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.

- 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.

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

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, ..., $x_n \in \{0, 1, \ldots, 10\}$
- Statistical model: X_1, ..., X_n i.i.d $\sim \mathit{Binomial}(10, 0.5)$.

Let X_1, ..., 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**.

Example 3

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

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

Let X_1, ..., 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, ..., X_n

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

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

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

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

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

Recall that if $X \sim N(\mu, \sigma^2)$, then its MGF is $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$. Then

$$M_{aX}(t) = M_X(at) = e^{\mu a t + \sigma^2 a^2 t^2/2},$$

that is

$$aX \sim N(a\mu, a^2\sigma^2).$$

Hence, for each i,

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

Example 7 of Week 6 Slides: If $Y_i \sim N(\mu_i, \sigma_i^2)$ and $Y_j \sim N(\mu_j, \sigma_j^2)$ are independent, then $Y_i + Y_j \sim N(\mu_i + \mu_j, \sigma_i + \sigma_j)$.

Since the X_i are i.i.d, it follows that

$$\begin{aligned} \overline{X} &= \frac{1}{n}(X_1 + \cdots + X_n) \sim N\left(\frac{1}{n}\sum_{i=1}^{n}\mu, \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right) \\ &= N\left(\mu, \frac{\sigma^2}{n}\right). \end{aligned}$$

□

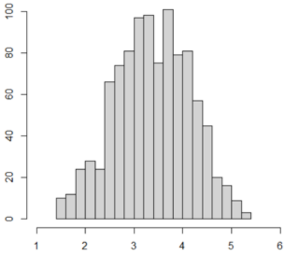
Remark: For increasing sample size n, the variance $\frac{\sigma^2}{n}$ tends to 0, and so the distribution of the sample mean $\overline{X}$ tends to the distribution of the constant μ. It turns out that this is true even if the random sample is not from a normal distribution!

Law of large numbers and CLT

An experiment:

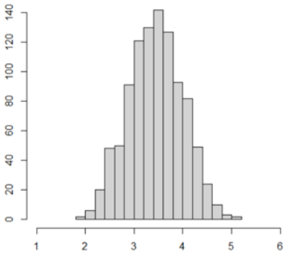
- Roll a fair dice n times.
- Compute average $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where $x_i \in \{1, 2, \ldots, 6\}$ is the outcome of the ith roll.
- Repeat this 1000 times to get 1000 observations for $\overline{X}$.
- Plot a histogram of these 1000 observations to visualize the distribution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

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



$n = 5$

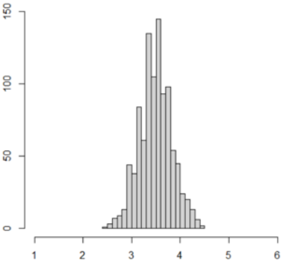
(1000 observations of $\frac{1}{5}\sum_{i=1}^{5} x_i$)



$n = 10$

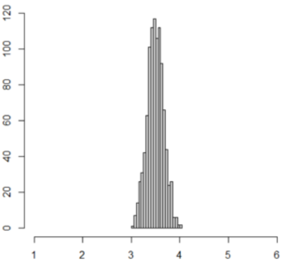
(1000 observations of $\frac{1}{10}\sum_{i=1}^{10} x_i$)

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



$n = 25$

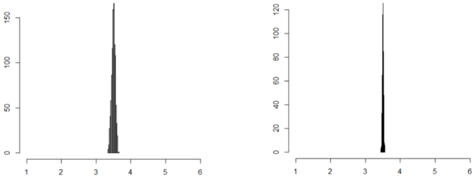
(1000 observations of $\frac{1}{25}\sum_{i=1}^{25} x_i$)



$n = 100$

(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 location of the sample mean $\overline{X}$ tends to get closer and closer to 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

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

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.

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

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

by CLT when n is large.

Example 9

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

Solution. $\mu = \mathbb{E}[X_i] = 0.8$, $\sigma = \sqrt{0.8 \cdot 0.2} = 0.4$. Thus,

$$\begin{aligned} \mathbb{P}(70 \leq \sum_{i=1}^{100} X_i \leq 90) &\approx \Phi\left(\frac{90 - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{70 - \mu n}{\sigma\sqrt{n}}\right) \\ &\approx \Phi(2.5) - \Phi(-2.5) \end{aligned}$$

$\square$

Example 10

Let X_1, X_2, ..., 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 $i = 1, \ldots, 20$.

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

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

$\square$

Example 11

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

Let $Y \sim Poisson(20)$.

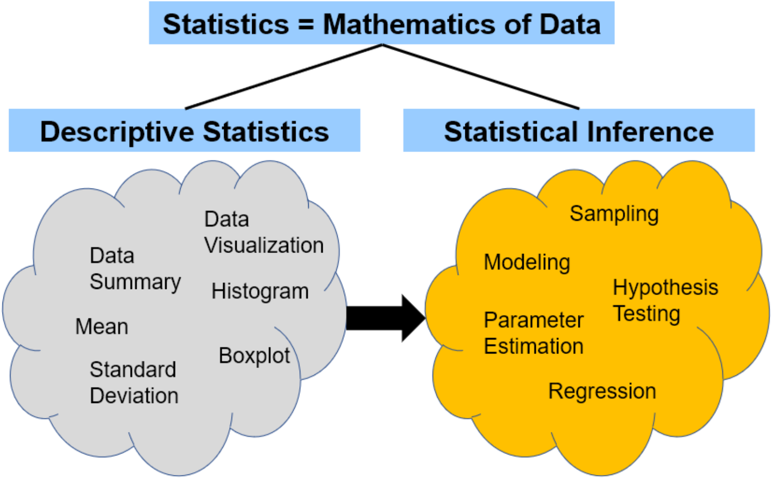
Consider 20 i.i.d Poisson random variables Y_1, ..., Y_{20}, where each $Y_i \sim Poisson(1)$.

Can think of Y as the sum of Y_i. By CLT,

$$\frac{Y - 20}{\sqrt{20}} = \frac{\frac{1}{20}\sum_{i=1}^{20} Y_i - 1}{1/\sqrt{20}}$$

has a distribution which is approximately $N(0, 1)$. $\square$

Parameter Estimation: Point Estimation


Statistics = Mathematics of Data
Descriptive Statistics
Statistical Inference
Data Summary
Data Visualization
Histogram
Mean
Boxplot
Standard Deviation
Sampling
Modeling
Hypothesis Testing
Parameter Estimation
Regression

Random sample $X_1, \ldots, X_n$ i.i.d.

Often the type of distribution ($N(\mu, \sigma^2)$, $Exp(\theta)$ etc.) of X_i is known, but its parameters μ, σ, θ etc. are unknown.

Parameter estimation: Extract information from $X_1, \ldots, X_n$ on these parameters.

- A **point estimator** is a random variable that provides a "best guess" for a parameter.
- An **interval estimate** produces an interval with random endpoints such that the true parameter (hopefully) with high probability is contained in the interval

Process of Point Estimation.

Given data/observations $x_1, \ldots, x_n$ as realizations of $X_1, \ldots, X_n$.

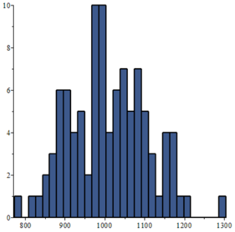
- Modeling: Identify a suitable type of distribution for X_i, which depends on parameter θ.
- Point estimation: Find functions $\widehat{\theta}(X_1, \ldots, X_n)$ which approximates θ.
- Substitute data $X_1 = x_1, \ldots, X_n = x_n$ into these functions to get estimates for θ.

Example: Data $x_1, \ldots, x_{100}$ (measurements in a physics experiment)

[1067., 773.2, 1119., 938.2, 1166., 1006., 881.4, 995.9, 1102., 1056.,
1045., 1091., 1170., 1085., 893.9, 1097., 1054., 959.3, 975.3, 969.4,
971.6, 1024., 984.2, 929.4, 1061., 998.4, 1209., 901.8, 864.2, 978.0,
1025., 1143., 858.0, 890.2, 1110., 1195., 944.0, 846.7, 872.7, 925.9,
1028., 980.5, 870.3, 1071., 1057., 1044., 987.0, 999.8, 981.4, 911.6,
1014., 1012., 825.4, 991.1, 1034., 944.8, 1001., 1097., 1149., 929.0,
1081., 994.1, 1174., 1050., 1162., 1081., 976.1, 1109., 1127., 1053.,
899.9, 1080., 941.4, 947.5, 1033., 912.1, 912.5, 1077., 1072., 1082.,
1005., 914.0, 1054., 883.9, 1164., 925.0, 1305., 1036., 998.7, 885.4,
998.2, 955.3, 883.7, 1155., 1095., 827.5, 993.0, 1152., 968.4, 976.6]

Step 1: Modeling: Normal distribution $N(\mu, \sigma^2)$ seems appropriate for this data. Want to estimate two parameters: μ, σ.

Histogram:



Step 2: Find functions to estimate

- Use sample mean to estimate μ (in view of Law of Large Number):

$$\mu \approx \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Use sample variance to estimate σ (not clear at this point why this is a good estimate):

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

Step 3: Sub in the data

- $\mu \approx \overline{x} = \frac{1}{100}\sum_{i=1}^{100} x_i = 1010.45.$
- $\sigma \approx S = \sqrt{\frac{1}{99}\sum_{i=1}^{100}(x_i - 1010.45)^2} = 98.54.$

$\square$

Idea:

When estimating a parameter θ which can be expressed as a **function of mean or variance**, we expect

$$\text{(sample mean)} \;\; \overline{X} \approx \mathbb{E}[X_i] \;\; \text{(population mean)};$$
$$\text{(sample variance)} \;\; S^2 \approx \mathrm{Var}[X_i] \;\; \text{(population variance)}.$$

From the above, we then deduce an estimate $\widehat{\theta}$ of the parameter.

Example 12

Let $X_1, \ldots, X_m$ be i.i.d $\sim \mathit{Binomial}(n, p)$, where n and p are both unknown.

- Given: Observations $x_1, \ldots, x_m$
- Goal: Estimate n and p from $x_1, \ldots, x_m$

Idea: We expect:

$$\overline{X} \approx \mathbb{E}[X_i] = np \tag{1}$$

$$S^2 \approx \mathrm{Var}[X_i] = np(1-p) \tag{2}$$

$\Longrightarrow 1 - p \approx S^2/\overline{X} \Longrightarrow p \approx 1 - S^2/\overline{X};$

$\Longrightarrow n \approx \frac{\overline{X}}{1-S^2/\overline{X}}$

Sample data: $x_1, \ldots, x_{1000}$ drawn from $Binomial(n, p)$ with unknown n and p.

12 13 7 10 6 11 12 6 7 10 10 9 10 15 11 14 9 8 9 12 10 16 16 135 11 7 6 9 19 7 16 9 11 14 9 8 13 16 8 7 11 10
10 7 12 11 1312 18 8 8 12 16 11 14 7 7 10 13 4 8 7 6 8 9 12 12 11 14 8 3 8 9 12 9 7 10 8 10 8 5 10 9 13 10
10 8 12 6 8 9 11 10 14 7 7 9 8 12 12 7 9 5 9 9 10 6 8 10 12 13 13 11 10 15 9 9 12 12 10 12 9 6 8 12 3 7 9
11 8 7 10 10 8 15 9 11 12 11 6 7 9 7 8 11 13 3 12 10 11 9 6 11 13 14 10 10 10 8 18 7 15 11 7 6 10 710 7 7 13
7 9 14 12 7 14 10 15 13 12 7 5 14 13 8 8 9 8 9 8 9 8 8 7 12 14 12 6 8 12 6 4 9 10 11 14 7 6 13 9 10 10 8 8
7 9 9 10 8 10 9 13 10 11 13 11 7 9 6 11 8 13 13 9 16 15 9 11 6 11 13 12 12 16 8 10 10 17 7 9 11 9 9 10 12 8 12
3 12 6 10 10 12 14 5 10 3 9 9 14 10 12 14 7 13 11 15 8 4 7 8 8 14 8 11 9 11 8 10 11 8 9 10 11 14 9 13 7 6 8
9 11 7 9 7 9 10 5 9 11 11 7 8 11 3 13 9 8 9 5 11 15 11 4 8 11 14 13 8 14 10 5 10 13 12 5 17 10 8 14 10 10 11
13 8 16 9 10 11 9 6 6 12 10 12 12 9 8 11 11 16 12 8 7 11 12 14 10 16 10 9 8 11 15 11 7 10 8 12 12 9 12 11 9 8
11 9 10 12 7 11 12 12 10 12 6 8 9 13 5 4 15 11 10 10 11 10 12 15 12 11 11 7 8 14 7 14 9 9 7 12 6 10 10 6 12 11
10 9 9 11 11 11 4 12 13 11 11 10 8 7 8 14 9 12 12 13 13 4 10 8 8 10 6 10 16 12 13 10 8 12 9 13 11 9 8 7 8 6
18 6 9 6 10 12 10 9 12 10 7 11 6 8 4 11 9 9 16 10 8 10 8 12 11 13 11 14 14 5 11 7 11 7 9 10 10 9 10 12 12 8 8
9 11 12 9 9 14 8 11 10 12 11 11 12 15 10 11 16 7 11 14 15 12 10 9 12 11 8 17 11 13 10 7 10 12 12 9 11 7 9 11 13
10 10 13 9 6 12 9 12 8 9 14 11 5 8 13 13 15 8 11 5 9 5 14 14 11 14 10 11 11 9 13 10 8 12 7 9 11 7 7 7 12 10
10 9 6 11 13 11 10 10 12 12 9 9 11 12 13 9 7 7 8 8 8 9 13 10 11 10 8 7 10 10 9 11 6 12 8 10 14 8 11 8 17 18
11 9 9 11 7 12 12 16 15 7 7 6 14 15 8 5 9 12 10 7 9 13 8 7 11 11 15 9 8 11 8 9 6 5 11 6 8 5 11 15 7 8 6 7
9 10 10 5 4 4 10 9 7 7 5 6 8 9 4 11 12 13 9 9 8 10 9 9 10 11 11 7 13 13 12 5 10 6 11 10 9 11 12 10 8 11 9 8
14 8 14 7 8 6 16 10 8 7 11 13 12 12 12 9 6 10 7 10 9 12 14 7 7 8 8 6 8 13 8 12 11 8 12 10 7 7 12 13 10 10 13
5 13 13 9 15 10 10 11 13 13 11 9 12 10 12 7 9 10 6 9 16 9 16 8 11 7 13 4 7 10 11 18 10 12 4 10 9 15 11 9 8 15
10 9 16 11 7 6 10 6 9 12 14 11 11 9 10 15 8 13 11 8 9 10 8 12 8 11 9 7 10 6 10 7 9 10 9 7 10 9 6 8 12 11 13
8 14 8 8 7 12 10 15 7 7 10 9 8 8 12 14 4 12 2 10 10 10 13 4 12 12 9 9 12 11 9 13 613 8 8 10 9 9 10 13 9 14
12 7 2 9 12 7 10 15 14 7 12 10 12 10 12 9 13 12 7 14 10 13 13 6 12 9 7 6 7 13 6 6 10 7 11 15 9 15 11 12 9 11
16 9 13 10 10 11 12 11 6 9 10 10 9 10 10 5 8 10 12 10 14 7 5 7 14 9 9 10 7 14 10 9 12 12 13 8 7 10 12 12 13 7
10 12 17 9 7 8 12 12 13 11 10 9 6 14 13 13 11 11

- $\overline{x} = \frac{1}{1000} \sum_{i=1}^{1000} x_i = 9.959$
- $s^2 = \frac{1}{999} \sum_{i=1}^{1000} (x_i - \overline{x})^2 = 7.749068$
- $p \approx 1 - s^2/\overline{x} \approx 0.22$
- $n \approx \frac{\overline{x}}{1 - s^2/\overline{x}} \approx 44.88.$

Observations where actually drawn from *Binomial*$(50, 0.2)$. $\square$

Conclusion:

- Sample mean and variance can be useful to estimate unknown parameters of the populations distribution.
- However, the arguments used so far are "ad-hoc" and we do not yet have a way to measure the accuracy of the estimation.
- More systematics methods are needed for parameter estimation in general.