INTRODUCTION TO STATISTICAL MODELING

STAT2507D

Chapter 7 – 2

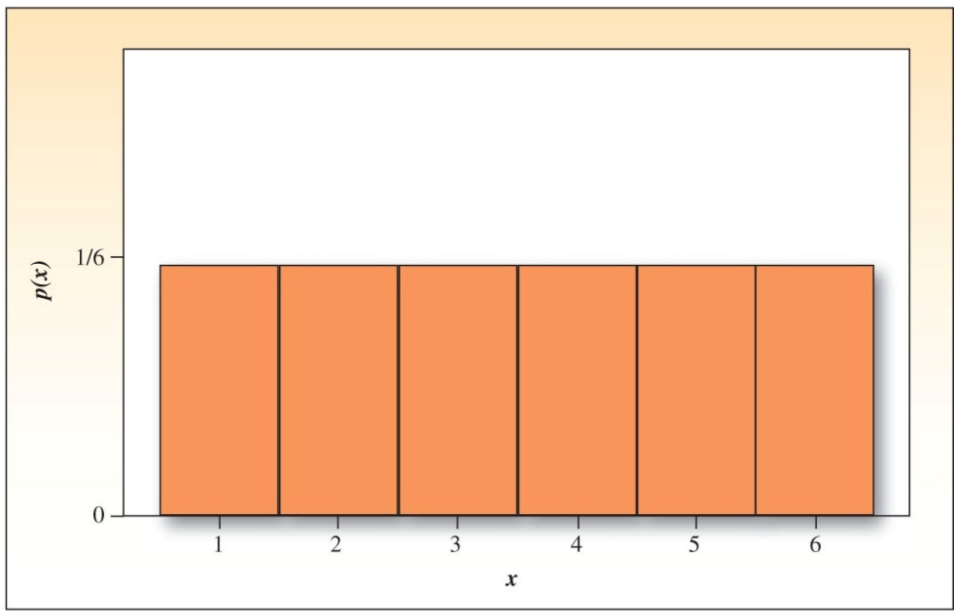
Sampling Distributions

THE CENTRAL LIMIT THEOREM

- Sampling distributions for statistics can be
  - Approximated with simulation techniques
  - Derived using mathematical theorems
- The central Limit Theorem is one such theorem
- Example: We will observe the sampling distribution of average value (x) of numbers observed on the top face of a die when tossed n times ($n = 1, 2, 3, 4, \cdots$).

THE CENTRAL LIMIT THEOREM

- Toss a fair die $n = 1$ time; the distribution of x the number on the upper face is flat or uniform



EXAMPLE CONT'D

$$\mu = \sum xp(x)$$

$$= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = 3.5$$

$$\sigma = \sqrt{\sum (x-\mu)^2 p(x)}$$

$$= \sqrt{(1-3.5)^2\left(\frac{1}{6}\right) + \cdots + (6-3.5)^2\left(\frac{1}{6}\right)} = 1.71$$

EXAMPLE CONT'D

- Toss a fair die n = 2 times. Total of numbers observed on the top faces of two dices

	First Die					
Second Die	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

EXAMPLE CONT'D

- Average of numbers observed on the top faces of two dice

Second Die	First Die					
	1	2	3	4	5	6
1	1	1.5	2	2.5	3	3.5
2	1.5	2	2.5	3	3.5	4
3	2	2.5	3	3.5	4	4.5
4	2.5	3	3.5	4	4.5	5
5	3	3.5	4	4.5	5	5.5
6	3.5	4	4.5	5	5.5	6

Example cont'd

X (Average on two dice)	P(X)
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36

EXAMPLE CONT'D

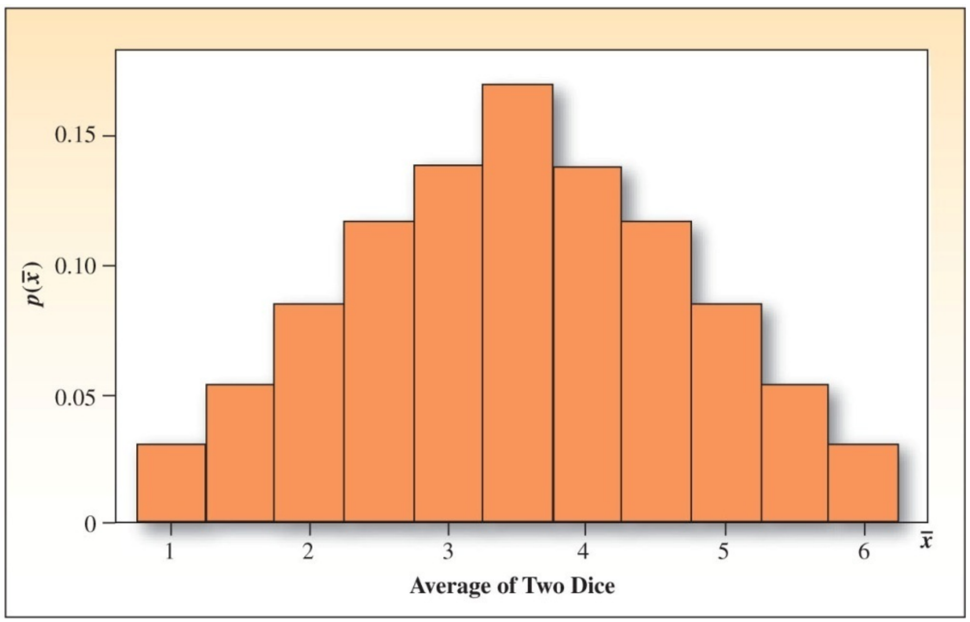
- The distribution of x the average number on the two upper faces is mound-shaped

$$Mean\text{: } \mu = 3.5$$

$$StdDev\text{:} \frac{\sigma}{\sqrt{2}} = \frac{1.71}{\sqrt{2}} = 1.21$$

EXAMPLE CONT'D

- Distribution of average of numbers observed on the face of the dice when n=2.



EXAMPLE CONT'D

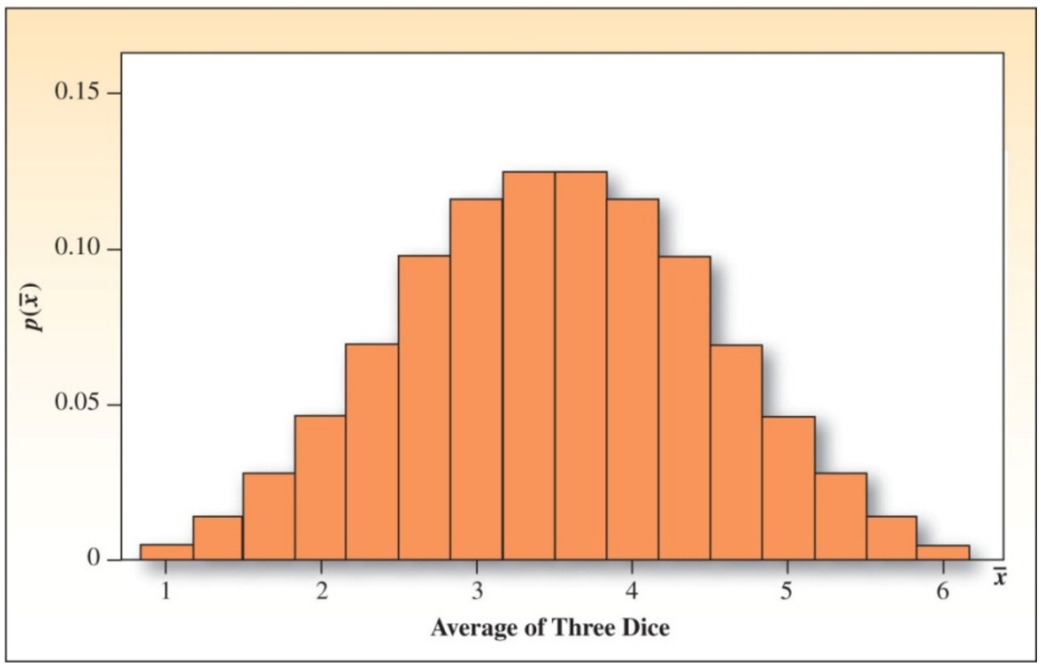
- Toss a fair die *n* = 3, 4 times; the distribution of *x* the average number on the two upper faces is **approximately normal**

$$Mean\text{: } \mu = 3.5$$

$$StdDev\text{:} \frac{\sigma}{\sqrt{3}} = \frac{1.71}{\sqrt{3}} = 0.987$$

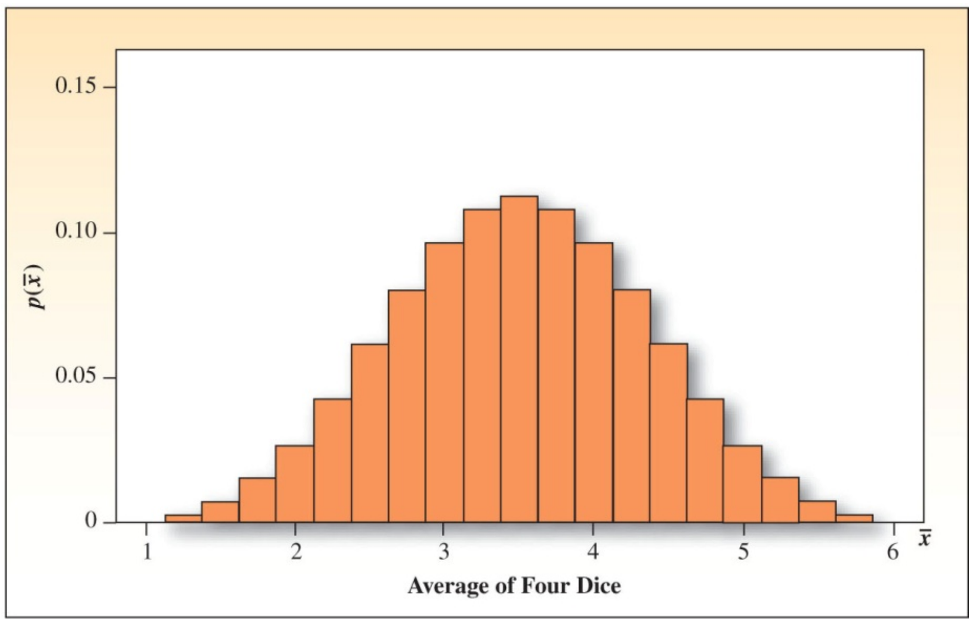
EXAMPLE CONT'D

- Distribution of average of numbers observed on the face of the dice when n=3.



EXAMPLE CONT'D

- Distribution of average of numbers observed on the face of the dice when n=4



EXAMPLE CONT'D

- Summary of mean and standard deviation for $n = 1,2,3,4,5$ in a die tossing experiment.

n	Mean	Standard deviation
1	3.5	$\frac{1.71}{\sqrt{1}} = 1.71$
2	3.5	$\frac{1.71}{\sqrt{2}} = 1.21$
3	3.5	$\frac{1.71}{\sqrt{3}} = 0.99$
4	3.5	$\frac{1.71}{\sqrt{4}} = 0.86$
5	3.5	$\frac{1.71}{\sqrt{5}} = 0.77$

REMARKS

- Regardless of its shape, the sampling distribution of $\bar{x}$ always has a mean identical to the mean of sampled population and standard deviation equal to the population standard deviation σ divided by $\sqrt{n}$.

$$E(\bar{X}) = \mu; Var(\bar{X}) = \sigma/\sqrt{n}$$

- Consequently, the spread of the distribution of sample mean is considerably less than the spread of the sampled population.

Central Limit Theorem

- If a random samples of n observations are drawn from a non-normal population with finite μ and standard deviation σ, then, when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The approximation becomes more accurate as n becomes large.

CENTRAL LIMIT THEOREM

- Central limit theorem can restated to apply to the sum of the sample measurements $\sum_{i=1}^{n} x_i$, which as n becomes large, also has an approximately normal distribution with mean nμ and standard deviation $\sqrt{n}\sigma$.

Central Limit Theorem

- How large sample size is needed
  - If the sampled population is normal, then the sampling distribution of $\bar{x}$ will also be normal regardless of the sample size
  - When the sample population is approximately symmetric, the sampling distribution of $\bar{x}$ becomes approximately normal for relatively small values of n.

CENTRAL LIMIT THEOREM

- How large sample size is needed
  - When sampled population is skewed, the sample size n must be larger, with n at least 30 before the sampling distribution of $\bar{x}$ becomes approximately normal.

POINTS TO CONSIDER WHEN CHOOSING ESTIMATOR OF μ

- If the population mean μ is unknown, several statistics can be used as an estimator
  - Sample mean
  - Sample median

POINTS TO CONSIDER WHEN CHOOSING ESTIMATOR OF μ

- Criteria for choosing the estimator for μ
  - Is it easy or hard to calculate $\bar{x}$?
  - Does it produce estimates that are consistently too high or too low?
  - Is it more or less variable than other possible estimators?
- In many situations, sample mean has desirable properties as an estimator.

THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

- A random sample of size n is selected from a population with mean μ and standard deviation σ
- The sampling distribution of the sample mean $\bar{x}$ will have mean μ and standard deviation $\sigma/\sqrt{n}$

THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

- The standard deviation of $\bar{x}$ is sometimes called the STANDARD ERROR of the mean (SE or SE($\bar{x}$))
- Probabilities are calculated using the standard normal random variable

$$Z = \frac{Estimator - Mean}{Standard\ Error}$$

FINDING PROBABILITIES FOR SAMPLE MEAN

- If the sampling distribution of $\bar{x}$ is normal or approximately normal
  - Find μ and calculate $SE(\bar{x}) = \sigma/\sqrt{n}$
  - Write down the event of interest in terms of $\bar{x}$, and locate the appropriate area on the normal curve
  - Convert the necessary values of $\bar{x}$ to z-values using
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$
  - Use the normal table to calculate the probability

EXERCISE

7.2) A random sample of size n =40 is selected from a population with mean μ =100 and standard deviation σ = 20

a) What will be the approximate shape of the sampling distribution of $\bar{x}$?
b) What will be the mean and standard deviation of the sampling distribution of $\bar{x}$?
c) Find the probability that the sample mean is between 105 and 110?
d) What is the probability that sample mean exceeds 140?

EXERCISE

7.3)Suppose that university faculty with the rank of assistant professor earn an average of \$74,000 per year with a standard deviation of \$6000. In an attempt to verify this salary level, a random sample of 60 assistant professors was selected from a personnel database for all universities in Canada.

EXERCISE

a) Describe the sampling distribution of the sample mean
b) Within what limits would you expect the sample average to lie, with probability 0.95?
c) Calculate the probability that the sample mean $\bar{x}$ is greater than \$78000
d) If your random sample actually produced a sample mean of \$78000, would you consider this unusual? What conclusion might you draw?

EXERCISE

7.4) The maximum load (with a generous safety factor) for the elevator in an office building is 900 kg. The relative frequency distribution of the weights of all men and women using the elevator is mound-shaped with mean μ=65 kg and σ=16kg. What is the largest number of people you can allow on the elevator if you want their total weight to exceed the maximum weight with a small probability (say, near 0.01)?

THE SAMPLING DIST'N OF THE SAMPLE PROPORTION

- The central limit theorem can be used to conclude that the binomial random variable x is approximately normal when n is large, with mean np and standard deviation $\sqrt{np(1-p)}$
- Requirement: $np > 5 \; and \; nq > 5$

THE SAMPLING DIST'N OF THE SAMPLE PROPORTION

- The sample proportion, $\hat{p} = x/n$ is simply a rescaling of the binomial random variable x, dividing it by n
- From central limit theorem, the sampling distribution of $\hat{p}$ will also be approximately normal, with rescaled mean and standard deviation

THE SAMPLING DIST'N OF THE SAMPLE PROPORTION

- A random sample of size n is selected from a binomial population with parameter p
- The sampling distribution of the sample $\hat{p}$ proportion will have mean p and standard deviation $\sqrt{pq/n}$
- If n is large, and p is not too close to zero or one, the sampling distribution of $\hat{p}$ will be approximately normal

FINDING PROBABILITIES FOR SAMPLE PROPORTION

- Find the necessary values of n and p
- Check whether the normal approximation to the binomial distribution is appropriate ($np > 5$ and $nq > 5$)
- Write down the event of interest in terms of $\widehat{p}$, and locate the appropriate area on the normal curve

FINDING PROBABILITIES FOR SAMPLE PROPORTION

- Convert the necessary values of $\hat{p}$ to z-values using

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

- Use normal table to calculate the probability

EXERCISE

7.5) Random sample of size n = 75 were selected from binomial population with p=0.4. Use normal distribution to approximate the following probabilities

a. $P(\hat{p} \leq 0.43)$

b. $P(0.35 \leq \hat{p} \leq 0.43)$

EXERCISE

- 7.6) A soda bottler claims that only 5% of the soda cans are under filled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are under filled?

SUMMARY

- Sampling Plans and Experimental Design
  - Survey
    - Survey Errors
    - Questionnaire design
    - Survey delivery
  - Simple random sampling

SUMMARY

- Sampling Plans and Experimental Design
  - Other sampling plans
    - Stratified sampling
    - Cluster sampling
    - Systematic 1-in-k sampling
  - Non-random sampling: Convenience sampling, Judgement sampling, Quota sampling

SUMMARY

- Statistics and Sampling Distributions
  - Sampling distributions describe the possible values of a statistic and how often they occur in repeated sampling
  - Sampling distributions can be derived mathematically, approximated empirically, or found using statistical theorems

SUMMARY

- Statistics and Sampling Distributions
  - The Central Limit Theorem states that sums and averages of measurements from a non-normal population with finite mean and standard deviation have approximately normal distributions for large samples of size n.

SUMMARY

- Sampling Distribution of the sample mean
  - When samples of size *n* are drawn from a normal population with mean μ and variance σ^2, the sample mean has a normal distribution with mean μ and variance $\sigma/\sqrt{n}$

SUMMARY

- Sampling Distribution of the sample mean
  - When samples of size *n* are drawn from a non-normal population with mean μ and variance σ^2, the Central Limit Theorem ensures that the sample mean $\bar{x}$ will have an approximately normal distribution with mean μ and variance σ^2/n when n is large ($n \geq 30$)
  - Probabilities involving the sample mean can be calculated by standardizing the value of $\bar{x}$ using $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

SUMMARY

- Sampling Distribution of the sample proportion
  - When samples of size *n* are drawn from a binomial population with parameter *p*, the sample proportion will have an approximately normal distribution with mean *p* and varian$\hat{p}$ce pq/n as long as $np > 5$ and $nq > 5$
  - Probabilities involving the sample proportion can be calculated by standardizing the value $\hat{p}$ using

SUMMARY

- Sampling Distribution of the sample proportion
  - Probabilities involving the sample proportion can be calculated by standardizing the value $\hat{p}$ using

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$