Inspiring Greatness
INDEPENDENCE SCHOOL DISTRICT

Math Virtual Learning

Probability and Statistics

April 29, 2020

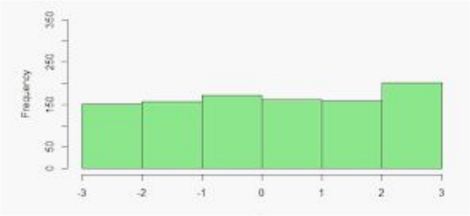
Probability and Statistics
Lesson: April 29, 2020

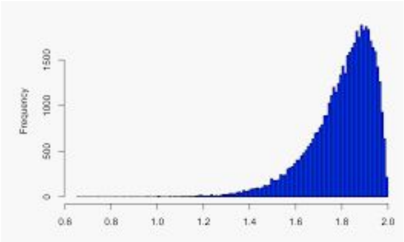
Objective/Learning Target:
Students will learn the meaning of the Central Limit Theorem and how it can be used when data is not normal

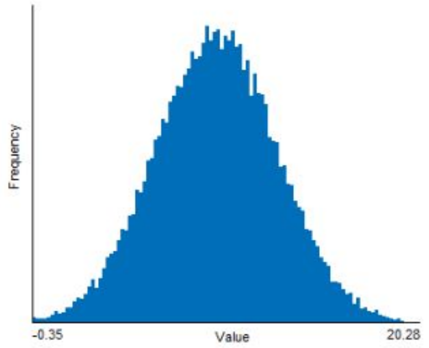
Let's Get Started!

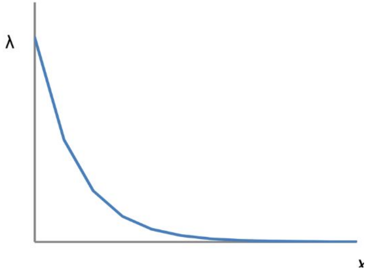
Match the graphs below to their distribution type

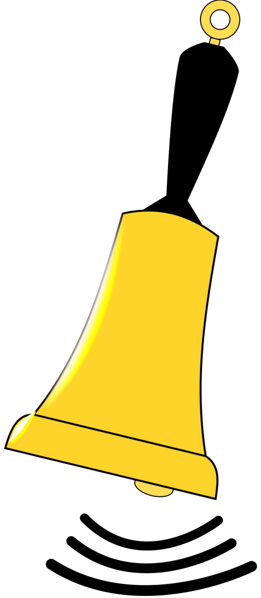
Normal | Skewed Left | Exponential | Uniform



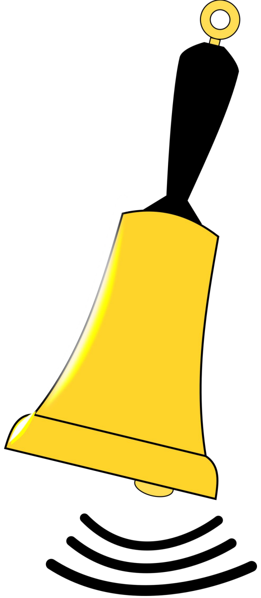








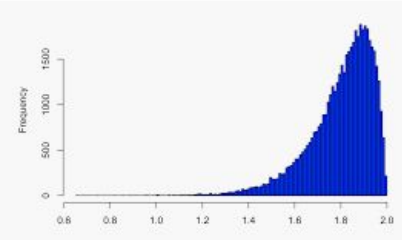
Let’s Get Started! ANSWER



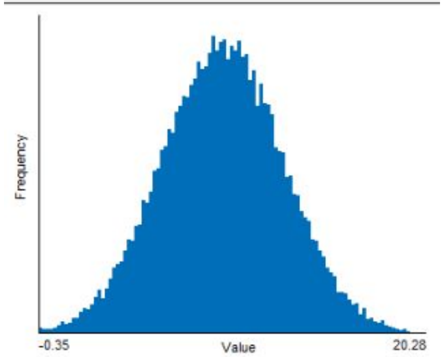
Match the graphs below to their distribution type

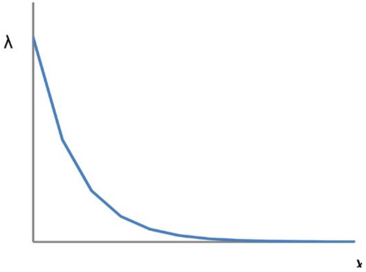
UNIFORM

SKEWED LEFT



NORMAL

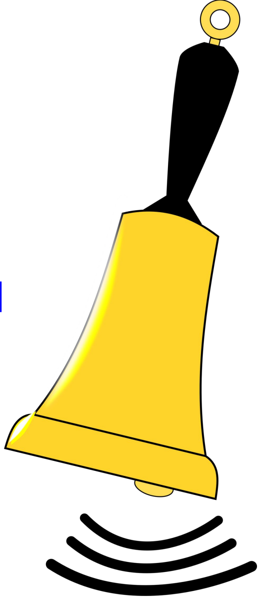




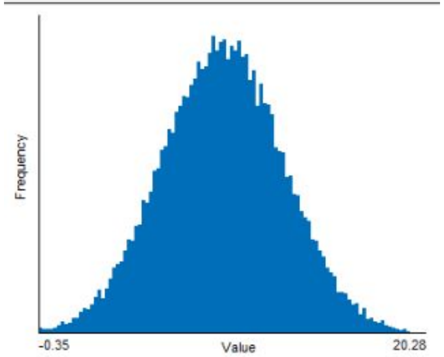
EXPONENTIAL

Let's Get Started! Follow Up Question

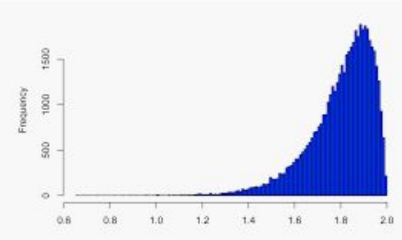
Based on the graphs, which is the only experiment that we can find z-scores and percentiles for? Why would this be a problem?



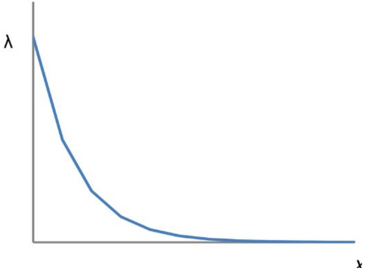
NORMAL



UNIFORM

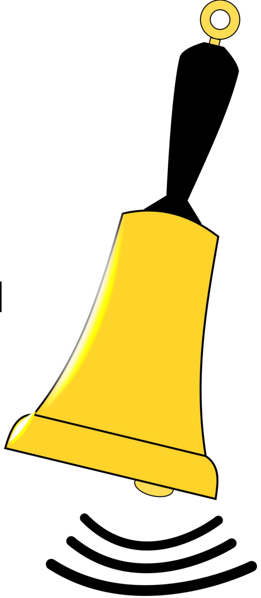


SKEWED LEFT

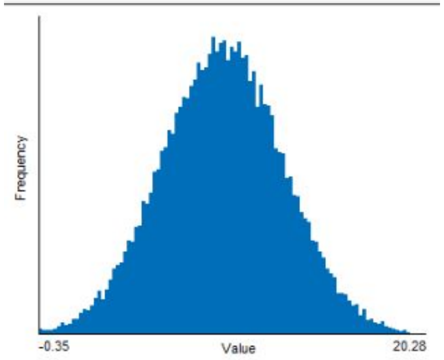


EXPONENTIAL

Let's Get Started! Follow Up Question ANSWER



Based on the graphs, which is the only experiment that we can find z-scores and percentiles for? Why would this be a problem?



NORMAL

Only data with a Normal Distribution can be used for Statistical analysis.

This is a problem, because finding trends of other distributions is still important.

Today we are going to learn how to overcome this obstacle.

Central Limit Theorem

The Central Limit Theorem is a very simple concept, but it is often difficult for students to wrap their head around because we are going to be taking a given distribution of data/graph and then taking samples and averaging them to create a new graph(distribution).

This concept is often called the “average of the averages” because you will see that we are going to create a Sample Distribution of the Sample Means or in other words, create a bell curve graph from the means of all the samples we take from the original data.

See how this can be confusing? Don’t overthink it though, it’s really not as hard as it sounds. Let’s break it down.

So why are we doing this???

Not all of your graphs in life will be normally distributed (bell curve) so in order to find percentages and other standard statistic measures, we need it to be normally distributed.

The Central Limit Theorem (CLT) says that if you take a bunch of samples and find the Mean of each one and then graph them, you will get approximately a normal distribution (bell curve) EVERY time.

Normal vs Not Normal to begin with...

Central Limit Theorem says that if your original data is:

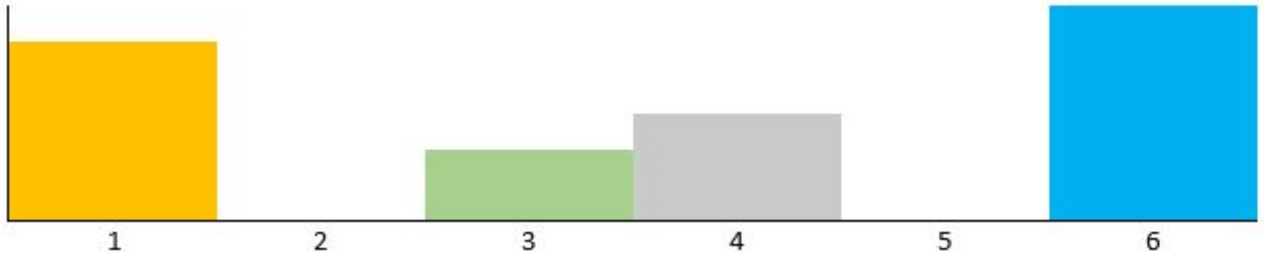
<u>Normal</u> - then your sample averages will automatically be a normal distribution (bell curve) also, using any sample size small or large.

<u>Not Normal</u> - then your sample averages will make a normal distribution (bell curve) **<u>only</u>** if your sample size is 30 or greater. In other words, you have to use at least 30 numbers when finding the mean...no less!

Let's start with an example...

This example is a basic one from Khan Academy but it is just what we need to demonstrate how simple this really is.

Let's say you have a weighted dice of some kind, where it has 6 sides (1-6) but because of the weighting it won't land on a 2 or a 5. Let's say someone rolls the dice a bunch of times (we don't know how many times) and comes up with a graph that looks like this…



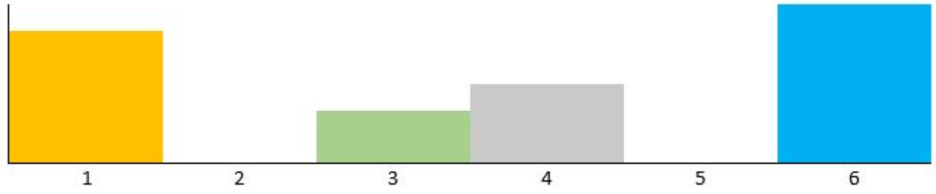
Population Distribution Graph

This graph is called the Population Distribution Graph (or our ORIGINAL graph of the data from the dice rolls)

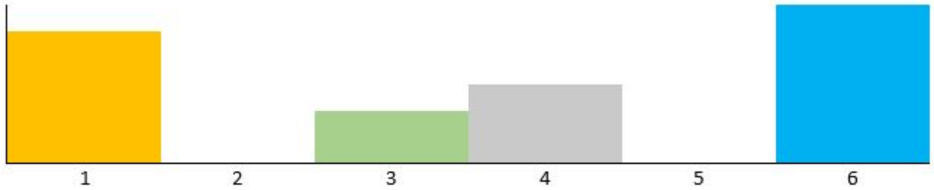
We can see that it is NOT a normal distribution (bell curve).

This is not good because we need a normal distribution so that we can make some standard inferences about the population.

So this is where the Central Limit Theorem comes into play. We can take samples (> 30) of this data and find the mean of our sample. Then graph the means to create a bell curve.



Taking Samples...



We CAN find the mean of our samples and the CLT says that the mean of our samples is always equal to the mean of the population. So to find our "true mean" (of the population) we need to first find the mean of our samples.

So let's take some samples….

Remember that we don't have any 2's or 5's so those cannot be in our samples.

Let's start with a small sample size (I know that we will need >30) but let's look at what happens when the sample size is smaller first.

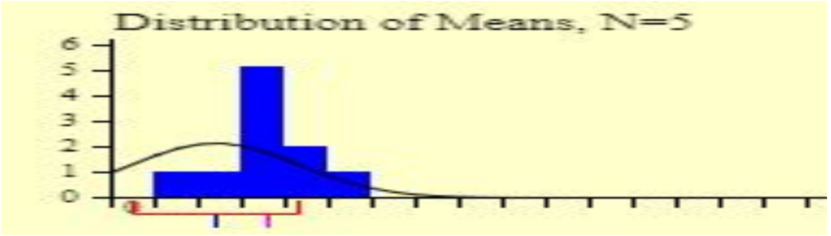
Sample 1: 1, 1, 3, 4, 4 = 2.6 (average)
Sample 2: 3, 4, 6, 6, 1 = 4
Sample 3: 3, 4, 3, 4, 6 = 4
| | |
| | |
Sample **10,000**: 3, 1, 3, 6, 4 = 3.4

Now let's graph all of our sample means...

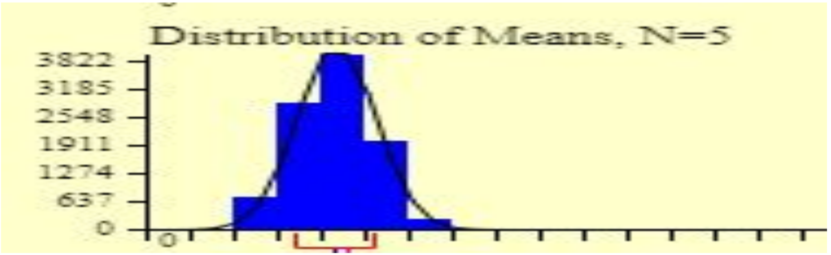
Graphing the Sample Means…

Obviously I didn't really take samples, but I used a computer program to generate the samples for me from *"onlinestatbook.com"*

If I choose 10 samples of 5 and plot the means, I get a graph that looks like this (clearly skewed and not "normal")

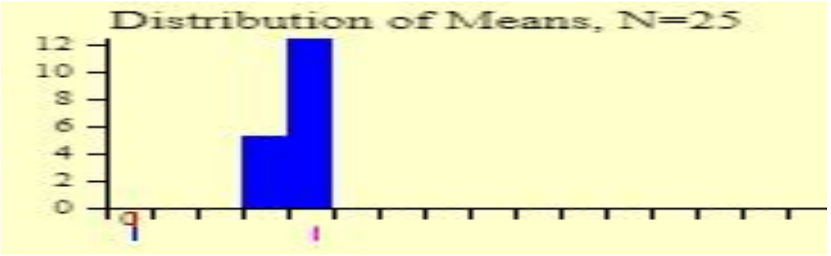


If I choose 10,000 samples of 5 and plot the means, I get a graph that looks like this (much closer to "normal")

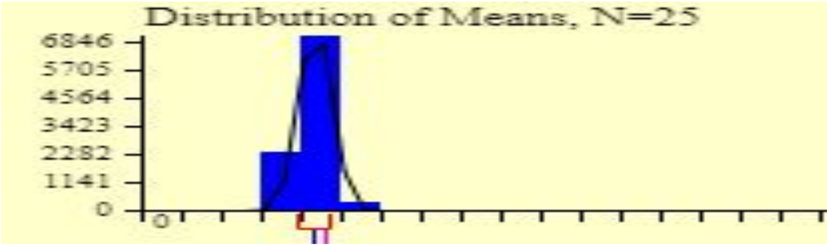


Graphing the Sample Means…

If I choose 10 samples of 25 and plot the means, I get a graph that looks like this (clearly skewed and not “normal”)



If I choose 10,000 samples of 25 and plot the means, I get a graph that looks like this (much closer to “normal”)



The Mean on this particular graph is 3.78 but if we use a sample size of 30 instead of 25 and did the experiment say 100,000 times, we would get closer to 3.5.

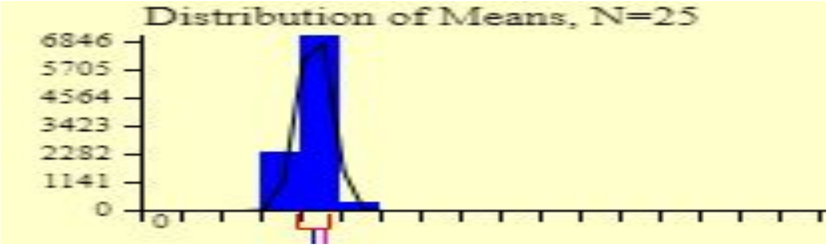
Back to Population Mean...

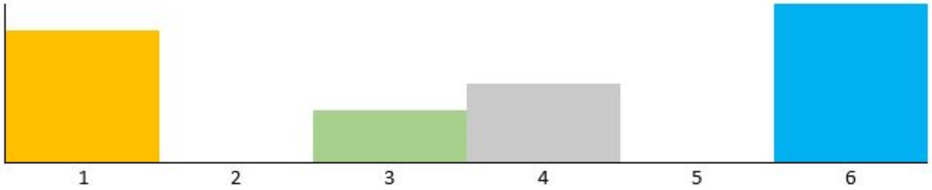
So if the mean of our samples eventually goes to 3.5 (with N=30, and 100,000 reps)

Then…..

We can say that the “true mean” or the population mean is 3.5 as well.

The whole point of doing the samples was to set up a bell curve that gets the same mean as the population mean. The population mean is often times refered to as the “True Mean”





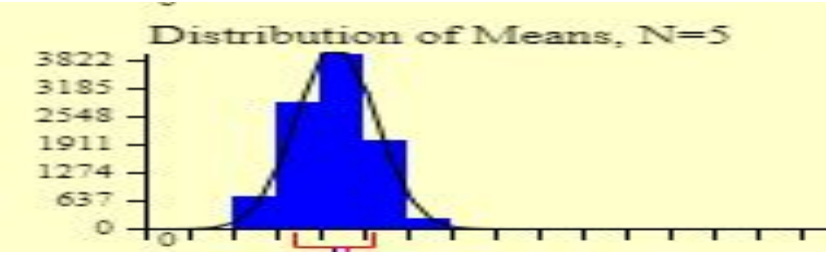
Now let’s go back to the Sample Means…

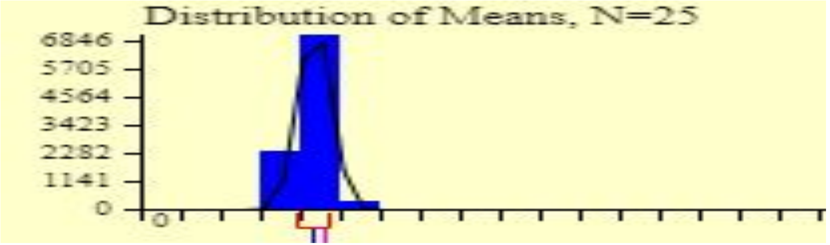
Notice the difference between these to graphs of our sample means

N=5: wider

N=25: narrower

The spread is smaller so the standard deviation will also be smaller.





When N gets larger, the standard deviation gets smaller.

Why does that happen?

When N gets larger, the standard deviation gets smaller.

That's because the SAMPLE standard deviation (called the Standard Error of the Mean) is a derivative of the value of N.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The Standard Error (your Sample standard deviation) will always be smaller than the population standard deviation because it is always affected by the square root of N.

If N is large then you are dividing by the square root of a larger number, which in turn will give you a smaller answer.

So that is why the standard deviation on your sample mean graph is smaller the more the N value goes up.

That's it! Now let's recap and watch a short video to tie it all together...

Central Limit Theorem (CLT)

- Says that if I take samples from a population of data, and I find and graph the Mean of those Samples, I will end up creating a Bell Curve (Normal Distribution) that has the same Mean as the Population Graph did

Standard Error of the Mean:

- Is the same thing as the Standard Deviation of your sample means graph.

Why are we learning this?

- To be able to take "non-normal" population data and turn it into a "normal distribution" graph
- Then we can use what we learned about Z-Scores and Percents to answer questions and make inferences (conclusions) about our population

Click on the next slide for a video that will help you grasp this concept better...

Additional Resources