

GE2262 Business Statistics

Topic 4 Sampling Distributions

Reference

Levine, D.M., Krehbiel, T.C. and Berenson, M.L., *Business Statistics: A First Course*, Pearson Education Ltd, Chapter 7

Outline

- Sampling Distributions
- Sampling Distribution of the Sample Mean
 - Standard Error of the Sample Mean
- Sampling from Normal Populations
- Sampling from Non-Normal Populations
 - The Central Limit Theorem

Sampling Distribution

- In earlier chapters, we discussed probability distribution (e.g. binomial distribution, and normal distribution) of a random variable
- Based on the assumed probability distribution, we can derive the descriptive statistics for the variables
 - Examples
 - The expected number of tagged invoices in 4 selected invoices
 - The expected amount filled in a 1-liter bottle

Sampling Distribution

Cont'd

- In order to derive these statistics, we need to know the true value of the parameters of the respective probability distribution
 - Examples
 - The expected value of a binomial distributed variable is $n\pi$
 - The expected value of a normally distributed variable is μ
- However, in practice, the values of these parameters (π and μ) are often unknown to us, and the expected values therefore need to be estimated

Sampling Distribution

Cont'd

- How to estimate the unknown expected value of a variable?
 - This is often done by :
 1. Take a random sample of the variable from the population
 - Selecting a representative sample for the population is critical for this activity. In this course, we will assume the sample is a **simple random sample** (i.e. each member of the population has an equal chance of being selected); and the sampling is done **with replacement** (i.e. the same member can be selected more than once) or from an infinite population without replacement; and it is a representative one
 2. Compute the sample mean of the observed values from the sample
 - The computed sample mean is considered as an estimate of the unknown expected value

Sampling Distribution

Cont'd

- Example: A sample of 100 people was asked for the amount they spent in their last visit to supermarket. The computed sample mean of amount spent based on the sample is calculated as \$203.345
 - We say, “the expected amount spent by the people in their last visit to supermarket is estimated as \$203.345.”

Sampling Distribution

Cont'd

- Is this estimate reliable?
 - If we select another 100 people from the population, we are likely to get a different sample mean for the amount spent, say \$210.05
 - A third sample is likely to get another mean amount spent
 - If you are going to take only 1 sample from the population, how can a conclusion be valid if we know the sample results (sample means) vary?
 - What assurance do we have that the limited information from a sample will not be misleading?

Sampling Distribution

Cont'd

- One possibility is to consider the accumulated information of “all possible samples” drawn from the population
 - Hold the sample size unchanged so that the sample size is not a complicating factor
 - The **sample results (e.g. the means)** from “all possible samples” can be organized into a distribution. This is called the **sampling distribution**
 - We will see this distribution assumes an important pattern for drawing inferences about the underlying population characteristics

Sampling Distribution of the Sample Mean

- Consider a small enterprise that has 4 staff, $N = 4$
- Age of individuals (X): 24, 26, 28, 30 measured in years



Sampling Distribution of the Sample Mean

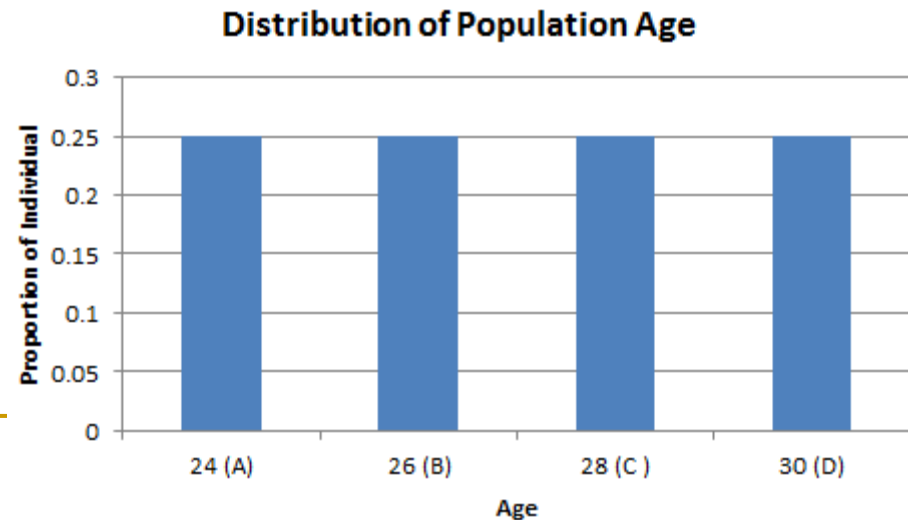
Cont'd

- Summary measures for the variable X in the population
 - The true (population) mean is:

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{24+26+28+30}{4} = 27$$

- The true (population) standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} = 2.236$$



Sampling Distribution of the Sample Mean

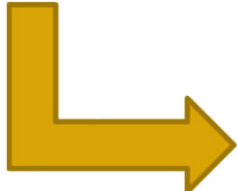
Cont'd

- Consider all possible samples (with replacement) of size $n = 2$

16 possible samples

Respondent	A	B	C	D
A	24, 24	24, 26	24, 28	24, 30
B	26, 24	26, 26	26, 28	26, 30
C	28, 24	28, 26	28, 28	28, 30
D	30, 24	30, 26	30, 28	30, 30

16 sample means



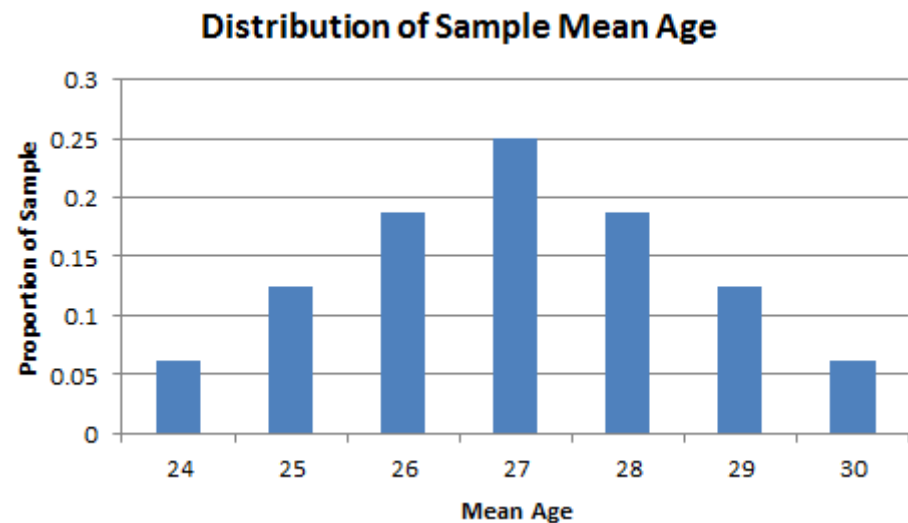
Respondent	A	B	C	D
A	24	25	26	27
B	25	26	27	28
C	26	27	28	29
D	27	28	29	30

Sampling Distribution of the Sample Mean

Cont'd

- Sample mean age is a random variable
- The sampling distribution of the mean is:

Sample Mean (\bar{X})	Frequency	Proportion
24	1	0.0625
25	2	0.125
26	3	0.1875
27	4	0.25
28	3	0.1875
29	2	0.125
30	1	0.0625



Sampling Distribution of the Sample Mean

Cont'd

- Summary measures for the sampling distribution of the sample mean
 - The true (population) mean of the sample mean is:

$$\begin{aligned}\mu_{\bar{X}} &= \sum \bar{X}_i P(\bar{X}_i) \\ &= 24 \left(\frac{1}{16} \right) + \cdots + 30 \left(\frac{1}{16} \right) = 27\end{aligned}$$

- The true (population) standard deviation of the sample mean is:

$$\begin{aligned}\sigma_{\bar{X}} &= \sqrt{\sum (\bar{X}_i - \mu_{\bar{X}})^2 P(\bar{X}_i)} \\ &= \sqrt{(24 - 27)^2 \left(\frac{1}{16} \right) + \cdots + (30 - 27)^2 \left(\frac{1}{16} \right)} = 1.5811\end{aligned}$$

What do you notice?

Population Distribution vs. Sampling Distribution of the Sample Mean

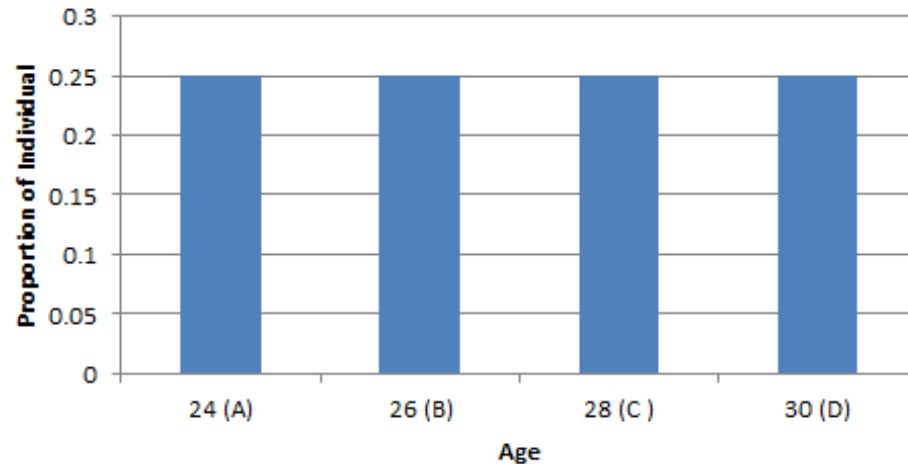
- Comparing the distribution between X and \bar{X} , we observe

Distribution of X

$$N = 4$$

$$\mu = 27 \quad \sigma = 2.236$$

Distribution of Population Age

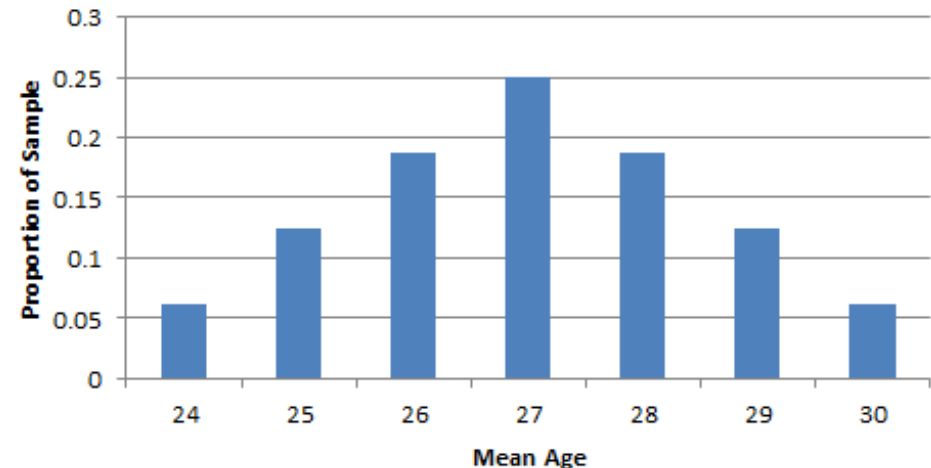


Distribution of \bar{X}

$$n = 2$$

$$\mu_{\bar{X}} = 27 \quad \sigma_{\bar{X}} = 1.5811$$

Distribution of Sample Mean Age



Sampling Distribution of the Sample Mean

Cont'd

- Comparing the distribution between age (X) and mean age (\bar{X}), we observe:
 - X is a random variable, so is \bar{X}
 - The population mean of X and the population mean of \bar{X} are identical
 - The population standard deviation of X is larger than that of \bar{X}
 - While the shape of the distribution of X is uniform (even), the distribution of \bar{X} appears to be in bell shape
- These features did not happen by pure chance
- The behaviour of \bar{X} is in fact governed by the theory of sampling distribution

Properties of Sampling Distribution of the Sample Mean

■ Mean of sample means

- $\mu_{\bar{X}} = \mu$
- Works for sampling with and without replacement if the samples are unbiased

■ Standard deviation of sample means

- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- Also called standard error of the mean
- Works for sampling with replacement, or sampling from large populations without replacement
- As n increases, $\sigma_{\bar{X}}$ decreases

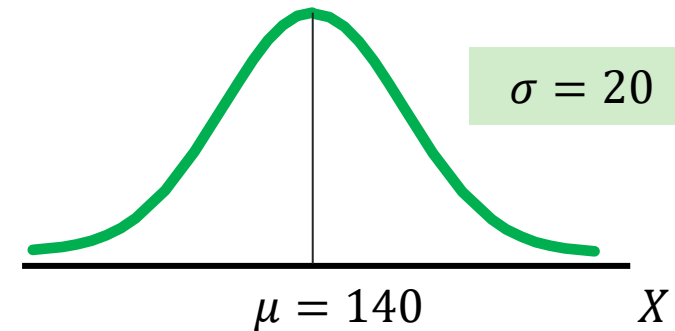
Sampling from Normal Populations

- For $X \sim N(\mu, \sigma^2)$,
 $\bar{X} \sim N(\mu_{\bar{X}}, (\sigma_{\bar{X}})^2)$
 - As $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$,

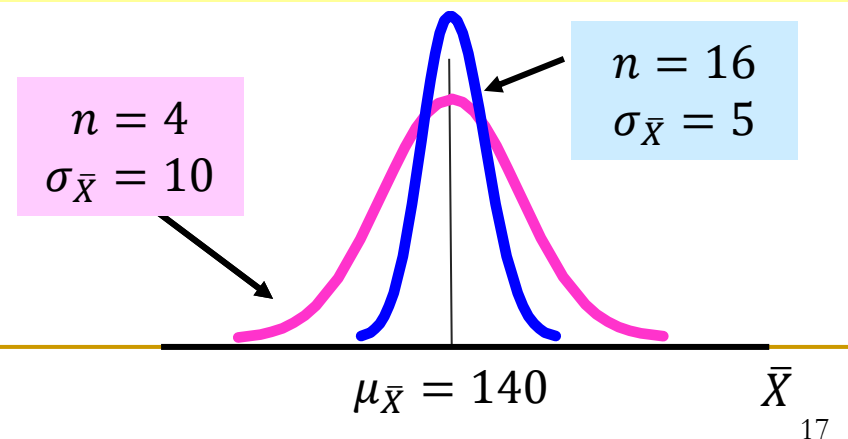
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- This is a property of sampling from a Normally distributed population

Population Distribution



Sample Mean Distributions



Sampling from Normal Populations – Example

Cont'd

- Suppose the packing equipment in a manufacturing process that is filling 350-gram boxes of cereal is set so that the amount of cereal in a box is normally distributed with a mean of 350 grams. From past experience, the population standard deviation for this filling process is known to be 15 grams. If a sample of 25 boxes is randomly selected from the many thousands that are filled in a day, what is the probability that the sample mean is in between 345 grams and 355 grams?

Sampling from Normal Populations – Example

Cont'd

- Let X be the amount of cereal in a box, and \bar{X} be the sample mean of the amount of cereal in the sample of 25 boxes respectively

- Since $X \sim N(350, 15^2)$, then $\bar{X} \sim N(350, \frac{15^2}{25})$

$$\begin{aligned} P(345 < \bar{X} < 355) &= P\left(\frac{345 - 350}{15/5} < \frac{\bar{X} - 350}{15/5} < \frac{355 - 350}{15/5}\right) \\ &= P(-1.6667 < Z < 1.6667) \\ &= 0.9050 \end{aligned}$$

- When the manufacturing process is operating properly, there is a good possibility that the mean amount cereal in 25 randomly selected boxes is within 345 grams and 355 grams
- If the observed sample mean is in fact outside this range, then the manufacturing process needs to be adjusted

Sampling from Non-Normal Populations

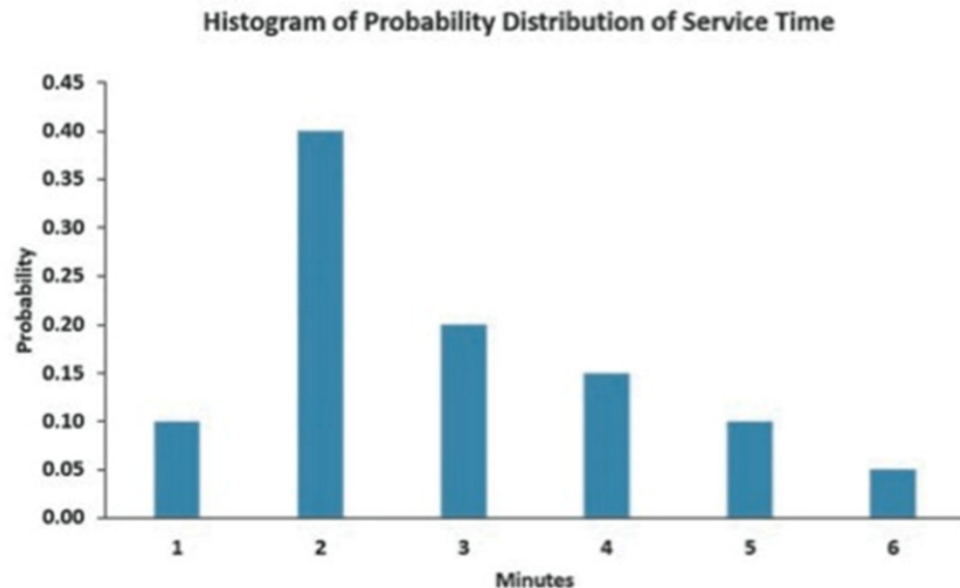
- In many instances either we will know the population is not normally distributed or we may believe that it is unrealistic to assume a normal distribution
 - For example, the monthly salary of people in Hong Kong
- What is the sampling distribution of the mean for populations that are not normally distributed?

Sampling from Non-Normal Populations – Example

Cont'd

- Consider the distribution of time it takes to fill orders at a fast-food chain counter. The population mean service time is 2.9 minutes and the population standard deviation is 1.34 minutes

Service Time (minutes)	Probability
1	0.10
2	0.40
3	0.20
4	0.15
5	0.10
6	0.05

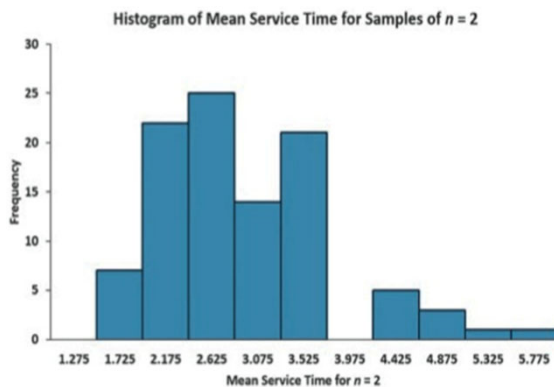


Sampling from Non-Normal Populations – Example

Cont'd

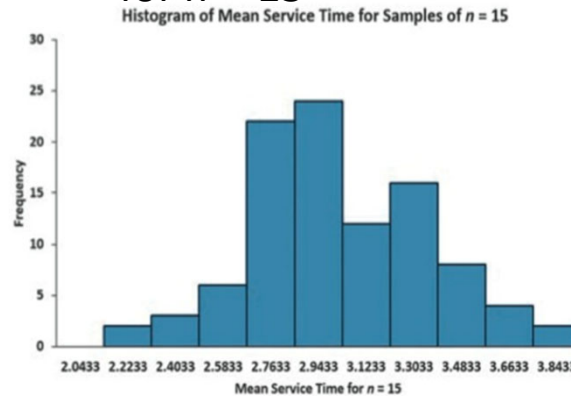
- Suppose 100 samples of $n = 2$, $n = 15$, $n = 30$ are selected. For each sample, the sample mean is calculated

Mean Service Time
for $n = 2$



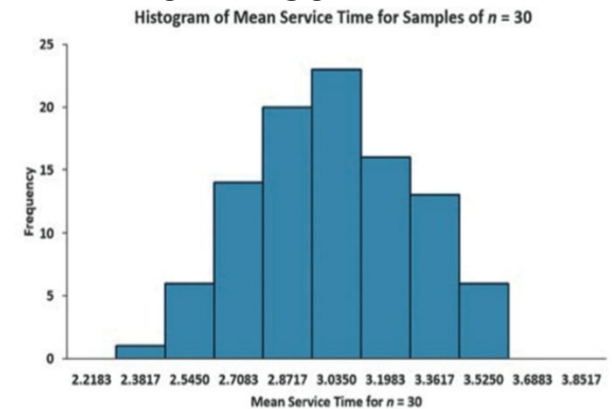
Skewed, but not as skewed as the population distribution

Mean Service Time
for $n = 15$



Somewhat symmetrical distribution that contains a concentration of values in the center of the distribution

Mean Service Time
for $n = 30$



Approximately bell-shaped with a concentration of values in the center of the distribution

Sampling from Non-Normal Populations – Example

Cont'd

- The population mean service time is 2.9 minutes and the population standard deviation is 1.34 minutes

n	Mean of Sample Means (based on the 100 selected samples)	Standard error of the mean (based on the 100 selected samples)	Standard error of the mean (based on theory, σ/\sqrt{n})
2	2.825	0.883	0.9475
15	2.9313	0.3458	0.3460
30	2.9527	0.2701	0.2446

- The mean of sample means approaches the value of the population mean
- As n increases, the standard error of the mean decreases
- The standard error of the mean is close to the value of $\frac{\sigma}{\sqrt{n}}$

Sampling from Non-Normal Populations

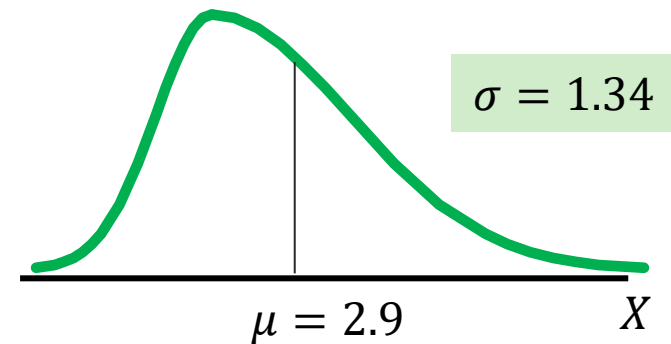
Cont'd

- For X follows non-normal distribution, even

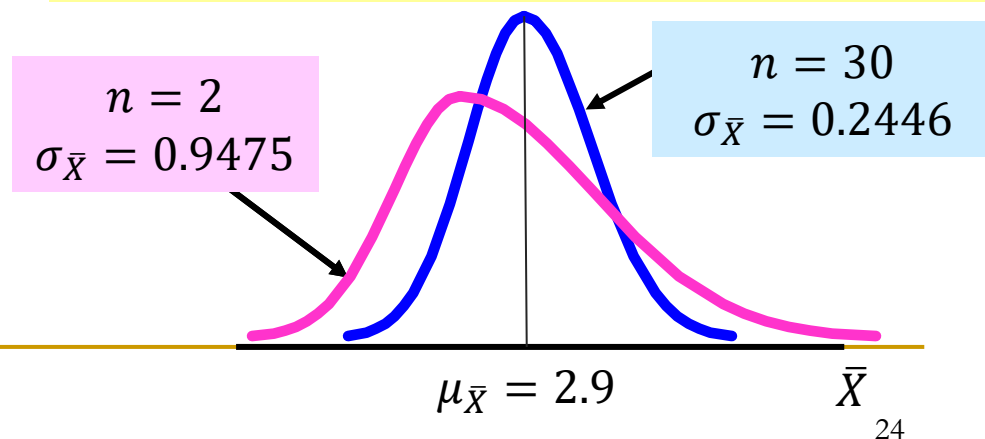
$$\mu_{\bar{X}} = \mu \text{ and } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

the distribution of \bar{X} will vary from sample sizes

Population Distribution



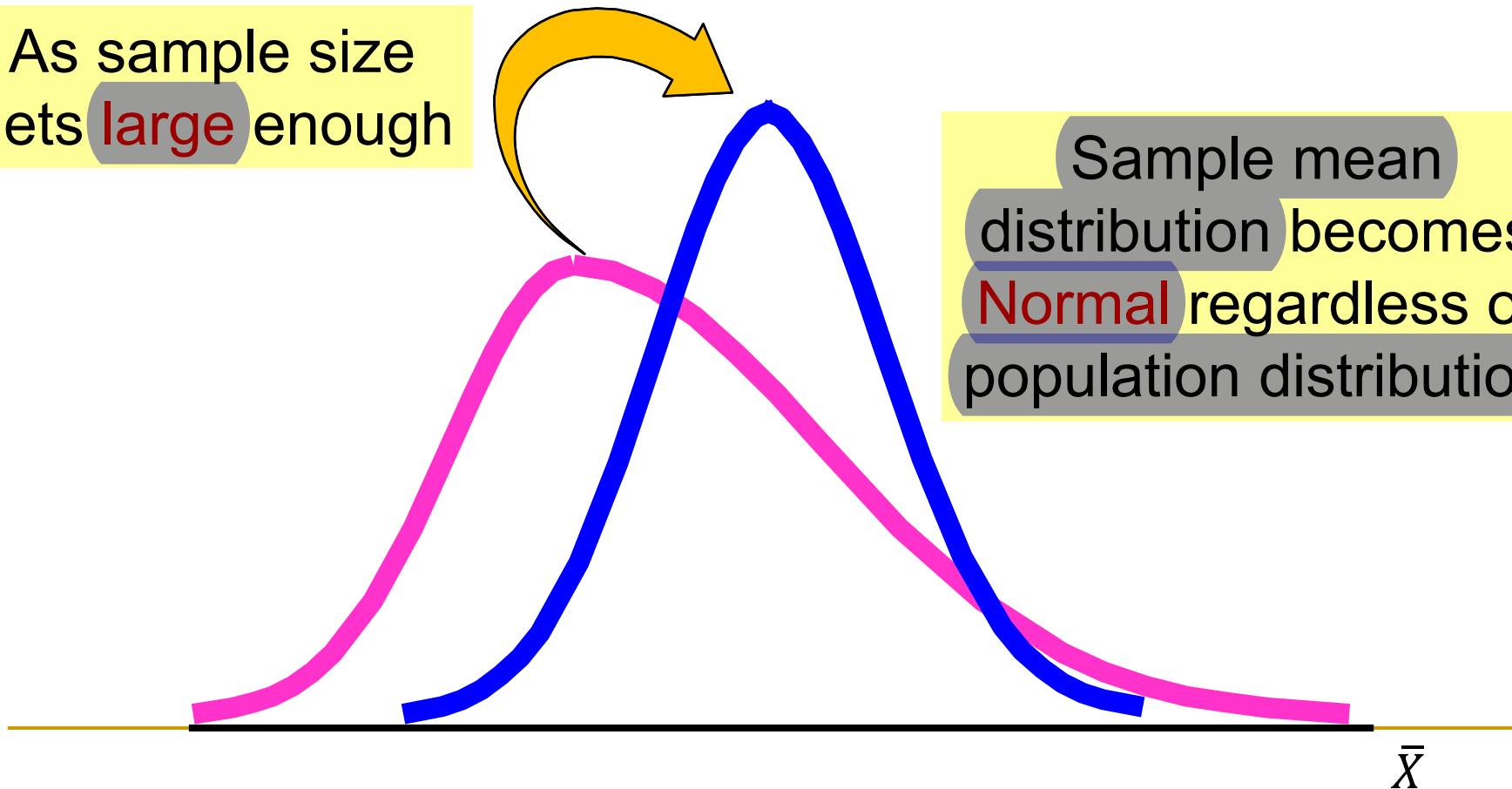
Sample Mean Distributions



Central Limit Theorem

As sample size
gets **large** enough

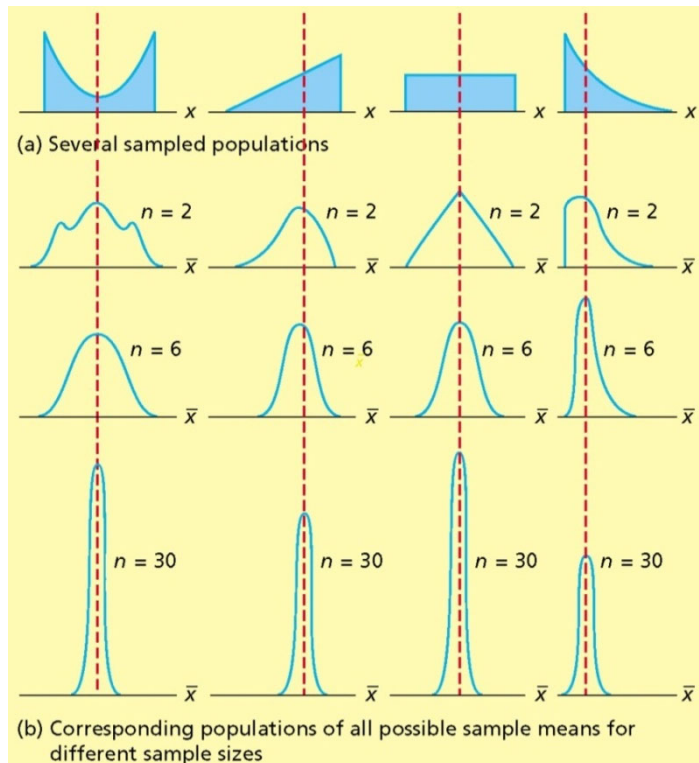
Sample mean
distribution becomes
Normal regardless of
population distribution



Central Limit Theorem

Cont'd

- But how large is large enough?
 - For most distributions, $n \geq 30$ is considered as a large sample



The larger the sample size, the more nearly normally distributed is the sampling distributions of the means

Sampling Distribution – Example

- Weights of a certain population can be assumed normal with mean 140 lb. and standard deviation 20 lb.
- 1. What is the chance that the mean weight of a sample of 20 exceeding 150 lb.?

Let the weight of an individual be X , $X \sim N(140, 20^2)$,

$$\bar{X} \sim N\left(140, \left(\frac{20}{\sqrt{20}}\right)^2\right)$$

$$P(\bar{X} > 150) = P\left(Z > \frac{150 - 140}{20/\sqrt{20}}\right) = P(Z > 2.236) = 0.0125$$

Sampling Distribution – Exercise

Cont'd

2. Will the chance be the same if the sample consists of 30 individuals? Why? If your answer is “NO”, what the chance should be?

Sampling Distribution – Exercise

Cont'd

3. If the 20 individuals are randomly selected from a non-Normal population, what is the probability that its mean weight exceeds 150lb.?

Sampling Distribution – Exercise

Cont'd

4. For a sample consists of 30 individuals from a non-normal population, what is the probability that its sample mean weight falls between 135 lb. and 150 lb.?

Sampling Distribution – Exercise

Cont'd

2. Will the chance be the same if the sample consists of 30 individuals? Why? If your answer is "NO", what the chance should be?

No. As by increasing the sample size, the standard error will drop, leading to a larger Z value, and a smaller upper-tail area.

$$\begin{aligned}\bar{X} &\sim N\left(140, \left(\frac{20}{\sqrt{30}}\right)^2\right) \\ P(\bar{X} > 150) &= P\left(Z > \frac{150 - 140}{20/\sqrt{30}}\right) = P(Z > 2.739) \\ &= 1 - 0.9969 = 0.0031\end{aligned}$$

28

Sampling Distribution – Exercise

Cont'd

3. If the 20 individuals are randomly selected from a non-Normal population, what is the probability that its mean weight exceeds 150lb.?

Since the population is non-Normal, and the sample size is small, we are unable to tell how the sample mean is being distributed, and the corresponding probability.

29

Sampling Distribution – Exercise

Cont'd

4. For a sample consists of 30 individuals from a non-normal population, what is the probability that its sample mean weight falls between 135 lb. and 150 lb.?

The population distribution is non-normal, but the sample size is large ($n \geq 30$), we can conclude that

$\bar{X} \sim N(140, (\frac{20}{\sqrt{30}})^2)$ according to the Central Limit Theorem.

$$\begin{aligned}P(135 < \bar{X} < 150) &= P\left(\frac{135-140}{20/\sqrt{30}} < Z < \frac{150-140}{20/\sqrt{30}}\right) \\ &= P(-1.369 < Z < 2.739) \\ &= 0.9969 - 0.0853 = 0.9116\end{aligned}$$

30