MH1820 Week 9

1. Population, random samples, statistics and sampling distribution
2. Law of large numbers and CLT
3. Parameter Estimation: Point Estimation

Population, random samples, statistics and sampling distribution

In statistics, a **population** is a set of objects or a certain kind of experiment that generates certain outcomes. A specific property of these objects is analyzed statistically.

Examples:

Population	Property
Undergraduate students in NTU	CGPA
Stars in the universe	Luminosity
Chess players in Singapore	Elo rating
Rolling a dice repeatedly	outcomes of rolls

- Instead of the whole population, often only a random subset is selected (easier, more efficient) for measurements of the property of interest.

  n = sample size.

- These measurements $x_1, x_2, \ldots, x_n$ (also called **observations/data**) can be modelled by random variables $X_1, X_2, \ldots, X_n$ (called **random sample**), which are assumed to be i.i.d (identically independently distributed),

- The distribution of the random variables X_i is called **population distribution**. ($\mathbb{E}[X_i]$ is called the **population mean**; $\mathrm{Var}[X_i]$ is called the **population variance**).

...*[continued]*

- n is called the **sample size**.
- x_1, ..., x_n can be viewed as realizations of i.i.d random variables X_1, ..., X_n.

Population

X

$\underline{\underline{\mu}}, \underline{\underline{\sigma^2}}$

random sample.

$X_1, X_2, \ldots, X_n$

i.i.d

estimate parameters like μ, σ^2, median, . etc.

statistic $T(X_1, X_2, .., X_n)$

Example 1

- Population: Undergraduate students at NTU
- Property: CGPA
- Population Distribution: $N(\mu, \sigma^2)$
- Random sample: n randomly chosen NTU students $X_1, \ldots, X_n$
- Observation/Data: $x_1, \ldots, x_n \in [0, 5]$
- Statistical model: $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$.

e.g If μ is unknown, we may estimate μ using the sample mean $\frac{X_1 + X_2 + \ldots + X_n}{n}$.

Example 2

- Population: Tossing a fair coin 10 times
- Property: Number of heads among the 10 tosses.
- Population Distribution: $\mathit{Binomial}(10, 0.5)$
- Random sample: n repetitions of 10 tosses.
- Observation/Data: $x_1, \ldots, x_n \in \{0, 1, \ldots, 10\}$
- Statistical model: $X_1, \ldots, X_n$ i.i.d $\sim \mathit{Binomial}(10, 0.5)$.

Let $X_1, \ldots, X_n$ be a random sample.

- A real valued function $T(X_1, \ldots, X_n)$ is called a **statistic**.
- The distribution of a statistic is called a **sampling distribution** of T.

Example 3

Let $X_1, \ldots, X_n$ be a random sample. Some examples of statistics.

- $T_1 = \sum_{i=1}^{n} X_i^2$
- $T_2 = \min\{X_1, \ldots, X_n\}$
- $T_3 = X_1$

$$T_4 = \frac{X_1 + X_2 + \cdots + X_n}{n} \quad \checkmark \qquad T_5 = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

Let $X_1, \ldots, X_n$ be an i.i.d random sample.

- **Population distribution**: distribution of X_i
- **Sampling distribution**: distribution of a statistic based on $X_1, \ldots, X_n$

$\downarrow$

$T(X_1, \ldots, X_n)$

$\downarrow$

estimate of some parameter like mean, variance, median etc.

$X_i \sim$ Population distribution.

Example 4

Let $X_1, \ldots, X_n$ be a random sample.

- **Sample mean**: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **Sample variance**: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Note that $\overline{X}$ and S^2 are statistics. Their distributions are examples of sampling distributions.

Theorem 5 (Random sample from Normal distribution)

Let $X_1, \ldots, X_n$ be observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$. Then the sample mean $\overline{X}$ and sample variance S^2 are independent, and

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Recall: $Z_i \sim N(0,1)$ i.i.d

$Z_i^2 \sim \chi^2(1)$

$\sum_{i=1}^n Z_i^2 \sim \chi^2(n).$

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

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2$$

$$= \sum_{i=1}^{n} z_i^2 \qquad z_i \sim N(0,1).$$

$$\sim \chi^2(n-1) \qquad \sum_{i=1}^{n} z_i = 0$$

$$\sum\left(\frac{x_i-\bar{x}}{\sigma}\right) = \frac{1}{\sigma}\sum_{i=1}^{n}(x_i-\bar{x})$$

$$= \frac{1}{\sigma}\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}\right)$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$= \frac{1}{\sigma}\left(n\bar{x} - n\bar{x}\right)$$

$$= 0.$$

$$\underline{z_1+z_2+z_3=0}$$

$$\underbrace{z_1+z_2+\dots}+z_n=0.$$

$X_i \sim N(\mu, \sigma^2)$ i.e $M_{X_i}(t) = e^{\mu t + \sigma^2 t/2}$.

Theorem 6 (Random sample from Normal distribution)

Let $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$. The sampling distribution of the sample mean $\overline{X}$ is given by

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

q

This implies that the **standardized sample mean** $\boxed{\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1).}$

Recall: $X \sim N(\mu, \sigma^2)$

MGF of X = $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$

- $M_{aX}(t) \overset{\star}{=} M_X(at)$
- $M_{X+Y}(t) \overset{\star\star}{=} M_X(t)\, M_Y(t)$

$$M_{\bar{X}}(t) = M_{\frac{1}{n}(X_1+\dots+X_n)}(t)$$

$$= M_{X_1+\dots+X_n}\left(\frac{1}{n}t\right) \quad \text{by } \ast.$$

$$= M_{X_1}\left(\frac{1}{n}t\right) M_{X_2}\left(\frac{1}{n}t\right) \cdots\cdots M_{X_n}\left(\frac{1}{n}t\right)$$

by $\ast\ast$

$$= \left(e^{\mu\frac{t}{n}+\sigma^2\frac{t^2}{2n^2}}\right)\cdot\left(e^{\mu\frac{t}{n}+\sigma^2\frac{t^2}{2n^2}}\right)\cdots\left(e^{\mu\frac{t}{n}+\sigma^2\frac{t^2}{2n^2}}\right)$$

$$= e^{\left(\mu\frac{t}{n}+\sigma^2\frac{t^2}{2n^2}\right)n} = e^{\mu t+\frac{\sigma^2}{n}\frac{t^2}{2}}.$$

So $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

or equivalently,

$$\boxed{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)}$$

Summary:

Population

X

$N(\mu, \sigma^2)$

$X_1, X_2, \ldots, X_n$.

- $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$
- $S^2 = \dfrac{\sum (X_i - \bar{X})^2}{n-1}$

Now know:

$$\begin{cases} \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) & \checkmark \\[2ex] \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) & \checkmark \end{cases}$$

Q: what if population distribution is not normal ?

is unknown ?

Remark: For increasing sample size n, the variance $\frac{\sigma^2}{n}$ tends to 0, and so the distribution of the sample mean $\overline{X}$ tends to the distribution of the constant μ. It turns out that this is true even if the random sample is not from a normal distribution!

(1) we speak of sampling distribution

of statistic $T(X_1, .., X_n)$ random sample.

if $T(X_1,..,X_n)$ = sample mean = $\bar{X}$

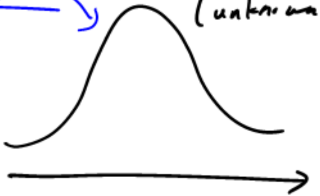
then sampling distribution of $\bar{X}$ = distribution of $\bar{X}$

②

random sample $\bar{X}$

sample size $= n$.

sampling distribution (unknown).



$$R_1 = (x_1^1, x_2^1, \ldots, x_n^1) \to \overline{x^1}$$

$$R_2 = (x_1^2, x_2^2, \ldots, x_n^2) \to \overline{x^2}$$

$$\vdots$$

$$R_m = (x_1^1, x_2, \ldots, x_n^m) \to \overline{x^{1000}}$$

$\bar{X}$ (depends on n)

Law of large numbers and CLT

True mean $\mu = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6}$

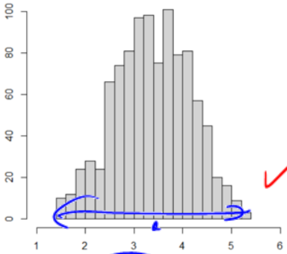
$= 3.5$.

An experiment:

- Roll a fair dice n times.
- Compute average $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where $x_i \in \{1, 2, \ldots, 6\}$ is the outcome of the ith roll.
- Repeat this 1000 times to get 1000 observations for $\overline{X}$.
- Plot a histogram of these 1000 observations to visualize the distribution of the average.

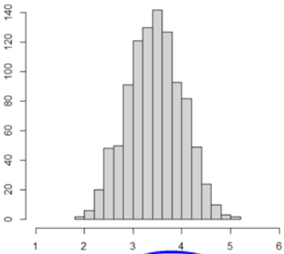
generate sampling distribution of $\overline{X}$.

Distribution of $\bar{X}$ (average result of rolling dice n times)



$n = 5$

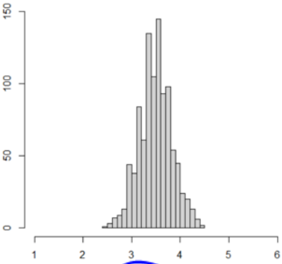
(1000 observations of $\frac{1}{5}\sum_{i=1}^{5} x_i$)



$n = 10$

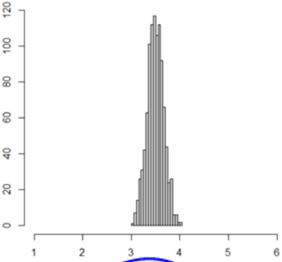
(1000 observations of $\frac{1}{10}\sum_{i=1}^{10} x_i$)

Distribution of $\bar{X}$ (average result of rolling dice n times)



$n = 25$

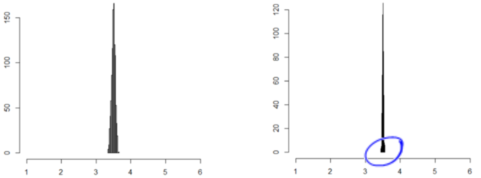
(1000 observations of $\frac{1}{25}\sum_{i=1}^{25} x_i$)



$n = 100$

(1000 observations of $\frac{1}{100}\sum_{i=1}^{100} x_i$)

Distribution of $\bar{X}$ (average result of rolling dice n times)



$n = 1000$
(1000 observations of $\frac{1}{1000}\sum_{i=1}^{1000} x_i$)

$n = 10000$
(1000 observations of $\frac{1}{10000}\sum_{i=1}^{10000} x_i$)

From this experiment, when n increases, the probability that $\overline{X}$ is close to the population mean $\mathbb{E}[X_i] = 3.5$ is getting higher. This fact is formalized by the Law of Large Numbers.

Theorem 7 (Law of Large Numbers)

Let $X_1, \ldots, X_n$ be i.i.d such that $\mu = \mathbb{E}[X_i]$ exists. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\mathbb{P}(|\overline{X} - \mu| < \epsilon) \to 1, \text{as } n \to \infty,$$

for all $\epsilon > 0$.

In other words, for increasing sample size, the distribution of the sample mean $\overline{X}$ tends to the distribution of the constant μ.

- In practice, we often encounter i.i.d random samples which are not normally distributed.
- The population distribution may even be totally unknown.
- In this situation, the exact distribution of $\overline{X}$ cannot be determined.
- For large samples, however, the Central Limit Theorem provides an approximation to the distribution of $\overline{X}$.

Theorem 8 (Central Limit Theorem (CLT))

Let $X_1, \ldots, X_n$ i.i.d with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then

$$\mathbb{P}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \Phi(x) \quad \text{for } n \to \infty.$$

Here, $\Phi(x)$ is the CDF of standard normal.

This means for large n, the standardized sample mean $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ approximately has a standard normal distribution.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

n large

The CLT is often used to approximate probabilities of sum of i.i.d:

$$
\begin{aligned}
\mathbb{P}(a \le \sum_{i=1}^{n} X_i \le b) &= \mathbb{P}\left(\frac{a}{n} \le \overline{X} \le \frac{b}{n}\right) \\
&= \mathbb{P}\left(\frac{a-n\mu}{n} \le \overline{X} - \mu \le \frac{b-n\mu}{n}\right) \\
&= \mathbb{P}\left(\frac{a-n\mu}{\sqrt{n}} \le \sqrt{n}(\overline{X}-\mu) \le \frac{b-n\mu}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\frac{a-n\mu}{\sigma\sqrt{n}} \le \boxed{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}} \le \frac{b-n\mu}{\sigma\sqrt{n}}\right) \\
&\approx \Phi\left(\frac{b-n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a-n\mu}{\sigma\sqrt{n}}\right),
\end{aligned}
$$

$\sim N(0,1)$

by CLT when n is large.

$\mu = p = 0.8$

$\sigma^2 = 0.8(0.2)$

Example 9

$X_1, \ldots, X_{100}$ i.i.d $\sim$ *Bernoulli*(0.8). Approximate $\mathbb{P}(70 \le X_1 + \cdots + X_{100} \le 90)$.

$n = 100$

Solution. $\mathbb{P}\left(70 \le \sum_{i=1}^{100} X_i \le 90\right)$

$$= \mathbb{P}\left(\frac{70}{100} \le \underbrace{\frac{1}{100}\sum_{i=1}^{100} X_i}_{\bar{X}} \le \frac{90}{100}\right)$$

$$= \mathbb{P}\left(\frac{0.7-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{0.9-\mu}{\sigma/\sqrt{n}}\right)$$

(by CLT)

$$\approx \mathbb{P}\left(\frac{0.7-0.8}{\sqrt{0.8(0.2)}/\sqrt{100}} \le Z \le \frac{0.9-0.8}{\sqrt{0.8(0.2)}/\sqrt{100}}\right)$$

$$\approx \mathbb{P}(-2.5 \leq Z \leq 2.5)$$

$$= \Phi(2.5) - \Phi(-2.5)$$

$n = 20.$

Example 10

Let $X_1, X_2, \ldots, X_{20}$ be a i.i.d random sample of size 20 from the uniform distribution $U(0,1)$. Let $Y = X_1 + X_2 + \cdots + X_{20}$. Use CLT to approximate the following probabilities.

(a) $\mathbb{P}(Y \leq 9.1)$;

(b) $\mathbb{P}(8.5 \leq Y \leq 11.7)$.

Solution. Note that $\mathbb{E}[X_i] = 1/2$ and $\mathrm{Var}[X_i] = 1/12$ for $i = 1, \ldots, 20$.

$\mu = \frac{1}{2}$

$\sigma^2 = \frac{1}{12} = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$= \int_0^1 x^2\,dx - \left(\frac{1}{2}\right)^2$

$\frac{1}{3} - \frac{1}{4} = \left[\frac{x^3}{3}\right]_0^1 - \frac{1}{4}$

$$\mathbb{P}(Y \leq 9.1) = \mathbb{P}\left(\sum_{i=1}^{20} X_i \leq 9.1\right)$$

$$= \mathbb{P}\left(\frac{1}{20}\sum_{i=1}^{20} X_i \leq \frac{9.1}{20}\right)$$

$$= \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{\frac{9.1}{20}-\frac{1}{2}}{\frac{1}{12}/\sqrt{20}}\right)$$

$$= \mathbb{P}\left(Z \leq \frac{\frac{9.1}{20}-\frac{1}{2}}{\frac{1}{12}/\sqrt{20}}\right).$$

$$\mathbb{P}(Y \leq 9.1) \approx \Phi\left(\frac{9.1 - 20(1/2)}{\underline{\sqrt{1/12}\sqrt{20}}}\right) = \Phi(-0.6971)0.2429.$$

$$\begin{aligned}\mathbb{P}(8.5 \leq Y \leq 11.7) &\approx \Phi\left(\frac{11.7 - 20(1/2)}{\sqrt{1/12}\sqrt{20}}\right) - \Phi\left(\frac{8.5 - 20(1/2)}{\sqrt{1/12}\sqrt{20}}\right) \\ &= \Phi(-1.162) - \Phi(1.317) = 0.7835.\end{aligned}$$

$\square$

Recall: MGF $X \sim \text{Poisson}(\lambda)$

$$M_X(t) = e^{\lambda(e^t - 1)}$$

Example 11

Explain how a Poisson distribution with mean $\lambda = 20$ can be approximated with the use of a normal distribution.

Let $Y \sim \text{Poisson}(\lambda = 20)$.

$$= \underbrace{\text{Poisson}(1) + \ldots + \text{Poisson}(1)}_{20 \text{ times.}}$$

Let $Y_i \sim \text{Poisson}(\lambda = 1)$

$$M_{Y_1 + \dots + Y_{20}}(t) = M_{Y_1}(t) \cdots M_{Y_{20}}(t)$$

$$= e^{1 \cdot (e^t - 1)} \cdot \ldots \cdot e^{1 \cdot (e^t - 1)}$$

$$= e^{(e^t - 1) \times 20}$$

$$= M_Y(t) \qquad Y \sim \text{Poisson}(20)$$

By CLT:

$Y_i \sim \text{Poisson}(\lambda = 1)$

$$Y = Y_1 + \cdots + Y_{20}.$$

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

$\mu = \mathbb{E}[Y_i] = 1$ ✓

$\sigma^2 = \text{Var}[Y_i] = 1$

$$\frac{\frac{Y}{20} - 1}{1/\sqrt{20}} \approx N(0,1)$$

$$\boxed{\sqrt{20}\left(\frac{Y}{20} - 1\right) \approx N(0,1).}$$

e.g. $\sqrt{20}\left(\frac{Y}{20}-1\right) \sim N(0,1)$

$$\mathbb{P}(Y \leq a) = \mathbb{P}\left(\frac{Y}{20} \leq \frac{a}{20}\right)$$

$$= \mathbb{P}\left(\frac{Y}{20}-1 \leq \frac{a}{20}-1\right)$$

$$= \mathbb{P}\left(\underbrace{\sqrt{20}\left(\frac{Y}{20}-1\right)} \leq \sqrt{20}\left(\frac{a}{20}-1\right)\right)$$

$$\approx \underline{\underline{\Phi}}\left(\sqrt{20}\left(\frac{a}{20}-1\right)\right). \checkmark$$

Take-away :

① $\bar{X} \sim N(\mu, \sigma^2/n)$

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

} assume population is $N(\mu, \sigma^2)$

i.e $X_i \sim N(\mu, \sigma^2)$

② CLT : $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ n large

use CLT to approximate sum of i.i.d

$\mu = \mathbb{E}[X_i]$

$\sigma^2 = Var[X_i]$