# Statistics I – Chapter 7
# Sampling Distributions (Part 1)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

November 14, 2012

# Introduction

- In this chapter, we will study **sampling techniques** and **sampling distributions**.
  - Different sampling techniques may be applied in different environments.
  - Once we obtain a statistic, we need to know its distribution to understand its behavior and make inferences.
- Two particular statistics we will study in this chapter are the **sample mean** and **sample proportion**.
- The **central limit theorem** is the foundation of many statistical inference processes.

# Road map

- **Sampling techniques**.
- Sampling distributions.
- Distribution of the sample mean.

# Sampling vs. census

- We have compared three pairs of concepts in Chapter 1:
  - Populations vs. samples.
  - Parameters vs. statistics.
  - Census vs. **sampling**.
- If we can always conduct a census, we will not need statistical inferences at all. So **why** sampling?
  - Saving money and time.
  - More detailed information under the same resources.
  - Destructive research processes.
  - Impossibility of a census.

# Frames

- When sampling from a population, we need a **list**, **map**, **directory**, or some other sources that represent the population.
- Such a source is called a **frame**.
  - A list of all students in NTU.
  - A list of all professors in Taiwan.
  - A list of all telephone numbers registered in Taipei.
- A frame may not be 100% accurate.
  - Frames with **overregistration** contain the target population plus some additional units.
  - Frames with **underregistration** have some units missing.

# Random vs. nonrandom sampling

- Sampling is the process of selecting a **subset** of entities from the whole population.
- Sampling can be **random** or **nonrandom**.
- If random, whether an entity is selected is **probabilistic**.
  - Randomly select 1000 phone numbers on the telephone book and then call them.
- If nonrandom, it is **deterministic**.
  - Ask all your classmates for their preferences on iOS/Android.
- Most statistical methods are **only** for random sampling.

# Random sampling techniques

- We will introduce four basic random sampling techniques:
  - Simple random sampling.
  - Stratified random sampling.
  - Systematic random sampling.
  - Cluster (or area) random sampling.

# Simple random sampling

- In simple random sampling, each entity has **the same probability** of being selected.
- Each entity is assigned a label (from 1 to $N$). Then a sequence of $n$ random numbers, each between 1 and $N$, are generated.
- One needs either a **table of random numbers** or a **random number generator**.
  - A table with many random numbers.
  - A software function that generate random numbers.

# Simple random sampling

- Suppose we want to study all students graduated from NTU IM regarding the number of units they took before their graduation.
  - $N = 1000$.
  - For each student, whether she/he double majored, the year of graduation, and the number of units are recorded.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | ... | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Double major | Yes | No | No | No | Yes | No | No | | Yes |
| Class | 1997 | 1998 | 2002 | 1997 | 2006 | 2010 | 1997 | ... | 2011 |
| Unit | 198 | 168 | 172 | 159 | 204 | 163 | 155 | | 171 |

  - Suppose we want to sample $n = 200$ students.

# Simple random sampling

- To run simple random sampling, we first generate a sequence of 200 random numbers:
  - Suppose they are 2, 198, 7, 268, 852, ..., 93, and 674.
- Then the corresponding 200 students will be sampled. Their information will then be collected.

| $i$ | 1 | **2** | 3 | 4 | 5 | 6 | **7** | ... | 1000 |
|---|---|---|---|---|---|---|---|---|---|
| Double major | Yes | **No** | No | No | Yes | No | **No** | | Yes |
| Class | 1997 | **1998** | 2002 | 1997 | 2006 | 2010 | **1997** | ... | 2011 |
| Unit | 198 | **168** | 172 | 159 | 204 | 163 | **155** | | 171 |

# Simple random sampling

- The good part of simple random sampling is **simple**.
- However, it may result in **nonrepresentative** samples.
- In simple random sampling, there are some possibilities that **too much** data we sample fall in **the same stratum**.
  - They have the same property.
  - For example, it is possible that all 200 students in our sample did not double major.
  - The sample is thus nonrepresentative.

# Simple random sampling

- As another example, suppose we want to sample 1000 voters in Taiwan regarding their preferences on two candidates. If we use simple random sampling, what may happen?
  - It is possible that 65% of the 1000 voters are men while in Taiwan only around 51% voters are men.
  - It is possible that 40% of the 1000 voters are from Taipei while in Taiwan only around 28% voters live in Taipei.
- How to fix this problem?

# Stratified random sampling

- We may apply **stratified random sampling**.
- We first split the whole population into several **strata**.
  - Data in **one** stratum should be (relatively) **homogeneous**.
  - Data in **different** strata should be (relatively) **heterogeneous**.
- We then use simple random sampling for each stratum.
- Suppose 100 students double majored, then we can split the whole population into two strata:

| Stratum | Strata size |
|---|---|
| Double major | 100 |
| No double major | 900 |

# Stratified random sampling

- Now we want to sample 200 students.
- If we sample $200 \times \frac{100}{1000} = 20$ students from the double-major stratum and 180 ones from the other stratum, we have adopted **proportionate stratified random sampling**.

| Stratum | Strata size | Number of samples |
|---|---|---|
| Double major | 100 | 20 |
| No double major | 900 | 180 |

- If the opinions in some strata are more important, we may adopt **disproportionate stratified random sampling**.
  - E.g., opening a nuclear power station at a particular place.

# Stratified random sampling

- We may further split the population into more strata.
  - Double major: Yes or no.
  - Class: 1994-1998, 1999-2003, 2004-2008, or 2009-2012.
  - This stratification makes sense **only if** students in different classes tend to take different numbers of units.
- Stratified random sampling is typically good in **reducing sample error**.
- But it can be hard to identify a reasonable stratification.
- It is also more **costly** and **time-consuming**.

# Systematic random sampling

- When even simple random sampling is too time-consuming, we may use **systematic random sampling**.
  - In simple random sampling, we need **at least** $n$ different random numbers.
  - In systematic random sampling, we need only **one**.
- We first determine a number $k$:

$$k = \left\lfloor \frac{N}{n} \right\rfloor.$$

- Then we generate one random number $s \in \{1, 2, ..., k\}$.
- The data we will sample are those with labels $s$, $s + k$, $s + 2k$, ..., and $s + nk$.

# Systematic random sampling

- As we want to sample $n = 200$ students from $N = 1000$ students, $k = \left\lfloor \frac{1000}{200} \right\rfloor = 5$.
- Suppose the random number is $s = 3$.
- Then we will sample:

| $i$ | 3 | 8 | 13 | 18 | 23 | 28 | ... | 993 | 998 |
|---|---|---|---|---|---|---|---|---|---|
| Double major | No | No | No | Yes | No | No | | No | Yes |
| Class | 2002 | 2000 | 1997 | 1998 | 2002 | 2005 | ... | 1999 | 2001 |
| Unit | 172 | 168 | 155 | 156 | 171 | 159 | | 180 | 183 |

# Systematic random sampling

- Systematic random sampling is **extremely simple**.
- In some cases, its quality is not lower than that of simple random sampling.
- However, if the data are labeled base on some periodicity and the sampling is in a similar **periodicity**, there will be a huge sample error.
- Also the possible outcomes of sampling is quite limited.

# Cluster (or area) random sampling

- Imagine that you are going to introduce a new product into all the retail stores in Taiwan.
- If the product is actually unpopular, an introduction with a large quantity will incur a huge lost.
- How to get an idea about the popularity?
- Typically we first try to introduce the product **in a small area**. We put the product on the shelves only in those stores in the specified area.
- This is the idea of **cluster (or area) random sampling**.
  - Those consumers in the area form a sample.

# Cluster (or area) random sampling

- In stratified random sampling, we define strata.
- Similarly, in cluster random sampling, we define **clusters**.
- However, instead of doing simple random sampling in each strata, we will only choose **one or some clusters** and then collect **all** the data in these clusters.
  - If a cluster is too large, we may further split it into multiple **second-stage clusters**.
- Therefore, we want data in a cluster to be **heterogeneous**.

# Cluster (or area) random sampling

- In the example of sampling 200 students, we may define clusters based on classes.
- Then we randomly select four classes and sample the 200 students in the four classes.
- This may or may not be representative.
  - Do students in a single class tend to be heterogeneous?

# Cluster (or area) random sampling

- In practice, the main application of cluster random sampling is to understand the popularity of **new products**. Those chosen cities (counties, states, etc.) are called **test market cities** (counties, states, etc.).
- People use cluster random sampling in this case because of its feasibility and convenience.
  - Is it easy to deliver the product to consumers selected by the other random sampling techniques?
- We should select test market cities whose population profiles are similar to that of the entire country.

# Nonrandom sampling

- **Convenience sampling**.
  - The researcher sample data that are easy to sample.
- **Judgment sampling**.
  - The researcher decides who to ask or what data to collect.
- **Quota sampling**.
  - In each stratum, we use whatever method that is easy to fill the quota, a predetermined number of samples in the stratum.
- **Snowball sampling**.
  - Once we ask one person, we ask her/him to suggest others.
- Nonrandom sampling **cannot** be analyzed by the statistical methods we introduce in this course.

# Road map

- Sampling techniques.
- **Sampling distributions**.
- Distribution of the sample mean.

# Distributions

- To describe a **random** variable or an experiment, we need to specify two things,
  - all the possible **outcomes** and
  - the **probability** (**density**) for each outcome to occur.
- E.g., rolling a dice:

| Outcome | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Probability | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

# Distributions

- E.g., drawing one ball from a box containing three white balls and two black balls.

| Outcome | White | Black |
|---|---|---|
| Probability | $\frac{2}{5}$ | $\frac{2}{3}$ |

- E.g., drawing two balls from a box containing three white balls and two black balls.

| Outcome | White and white | White and black | Black and black |
|---|---|---|---|
| Probability | $(\frac{2}{5})^2 = \frac{4}{25}$ | $2(\frac{2}{5})(\frac{3}{5}) = \frac{12}{25}$ | $(\frac{3}{5})^2 = \frac{9}{25}$ |

# Distributions

- Suppose we are facing a population and we want to randomly draw one item.
  - E.g., rolling a dice: Population $= \{1, 2, 3, 4, 5, 6\}$; the probability of drawing each of them is $\frac{1}{6}$.
- The outcome of "**drawing one item**" is certainly random.
- Suppose we are facing a population and we want to randomly draw $n < N$ item.
  - E.g., rolling $n$ dice: The outcome is an $n$-dimensional vector $(X_1, X_2, ..., X_n)$, where $X_i \in \{1, 2, 3, 4, 5, 6\}$ is the outcome of the $i$th dice.
- The outcome of "**drawing $n < N$ items**" is also random.

# Sampling distributions

- The outcome of drawing $n$ items forms a **sample**.
- A sample with $n > 1$ and $n < N$ is a **random vector**.
- The distributions of samples are **sampling distributions**.
- In Statistics, we typically do not care about the distributions of a sample directly. Instead, we care about the distribution of a **statistic**, which is a **function of the sample**.
  - A sample: $(X_1, X_2, ..., X_n)$.
  - A statistic: the **sample mean**: $\overline{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$.
  - Other statistics: the sample variance, sample median, sample range, sample max, etc.

# Sampling distributions

- The distributions of statistics, as they are derived from the distributions of samples, are also called sampling distributions.
- The reason to care about sampling distributions:
  - We will use a statistic to **infer** a **parameter**.
  - We can **scientifically** describe or estimate the parameter only if we know the distribution of the statistic.
- Some concrete examples will be given in Chapters 8 and 9.
- In Chapter 7, let's derive some sampling distributions.

# Sampling distributions

- What are those sampling distributions we will derive?
- In Chapter 7 of the textbook:
  - Sample mean.
  - Sample proportion.
- In Chapter 8 of the textbook:
  - Sample variance.
- Outside the textbook:
  - Sample minimum.
- Before we derive those distributions, let's first get more general ideas about sampling distributions.

# Sampling distributions of rolling dices

- We know how to describe the experiment of rolling a dice:

| Outcome | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Probability | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |

- Suppose we roll a dice twice. How to describe this?

| Outcome | $(1,1)$ | $(1,2)$ | $(1,3)$ | $\cdots$ | $(6,5)$ | $(6,6)$ |
|---|---|---|---|---|---|---|
| Probability | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |

# Sampling distributions of rolling dices

- Let
  - $X_1$ be the outcome of rolling **the first dice** and
  - $X_2$ be the outcome of rolling **the second dice**.
- We have derived the distributions of $X_1$ and $(X_1, X_2)$.
- What is the distribution of $X_1 + X_2$?
- First we need to have the set of all possible outcomes:
  - $\{2, 3, 4, ..., 11, 12\}$.
- Then we need to know the probability for each outcome to occur. How?

# Distributions of sum of two dices

- The distribution of $X_1 + X_2$ comes from that of $(X_1, X_2)$.
  - For the outcome 2, we have

$$\begin{aligned}\Pr(X_1 + X_2 = 2) &= \Pr(X_1 = 1, X_2 = 1) \\ &= \Pr(X_1 = 1)\Pr(X_2 = 1) = \tfrac{1}{36}.\end{aligned}$$

  - For the outcome 3, we have

$$\begin{aligned}&\Pr(X_1 + X_2 = 3) \\ = \ &\Pr(X_1 = 1, X_2 = 2 \cup X_1 = 2, X_2 = 1) \\ = \ &\Pr(X_1 = 1, X_2 = 2) + \Pr(X_1 = 2, X_2 = 1) \\ = \ &\Pr(X_1 = 1)\Pr(X_2 = 2) + \Pr(X_1 = 2)\Pr(X_2 = 1) = \tfrac{2}{36}.\end{aligned}$$

- The probabilities of all outcomes can be derived similarly.

# Distributions of sum of two dices

- It may be easier to look at the table:

| $X_1$ | $X_2$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | $\left\{\frac{1}{36}\right\}$ | $\left[\frac{1}{36}\right]$ | $\left(\frac{1}{36}\right)$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 2 | $\left[\frac{1}{36}\right]$ | $\left(\frac{1}{36}\right)$ | $\frac{1}{36}$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 3 | $\left(\frac{1}{36}\right)$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 4 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 5 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |
| 6 | $\frac{1}{36}$ | $\frac{1}{36}$ | $\frac{1}{36}$ | $\cdots$ | $\frac{1}{36}$ | $\frac{1}{36}$ |

- { }: $X_1 + X_2 = 2$; [ ]: $X_1 + X_2 = 3$; ( ): $X_1 + X_2 = 4$.

# Distributions of sum of two dices

- The distribution of sum of two dices, $X_1 + X_2$, is:

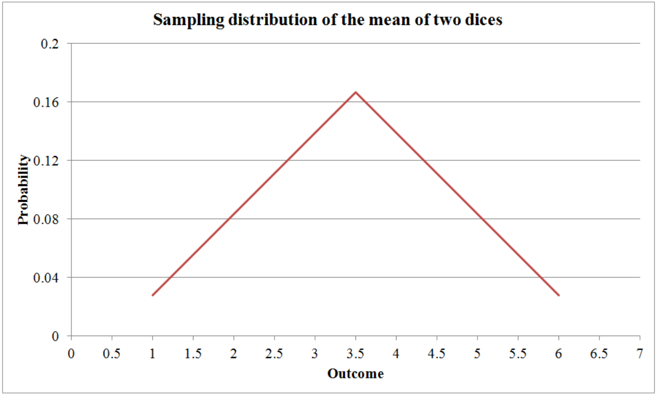
| Outcome | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Probability | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

- It then follows that the distribution of the sample mean of sample size 2, $\frac{1}{2}(X_1 + X_2)$, is:

| Outcome | 1 | $\frac{3}{2}$ | 2 | $\frac{5}{2}$ | 3 | $\frac{7}{2}$ | 4 | $\frac{9}{2}$ | 5 | $\frac{11}{2}$ | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Probability | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

# Distributions of sum of two dices

- The distribution of the sample mean of sample size 2:




- Why most occurrences are around the mean?

# Sampling distributions

- The distribution of $X_1$ or $X_2$ is a population distribution.
  - Or a sampling distribution with sample size 1.
- The distributions of $(X_1, X_2)$, $X_1 + X_2$, and $\frac{1}{2}(X_1 + X_2)$ are sampling distributions.
- **Analytically**, we may derive the distribution of the sample mean of rolling $n$ dices for any $n \in N$.
  - Nevertheless, the derivation will be tedious and costly for large sample sizes and general population distributions.
- To make our lives easier and to give you some ideas about random sampling, let's find the distributions **numerically**:
  - Roll dices for many times and then draw a histogram.

# Numerical sampling distributions

- Let's do the experiment of rolling two dices for 500 times.
- Think in this way:
  - Tomorrow I will roll two dices and get $\overline{X}^1 = \frac{1}{2}(X_1^1 + X_2^1)$.
  - Two days later I will do it again and get $\overline{X}^2 = \frac{1}{2}(X_1^2 + X_2^2)$.
  - Three days later I will get $\overline{X}^3 = \frac{1}{2}(X_1^3 + X_2^3)$.
  - 500 days later I will get $\overline{X}^{500} = \frac{1}{2}(X_1^{500} + X_2^{500})$.
- Each of $X^i$s is a **sample**. At this time, they are all **random**.

# Numerical sampling distributions

- We may apply the same idea to realistic sampling. Suppose I want to know the average height of all NTU students:
  - Tomorrow I will ask one hundred students and get

$$\overline{X}^1 = \frac{1}{2}(X_1^1 + \cdots + X_{100}^1).$$

  - 500 days later I will get $\overline{X}^{500}$.
- Each of $X^i$s is a **sample**.
  - They are **random** now but will be **known** after 500 days.
- Because I do not know the population distribution, I cannot analytically derive the sampling distribution.
- But I can numerically draw a histogram for the 500 values.
  - That histogram will "describe" the distribution of $\overline{X}$.

# Numerical sampling distributions

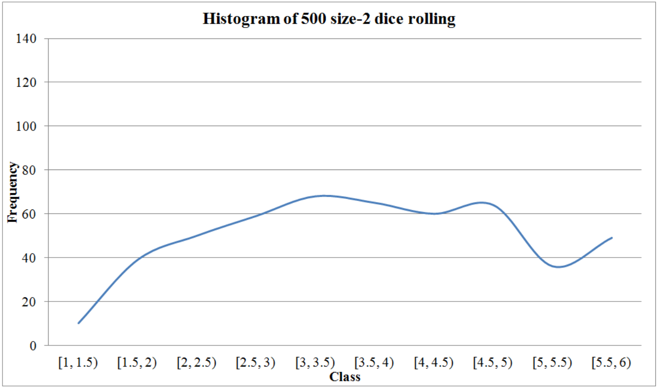
- Let's focus on rolling dices now.
- Suppose the data I collected are:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | $\cdots$ | 500 |
|---|---|---|---|---|---|---|---|---|---|
| $x_1^i$ | 6 | 3 | 1 | 1 | 6 | 6 | 3 | | 5 |
| $x_2^i$ | 3 | 1 | 4 | 4 | 3 | 6 | 2 | $\cdots$ | 3 |
| $\overline{x}$ | 4.5 | 2 | 2.5 | 2.5 | 4.5 | 6 | 2.5 | | 4 |

  - They are $(x_1^i, x_2^i)$, not $(X_1^i, X_2^i)$; they are known, not random.
- Let's draw a histogram for these 500 values.

# Numerical sampling distributions

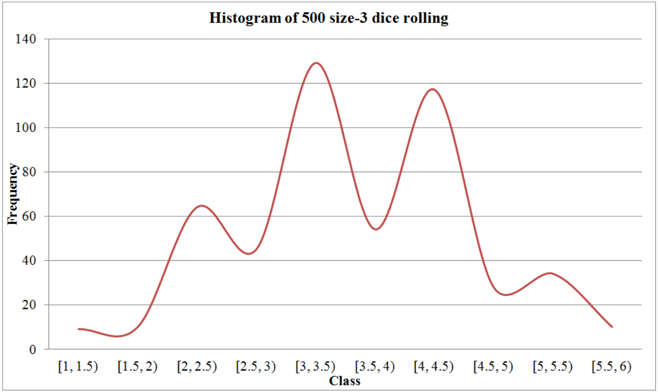
- The sampling distribution of $\frac{1}{2}(X_1 + X_2)$ looks like




- It **slightly deviates** from the population distribution (a discrete uniform distribution).

# Numerical sampling distributions

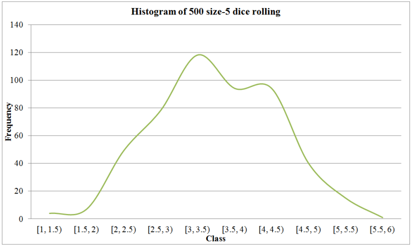
- What if each time we roll three dices and then get the mean?

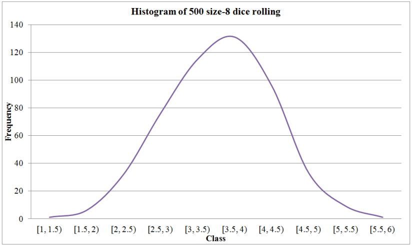



- It deviates from the population distribution more.

# Numerical sampling distributions

- If we roll five or eight dices at each time:




- As the sample size becomes larger:
  - It **deviates** from the population distribution more.
  - It gradually becomes a **bell-shaped** distribution.

# Sampling distributions: summary

- The population has its **population distribution**.
  - Rolling one dice.
  - Randomly selecting one student in NTU.
- Note that these are two interpretations of a population!
  - Alternatively, you may think in this way: I am not rolling a dice. Instead, someone has rolled a dice for 1000000 times, then I randomly draw one. What is the distribution of the 1000000 rolls?
- A statistic, which is random, has its **sampling distribution**.
  - Mean of rolling $n$ dices.
  - Mean of $n$ randomly selected NTU students heights.

# Sampling distributions: summary

- Sometimes we may **analytically** derive sampling distributions.
  - Mean of rolling $n$ dices.
- Sometimes we may not:
  - What's the population distribution of NTU students' heights?
- If we want to **numerically** depict a sampling distribution, we may repeat the sampling for many times, recording the value of the statistic each time, and then draw a histogram.
  - E.g., rolling two dices for 500 times.
- When we do this:
  - The **sample size** is 2, not 500!

# Road map

- Sampling techniques.
- Sampling distributions.
- **Distribution of the sample mean**.

# Sample means

- The sample mean is one of the most important statistics.

**Definition 1**

*Let $\{X_i\}_{i=1,\ldots,n}$ be a sample from a population, then*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

*is the sample mean.*

- Unless otherwise specified, a sample mean comes from an **independent sum**.
  - $X_i$ and $X_j$ are independent for all $i \neq j$.

# Means and variances of sample means

- A sample mean is also a random variable.
- No matter what the population distribution is, as long as the population mean is $\mu$ and the population variance is $\sigma^2$, the mean and variance of the sample mean of size $n$ are:
  - $\mathbb{E}[\overline{X}] = \mu$.
  - $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

# Means and variances of sample means

- Do the terms confuse you?
  - The sample mean vs. the mean of the sample mean.
  - The sample variance vs. the variance of the sample mean.
- By definition, they are:
  - $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$; a random variable.
  - $\mathbb{E}[\overline{X}]$; a constant.
  - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$; a random variable.
  - $\text{Var}(\overline{X})$; a constant.
- How about the mean and variance of the sample variance?

# Distribution of the sample mean

- If we **do not know** the population distribution, we cannot explicitly derive the distribution of the sample mean.
  - But at least we know its mean and variance.
- If we **know** the population distribution, what can we say?
  - When we are rolling dices?
  - When the population follows a normal distribution?
- Let's focus on sampling from a normal population first.

# Sampling from a normal population

- In the last homework you have proved the following:

**Proposition 1**

*Let $\{X_i\}_{i=1,\dots,n}$ be a sample from a normal population with mean $\mu$ and standard deviation $\sigma$. Then*

$$\overline{X} \sim \mathrm{ND}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- Let's see some examples.

# Sampling from a normal population

- Suppose we sampled 4 values from a normal population with mean 80 and standard deviation 10.
  - What is the mean of the sample mean?
  - What is the standard deviation of the sample mean?
  - What is the distribution of the sample mean?
  - What is the probability that the sample mean is above 82?
  - What is the probability that the sample mean is below 76?

# Sampling from a normal population

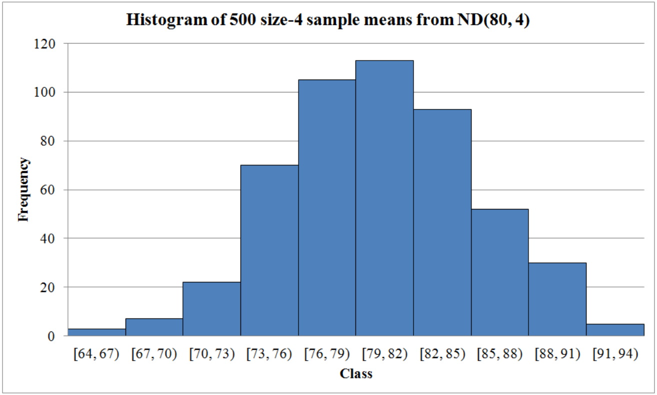
- What is the mean of the sample mean?
  - $\mathbb{E}[\overline{X}] = \mu = 80$.
- What is the standard deviation of the sample mean?
  - $\text{Var}(\overline{X}) = \frac{\sigma^2}{n} = \frac{100}{4} = 25$. The standard deviation is $\sqrt{25} = 5$.
- What is the distribution of the sample mean?
  - $\text{ND}(80, 5)$.
- What is the probability that the sample mean is above 82?
  - $\Pr(\overline{X} > 82) = \Pr(Z > 0.4) \approx 0.345$.
- What is the probability that the sample mean is below 76?
  - $\Pr(\overline{X} < 76) = \Pr(Z < -0.8) \approx 0.212$.

# Sampling from a normal population

- May we verify whether the theory is true?
  - At least we can verify it numerically for this example.
- The process:
  - We first generate 1000 values from $\text{ND}(80, 4)$.
  - Then randomly select 4 values and calculate the sample mean.
  - Repeat the size-4 sampling for 500 times.
  - Calculate the mean and standard deviation for the 500 values.
  - Finally, draw the histogram.

# Sampling from a normal population

- Mean = 80.24. Standard deviation = 4.97.



# Distribution of the sample mean

- So now we have one general conclusion: When we sample from a normal population, the sample mean is also normal.
- What if the population is **non-normal**?
  - In general, it is hard to analytically derive the distributions of sample means from non-normal populations.
  - Numerically we can do anything, but each time we get different results and conclusions.
- Fortunately, we have a very powerful theorem, the **central limit theorem**, which applies to **any** population distribution.

# Central limit theorem

- The theorem says that a sample mean is **approximately normal** when the sample size is large enough.

**Proposition 2 (Central limit theorem)**

*Let* $\{X_i\}_{i=1,\ldots,n}$ *be an independent sample from a population with mean* $\mu$ *and standard deviation* $\sigma$*, i.e.,* $\mathbb{E}[X_i] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2$*. Let* $\overline{X}$ *be the sample mean. If* $\sigma < \infty$*, then*

$$Z_n \equiv \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

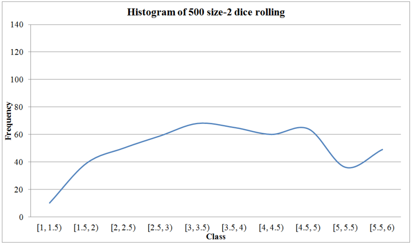
*converges to* $Z \sim ND(0, 1)$ *as* $n \to \infty$*.*

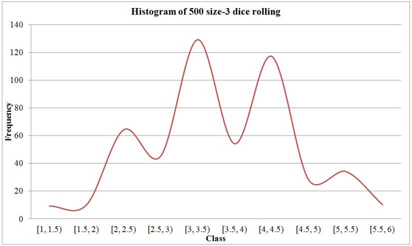
- Before we prove it, that see how it works.

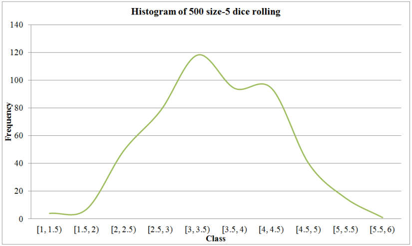
# Central limit theorem

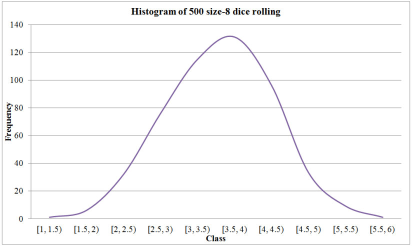
- Suppose we roll a dice (again). Let $X_i$ be the outcome of the $i$th roll.
  - $\Pr(X_i = x) = \frac{1}{6}$ for all $x \in \{1, 2, ..., 6\}$.
- What is the distribution of $\overline{X}$ when $n$ is large?
- The central limit theorem says: As $n$ is large enough, $\overline{X}$ follows a normal distribution (approximately).
- Is this true?

# CLT for rolling dices



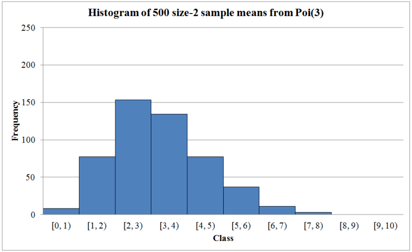



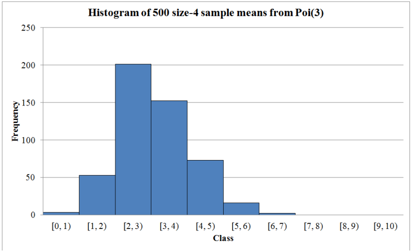
# CLT for Poisson population

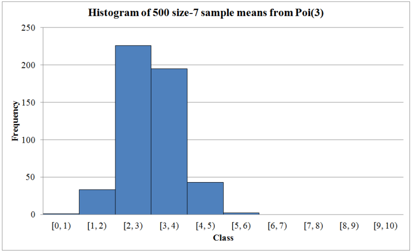
- As another example, let's consider a population following the Poisson distribution with rate $\lambda = 3$: $X_i \sim \text{Poi}(3)$.
  - The population mean and variance are both 3.
- We try four sample sizes: $n = 2$, 4, 7, and 10.
- For each sample size, we run 500 times of sampling.

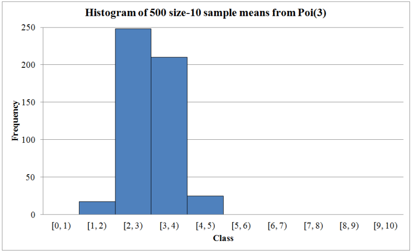
| $n$ | $\mathbb{E}[\overline{X}]$ | $\text{Var}(\overline{X})$ | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ | $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ |
|---|---|---|---|---|
| 2 | 3 | $\frac{3}{2} = 1.5$ | 2.972 | 1.702 |
| 4 | 3 | $\frac{3}{4} = 0.75$ | 2.966 | 0.804 |
| 7 | 3 | $\frac{3}{7} \approx 0.429$ | 2.947 | 0.485 |
| 10 | 3 | $\frac{3}{10} = 0.3$ | 2.950 | 0.328 |

# CLT for Poisson population



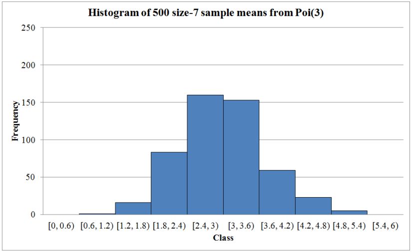



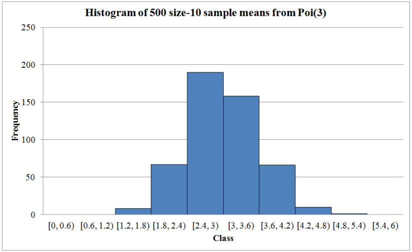
# CLT for Poisson population

- So indeed
  - The means of sample means are all close to 3.
  - The variance of sample means are all close to $\frac{3}{n}$.
  - The distribution of sample mean becomes more centered when $n$ becomes larger.
- Does it really approach a normal distribution?
  - The two histograms for n = 7 and n = 10 are not like normal!

# CLT for Poisson population

- Do not forget to adjust the interval length:



# Timing for central limit theorem

- In short, the central limit theorem says that, for any population, the sample mean will be approximately normally distributed as long as the sample size is large enough.
- How large is "large enough"?
- In practice, typically $n \geq 30$ is believed to be large enough.
- Do not forget that the central limit theorem **only applies** to the sample mean. It does not applies to other statistics.