Why the Mean Matters

References:

Chapter 14

https://www.inferentialthinking.com/chapters/14/Why_the_Mean_Matters.html

Why the Mean Matters

- Several different statistics,
  - total variation distance,
  - maximum,
  - median, and
  - mean.
- Under clear assumptions about randomness, we have drawn empirical distributions of all of these statistics.
- Some, like the maximum and the total variation distance, have distributions that are clearly skewed in one direction or the other.
- The empirical distribution of the sample mean has almost always turned out close to bell-shaped, regardless of the population being studied

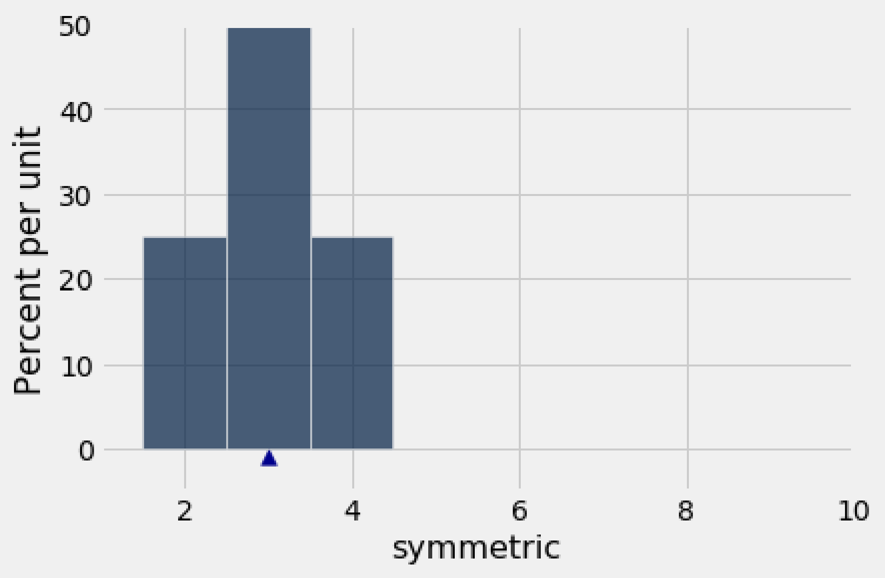
Why the Mean Matters

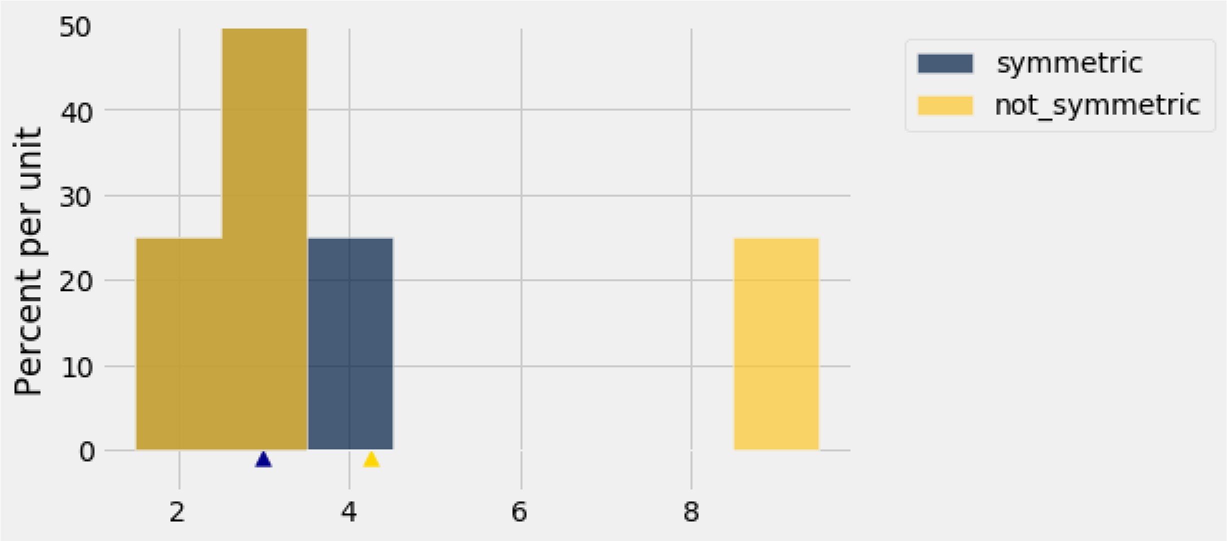
- If a property of random samples is true *regardless of the population,* it becomes a powerful tool for inference because we rarely know much about the data in the entire population.
- The distribution of the mean of a large random sample falls into this category of properties.
- That is why random sample means are extensively used in data science.
  - What exactly does the mean measure?
  - How close to the mean are most of the data?
  - How is the sample size related to the variability of the sample mean?
  - Why do empirical distributions of random sample means come out bell shaped?
  - How can we use sample means effectively for inference?

Definition and Basic Properties of Mean

- **Definition.** The *average* or *mean* of a collection of numbers is the sum of all the elements of the collection, divided by the number of elements in the collection.
- **Basic Properties**
  - It need not be an element of the collection.
  - It need not be an integer even if all the elements of the collection are integers.
  - It is somewhere between the smallest and largest values in the collection.
  - It need not be halfway between the two extremes; it is not in general true that half the elements in a collection are above the mean.
  - If the collection consists of values of a variable measured in specified units, then the mean has the same units too.

- **The Mean is a "Smoother"**
- If a collection consists only of ones and zeroes, then the sum of the collection is the number of ones in it, and the mean of the collection is the proportion of ones. Proportions are a special case of means, results about random sample means apply to random sample proportions as well.
- Therefore, **if two collections have the same distribution, then they have the same mean.**
- **The mean is the center of gravity or balance point of the histogram.**
- In general, **for symmetric distributions, the mean and the median are equal.**
- In general, **if the histogram has a tail on one side (the formal term is "skewed"), then the mean is pulled away from the median in the direction of the tail.**
- Distributions of incomes of large populations tend to be right skewed. When the bulk of a population has middle to low incomes, but a very small proportion has very high incomes, the histogram has a long, thin tail to the right.
- Economists often summarize income distributions by the median instead of the mean.





Variability

- The mean tells us where a histogram balances.
- But in almost every histogram we have seen, the values spread out on both sides of the mean.
- How far from the mean can they be?
- To answer this question, we will develop a measure of variability about the mean.

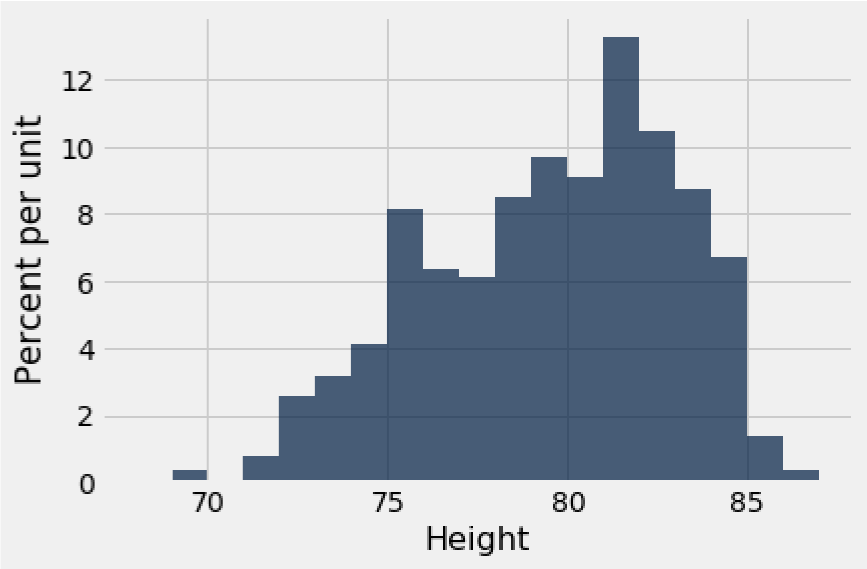
Variability

	Value	Deviation from Average	Squared Deviations from Average
	1	-2.75	7.5625
	2	-1.75	3.0625
	2	-1.75	3.0625
	10	6.25	39.0625
Mean	**3.75**	**0**	**13.1875**

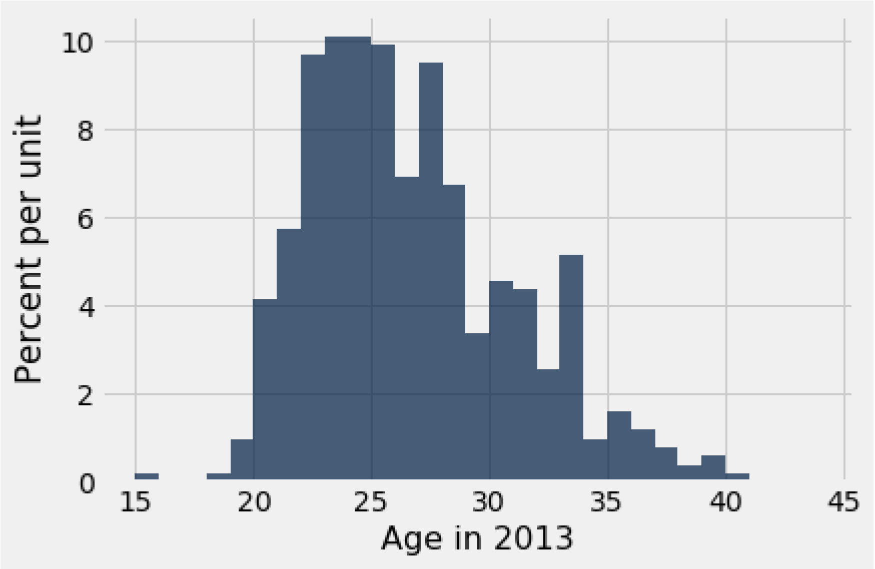
- The sum of the deviations from average is zero
- **Variance:** The mean squared deviation calculated above is called the *variance* of the values.
- The SD of a list is defined as the *root mean square of deviations from average*. read it from right to left and you have the sequence of steps in the calculation.
- No matter what the shape of the histogram, the average and the SD together tell you a lot about where the histogram is situated on the number line.

First main reason for measuring spread by the SD

- In all numerical data sets, the bulk of the entries are within the range "average ± a few SDs".
- In the nba2013.csv,



	Mean	SD	Tallest	(x- μ) / σ	Shortest	(x- μ) / σ
Height	79.06	3.45	87	2.29	69	-2.91
Age	26	4.3	40	3.19	15	-2.58



For all lists, the bulk of the entries are no more than 2 or 3 SDs away from the average.

Chebychev's Bounds

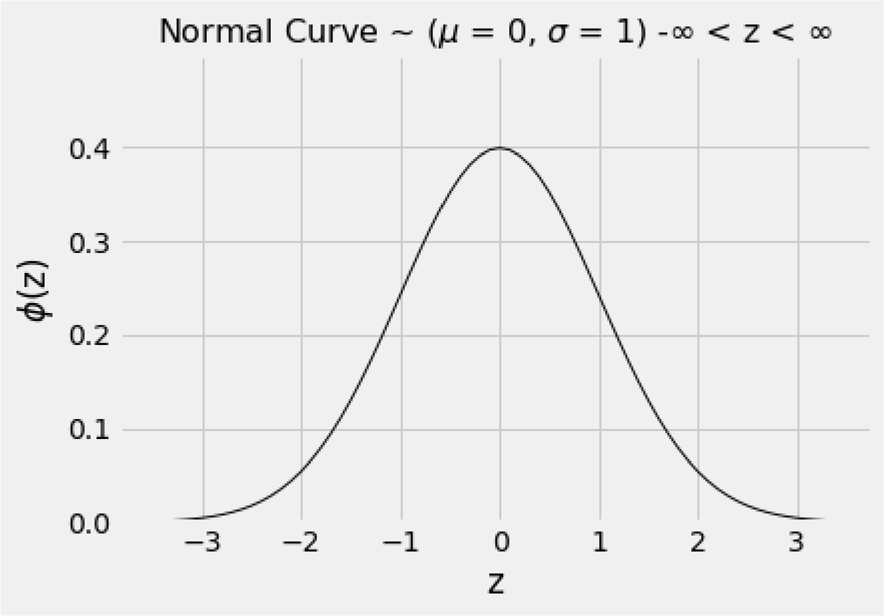
- The Russian mathematician Pafnuty Chebychev (1821-1894) proved a result that makes our rough statements precise.
- For all lists, and all numbers z , the proportion of entries that are in the range "average ±z SDs" is at least $1-1/z^2$.
  - **the proportion in the range "average ± 2 SDs" is at least 1 - 1/4 = 0.75**
  - **the proportion in the range "average ± 3 SDs" is at least 1 - 1/9 ≈ 0.89**
  - **the proportion in the range "average ± 4.5 SDs" is at least $1 - 1/(4.5)^2$ ≈ 0.95**
- It is important to note that the result gives a bound, not an exact value or an approximation.
- What makes the result powerful is that it is true for all lists – all distributions, no matter how irregular.
- For example, the percent of entries in the range "average ± 2 SDs" might be quite a bit larger than 75%. But it cannot be smaller.

Standard units

- In the calculations above, the quantity z measures standard units, the number of standard deviations above average.
- Some values of standard units are negative, corresponding to original values that are below average.
- Other values of standard units are positive.
- But no matter what the distribution of the list looks like, Chebychev's bounds imply that standard units will typically be in the (-5, 5) range.

The SD and the Normal Curve

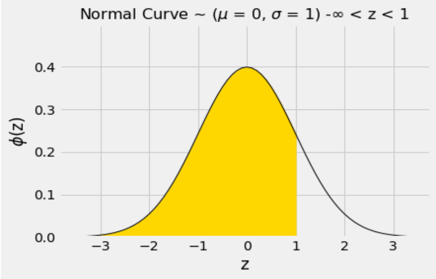
- In general, **for bell-shaped distributions, the SD is the distance between the mean and the points of inflection on either side.**
- we will use the units into which we can convert every list: standard units. The resulting curve is therefore called the *standard normal curve*
    - The total area under the curve is 1. So you can think of it as a histogram drawn to the density scale.
    - The curve is symmetric about 0. So if a variable has this distribution, its mean and median are both 0.
    - The points of inflection of the curve are at -1 and +1.
    - If a variable has this distribution, its SD is 1. The normal curve is one of the very few distributions that has an SD so clearly identifiable on the histogram.



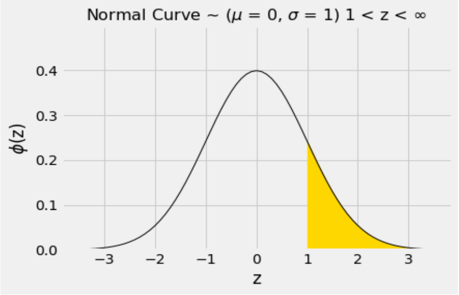
- Since we are thinking of the curve as a smoothed histogram, we will want to represent proportions of the total amount of data by areas under the curve. However, that the standard normal curve cannot be integrated in any of the usual ways of calculus.
- Therefore, areas under the curve have to be approximated.
- That is why almost all statistics textbooks carry tables of areas under the normal curve. It is also why all statistical systems, including a module of Python, include methods that provide excellent approximations to those areas

The standard normal "cdf"

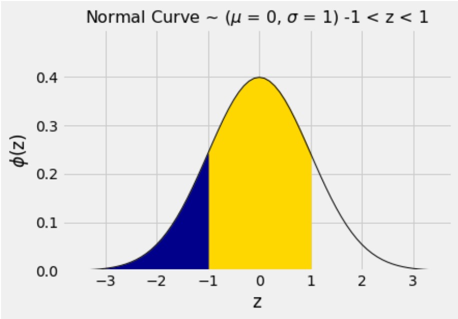
- The fundamental function for finding areas under the normal curve is stats.norm.cdf.
- It takes a numerical argument and returns all the area under the curve to the left of that number
- Formally, it is called the "cumulative distribution function" of the standard normal curve.



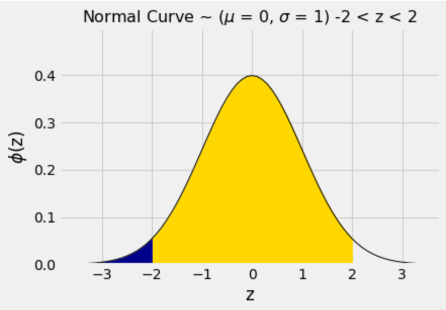
stats.norm.cdf(1) = 0.84



1 - stats.norm.cdf(1) = 0.16



stats.norm.cdf(1) - stats.norm.cdf(-1) = 0.68



stats.norm.cdf(2) - stats.norm.cdf(-2) = 0.95

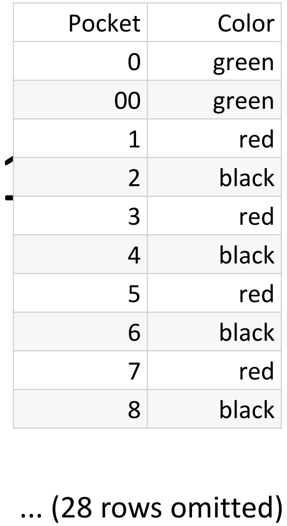
Chebychev’s Bound and Normal Distributions

Percent in Range	All Distributions: Bound	Normal Distribution: Approximation
Average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

- Chebychev's bound is weaker because it has to work for all distributions.
- If we know that a distribution is normal, we have good approximations to the proportions, not just bounds.

Discussions leading to The Central Limit Theorem

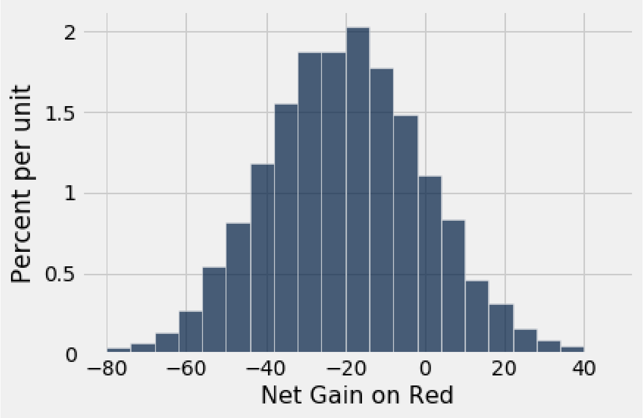
- Very few of the data histograms that we have seen in this course have been bell shaped.
- When we have come across a bell shaped distribution, it has almost invariably been an empirical histogram of a statistic based on a random sample
- The bell appeared as the rough shape of the total amount of money we would make if we placed the same bet repeatedly on different spins of a roulette wheel.
- Specifically, the function takes a color as its argument and returns 1 the color is red. For all other colors it returns -1.

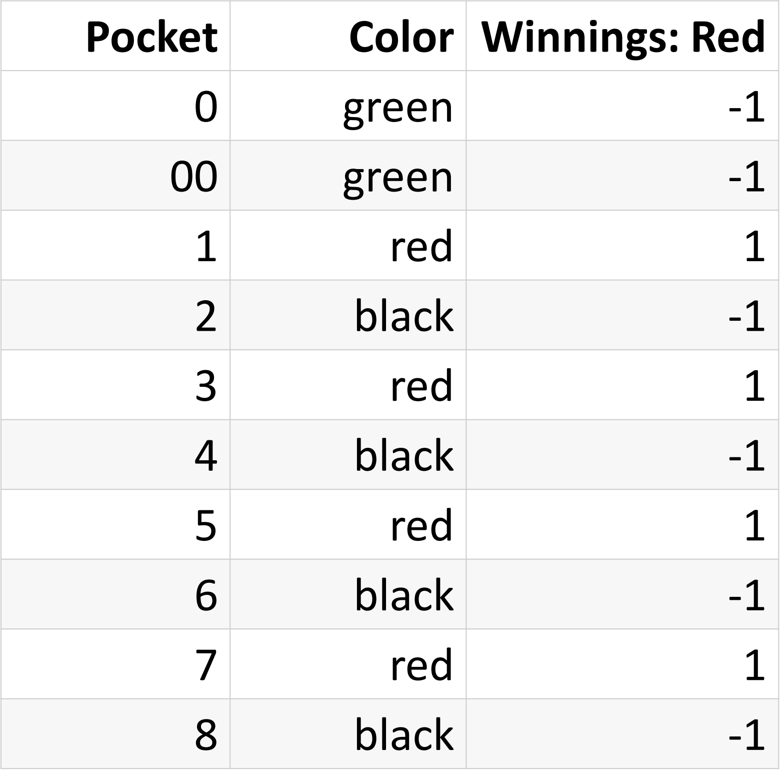


Pocket	Color
0	green
00	green
1	red
2	black
3	red
4	black
5	red
6	black
7	red
8	black

... (28 rows omitted)

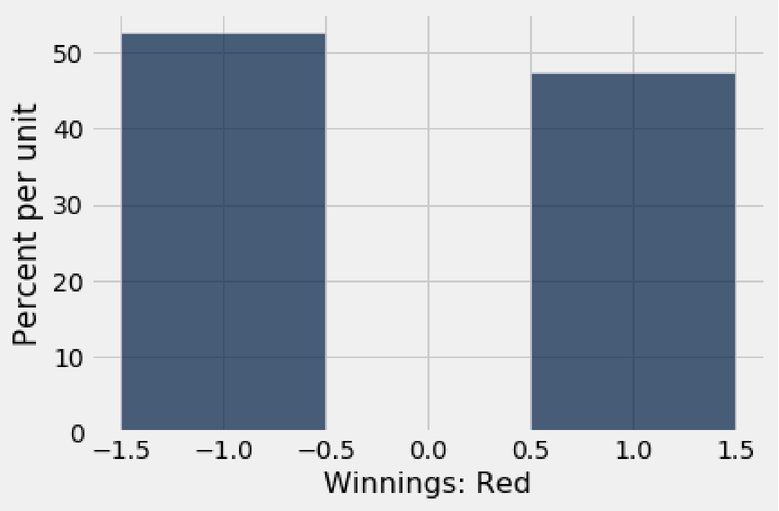
- Your net gain on one bet is one random draw from the Winnings: Red column. There is an 18/38 chance making $1, and a 20/38 chance of making -$1. This probability distribution is shown in the histogram below.
- Now suppose you bet many times on red. Your net winnings will be the sum of many draws made at random with replacement from the distribution above.
- It will take a bit of math to list all the possible values of your net winnings along with all of their chances. We won't do that; instead, we will approximate the probability distribution by simulation, as we have done all along in this course.
- We simulate your net gain if you bet $1 on red on 400 different spins of the roulette wheel





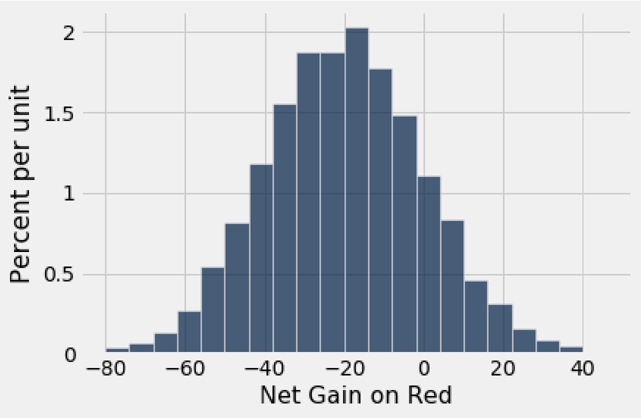
Pocket	Color	Winnings: Red
0	green	-1
00	green	-1
1	red	1
2	black	-1
3	red	1
4	black	-1
5	red	1
6	black	-1
7	red	1
8	black	-1

... (28 rows omitted)



Discussions leading to The Central Limit Theorem

- **Center**. The distribution is centered near -20 dollars, roughly.
- To see why, note that your winnings will be $1 on about 18/38 of the bets, and -$1 on the remaining 20/38. So your average winnings per dollar bet will be roughly -5.26 cents
- average_per_bet = 1*(18/38) + (-1)*(20/38) = -0.0526
- So in 400 bets you expect that your net gain will be about -$21:
- **Spread.** The center is roughly -$20, which means that the SD of the distribution is around $20.

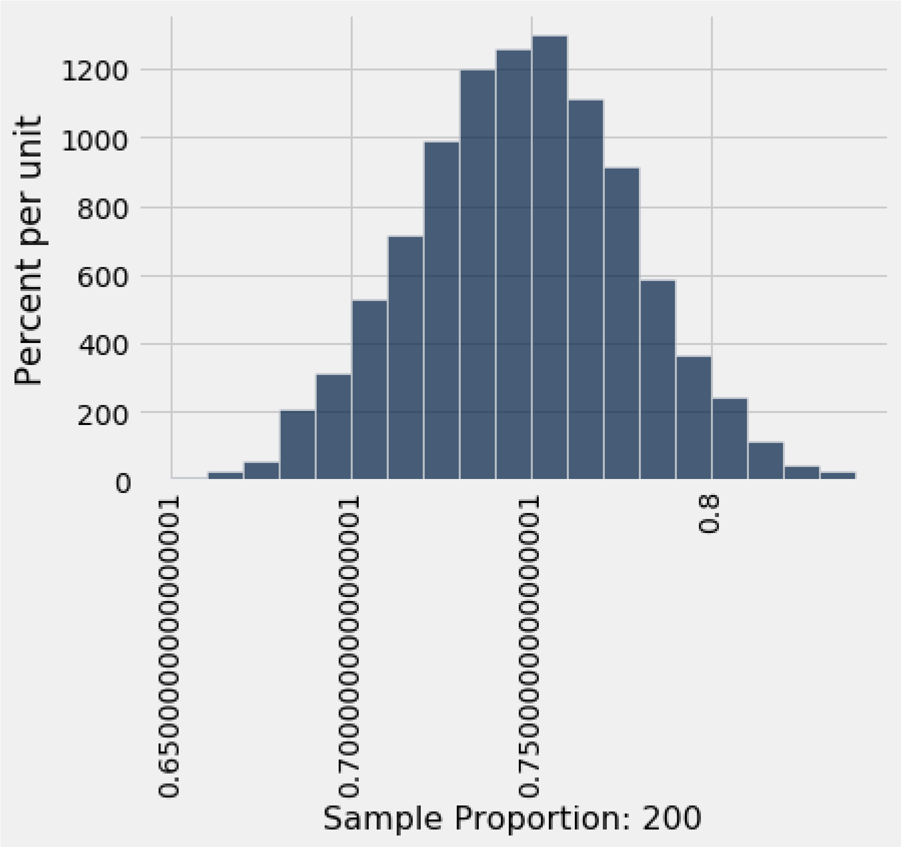


Central Limit Theorem

- The reason why the bell shape appears in such settings is a remarkable result of probability theory called the **Central Limit Theorem**.
- **The Central Limit Theorem says that the probability distribution of the sum or average of a large random sample drawn with replacement will be roughly normal, *regardless of the distribution of the population from which the sample is drawn*.**
- As we noted when we were studying Chebychev's bounds, results that can be applied to random samples *regardless of the distribution of the population* are very powerful, because in data science we rarely know the distribution of the population.
- The Central Limit Theorem makes it possible to make inferences with very little knowledge about the population, provided we have a large random sample. That is why it is central to the field of statistical inference.

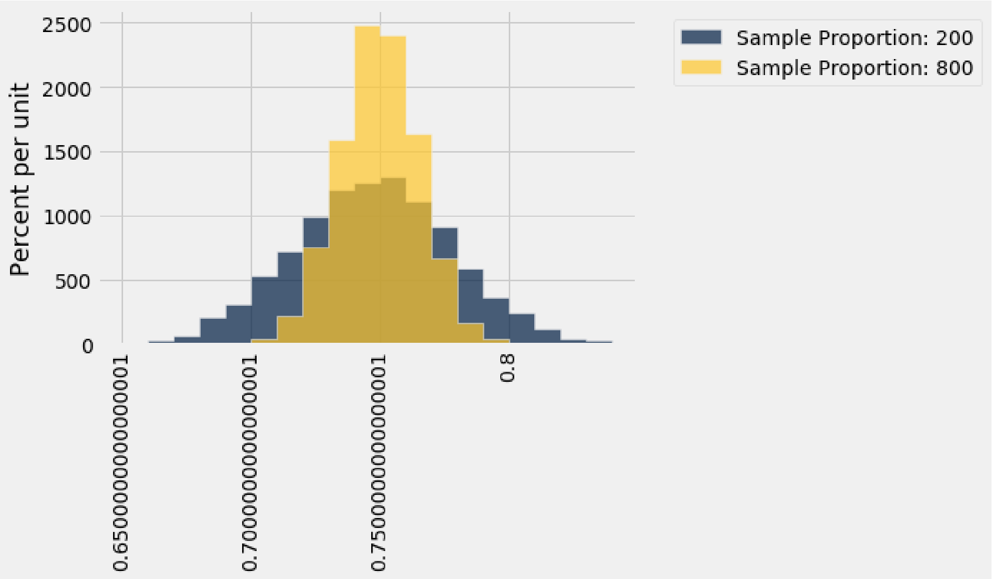
Proportion of Purple Flowers

- Recall Mendel's probability model for the colors of the flowers of a species of pea plant. The model says that the flower colors of the plants are like draws made at random with replacement from {Purple, Purple, Purple, White}.
- In a large sample of plants, about what proportion will have purple flowers?
- We would expect the answer to be about 0.75, the proportion purple in the model. And, because proportions are means, the Central Limit Theorem says that the distribution of the sample proportion of purple plants is roughly normal.
- We can confirm this by simulation. Let's simulate the proportion of purple-flowered plants in a sample of 200 plants.



Proportion of Purple Flowers

- There's that normal curve again, as predicted by the Central Limit Theorem, centered at around 0.75 just as you would expect.
- How would this distribution change if we increased the sample size?
- Let's run the code again with a sample size of 800, and collect the results of simulations in the same table in which we collected simulations based on a sample size of 200.
- We will keep the number of repetitions the same as before so that the two columns have the same length.
- Both distributions are approximately normal but one is narrower than the other.
- The proportions based on a sample size of 800 are more tightly clustered around 0.75 than those from a sample size of 200. Increasing the sample size has decreased the variability in the sample proportion.

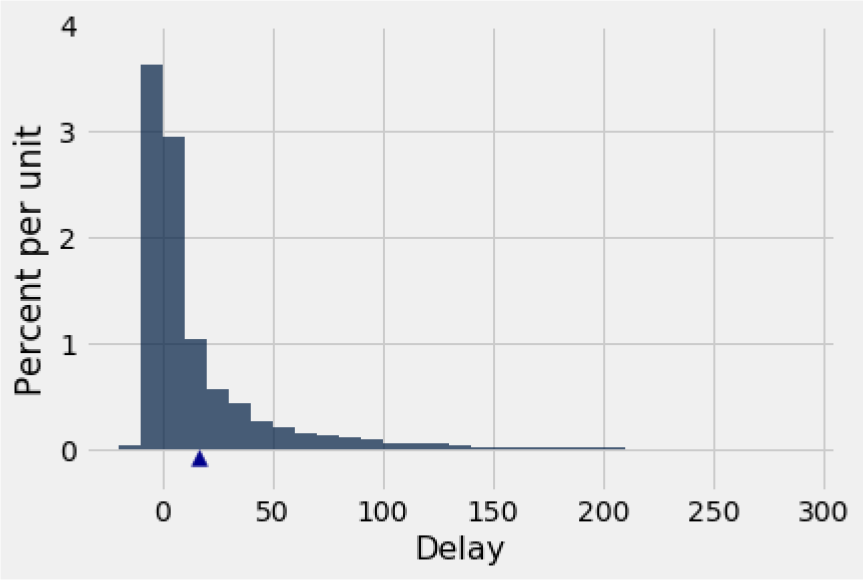


- We have leaned many times on the intuition that a larger sample size generally reduces the variability of a statistic. However, in the case of a sample average, we can *quantify* the relationship between sample size and variability.
- Exactly how does the sample size affect the variability of a sample average or proportion

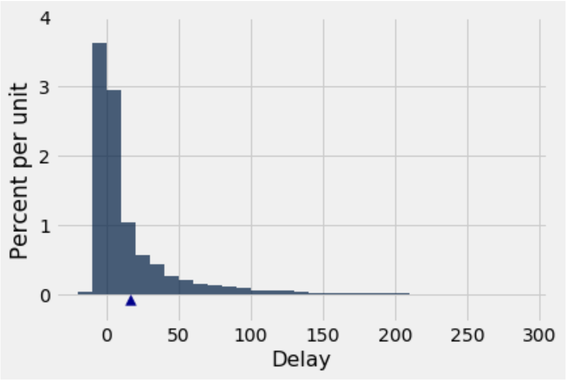
The Variability of the Sample Mean

- By the Central Limit Theorem, the probability distribution of the mean of a large random sample is roughly normal.
- The bell curve is centered at the population mean. Some of the sample means are higher, and some lower, but the deviations from the population mean are roughly symmetric on either side, as we have seen repeatedly.
- Formally, probability theory shows that the sample mean is an *unbiased* estimate of the population mean.
- In our simulations, we also noticed that the means of larger samples tend to be more tightly clustered around the population mean than means of smaller samples.
- In this section, we will quantify the variability of the sample mean and develop a relation between the variability and the sample size.

- In our simulations, we also noticed that the means of larger samples tend to be more tightly clustered around the population mean than means of smaller samples.
- In this section, we will quantify the variability of the sample mean and develop a relation between the variability and the sample size.
- The mean delay is about 16.7 minutes, and the distribution of delays is skewed to the right.
- Now let's take random samples and look at the probability distribution of the sample mean. As usual, we will use simulation to get an empirical approximation to this distribution.

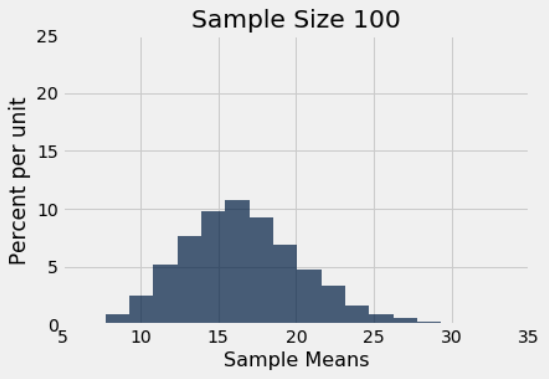


Central Limit Theorem

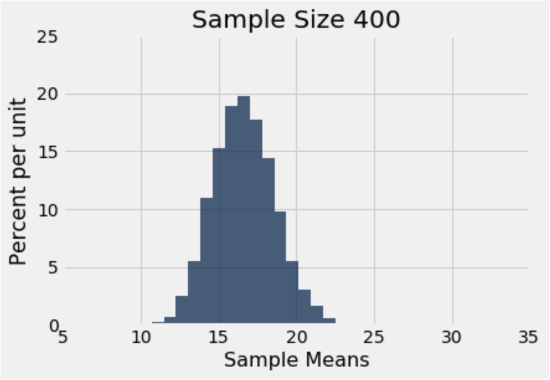


- Population Size: 13,825
- Population mean: 16.658
- Population SD: 39.48
- Let us simulate the mean of a random sample of 100 delays, then of 400 delays, and finally of 625 delays.
- We will perform 10,000 repetitions of each of these process.

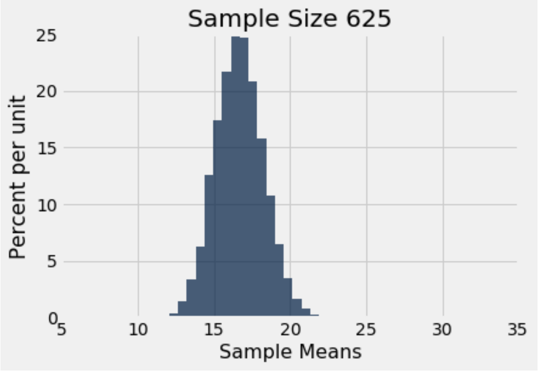
You can see the Central Limit Theorem in action – the histograms of the sample means are roughly normal, even though the histogram of the delays themselves is far from normal.



Average of sample means: 16.61
SD of sample means: 3.89



Average of sample means: 16.65
SD of sample means: 1.97



Average of sample means: 16.68
SD of sample means: 1.57

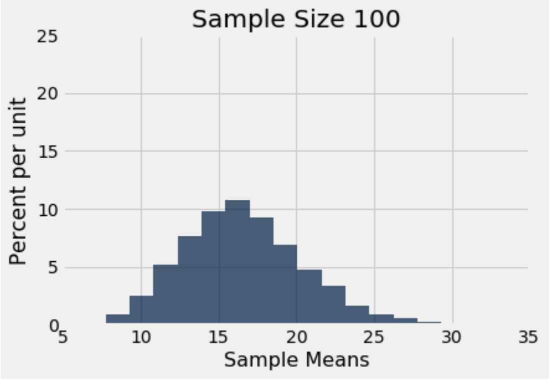
The SD of All the Sample Means

- You can also see that the histograms get narrower, and hence taller, as the sample size increases.
- We have seen that before, but now we will pay closer attention to the measure of spread.
- The SD of the population of all delays is about 40 minutes. (39.48)

- When the sample size is 100, the SD (3.89) is about one-tenth of the population SD (39.48)

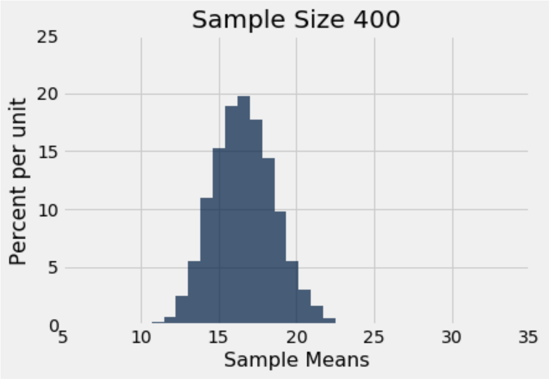
- When the sample size is 400, the SD (1.97) is about one-twentieth of the population SD (39.48)

- When the sample size is 625, the SD (1.57) is about one-twentyfifth of the population SD (39.48)



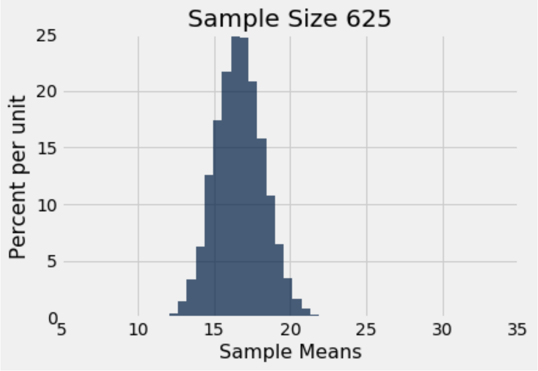
Average of sample means: 16.61

SD of sample means: 3.89



Average of sample means: 16.65

SD of sample means: 1.97

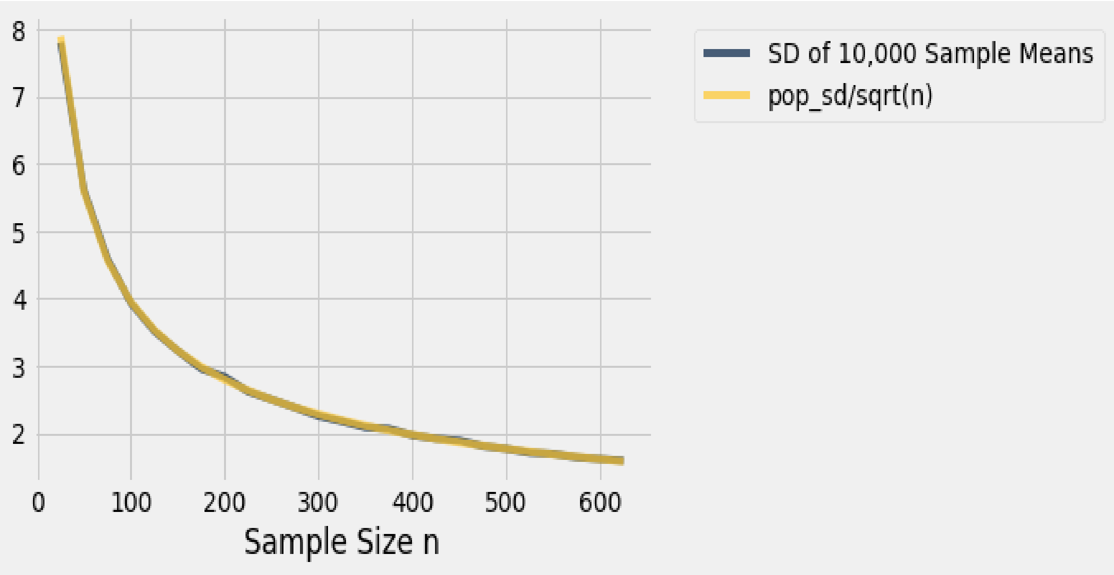


Average of sample means: 16.68

SD of sample means: 1.57

Standard Deviation Comparison

Sample Size n	SD of 10,000 Sample Means	pop_sd/sqrt(n)
25	7.80558	7.89604
50	5.60588	5.58334
75	4.58432	4.55878
100	3.93334	3.94802
125	3.51501	3.53122
150	3.22156	3.22354
175	2.95652	2.98442
200	2.83387	2.79167
225	2.61634	2.63201
250	2.49718	2.4969



- There really are two curves there. But they are so close to each other that it looks as though there is just one.
- The probability distribution of the sample mean is based on the means of *all possible samples* of a fixed size.

Sample Size

- Fix a sample size. If the samples are drawn at random with replacement from the population, then

  SD of all possible sample means = Population SD/(Sample Size)$^{1/2}$

- This is the standard deviation of the averages of all the possible samples that could be drawn.
- **It measures roughly how far off the sample means are from the population mean.**

The Central Limit Theorem for the Sample Mean

- If you draw a large random sample with replacement from a population, then, regardless of the distribution of the population, the probability distribution of the sample mean is roughly normal, centered at the population mean, with an SD equal to the population SD divided by the square root of the sample size.

The Accuracy of the Sample Mean

- The SD of all possible sample means measures how variable the sample mean can be.
- As such, it is taken as a measure of the accuracy of the sample mean as an estimate of the population mean.
- The smaller the SD, the more accurate the estimate.
- The formula shows that:
  - The population size doesn't affect the accuracy of the sample mean. The population size doesn't appear anywhere in the formula.
  - The sample size can be varied. Because the sample size appears in the denominator, the variability of the sample SD *decreases* as the sample size increases, and hence the accuracy increases.

The Square Root Law

- From the table of SD comparisons, you can see that the SD of the means of random samples of 25 flight delays is about 8 minutes.
- If you multiply the sample size by 4, you'll get samples of size 100. The SD of the means of all of those samples is about 4 minutes.
- That's smaller than 8 minutes, but it's not 4 times as small; it's only 2 times as small.
- That's because the sample size in the denominator has a square root over it.
- The sample size increased by a factor of 4, but the SD went down by a factor of 2 (square root of 4). In other words accuracy went up by factor of 2.
- In general, when you multiply the sample size by a factor, the accuracy of the sample mean goes up by the square root of that factor.
- So to increase accuracy by a factor of 10, you have to multiply sample size by a factor of 100. Accuracy doesn't come cheap!

Sample Size n	SD of 10,000 Sample Means	pop_sd/sqrt(n)
25	7.80558	7.89604
50	5.60588	5.58334
75	4.58432	4.55878
100	3.93334	3.94802
125	3.51501	3.53122
150	3.22156	3.22354
175	2.95652	2.98442
200	2.83387	2.79167
225	2.61634	2.63201
250	2.49718	2.4969

Choosing a Sample Size

- Candidate A is contesting an election.
- A polling organization wants to estimate the proportion of voters who will vote for her.
- Let's suppose that they plan to take a simple random sample of voters, though in reality their method of sampling would be more complex.
- How can they decide how large their sample should be, to get a desired level of accuracy?
- We are now in a position to answer this question, after making a few assumptions:
  - The population of voters is very large and that therefore we can just as well assume that the random sample will be drawn with replacement.
  - The polling organization will make its estimate by constructing an approximate 95% confidence interval for the percent of voters who will vote for Candidate A.
  - The desired level of accuracy is that the width of the interval should be no more than 1%. That's pretty accurate! For example, the confidence interval (33.2%, 34%) would be fine but (33.2%, 35%) would not.
- We will work with the sample proportion of voters for Candidate A.
- Recall that a proportion is a mean, when the values in the population are only 0 (the type of individual you are not counting) or 1 (the type of individual you are counting)

Width of Confidence Interval

- If we had a random sample, we could go about using the bootstrap to construct a confidence interval for the percent of voters for Candidate A.
- But we don't have a sample yet – we are trying to find out how big the sample has to be so that our confidence interval is as narrow as we want it to be.
- In situations like this, it helps to see what theory predicts.
- The Central Limit Theorem says that the probabilities for the sample proportion are roughly normally distributed, centered at the population proportion of 1's, with an SD equal to the SD of the population of 0's and 1's divided by the square root of the sample size.
- So the confidence interval will still be the "middle 95%" of a normal distribution, even though we can't pick off the ends as the 2.5th and 97.5th percentiles of bootstrapped proportions.
- Is there another way to find how wide the interval would be?
    - Yes, because we know that for normally distributed variables, the interval "center ± 2 SDs" contains 95% of the data.

Choosing a Sample Size

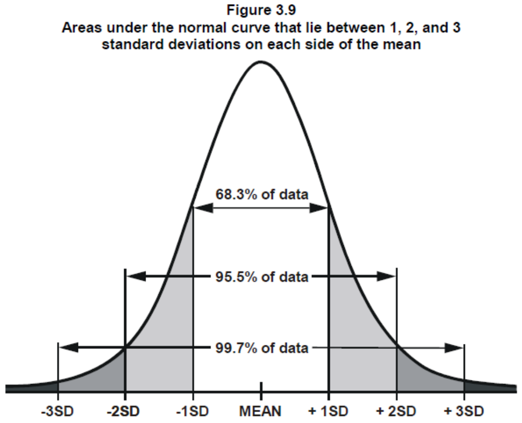


Figure 3.9
Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean

- The confidence interval will stretch for 2 SDs of the sample proportion, on either side of the center. So the width of the interva will be 4 SDs of the sample proportion.

- We are willing to tolerate a width of 1% = 0.01. So, using the formula developed in the last section,
  - $4 \times \text{SD of Sample} \leq 0.01$
  - $4 \times \frac{SD\ of\ 0{-}1\ Population}{\sqrt{Sample\ Size}} \leq 0.01$
  - $\sqrt{Sample\ Size} \geq 4 \times \frac{SD\ of\ 0{-}1\ Population}{0.01}$

Img Source: http://my.ilstu.edu/~gjin/hsc204-hed/Module-5-Summary-Measure-2/Module-5-Summary-Measure-28.html

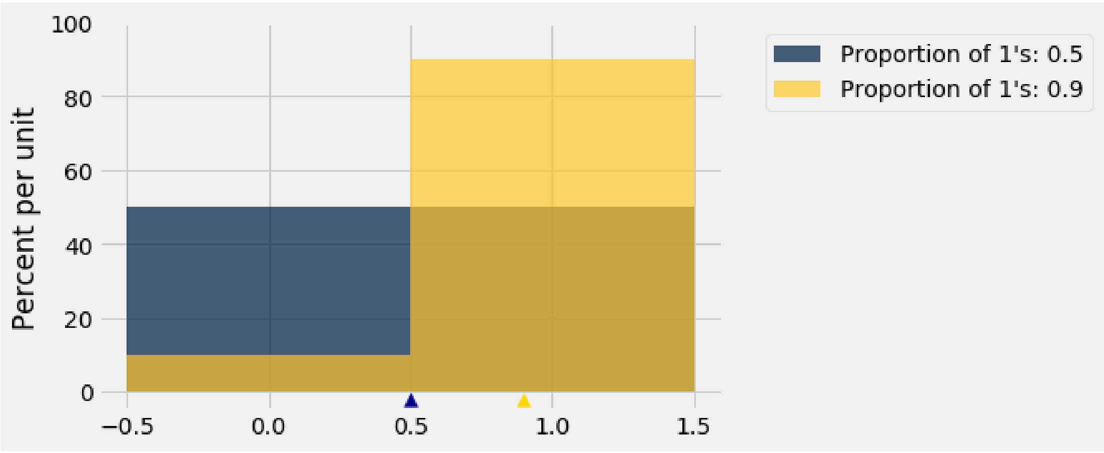
The SD of a collection of 0's and 1's

- If we knew the SD of the population, we'd be done.
- We could calculate the square root of the sample size, and then take the square to get the sample size.
- But we don't know the SD of the population. The population consists of 1 for each voter for Candidate A, and 0 for all other voters, and *we don't know what proportion of each kind there are.* That's what we're trying to estimate.
- So are we stuck? No, because we can *bound* the SD of the population.

Choosing a Sample Size

- Here are histograms of two such distributions, one for an equal proportion of 1's and 0's, and one with 90% 1's and 10% 0’s.
- Which one has the bigger SD?
- Remember that the possible values in the population are only 0 and 1.
- The blue histogram (50% 1's and 50% 0's) has more spread than the gold. The mean is 0.5. Half the deviations from mean are equal to 0.5 and the other half equal to -0.5, so the SD is 0.5.
- In the gold histogram, all of the area is being squished up around 1, leading to less spread. 90% of the deviations are small: 0.1. The other 10% are -0.9 which is large, but overall the spread is smaller than in the blue histogram.
- The same observation would hold if we varied the proportion of 1's or let the proportion of 0's be larger than the proportion of 1's

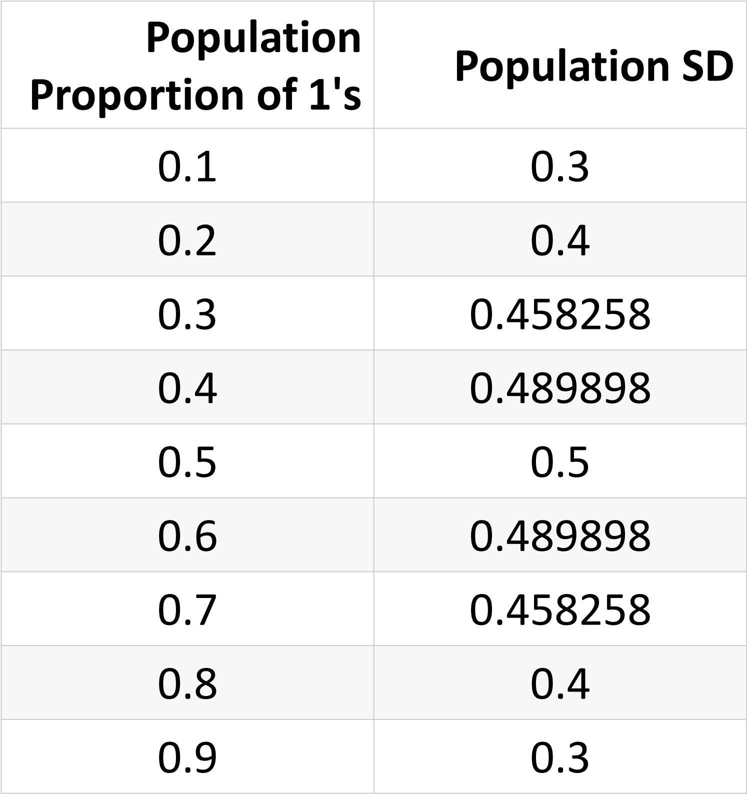
#	pop_50	Deviation_50	pop_90	Deviation_90
1	0	-0.5	1	0.1
2	0	-0.5	1	0.1
3	0	-0.5	1	0.1
4	0	-0.5	1	0.1
5	0	-0.5	1	0.1
6	1	0.5	1	0.1
7	1	0.5	1	0.1
8	1	0.5	1	0.1
9	1	0.5	1	0.1
10	1	0.5	0	-0.9
Mean	0.5	0	0.9	0
SD	0.5		0.3	



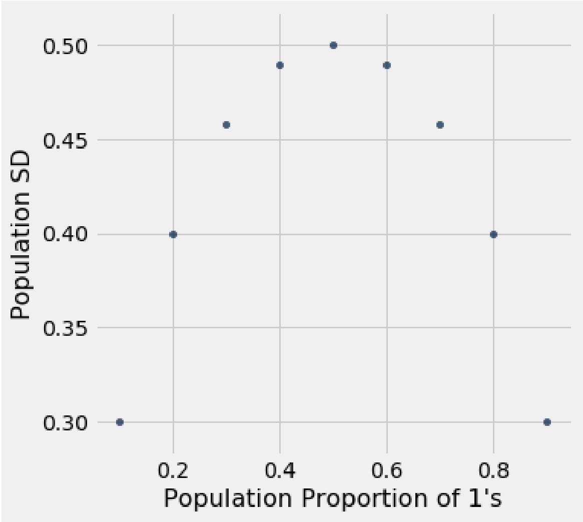
Choosing a Sample Size

- Let's check this by calculating the SDs of populations of 10 elements that only consist of 0's and 1's, in varying proportions.
- Not surprisingly, the SD of a population with 10% 1's and 90% 0's is the same as that of a population with 90% 1's and 10% 0's. That's because you switch the bars of one histogram to get the other; there is no change in spread.
- More importantly for our purposes, the SD increases as the proportion of 1's increases, until the proportion of 1's is 0.5; then it starts to decrease symmetrically.

Summary: The SD of a population of 1's and 0's is at most 0.5. That's the value of the SD when 50% of the population is coded 1 and the other 50% are coded 0.



Population Proportion of 1's	Population SD
0.1	0.3
0.2	0.4
0.3	0.458258
0.4	0.489898
0.5	0.5
0.6	0.489898
0.7	0.458258
0.8	0.4
0.9	0.3



Choosing a Sample Size

- We know that
  - $\sqrt{Sample\ Size} \geq$ 4 x $\frac{SD\ of\ 0-1\ Population}{0.01}$
- and that the SD of the 0-1 population is at most 0.5, regardless of the proportion of 1's in the population. So it is safe to take
  - $\sqrt{Sample\ Size} \geq$ 4 x $\frac{0.5}{0.01}$ = 200
- So the sample size should be at least 200^2 = 40,000 .
- That's an enormous sample! But that's what you need if you want to guarantee great accuracy with high confidence no matter what the population looks like.