# Lecture 16 Sampling Distribution of The Sample Variance

BIO210 Biostatistics

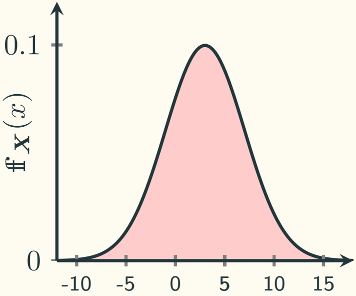
Xi Chen

Fall, 2024

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

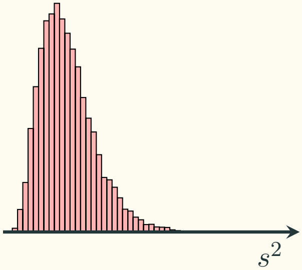
## Sampling Distribution of The Sample Variance




*POU5F1* $\sim \mathcal{N}(\mu = 3, \sigma^2 = 4^2)$

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

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




$\boldsymbol{S}^2 \sim$ ?

**Sampling distribution of the sample variance**

# Start With The Special Case

**Task:** We draw a sample of size $n$ ($\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$) from a population ($\boldsymbol{X} \sim \boldsymbol{\mathcal{D}}$), where $\mathbb{V}\mathrm{ar}\left(\boldsymbol{X}\right) = \sigma^2$, we want to figure out:

$$\boldsymbol{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^2 = \frac{1}{n-1}\Big[(\boldsymbol{X}_1 - \bar{\boldsymbol{X}})^2 + (\boldsymbol{X}_2 - \bar{\boldsymbol{X}})^2 + \cdots + (\boldsymbol{X}_n - \bar{\boldsymbol{X}})^2\Big]$$

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

$$\boldsymbol{S}^2 = \frac{1}{n-1}\Big[(\boldsymbol{X}_1 - \bar{\boldsymbol{X}})^2 + (\boldsymbol{X}_2 - \bar{\boldsymbol{X}})^2 + \cdots + (\boldsymbol{X}_n - \bar{\boldsymbol{X}})^2\Big]$$

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

## The Standard Normal Squared

Let $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \cdots, \boldsymbol{Z}_n$ be **i.i.d.** standard normal random variables: $\boldsymbol{Z}_i \sim \mathcal{N}(0, 1)$, then

- $\boldsymbol{Z}_1^2 \sim ?$
- $\boldsymbol{Z}_1^2 + \boldsymbol{Z}_2^2 \sim ?$
- $\vdots$
- $\sum_{i=1}^{n} \boldsymbol{Z}_i^2 \sim ?$

# The Chi-squared ($\chi^2$) Distribution

Friedrich Robert Helmert in 1876:

| Number of $\boldsymbol{Z}_i^2$ | 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 | $\frac{1}{4} x e^{-\frac{x}{2}} : \chi^2(4)$ |
| 5 | $\frac{1}{3\sqrt{2\pi}} x^{\frac{3}{2}} e^{-\frac{x}{2}} : \chi^2(5)$ |
| $\vdots$ | $\vdots$ |

by induction:

$$\chi^2(n) : \mathbb{f}_{\boldsymbol{X}}(x) = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, x \geqslant 0$$

where:

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

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

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

**One parameter - the degree of freedom:**

the number of independent $\boldsymbol{Z}^2$ in the sum

# The Distribution of $S^2$

**By definition:**

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

**Replacing $\mu$ with $\bar{X}$:**

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

**Manipulate to get the sample variance:**

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

# Why $n-1$? part 1

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

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, $\nu$)

**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:**

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

**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 |
|---|---|---|
| | $x_1$ | |
| | $x_2$ | |
| | $x_3$ | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ |
| | $\vdots$ | |
| | $x_n$ | |
| $df$ | $n$ | $1$ |