Point and Interval Estimation

Mathematics Education Section
CDI, EMB

Email： stchan@emb.gov.hk

Normal Distribution

6. **Normal Distribution**	
6.1 Basic definition and properties	6.1.1 recognise the concepts of a continuous random variable and a continuous probability distribution, with reference to the normal distribution

Continuous random variables and normal distribution

Let X be a continuous random variable defined on an interval Ω.

The probability distribution of X is described by a function $f(x)$, where $x \in \Omega$, such that

(a) $f(x) \geq 0, \quad x \in \Omega$.

(b) $\int_{\Omega} f(x)dx = 1$

(c) $P(a \leq X \leq b) = \int_a^b f(x)dx$

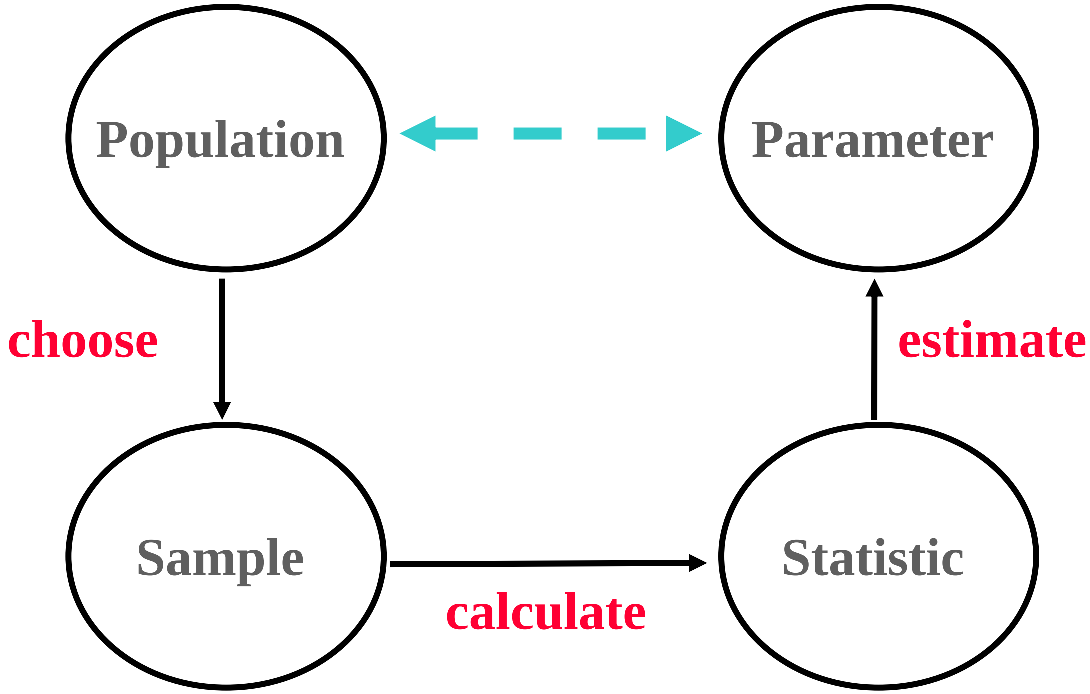
This function f(x) is then called the probability density function (*pdf*) of the random variable X.

Point and Interval Estimation

7.1 Sampling distribution and point estimates	7.1.1 recognise the meaning of sample statistics and population parameters
	7.1.2 recognise the sampling distribution of the sample mean from a random sample of size n
	7.1.3 recognise the meaning of point estimates including the sample mean, sample variance and sample proportion
	7.1.4 recognise the concept of Central Limit Theorem

Parameter and Statistic

- A number that describes a population is called a parameter
- A number that describes a sample is a statistic
- If we take a sample and calculate a statistic, we often use that statistic to infer something about the population from which the sample was drawn



The sample mean
= the population mean ?

Rolling a fair dice

- Cast a fair dice and take X to be the uppermost number, the population mean is $\mu = 3.5$, and that the population median is also $m = 3.5$
- Take a sample of four throws, the mean may be far from 3.5

The results of 5 such samples of 4 throws

	X_1	X_2	X_3	X_4	$\overline{X}$
Sample 1	6	2	5	6	4.75
Sample 2	2	3	1	6	3
Sample 3	1	1	4	6	3
Sample 4	6	2	2	1	2.75
Sample 5	1	5	1	3	2.75

The sample size $n = 4$

Random samples

$\{X_1, X_2, \ldots, X_n\}$ is said to be a **random sample** taken from a population of (the values of a random variable) X with pdf $f(x)$ if

(a) $X_1, X_2, \ldots, X_n$ are **independent**;

(b) All X_i ($i = 1, 2, \ldots, n$) have the **same pdf** $f(x)$.

Sampling distribution: Basic theorem

1. If $\{X_1, X_2, \ldots, X_n\}$ is a random sample taken from a population of X with mean μ and variance σ^2, then

$$E(\overline{X}) = \mu \quad \text{and} \quad V(\overline{X}) = \frac{\sigma^2}{n}.$$

2. If $\{X_1, X_2, \ldots, X_n\}$ is a random sample from $N(\mu, \sigma^2)$, then

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

Central Limit Theorem (CLT)

3. (The **Central Limit Theorem**)

If $\{X_1, X_2, \ldots, X_n\}$ is a random sample taken from a population of X with mean μ and variance σ^2, then when n is large ($n \geq 30$)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \quad \text{approximately.}$$

Central Limit Theorem (CLT)

- For ANY population (regardless of its shape) the distribution of sample means will approach a normal distribution as *n* approaches *infinity*

- The variability of a sample mean decreases as the sample size increases

Central Limit Theorem **(CLT)**

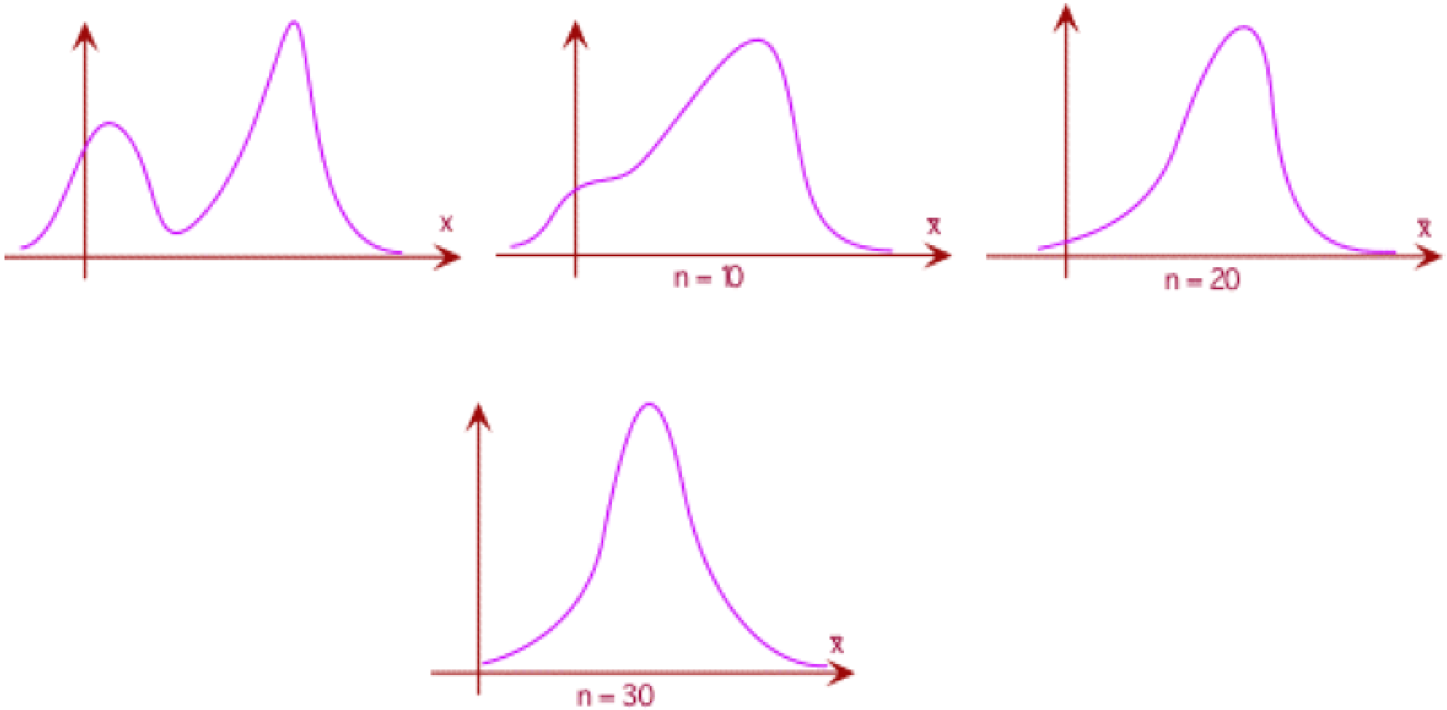
How large is a "large sample"?

It depends upon the form of the distribution from which the samples were taken. If the population distribution deviates greatly from normality larger samples will be needed to approximate normality.

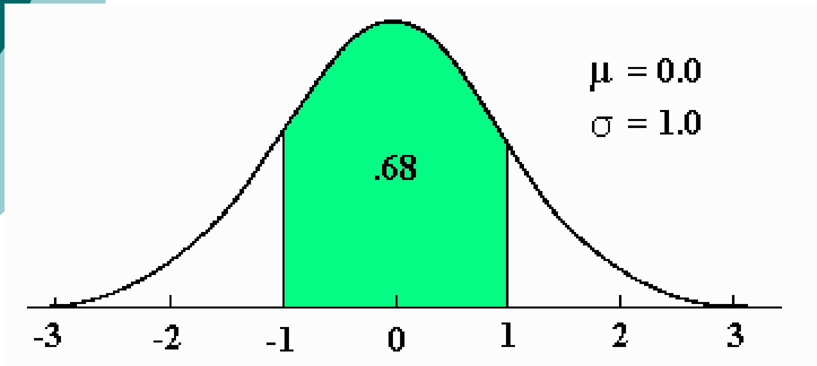
$$n \geq 30$$

Central Limit Theorem (CLT)

The following illustration shows how the sample size effects the shape of the sampling distribution.

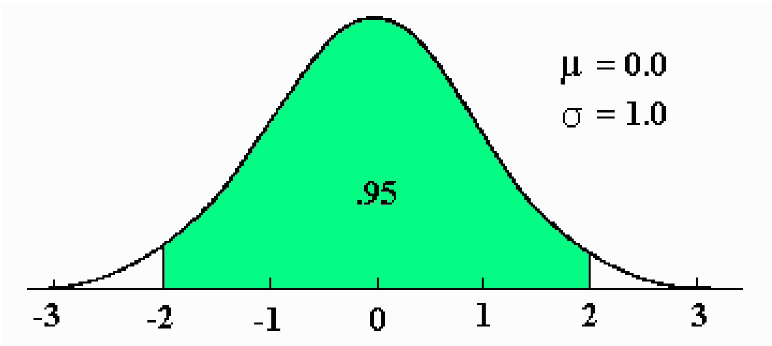


The Central Limit Theorem



$$\overline{X} : N(\mu, \frac{\sigma}{\sqrt{n}})$$

68% probability that our $\overline{X}$ will be in this region



95% probability that our $\overline{X}$ will be in this region

POINT AND INTERVAL ESTIMATION

- **Point Estimate**
  - is a single number, calculated from available sample data, that is used to estimate the value of an unknown population parameter.

- **Interval Estimate (Confidence Interval)**
  - is an interval that provides an upper and lower bound for a specific unknown population parameter.

-> arguably the most useful type of inference.

Estimators

1. An estimator $\hat{\theta}$ is said to be **unbiased** for the parameter θ if $E\left(\hat{\theta}\right) = \theta$

2. If this equality does not hold, $\hat{\theta}$ is said to be a **biased estimator** of θ, with $\text{Bias} = E\left(\hat{\theta}\right) - \theta$

Questions:

- An unfair coin has a 75% chance of landing heads-up.

  Let X = 1 if it lands heads-up,

  and X = 0 if it lands tails-up.

- Find the sampling distribution of the mean $\overline{X}$ for samples of size 3.

Solution:

- The experiment consists of tossing a coin 3 times and measuring the sample mean $\overline{X}$

	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Probability	27/64	9/64	9/64	3/64	9/64	3/64	3/64	1/64
$\overline{X}$	1	2/3	2/3	1/3	2/3	1/3	1/3	0

The possible values of $\overline{X}$ are 0, 1/3, 2/3 and 1.

The distribution of the sample mean is a binomial distribution

$\overline{X}$	0	1/3	2/3	1
P($\overline{X} = \overline{x}$)	1/64	9/64	27/64	27/64

Is the Sample Mean an Unbiased estimator of the Population Mean?

Example 1

An engineer wish to estimate the mean yield of a chemical process based on the yield measurements X_1, X_2, X_3 from three independent runs of an experiment.

Consider the following two estimators of the mean yield θ:

- $\hat{\theta}_1 = \dfrac{X_1 + X_2 + X_3}{3}$ and $\hat{\theta}_2 = \dfrac{X_1 + 2X_2 + X_3}{4}$.
- Which one should be preferred?

Solution:

- Since $E\left(\hat{\theta}_1\right) = E\left(\overline{X}\right) = \theta$ and

$$E\left(\hat{\theta}_2\right) = \frac{E\left(X_1\right) + 2E\left(X_2\right) + E\left(X_3\right)}{4} = \frac{\theta + 2\theta + \theta}{4} = \theta,$$

both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased for θ.

Solution:

Denoting the population variance by σ^2, we have

$$V\left(\hat{\theta}_1\right) = V\left(\overline{X}\right) = \frac{\sigma^2}{3}$$ and

$$V\left(\hat{\theta}_2\right) = \frac{V\left(X_1\right) + 4V\left(X_2\right) + V\left(X_3\right)}{16} = \frac{3\sigma^2}{8}.$$

$\frac{\sigma^2}{3} < \frac{3\sigma^2}{8}$, $\hat{\theta}_1$ is better than $\hat{\theta}_2$.

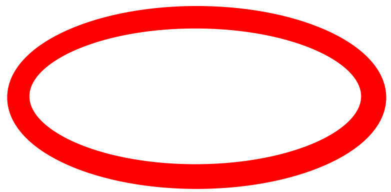
Confidence Interval

7.2 Confidence interval for a population mean	7.2.1 recognise the meaning of confidence interval 7.2.2 find a confidence interval for a population mean

Confidence Interval

'… perhaps the most obvious difficulty with confidence intervals lies in how we interpret what the confidence statement means' (Smithson, 2003, p.16)

Explanation:



Confidence Interval

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

$$P(-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

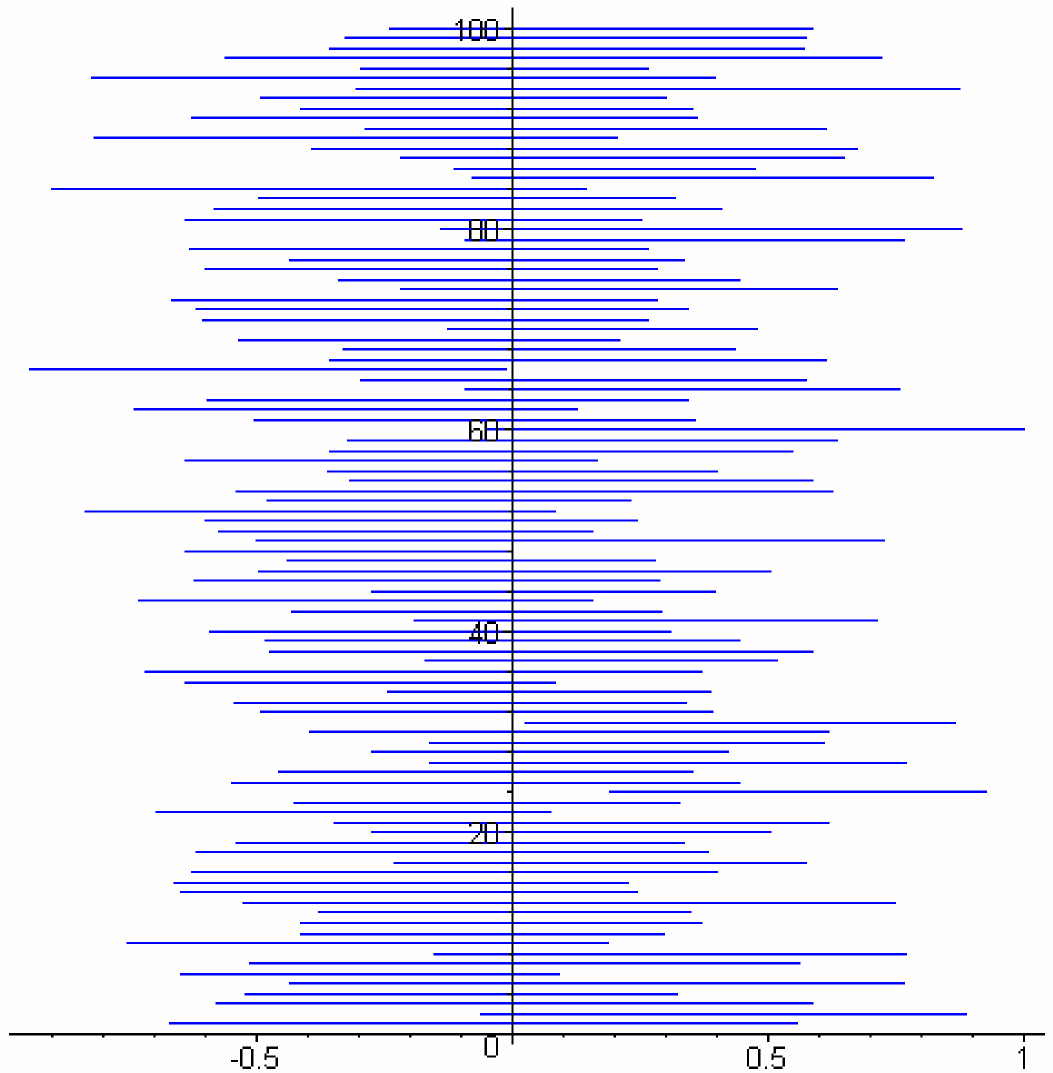
$$Z = \frac{\left(\overline{X} - \mu\right)}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$$P(-z_{\frac{\alpha}{2}} < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$P(\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$(\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \ , \ \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

a 100(1-α)% confidence interval of μ



a 100(1-α)% confidence interval of μ

- Large samples have narrower widths than small samples
- Higher confidence levels have wider intervals than lower confidence levels
- Narrow widths and high confidence levels are desirable, but these two things affect each other

a 100(1-α)% confidence interval of μ

Estimating the mean μ

1. Suppose that $\{X_1, X_2, \ldots, X_n\}$ is a random sample taken from the normal population $N(\mu, \sigma^2)$, where σ^2 is known.

   A 100(1 − α)% confidence interval (CI) of μ is given by

$$\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

   where $Z_{\frac{\alpha}{2}}$ is the value such that $P\left(Z \geq Z_{\frac{\alpha}{2}}\right) = \dfrac{\alpha}{2}$.

a 100(1-α)% confidence interval of μ

What sample size will ensure that the estimate is within $\pm\,0.5$ of the true mean in 90% of the cases?

The required sample size n is determined by solving the inequality

$$1.645\frac{2}{\sqrt{n}} \le 0.5.$$

i.e. $$n \ge \left(\frac{1.645\times 2}{0.5}\right)^2 \approx 44\text{, when rounded up.}$$

In general $$n \ge \left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2,$$

where E is the maximum error of estimate.

a 100(1-α)% confidence interval of μ

If σ^2 is unknown, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{X} \pm t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},$$

Confidence interval for a population proportion

7.3 Confidence interval for a population proportion	7.3.1 find an approximate confidence interval for a population proportion

Confidence interval for a population proportion *p*

Assumptions:

1. **The sample is a random sample**
2. **The conditions for the binomial distribution are satisfied**
3. **The normal distribution can be used to approximate the binomial distribution**

Confidence interval for a population proportion *p*

$$
\begin{aligned}
\mu_{\hat{p}} &= \mathrm{E}(\hat{p}) \\
&= \mathrm{E}\left(\frac{X}{n}\right) \\
&= \frac{1}{n}\mathrm{E}(X) \\
&= \frac{np}{n} \\
&= p
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{\hat{p}}^{\;2} &= \sigma_{\frac{X}{n}}^{\;2} \\
&= \frac{1}{n^2}\sigma_X^{\;2} \\
&= \frac{1}{n^2}[np(1-p)] \\
&= \frac{p(1-p)}{n}
\end{aligned}
$$

Confidence interval for a population proportion *p*

Estimating the proportion p

Let $\{X_1, X_2, \ldots, X_n\}$ be a large random sample ($n \geq 30$) from a Bernoulli population Ber(p).

Since the sample proportion

$$\hat{p} = \overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ approximately,}$$

an estimated $100(1-\alpha)\%$ CI for p is given by

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Thank You