Unit 2B - Estimating a Probability (Confidence Interval for p)

Adam Petrie
Department of Business Analytics
University of Tennessee

June 12, 2021

Probability Review

Thinking in terms of frequency

In business analytics, we're often concerned with how *often* some event of interest occurs

- How often will a stockout occur if inventory is "reset" to 20 at the beginning of each week?
- How often will someone click on an ad when placed on the top-center of a webpage?
- How often will an Uber get a request for pickup from Market Square?
- How often will someone who buys dry pasta also buy tomato sauce?

Sometimes, the event happens. Other times it does not. The *frequentist view* of probability defines the probability of some event of interest as the long-run fraction (or proportion) of trials when that event occurs.

Frequentist Definition

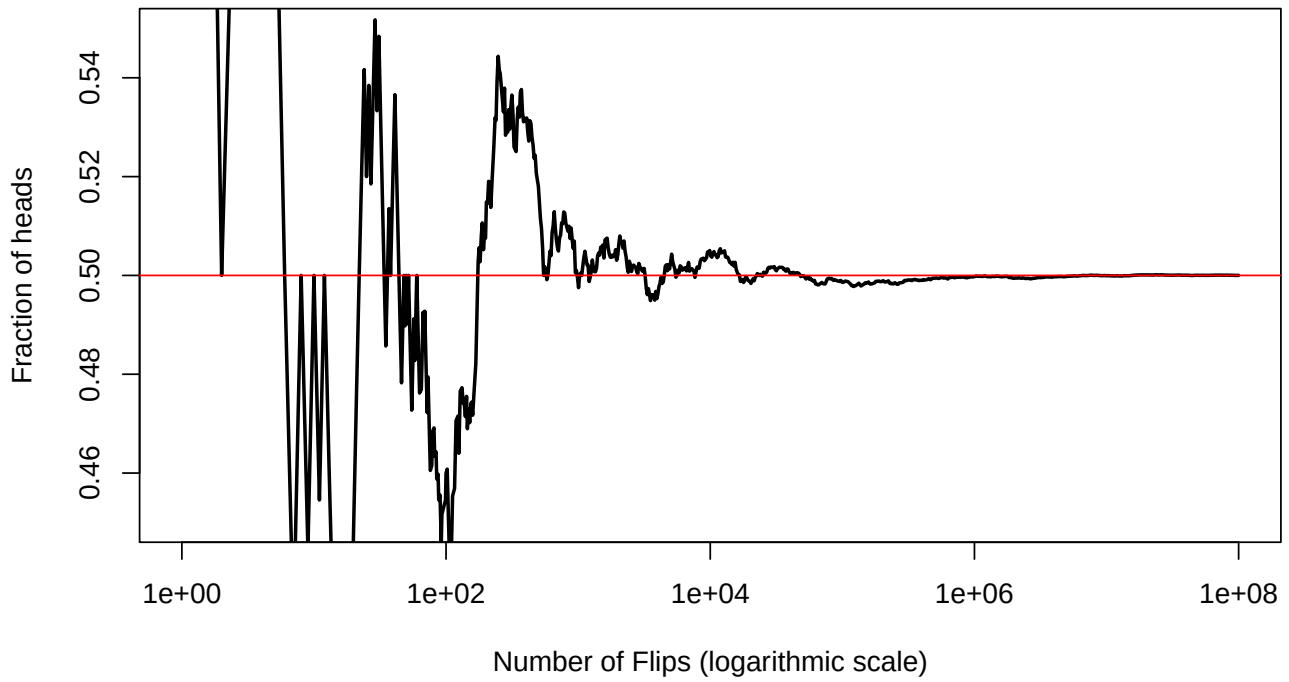
$$P(event) = \lim_{\#trials \to \infty} \frac{\#\text{ trials in which event occurs}}{\#\text{ trials}}$$

For example, we know that the probability of a fair coin coming up heads is $P(heads) = 50\%$.

- In the long run, about 50% of flips turn up heads. The more trials we consider, the closer to 50% the fraction of heads tends to be.
- This does *not* mean that the number of heads after n flips edges progressively closer to $n/2$ (in fact, the difference between the number of heads flipped and $1/2$ the number of flips tends to grow bigger as the number of flips increases).

The probability of a coin coming up heads is 50%

Demo: as the number of coin flips grows, the fraction of trials (flips) that have come up heads edges closer and closer to 50%. However, it tends to always be "a little off". The fraction never lands on 50% and stays there forever after some magic number of trials.



Key Results

Key Results (p vs. $\hat{p}$ and standard error)

Let p be the (true) probability of some event of interest. Denote $\hat{p}$ (our best guess for p) as the fraction of n trials where the event occurred during a Monte Carlo simulation or during data collection.

- $\hat{p}$ will be "close" to the true value of p
- The typical difference between $\hat{p}$ and p is called the **standard error** (SE) of $\hat{p}$. It's the typical size of the error we make when we use $\hat{p}$ to estimate p.
- Surprisingly, there is a mathematical formula that gives us the SE

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Key Results (95% confidence interval)

- To come up with a range of "plausible values" for p that are consistent with our Monte Carlo simulation (or data we've collected), we can construct a "95% confidence" interval:

$$\hat{p} \pm 2SE$$

- There are other procedures for constructing 95% confidence intervals; using `binom.test(x,n)` is an alternative (`n` in the number of trials, `x` is the number of times the event occurred).

Key Results (meaning of confidence)

- The level of confidence of a confidence interval refers to the probability that the *procedure* used to construct the interval produces a range of values that includes p.
- Any particular 95% confidence interval may or may not include p (since we don't know p, we'll never know which are right and which are wrong). But in the long run, about 95% of Monte Carlo simulations you'd run will yield 95% confidence intervals that cover p.
- Different procedures will yield different-looking 95% confidence intervals. This is not a contradiction nor problematic. In the long run, *each* of these procedures will produce intervals that cover p for about 95% of Monte Carlo simulations you'll run.

Key Results: Margin of Error

The "margin of error" is just double the standard error (and is thus the half-width of a confidence interval).

$$95\%\ \text{CI:} \qquad \hat{p} \pm ME$$

To determine the number of trials of a Monte Carlo simulation required to achieve a certain margin of error, use:

$$n = \frac{4\hat{p}(1-\hat{p})}{ME^2}$$

- $\hat{p}$ above is a "guess" for p (perhaps from a small pilot study)
- If you have no guess for p, plugging in $p = 0.5$ gives the *largest* n will ever need to be to achieve the desired margin of error.

How far off is our estimate of p?

How far off are our estimates?

Consider the following scenario:

- We run a Monte Carlo simulation that mimics us playing roulette with a specific betting strategy.
- We record whether we were able to double our money before going broke.
- We repeat the simulation with a different stream of random numbers a bunch of times.
- Out of the 1000 simulations, we doubled our money 512 times.
- The estimated probability is $512/1000 = 51.2\%$.
- How wrong are we? How far off is our estimate from the actual probability of doubling our money using the strategy?

Notation

- n - the number of trials (or number of observations in our data)
- p - the actual probability of the event of interest occurring
- $\hat{p}$ (pronounced "p-hat") - the estimated probability; the fraction of trials where the event of interest occurred
- SE - the **standard error** of $\hat{p}$ (our "best guess" of how far $\hat{p}$ is from p)

Question: can we come up with a number for SE based on our data, i.e., the measured value of $\hat{p}$ and the number of trials?

Intuition and Motivation

Yes! It turns out we can be fairly confident about just "how wrong" our guess is.

Let's develop some intuition as to the behavior of the size of the standard error via a series of investigations.

While this "deep dive" is not something we will test on (and you won't need to reproduce it), some of the insights we will stumble upon are invaluable to statistics as a whole. Sit back and enjoy the ride!

Best guess for how far p might be from $\hat{p}$

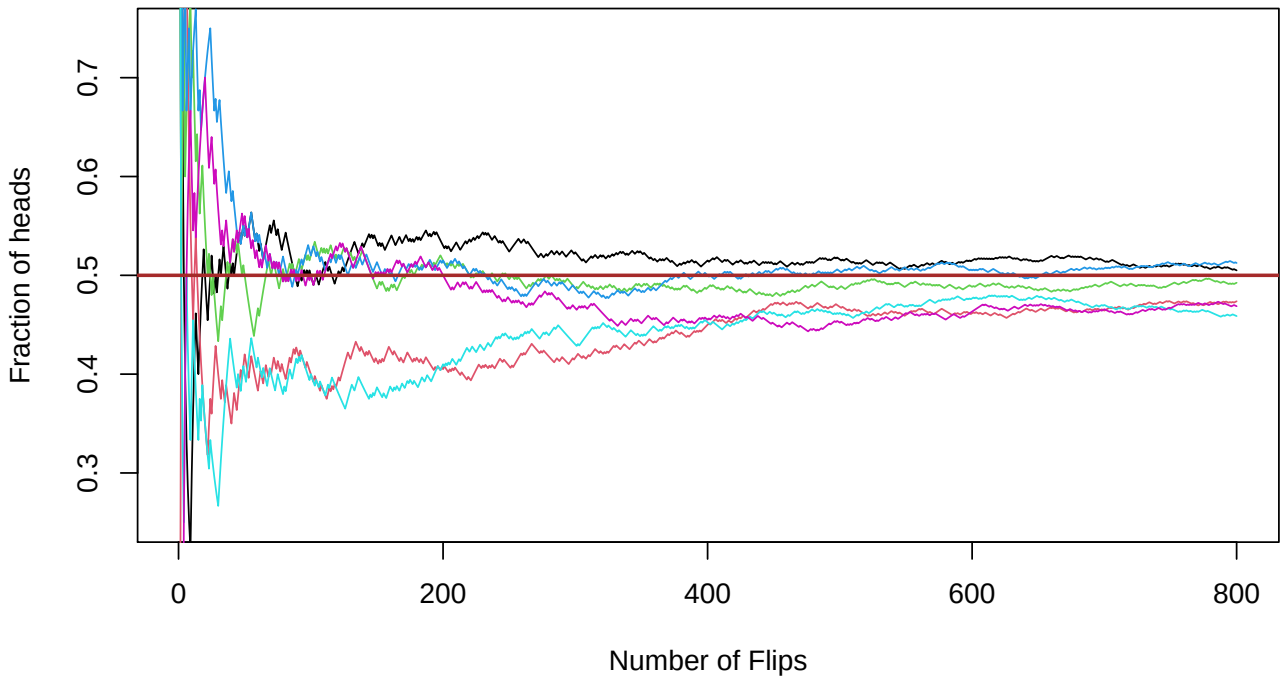
After we flip a fair coin 800 times, how close will $\hat{p}$, the proportion of flips that came up heads, be to 50%?

That's going to depend on the exact sequences of flips that happen to be generated.

$\hat{p}$ will be closer to 50% for some sequences than others. For some sequences, it's possible that after 800 flips *exactly* 50% will be heads.

Best guess for how far p might be from $\hat{p}$

For example, the plot below shows the trajectory of the fraction of flips that have come up heads over the course of the first 800 flips for each of six students conducting a Monte Carlo simulation with different random number seeds.



General Observations

Observations regarding the simulations of these 6 students:

- About half of the trajectories end with $\hat{p}$ above p (the horizontal line at 50%), while the other half end with $\hat{p}$ below.
- Sometimes $\hat{p}$ is quite close to p), and other times it's rather "far".
- The difference between p and $\hat{p}$ seems to get smaller as the number of flips increases.
- Some trajectories feel "out of whack" (looking at you red/turquoise; *always* below 50%), but they are legitimate!

Typical difference

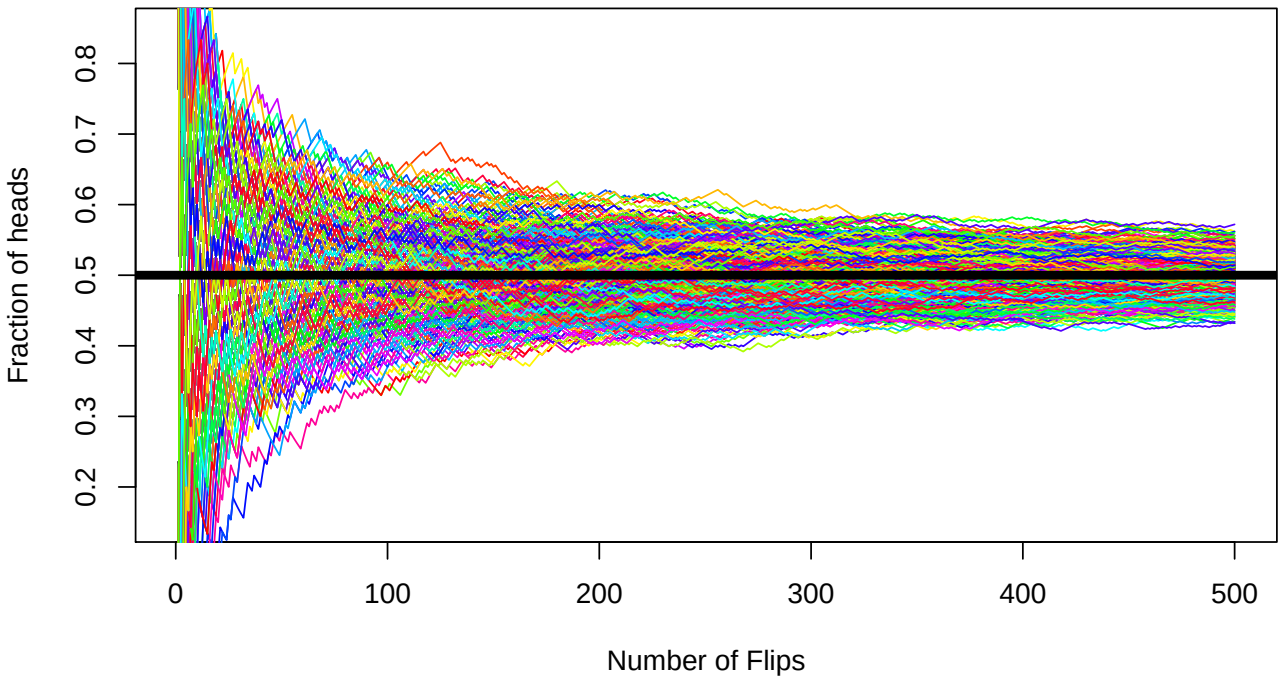
Since the difference between p and $\hat{p}$ is sometimes big and sometimes small, we really first need to decide what we even mean by the "typical" difference between them.

Average difference? Average absolute difference? Something else?

Once we agree on exactly how to measure this "typical" difference, we can try to discover how it depends on n, p, and/or $\hat{p}$.

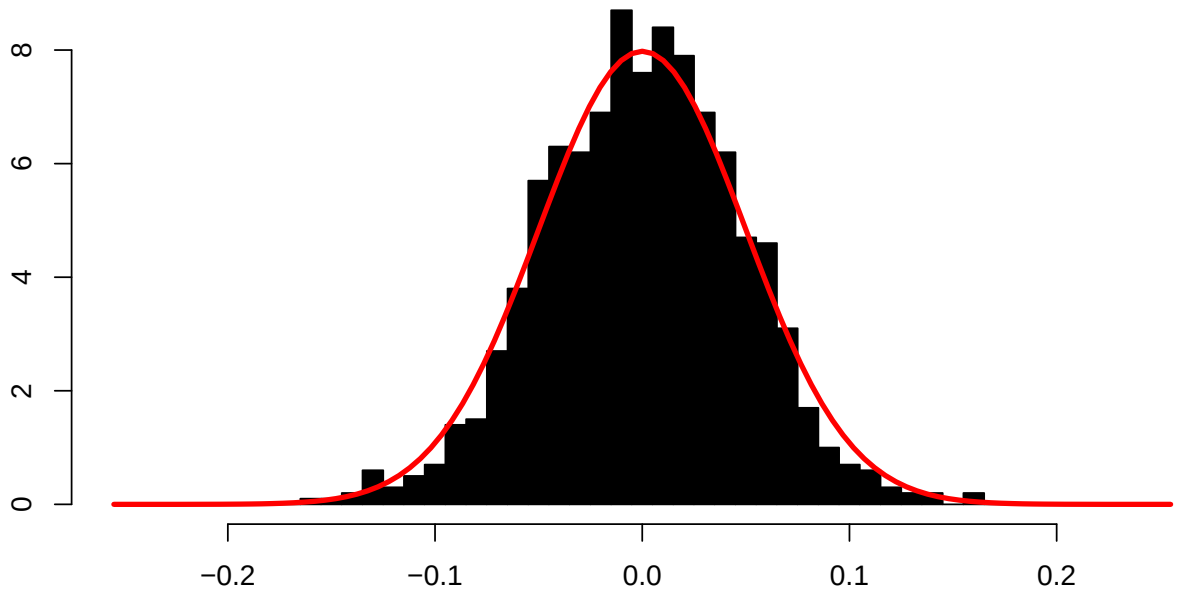
Quantifying the "typical difference"

Let's now have 1000 students each run a Monte Carlo simulation where they mimic 800 flips of a fair coin. Let's look at each of the 1000 trajectories of $\hat{p}$.



Quantifying the "typical difference"

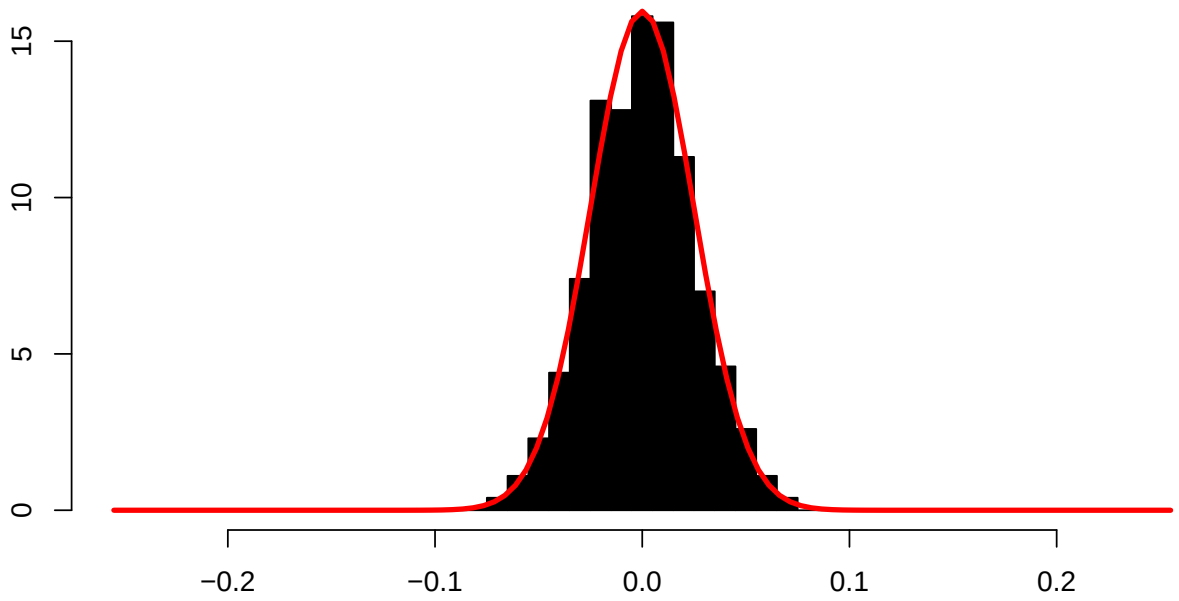
Let's look at how far $\hat{p}$ is from p for each of these 1000 students' simulations after 100 flips with a histogram. This is a shape you might recognize from STAT 201 – the famous "bell curve", so let's superimpose that in "red".



Distances between p and 1000 p−hats calculated after 100 flips

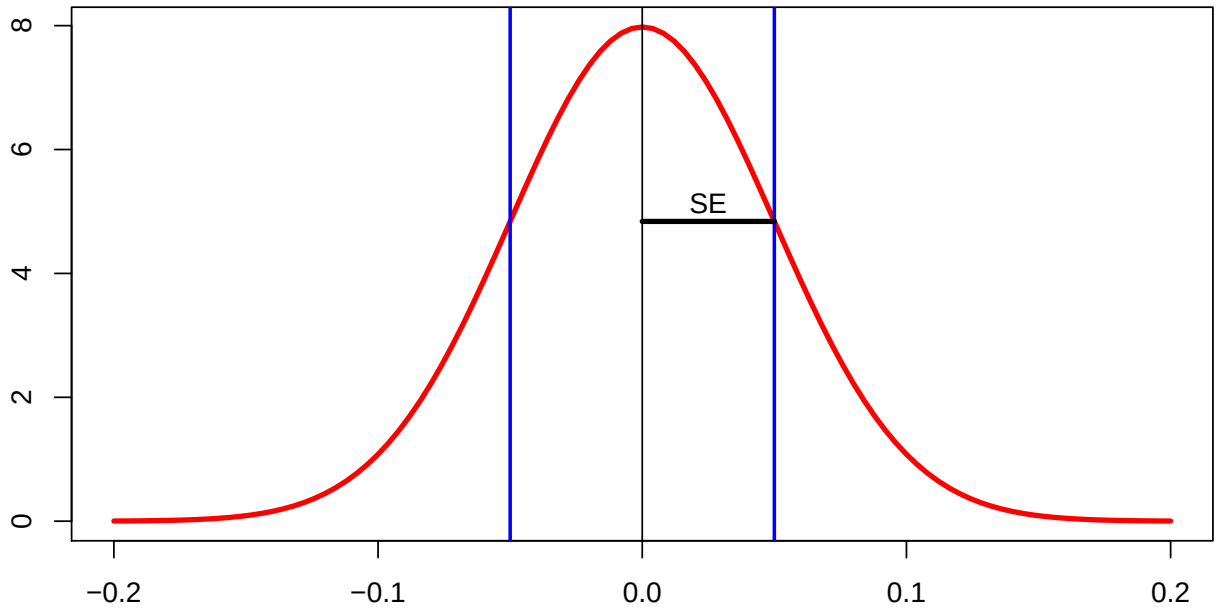
Quantifying the "typical difference"

A skinnier bell-curve looks appropriate if we instead look at how far $\hat{p}$ is from p for each of these 1000 students' simulations after 400 flips with a histogram.



Quantifying the "typical difference"

See where the curve goes from "curving down" to "curving out" (intersection of red and blue lines)? The curve goes from decreasing ever more quickly to ever more slowly here. **Let us agree to define the standard error *SE* based on where this "inflection point" occurs**. For this specific example, it looks like the SE equals 0.05. "Most of the time", $\hat{p}$ is within 1 SE of p.



Quantifying the "typical difference"

Ok but what does "most of the time" mean? I don't want to pay too close attention to the specific number just yet, so let's say "about 2/3rds of the time p will be within 1 SE of $\hat{p}$".

```
p <- 0.5
#What fraction of these 1000 people had p within 1 SE (0.05 in this case) after 100 flips?

mean( p - p.hat > -0.05 & p - p.hat < 0.05)

## [1] 0.698
```

Size of the *SE* and the number of trials

After 100 flips of a fair coin, our investigation reveals that "usually" (about 2/3rd of the time) $\hat{p}$ is within 0.05 of p. While the size of the *SE* is 0.05 after 100 flips, our investigation shows that the SE gets smaller and smaller as the number of trials grows (skinnier bell curve when $n = 400$).

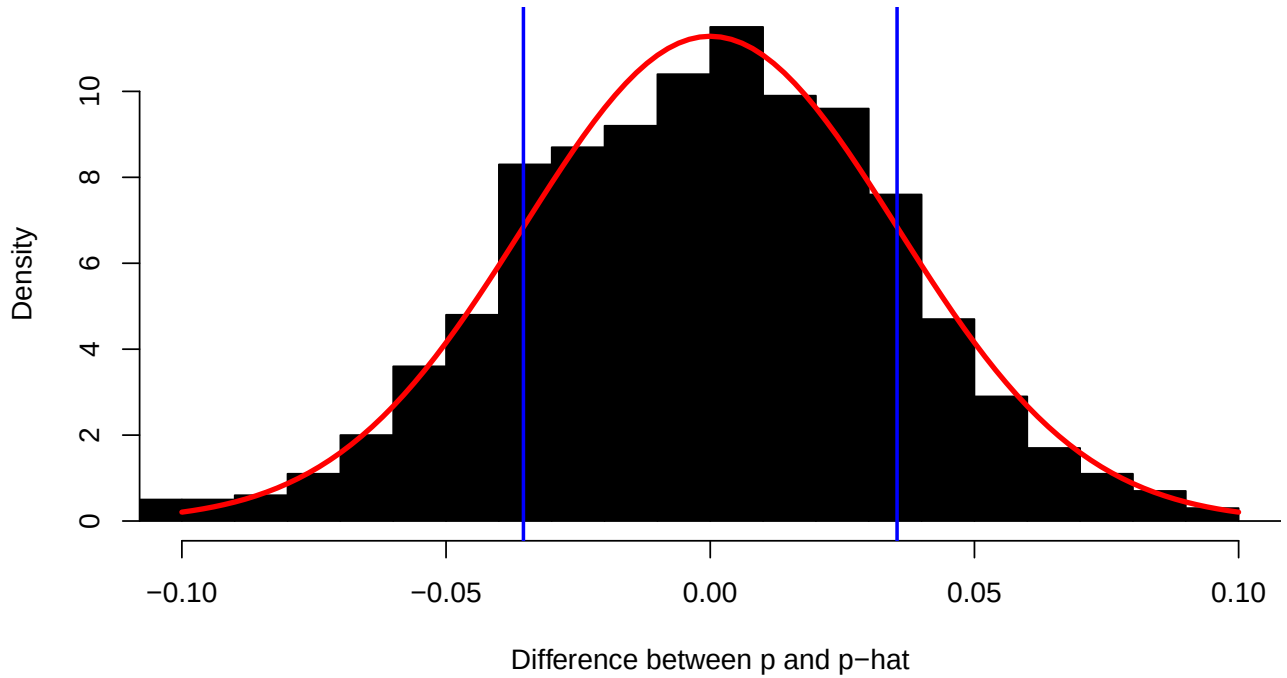
What if we looked at the differences between p and the $\hat{p}s$ after 200 flips? After 400? After 800? Maybe a pattern will emerge and we can make a guess as to the general behavior of SE with the number of flips.

Size of the *SE* and the number of trials

After 200 flips, it looks like the SE is about 0.035.

```
mean( p - p.hat > -0.035 & p - p.hat < 0.035)

## [1] 0.682
```

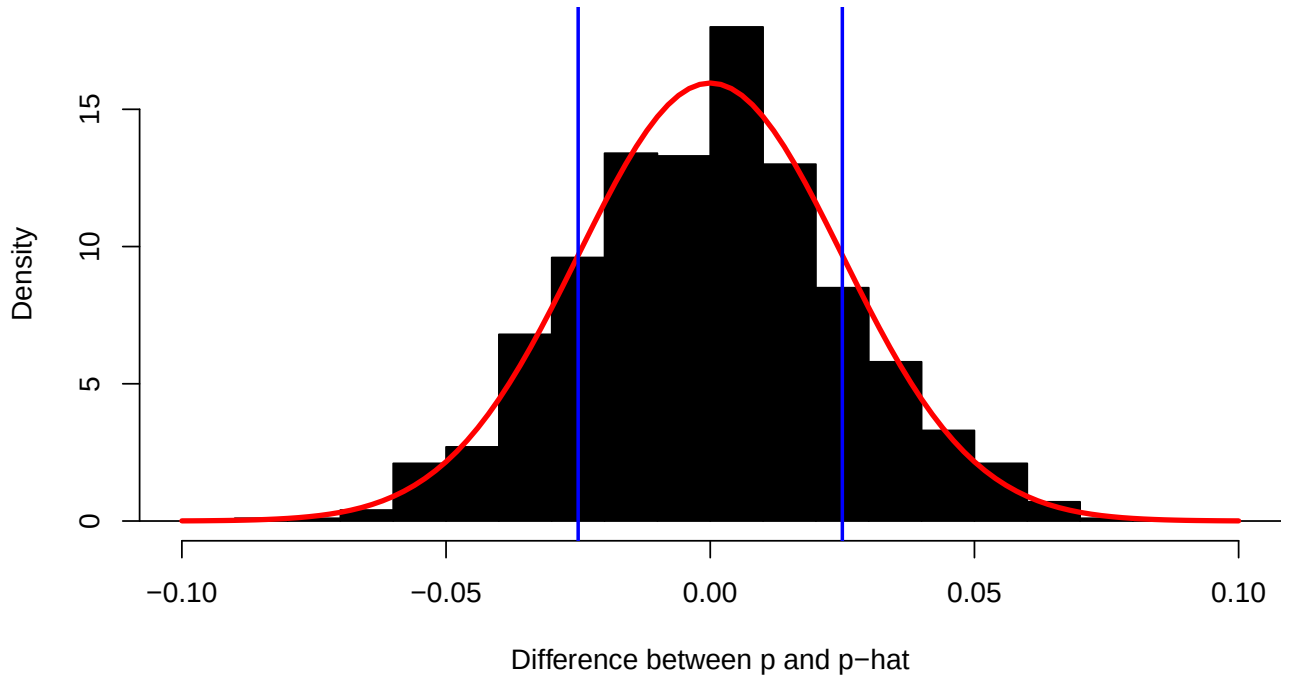


Size of the *SE* and the number of trials

After 400 flips, it looks like the SE is right at 0.025.

```
mean( p - p.hat > -0.025 & p - p.hat < 0.025)

## [1] 0.662
```

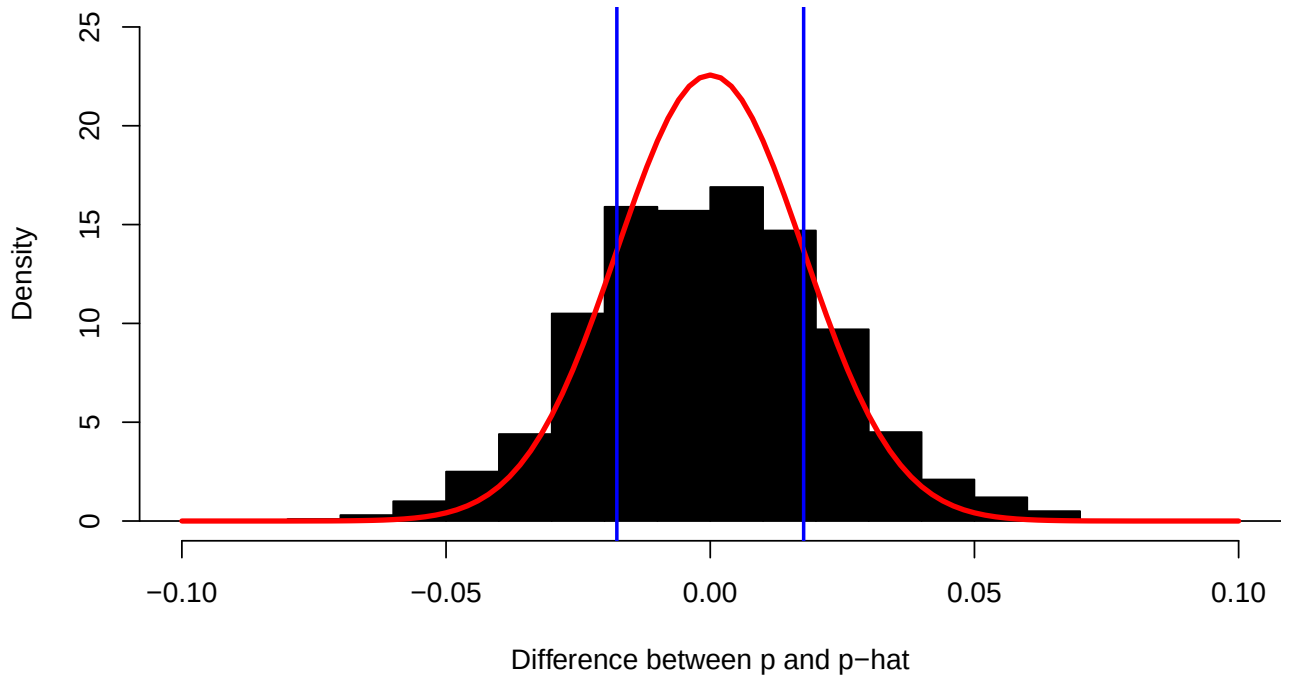


Size of the *SE* and the number of trials

After 800 flips, it looks like the SE is about at 0.0177.

```
mean( p - p.hat > -0.0177 & p - p.hat < 0.0177)

## [1] 0.556
```



Observations regarding size of the *SE* and the number of trials

Flips	100	200	400	800
SE	0.05	0.035	0.025	0.0177

Notice that when the # flips is *quadrupled* (100 $\rightarrow$ 400, or 200 $\rightarrow$ 800), the SE is *halved* (0.05 $\rightarrow$ 0.025, or 0.035 $\rightarrow$ 0.0177).

Observations regarding size of the *SE* and the number of trials

Flips	100	200	400	800
SE	0.05	0.035	0.025	0.0177

More generally, when the # flips goes up by a factor of 2, the SE goes down by a factor of $\sqrt{2}$. $0.05/\sqrt{2} = 0.03535534$, then $0.03535534/\sqrt{2} = 0.025$, then $0.025/\sqrt{2} = 0.01767767$.

Observations regarding size of the SE and the number of trials

We've discovered a fundamental truth (which is true regardless of p or $\hat{p}$).

- $\hat{p}$ is typically off from p by about a standard error so (about 2/3rds of the time, $\hat{p}$ is within 1 SE of p); there's an equal chance that it's larger/smaller than p.
- The size of the SE "goes as" $1/\sqrt{n}$. Meaning that if we increase the number of trials n by a factor of 4, the size of the SE goes down by a factor of 2; if we increase n by a factor of 25, the size of the SE goes down by a factor of 5, etc.

In fact, when $p = 0.5$, the formula for the standard error is $SE = \frac{0.5}{\sqrt{n}}$

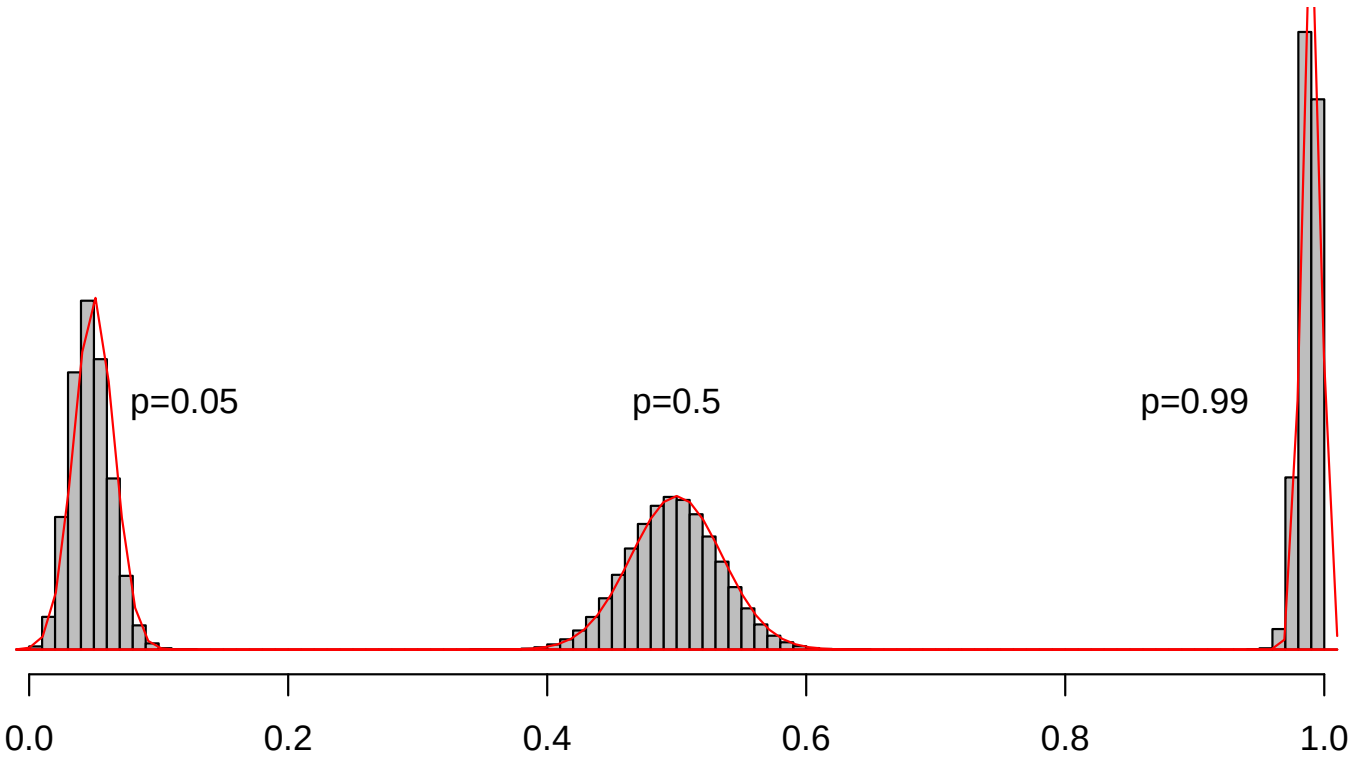
Standard Error for any p

We've discovered that when $p = 0.5$, the formula for the standard error is $SE = \frac{0.5}{\sqrt{n}}$. However, this formula won't generalize to other values of p.

- What if $p = 0.999$ and $n = 40000$? Then the formula gives $SE = \frac{0.5}{\sqrt{40000}} = 0.0025$. However, we can't measure a value for $\hat{p} > 1$, so at most p can be 0.001 below $\hat{p}$ (much less than 0.0025).
- What if $p = 0.001$ and $n = 100$? Then the formula gives $SE = \frac{0.5}{\sqrt{100}} = 0.005$. However, we can't measure a value for $\hat{p} < 0$, so at most p can be 0.001 above $\hat{p}$ (much less than 0.005 above).

Behavior of SE with p

The size of the SE depends on p as well as n: it's harder to be off by "a lot" if $p = 0.05$ or $p = 0.99$ as opposed to when $p = 0.5$.



p.hats from 1000 students after trials flips when p=0.05, 0.5, 0.99

Standard Error for any p

Thus, logic dictates that our formula is incomplete. The size of the standard error and thus how far $\hat{p}$ might be from p is going to depend on p!

What does this dependency look like?

Let's investigate!

Standard Error for any p

- Take $p = 0.01$ and consider a Monte Carlo simulation with 800 trials.
- We'll let 1000 people run their own simulations, and we'll compare the 1000 $\hat{p}$'s that emerge with $p = 0.01$ and find the size of the SE in that case (we'll superimpose the bell-curve and see where it goes from curving down to curving out).
- Do this again but for $p = 0.02$, $p = 0.03$, $p = 0.04$, etc., all the way up to 0.99.
- Are there any recognizable patterns between SE and p?

Aside: Code for standard error for any p investigation

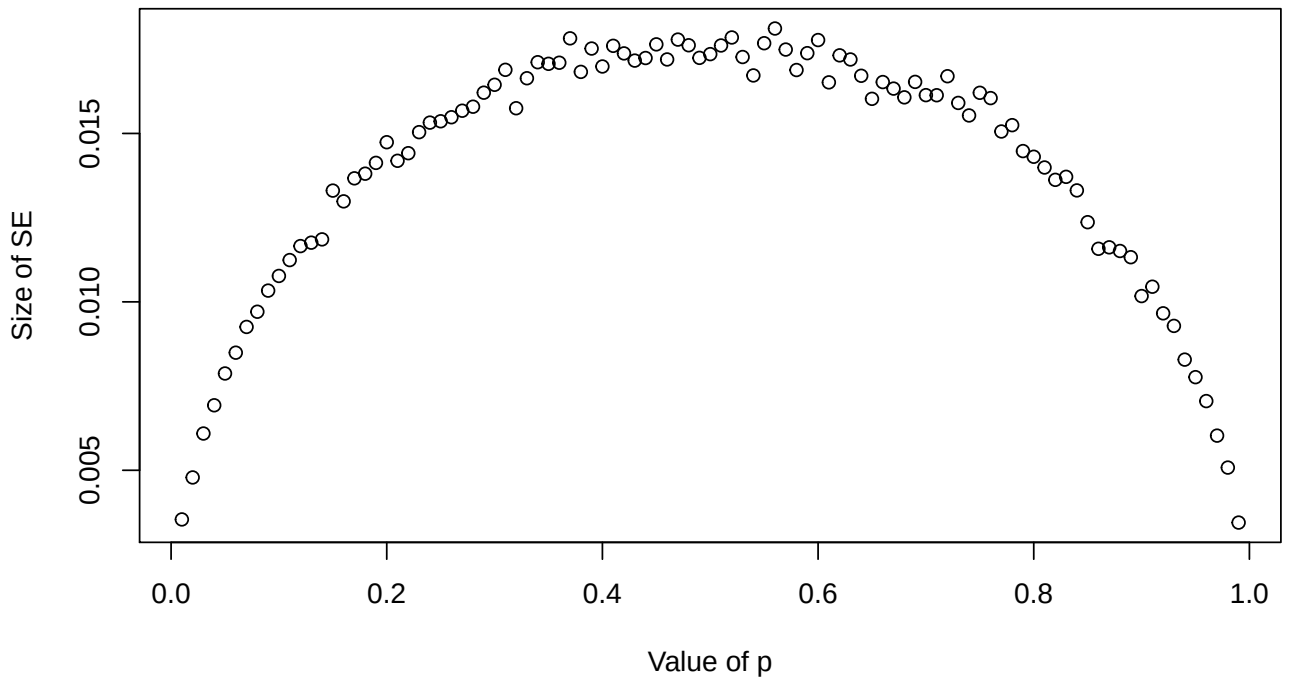
In English, consider values of $p = 0.01, 0.02, \ldots, 0.99$. For each value of p, have 1000 individuals conduct a simulation of 800 trials where "heads" comes up with a probability p. The proportion of heads that occur after 800 flips is that individual's $\hat{p}$. The standard deviation of those 1000 $\hat{p}$s is how we'll compute the SE for that value of p.

```
p <- seq(from=0.01,to=0.99,by=0.01)  #all the ps we want to consider
SE <- rep(0,length(p))
for( i in 1:length(p) ) {
  phat <- rep(0,1000) #initialize measured p-hat as 0 for each of the 1000 people doing the simulation
  truep <- p[i]  #truep will contain the value of p under consideration
  for (person in 1:1000) {
    observed <- sample( c(1,0), prob=c(truep,1-truep), size=800, replace=TRUE )
    total_observed <- cumsum(observed)
    phat[person] <- total_observed[800]/800
  }
  SE[i] <- sd(phat)  #turns out sd() of the measured p-hats give a good idea of size of SE
}
```

Standard Error for any p investigation

Plotting the SE vs. p, we definitely see a pattern. Looks like the size of the SE is biggest when $p = 0.5$ and smallest when p is close to 0 or 1.

```
plot(SE~p,xlab="Value of p",ylab="Size of SE")
```



Standard Error for any p investigation

This looks like a parabola, so maybe the SE can be *predicted* from p and p^2? We all took BAS 320, so let's fit the regression $SE = b_0 + b_1 p + b_2 p^2$. Nice fit, but room for improvement

```
summary( lm(SE~p+I(p^2)) )  #to make p^2 a predictor have to put it inside I()

##
## Call:
## lm(formula = SE ~ p + I(p^2))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0026268 -0.0004445  0.0001302  0.0005503  0.0013060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0056187  0.0002337   24.05   <2e-16 ***
## p            0.0501619  0.0010786   46.51   <2e-16 ***
## I(p^2)      -0.0502043  0.0010450  -48.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0007594 on 96 degrees of freedom
## Multiple R-squared:  0.9601,Adjusted R-squared:  0.9592
## F-statistic:  1154 on 2 and 96 DF,  p-value: < 2.2e-16
```

Standard Error for any p investigation

On a whim, let's predict SE^2 from p and p^2. Even better fit! The intercept term b_0 is basically 0 (notice the p-value being larger than 5%), and the coefficients of p and p^2 are basically the same just opposite signs Coincidentally (?), these coefficients are basically 1/800, which is one over the number of trials!

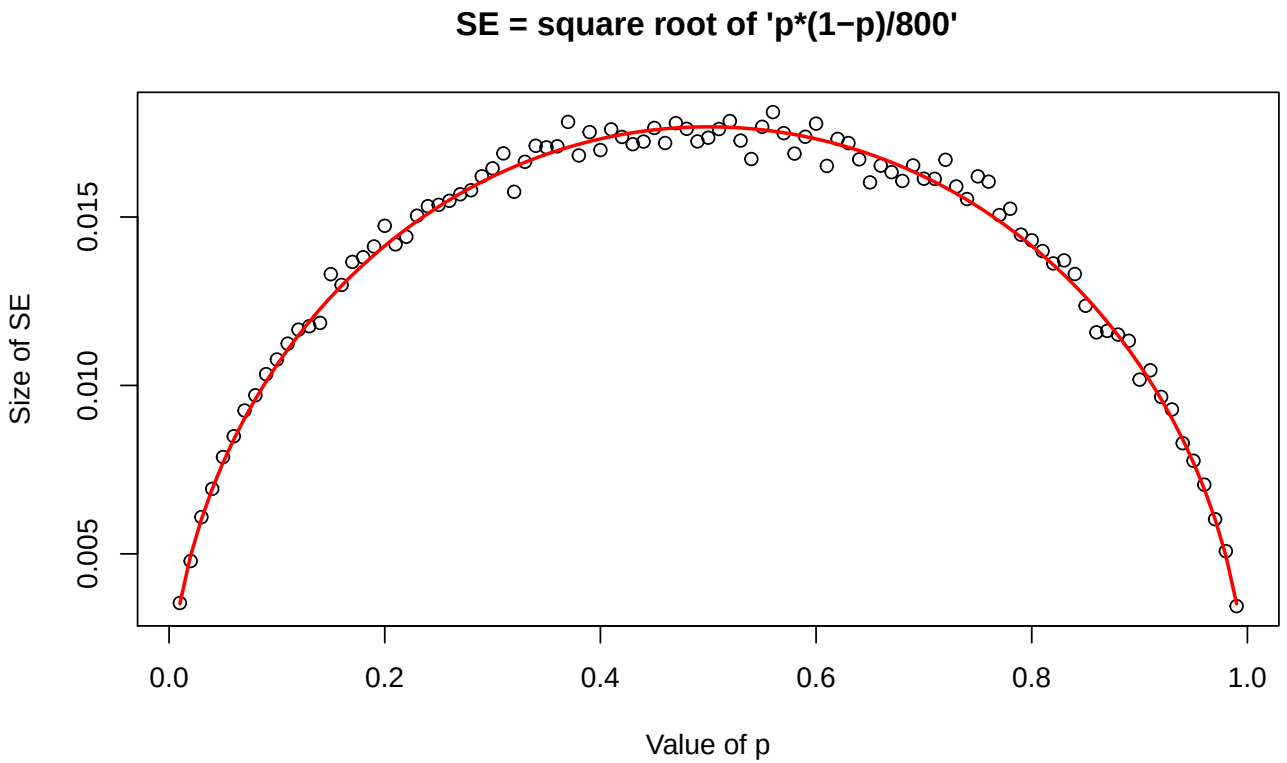
```
summary( lm(SE^2~p+I(p^2)) )

##
## Call:
## lm(formula = SE^2 ~ p + I(p^2))
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.035e-05 -5.513e-06  4.630e-08  5.211e-06  2.912e-05
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.405e-06  3.407e-06   1.293    0.199
## p            1.231e-03  1.572e-05  78.293   <2e-16 ***
## I(p^2)      -1.232e-03  1.524e-05 -80.866   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.107e-05 on 96 degrees of freedom
## Multiple R-squared:  0.9855,Adjusted R-squared:  0.9852
## F-statistic:  3270 on 2 and 96 DF,  p-value: < 2.2e-16

#1.25e-03 = 1/800 !
```

Standard Error for any p investigation

Looks like quite a good fit!



Standard Error for any p investigation results

Let the number of trials be n (instead of 800). The equation we've discovered is:

$$SE^2 = \frac{p}{n} - \frac{p^2}{n}$$

Flexing our algebra skills:

$$SE^2 = \frac{p(1-p)}{n}$$

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Remember our investigation when $p = 0.5$? Well, if we plug in $p = 0.5$ we get $SE = 0.5/\sqrt{n}$, exactly what we figured out before!

Cool formula bro, but . . .

Mathematics and probability theory can be used to show this formula is "correct" (in the sense that it describes where the bell-curve goes from sloping down to sloping out, at least when the number of trials n is large). But is the formula *useful*?

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

To know how far off p might be from $\hat{p}$, it looks like we have to *know* what p is. We've been writing Monte Carlo simulations to estimate p because its value is exactly what we're trying to ascertain! So this formula is unusable.

Our best guess for SE

However, if we've conduct a "large" number of trials, then $\hat{p}$ should be pretty close to p. So what if we just substituted $\hat{p}$ into the equation and are honest that it's our "best guess" of the value of SE (p might be more than a SE above or below $\hat{p}$ anyway).

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The value output by this formula should be "pretty accurate" (i.e. about 2/3rd of the time we'll find that p is within 1 SE of $\hat{p}$) assuming that n is large. Experiments show as long as $n\hat{p} \geq 10$ or so, or $n(1-\hat{p}) \geq 10$ or so, the formula is pretty good.

Summary

- After we've conducted a Monte Carlo simulation, it's unlikely that the proportion of trials where our event of interest occurred ($\hat{p}$) equals p, but it should be "close".
- The typical difference between p and $\hat{p}$ is referred to as the "standard error" of $\hat{p}$, and we have figured out a formula for it:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Usually (meaning about 2/3rds of the time), the actual value of p will be at most one standard error away from $\hat{p}$.

Confidence Interval for p

Can we come up with a range of plausible values for p?

The whole goal of a Monte Carlo simulation is to estimate p. The more trials we perform, the closer $\hat{p}$ typically is to p. In fact we even have a formula that tells us how far p typical is from $\hat{p}$.

Can we come up with a *range* of values for p that seem plausible based on our measured value of $\hat{p}$?

Let's reason our way through it.

Can we come up with a range of plausible values for p?

Consider the following **procedure**:

- Run a Monte Carlo simulation with a "large" number of trials n
- Find the fraction of trials where the event of interest occurred $\hat{p}$
- Calculate $\hat{p} \pm SE$, in other words make the interval

$$\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- Let the values inside the interval be a set of "plausible" values for p based on the simulation.

For example, if $n = 10000$ and $\hat{p} = 0.420$, then $SE = \sqrt{0.42 \cdot 0.58/10000} = 0.004935585$. Any value inside the interval $(0.415, 0.425)$ is a "plausible" value for p.

Confidence interval for p

What is the probability that the described procedure will provide an interval that covers p, i.e. that "gets it right"?

We addressed this in a previous slide, but in about 2/3rd of Monte Carlo simulations, p is within one standard error of $\hat{p}$, so the procedure should have about a 2/3rd chance of making an interval that covers p.

We can easily verify that with a simulation.

Confidence interval for p illustration

Let's take p to be 0.32 and n to be 800. We'll run a simple Monte Carlo simulation to find $\hat{p}$ after 800 trials, calculate the SE by plugging in $\hat{p}$ into our formula, construct the interval of plausible values for p, and see if that interval contains 0.32. We'll repeat the Monte Carlo simulation with different random number streams a total of 10000 times and look at the fraction of them where the interval was correct.

```
counter <- 0
for (i in 1:10000) {
  observed <- sample( c(1,0), prob=c(0.32,0.68), size=800, replace=TRUE )
  phat <- sum(observed)/800
  SE <- sqrt( phat*(1-phat)/800 )
  if( phat - SE <= 0.32 & 0.32 <= phat + SE ) { counter <- counter + 1 }  #Was .32 in interval?  If so, a
}
counter/10000

## [1] 0.6973
```

As claimed, the procedure constructing the interval has about a 2/3rds chance of covering p.

Confidence interval for p definition

The probability that our procedure (constructing a range of plausible values by calculating $\hat{p} \pm 2SE$) provides an interval that covers p looks to be about 2/3rds.

In probability terms, we would say that $\hat{p} \pm SE$ is a 68% **confidence interval** because the **procedure** making the interval has a 68% chance of covering p.

Other levels of confidence

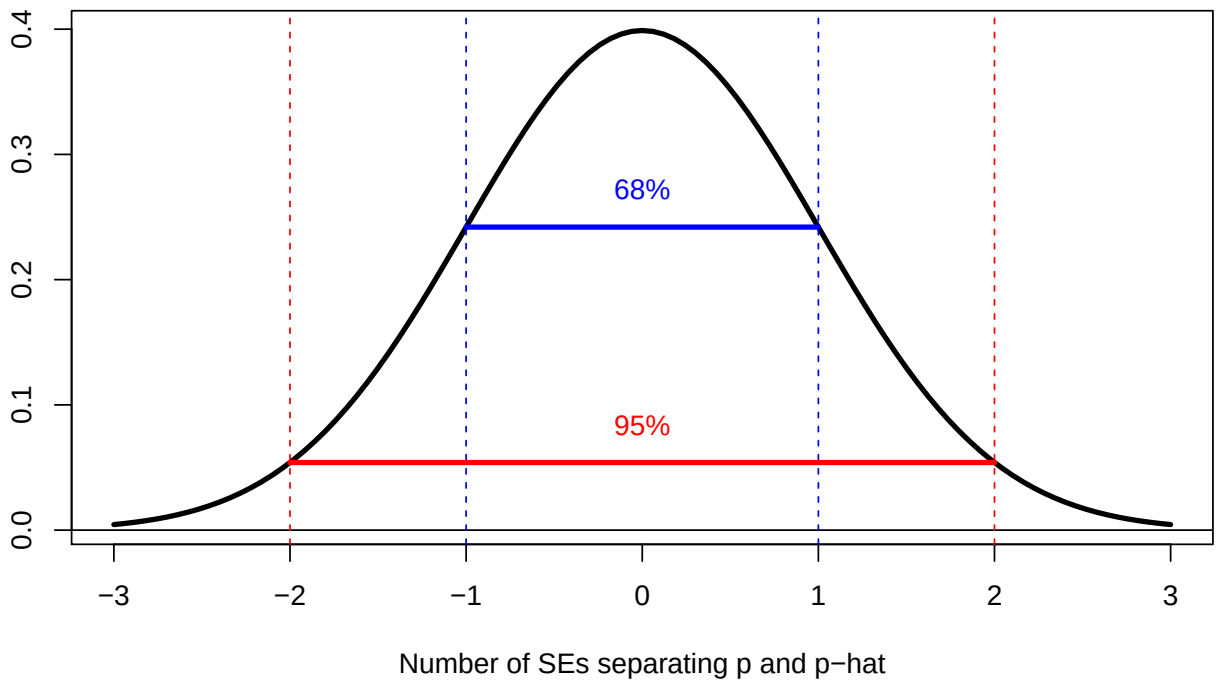
You might not be comfortable with a procedure that "only" has about a 2 in 3 chance of "getting it right" (providing a range of values that covers p).

To construct a confidence interval with a higher level of confidence, i.e., design a procedure that gives an interval with a higher probability of covering p, we need to move *more* than $\pm SE$ away from $\hat{p}$ so that the interval includes more values.

But by how much? Our previous investigation showed that the number of standard errors separating $\hat{p}$ and p was well-described by a bell-curve. Properties of this curve are well known!

Higher levels of confidence

We can determine how many SEs are required for a specific level of confidence by finding the "right" location on the bell-curve. For example, $\hat{p} \pm 2SE$ should yield a 95% confidence interval since it turns out that 95% of the area under the curve lies within 2 SEs of the peak.



Confidence Interval Logic Summary

- If we asked a bunch of people to each run a Monte Carlo simulation to estimate the probability of some event, everyone would come back with slightly different $\hat{p}$s.
- Some $\hat{p}$s would be bigger than p, some would be smaller, but the distribution of the different measured $\hat{p}$s will resemble a bell-curve.
- The bell-curve's shape is predictable. We'll use the distance between the peak and where the curve changes from sloping down to sloping out as the "typical difference" (standard error SE) between $\hat{p}$ and p.
- We've see that about 68% of those people who ran Monte Carlo simulations will measure a $\hat{p}$ within 1 SE of p. $\hat{p} \pm SE$ is a 68% confidence interval for p.
- It turns out about 95% of those people who ran Monte Carlo simulations will measure a $\hat{p}$ within 2 SE of p. $\hat{p} \pm 2SE$ is a 95% confidence interval for p.

A 95% confidence interval for p

The "industry standard" is a 95% confidence interval. "95% confidence" means that the procedure we're using to generate the range of plausible values for p has a probability of *about* 0.95 of covering p (the *about* is key here; we've made a few guesses and approximations). In other words, "95% confidence" means that, in the long run, 95% of 95% confidence intervals cover p.

95% confidence interval for p

$$\hat{p} \pm 2SE$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Meaning of 95% confidence illustrated

The level of confidence of a confidence interval refers to the probability that the *procedure* used to construct the interval produces a range of values that includes p.

Let's conduct a bunch of Monte Carlo simulations and see how often the 95% confidence intervals for p constructed on them actually do include p.

Meaning of 95% confidence illustrated

Each Monte Carlo simulation here consists of 900 trials where our event occurs with $p = 0.25$ by design. Once the simulation concluded we can calculate $\hat{p}$ and SE and see if the 95% confidence interval covers p. The `for` loop here goes through 10000 such Monte Carlo simulations are reports the fraction whose 95% confidence intervals cover p: just about 95% as expected!

```
counter <-0  #To count up number of confidence intervals that cover p
for (i in 1:10000) {
  #Randomly generate 900 1s/0s (1=event happened) where p=0.25 and measure p-hat
  p.hat <- mean( sample( c(0,1), size=900, replace=TRUE, prob=c(.75,.25) ) )
  SE <- sqrt(p.hat*(1-p.hat)/900)
  #Check to see if p=0.25 is in the interval
  if( p.hat - 2*SE <= 0.25 & 0.25 <= p.hat + 2*SE ) { counter <- counter + 1 }
}
counter/10000

## [1] 0.9525
```

General levels of confidence

Although there is rarely a reason to construct an interval with other than 95% confidence, is there a way to know how many standard errors you would have to go from $\hat{p}$? Yes! We'll use the theoretical properties of the bell curve, which can be accessed using the `qnorm` function in `R`. For a desired level of confidence (written a a number between 0-1), run `qnorm( 1 - (1-confidence)/2)` .

```
qnorm( 1 - (1-0.95)/2 )  #95% confidence, go plus or minus 2 SEs
```

```
## [1] 1.959964
```

```
qnorm( 1 - (1-0.68)/2 )  #68% confidence, go plus or minus 1 SEs
```

```
## [1] 0.9944579
```

```
qnorm( 1 - (1-0.99)/2 )  #99% confidence, go plus or minus 2.58 SEs
```

```
## [1] 2.575829
```

Master formula for confidence interval for p

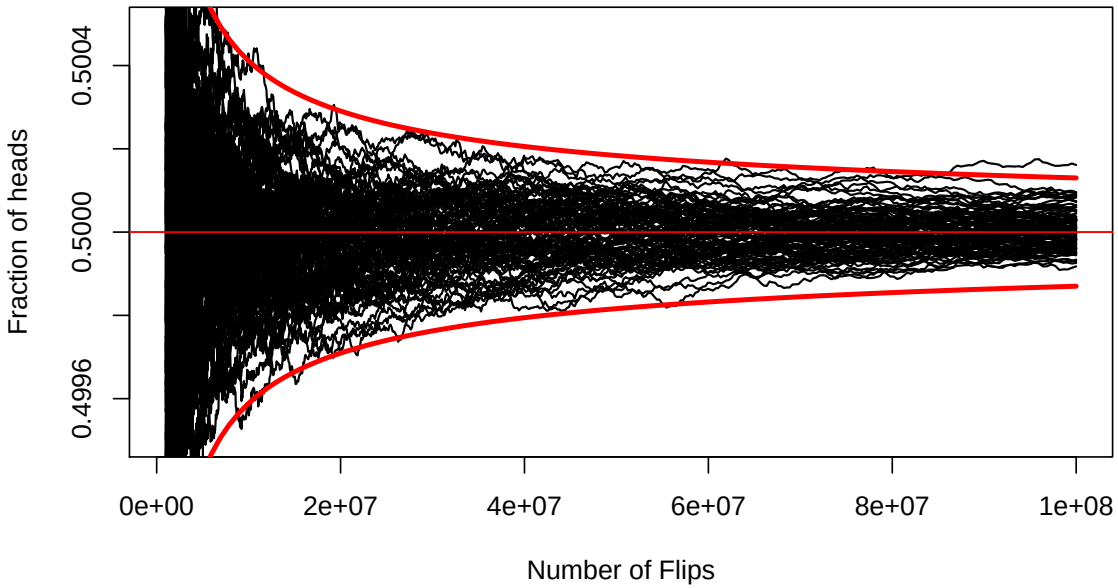
Writing the desired confidence as a number between 0-1

Arbitrary confidence interval for p

$$\hat{p} \pm \texttt{qnorm(1-(1-confidence)/2)} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

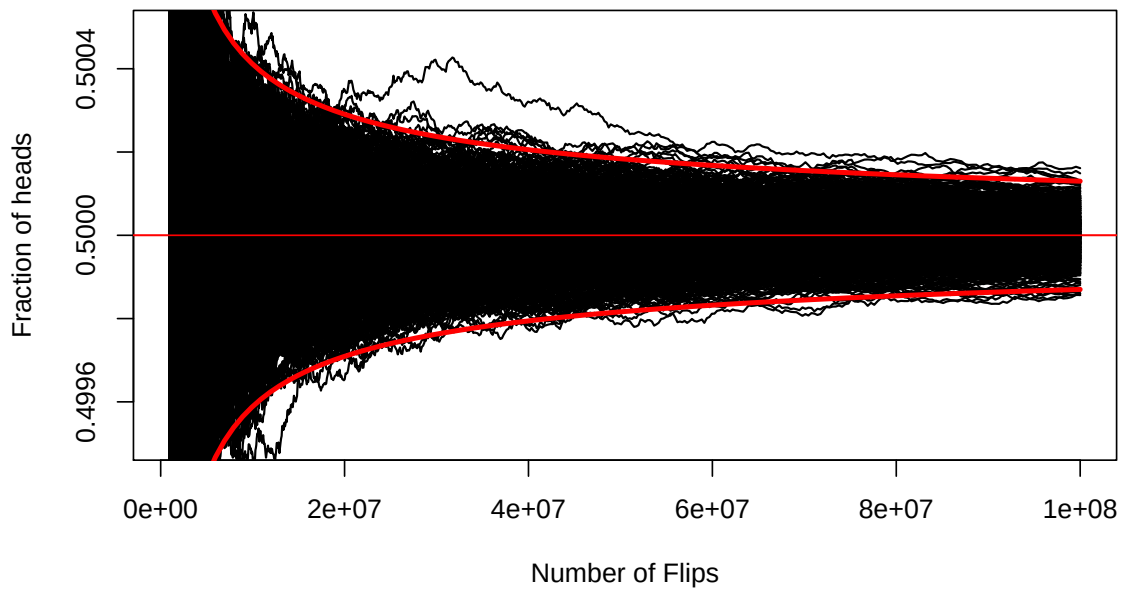
Size of confidence band vs. n

Just so you can see the formula working in action, the upper and lower red bands are the limits of a 99% confidence interval. Loosely, this means that $\hat{p}$ should be at most 2.6 SEs from p (the horizontal line) 99% of the time. Plotted are the trajectories of 100 different Monte Carlo simulations. The curve is doing a great job mimicking the behavior, and only 1 happens to stray outside the bands.



Size of confidence band vs. n

Now including the results from 1000 different Monte Carlo simulations, you can appreciate how well the formula for the upper and lower confidence limits are providing a fence that "most" simulations live within (check out that one oddball trajectory though).



Caution: confidence intervals are often misinterpreted

Unfortunately, despite their widespread use in all aspects of business, engineering, and science, confidence intervals remain one the most misinterpreted quantities in statistics. The goal in introducing confidence intervals in this (rather unorthodox) manner here in the notes is to try to avoid some of the common misunderstandings.

One of recruiters favorite interview questions is about confidence intervals and what they mean!

Let's see where misunderstandings occur and clear them up.

Misconception: the probability that p is inside the confidence interval is 95%

It's very tempting to construct a 95% confidence interval, say it turns out to be (0.548, 0.612), and to say "the probability that p is between 0.548 and 0.612 is 95%". Why is that wrong?

- p is not a random quantity; we just don't know it's value. As we have discussed at length, under the frequentist definition, we don't use probability to quantify our ignorance of reality. If we had a way to "see" the value of p, it would be the same value every time we checked!
- Sometimes our confidence intervals cover p, sometimes they don't. In developing the confidence interval, we invented a procedure to determine the end points: namely taking $\hat{p}$ and going $\pm 2SE$. The probability that the *procedure* creates an interval that covers p is 95%.

Misconception: the probability that p is inside the confidence interval is 95%

It might take a while for it to sink in that "the probability that p is inside the confidence interval is 95%" and "there is a 95% chance that the procedure which created the confidence interval manages to cover p" are two *very* different statements (the former being wrong, the later being correct), but take the time to do so.

You'll never know if a particular confidence interval actually covers p

Since the *procedure* generating the interval has a 95% chance of coming up with a range of values that covers p, that means some of your confidence intervals will contain p and others won't.

Since you don't know what p is (and never will), you will never know whether a particular confidence interval you made actually covers p!

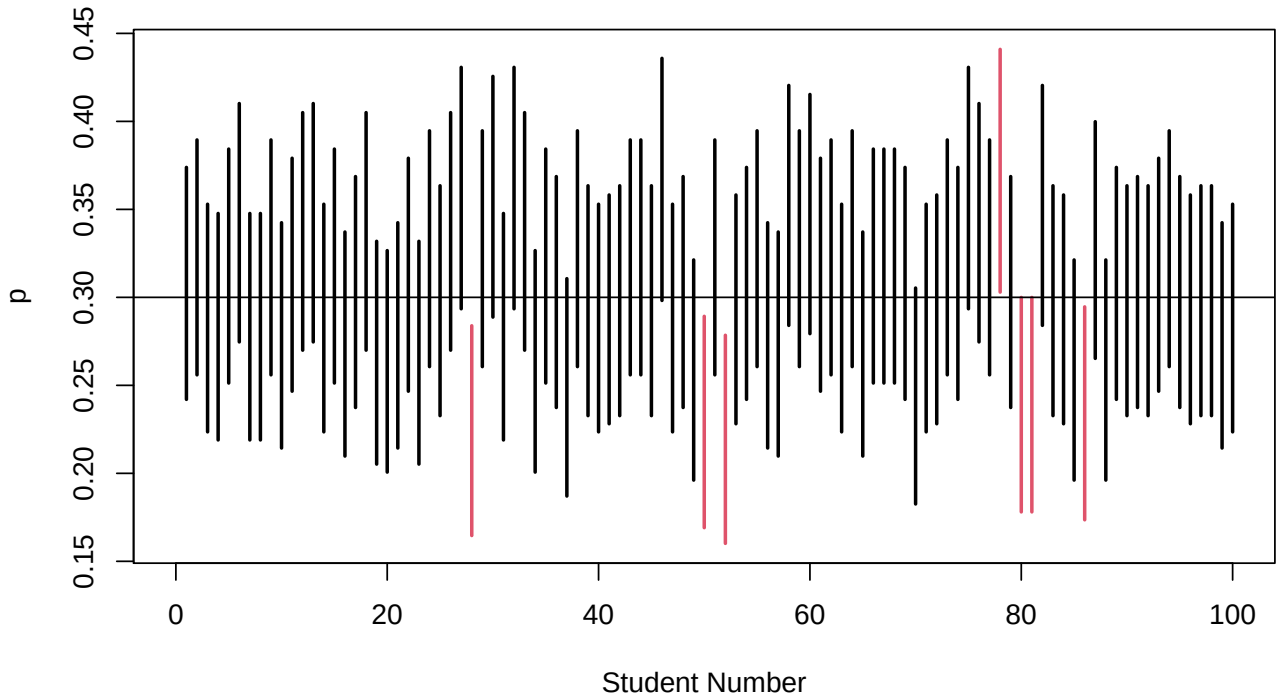
This is kind of depressing! However, what other job out there (besides weatherman) can you be praised for getting it right "only" 95% of the time?

You'll never know if your confidence interval actually covers p illustration

Imagine 100 different students run a Monte Carlo simulation to estimate the probability that some event occurs (which *we* happen to know is exactly $p = 0.30$). Each does so by running 200 trials, finding $\hat{p}$, and constructing a 95% confidence interval.

You'll never know if your confidence interval actually covers p illustration

Each vertical line represents the confidence interval of one of the students. We see that most end up with an interval that covers 0.3. Others (in red) come up with intervals that miss 0.3 entirely (with the entire range being too low or too high). Unfortunately for the students, they'll never know if *their* interval covered p (only we are privy to that information since *we* know $p = 0.3$).



Alternative (recommended) procedure: use `binom.test`

Hopefully, after a few readings of the notes, you'll "get" the intuition behind a confidence interval. The formula that we've derived was used for decades to generate confidence intervals ... but there's a better way now.

The `binom.test` command outputs a confidence interval whose endpoints are chosen a little bit more carefully (specifically, instead of using the bell curve, it uses exact probabilities for observing a certain number of events after a given number of trials), and it should be used whenever possible.

Syntax: `binom.test(x,n,conf.level)`

- `x` - number of times event of interest occurred
- `n` - number of trials
- `conf.levels` - desired level of confidence (a number between 0 and 1)

`binom.test` example

Out of $n = 1000$ trials, the event occurred 312 times.

- $\hat{p} = 312/1000 = 0.312$
- $SE = \sqrt{0.312 \cdot 0.688/1000} = 0.01465114$
- "Classic" 95% CI is $\hat{p} \pm 2SE$ or (0.297, 0.327)

The confidence interval from `binom.test` is:

```
binom.test(312,1000,conf.level=0.95)$conf.int

## [1] 0.2833732 0.3417384
## attr(,"conf.level")
## [1] 0.95
```

Contradiction?

How can *both* $(0.297, 0.327)$ and $(0.283, 0.342)$ simultaneously be valid 95% confidence intervals?

- Both are valid because *each* procedure making the interval has a 95% chance of producing an interval that cover p.
- Perhaps both of these intervals cover p, or maybe only one does, or maybe both don't. Since we don't know p, who knows!
- Although it's tempting to use the values where these intervals overlap to further narrow down the set of plausible values of p, that's not how it works – don't do it.

Preference for `binom.test`

The $\hat{p} \pm 2SE$ is a classic and time tested procedure for making a 95% confidence interval, but in this day and age it is, in my opinion, obsolete. The `binom.test` command in R provide an alternative procedure for making a 95% confidence interval that relies on fewer approximations and guesses.

What practical difference is there between the two?

- The "classic" way provides a procedure with *about* 95% confidence (could be higher, could be lower).
- `binom.test` provides a procedure with *at least* 95% confidence.

How much confidence does the procedure *really* have?

By design, a 95% confidence interval *should* provide a range of values that covers p for 95% of random samples. We can compare how often each procedure actually *does* provide such an interval with the target of 95%.

- Simulate 87 visits to a stoplight that has a 20% chance of being green and get $\hat{p}$, the proportion of visits where a green light occurs
- Run `binom.test` and also construct $\hat{p} \pm 2SE$. Check to see if these intervals cover 0.2.
- Repeat the first two steps a million times. See what percentage of confidence intervals made with `binom.test` covered 0.2 and see what percentage of confidence intervals made with $\hat{p} \pm 2SE$ covered 0.2.

`binom.test` vs. $\hat{p} \pm 2SE$

```
set.seed(471)
n <- 87
counter.classic <- counter.binom <- 0
for( trial in 1:1e6 ) {
  n.green <- sum( sample( c(0,1), size=n, prob=c(.8,.2), replace=TRUE) ) #number of green lights
  p.hat <- n.green / n  #proportion of 100 visits that came up with green lights
  BINOM <- binom.test(n.green,n)$conf.int
  lower.binom <- BINOM[1]; upper.binom <- BINOM[2]  #lower and upper limits from binom.test
  lower.classic <- p.hat - 2*sqrt(p.hat*(1-p.hat)/n) #lower limit from phat +/ 2 SE
  upper.classic <- p.hat + 2*sqrt(p.hat*(1-p.hat)/n) #upper limit from phat +/ 2 SE
  #if 0.2 in interval, take note!
  if( lower.classic <= 0.2 & 0.2 <= upper.classic ) { counter.classic <- counter.classic + 1 }
  if( lower.binom <= 0.2 & 0.2 <= upper.binom ) { counter.binom <- counter.binom + 1 }
}
counter.classic/1e6
#0.929994
counter.binom/1e6
#0.967286
```

The classic method provides a "95% confidence interval" that really has something more like 93.0% confidence. The `binom.test` method provides a "95% confidence interval" that really has something more like 96.7% confidence.

`binom.test` vs. $\hat{p} \pm 2SE$

In our experiment:

- `binom.test` achieved a level of confidence of 96.9% (since 96.9% of the time it provided an interval that covered 0.2).
- The classic $\hat{p} \pm 2SE$ achieved a level of confidence of 93.0%

The achieved level of confidence for `binom.test` is just a little bit closer to the target of 95% than the classic interval!

`binom.test` vs. $\hat{p} \pm 2SE$

Problem: the achieved level of confidence of the classic interval is *below* the target!

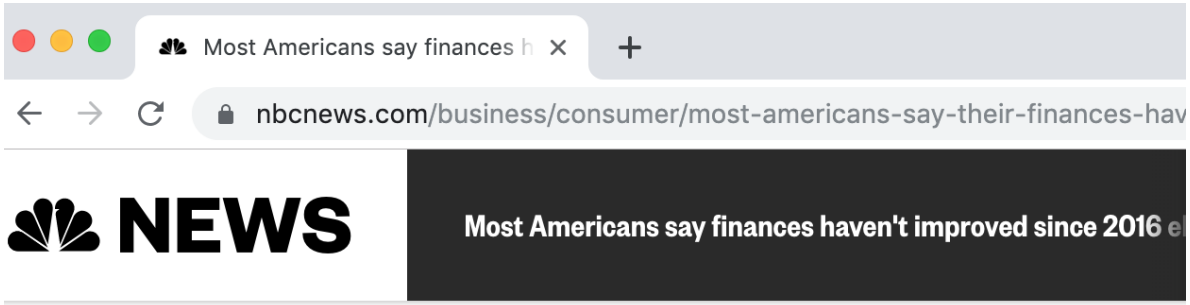
From my point of view, I prefer being wrong less frequently than expected instead of more frequently than expected. Since `binom.test` provides an interval with *at least* the target level, it provides a range of values that *doesn't* cover p at most 5% of the time.

The classic interval doesn't have such a guarantee: the confidence interval may fail to cover p a little less than 5% of the time, or it may fail to cover p substantially more than 5% of the time.

Margin of Error and Sample Size Calculation

Margin of Error vs. Standard Error

In the news, you might be more accustomed to hearing the phrase "margin of error" when talking about the results of a poll instead of "standard error". What's the difference?



The survey was conducted from Sept. 25-30 among 1,001 respondents and has a margin of error of plus or minus 3.1 percentage points at the 95 percent confidence level.

Margin of Error vs. Standard Error

Because the industry standard is to construct a 95% confidence interval, the "margin of error" is just the half-width of the 95% confidence interval.

$$95\%\ \text{CI:} \qquad \hat{p} \pm ME$$

In other words, the margin of error is just double the standard error (since a 95% CI is just $\hat{p} \pm 2SE$).

Margin of Error vs. Standard Error

In the quoted poll, 62% of Americans don't believe their financial situation has improved since the last presidential election in 2016. The percentage comes from from 1001 individuals.

$$2 \times \sqrt{0.62 \cdot 0.38/1001} = 0.031$$

This is consistent with the article, which quotes a margin of error of 3.1 percentage points.

Sample Size Calculations

It can be the case that trials are difficult to perform and each one comes at a substantial cost in time or dollars. You may want to estimate the number of trials (sample size) required to obtain a particular margin of error. We derived all the math for that earlier!

$$ME = 2SE = 2\sqrt{\frac{p(1-p)}{n}}$$

Solving for n:

$$n = \frac{4p(1-p)}{ME^2}$$

We run into the issue that we need to know p to find the required sample size for a given margin of error.

Sample Size Calculations

Two options:

- Conduct a small pilot study to get an estimate of p and use that in the formula.
- Just plug in $p = 0.5$, since the *largest* n will ever need to be to achieve the desired margin of error is when $p = 0.5$. In this case you can show that $n \leq 1/ME^2$

Sample Size Calculation Example

Example: it's desired that a 95% confidence interval should have a ME of 0.01 (1 percentage point).

1) If no information is available as to what p might be, how many trials do we need to perform to achieve that margin of error?

Solution: plug in $p = 0.5$ into the formula to get an upper bound on the number of trials.

$$n \leq \frac{1}{ME^2} \qquad n \leq \frac{1}{0.01^2} \qquad n \leq 10000$$

2) If we think $p \approx 7\%$, about how many trials do we need to perform to achieve that margin of error?

Solution: plug in $p = 0.07$ into the formula to get a rough idea of the number of trials.

$$n \approx \frac{4p(1-p)}{ME^2} \qquad n \approx \frac{4 \cdot 0.07 \cdot 0.93}{0.01^2} \qquad n \approx 2604$$