GE2262 Business Statistics
Topic 4 Sampling Distribution

Lecturer: Dr. Iris Yeung

Room : LAU-7239

Tel No.: 34428566

E-mail: msiris@cityu.edu.hk

Outline

- Sampling Distribution of the Sample Mean
- Sampling Distribution of the Sample Proportion

Reference

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

Part One

- Sampling Distribution of the Sample Mean
- Sampling Distribution of the Sample Proportion

Population Parameter and Sample Statistic

- A population contains **all** the items or individuals about which we want to study
- A sample contains only a **portion** of the population of items or individuals
- A variable is a characteristic of an item or individual
- A population parameter summarizes the value of a specific variable for a population
  - Population proportion (p) is the proportion of items in the entire population that have a certain characteristic
- A sample statistic summarizes the value of a specific variable for sample data
  - Sample proportion ($\hat{p}$) is the proportion of items in the sample that have a certain characteristic

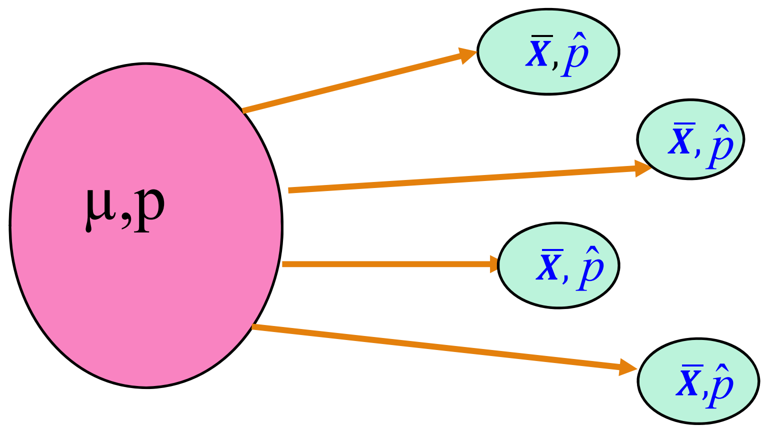
Measure	Population parameter	Sample statistic
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	p	$\hat{p}$

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Sampling Distribution of Sample Statistic

- Different samples give different sample statistic values
- The values of the sample statistic from "all possible samples" can be organized into a distribution, called sampling distribution
- The sampling distribution of a sample statistic is the probability distribution of all of the possible values of a sample statistic for a given sample size (*n)* selected from a population

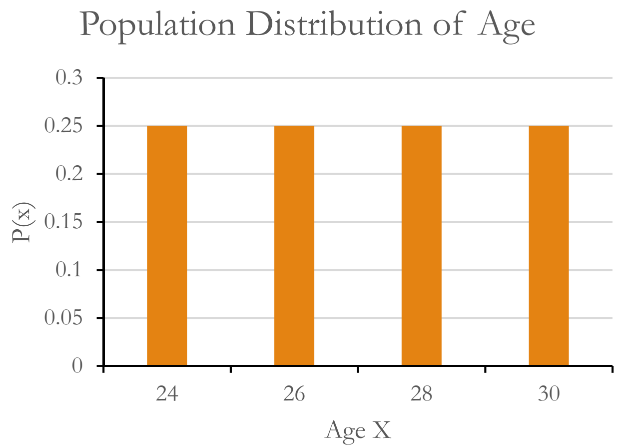


Example 1: Sampling Distribution of the Sample Mean

- Consider a small enterprise that has 4 staff (A, B, C, D) with age measured in years : 24, 26, 28, 30
- Population size $N = 4$, X = Age of individuals
- Population mean $\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{24+26+28+30}{4} = 27$
- Population standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}} = 2.236$

Probability Distribution of X

X	$P(X = x)$
24	1/4 = 0.25
26	1/4 = 0.25
28	1/4 = 0.25
30	1/4 = 0.25
Total	1



Example 1: Sampling Distribution of the Sample Mean

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

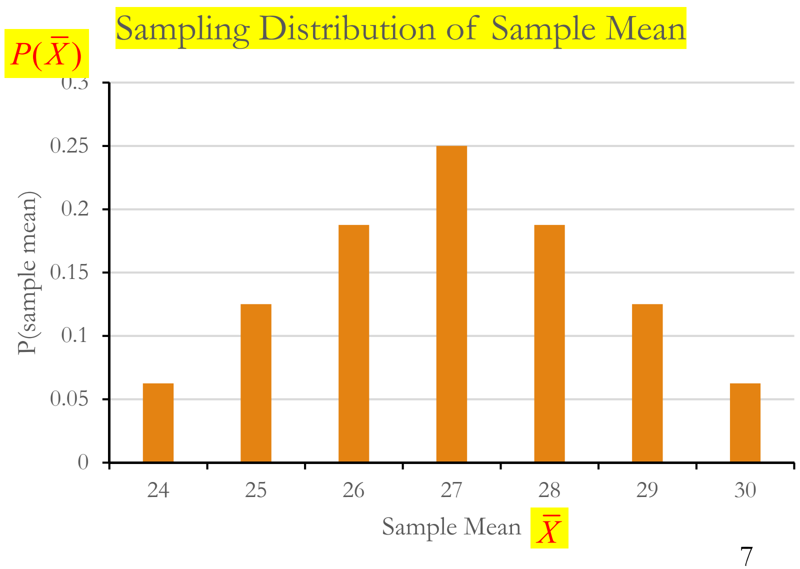
16 possible samples

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

16 sample means

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

Sample Mean $(\bar{X})$	Frequency	Probability $P(\bar{X})$
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
Total	16	1.000



Example 1: Sampling Distribution of the Sample Mean

Sample Mean ($\bar{X}$)	Probability $P(\bar{X})$	$\bar{X}P(\bar{X})$	$(\bar{X}-\mu_{\bar{X}})^2P(\bar{X})$
24	0.0625	1.5	0.5625
25	0.125	3.125	0.5
26	0.1875	4.875	0.1875
27	0.25	6.75	0
28	0.1875	5.25	0.1875
29	0.125	3.625	0.5
30	0.0625	1.875	0.5625
Total	1.000	27	2.5

- Mean value (expected value) of the sampling distribution of sample mean:

$$\mu_{\bar{X}} = \sum \bar{X}P(\bar{X})$$

$$= 24\left(\frac{1}{16}\right) + \cdots + 30\left(\frac{1}{16}\right) = 27$$

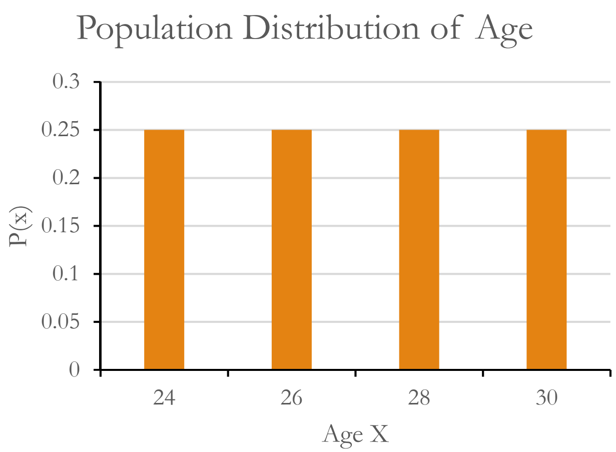
- Standard deviation of the sampling distribution of of the sample mean:

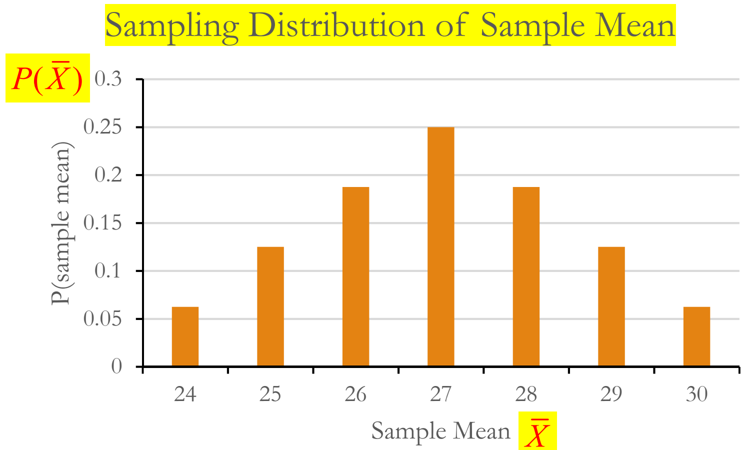
$$\sigma_{\bar{X}} = \sqrt{\sum(\bar{X}-\mu_{\bar{X}})^2P(\bar{X})}$$

$$= \sqrt{(24-27)^2\left(\frac{1}{16}\right) + \cdots + (30-27)^2\left(\frac{1}{16}\right)} = \sqrt{2.5} = 1.5811$$

Example 1: Population Distribution vs Sampling Distribution of the Sample Mean

Population Distribution of X	Sampling Distribution of $\bar{X}$
$N = 4$	$n = 2$
$\mu = 27;\ \sigma = 2.236$	$\mu_{\bar{X}} = 27;\ \sigma_{\bar{X}} = 1.5811$





- Mean of the sampling distribution of $\bar{X}$ ($\mu_{\bar{X}}$) = Mean of the population (μ)
- Standard deviation of the sampling distribution of $\bar{X}$ < Standard deviation of the population (σ)
- The distribution of $\bar{X}$ appears to be bell shaped even though the population is not

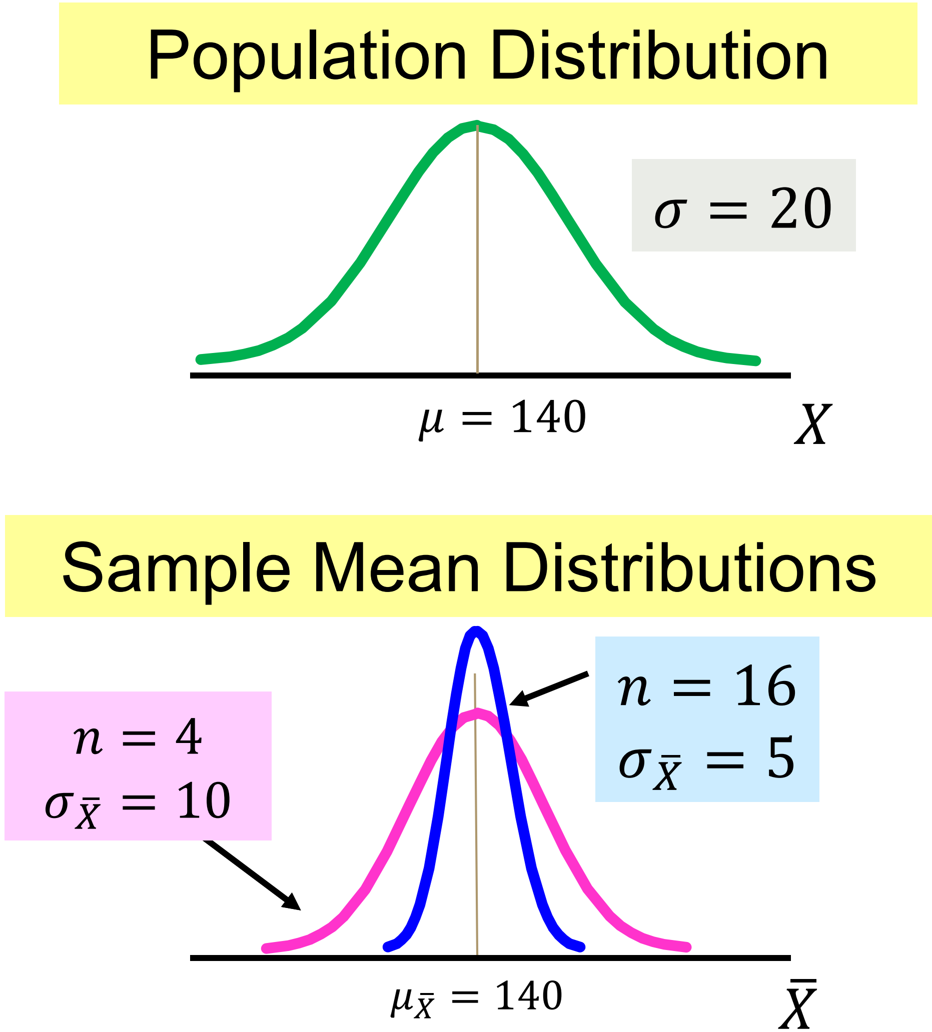
Properties of Sampling Distribution of the Sample Mean

- For sampling with replacement, or sampling from large population without replacement
  - Mean of the sampling distribution of $\bar{X}$ = Mean of the population
    - $\mu_{\bar{X}} = \mu$
  - Standard deviation of the sampling distribution of $\bar{X}$ (also called standard error of the mean)
    - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
    - As n increases, $\sigma_{\bar{X}}$ decreases

Shape of Sampling Distribution of the Sample Mean

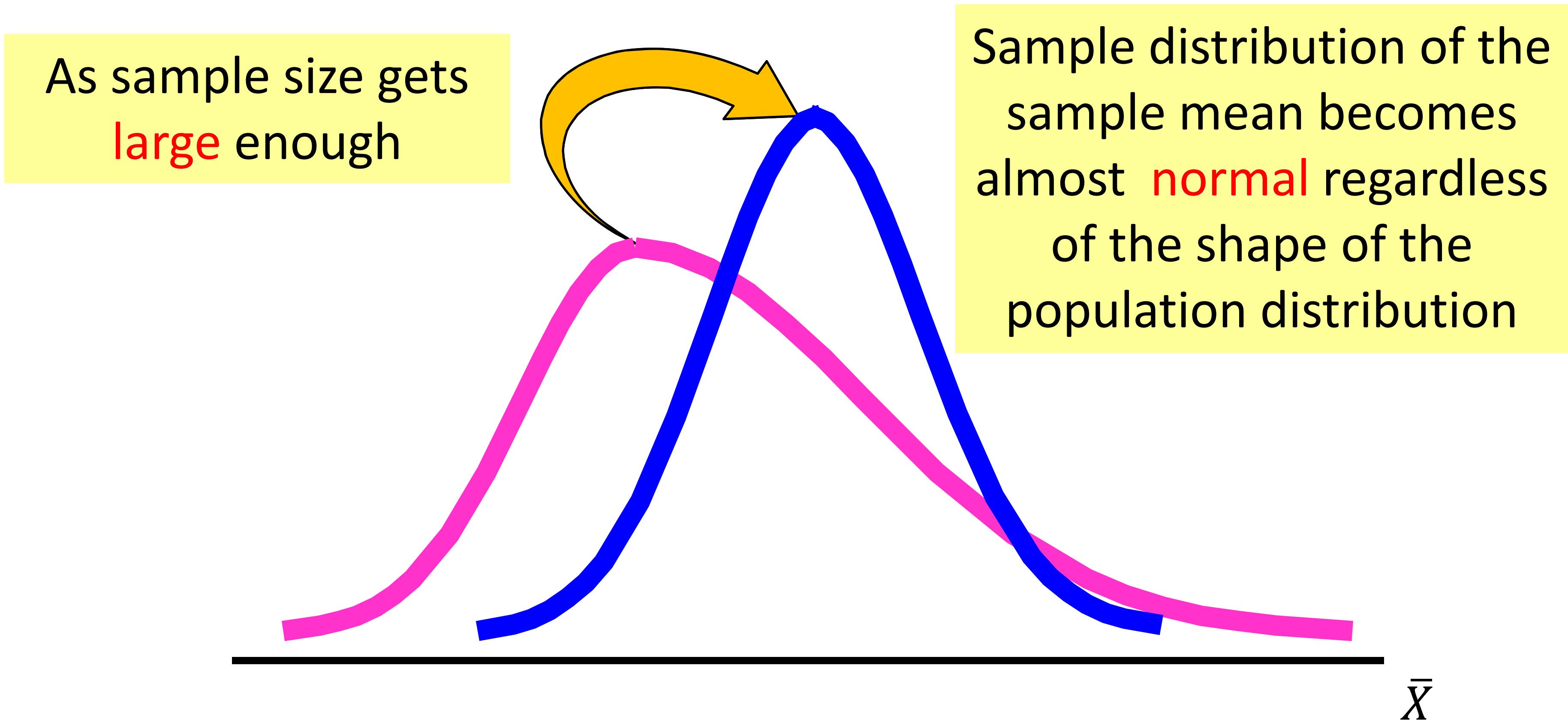
- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- $X \sim N(\mu, \sigma^2) \rightarrow$

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



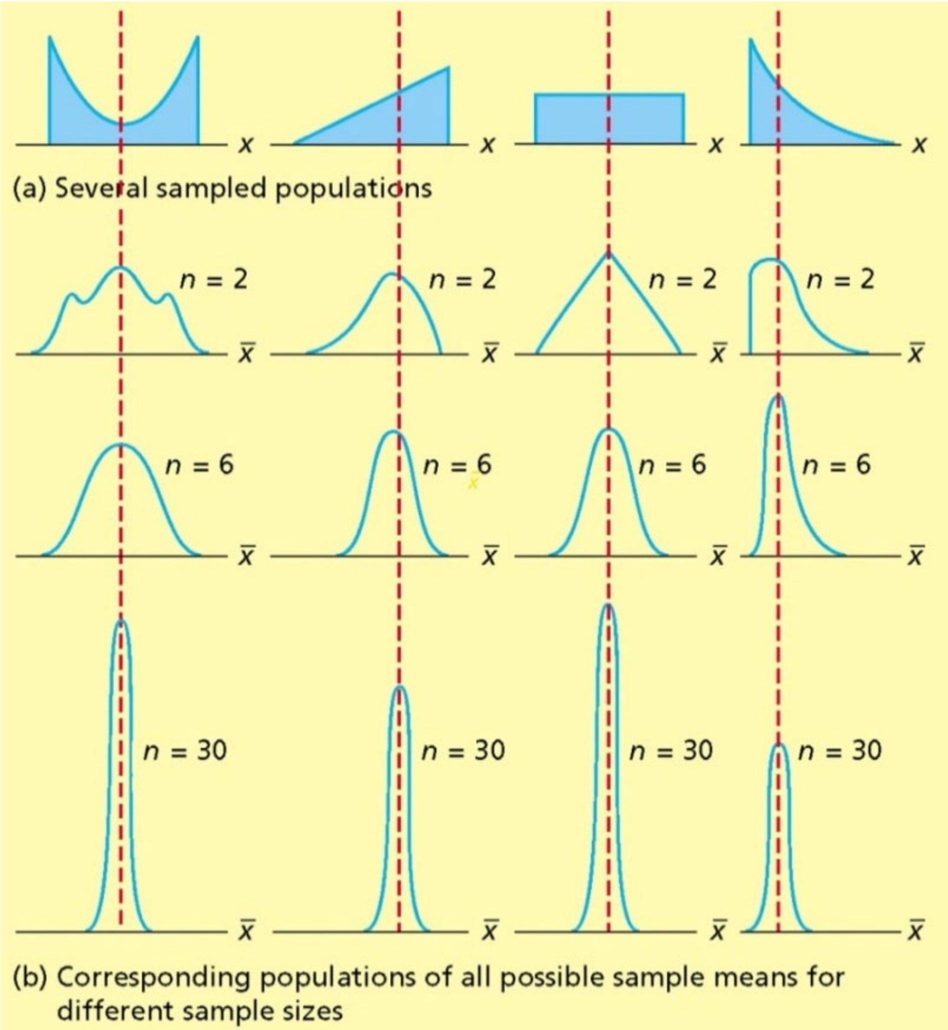
Central Limit Theorem

- If the population is not normal, sampling distribution of the sample means will be approximately normal as long as the sample size is large enough.



Shape of Sampling Distribution of the Sample Mean

- For most distributions, $n \geq 30$ will give a sampling distribution that is nearly normal



The larger the sample size, the more nearly normally distributed is the sampling distribution of the sample mean

Example 2: Sampling from Normal Populations

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

$$P(345 < \bar{X} < 355) = P\left(\frac{345-350}{15/5} < Z < \frac{355-350}{15/5}\right)$$
$$= P(-1.6667 < Z < 1.6667)$$
$$= 0.9525 - 0.0475 = 0.9050$$

- 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

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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

Example 3: Sampling from Normal Population

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

Let X = weight of an individual

$$X \sim N(140, 20^2),$$

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

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

$$= 1 - 0.9875 = 0.0125$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

Example 3: Sampling from Normal Population

2. Suppose the weights of a certain population is normal with mean 140 lb. and standard deviation 20 lb.
   - 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 deviation of the sampling distribution of $\bar{X}$ will drop, leading to a larger Z value, and a smaller upper-tail area.

$$\bar{X} \sim N\left(140, \left(\frac{20}{\sqrt{30}}\right)^2\right)$$

$$P(\bar{X} > 150) = P\left(Z > \frac{150 - 140}{20/\sqrt{30}}\right) = P(Z > 2.739)$$

$$= 1 - 0.9969 = 0.0031$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981

Example 4: Sampling from Non-Normal Population

3. Suppose the weights of a certain population is non-normal with mean 140 lb. and standard deviation 20 lb.
   - If the 20 individuals are randomly selected from the 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.

Example 4: Sampling from Non-Normal Population

4. Suppose the weights of a certain population is non-normal with mean 140 lb. and standard deviation 20 lb.
   - For a sample of 30 individuals, 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.

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

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

Part Two

- Sampling Distribution of the Sample Mean
- Sampling Distribution of the Sample Proportion

Binomial Distribution

- Recall Bernoulli/Binomial Experiment in Topic 3
  - The experiment is repeated n times (n trials).
  - Each trial has only two possible outcomes (denoted as success S and failure F).
  - The probability of success, denoted by p, is the same for each trial.
    - The probability of failure for each trial is equal to $q=1-p$.
  - The trials are independent (the outcome of a trial does not depend on the outcomes of previous trials).
  - We are interested in the random variable X where X is the number of successes observed in n trials. Note the possible values of X are $0,1,2,\ldots,n$.
- Examples of Bernoulli/Binomial experiment
  - Toss a coin is a Bernoulli trial – in each toss, there are 2 outcomes: a head or a tail. We can designate head as a success and tail as a failure. For a balanced coin, the probability of getting a head in each trial is 0.5.
  - Salesman contacting a customer is a Bernoulli trial -- In each contact, there are 2 possible outcomes: the customer purchases the product (success), or the customer does not purchase the product (failure). From past experience, the salesperson knows that the probability of buying a product in each contact is 0.1.

Binomial Distribution

Let n = number of independent identical trials

p = the probability of "success" in a trial

X = the number of "successes" out of n trials

Probability distribution of X

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

$$X \sim BIN(n, p)$$

$$E(X) = np, Var(X) = np(1-p)$$

Binomial Distribution

Example

In the salesman example, suppose the salesman contacts 147 customers $\rightarrow n$=147. From past experience, the salesperson knows that the probability of buying a product in each contact is 0.1 $\rightarrow p$ =P(buy)=0.1.

let X = number of customers (out of 147) who will purchase the product.

What is the probability that exactly twenty customers will purchase the product?

$$P[X=20]=\frac{147!}{20!127!}(0.1)^{20}(0.9)^{127}=0.036$$

What is the probability that at most twenty customers will purchase the product?

If n is large such that $np \geq 5, n(1-p) \geq 5$, the Binomial distribution can be approximated by Normal distribution with $\mu = np, \sigma^2 = npq$, where $q=(1-p)$

$$P[X \leq 20] = P(Z \leq \frac{20.5-147*0.1}{\sqrt{147*0.1*0.9}}) = P(Z \leq 1.595) = 0.9446$$

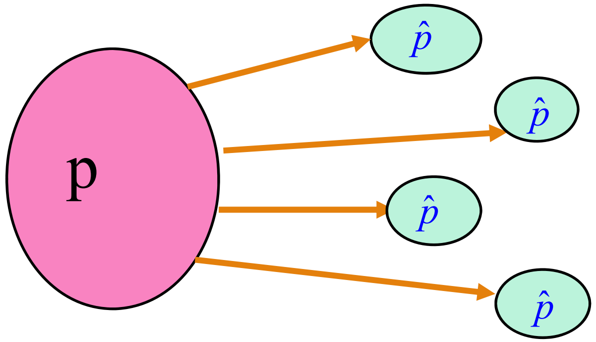
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

Sample Proportion

- Now we are interested in the sample proportion

$$\hat{p} = \frac{X}{n} = \frac{\text{number of successes in } n \text{ trials}}{n \text{ trials}}$$

  not the number of successes X



- Example
  - Suppose the salesman contacts 147 customers and 20 of them buy the product. Sample proportion is

$$\hat{p} = \frac{X}{n} = \frac{\text{no. of customers who buy}}{\text{sample size}} = \frac{20}{147} = 0.1361$$

  - If the salesman contacts another 147 customers, the sample proportion of customers buying the product will vary
  - From past experience, the salesman knows that the probability of buying a product in each contact is 0.1. We can treat this probability of success (buy the product) as population proportion of success *p*.
  - The probability of success is the same in each trial. Population proportion of success *p* is a fixed value, not variable
- Different sample gives different sample proportion. We want to find the distribution of sample proportion

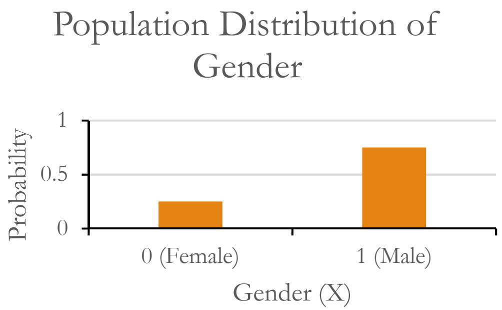
Example 5: Sampling Distribution of the Sample Proportion

- Suppose the small enterprise has 3 male and 1 female staff
- Population size $N = 4$
- Suppose we are now interested in the gender variable (2 values: male, female).
- Let *X*=1 if male, *X*=0 if female →the X value for staff A, B, C and D is 1,1,0,1 respectively
- Population mean $\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{1+1+0+1}{4} = 0.75,$
  - Population proportion (*p*) of male staff in the enterprise is ¾=0.75. Population mean of X variable where X=0,1 is actually the population proportion of success ("success" in this example refers to being a male staff)
- Population standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}} = \sqrt{\frac{3*(1-0.75)^2+(0-0.75)^2}{4}} = \sqrt{0.1875} = 0.433$$

Probability Distribution of X

X	$P(X = x)$
0	1/4 = 0.25
1	3/4 = 0.25
Total	1



Example 5: Sampling Distribution of the Sample Proportion

- Consider all possible samples with replacement of size *n*=2 and calculate sample mean (i.e. sample proportion of male staff for each sample)
  - sample mean of X variable where *X*=0,1 is actually the sample proportion of success

16 possible samples

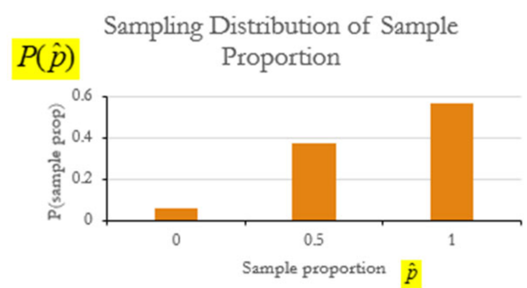
Respondent	A (X=1)	B (X=1)	C (X=0)	D (X=1)
A (X=1)	1,1	1,1	1,0	1,1
B (X=1)	1,1	1,1	1,0	1,1
C (X=0)	0,1	0,1	0,0	0,1
D (X=1)	1,1	1,1	1,0	1,1

16 possible sample proportions

Respondent	A (X=1)	B (X=1)	C (X=0)	D (X=1)
A (X=1)	2/2 = 1	2/2 = 1	1/2 = 0.5	2/2 = 1
B (X=1)	2/2 = 1	2/2 = 1	1/2 = 0.5	2/2 = 1
C (X=0)	1/2 = 0.5	1/2 = 0.5	0/2 = 0	1/2 = 0.5
D (X=1)	2/2 = 1	2/2 = 1	1/2 = 0.5	2/2 = 1

- Probability distribution of sample proportion $\hat{p}$ is:

$\hat{p}$	0	0.5	1
$P(\hat{p})$	1/16	6/16	9/16



Example 5: Sampling Distribution of the Sample Proportion

$\hat{p}$	$P(\hat{p})$	$\hat{p}[P(\hat{p})]$	$(\hat{p}-\mu_{\hat{p}})^2[P(\hat{p})]$
0	0.0625	0	0.0351563
0.5	0.375	0.1875	0.0234375
1	0.5625	0.5625	0.0351563
Total	**1**	**0.75**	**0.09375**

- Mean, variance and standard deviation of the sampling distribution of sample proportion

$$\mu_{\hat{p}} = \sum \hat{p}P(\hat{p}) = 0(1/16) + 0.5(6/16) + 1(9/16) = 0.75$$

$$Var(\hat{p}) = \sum (\hat{p} - \mu_{\hat{p}})^2 P(\hat{p})$$
$$= (0-0.75)^2(1/16) + (0.5-0.75)^2(6/16) + (1-0.75)^2(9/16) = 0.09375$$

$$\sigma(\hat{p}) = \sqrt{0.09375} = 0.3062$$

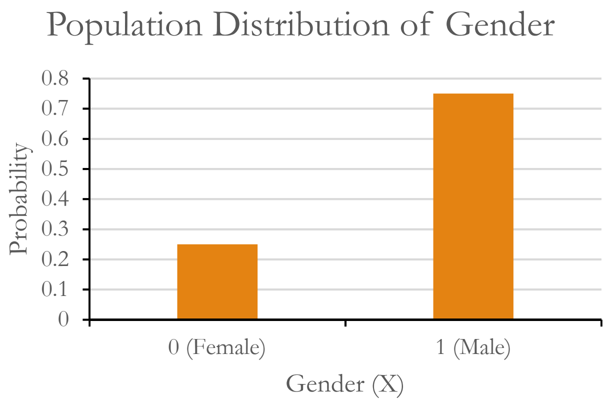
Example 5: Population Distribution vs Sampling Distribution of the Sample Proportion

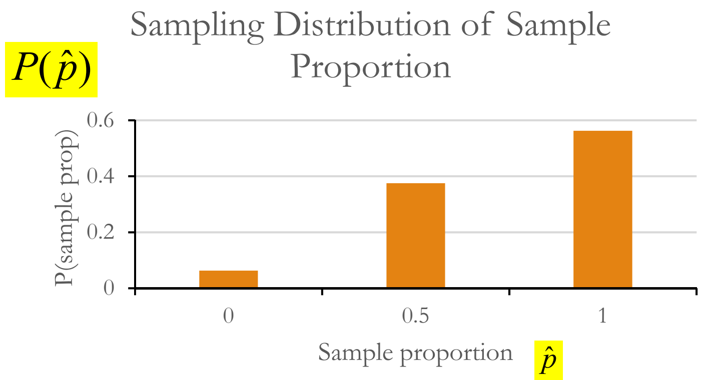
Population Distribution of Gender
(X=1 for Male, 0 for Female)
$N = 4$

$\mu = E(X) = p = 0.75; \quad \sigma = 0.433$

Sampling Distribution of sample proportion $\hat{p}$
$n = 2$

$\mu_{\hat{p}} = 0.75, \quad \sigma_{\hat{p}} = 0.3062$





- Mean of the sampling distribution of $\hat{p}$ = Population proportion $\mu_{\hat{p}} = p$
- Standard deviation of the sampling distribution of $\hat{p}$ < Standard deviation of the population $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.75(0.25)}{2}} = 0.3062$$

Sampling Distribution of the Sample Proportion

- Recall central limit theorem
  - if *n* is large ($n \geq 30$), sampling distribution of the sample mean is nearly normal even though the population is not normal
- Sample proportion is a special case of sample mean
  - If *n* is large such that *np≥5, nq≥5*, the sampling distribution of the sample proportion is approximately normal with mean *p* and variance *pq/n*

$$\hat{p} \sim N(p, \frac{pq}{n}) \quad \Rightarrow \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

- Recall in Topic 3 binomial distribution, we are interested in $X = n\hat{p}$= number of successes in *n* trials
- If *n* is large such that *np≥5, nq≥5, X~BIN (n,p)* can be approximated by *X~N(np,npq)*

$$E(X) = E(n\hat{p}) = nE(\hat{p}) = np$$

$$Var(X) = Var(n\hat{p}) = n^2 Var(\hat{p}) = \frac{n^2 pq}{n} = npq$$

Example 6: Sampling Distribution of Sample Proportion

- Suppose that the manager of the local bank determines that 40% of all depositors have multiple accounts at the bank. If you select a random sample of 200 depositors, what is the probability that the sample proportion of depositors with multiple accounts is less than 0.3 ?
- Let *p* and $\hat{p}$ denote the population proportion and sample proportion of depositors with multiple accounts respectively
- *n*=200, $p = 0.4$. As $np = 200(0.4) = 80 > 5, n(1-p) = 200(0.6) = 120 > 5$ → The sampling distribution of sample proportion follows Normal distribution approximately

$$\hat{p} \sim N(p, \frac{pq}{n})$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

$$P(\hat{p} < 0.3) = \mathrm{P}\left(\frac{\hat{p} - 0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right)$$

$$= \mathrm{P}\left(\mathrm{Z} < \frac{0.3-0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right) = P(Z < -2.89) = 0.0019$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026