

Lecture 16 Sampling Distribution of The Sample Variance

BIO210 Biostatistics

Xi Chen

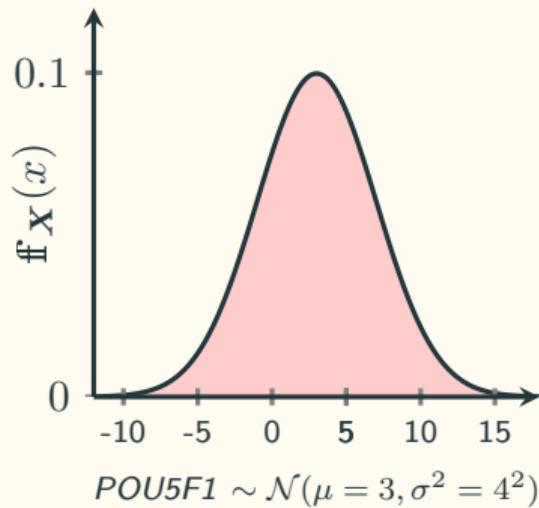
Spring, 2025

School of Life Sciences
Southern University of Science and Technology



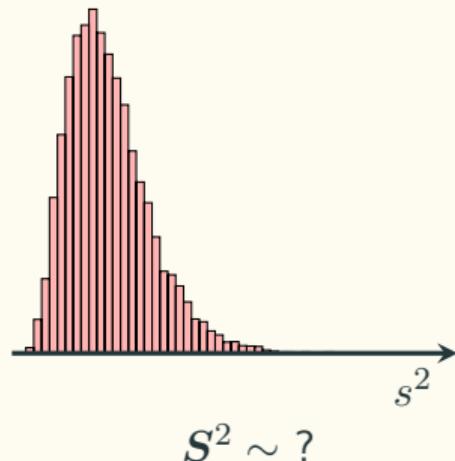
南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Sampling Distribution of The Sample Variance



$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Sample	Sample variance
$n = 96$	$\rightarrow s^2 = 15.11$
$n = 96$	$\rightarrow s^2 = 15.75$
$n = 96$	$\rightarrow s^2 = 17.44$
$n = 96$	$\rightarrow s^2 = 16.85$
$n = 96$	$\rightarrow s^2 = 13.31$
$n = 96$	$\rightarrow s^2 = 16.36$
\vdots	\vdots
$n = 96$	$\rightarrow s^2 = 17.29$
\vdots	\vdots



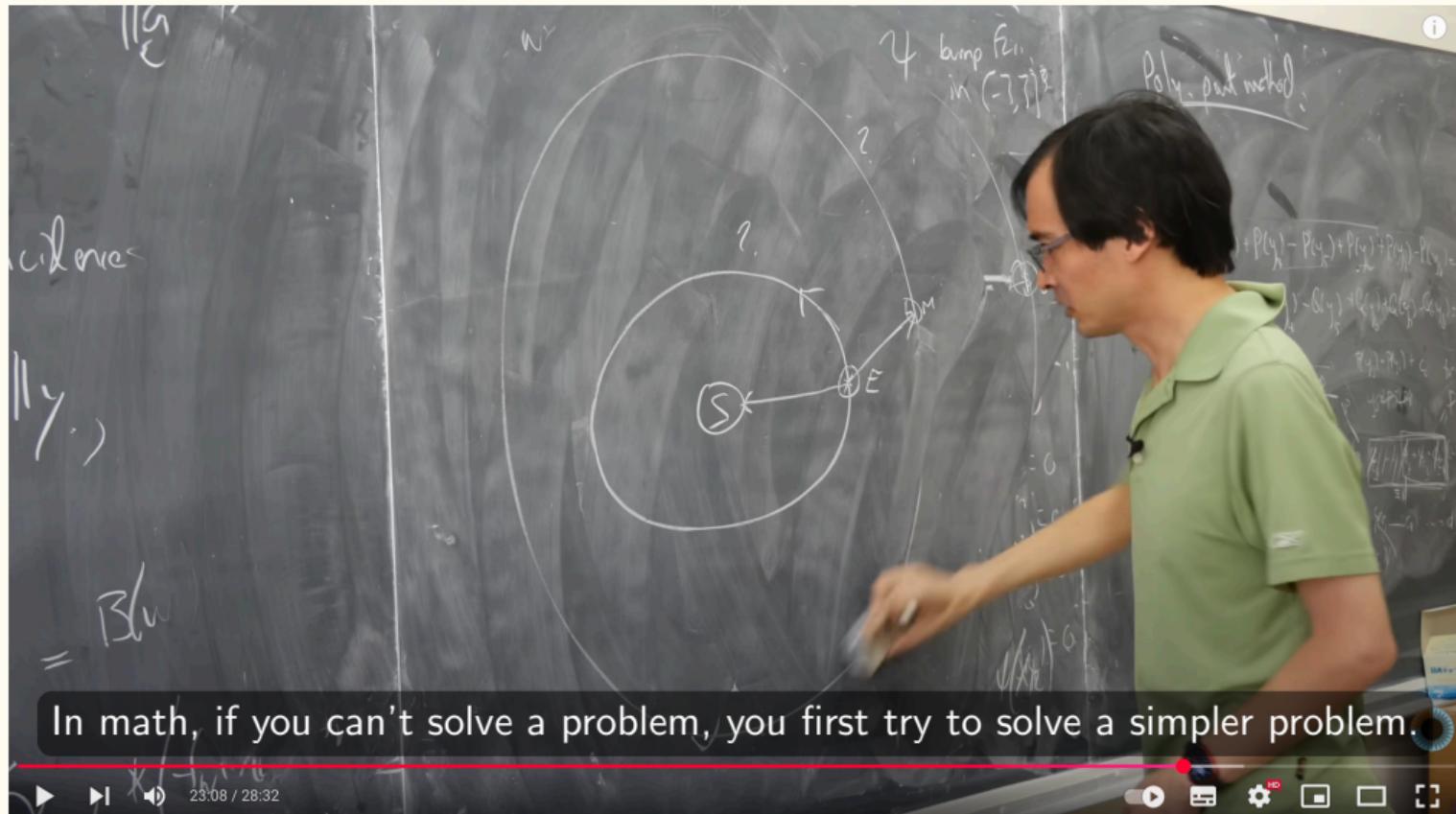
**Sampling distribution
of the sample variance**

Figuring Out The Distribution of S^2

Task: We draw a sample of size n ($\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$) from a population ($\mathbf{X} \sim \mathcal{D}$), where $\text{Var}(\mathbf{X}) = \sigma^2$, we want to figure out:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^2 = \frac{1}{n-1} \left[(\mathbf{X}_1 - \bar{\mathbf{X}})^2 + (\mathbf{X}_2 - \bar{\mathbf{X}})^2 + \dots + (\mathbf{X}_n - \bar{\mathbf{X}})^2 \right]$$

Problem Solving Advice By Terrence Tao



Start With The Special Case

Simplify: Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be **i.i.d.** random variables from a normal population $\mathcal{N}(\mu, \sigma^2)$

$$S^2 = \frac{1}{n-1} \left[(\mathbf{X}_1 - \bar{\mathbf{X}})^2 + (\mathbf{X}_2 - \bar{\mathbf{X}})^2 + \dots + (\mathbf{X}_n - \bar{\mathbf{X}})^2 \right]$$

The question becomes: what is the sum of a bunch of squared normal random variables?

The Standard Normal Squared

Simplify even further:

Let $Z_1, Z_2, Z_3, \dots, Z_n$ be **i.i.d.** standard normal random variables: $Z_i \sim \mathcal{N}(0, 1)$, then

- $Z_1^2 \sim ?$
- $Z_1^2 + Z_2^2 \sim ?$
- \vdots
- $\sum_{i=1}^n Z_i^2 \sim ?$

The Chi-squared (χ^2) Distribution

Friedrich Robert Helmert in 1876:

Number of Z_i^2	The PDF of the sum
1	$\frac{1}{\sqrt{2\pi}}x^{-\frac{1}{2}}e^{-\frac{x}{2}} : \chi^2(1)$
2	$\frac{1}{2}e^{-\frac{x}{2}} : \chi^2(2)$
3	$\frac{1}{\sqrt{2\pi}}x^{\frac{1}{2}}e^{-\frac{x}{2}} : \chi^2(3)$
4	$\frac{1}{4}xe^{-\frac{x}{2}} : \chi^2(4)$
5	$\frac{1}{3\sqrt{2\pi}}x^{\frac{3}{2}}e^{-\frac{x}{2}} : \chi^2(5)$
:	:

by induction:

$$\chi^2(n) : f_X(x) = \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x \geq 0$$

where:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \alpha > 0$$

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

$$\Gamma(k) = (k - 1)! , \text{ when } k \text{ is an integer}$$

One parameter - the degree of freedom:
the number of independent Z^2 in the sum

The Distribution of S^2

By definition:

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

Replacing μ with \bar{X} :

$$\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$$

Manipulate to get the sample variance:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Why $n - 1$? part 1

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^2 \leq \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \mu)^2$$

Why? Because:

$$\sum_{i=1}^n (x_i - m)^2 = n \cdot m^2 - \left(2 \sum_{i=1}^n x_i \right) \cdot m + \sum_{i=1}^n x_i^2$$

But why exactly $n - 1$? Wait until **part 2** in **Lecture 18**

The Degree of Freedom (DF, DOF, ν)

Typical definition: the number of values in the final calculation of a statistic that are free to vary; the number of independent pieces of information used to calculate the statistic.

There are two types of degrees of freedom:

$$\begin{cases} df \text{ of the data} & - df \text{ left (statistical cash)} \\ df \text{ of the statistical model} & - df \text{ spent (buy with cash)} \end{cases}$$

A statistical model: a mathematical process that attempts to describe the sample data that come from a population, allowing us to make predictions.

Different Types of df

Intuitive thinking: the number of cells that can vary in a Spreadsheet.

	Data	Model
	x_1	
	x_2	
	x_3	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
	:	
	x_n	
df	n	1