# MA/CS 109 Lecture 12

Central Limit Theorem

Midterm Thursday 26th

# Last Time:

Started Statistics…

Descriptive statistics…

representing data in a useful form.

Mean (average), Median (middle)

Find the “middle” of the data

# Standard Deviation

Measure the “spread” of the data from the middle.

Standard deviation measures how much the data is spread out from the mean

(It is the square root of the average of the  of the squares of the differences from the mean.)

# What to remember?

If you have two sets of data with the same mean and the standard deviation of one set of data is larger, then we say that data is more “spread out from the mean”.

# Vocabulary

The set of individuals you are interested in is the **population**, (in the examples from last time, the populations are all 1,000,000 light bulbs and all U.S. citizens over 18)

A **sample** is a subset of the population. How you choose the sample is key, but the word refers to just some subset of the population you obtained somehow. (The 1000 light bulbs and the 1000's of people asked in the examples from last time.)

# Vocabulary

A **parameter** is a number representing a quality of the entire population (e.g., the proportion of the 1,000,000 bulbs that are defective or the percent of the U.S. citizens over 18 that support a particular piece of legislation).

A **statistic** is a number representing a quality of a sample (e.g., the proportion of the 1000 tested bulbs that are defective or the percent of the people asked who support a particular piece of legislation).

# Goal

Using this vocabulary, our goal is easy to state.

Predict a parameter of the population from a statistic of a sample.

This is actually easy…if all the information we have about the parameter is from the sample, then the best prediction is that the population parameter is the same as the sample statistic!

# Goal

But we have only used a sample and it is unlikely the sample is exactly the same as the population!! Hopefully it is close…but how close and how sure are we it is that close?

So our real goal is: Predict a parameter of the population from a statistic of a sample, and quantify the uncertainty of the prediction.

# Looked at some “Toy” examples

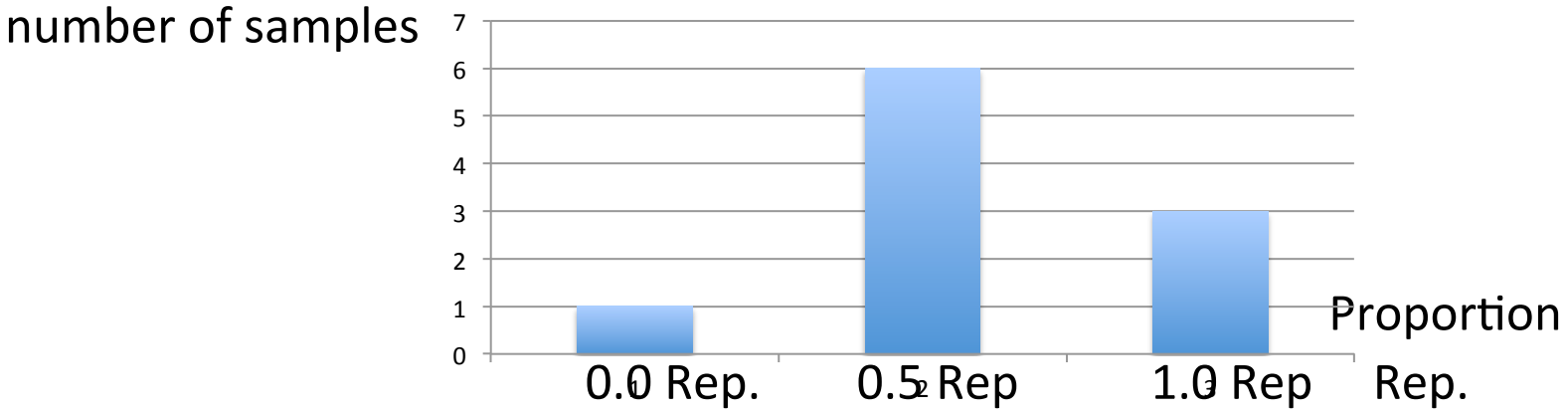
Two little towns…one with 5 people, 3 Republicans and 2 Democrats, the other with 10 people,  6 Republicans and 4 Democrats.

In the 5 person town a reporter decided to take a 2 person sample “at random”.

In the 10 person town a reporter decided to take a 3 person sample at random.

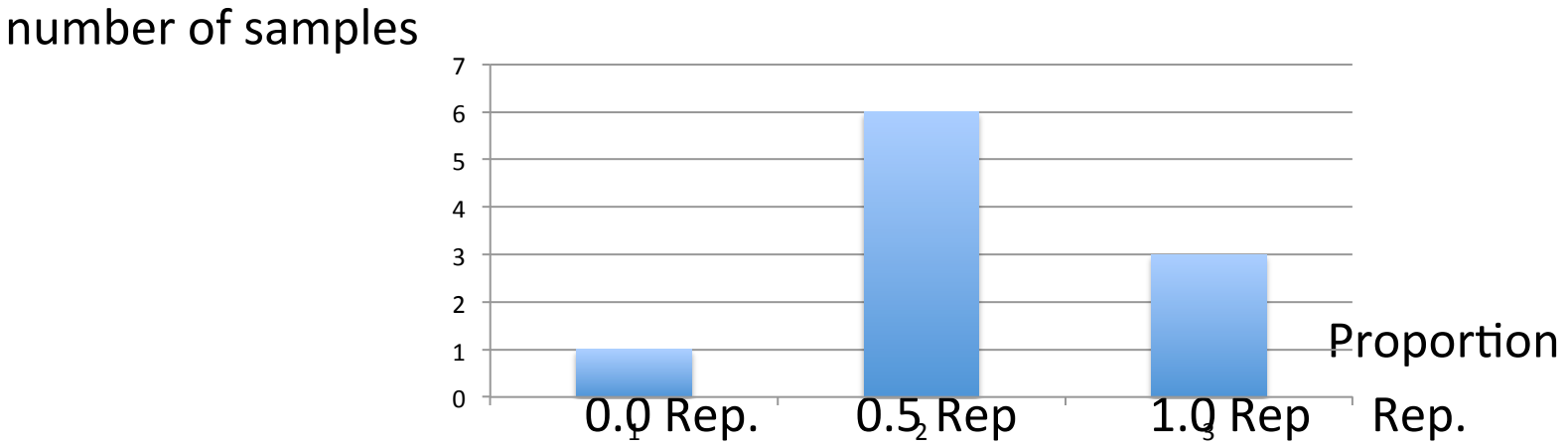
# Since these are little toys…

Since these are small towns—we can look at ALL 2 person samples in the 5 person town and see what is most likely to happen when the reporter takes one 2 person sample…we get a "distribution of sample proportions".



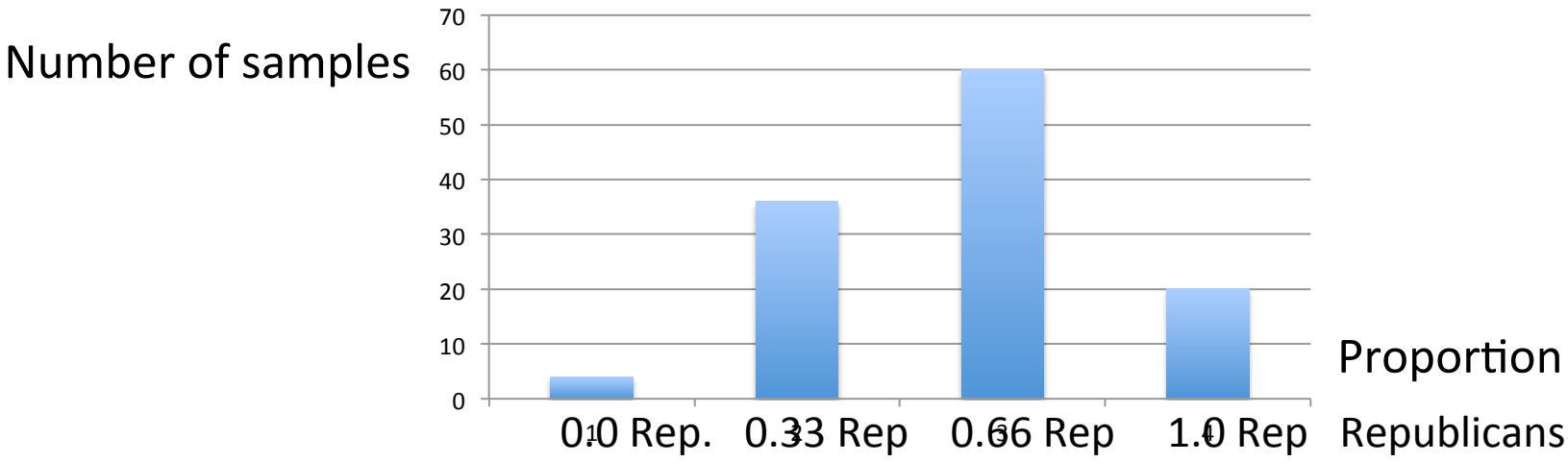
# What we noticed

The most common samples were the closest to the population proportion (0.6 Republicans). The samples with all Democrats were very rare and those with all Republicans were pretty rare.



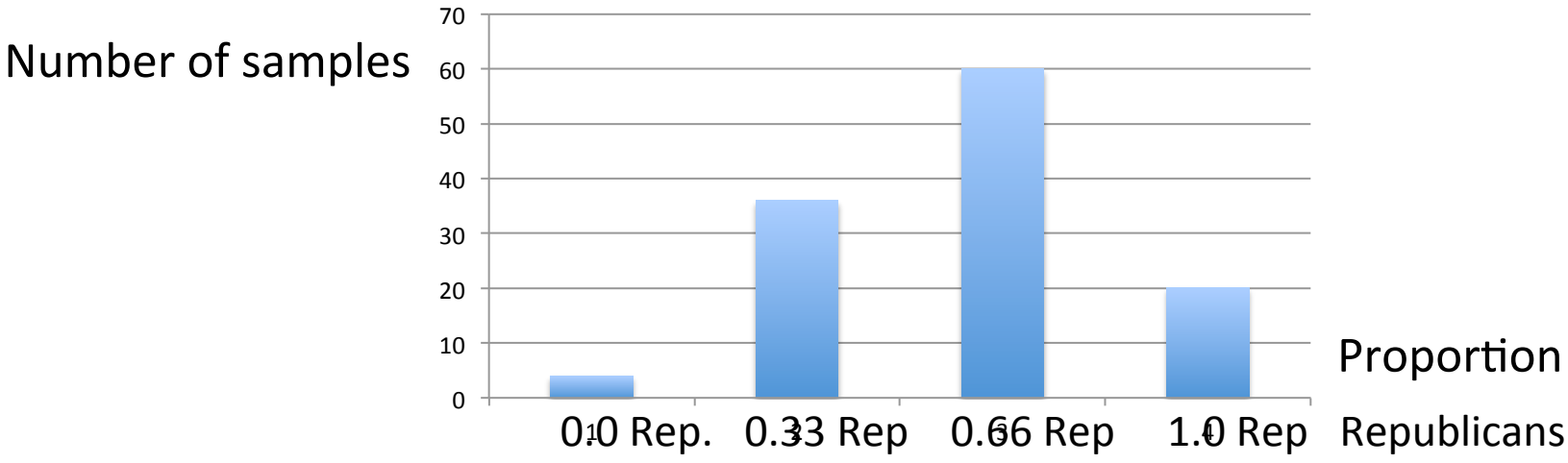
# Larger Town…

In the larger town with 10 people (still 0.6 proportion of Republicans) we looked at ALL three person samples…



# What we noticed

Same thing…Most common samples are those with proportion Republicans nearest that of the population.



# Conclude?

Well,  while the reporter might be unlucky and choose a sample that happens to have a proportion of Republicans far from that of the population, this is not very likely.

AT LEAST IF WE USE THE EQUALLY LIKELY OUTCOME MODEL TO CHOOSE THE SAMPLE!

# KEY POINT

When we say “choose a sample with n people at random”, what we mean is that every n person sample has the same probability of being chosen. (So we use the Equally likely outcomes model to compute the probability that a particular sample is chosen.)

Remember this point(!) This is key...AND in practice, it is very difficult (more on this later).

# Go Big…

Now, what happens when we have a huge city, and a large (say n>30 people) sample?

We can imagine (at least) doing the same analysis—that is, looking at ALL the n person samples, counting up the proportion of Republicans for each one and then making the histogram.

# Amazing Beautiful Fact!

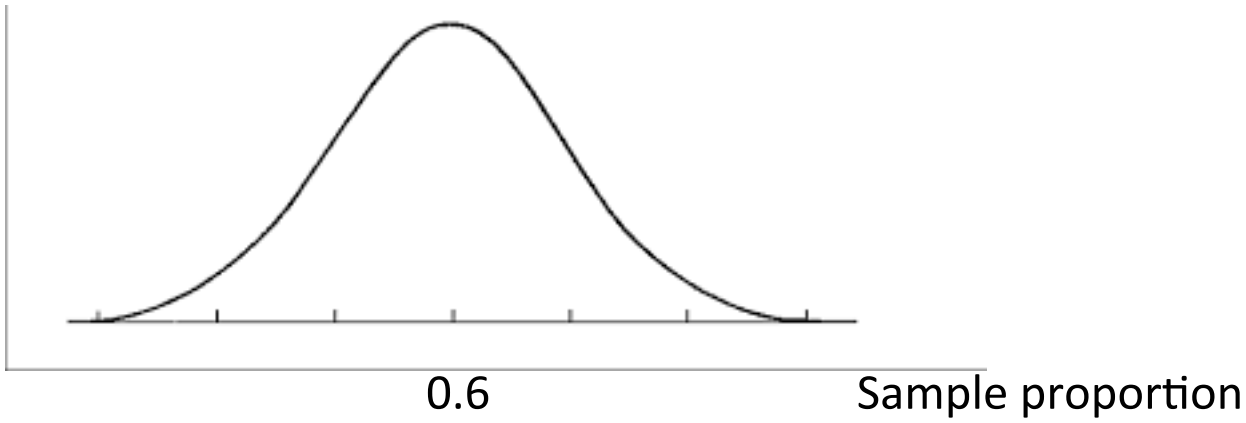
If we do this for a large city with a big sample size (say n>30) then we know what the histogram of sample proportions will be…

We don't just know what it will "look like"—lots of samples with proportion near the population proportion, just a few far away.

We actually know EXACTLY what the distribution will be!!!

# CENTRAL LIMIT THEOREM

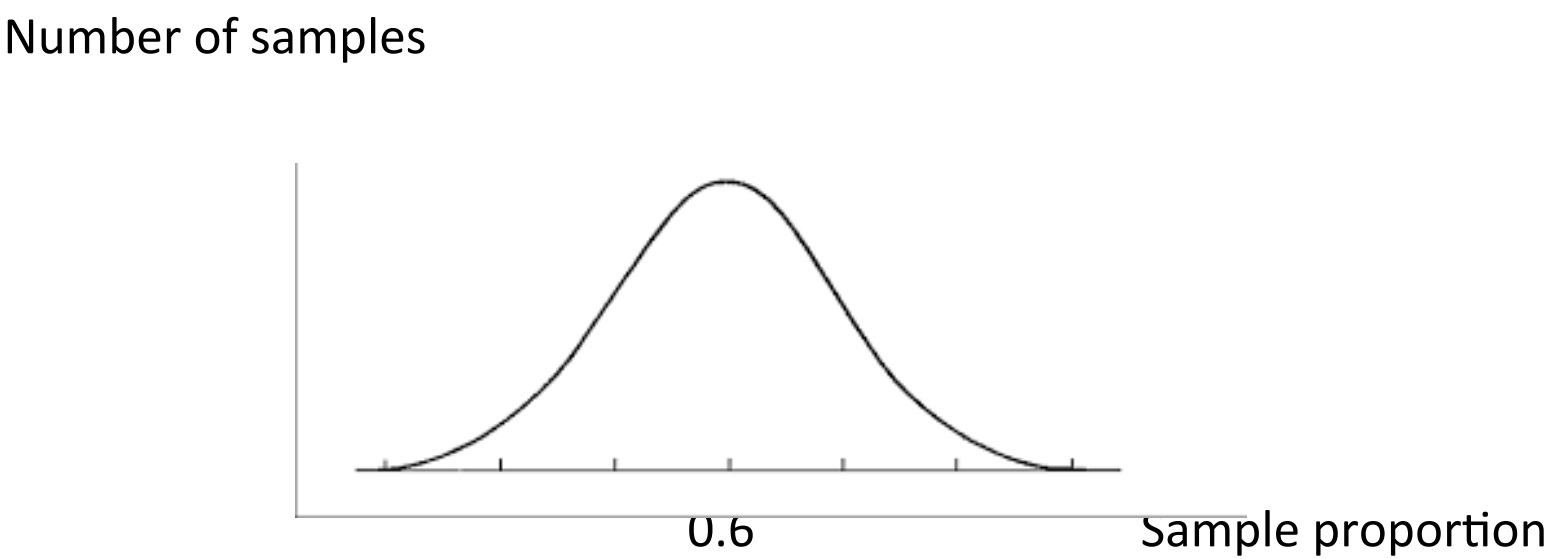
The Central Limit Theorem says the distribution of sample proportions will be ***THE*** NORMAL curve with mean equal to the population proportion (high point in the middle is at the population proportion).



# CENTRAL LIMIT THEOREM

And we even know the standard deviation (spread) of this distribution, it is $1/(2\sqrt{n})$

So we know exactly what curve we get!



# Why is this so Great?

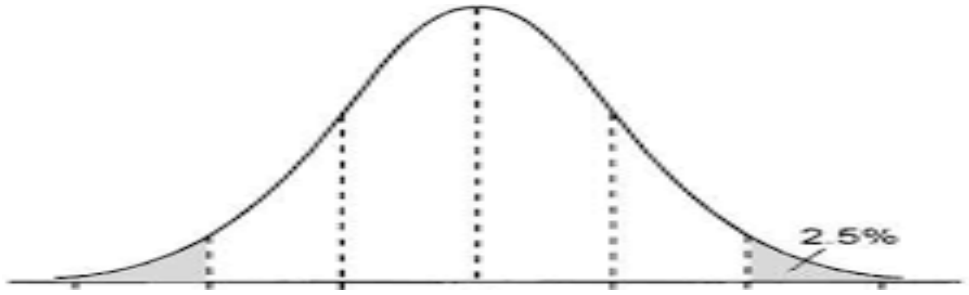
Now we can use a statistic from a sample to predict the parameter of a population with a known degree of confidence…

And because it lets us predict the future…..

Let's see how…

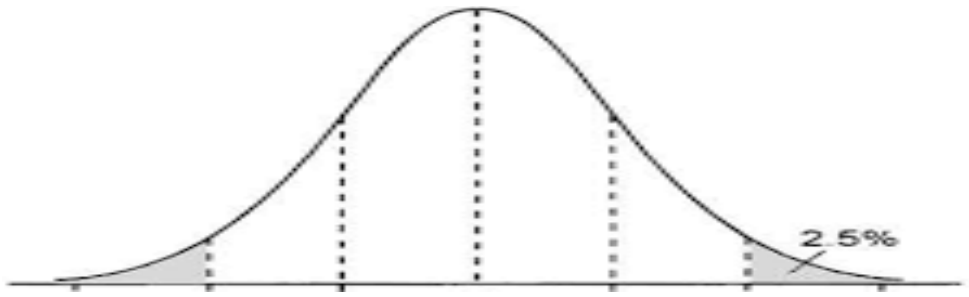
# Predicting a population parameter

One thing we know about the normal distribution is exactly how many samples are represented in each part of the curve, so exactly what fraction of the n person samples are in a given range.



# Predicting a population parameter

For example, we know that 95% of the samples of size n have proportion of Republicans that is within 2 standard deviations (or $1/\sqrt{n}$ ) of the population proportion.



# For the reporter

This means for our reporter, who picks just one sample of size n, there is a 95% chance of having the proportion of Republicans in the sample be within $1/\sqrt{n}$ of the proportion of Republicans in the town!

# Example

So, a reporter comes to a large city. She uses the equally likely outcomes model to choose a sample of size 100. She asks all 100 people if they are Democrat or Republican and 57 say Republican.

So, of course, she reports that the city has a proportion of Republicans of 0.57 (or 57%).

# BUT

But she is honest! So she reports that this is the result of a survey of 100 people, so she isn't sure the proportion of Republicans in the town is exactly 0.57. In fact, she is pretty sure that the proportion of Republicans ISN'T exactly 0.57.

But she knows her sample has a good chance of having proportion near the population proportion.

# What she knows…

Because she knows the Central Limit Theorem, she knows that 95% of the 100 person samples have proportion Republicans within $1/\sqrt{100}$ $=1/10=0.1$ of the population proportion.

So there is a 95% chance that the sample she chose has proportion within 0.1 of the population proportion.

# What she knows…

Since her sample has a proportion of Republicans of 0.57 (or 57%), she knows there is a probability of 0.95 or a 95% chance that the population has between 0.57-0.1=0.47 and

0.57 +0.1=0.67 Republicans, or between 47% and 67% Republicans.

She reports that she is “95% confident that the town is betweeh 47% and 67% Republican”.

The reporter has made a prediction about the population proportion from the sample proportion AND quantified her uncertainty about the prediction…Note that the prediction is that the population proportion is in an interval, not a specific number.

If less uncertainty is desired, a larger sample is required so that $1/\sqrt{n}$ is smaller (because n is bigger) and the spread of the samples in the distribution of sample proportions is less.

# Vocabulary

The **confidence level** is the probability that a statement is correct.

The **confidence interval** is an interval that we claim a population parameter is in—it always comes with a confidence level (a probability that the parameter is in the interval).

The **margin of error** is half the width of the confidence interval (so it also comes with a confidence level).

# Reporters example

So for the example we just did where a 100 person sample had a proportion of Republicans of 0.57, we predict with 95% confidence (or probability of 0.95) that the population proportion is within the margin of error of $0.1=1/\sqrt{100}$ of the sample proportion or we are 95% confident that the population proportion is in the interval 0.47 to 0.67.

# More information

This is still a wide spread. If we wanted a better prediction, we could use a sample of size 1000 so the margin of error is $1/\sqrt{1000} = 0.03$ (so 3%).

This is typical of polls in news papers.

Note that increasing the sample size from 100 to 1000 decreased the margin of error from 0.1 to 0.03…This is the effect of the square root!

# Predicting the future?

Sure…This is what casinos use to take our money…(or why we end up giving them our money).

If you flip a "fair" coin 100 times it is like taking a sample of size 100 from all the possible fair coin flips. Since the coin is fair, the proportion of heads will be 0.5 in the long run.

# Predicting the future

Since the population proportion is known to be 0.5, the sample proportions (your 100 flips) will cluster near 0.5, but only sometimes will they come out with exactly 50 heads.

However, the casino knows the Central Limit Theorem, so they know that the 95% of the samples (100 coin flips) will have a proportion of heads within $1/\sqrt{100}$ =0.1 of 0.5.

# Predicting the Future

So the casino knows that 95% of the time, 100 coin flips comes out with between 40 and 60 heads.

So they propose a game. They say:

Pay us $10. Flip a coin 100 times. If it comes out with less than 40 heads or more than 60 heads, we'll pay you $100. Who wins??

# Casino wins!

The casino can't predict who will win. But the Central Limit Theorem guarantees them that 95% of the players will lose over the long term, so they will only have to pay out 5% of the games or once in 20 games over the long-term.

From 20 games they make $200, but only pay out once for $100…they win…always.

# Why you shouldn't gamble

All gambling games work on basically this process…the casino knows the "population proportion" of winning the game. So the Casino can accurately predict (using the Central Limit Theorem) that over the long run, the proportion of the time they pay out.

So, my advice…don't gamble. But if you do and you win…stop gambling!

# Quality Control

Remember our example of 1,000,000 light bulbs? We chose 1000 light bulbs "at random"

(so each sample of 1000 light bulbs was as likely to be chose as any other).

We tested and found 100 light bulbs were defective from the 1000 we chose. So the sample statistics is 0.1 proportion of defective light bulbs.

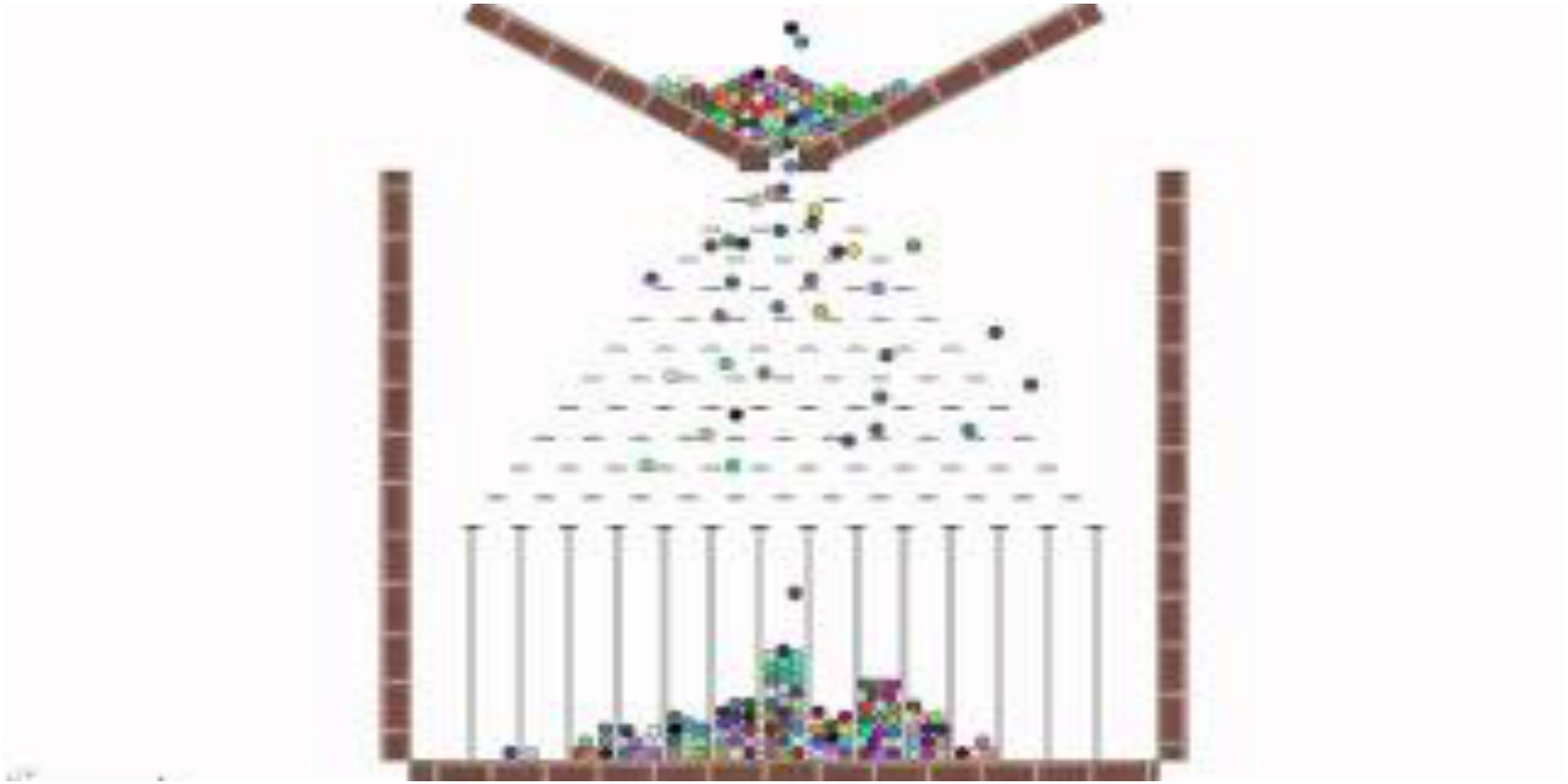
# Quality Control

So…we predict the population parameter of 0.1 proportion of the 1,000,000 light bulbs are defective or about 10% are defective. But how sure are we of this prediction?

Since $1/\sqrt{1000} = 0.03$, we can, with 95% confidence (probability 0.95), predict that between 7% and 13% of the light bulbs are defective.

(The 95% margin of error is 0.03 so the 95% confidence interval is 0.07 to 0.13).

# See CLT in Action



# Predicting the Future!
# Galton Machine

At each level, the falling marble makes a choice of left or right. The marbles are governed by the laws of physics…but we have a very hard time predicting which way any particular marble will go when it hits a peg. So we model each "choice" as a coin flip with probability ½ or left and ½ of right..

We can not predict the path of any particular marble…

But, amazingly, the CLT allow us to predict the future!  It predicts that the distribution of marbles (just like the distribution of answers to a poll or flips of a coin) will, over the long run, distribute in a normal or bell shaped curve—

And this is what happens!!

# Some demos…

https://www.youtube.com/watch?v=VQepRP-f6GA

https://www.youtube.com/watch?v=Bampgm0HKDU

https://www.youtube.com/watch?v=9xUBhhM4vbM

# Some demos…

https://www.youtube.com/watch?v=VQepRP-f6GA

https://www.youtube.com/watch?v=Bampgm0HKDU

https://www.youtube.com/watch?v=9xUBhhM4vbM

# Things to Remember…

1. Confidence intervals and margins of error always come with a confidence level. The larger the confidence level (closer to 100% confidence) the larger the confidence interval (to get more confidence you need more "wiggle room" in your prediction). The 95% confidence level is standard in polls and social sciences. In physics it is 99.999% confidence…

# Things to Remember…

Beware those who chose their confidence level AFTER doing the experiment or taking the poll.

2. We MUST choose our sample in such a way that the probability of every sample of the same size being chosen is the same! (Equally likely outcomes on the space of samples.)

# Things to Remember…

This is harder than it sounds…choosing every fifth person to walk past you between 12 and 1 outside of Warren Towers on Tuesday, for example, will not work for a sample of BU students! Those who live on west campus will be under represented. Those with classes only on MWF will be underrepresented…adding in some people with MWF classes won't work because Equally likely outcomes says you might not get anybody with MWF classes……

# Things to Remember…

And, if you come up with a sample by choosing names at random from a list, then you have to find all the people, and you have to get them to answer…

3. Even when you find someone on your list and convince them to talk to you…will they tell you the truth, or give the answer they think you want?

# Things to Remember…

4. Even when you do everything right and find everyone in your sample and convince them all to answer (with the truth) and do all the arithmetic correctly…you still are not sure of your prediction. You are making a prediction with a probability (confidence level)…If you use the 95% confidence level then over the long run 1 time out of 20 you will be wrong.

# ESP

Suppose you set up an experiment as follows

You have strangers where they can not see each other. Whenever a bell rings, one of the strangers flips a fair coin and stares at the result. The other stranger tries to read the first person's mind and write down the result of the flip.

# ESP

On each flip, there is a probability of 0.5 that person who did not flip the coin will write the correct flip.

If you repeat this experiment 100 times and the pair of people get more than 60 correct, you are amazed! This is a good result. There is a 95% chance that only 40 to 60 predictions will be correct (remember $1/\sqrt{100} = 0.1$ so the 95% confidence interval is 0.4 to 0.6 or 40 to 60 correct)

# ESP

Do you report this? Do you announce that you have discovered ESP abilities?...you are 95% confident that you have…

Except that this experiment gets done many times. You expect more than 60 or less than 40 correct predictions 5% of the time.

# Scientific Retractions

A phenomenon in science is the “non-repeatability” of exciting new results. Somebody makes an exciting discovery and tells people about it (in a paper or at a conference). Others try to repeat the work, but fail…even the original researcher tries again and the result does not “work”…This seems to happen a lot—why?

# Scientific Retractions

If you try an experiment and get a "surprising" result—at the 95% or even 99% confidence level, then you should take it seriously. But you should also remember that perhaps 100 other researchers have tried this experiment and gotten an "unsurprising" result. You might have gotten the surprising result just by chance…

Why science must be tested over and over (and why it is is so nice to do math and have Proofs.)

# Next Time

A story about Causation vs. Correlation…