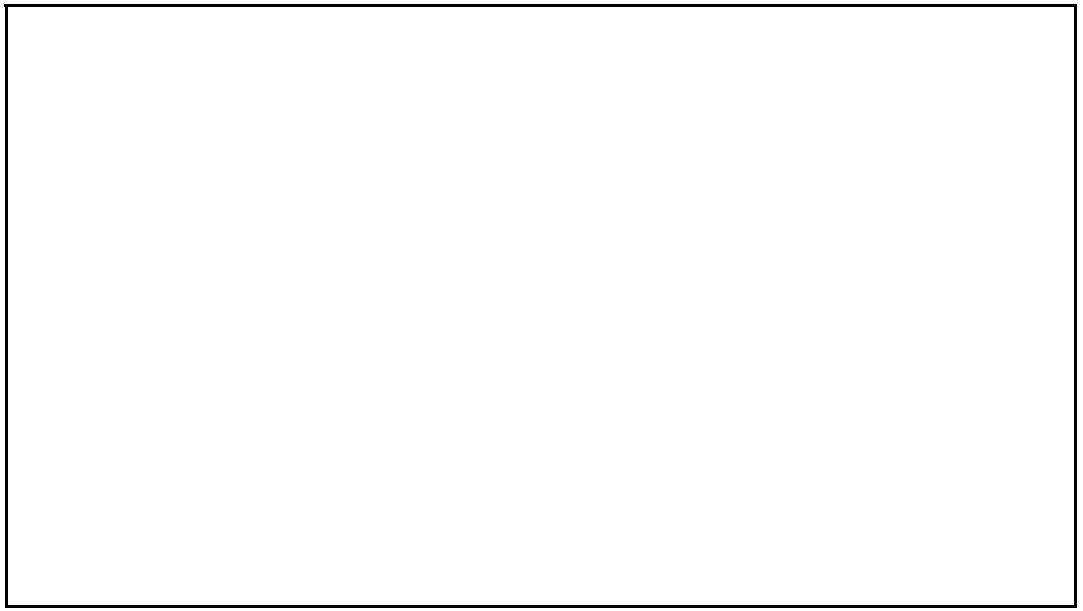
THE T-DISTRIBUTION



0:00 / 10:17 1.0x

SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

In this video I'd like to introduce you to a very important statistician:

William Gosset.

Why is he important?

Because he was an employee of Guinness Beer in the early 20th century

- but also because he came up with what's

known today as the Student's T Distribution.

This is another reason why I love statistics.

not only

are we building a set of tools to answer questions about data,

but we're also learning some absolutely interesting stories.

Help

William Gosset was a chemist and mathematician

and an employee of Guniess.

He was interested in the yield of different types of barley.

Now, working at Guinness led to two small problems.

First: small sample sizes.

You see Gosset was working during a time when

large sample theory and the "normal" or z-distiribution was the norm.

But, when experimenting withthe changes to batches of beer,

or measuring the yield of a crop, the numbers

of observations in a single sample are unfortunately just small.

Imagine: you're tying to alter the outcome - a measurable quality of beer,

say nitrogen content.

But the process of making the alteration is time consuming and expensive.

Each "experiment" of the alteration provides only a single data

point for the sample.

Now, to collect good data, you'd want to replicate the experiment over and over

and over again, thus adding single values to your experiment.

Help

Gosset realized that this kind of data just

didn't fit the ideas of large sample theory.

He stated that "There are other experiments, however,

which cannot easily be repeated...Some chemical, many biological,

and most agricultural and large-scale experiments belong to this class,

which has hitherto been almost outside the range of statistical inquiry."

Gosset discovered that as sample size changed,

the relationship between the population mean,

sample mean and the Standard Deviation was not a normal distribution.

Now, through a series of events - work with Carl Pearson of the Pearson

Correlation fame, and academic and personal communication with Sir Ronald

A. Fisher - another important person in modern statistics -

a proposed "correction" to the z-distribution or normal distribution

was discovered.

Eventually, this work would lead to the creation of what is now known

as the t-distribution - a distribution that actually takes into account

the number of "degrees of freedom" in the statistical test.

We'll get to that definition in a little bit.

Here's why a z-test - and the "normal distribution" just don't work:

Small sample sizes make the distribution of samples not normal;

the normal distribution is based on knowledge of sigma (the Standard

Deviation of the population).

Now, in large sample theory, that second point sort of got overlooked.

Remember, the Standard Deviation of the sample

is supposed to be an estimate of the Standard

Deviation of the population (or sigma).

At the time that Gosset was writing his breakthrough paper

"The Probable Error of a Mean" in 1908.

the standard Standard Deviation calculation

was to use n in the denominator.

Now, in large sample theory - the prominent statistical methods

of the time - using n in the denominator didn't raise any red flags.

Look what happens if we use some simulation:

Let's take 100,000 samples of size 500 from a population

with a Standard Deviation of 10.

The "average" Standard Deviation (using the denominator of n)

from all our samples is 9.98.

Doesn't seem too far off from 10, does it?

With samples of size 100, the average Standard Deviation of all the samples

is 9.92.

And, now using samples of size 30, we get an average Standard Deviation 9.74.

As sample size decreases, the Standard Deviation of each sample

get's farther and farther away from the actual population Standard

Deviation of sigma.

Using just n in the denominator causes an under-estimation of sigma.

Unfortunately the normal distribution is based on sigma,

and we've just seen that our estimate of sigma

will be off if we have small samples.

The fix?

Well, it's two fold: first, we use a better estimate of sigma.

Fisher actually shared a proof with Gosset in 1912 for using (n-1)

in the denominator for the calculation of the sample Standard Deviation

to help with this under estimation of sigma problem.

Second, we apply the t-distributuion, which

takes into account the size of the sample.

And as sample size increases, the resulting distribution

for the relationship between population mean and sample

mean and the new Standard Deviation (now called t instead of z)

becomes more and more like the normal distribution.

So - back to the t-distribution.

In order to use it effectively, we first need

to determine what's called the "Degrees of Freedom."

Each type of test that uses the t-distribution

has a different way to calculate its degrees

of freedom,

so we'll address degrees of freedom calculation

separately for each statistical test we'll use.

But, it will help if we can define degrees of freedom to get us started.

Help

There are a few ways to describe or define degrees of freedom,

but the simplest way that I like to define degrees of freedom

is as "the number of observations that are free to vary."

Here's a simple a explaination of concept "free to vary."

Imagine we have five observations in a single sample

- say, students in a class.

And we know the average of their test scores on a recent exam is 75.

So, if we know we NEED to end up with 75 as a final mean,

how many of the students' scores are in fact "free to vary" to any number?

Let's try 5.

Randomly picking 5 numbers (or 5 degrees of freedom) won't work.

We won't be able to get to the mean of 75.

But what if we allow 4 student scores to

freely vary -

could we still get a mean of 75?

Absolutely.

Say the 4 freely varying scores are all 10 - then to get to 75,

the fifth (the score that's NOT free to vary would need to be 335.

So, in our example, there are a maximum of 4 scores that can freely vary

- for 4 degrees of freedom.

For a sample of 10 students - how many are free to vary?

Well the answer is 9.

Let's take a look at the t-distribution in action.

Here's a t-distribution for 2 degrees of freedom.

now it looks pretty much like we're used to right?

But here's the normal distribution.

Notice the difference?

The t-distribution peak is shorter and the tails are fatter.

Let's see how this affects the promotion under the curve concept

by using our idea of critical values.

We'll use an alpha of 0.05.

For a t-distribution with 2 degrees of

freedom,

the critical values are -4.30 AND 4.30.

That's the effect of the change in shape - the fatter tails.

Here's the shape with 5 degrees of freedom and the new critical values.

Help

This new shape has a taller peak than the 2 degrees of freedomthat shape,

with correspondingly thinner tails - so the critical values

are a little less - a little closer to the middle of the distribution.

Here's another t-distribution with 10 degrees of freedom - again,

thinner tails than what we had, and a smaller critical value.

We should hopefully be seeing that as our n increases,

the t-distribution becomes more and more like the normal distribution.

And our critical values for this same alpha level - what ever that value

happens to be - will get closer and closer

to the critical values of the normal distribution.

And that's the t-distribution.

Where it came from, and how it differs from the normal distribution.

So, when should we use the t-distribution

over the normal distribution.

For me, it's very simple:.

There's only one rule.

Do we KNOW sigma - the Standard Deviation of the population?

If we do, then we get to use the normal distribution, or the z-statistic.

If we don't know sigma, however, then we must use the t-distribution.

Help

Comprehension Check

1. Here is the formula for calculating the t-statistic for a single sample:

$$t = \frac{sample\ mean\ -\ hypothesized\ mean}{standard\ error}$$

Help

(1 point possible)

1a. Which part of the formula tells you how far your sample mean is from the center of the sampling distribution?

Numerator

Hide Answer

(1 point possible)

1b. Which part of the formula tells you how far a sample mean is expected to be from the center of the distribution?

Denominator

Hide Answer

2. Which of the following is true of the t-distribution?

(1 point possible)

- Increasingly resembles the normal distribution as degrees of freedom increase.
- Assumes the population is normally distributed
- It has a greater spread than the normal distribution.
- All of the above. ✔

Help

Hide Answer

Help

About edX

About

News

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Follow Us

Twitter

Facebook

Meetup

LinkedIn

Google+