Lecture 15 Sampling Distribution And The Central Limit Theorem

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

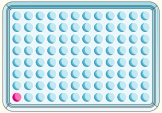
Use Sample Statistics To Estimate Population Parameters

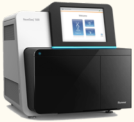
$$\text{Sample(s)} \xrightarrow[\text{Sampling distribution}]{?} \text{Population}$$

The scenario: we draw a sample of size n from the population. We observe the sample mean is $\bar{x}$ and the sample variance is s^2. **We want to answer the following type of questions:**

- **If the population mean were μ_0:**
  - what would be the probability of observing a sample of size n with a mean of $\bar{x}$?
  - what would be the probability of observing a sample of size n with a mean falling into $[a, b]$?
- **If the population variance were σ_0^2:**
  - what would be the probability of observing a sample of size n with a variance of s^2?
  - what would be the probability of observing a sample of size n with a variance of $[a, b]$?

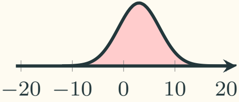
Intuition of Sampling Distribution



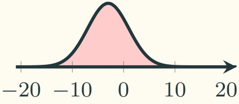


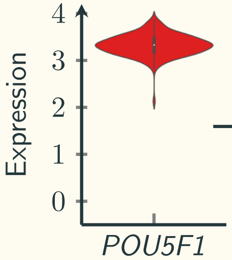
Experiment_name
- plate_1
  - cell_1
  - cell_2
  - cell_3
  - ⋮
- plate_2
  - cell_1
  - cell_2
  - cell_3
  - ⋮

mESC: $\mathcal{N}(3, 4^2)$



Tc: $\mathcal{N}(-3, 4^2)$

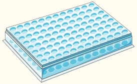




→ Which type?

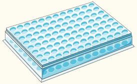
Intuition of Sampling Distribution

100 plates (samples)



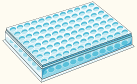
$n = 96$

$\bar{x} = 3.07$



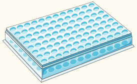
$n = 96$

$\bar{x} = 2.97$



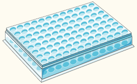
$n = 96$

$\bar{x} = 2.80$



$n = 96$

$\bar{x} = 4.95$

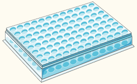


$n = 96$

$\bar{x} = 3.60$

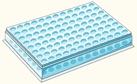
$\vdots$

$\vdots$



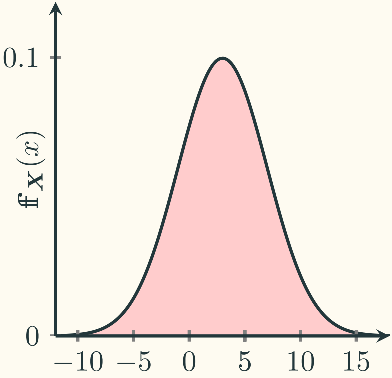
$n = 96$

$\bar{x} = 1.05$



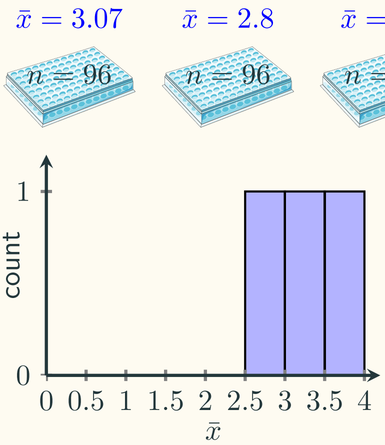
$n = 96$

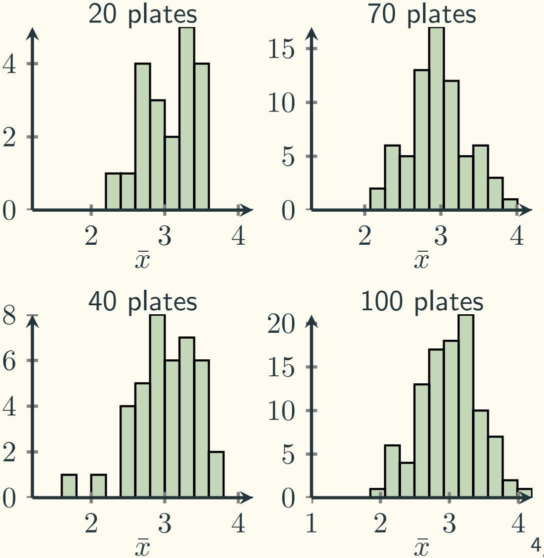
$\bar{x} = 3.27$



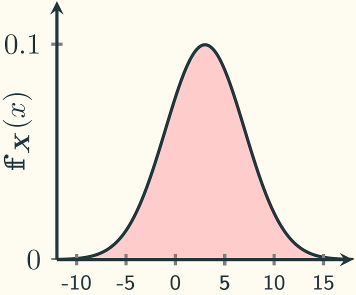
POU5F1 $\sim \mathcal{N}(\mu = 3, \sigma^2 = 4^2)$

Intuition of Sampling Distribution



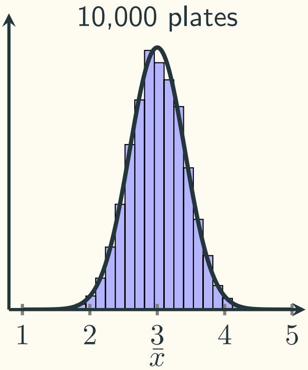


Sampling Distribution of The Sample Mean



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

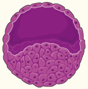
Sample	**Sample mean**
$n = 96$	$\longmapsto \bar{x} = 3.05$
$n = 96$	$\longmapsto \bar{x} = 3.13$
$n = 96$	$\longmapsto \bar{x} = 3.15$
$n = 96$	$\longmapsto \bar{x} = 2.95$
$n = 96$	$\longmapsto \bar{x} = 2.87$
$n = 96$	$\longmapsto \bar{x} = 3.20$
$\vdots$	$\vdots \qquad \vdots$
$n = 96$	$\longmapsto \bar{x} = 2.50$
$\vdots$	$\vdots \qquad \vdots$



$\bar{X} \sim$

Sampling distribution of the sample mean

i.i.d. Random Variables



Gene A expression $\sim \mathcal{D}(\mu, \sigma^2)$

$$\boldsymbol{X}_1 \quad \boldsymbol{X}_2 \quad \boldsymbol{X}_3 \quad \cdots \quad \cdots \quad \boldsymbol{X}_{n-1} \quad \boldsymbol{X}_n \qquad \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

$$x_1 \quad x_2 \quad x_3 \quad \quad \quad x_{n-1} \quad x_n \qquad \bar{x}$$

$$3.0 \quad 3.2 \quad 3.2 \quad \cdots \quad \cdots \quad 2.2 \quad 4.0 \qquad 3.0$$

$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are independent and identically distributed (**i.i.d.**) random variables.

$$\boldsymbol{X}_1 \sim \mathcal{D}(\mu, \sigma^2)$$
$$\boldsymbol{X}_2 \sim \mathcal{D}(\mu, \sigma^2)$$
$$\boldsymbol{X}_3 \sim \mathcal{D}(\mu, \sigma^2)$$
$$\vdots$$
$$\boldsymbol{X}_{n-1} \sim \mathcal{D}(\mu, \sigma^2)$$
$$\boldsymbol{X}_n \sim \mathcal{D}(\mu, \sigma^2)$$

$$\bar{\boldsymbol{X}} \sim ?(?, ?)$$

The Central Limit Theorem

By Pierre Simon de Laplace in 1810.

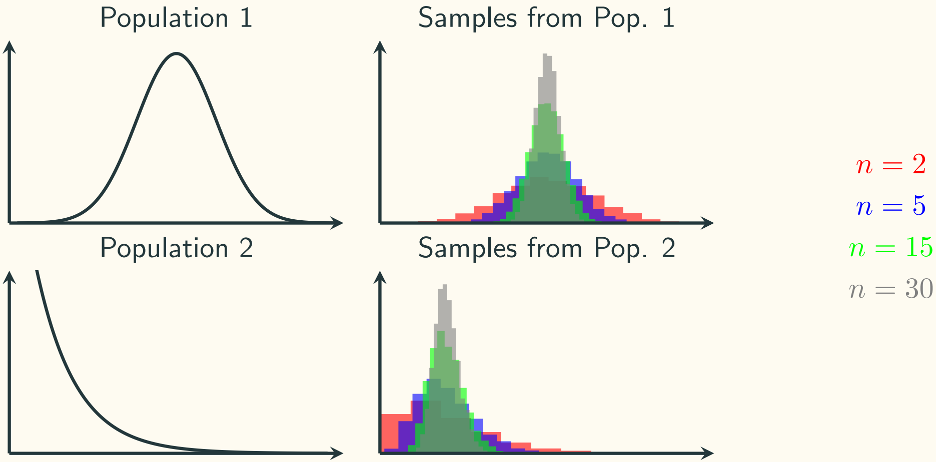
Theorem

The sampling distribution of the sample mean of n independent and identically distributed (i.i.d.) random variables is approximately normal, even if original variables themselves are not normally distributed, provided that n is large enough.

$$\bar{\boldsymbol{X}} \dot{\sim} \mathcal{N}(\mu_{\bar{\boldsymbol{X}}}, \sigma^2_{\bar{\boldsymbol{X}}}), \text{ where } \mu_{\bar{\boldsymbol{X}}} = \mu, \sigma^2_{\bar{\boldsymbol{X}}} = \frac{\sigma^2}{n}$$

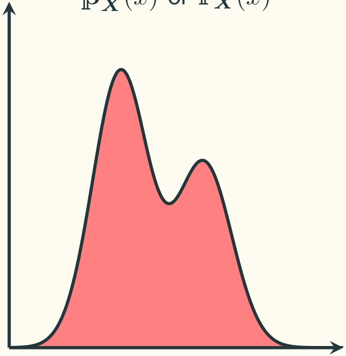
$\sigma_{\bar{\boldsymbol{X}}} = \frac{\sigma}{\sqrt{n}}$: **standard error**.

The Central Limit Theorem



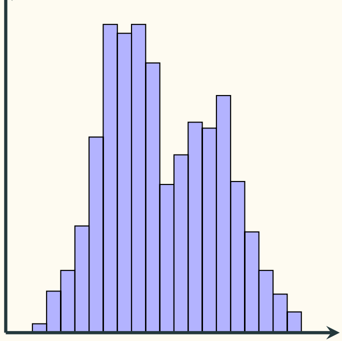
Three Distributions

Population Distribution
$\mathbb{p}_{\boldsymbol{X}}(x)$ or $\mathbb{f}_{\boldsymbol{X}}(x)$



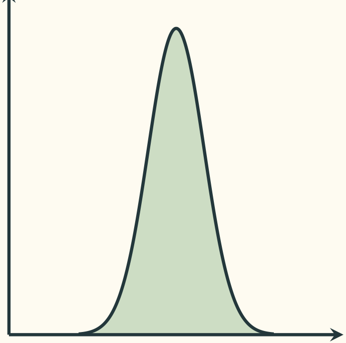
Population parameters
$\mathbb{E}\left[\boldsymbol{X}\right] = \mu,\ \mathbb{V}\mathrm{ar}\left(\boldsymbol{X}\right) = \sigma^2$

Sample
Distribution



Sample statistics
$\bar{x},\ s^2$

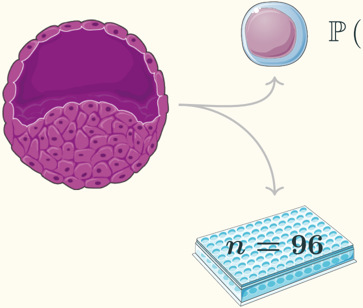
Sampling Distribution
of The Sample Mean



Parameters
$\mathbb{E}\left[\bar{\boldsymbol{X}}\right] = \mu_{\bar{\boldsymbol{X}}},\ \mathbb{V}\mathrm{ar}\left(\bar{\boldsymbol{X}}\right) = \sigma^2_{\bar{\boldsymbol{X}}}$

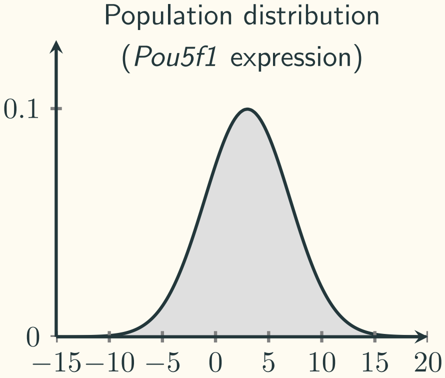
Practice: $Pou5f1$ Expression

Based on the previous research, the expression of $Pou5f1$ in all ES cells follow a normal distribution with $\mu = 3$ and $\sigma^2 = 4^2$.

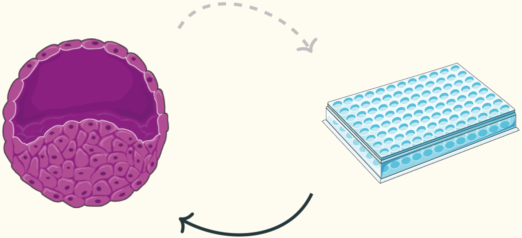


$\mathbb{P}(x \geqslant 8)$?

$\mathbb{P}(\bar{x} \geqslant 3.8)$?



Estimation



Use info. from the sample
to do a point estimation

Population parameter
μ, σ^2

Sample statistics
$\bar{x}, s^2$

- **Estimator**

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

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

- **Estimate**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Unbiased Estimator

We say the following estimators are unbiased estimators:

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

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

Because:

$$\mathbb{E}\left[\bar{\boldsymbol{X}}\right] = \mu$$

$$\mathbb{E}\left[\boldsymbol{S}^2\right] = \sigma^2$$