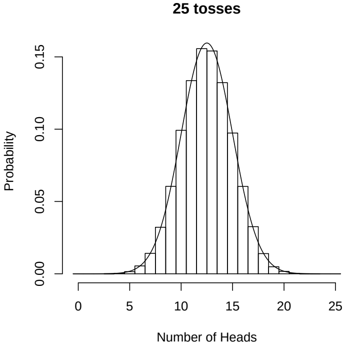


Expected Value = $4 \times .5 = 2$
Standard Error = $\sqrt{4} \times .5 = 1$



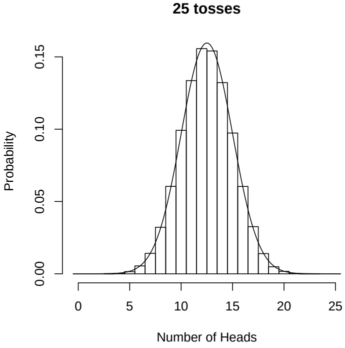
Expected Value = $25 \times .5 = 12.5$
Standard Error = $\sqrt{25} \times .5 = 2.5$

P(exactly 15 heads) = 0.0974 using binomial formula
P(exactly 15 heads) = 0.0967 using normal approximation



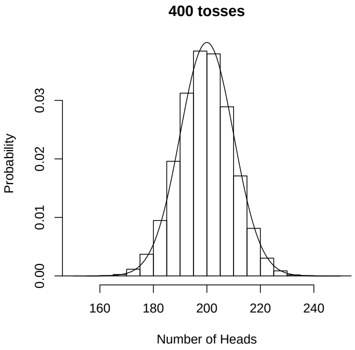
Expected Value = $25 \times .5 = 12.5$
Standard Error = $\sqrt{25} \times .5 = 2.5$

P(exactly 15 heads) = 0.0974 using binomial formula
P(exactly 15 heads) = 0.0967 using normal approximation



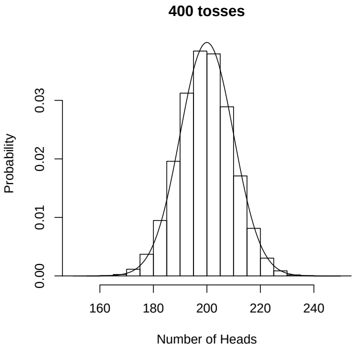
Expected Value = $25 \times .5 = 12.5$
Standard Error = $\sqrt{25} \times .5 = 2.5$

P(exactly 15 heads) = 0.0974 using binomial formula
P(exactly 15 heads) = 0.0967 using normal approximation



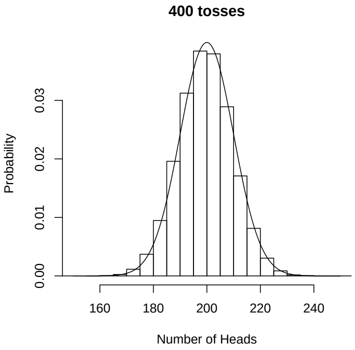
Expected Value = $400 \times .5 = 200$
Standard Error = $\sqrt{400} \times .5 = 10$

P(more than 220 heads) = 0.0201 using binomial formula
P(more than 220 heads) = 0.0227 using normal approximation



Expected Value = $400 \times .5 = 200$
Standard Error = $\sqrt{400} \times .5 = 10$

P(more than 220 heads) = 0.0201 using binomial formula
P(more than 220 heads) = 0.0227 using normal approximation



Expected Value = $400 \times .5 = 200$
Standard Error = $\sqrt{400} \times .5 = 10$

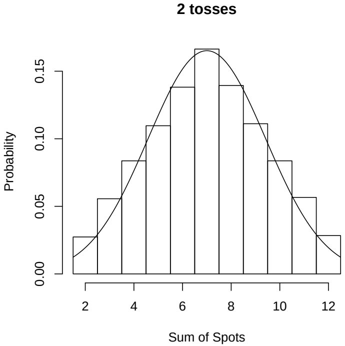
P(more than 220 heads) = 0.0201 using binomial formula
P(more than 220 heads) = 0.0227 using normal approximation

Consider rolling a fair die.

Box: [1 2 3 4 5 6]

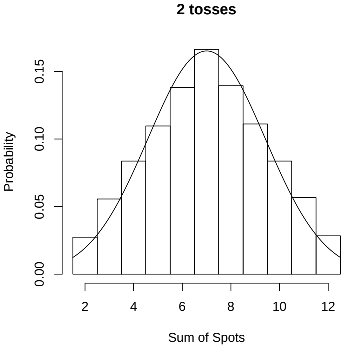
Mean: 3.5

SD: 1.71



Expected Value = $2 \times 3.5 = 7$
Standard Error = $\sqrt{2} \times 1.71 = 2.42$

You probably wouldn't use the normal approximation here. It's a little too far off, and the actual probabilities aren't too hard to sort out.



Expected Value = $2 \times 3.5 = 7$
Standard Error = $\sqrt{2} \times 1.71 = 2.42$

You probably wouldn't use the normal approximation here. It's a little too far off, and the actual probabilities aren't too hard to sort out.

Let's say you were silly and wanted to anyway. Let's find P(roll 8 or 9).

Actual probability = $\frac{5}{36} + \frac{4}{36} = 0.25$

lower end: $z = \frac{7.5-7}{2.42} = 0.20$

upper end: $z = \frac{9.5-7}{2.42} = 1.05$

area = $\frac{.7063-.1585}{2} = .2739$

really precise area using computer normal tables = 0.2677

Let's say you were silly and wanted to anyway. Let's find P(roll 8 or 9).

Actual probability = $\frac{5}{36} + \frac{4}{36} = 0.25$

lower end: $z = \frac{7.5-7}{2.42} = 0.20$

upper end: $z = \frac{9.5-7}{2.42} = 1.05$

area = $\frac{.7063-.1585}{2} = .2739$

really precise area using computer normal tables = 0.2677

Let's say you were silly and wanted to anyway. Let's find P(roll 8 or 9).

Actual probability = $\frac{5}{36} + \frac{4}{36} = 0.25$

lower end: $z = \frac{7.5-7}{2.42} = 0.20$

upper end: $z = \frac{9.5-7}{2.42} = 1.05$

area = $\frac{.7063-.1585}{2} = .2739$

really precise area using computer normal tables = 0.2677

Wait a minute! Why did we look at 7.5 to 9.5 instead of 8 to 9?

Recall: The probability histogram has 1 bin for each possible value. The value itself is in the center of the bin. So, the bin for 8 is 1 wide and has height equal to P(8). If we just went from 8 to 9 in the normal approximation, we'd miss the left half of the 8 bin, and the right half of the 9 bin.

This is called the continuity correction. It's especially important if you have relatively few possible values, or if you need a high level of precision.

Wait a minute! Why did we look at 7.5 to 9.5 instead of 8 to 9?

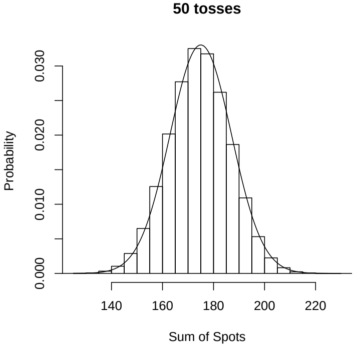
Recall: The probability histogram has 1 bin for each possible value. The value itself is in the center of the bin. So, the bin for 8 is 1 wide and has height equal to P(8). If we just went from 8 to 9 in the normal approximation, we'd miss the left half of the 8 bin, and the right half of the 9 bin.

This is called the continuity correction. It's especially important if you have relatively few possible values, or if you need a high level of precision.

Wait a minute! Why did we look at 7.5 to 9.5 instead of 8 to 9?

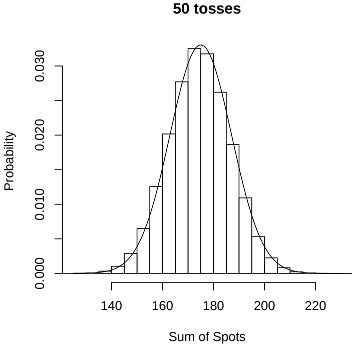
Recall: The probability histogram has 1 bin for each possible value. The value itself is in the center of the bin. So, the bin for 8 is 1 wide and has height equal to P(8). If we just went from 8 to 9 in the normal approximation, we'd miss the left half of the 8 bin, and the right half of the 9 bin.

This is called the continuity correction. It's especially important if you have relatively few possible values, or if you need a high level of precision.



Expected Value = $50 \times 3.5 = 175$
Standard Error = $\sqrt{50} \times 1.71 = 12.08$

Figuring out the exact probabilities would be quite challenging, but the normal approximation is pretty good.



Expected Value = $50 \times 3.5 = 175$
Standard Error = $\sqrt{50} \times 1.71 = 12.08$

Figuring out the exact probabilities would be quite challenging, but the normal approximation is pretty good.

Let's find P(200 or more total spots). To use the continuity correction, or not?

With correction: $z = \frac{199.5-275}{12.08} = 2.029$

Without correction $z = \frac{200-275}{12.08} = 2.07$

You'd use 2.05 on your normal table either way. There are so many bins that the width of half a bin doesn't make any difference.

area = $\frac{1-.9596}{2} = .0202$

Let's find P(200 or more total spots). To use the continuity correction, or not?

With correction: $z = \frac{199.5-275}{12.08} = 2.029$

Without correction $z = \frac{200-275}{12.08} = 2.07$

You'd use 2.05 on your normal table either way. There are so many bins that the width of half a bin doesn't make any difference.

area = $\frac{1-.9596}{2} = .0202$

Let's find P(200 or more total spots). To use the continuity correction, or not?

With correction: $z = \frac{199.5-275}{12.08} = 2.029$

Without correction $z = \frac{200-275}{12.08} = 2.07$

You'd use 2.05 on your normal table either way. There are so many bins that the width of half a bin doesn't make any difference.

area = $\frac{1-.9596}{2} = .0202$

Let's find P(200 or more total spots). To use the continuity correction, or not?

With correction: $z = \frac{199.5-275}{12.08} = 2.029$

Without correction $z = \frac{200-275}{12.08} = 2.07$

You'd use 2.05 on your normal table either way. There are so many bins that the width of half a bin doesn't make any difference.

area = $\frac{1-.9596}{2} = .0202$

Let's find P(200 or more total spots). To use the continuity correction, or not?

With correction: $z = \frac{199.5-275}{12.08} = 2.029$

Without correction $z = \frac{200-275}{12.08} = 2.07$

You'd use 2.05 on your normal table either way. There are so many bins that the width of half a bin doesn't make any difference.

area = $\frac{1-.9596}{2} = .0202$

Should I go half up, or half down?

"200 or more" indicates that 200 should be included... so we went with 199.5.

If I'd asked "more than 200", that indicates that 200 isn't included... we'd have gone with 200.5.

Consider whether the endpoint (in this example, 200) is meant to be included or not. That'll help you decide on which side of the value to use the continuity correction.

Should I go half up, or half down?

"200 or more" indicates that 200 should be included... so we went with 199.5.

If I'd asked "more than 200", that indicates that 200 isn't included... we'd have gone with 200.5.

Consider whether the endpoint (in this example, 200) is meant to be included or not. That'll help you decide on which side of the value to use the continuity correction.

Should I go half up, or half down?

"200 or more" indicates that 200 should be included... so we went with 199.5.

If I'd asked "more than 200", that indicates that 200 isn't included... we'd have gone with 200.5.

Consider whether the endpoint (in this example, 200) is meant to be included or not. That'll help you decide on which side of the value to use the continuity correction.

Should I go half up, or half down?

"200 or more" indicates that 200 should be included... so we went with 199.5.

If I'd asked "more than 200", that indicates that 200 isn't included... we'd have gone with 200.5.

Consider whether the endpoint (in this example, 200) is meant to be included or not. That'll help you decide on which side of the value to use the continuity correction.

Consider a single number bet in roulette. We want to count the number of times you win.

Box: [One 1 and Thirty-Seven 0s]

Mean: $\frac{1}{38} \approx 0.026$

SD: $(1 - 0) \times \sqrt{\frac{1}{38} \times \frac{37}{38}} \approx 0.16$

Consider a single number bet in roulette. We want to count the number of times you win.

Box: [One 1 and Thirty-Seven 0s]

Mean: $\frac{1}{38} \approx 0.026$

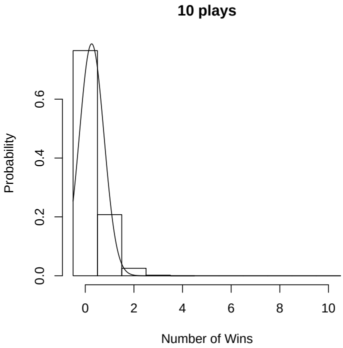
SD: $(1-0) \times \sqrt{\frac{1}{38} \times \frac{37}{38}} \approx 0.16$

Consider a single number bet in roulette. We want to count the number of times you win.

Box: [One 1 and Thirty-Seven 0s]

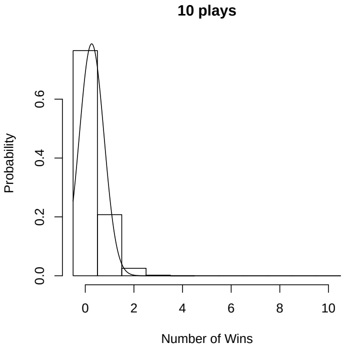
Mean: $\frac{1}{38} \approx 0.026$

SD: $(1 - 0) \times \sqrt{\frac{1}{38} \times \frac{37}{38}} \approx 0.16$



Expected Value = $10 \times 0.026 = 0.26$
Standard Error = $\sqrt{10} \times 0.16 = 0.51$

Using the normal approximation here would be very silly.



Expected Value = $10 \times 0.026 = 0.26$
Standard Error = $\sqrt{10} \times 0.16 = 0.51$

Using the normal approximation here would be very silly.

Let's do it anyway. Find P(win fewer than 0 times).

$z = \frac{-0.5-0.26}{0.51} \approx -0.5$

Area = $\frac{1-.3829}{2} \approx .31$

The normal approximation quite cheerfully tells you there's a 31% chance of winning a negative number of games. This is completely absurd.

Let's do it anyway. Find P(win fewer than 0 times).

$z = \frac{-0.5-0.26}{0.51} \approx -0.5$

Area = $\frac{1-.3829}{2} \approx .31$

The normal approximation quite cheerfully tells you there's a 31% chance of winning a negative number of games. This is completely absurd.

Let's do it anyway. Find P(win fewer than 0 times).

$z = \frac{-0.5-0.26}{0.51} \approx -0.5$

Area = $\frac{1-.3829}{2} \approx .31$

The normal approximation quite cheerfully tells you there's a 31% chance of winning a negative number of games. This is completely absurd.

Let's do it anyway. Find P(win fewer than 0 times).

$z = \frac{-0.5-0.26}{0.51} \approx -0.5$

Area = $\frac{1-.3829}{2} \approx .31$

The normal approximation quite cheerfully tells you there's a 31% chance of winning a negative number of games. This is completely absurd.

So when SHOULDN'T you use the normal approximation?

Figure out the expected value and standard error.

Consider what values your sum could possibly take. In this example, it was between 0 and 10.

If you take $EV \pm 2 \times SE$ and get a value outside the range of possible values, do NOT use the normal approximation.

In this example, $EV \pm 2 \times SE$ gives a range from -0.75 to 1.28. This includes values that are not possible, so the normal approximation should not be used.

So when SHOULDN'T you use the normal approximation?

Figure out the expected value and standard error.

Consider what values your sum could possibly take. In this example, it was between 0 and 10.

If you take $EV \pm 2 \times SE$ and get a value outside the range of possible values, do NOT use the normal approximation.

In this example, $EV \pm 2 \times SE$ gives a range from -0.75 to 1.28. This includes values that are not possible, so the normal approximation should not be used.

So when SHOULDN'T you use the normal approximation?

Figure out the expected value and standard error.

Consider what values your sum could possibly take. In this example, it was between 0 and 10.

If you take $EV \pm 2 \times SE$ and get a value outside the range of possible values, do NOT use the normal approximation.

In this example, $EV \pm 2 \times SE$ gives a range from -0.75 to 1.28. This includes values that are not possible, so the normal approximation should not be used.

So when SHOULDN'T you use the normal approximation?

Figure out the expected value and standard error.

Consider what values your sum could possibly take. In this example, it was between 0 and 10.

If you take $EV \pm 2 \times SE$ and get a value outside the range of possible values, do NOT use the normal approximation.

In this example, $EV \pm 2 \times SE$ gives a range from -0.75 to 1.28. This includes values that are not possible, so the normal approximation should not be used.

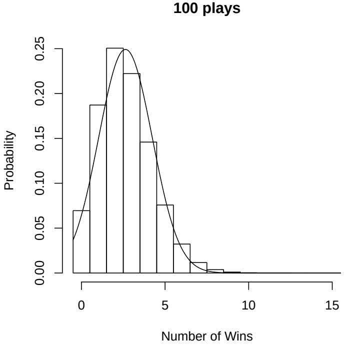
So when SHOULDN'T you use the normal approximation?

Figure out the expected value and standard error.

Consider what values your sum could possibly take. In this example, it was between 0 and 10.

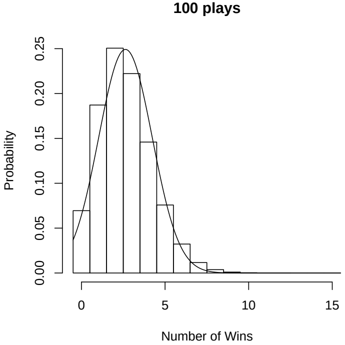
If you take $EV \pm 2 \times SE$ and get a value outside the range of possible values, do NOT use the normal approximation.

In this example, $EV \pm 2 \times SE$ gives a range from -0.75 to 1.28. This includes values that are not possible, so the normal approximation should not be used.



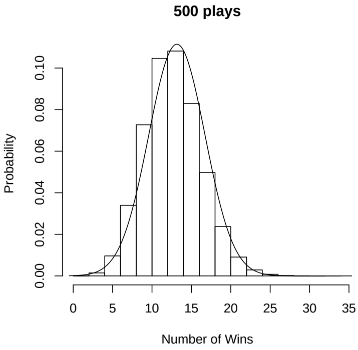
Expected Value = $100 \times 0.026 = 2.63$
Standard Error = $\sqrt{100} \times 0.16 = 1.6$

Even with 100 plays, it's still not very normal.



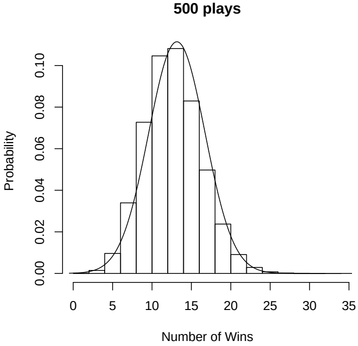
Expected Value = $100 \times 0.026 = 2.63$
Standard Error = $\sqrt{100} \times 0.16 = 1.6$

Even with 100 plays, it's still not very normal.



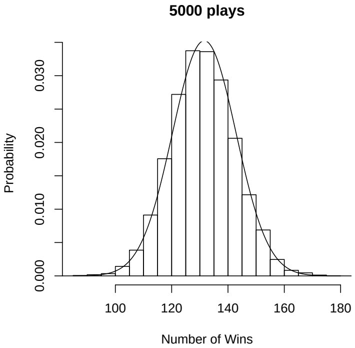
Expected Value = $500 \times 0.026 = 13.16$
Standard Error = $\sqrt{500} \times 0.16 = 3.6$

It still looks a little wonky, but you can use the normal approximation and get decent results.



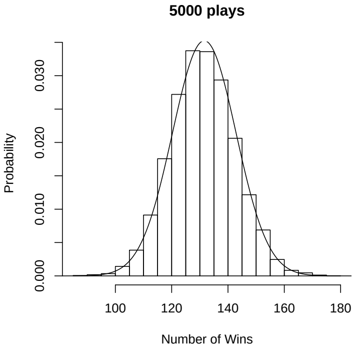
Expected Value = $500 \times 0.026 = 13.16$
Standard Error = $\sqrt{500} \times 0.16 = 3.6$

It still looks a little wonky, but you can use the normal approximation and get decent results.



Expected Value = $5000 \times 0.026 = 131.6$
Standard Error = $\sqrt{5000} \times 0.16 = 11.32$

Now it looks pretty normal.



Expected Value = $5000 \times 0.026 = 131.6$
Standard Error = $\sqrt{5000} \times 0.16 = 11.32$

Now it looks pretty normal.

Consider drawing from the following box, and adding the numbers you see.

Box [1 2 9]

Mean: 4

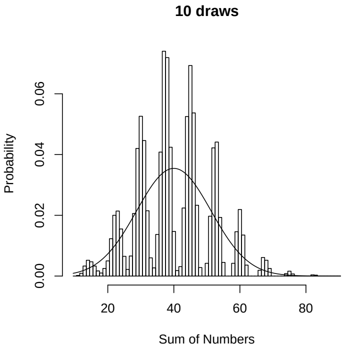
SD: 3.56

Consider drawing from the following box, and adding the numbers you see.

Box [1 2 9]

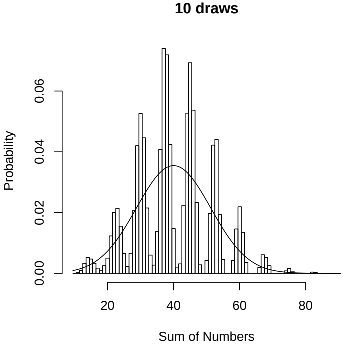
Mean: 4

SD: 3.56



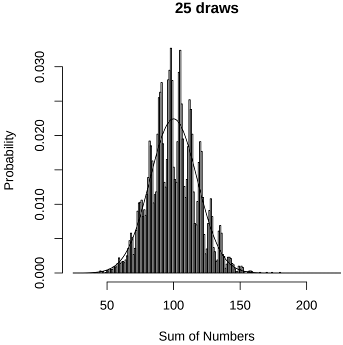
Expected Value = $10 \times 4 = 40$
Standard Error = $\sqrt{10} \times 3.56 = 13.78$

According to our standard, you can use the normal approximation. But be warned: it won't be very good. There are lots of spikes.



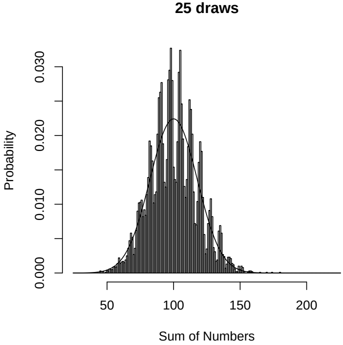
Expected Value = $10 \times 4 = 40$
Standard Error = $\sqrt{10} \times 3.56 = 13.78$

According to our standard, you can use the normal approximation. But be warned: it won't be very good. There are lots of spikes.



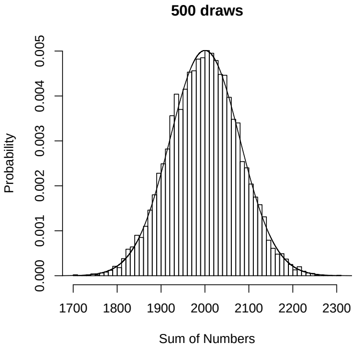
Expected Value = $25 \times 4 = 100$
Standard Error = $\sqrt{25} \times 3.56 = 17.80$

The spikes have smoothed out a bit...



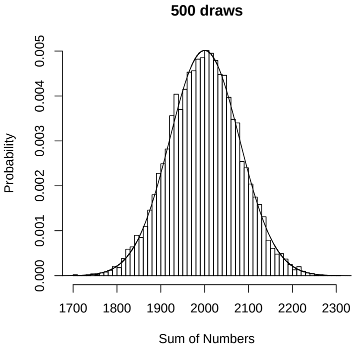
Expected Value = $25 \times 4 = 100$
Standard Error = $\sqrt{25} \times 3.56 = 17.80$

The spikes have smoothed out a bit...



Expected Value = $500 \times 4 = 2000$
Standard Error = $\sqrt{500} \times 3.56 = 79.58$

Much better.



Expected Value = $500 \times 4 = 2000$
Standard Error = $\sqrt{500} \times 3.56 = 79.58$

Much better.

Unfortunately, there's no standard like $EV \pm 2 \times SE$ to use when you have a histogram that will have lots of spikes. This happens when you're taking the sum of draws from a box that has gaps in the numbers. It's just something to keep in mind.

We looked at four situations: tossing a coin, rolling a die, playing a single number bet in roulette, and summing from a box with 1 2 9 in it. In every case, the probability histogram eventually converged to a normal distribution.

Some converged very quickly, like the coin flip. Some took quite a few repetitions, like playing the single number bet. However, they all resembled a normal curve eventually.

We looked at four situations: tossing a coin, rolling a die, playing a single number bet in roulette, and summing from a box with 1 2 9 in it. In every case, the probability histogram eventually converged to a normal distribution.

Some converged very quickly, like the coin flip. Some took quite a few repetitions, like playing the single number bet. However, they all resembled a normal curve eventually.

The Central Limit Theorem.

This is perhaps the single most important theorem in all of statistics. There's a very formal statement of it, with lots of math notation and lists of technical details that you never worry about until graduate school. The gist of it is:

When drawing at random with replacement from a box, the probability histogram for the sum will follow the normal curve if the number of draws is reasonably large.

The center is the expected value, and the spread is the standard error.

The Central Limit Theorem.

This is perhaps the single most important theorem in all of statistics. There's a very formal statement of it, with lots of math notation and lists of technical details that you never worry about until graduate school. The gist of it is:

When drawing at random with replacement from a box, the probability histogram for the sum will follow the normal curve if the number of draws is reasonably large.

The center is the expected value, and the spread is the standard error.

The Central Limit Theorem.

This is perhaps the single most important theorem in all of statistics. There's a very formal statement of it, with lots of math notation and lists of technical details that you never worry about until graduate school. The gist of it is:

When drawing at random with replacement from a box, the probability histogram for the sum will follow the normal curve if the number of draws is reasonably large.

The center is the expected value, and the spread is the standard error.

For tomorrow:

Chapter 19 - Sampling Surveys