Statistics ST2334 Topic 6: Sampling Distributions (part a)

Random Samples, Sample Distribution of Sample mean, Law of Large Numbers, Central Limit Theorem.

2011/2012 Semester 2

Recap: Population and Parameters

- Population can be finite or infinite
  - Infinite: impossible to observe all its values
  - Finite: may be impractical or uneconomical to observe all its values.
- Example:
  - Infinite population: products from a production line; outcomes of flipping a coin;
  - Finite population: monthly income of Singaporean. products of a factory in one day. scores of students in NUS.
- Numerical descriptive measures of a population are called **parameters** e.g., p, μ, σ.

Known population distribution with unknown parameters

- We often know the population belongs to a known family of distributions.
- However, the values of parameters that specify the distribution are unknown.
- Examples:
  - A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
  - An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.

Recap: Sampling

- You must rely on the **sample** to learn about these parameters and study the properties of the population.
- The sample should be representative of population. We have different types of sampling schemes attempting to do that.
- For the probability methods, it is possible to fully describe the quantitative properties of the sample.
- We will focus on the simple random sample. The textbook calls it a "random sample".

Definition of Random Sample

A set of observations $X_1, X_2, ..., X_n$ constitute a random sample of size n from

- a finite population of size N, if its values are chosen so that each subset of n, the N elements of the population has the same probability of being selected.
- an infinite population with distribution f(x) if
  1. each X_i is a random variable whose distribution is given by f(x).
  2. These n random variables are independent.

Estimator of μ: $\bar{X}$

- Suppose a random sample of n observations, i.e. $X_1, X_2, ..., X_n$, has been taken.
- We estimate the population mean μ by the sample mean $\bar{X}$.
  - X_1 is a random variable and so are $X_2, ..., X_n$.
  - $\bar{X}$ is a random variable as well.

Example: What does $\bar{X}$ look like?

- Let's look at bus waiting time again.
- Assume that the waiting time for each student every morning has the same uniform distribution with $\alpha = 0, \beta = 9$. Assume further that they are all independent.
- For the convenience of recording, only mins are reported. i.e. when I ask a students waiting time for the bus in a morning, the student answers 0min or 1min or 2mins, ...,
- Therefore, denote by X the waiting time for the bus in a morning. Then, $P(X = i) = 1/10$, for each $i = 0, 1, 2, ..., 9$.
- So we know X completely, but let's say we're interested in the average waiting time $\bar{X}$ for each student instead.

Empirical Sampling Distribution of the Mean

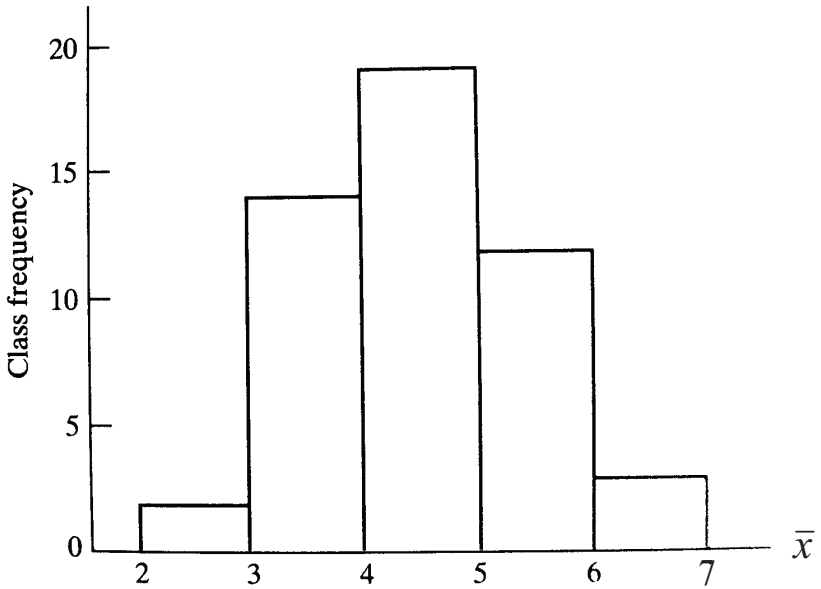
- I requested 50 randomly selected students to report their waiting time for a bus each morning in 10 randomly selected mornings over the semester.
- For each student, 10 observations $X_1, X_2, ..., X_{10}$, are obtained. The sample mean $\bar{X}$ is then computed.
- Now that we obtain 50 sample means (50 students) as follows

4.4	3.2	5.0	3.5	4.1	4.4	3.6	6.5	5.3	4.4
3.1	5.3	3.8	4.3	3.3	5.0	4.9	4.8	3.1	5.3
3.0	3.0	4.6	5.8	4.6	4.0	3.7	5.2	3.7	3.8
5.3	5.5	4.8	6.4	4.9	6.5	3.5	4.5	4.9	5.3
3.6	2.7	4.0	5.0	2.6	4.2	4.4	5.6	4.7	4.3

Empirical Sampling Distribution of the Mean

- Let's see what that looks like

$\bar{X}$	[2.0, 3.0]	[3.0, 4.0]	[4.0, 5.0]	[5.0, 6.0]	[6.0, 7.0]
frequency	2	14	19	12	3



Theoretical Sampling Distribution of $\bar{X}$

- If a random sample of size n is taken from a population with mean μ and variance σ^2, then $\bar{X}$ is a random variable with mean μ and
  - For samples from an infinite population, its variance is

$$\frac{\sigma^2}{n}.$$

  - For samples from a finite population with size N, its variance is

$$\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1},$$

where $\frac{N-n}{N-1}$ is called finite population correction factor

Validity of $\bar{X}$ as an estimator for μ

- The expectation of $\bar{X}$ is equal to the population mean μ.
- In "the long run", $\bar{X}$ does not introduce any systematic bias as an estimator of μ.
- $\bar{X}$ can serve as a valid estimator of μ.
- for infinite population, when n gets larger and larger, σ^2/n, the variance of $\bar{X}$, becomes smaller and smaller, that is, the accuracy of $\bar{X}$, as an estimator of μ keeps improving.
- for finite population, similar arguments apply, if the sample constitutes a substantial proportion of the population.

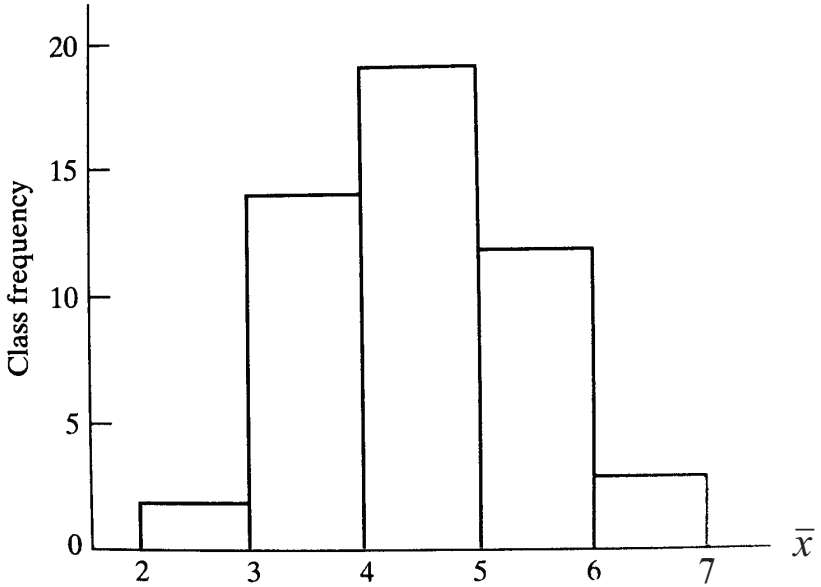
Law of Large Numbers (L.L.N.)

- If $X_1, X_2, ..., X_n$ are independent random variables with the same mean μ and variance σ^2 , then for any $\epsilon \in \mathbb{R}$,

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

- This is saying as the sample size increase, the probability that the sample mean is away from the population mean goes to zero.
- Or another way of looking at this is that it is increasingly likely that $\bar{X}$ is close to μ, as n gets larger.

Back to the bus example



- We see that even just for $n = 10$ it is likely that $\bar{X}$ is close to $\mu = 4.5$.
- Notice also, the histogram looks bell-shaped. Is this always the case?

Sample Mean of Normally Distributed Population

- $\{X_1, X_2, ..., X_n\}$ is a random sample from $N(\mu, \sigma^2)$.
- We know $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$.
- Fact: Linear combinations of normal random variables are themselves normal.
- Thus, $\bar{X}$ has distribution: $N(\mu, \sigma^2/n)$.
- and

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

- But our bus waiting time wasn't normal! For that we need.....

Central Limit Theorem (C.L.T.)

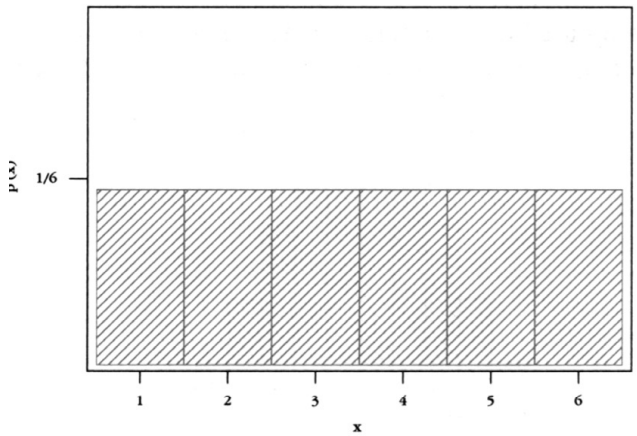
- If $\bar{X}$ is the mean of a random sample of size n taken from a population having mean μ and finite variance σ^2, then, as $n \to \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to Z \sim N(0, 1)$$

- The C.L.T. states that, under rather general conditions, for large n, sums and means of random samples drawn from a population follows the normal distribution closely.
- The bus waiting time was uniform, but with just $n = 10$, the distribution of $\bar{X}$ is close to normal.

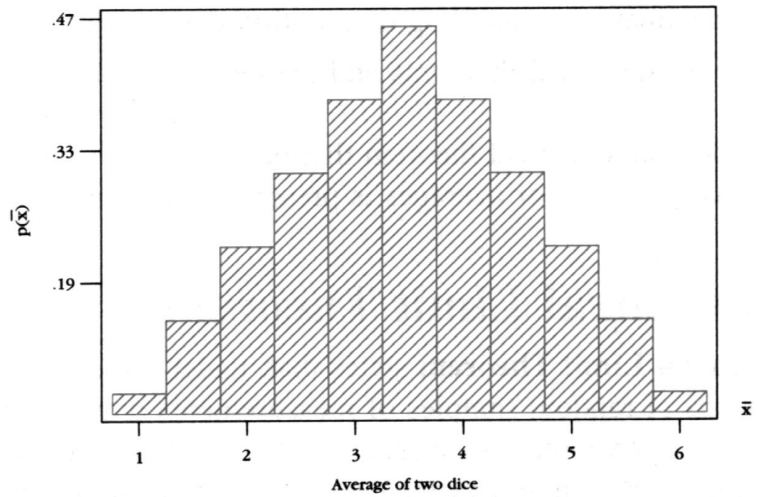
Example: Die Tossing

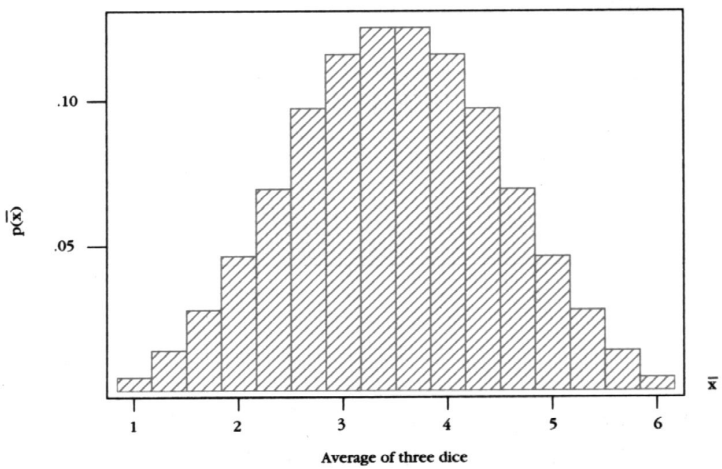
- The figure below shows the probability distribution of X, the number appearing on a single toss of a die.

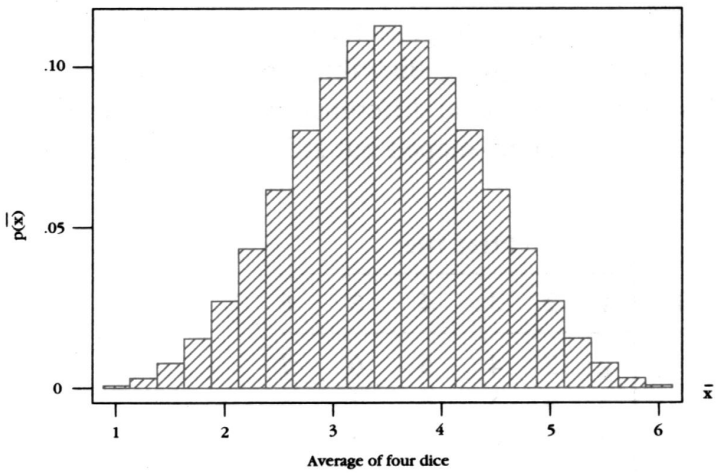


- Let's look at the sampling distributions of $\bar{X}$. That is the average of n dice rolls.
- From the figures below, we can see that the distribution of $\bar{X}$ gets more and more close to normal, as n gets larger and larger. The convergence speed is fairly high: $n = 4$, it is close enough to normal.

Example: Sampling Distribution of $\bar{X}$ for $n = 2, 3, 4$ dice







CLT: Rule of Thumb

- In practice, the normal distribution provides an excellent approximation to the sampling distribution of the mean $\bar{X}$ if $n \geq 30$.
- Like the dice example, for some distributions, the sampling distribution of $\bar{X}$ quickly converges to normal much sooner.
- Note also, if the random sample come from a normal population, $\bar{X}$ is normally distributed regardless of the value of n.

Example: Bowling League

In a bowling league season, bowlers bowl 50 games and the average score is ranked at the end of the season. Historically, Newt averages 175 a game with a standard deviation of 30. What is the probability that Newt will average more than 180 this season?

- Note $\mu = 175$, $\sigma = 30$ and $n = 50$. Let $\bar{X}$ be the sample mean. We need to find $P(\bar{X} > 180)$.
- By CLT, we can approximate $\bar{X}$ by $N(\mu, \sigma^2/n)$. Hence

$$\begin{aligned} P(\bar{X} > 180) &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{180 - \mu}{\sigma/\sqrt{n}}) \\ &= P(Z > 1.18) = 0.119 \end{aligned}$$