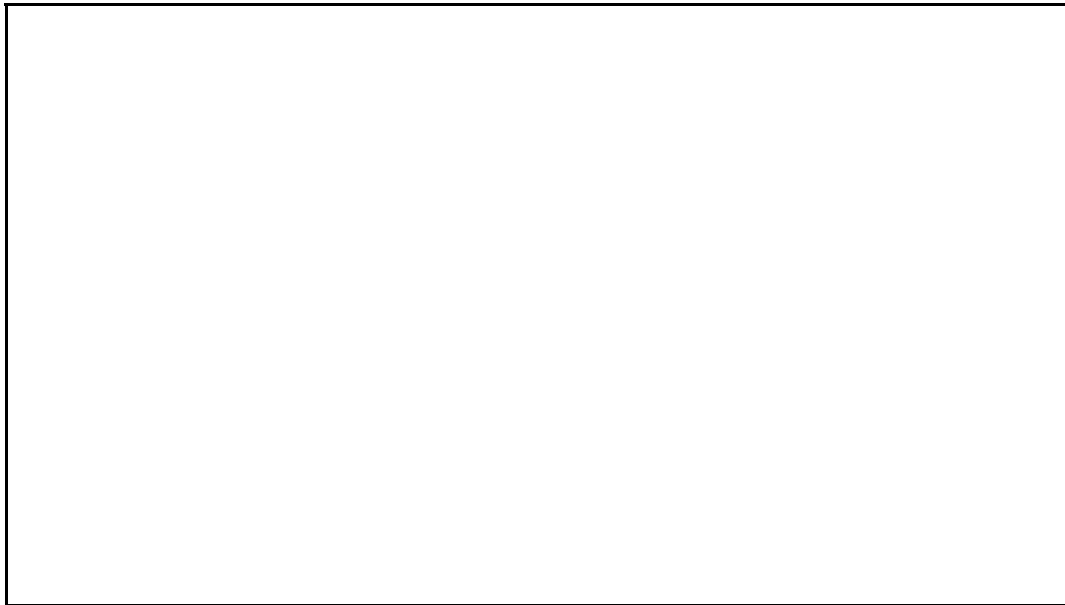


- Courseware
- Course Info
- Discussion
- Syllabus
- Download R and RStudio
- R Tutorials
- Readings
- Contact Us
- Progress
- Office Hours
- Community

THE SAMPLING DISTRIBUTION



	0:00 / 8:38	1.0x			
--	-------------	------	--	--	--

SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

We'll be starting a new discussion - one that revolves around the idea of inference.

When it comes to data, we have two types: population data and sample data.

Now in this time of Big Data, most all of the idea of data is population.

For example, a company that looks at all of its transactions

that have occurred on its website.

But sometimes - most of the time - it's just

not

feasible to collect all of the data of a population.

So there's another type of data, and that's sample data.

And with sample data, we need a way to point back to the population

- to infer what the population is behaving like based on our sample data.

And to do that, we need to understand something

called the Central Limit Theorem.

That's a great question: What is the Central Limit Theorem?

But to answer that question, we need to talk

about the population and the sample of our data.

Here's some data on professor salaries from the "car" package in R.

It contains 397 records, and we'll be treating this as our "population."

But, in actuality, these professors do not represent a population of all professors nationally, but instead

a population of professors at a specific U.S. college.

If we look at the distribution of the

we see some right or positive skew.

And we can calculate the population statistics.

Here's the mean and the standard deviation of the population of professor salaries at this single college.

Now, it was fairly easy to get all the professors salaries.

They are infact employees, so these records are readily available.

But what if the gathering this data wasn't so easy?

Could you still look at the idea of professor salaries?

Well sure - you could simply take a sample from the population.

Now, in a perfect world, you take a random sample.

Where each member of the population has an equal likelihood

of being chosen for the sample as any other member.

Let's draw one sample of say 10 members.

Our mean is an estimate of the population value.

And we know that this sample mean may not exactly match the population mean.

There's this thing called sampling error.

But here's the cool thing - if we draw multiple samples (each of the same size) from the same population those mean values that we get will in fact "bounce around" the true population mean,

eventually taking on the shape of a distribution - a NORMAL distribution.

We call this distribution of mean values a Sampling Distribution.

A distribution of sample values is a Sample Distribution,

while a theoretical distribution of sample means from a single population

is a Sampling Distribution.

Now let's see this in action to get an idea of what's going on.

I've taken 1,000 samples from our population of professor salaries, each at a size of 10, and recorded the mean for each sample.

At the end of the day, we end up with a vector of 1,000 mean values.

Let's visualize those mean values with a histogram.

Each mean value has been placed in a particular bin, indicated

by the red dot on the axis.

Now, notice what's happening as we add

more and more sample

means to our histogram.

The histogram - the Sampling Distribution -

looks more and more normal.

In fact, this is a major component of the Central Limit Theorem.

Regardless of the population shape in our case,

it's slightly skewed to the right regardless of the population

shape, regardless of the sample shape, the Sampling Distribution

(the distribution of sample means) will approach normal.

The greater the size of each sample - the number

of salaries in each of the 1,000 samples - the more normal

the Sampling Distribution will be.

For example, if each of my samples were only of size 5,

the Sampling Distribution will have a harder time being normal,

than say if each of my samples have a size of 20.

Now, what does all of this mean?

It means that we can assume a distribution of means

- the Sampling Distribution - can utilize
the normal distribution, and its properties.

It will have a stable symmetrical shape,
more instances in the middle,
trailing off to either side equally.

And what does this look like?

It's the normal distribution or the
z-distribution.

And now we can ask questions like "How
likely is

it that this particular sample with this
particular mean value

is like some proposed value?"

Do you see it?

We could use a z-test.

And here's another interesting aspect of the
Central Limit Theorem:

if we use this idea of a z-test, what do you
think we'll use for the denominator?

Well that's the Standard Deviation of the
Sampling Distribution.

Now the Standard Deviation of the
Sampling Distribution

is also defined by the Central Limit
Theorem.

The guy is called the Standard Error.

And notice that denominator.

The larger the n for each of the samples that we take to construct the distribution of sample means, the smaller this Standard Error.

Say the standard deviation of the population of salaries is 30,251.

If we take samples of size 10, then the Standard Error will be about 9,578.

So, according to the Empirical Rule, roughly 68%

of the sample means of size 10 will have an internal spread of 19,156,

the distance between 1 Standard Error below the mean

and 1 Standard Error above the mean of the Sampling Distribution.

But what if we take 30 cases for each sample?

Our new Standard Error is 5,530.

And our new 68% central spread is only 11,060.

Now if we think about this change, it actually makes sense.

A sample with a small amount of data in it

- a low n - may have a pretty hard time accurately estimating

the population mean.

But a larger sample does a better job
estimating the population mean.

Meaning the amount of error around the
central value of the Sampling

Distribution will be less, as reflected by a
smaller Standard Error.

And that is the Central Limit Theorem:
When

examining the distribution of thousands of
sample means of the same size

from the same population, normality
ensues regardless

of what the population or the sample looks
like.

And the spread of those mean values
around their central core

- the population mean - will get less and
less

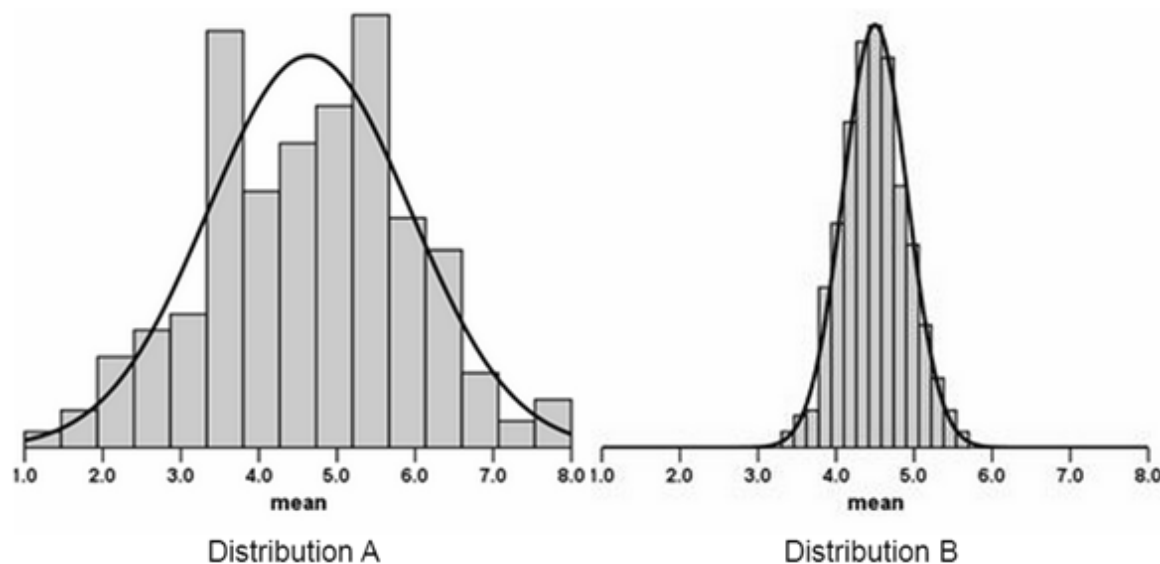
as the size within each of the samples gets
larger,

because the mean of larger samples will
more often fall closer

to the actual population value.

Comprehension Check


1. Below are two sampling distributions showing the average library fines owed by patrons of the Chicago public library system. They are drawn from the same population.



(1/1 point)


9 of 10. Which distribution shows means calculated from larger sized samples?

02/02/2015 11:21 AM

☐ Distribution A☒ Distribution B **Check****Hide Answer****Help**

(1/1 point)

1b. What is the likely average library fine owed by a Chicago Public Library patron?

☐ Between \$3 and \$6☒ About \$4.50 ☐ About \$5.25☐ We cannot tell from these distributions.**Check****Hide Answer**

(1/1 point)

1c. What can be said about the variability in the sample means for both these distributions?

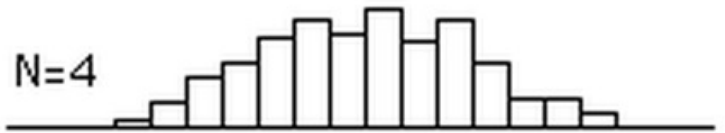
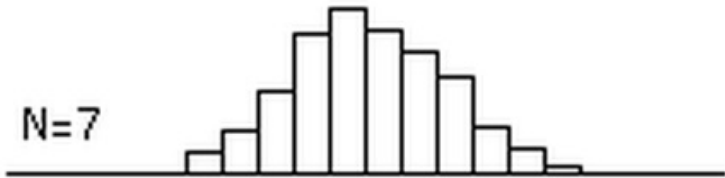
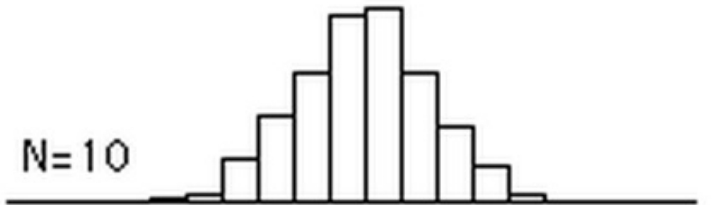
- ☐ The sample means in Distribution A are more realistic; they show the true variability in the population.
- ☒ The sample means in Distribution B are more similar to each other, resulting in a tighter cluster of values. ✓
- ☐ There is no difference in variability between the means in Distribution A and Distribution B because all the samples are drawn from the sample population.

(1/1 point)

1d. If we were to repeat the survey using **larger** sample sizes than either of those in Distribution A and Distribution B, what would you predict to be the shape of the new distribution?

- ☒ Bell-shaped, but taller and more narrow than the other distributions. ✓
- ☐ Bell-shaped, but wider and shorter than the other distributions.
- ☐ The shape would not change, but the value of the mean would probably change.

2. Below are four distributions of sample means drawn from the same population. The distributions differ by the size of the samples, which is noted by the value of N .

$N=1$  $N=4$  $N=7$  $N=10$ 

(1/1 point)

2a. Which sample size resulted in the tightest clustering of values around the "true" population mean?

☐ 1☐ 4☐ 7☒ 10

[Check](#)[Hide Answer](#)

(1/1 point)

2b. Which sample size resulted in the lowest variability in sample means?

Help

☐ 1☐ 4☐ 7☒ 10[Check](#)[Hide Answer](#)

(1/1 point)

2c. What happens to the shape of the sampling distribution as the sample size increases from 1 to 10?

Help

- ☐ It becomes less normal.
- ☒ It becomes more normal (bell-shaped). ✓
- ☐ It becomes more uniform.
- ☐ It does not change.

Check

Hide Answer

(1/1 point)

2d. How many of the sample means, when $N=10$, fall within 1 standard deviation of the true population mean?

- ☐ About 50%
- ☒ About 68% ✓
- ☐ About 95%
- ☐ About 99%

Check

Hide Answer



Help



EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

© 2015 edX, some rights reserved

[Terms of Service and Honor Code](#)

[Privacy Policy \(Revised 4/16/2014\)](#)

About edX

[About](#)

[News](#)

[Contact](#)

[FAQ](#)

[edX Blog](#)


[Donate to edX](#)

[Jobs at edX](#)

Follow Us

 [Twitter](#)

 [Facebook](#)

 [Meetup](#)

 [LinkedIn](#)

 [Google+](#)