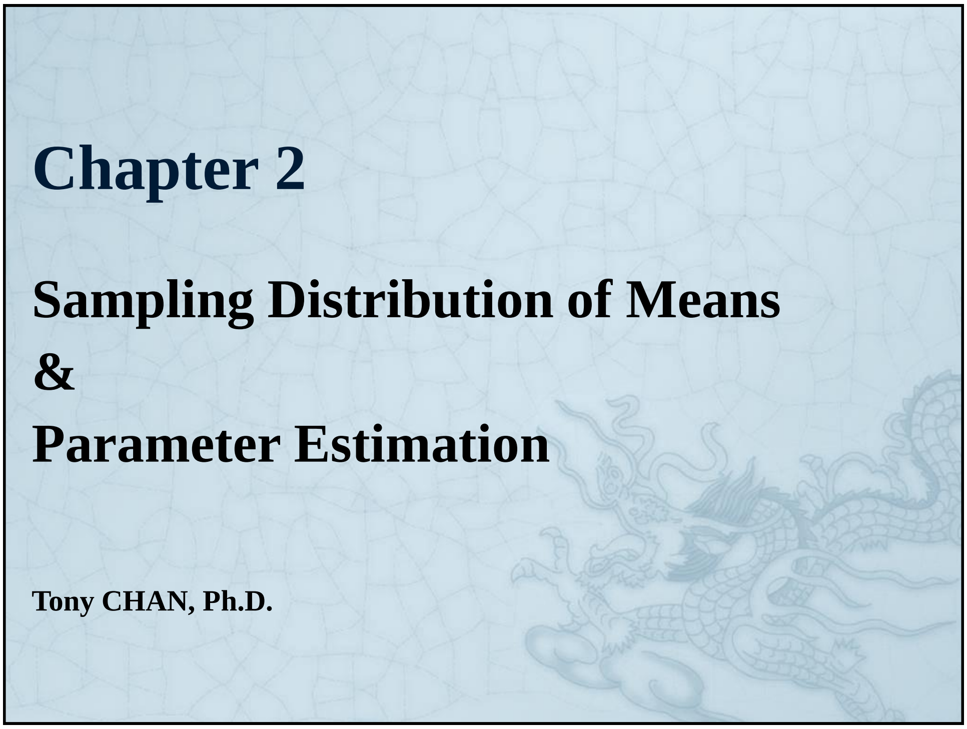
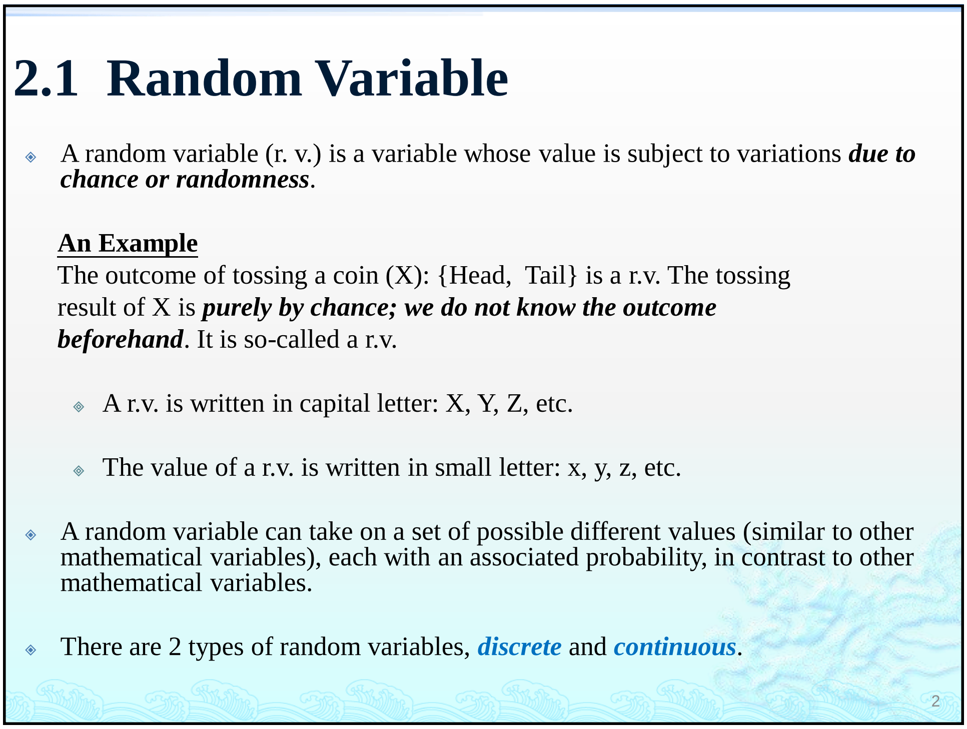


Chapter 2

Sampling Distribution of Means
&
Parameter Estimation

Tony CHAN, Ph.D.



2.1 Random Variable

- A random variable (r. v.) is a variable whose value is subject to variations ***due to chance or randomness***.

  <u>An Example</u>
  The outcome of tossing a coin (X): {Head, Tail} is a r.v. The tossing result of X is ***purely by chance; we do not know the outcome beforehand***. It is so-called a r.v.

  - A r.v. is written in capital letter: X, Y, Z, etc.
  - The value of a r.v. is written in small letter: x, y, z, etc.

- A random variable can take on a set of possible different values (similar to other mathematical variables), each with an associated probability, in contrast to other mathematical variables.

- There are 2 types of random variables, ***discrete*** and ***continuous***.

2.1.1 Discrete Random Variable

Definition 1.1

A ***discrete random variable***, X , is one which may take on only a countable number of distinct values such as 0,1,2,3,4,….

Examples

Number of children in a family

Attendance at a cinema

Number of patients in a doctor's surgery

Number of defective light bulbs in a box

2.1.1 Discrete Random Variable (Cont’d)

Let the discrete r.v. X takes on n discrete values: $x_1, x_2, \cdots, x_n$.

The probability mass function (pmf (for discrete r.v.)) is given by

$$P(x) = \begin{cases} P(X = x_i), & i = 1, 2, \cdots, n; \\ \quad 0, & otherwise, \end{cases}$$

where $P(X = x_i)$ denotes the probability when the r.v. X takes on the i-th data value x_i.

2.1.2 Continuous Random Variable

Definition 2.2

A *continuous random variable* , X , is one which takes an infinite number of possible values. Continuous random variables are usually measurements.

Examples

Height

Weight

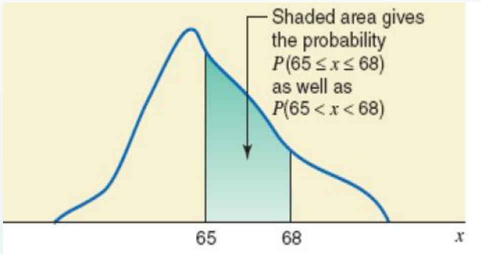
Amount of sugar in an orange

Time required to run a mile

2.1.2 Continuous Random Variable (Cont'd)

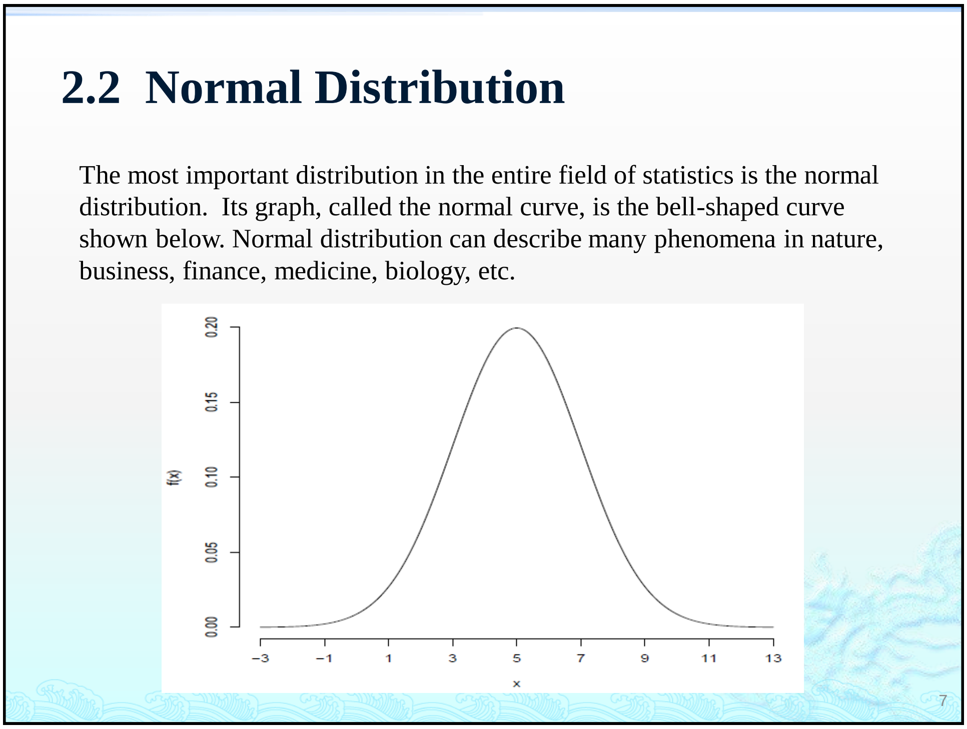
If a continuous r.v. X satisfies the following conditions:

- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1$
- $\int_{a}^{b} f(x) dx = P(a < x < b)$



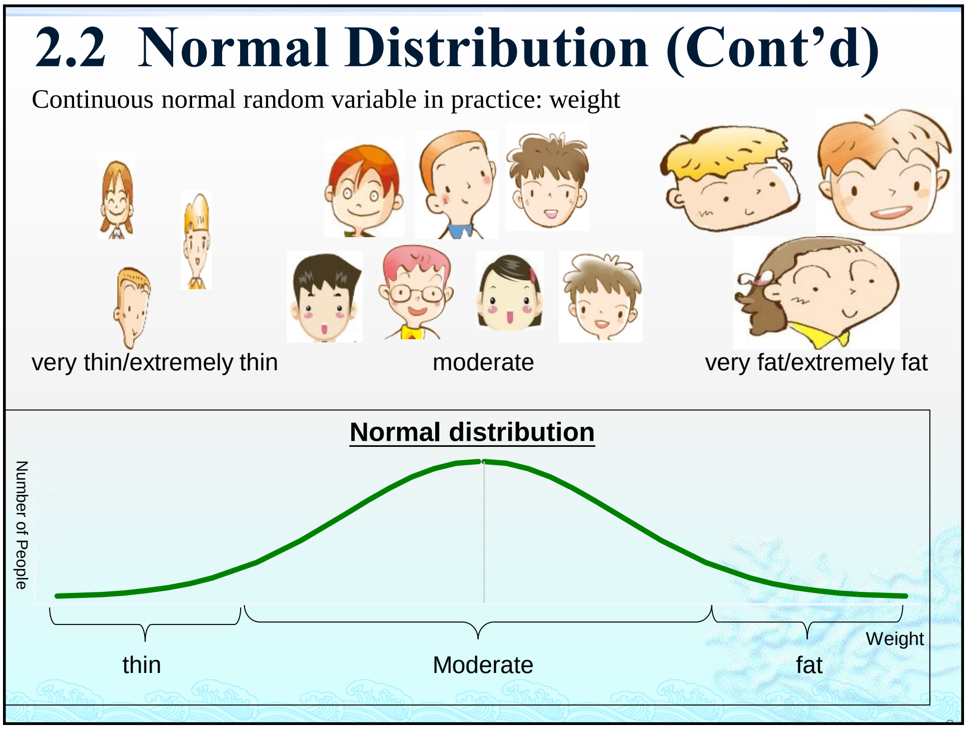
Then $f(x)$ is called the probability density function (pdf (for continuous r.v.)) of X

The notation $P(a < x < b)$ represents the probability (area) that X lies within (a, b).



2.2 Normal Distribution

The most important distribution in the entire field of statistics is the normal distribution. Its graph, called the normal curve, is the bell-shaped curve shown below. Normal distribution can describe many phenomena in nature, business, finance, medicine, biology, etc.



2.2 Normal Distribution (Cont'd)

Continuous normal random variable in practice: weight

2.2 Normal Distribution (Cont'd)

Importance of Normal Distribution

- Random variables occurring in practice often satisfy well a normal distribution.
- Large-sample statistics often turn out to be approximately normally distributed. This is a consequence of the central limit theorem.
- Most hypothesis testing that we're going to perform requires normality in some sense.

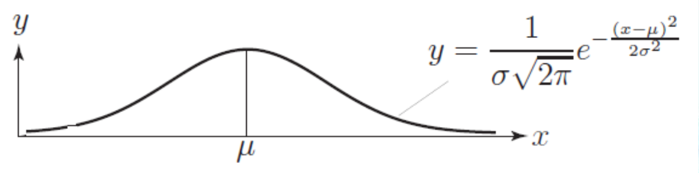
2.2 Normal Distribution (Cont'd)

The probability density function of normal distribution is given by

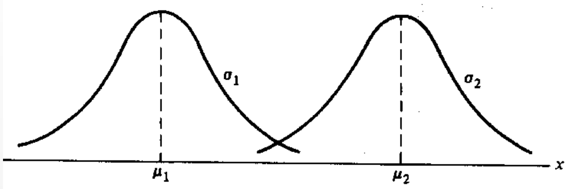
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty<x<+\infty, \text{ where } \pi=3.1416, e=2.71828\cdots.$$

Once μ and σ^2 are specified, the normal curve is completely determined. The 2 population parameters for normal distribution are given by

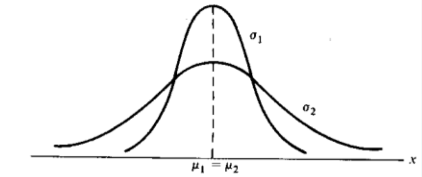
$$\begin{cases}\mu = \text{population mean}\\ \sigma = \text{population standard deviation}\end{cases}$$



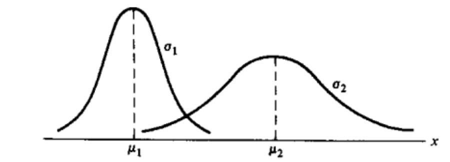
2.2 Normal Distribution (Cont’d)



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

2.2 Normal Distribution (Cont’d)

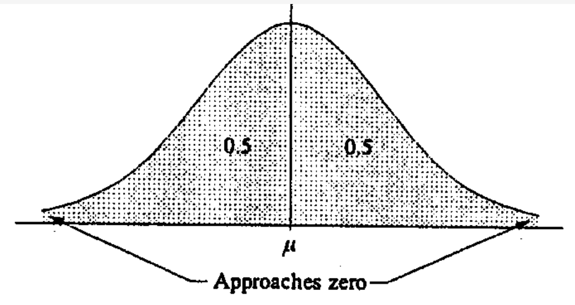
Characteristics of the Normal Curve:

- The mode (maximum) occurs at $x = \mu$.
- The curve is symmetric about a vertical axis through the mean μ. Hence, area on either side of mean is 0.5.
- The curve is unimodal.
- Mode = mean = median.
- As $x \rightarrow \pm\infty$, $f(x) \rightarrow \pm \infty$, i.e. x-axis is an asymptote.

2.2 Normal Distribution (Cont’d)

Features of the Normal Curve (Cont’d):

- The total area under the curve and above the horizontal axis is equal to 1.

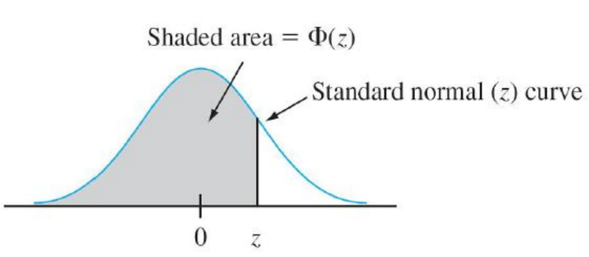


2.2 Standard Normal Distribution (Cont’d)

- We denote normal distribution as $N(\mu, \sigma^2)$, and standard normal distribution as $N(0,1)$.

- Comparing $N(\mu, \sigma^2)$ and $N(0,1)$, we have

$$\begin{cases} \mu = 0 \\ \sigma = 1 \end{cases}$$



for standard normal distribution $N(0,1)$.

2.2.1 Normal Random Variable

- If a random variable X follows a ***normal distribution***, we use the following symbol:

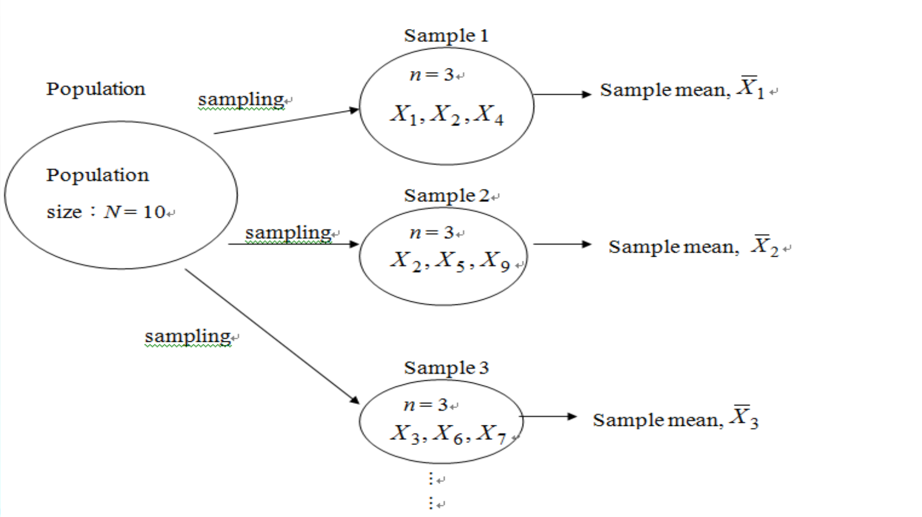
$$X \sim N(\mu, \sigma^2) \qquad \begin{cases} \text{pop mean} = \mu \\ \text{pop standard deviation} = \sigma \end{cases}$$

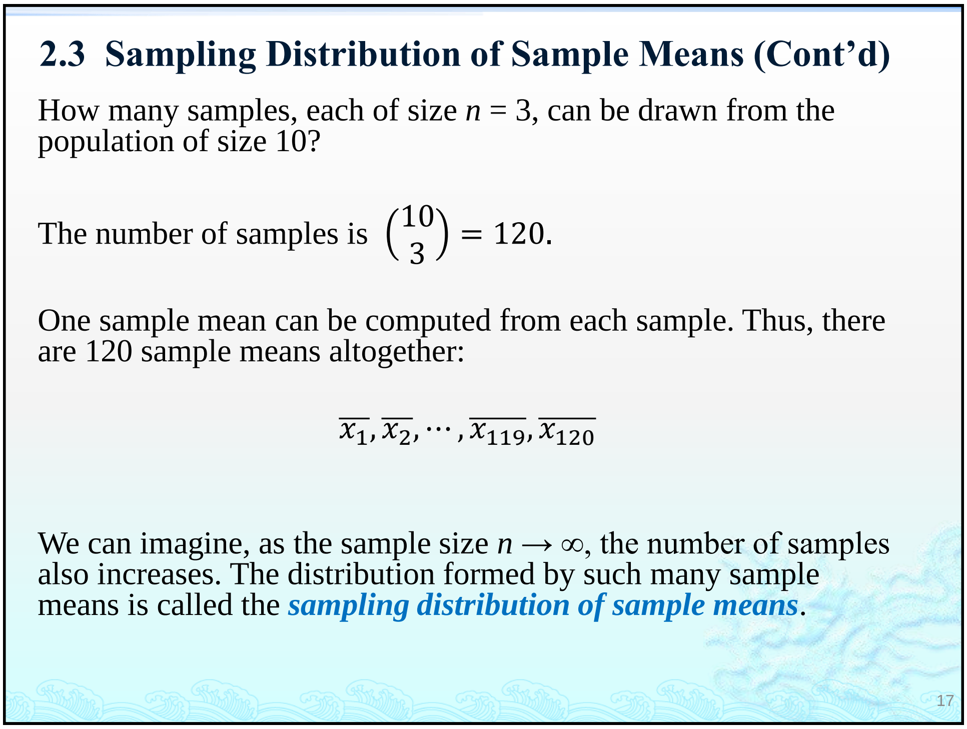
- If a random variable X follows a ***standard normal distribution***, we denote it as

$$X \sim N(0, 1) \qquad \begin{cases} \text{pop mean} = 0 \\ \text{pop standard deviation} = 1 \end{cases}$$

2.3 Sampling Distribution of Sample Means

Consider the following sampling problem:





2.3 Sampling Distribution of Sample Means (Cont'd)

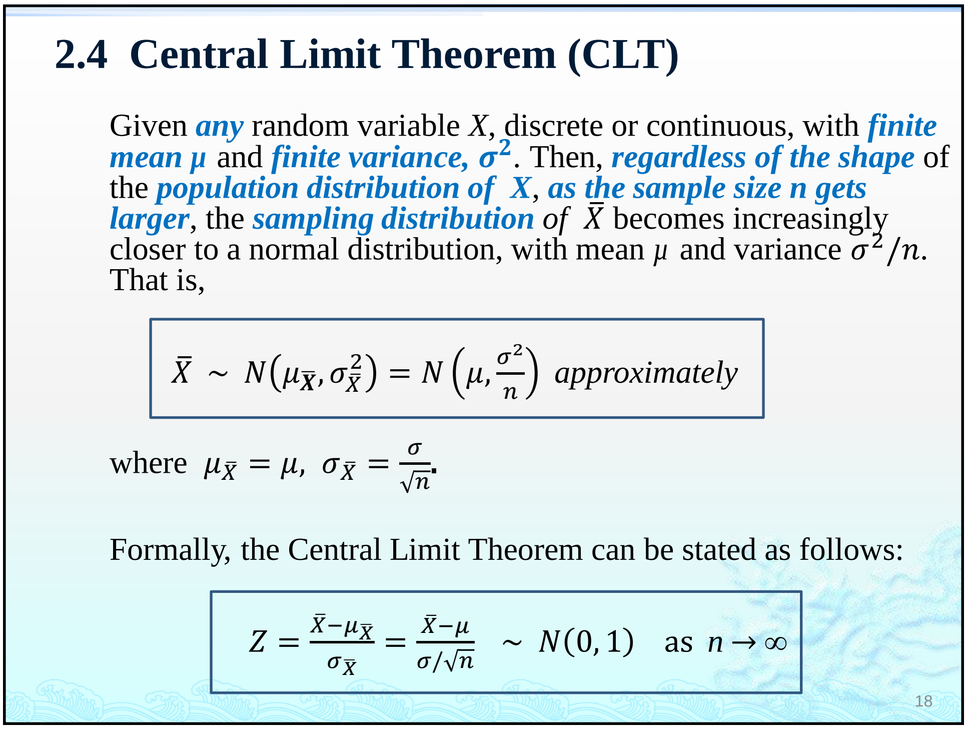
How many samples, each of size n = 3, can be drawn from the population of size 10?

The number of samples is $\binom{10}{3} = 120$.

One sample mean can be computed from each sample. Thus, there are 120 sample means altogether:

$$\overline{x_1}, \overline{x_2}, \cdots, \overline{x_{119}}, \overline{x_{120}}$$

We can imagine, as the sample size $n \rightarrow \infty$, the number of samples also increases. The distribution formed by such many sample means is called the ***sampling distribution of sample means***.



2.4 Central Limit Theorem (CLT)

Given ***any*** random variable X, discrete or continuous, with ***finite mean*** μ and ***finite variance,*** σ^2. Then, ***regardless of the shape*** of the ***population distribution of*** X, ***as the sample size n gets larger***, the ***sampling distribution*** *of* $\bar{X}$ becomes increasingly closer to a normal distribution, with mean μ and variance σ^2/n. That is,

$$\boxed{\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2) = N\left(\mu, \frac{\sigma^2}{n}\right) \text{ approximately}}$$

where $\mu_{\bar{X}} = \mu,\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Formally, the Central Limit Theorem can be stated as follows:

$$\boxed{Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{as } n \rightarrow \infty}$$

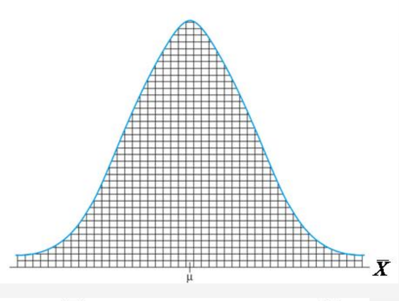
2.4 Central Limit Theorem (Cont’d)

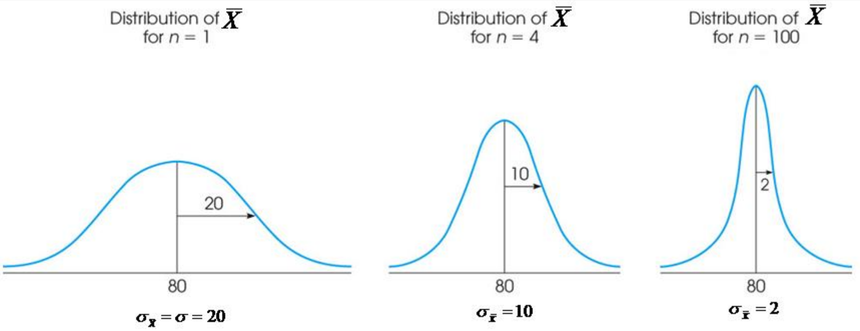
Summary of Central Limit Theorem:

- When a ***population is normally distributed***, the sampling distribution must be normally distributed no matter how small the sample size.
- When a ***population is non-normally distributed***, the sampling distribution will be ***approximately normally distributed*** when the sample size tends to infinity.
- In practice, when the ***sample size is at least 30*** (≥ 30), the sampling distribution of $\bar{X}$ is ***approximately normally distributed***.

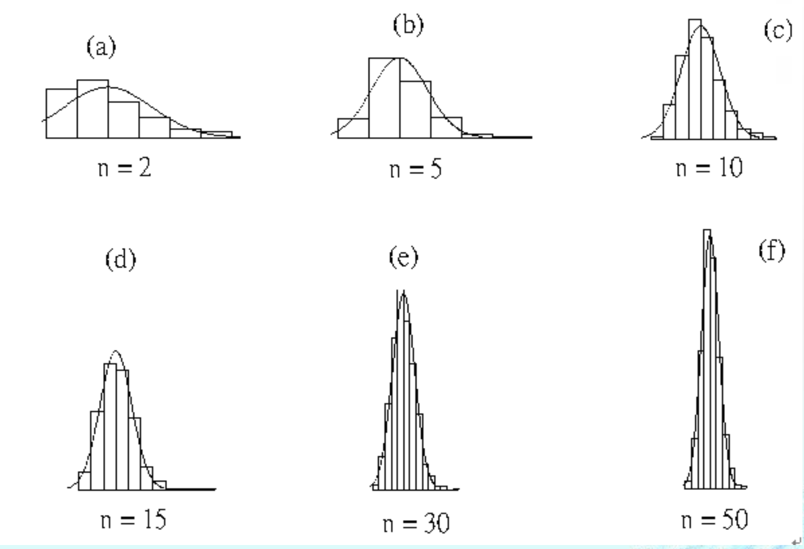
2.4 Central Limit Theorem (Cont’d)

As the sample size, n increases, the standard deviation of sample means, $\sigma_{\bar{X}}$ decreases.





2.4 Central Limit Theorem (Cont'd)



As the sample size increases, normality of the sampling distribution of sample means become more and more apparent.

2.4 Central Limit Theorem (Cont'd)

Notes:

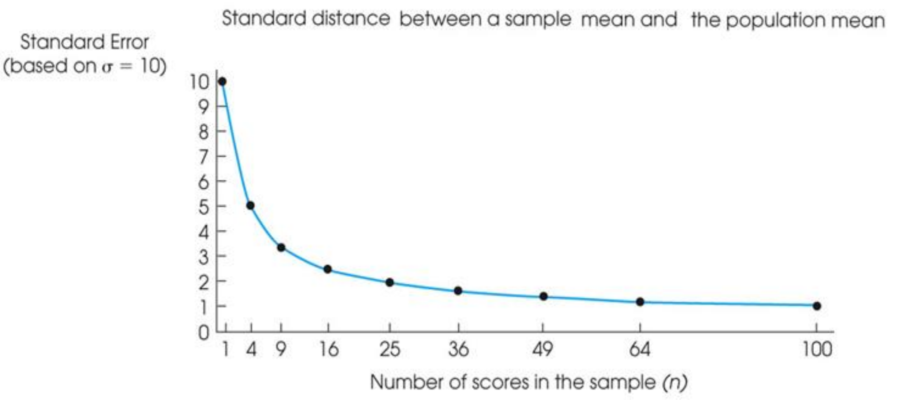
- $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is called the ***standard error of the mean***, and is denoted by s.e. , i.e.,

$$\text{s.e.} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

- If the sample size, *n* equals 1, then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{1}} = \sigma$$

2.4.1 Relationship between Sample Size and s.e.

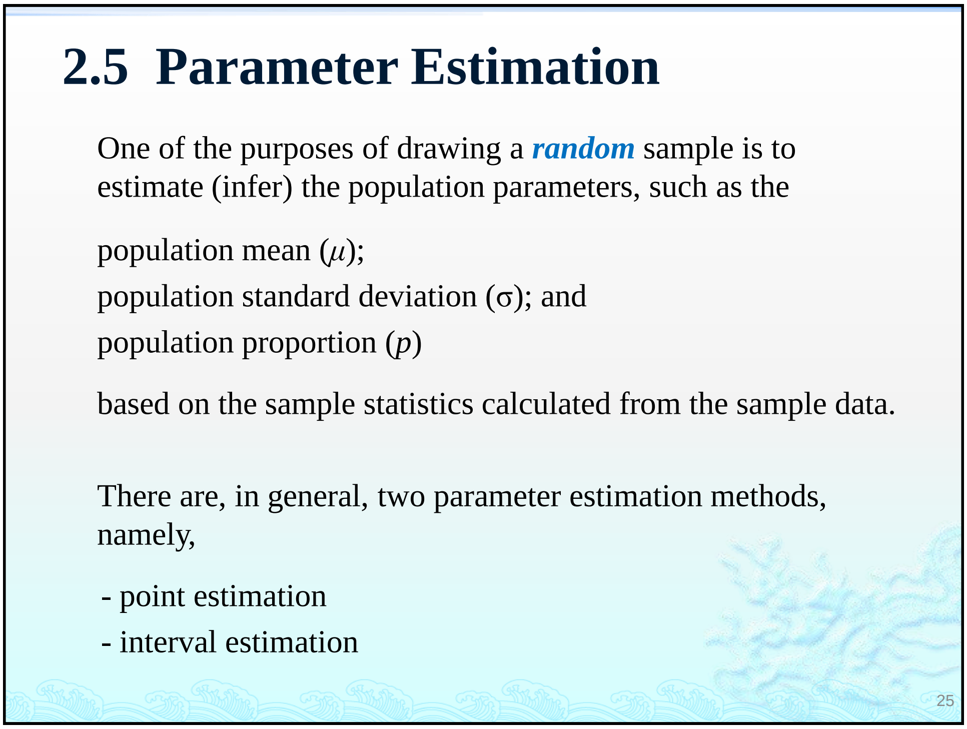


As the sample size increases, the s.e. decreases.

2.4.1 Relationship Between Sample Size and s.e. (Cont'd)

Sample Size (n)	*s.e.*	Sample Size (n)	*s.e.*
1	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{1}} = 10$	25	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$
4	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{4}} = 5$	49	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{49}} = 1.43$
9	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{9}} = 3.33$	64	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{64}} = 1.25$
16	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{16}} = 2.5$		

As the sample size increases, the s.e. decreases.



2.5 Parameter Estimation

One of the purposes of drawing a ***random*** sample is to estimate (infer) the population parameters, such as the

population mean (μ);

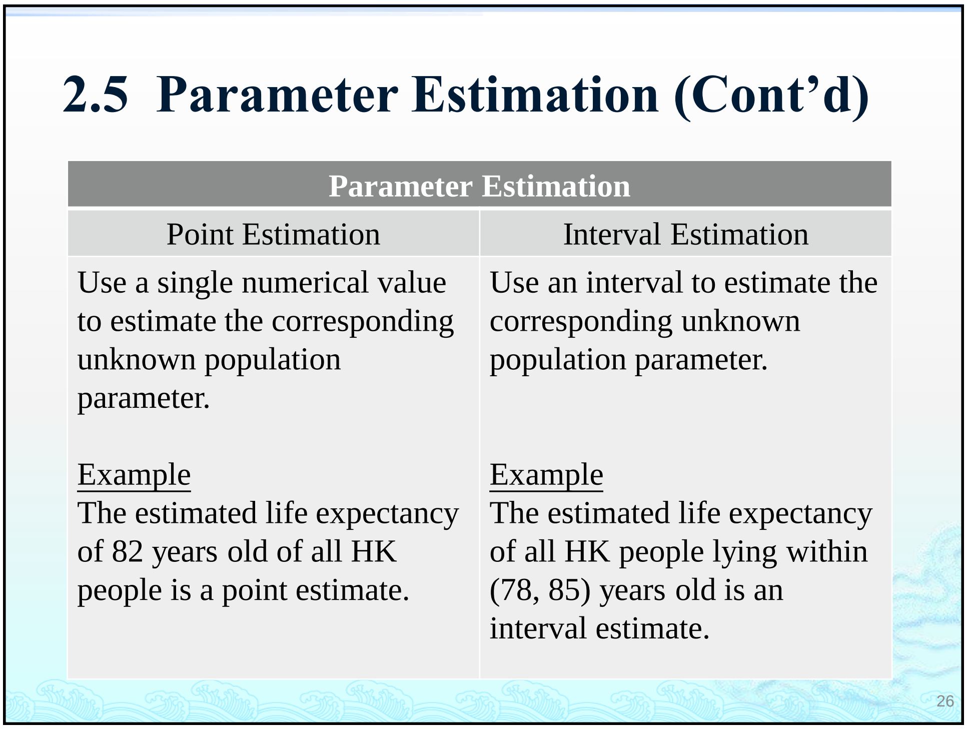
population standard deviation (σ); and

population proportion (p)

based on the sample statistics calculated from the sample data.

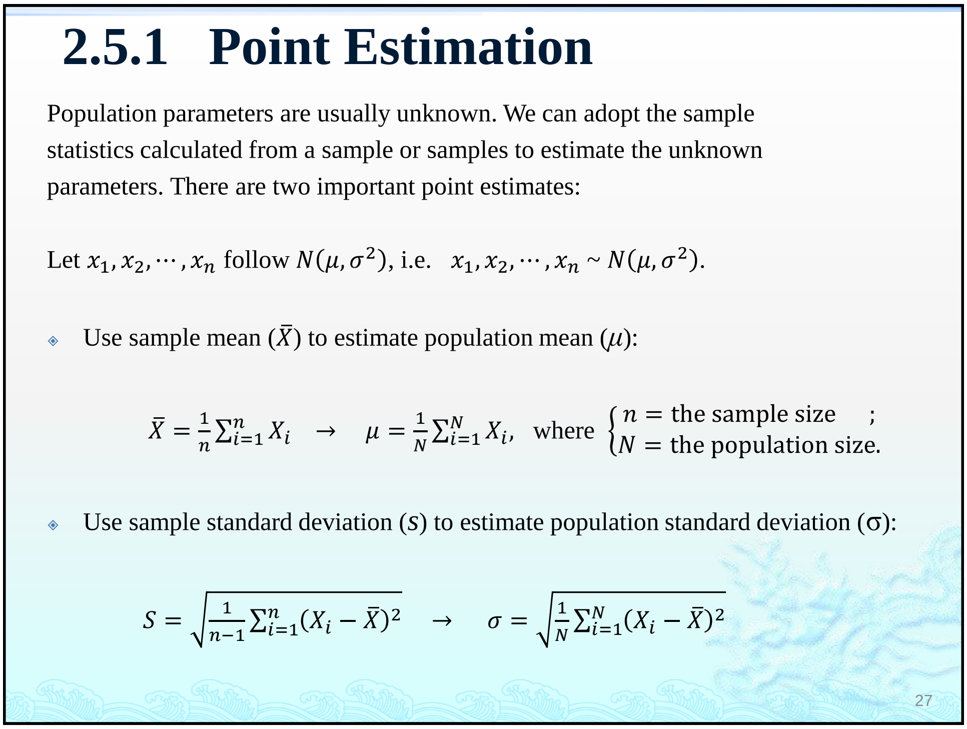
There are, in general, two parameter estimation methods, namely,

- point estimation
- interval estimation



2.5 Parameter Estimation (Cont'd)

Parameter Estimation	
Point Estimation	Interval Estimation
Use a single numerical value to estimate the corresponding unknown population parameter. Example The estimated life expectancy of 82 years old of all HK people is a point estimate.	Use an interval to estimate the corresponding unknown population parameter. Example The estimated life expectancy of all HK people lying within (78, 85) years old is an interval estimate.



2.5.1 Point Estimation

Population parameters are usually unknown. We can adopt the sample statistics calculated from a sample or samples to estimate the unknown parameters. There are two important point estimates:

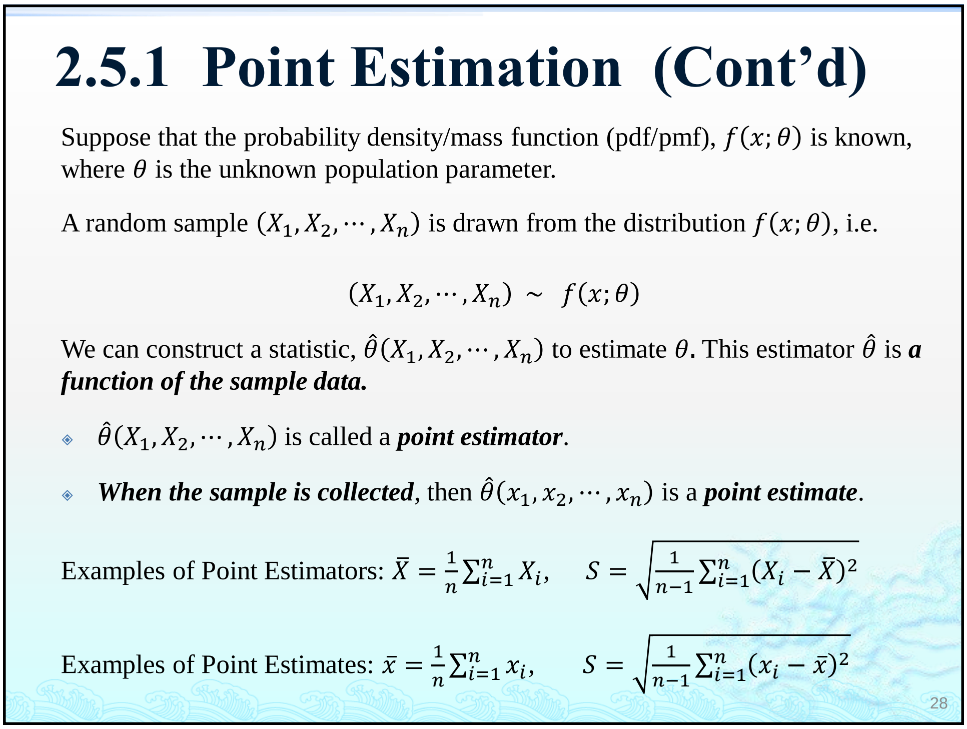
Let $x_1, x_2, \cdots, x_n$ follow $N(\mu, \sigma^2)$, i.e. $x_1, x_2, \cdots, x_n \sim N(\mu, \sigma^2)$.

- Use sample mean ($\bar{X}$) to estimate population mean (μ):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \rightarrow \quad \mu = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \text{where} \begin{cases} n = \text{the sample size} & ; \\ N = \text{the population size.} \end{cases}$$

- Use sample standard deviation (S) to estimate population standard deviation (σ):

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \rightarrow \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$



2.5.1 Point Estimation (Cont'd)

Suppose that the probability density/mass function (pdf/pmf), $f(x;\theta)$ is known, where θ is the unknown population parameter.

A random sample $(X_1, X_2, \cdots, X_n)$ is drawn from the distribution $f(x;\theta)$, i.e.

$$(X_1, X_2, \cdots, X_n) \sim f(x;\theta)$$

We can construct a statistic, $\hat{\theta}(X_1, X_2, \cdots, X_n)$ to estimate θ. This estimator $\hat{\theta}$ is ***a function of the sample data.***

- $\hat{\theta}(X_1, X_2, \cdots, X_n)$ is called a ***point estimator***.
- ***When the sample is collected***, then $\hat{\theta}(x_1, x_2, \cdots, x_n)$ is a ***point estimate***.

Examples of Point Estimators: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Examples of Point Estimates: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

2.5.1 Point Estimation (Cont'd)

Example 2.1

The pocket money per week for 30 students are as follows:

50, 60, 80, 100, 100, 200, 250, 260, 280, 280,
290, 290, 300, 300, 300, 300, 320, 340, 350, 350,
350, 360, 360, 360, 380, 380, 400, 400, 500, 550

Find the point estimates of the population mean pocket money (μ) and population standard deviation of pocket money (σ).

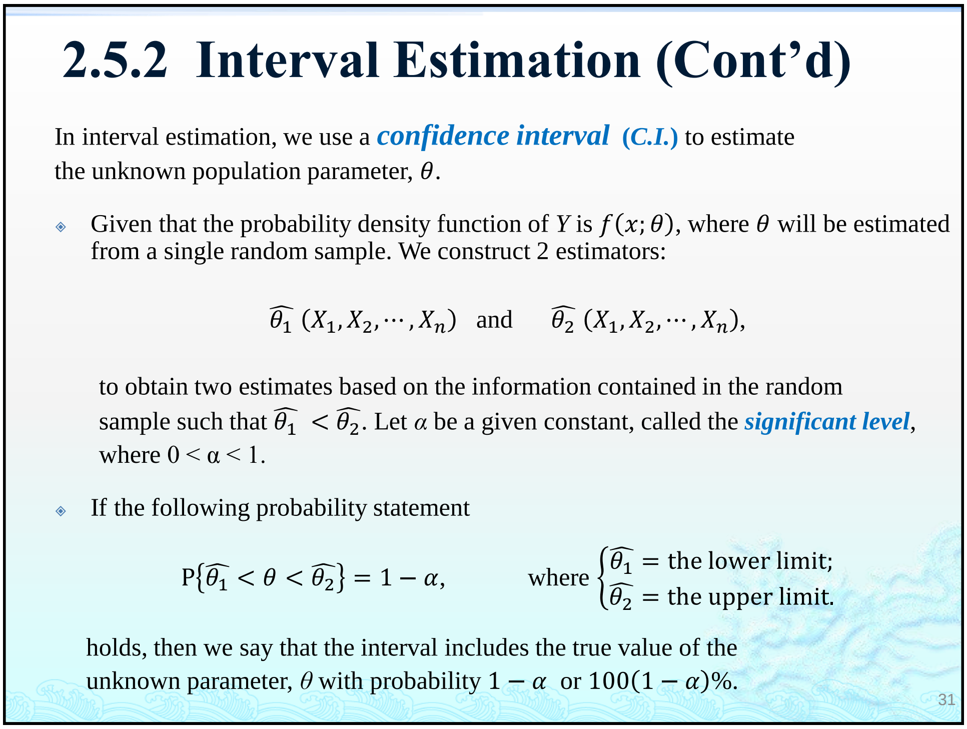
$$\begin{cases} \bar{x} = \hat{\mu} = \$294.67 \\ s = \hat{\sigma} = \$120.11 \end{cases}$$

The symbol $\hat{\mu}$ represents the estimate of μ.

2.5.2 Interval Estimation

Is a point estimate accurate without error?

- In practice, any population parameter θ is usually unknown. The error (e) induced is given by $e = \bar{x} - \theta$.
  The size of the error, e cannot be determined as θ is unknown.
- A single point estimate may not be close to the true value of the parameter, θ.
- It is therefore difficult to evaluate the closeness of the estimate to the true value of the parameter solely relying on point estimation.
- In other words, we can hardly evaluate the reliability of a point estimate through the use of point estimation approach.
- One way to solve such a problem is to employ interval estimation approach.



2.5.2 Interval Estimation (Cont'd)

In interval estimation, we use a ***confidence interval* (C.I.)** to estimate the unknown population parameter, θ.

- Given that the probability density function of Y is $f(x;\theta)$, where θ will be estimated from a single random sample. We construct 2 estimators:

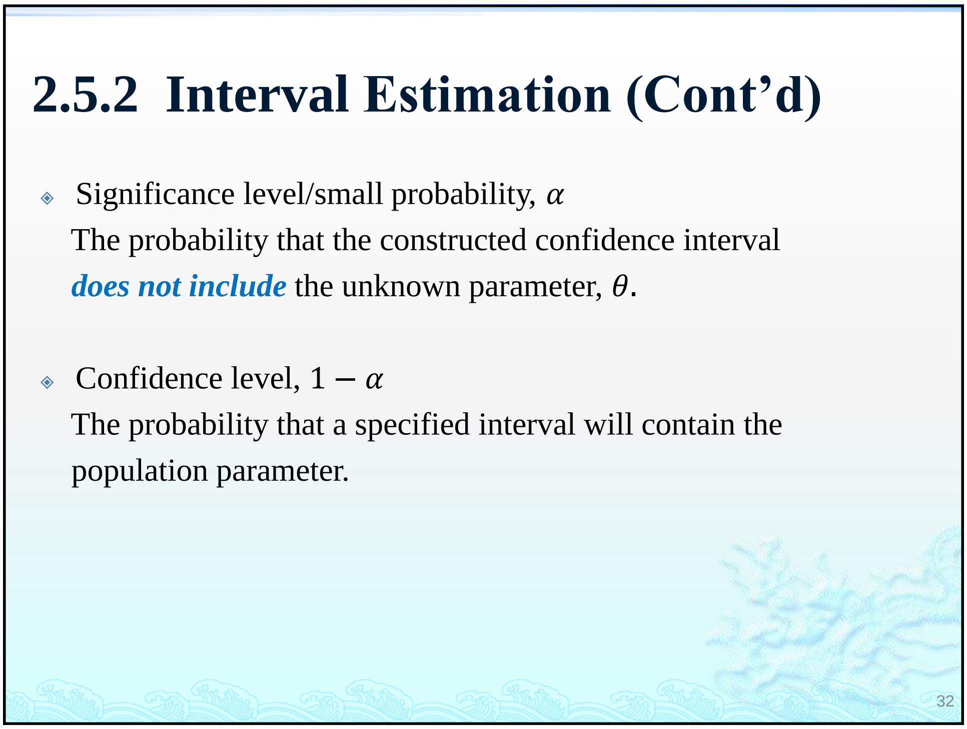
$$\widehat{\theta_1}\,(X_1, X_2, \cdots, X_n) \quad \text{and} \quad \widehat{\theta_2}\,(X_1, X_2, \cdots, X_n),$$

to obtain two estimates based on the information contained in the random sample such that $\widehat{\theta_1} < \widehat{\theta_2}$. Let α be a given constant, called the ***significant level***, where $0 < \alpha < 1$.

- If the following probability statement

$$\mathrm{P}\{\widehat{\theta_1} < \theta < \widehat{\theta_2}\} = 1 - \alpha, \qquad \text{where} \begin{cases} \widehat{\theta_1} = \text{the lower limit;} \\ \widehat{\theta_2} = \text{the upper limit.} \end{cases}$$

holds, then we say that the interval includes the true value of the unknown parameter, θ with probability $1-\alpha$ or $100(1-\alpha)\%$.



2.5.2 Interval Estimation (Cont'd)

- Significance level/small probability, α

  The probability that the constructed confidence interval ***does not include*** the unknown parameter, θ.

- Confidence level, $1-\alpha$

  The probability that a specified interval will contain the population parameter.

2.5.2 Interval Estimation (Cont'd)

Difference between Point Estimation & Interval Estimation:

- Interval estimation does not tell us directly about the actual value of the unknown parameter, θ.
- Rather, it tells us how large a probability to guarantee that a certain interval can include the true value of θ.

2.5.2 Interval Estimation (Cont'd)

A population parameter is a fixed but unknown parameter, i.e. a constant. Thus, for a particular <u>numerical</u> confidence interval, there are only 2 situations:

- The interval can contain the unknown parameter (probability = 1)
- The interval cannot contain the unknown parameter (probability = 0)

Thus, we cannot interpret a confidence interval using probabilistic point of view. Instead, we employ "confidence" for interpretation.

2.5.2 Interval Estimation (Cont'd)

Confidence vs. Probability

- ***BEFORE*** a sample is collected, probability can be used to describe the chance that the population mean will fall into a random C. I.

- ***AFTER*** the sample is collected, the population mean either fell within the constructed C. I. (probability = 1) or it did not (probability = 0). ***After the event, it does not make any sense to talk about probability.***

Analogy

Boys Scout Association sold 10,000 lottery tickets. Suppose that you own 9500 tickets. The drawing was held one hour ago, but you don't know the result as you are busy.

P(win) = 0 or 1, but you are very ***CONFIDENT that you have won the lottery.***

2.5.2 Interval Estimation (Cont'd)

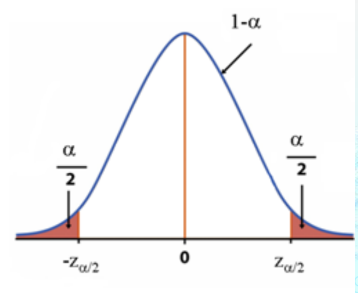
Confidence Interval Construction

For an interval with (1-α)100% confidence, its two end points on the left and right sides are $\pm z_{\alpha/2}$. The probability statement of the confidence interval for a ***normal*** population mean (μ) is given by

$$P\left[-z_{\alpha/2} \le z \le z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[-z_{\alpha/2} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right] = 1-\alpha \qquad \left(\because\ z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$$

$$P\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$



2.5.2 Interval Estimation (Cont’d)

(i) ***If the population SD (σ) is KNOWN***

The (1-α)100% confidence interval is then given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

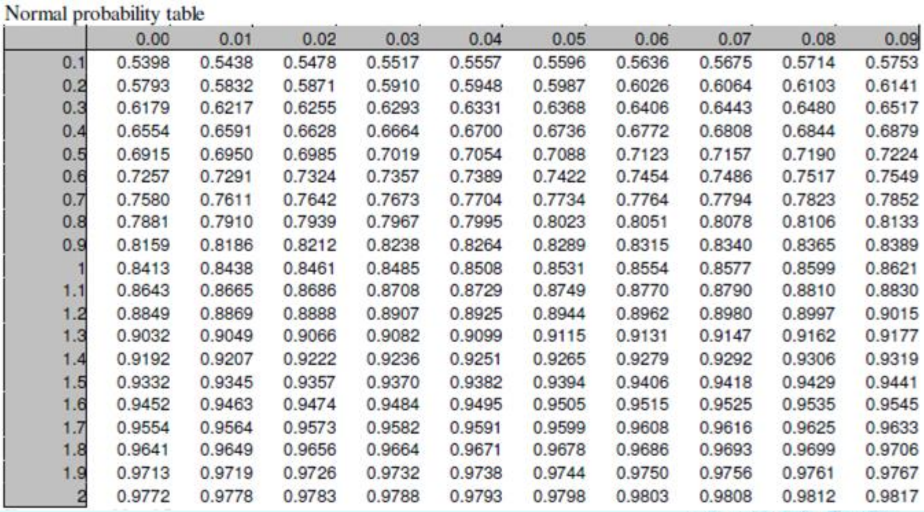
(ii) ***If σ is UNKNOWN but n ≥ 30***

The ***large-sample C.I.*** for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

where s is the sample SD.

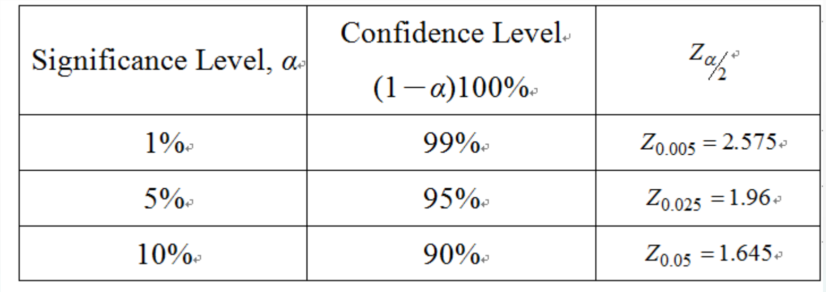
2.5.2 Interval Estimation (Cont’d)



Normal probability table

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

2.5.2 Interval Estimation (Cont’d)



Significance Level, α	Confidence Level $(1-\alpha)100\%$	$Z_{\alpha/2}$
1%	99%	$Z_{0.005} = 2.575$
5%	95%	$Z_{0.025} = 1.96$
10%	90%	$Z_{0.05} = 1.645$

2.5.2 Interval Estimation (Cont’d)

Example 2.2

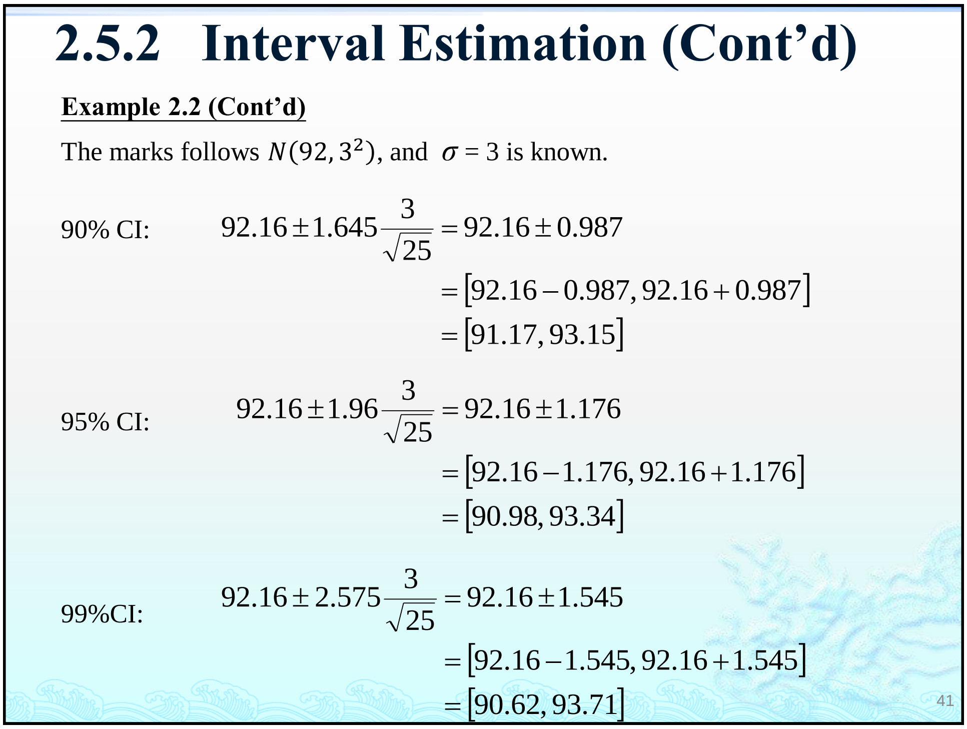
Based on past data, the marks of Statistical Analysis follows $N(92, 3^2)$. The marks of 25 students are displayed below:

87.8, 94.3, 92.2, 95.2, 89.0, 88.2, 87.9, 95.3, 92.1, 96.2,
92.0, 97.0, 93.6, 89.1, 93.9, 91.8, 97.6, 88.4, 90.1, 91.6,
90.7, 93.1, 96.5, 89.8, 90.5

Compute 90%, 95%, and 99% CI for the population mean (μ).

Solution

$$\hat{\mu} = \bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = 92.16$$



2.5.2 Interval Estimation (Cont’d)

Example 2.2 (Cont’d)

The marks follows $N(92, 3^2)$, and $\sigma = 3$ is known.

90% CI:

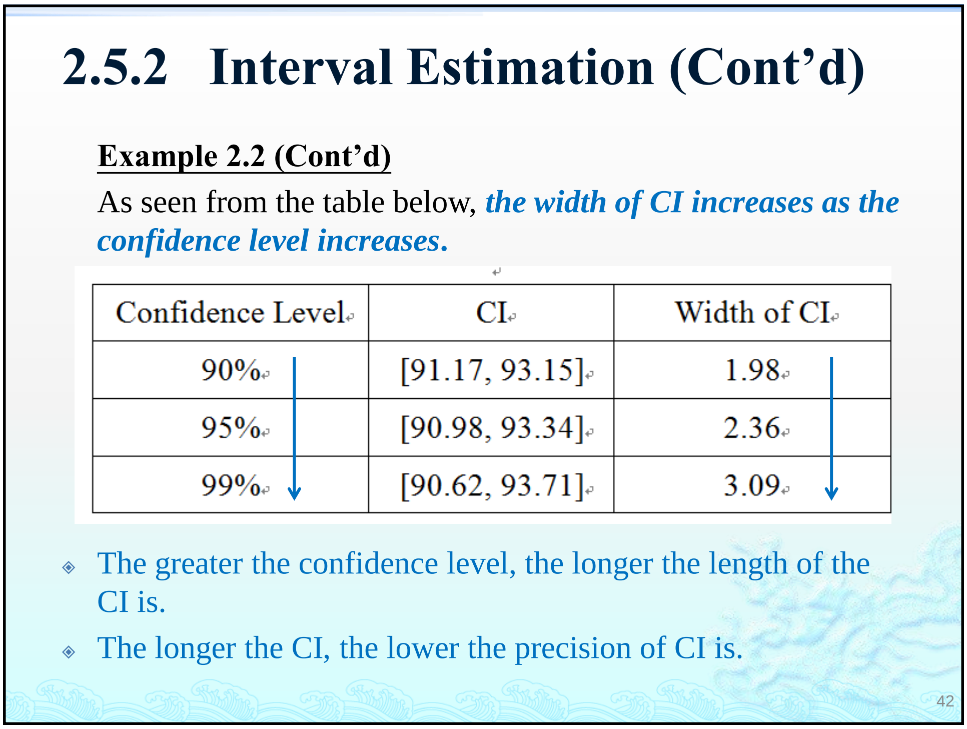
$$92.16 \pm 1.645\frac{3}{\sqrt{25}} = 92.16 \pm 0.987$$
$$= [92.16 - 0.987, 92.16 + 0.987]$$
$$= [91.17, 93.15]$$

95% CI:

$$92.16 \pm 1.96\frac{3}{\sqrt{25}} = 92.16 \pm 1.176$$
$$= [92.16 - 1.176, 92.16 + 1.176]$$
$$= [90.98, 93.34]$$

99%CI:

$$92.16 \pm 2.575\frac{3}{\sqrt{25}} = 92.16 \pm 1.545$$
$$= [92.16 - 1.545, 92.16 + 1.545]$$
$$= [90.62, 93.71]$$



2.5.2 Interval Estimation (Cont’d)

Example 2.2 (Cont’d)

As seen from the table below, ***the width of CI increases as the confidence level increases.***

Confidence Level	CI	Width of CI
90%	[91.17, 93.15]	1.98
95%	[90.98, 93.34]	2.36
99%	[90.62, 93.71]	3.09

- The greater the confidence level, the longer the length of the CI is.
- The longer the CI, the lower the precision of CI is.

2.5.2 Interval Estimation (Cont’d)

Example 2.3

The engineer of the Hong Kong Electricity Inc. wants to investigate the mean electricity consumption of the Tin Sum Village. He randomly sampled 100 households (HH), and got the following results:

sample mean electricity consumption per HH = 96 units
sample standard deviation of electricity consumption = 24 units

Construct a 95% CI for the population mean electricity consumption, μ.

Solution: σ is unknown; $\bar{X}$and s are given; n = 100 > 30. The large-sample CI can be used.

$$\begin{aligned}\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} &= 96 \pm 1.96\frac{24}{\sqrt{100}} \\ &= 96 \pm 4.70 \\ &= (96 - 4.70, 96 + 4.70) \\ &= (91.3, 100.70) \text{ units}\end{aligned}$$

2.5.2 Interval Estimation (Cont’d)

Meaning of a 95% confidence interval

- In conducting a sample survey, we can draw a random sample and then calculate a sample statistic (say, sample mean, $\bar{x}$) for the corresponding parameter (say, population mean, μ).
- Since the population parameter is unknown, therefore we cannot know whether or not the constructed CI can capture the unknown parameter. Based on the sample statistic, $\bar{x}$, a 95% confidence interval can be constructed.
- If, for example,1000 random samples are drawn, then we can construct 1000 confidence intervals from these samples. The meaning of a 95% confidence interval indicates:

of these 1000 intervals, about 95% (i.e., about 950) of them can capture the unknown parameter, whereas about 50 intervals cannot.

2.5.3 Student's t-distribution

- A t-distribution is similar to a standard normal distribution, $N(0,1)$, except that it has slightly fatter tails on both sides to reflect the uncertainty added by estimating σ.
- The larger the sample size is used to estimate σ, the closer t approaches Z will be, where Z is a $N(0, 1)$ variable.
- If $n > 100$, $t \approx Z$.

2.5.4 Student's t-distribution: Degree of Freedom

Selig (1994) stated that degrees of freedom (df) are ***lost for each parameter in a model that is estimated in the process of estimating another parameter***.

Examples

- 1 df is lost when we estimate the population mean, μ using the sample mean. Let $\{x_1, x_2, \cdots, x_n\}$ be random sample.

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ subject to the constraint } \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

- 2 df are lost when estimating the s. e. of parameter estimate in a regression model using $\hat{y}$:
  - 1 df lost for estimating the y-intercept;
  - 1 df lost for estimating the slope of the regression line.

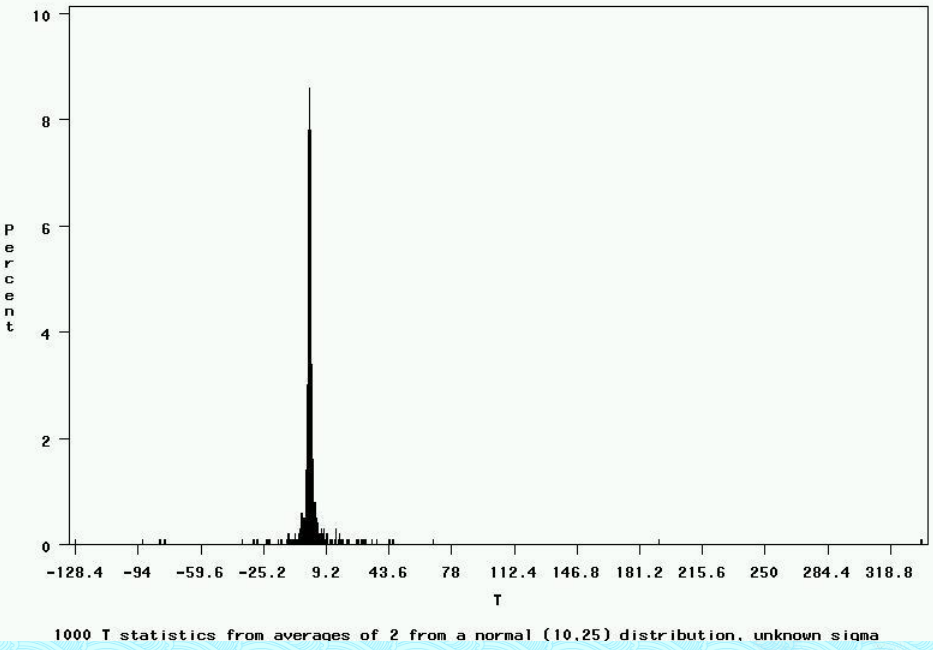
2.5.5 pdf of Student’s t-distribution

The probability density function of the Student’s t-distribution is given by

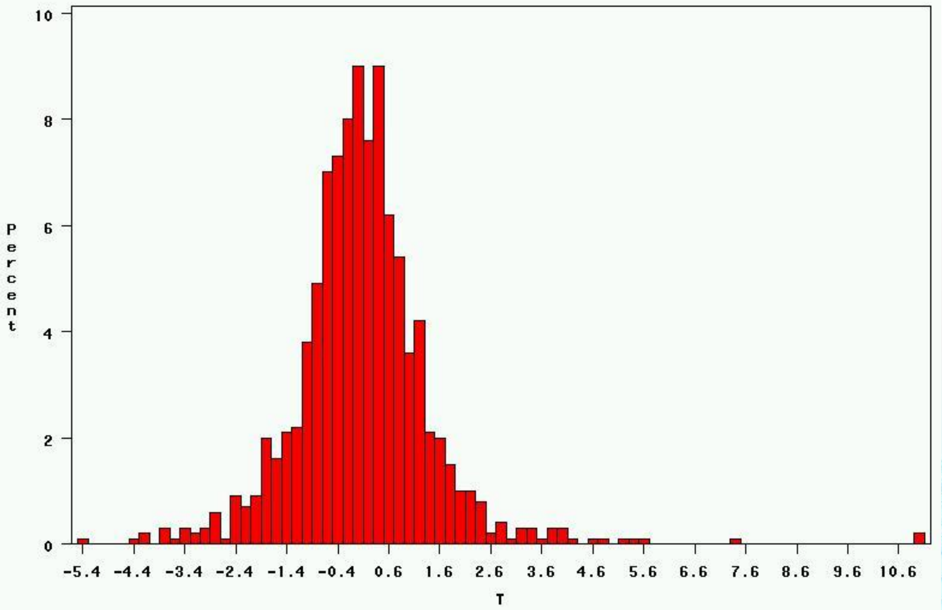
$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}, \quad -\infty < x < +\infty$$

where $\Gamma(s) = \int_0^{+\infty} x^{s-1}e^{-x}dx$ is the gamma function, and *v* is the degree of freedom of the t-distribution.

T-distribution with only 1 degree of freedom.

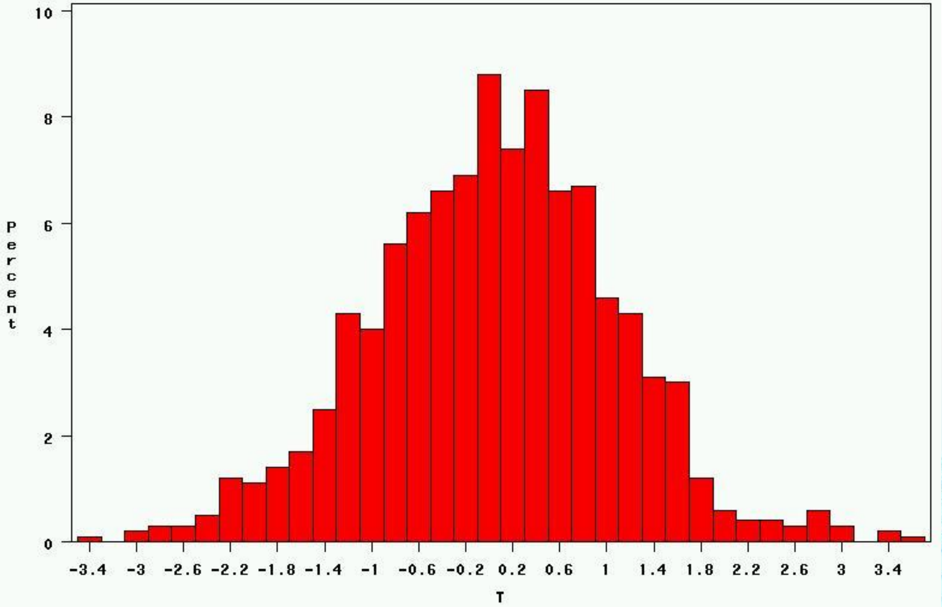


T-distribution with 4 degrees of freedom.

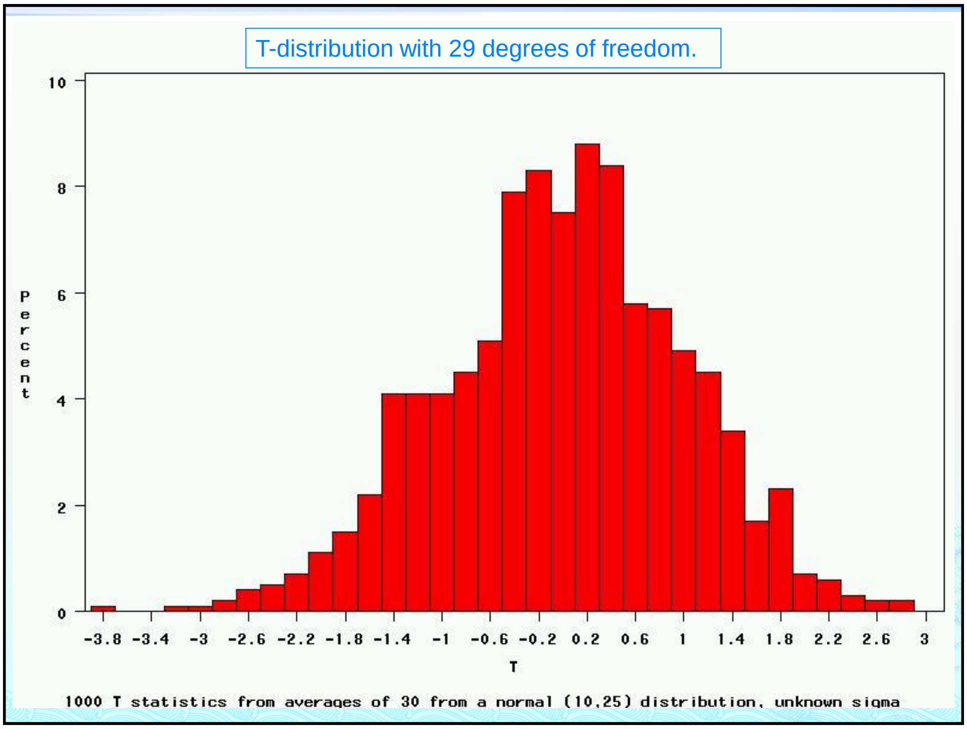


1000 T statistics from averages of 5 from a normal (10,25) distribution, unknown sigma

T-distribution with 9 degrees of freedom.

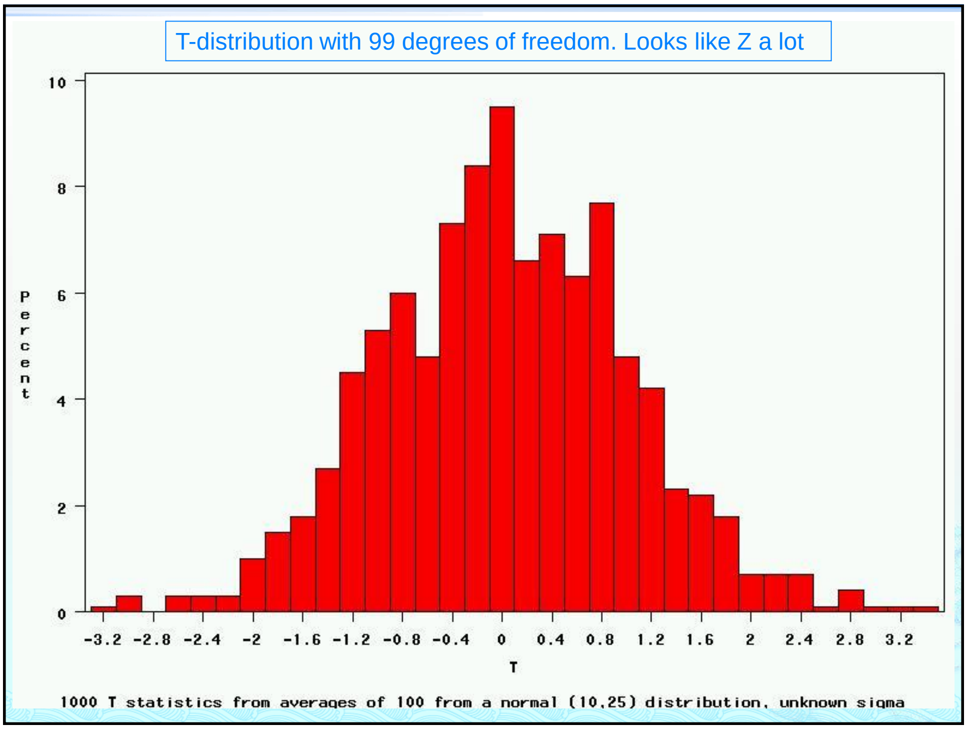


1000 T statistics from averages of 10 from a normal (10,25) distribution, unknown sigma



T-distribution with 29 degrees of freedom.

1000 T statistics from averages of 30 from a normal (10,25) distribution, unknown sigma



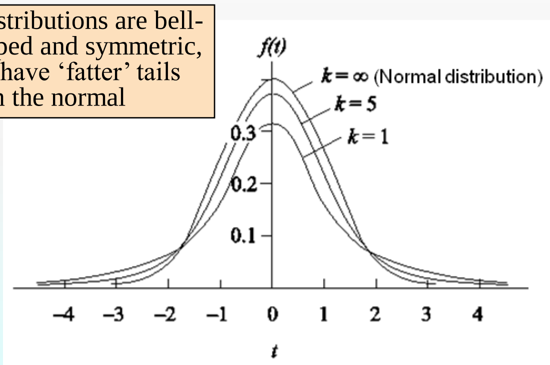
T-distribution with 99 degrees of freedom. Looks like Z a lot

1000 T statistics from averages of 100 from a normal (10,25) distribution, unknown sigma

2.5.6 Student’s t-distribution Curve

In practice, σ is usually unknown. ***Error must be induced when σ is replaced by s. The error will be larger for smaller sample sizes***. To overcome such difficulty, we use the ***Student’s t-distribution*** rather than the normal distribution.

t-distributions are bell-shaped and symmetric, but have ‘fatter’ tails than the normal

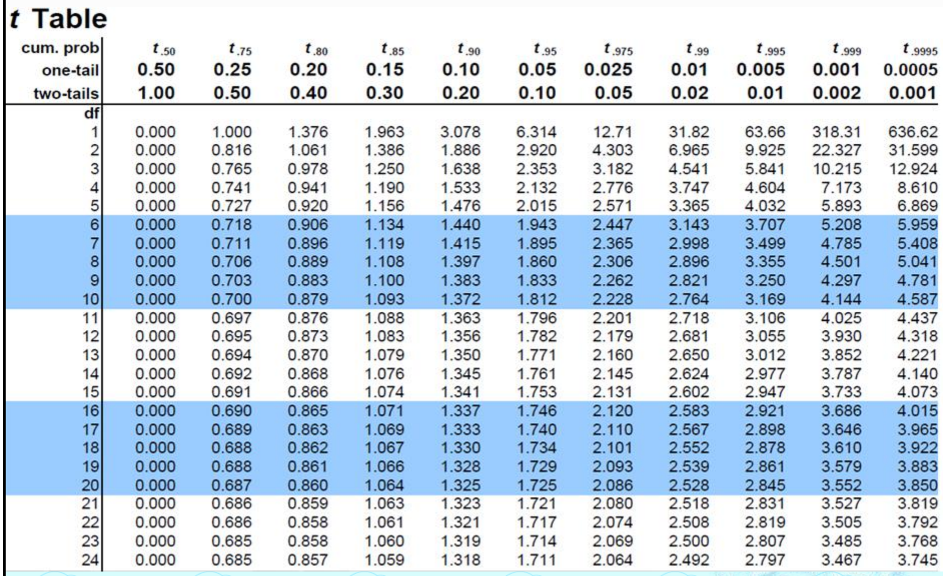


1. When $k \to \infty$, the asymptotic distribution of t-distribution is the standard normal distribution $N(0, 1)$.

2. If a random variable X follows a t-distribution with df equal to k, then we use the following symbol:

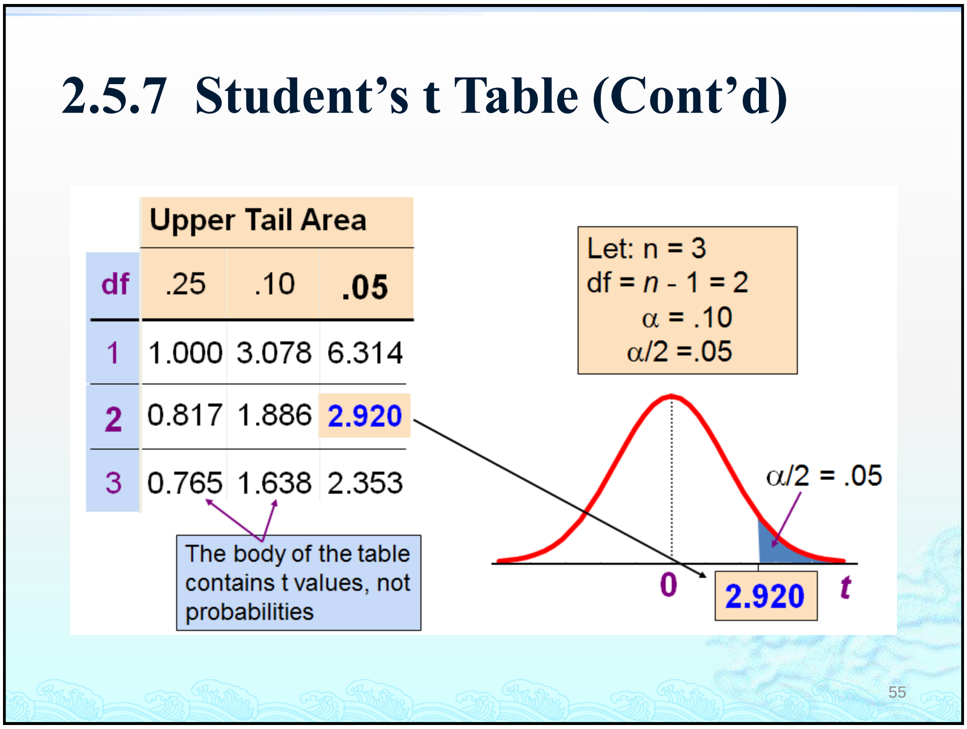
$$X \sim t(k)$$

2.5.7 Student’s t Table



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745

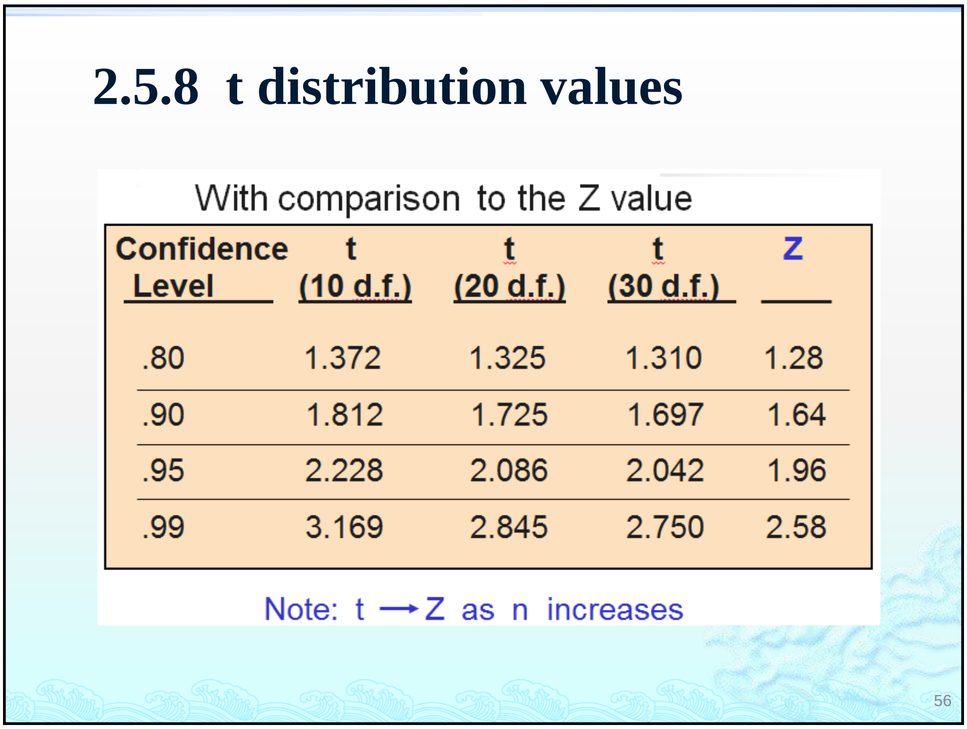


2.5.7 Student’s t Table (Cont’d)

	Upper Tail Area		
df	.25	.10	**.05**
1	1.000	3.078	6.314
2	0.817	1.886	**2.920**
3	0.765	1.638	2.353

The body of the table contains t values, not probabilities

Let: n = 3
df = *n* - 1 = 2
α = .10
α/2 =.05



2.5.8 t distribution values

With comparison to the Z value

Confidence Level	t (10 d.f.)	t (20 d.f.)	t (30 d.f.)	Z
.80	1.372	1.325	1.310	1.28
.90	1.812	1.725	1.697	1.64
.95	2.228	2.086	2.042	1.96
.99	3.169	2.845	2.750	2.58

Note: t → Z as n increases

2.5.9 Interval Estimation Using t-distribution

(iii) *If the population SD (σ) is UNKNOWN and n < 30*

In general, σ is unknown. We can use sample SD, s to replace the population SD, σ . That is,

$$\frac{s}{\sqrt{n}} \rightarrow \frac{\sigma}{\sqrt{n}}$$

The corresponding z-score transformation is changed accordingly:

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \rightarrow \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

2.5.9 Interval Estimation Using t-distribution (Cont'd)

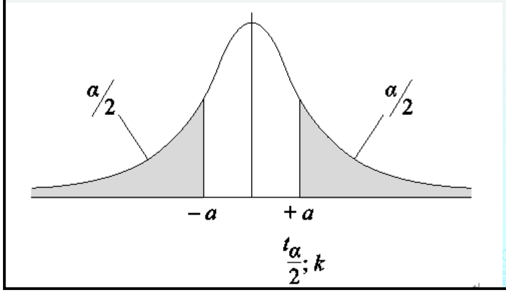
(iii) *If the population SD (σ) is UNKNOWN and n < 30, then the corresponding normal distribution is changed to the t-distribution.*

$$P\left[-t_{\alpha/2;k} \le z \le t_{\alpha/2;k}\right] = 1-\alpha,$$

$$P\left[-t_{\alpha/2;k} \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le t_{\alpha/2;k}\right] = 1-\alpha, \qquad \left(\because\ z = \frac{\bar{x}-\mu}{s/\sqrt{n}}\right)$$

$$P\left[\bar{x} - t_{\alpha/2;k}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2;k}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha,$$

k is the degrees of freedom
$k = n$ - 1



The (1-α)100% CI for μ is given by:

$$\bar{x} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}$$

$$= \boxed{\left[\bar{x} - t_{\alpha/2;n-1}\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2;n-1}\frac{s}{\sqrt{n}}\right]}$$

2.5.10 Interval Estimation Examples

Example 2.4

Ten packets of a particular brand of biscuits are chosen at random and their masses are noted. The results (in grams) are 397.3, 399.6, 401.0, 392.9, 396.8, 400.0, 397.6, 392.1, 400.8, 400.6. Assuming that the sample is taken from a normal population with mean mass μ. Construct a 99% CI for μ.

Solution:

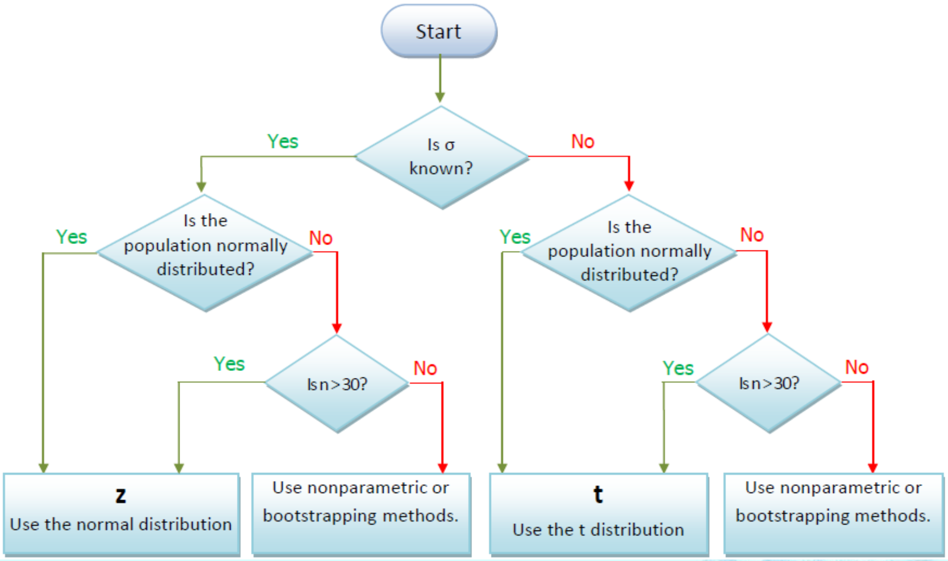
Let X be the mass of biscuits, and $X \sim N(\mu\ , \sigma^2)$. Since σ is not given (unknown) and n = 10 < 30, therefore the 99% CI for μ is given by

$$\bar{x} = 397.87, \quad s = 9.29$$

$$\begin{aligned}\bar{x} \pm t_{\alpha/2;n-1}\frac{s}{\sqrt{n}} &= 397.87 \pm 3.25\frac{\sqrt{9.29}}{\sqrt{10}} \quad (3.25 \text{ is from t - table})\\ &= 397.87 \pm 0.96\\ &= [397.87 - 0.96, 397.87 + 0.96]\\ &= [396.91, 398.83]\,\text{g}\end{aligned}$$

59

2.5.11 Interval Estimation - Summary



Start

Is σ known?

Yes

Is the population normally distributed?

Yes

No

Is n>30?

Yes

No

No

Is the population normally distributed?

Yes

No

Is n>30?

Yes

No

z

Use the normal distribution

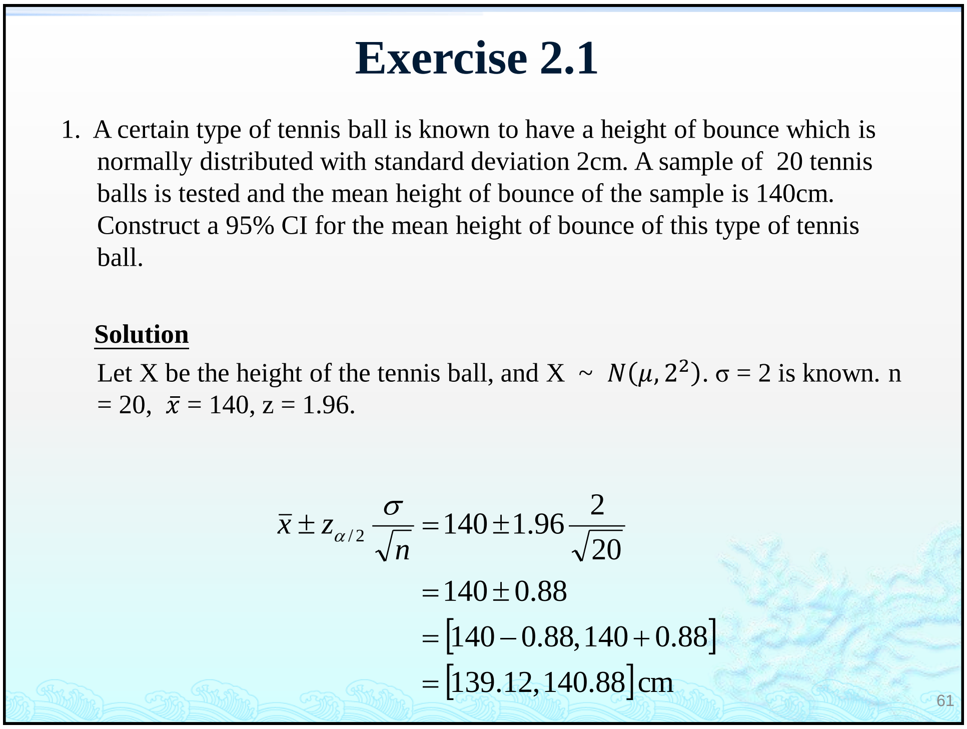
Use nonparametric or bootstrapping methods.

t

Use the t distribution

Use nonparametric or bootstrapping methods.

60



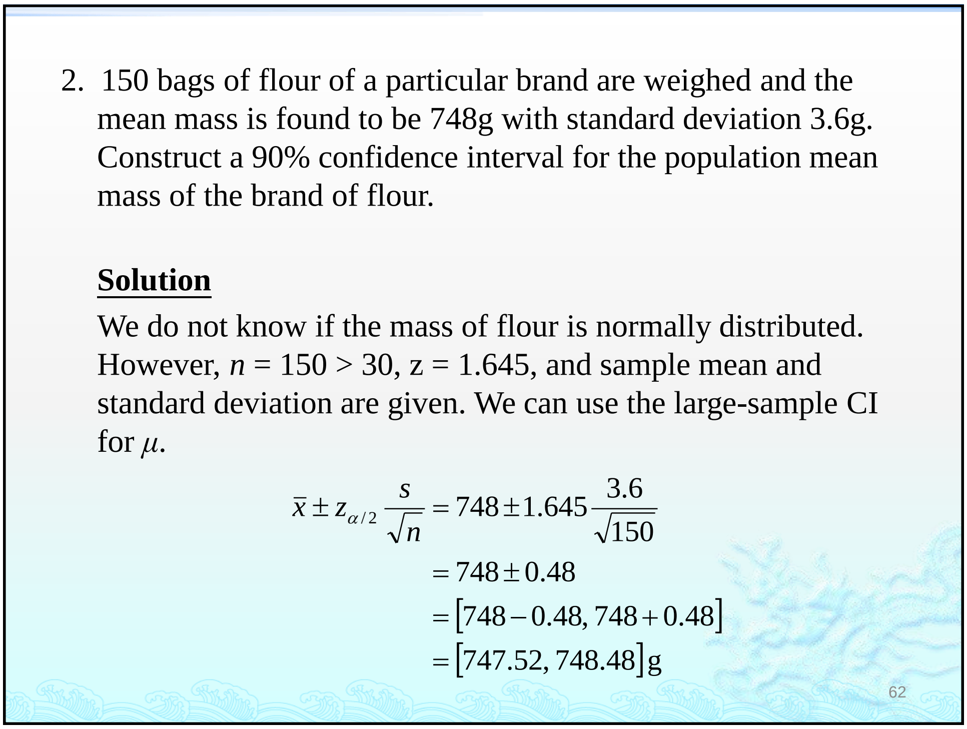
Exercise 2.1

1. A certain type of tennis ball is known to have a height of bounce which is normally distributed with standard deviation 2cm. A sample of 20 tennis balls is tested and the mean height of bounce of the sample is 140cm. Construct a 95% CI for the mean height of bounce of this type of tennis ball.

Solution

Let X be the height of the tennis ball, and X ~ $N(\mu, 2^2)$. σ = 2 is known. n = 20, $\bar{x}$ = 140, z = 1.96.

$$\begin{aligned}\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} &= 140 \pm 1.96\frac{2}{\sqrt{20}} \\ &= 140 \pm 0.88 \\ &= [140 - 0.88, 140 + 0.88] \\ &= [139.12, 140.88]\,\text{cm}\end{aligned}$$



2. 150 bags of flour of a particular brand are weighed and the mean mass is found to be 748g with standard deviation 3.6g. Construct a 90% confidence interval for the population mean mass of the brand of flour.

Solution

We do not know if the mass of flour is normally distributed. However, *n* = 150 > 30, z = 1.645, and sample mean and standard deviation are given. We can use the large-sample CI for μ.

$$\begin{aligned}\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} &= 748 \pm 1.645\frac{3.6}{\sqrt{150}} \\ &= 748 \pm 0.48 \\ &= [748 - 0.48, 748 + 0.48] \\ &= [747.52, 748.48]\,\text{g}\end{aligned}$$

3. The heights (measured in cm) of six policemen were as follows:

180, 176, 179, 181, 183, 179

Assume that the heights are normally distributed, construct a 99% confidence interval for the population mean height of all policemen.

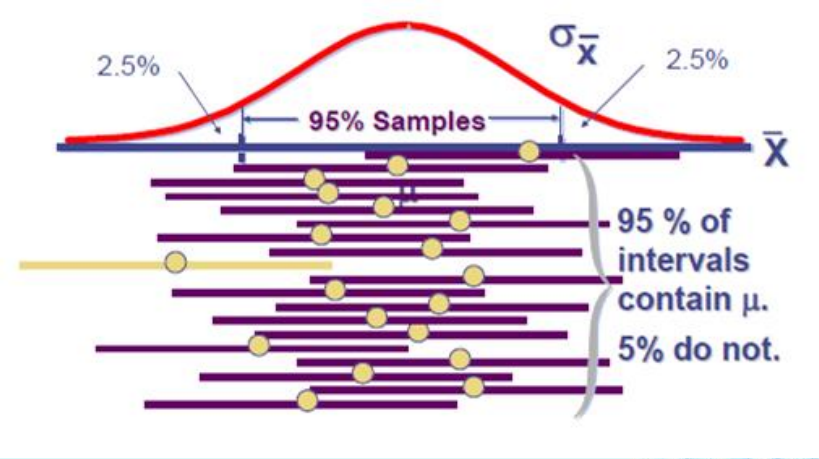
Solution

Since the data are normally distributed, but n = 6 < 30, t-distribution should be adopted. The sample mean and standard deviation are:

$$\bar{x} = 179.67,\ s = 2.34$$

$$\begin{aligned}\bar{x} \pm t_{\alpha/2;n-1}\frac{s}{\sqrt{n}} &= 179.67 \pm 4.03\frac{2.34}{\sqrt{6}} \quad (4.03\text{ is from t-table})\\ &= 179.67 \pm 3.85\\ &= [179.67 - 3.85, 179.67 + 3.85]\\ &= [175.82, 183.52]\text{cm}\end{aligned}$$

2.6 Sample Size Determination for the Mean



2.6 Sample Size Determination for the Mean (Cont'd)

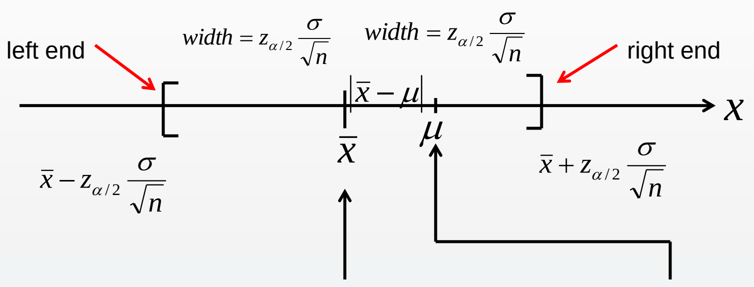
For normally distributed population, the (1-α)100% CI for μ is given by

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad (\sigma \text{ is known})$$

Factors Affecting Interval Width

1. Data dispersion – measured by σ or s
2. Sample size – affects standard error, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ or s
3. Confidence level, (1- α) – affects $z_{\alpha/2}$ or $t_{\alpha/2;n-1}$

2.6 Sample Size Determination for the Mean (Cont'd)



left end

$width = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

$width = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

right end

$|\bar{x} - \mu|$

x

$\bar{x}$

μ

$\bar{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

$\bar{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

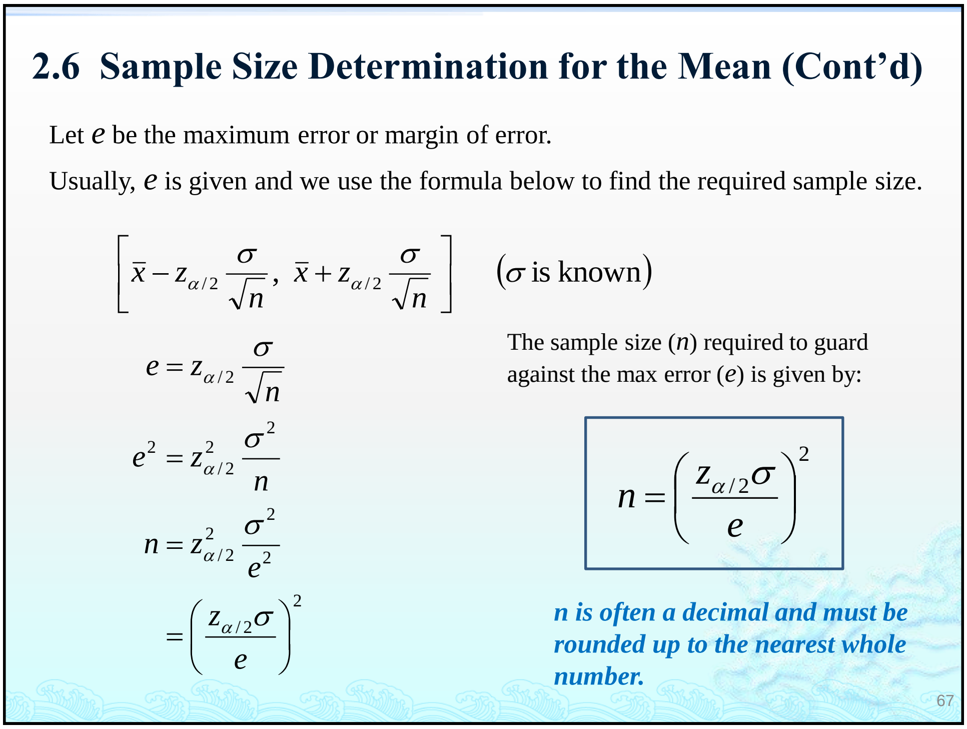
If μ lies at either end of the CI, then the error induced is the maximum error:

Max. error = margin of error

If μ lies at the same position of the sample mean, then no error is induced.

If μ lies here, then the error induced is

$$|\bar{x} - \mu|$$



2.6 Sample Size Determination for the Mean (Cont'd)

Let e be the maximum error or margin of error.

Usually, e is given and we use the formula below to find the required sample size.

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad (\sigma \text{ is known})$$

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$e^2 = z_{\alpha/2}^2\frac{\sigma^2}{n}$$

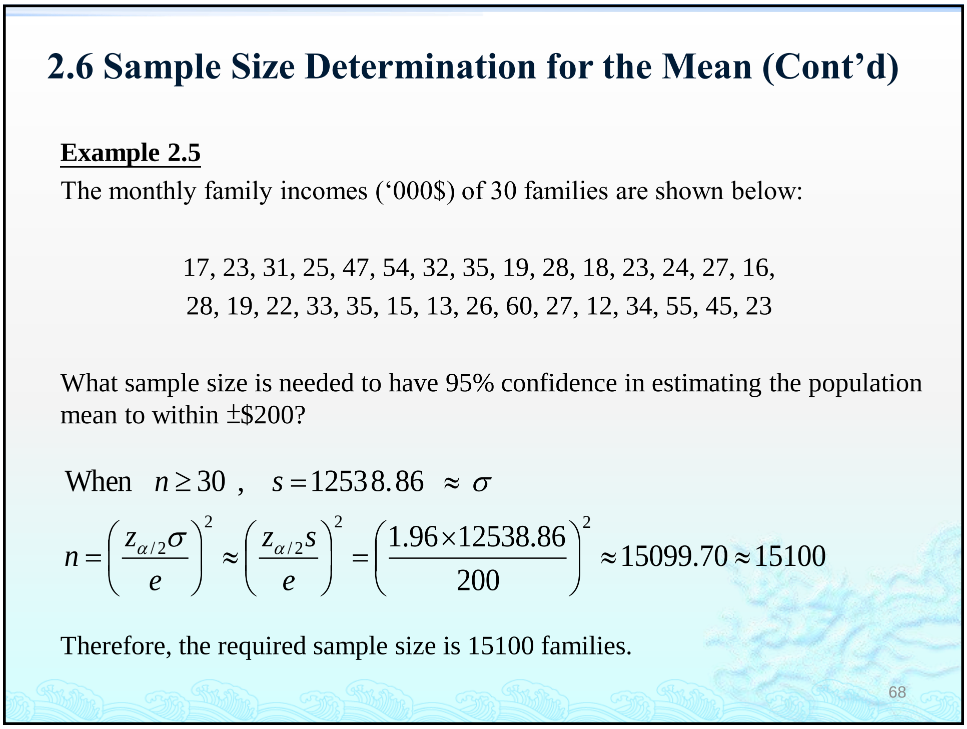
$$n = z_{\alpha/2}^2\frac{\sigma^2}{e^2}$$

$$= \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

The sample size (n) required to guard against the max error (e) is given by:

$$\boxed{n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2}$$

n is often a decimal and must be rounded up to the nearest whole number.



2.6 Sample Size Determination for the Mean (Cont'd)

Example 2.5

The monthly family incomes ('000$) of 30 families are shown below:

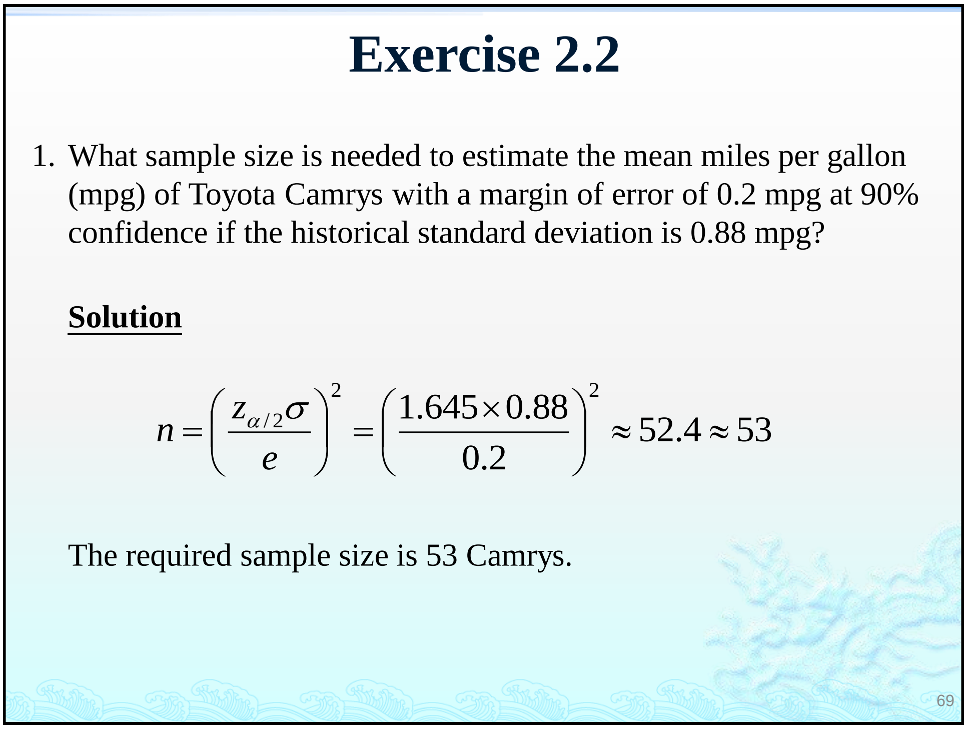
17, 23, 31, 25, 47, 54, 32, 35, 19, 28, 18, 23, 24, 27, 16,
28, 19, 22, 33, 35, 15, 13, 26, 60, 27, 12, 34, 55, 45, 23

What sample size is needed to have 95% confidence in estimating the population mean to within ±$200?

When $n \geq 30$, $s = 12538.86 \approx \sigma$

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2 \approx \left(\frac{z_{\alpha/2}s}{e}\right)^2 = \left(\frac{1.96 \times 12538.86}{200}\right)^2 \approx 15099.70 \approx 15100$$

Therefore, the required sample size is 15100 families.



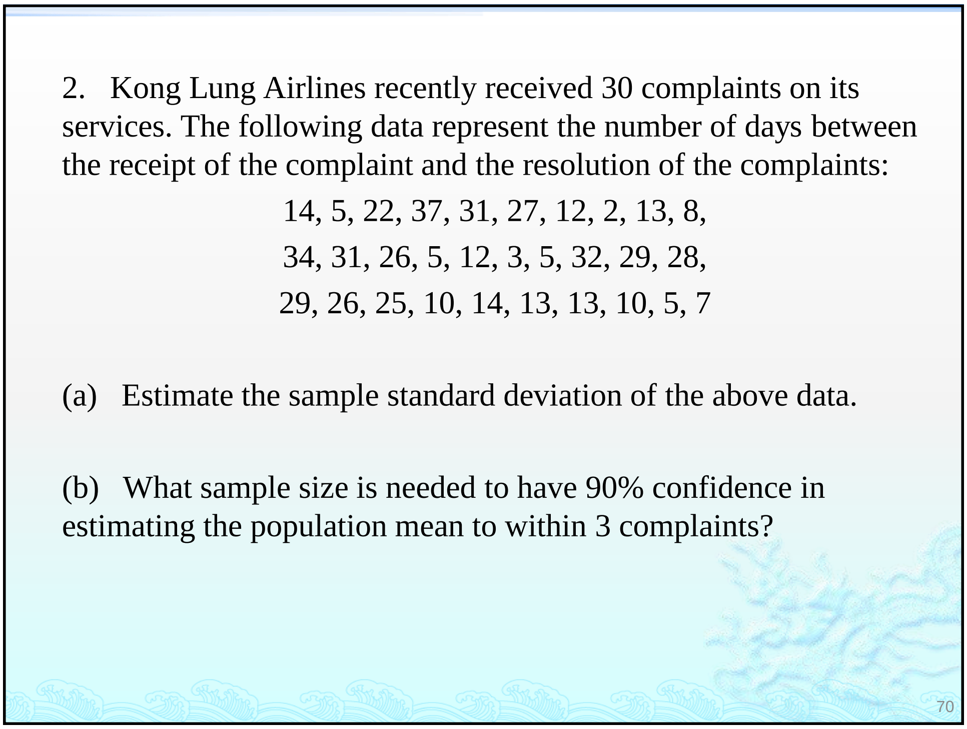
Exercise 2.2

1. What sample size is needed to estimate the mean miles per gallon (mpg) of Toyota Camrys with a margin of error of 0.2 mpg at 90% confidence if the historical standard deviation is 0.88 mpg?

Solution

$$n=\left(\frac{z_{\alpha/2}\sigma}{e}\right)^2=\left(\frac{1.645\times 0.88}{0.2}\right)^2\approx 52.4\approx 53$$

The required sample size is 53 Camrys.

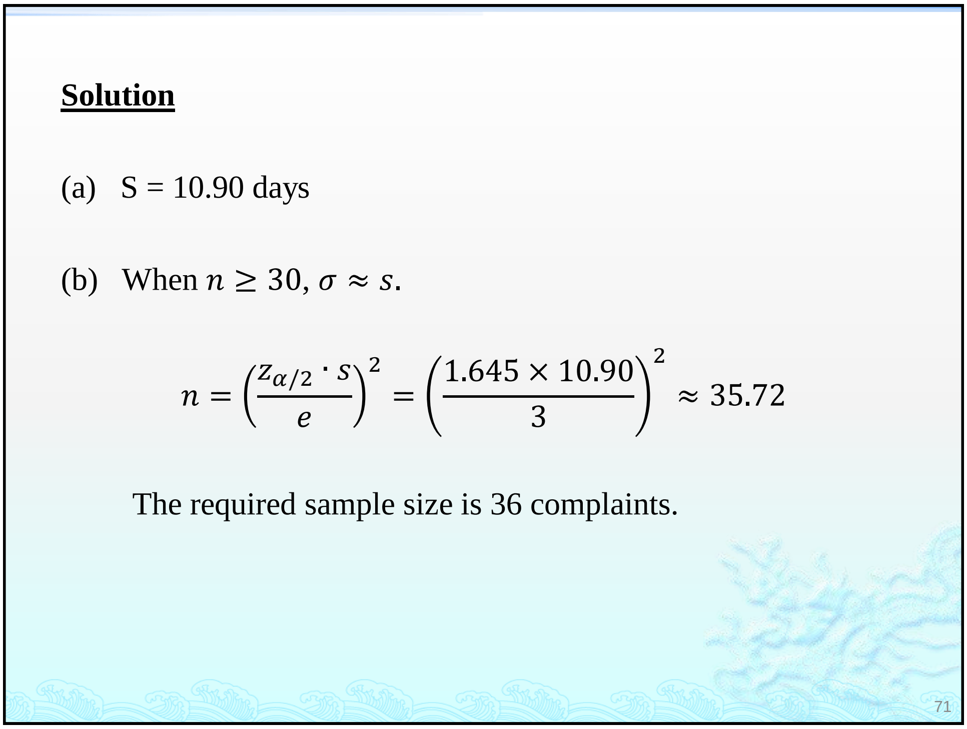


2. Kong Lung Airlines recently received 30 complaints on its services. The following data represent the number of days between the receipt of the complaint and the resolution of the complaints:

14, 5, 22, 37, 31, 27, 12, 2, 13, 8,
34, 31, 26, 5, 12, 3, 5, 32, 29, 28,
29, 26, 25, 10, 14, 13, 13, 10, 5, 7

(a) Estimate the sample standard deviation of the above data.

(b) What sample size is needed to have 90% confidence in estimating the population mean to within 3 complaints?



Solution

(a) S = 10.90 days

(b) When $n \geq 30$, $\sigma \approx s$.

$$n = \left(\frac{z_{\alpha/2} \cdot s}{e}\right)^2 = \left(\frac{1.645 \times 10.90}{3}\right)^2 \approx 35.72$$

The required sample size is 36 complaints.