BNAD 276
Lecture 7
Statistical Inference I
Sampling. Point Estimation. Sampling Distribution

Phuong Ho

May 28, 2017

Outline

Outline

Population vs. Sample

Population

The set of all elements of interest in a particular study

Sample

A subset of the population

- It is extremely hard to have a population data.
- It is safe to think that every data we can get is a sample not a population.

Data Point as Random Variable

- Data is nothing but a set of many numbers.
- A data point is a specific number among those many numbers.
- From now on, we give another interpretation on a data point rather than just a number.
- We can think a data point is a realization of the random variable that we are interested in.

Data point as a realization of a random variable

We may interpret a data point as a realization of the Random Variable that we are interested in.

Example 1: Rolling a die

- Suppose that our experiment is rolling a die once.
- Our random variable X is defined by the number we get in the experiment.
- The values that X can take are $\{1, 2, 3, 4, 5, 6\}$
- We will do this experiment N times.

Example 1 cont'd

- We will do this experiment N times.
- Let us call the random variable in the i^{th} experiment X_i.
- Since we are going to do N numbers of experiments, we have N numbers of random variables as follows.

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N

Example 1 cont'd

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N

- We roll a die and record the number we get under X_1. This is the first experiment. The number we get is a realized outcome of X_1.
- We roll a die again and record the realized number under X_2. This is the second experiment.
- We keep doing this until we finish the N_{th} experiment. Then, we finish our first job, collecting a sample data.

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N
Outcomes	4	1	3	$\cdots$	6

Example 1 cont'd

- Now suppose that we re-collect our data again.
- We will do the same thing as we did before: Roll a die and record each realized outcome under X_i.
- Probably, we will have a different numbers (data) pattern from the number pattern we had in the first data collection.

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N
Data 1	4	1	3	$\cdots$	6
Data 2	2	1	7	$\cdots$	5

Data Point as a Realization of Random Variable

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N
Data Point 1	4	1	3	$\cdots$	6
Data Point 2	2	1	7	$\cdots$	5

- From the above table, we can clearly see why we can interpret a data point as a realization of a random variable.
- Conceptually, X_1 could be either $1, 2, 3, 4, 5$ or 6. When we collect data, the number we observe is a **realized outcome** of a random variable.

Identically and Independently Distributed Random Variables

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N

- The above is the table we would fill with our data collection procedure.
- Before collecting data, X_is are random variable from i^{th} experiment.
- Note that even though we distinguish X_i from X_j, they have the IDENTICAL underlying experiment.

Identically and Independently Distributed Random Variables

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N

- Since X_i and X_j have the IDENTICAL underlying experiment, they are essentially IDENTICAL (they have the same probability function.)
- We say X_is are IDENTICALLY distributed when X_is have the same underlying experiment (or the same probability/density distribution).

Identically and Independently Distributed Random Variables

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N

- Furthermore, there is no reason to believe that a realization of X_1 and a realization of X_2 are related to each other.
- For example, the fact that we have 6 for X_1 has nothing to do with the number we will get for X_2.
- In this sense, X_i and X_j are INDEPENDENT.
- We say X_is are INDEPENDENTLY distributed when the experiments of X_is are independent.

Example 2

- Suppose that we collected data for the annual income of 100 people living in Tucson to study the income distribution of Tucson, or say, just the average income of the population in Tucson.
- We randomly selected 100 people and asked them to report their annual income. Finally we have the following data.
- Let us denote the i^{th} person's annual income by X_i.

X_is	X_1	X_2	X_3	$\cdots$	X_{100}
	80	30	15	$\cdots$	50

EXample 2

X_i	X_1	X_2	X_3	$\cdots$	X_i	$\cdots$	X_{100}
	80	30	15	$\cdots$	100	$\cdots$	50

- Our random variable, X_i, is the annual income of i^{th} person. X_i is random because the person i^{th} is randomly selected.

Example 2 cont'd

- Suppose that we conduct the same data collecting procedure again.
- We again randomly select 100 people and ask their annual income.
- Then, we will have a different numbers (data) than the number in the first data set.

X_i	X_1	X_2	X_3	$\cdots$	X_{100}
Data 1	80	30	15	$\cdots$	50
Data 2	26	39	35	$\cdots$	39

- It again illustrates that the numbers in a data set is a realized outcome of a random variable.

Example 2 cont'd

X_i	X_1	X_2	X_3	$\cdots$	X_{100}
Data 1	80	30	15	$\cdots$	50
Data 2	26	39	35	$\cdots$	39

- Since we randomly selected people in the same population (in Tucson), our data, random sample will give identically and independently random variables X_i and X_j.

Random Sampling

- In the previous example 2, we randomly select 100 people (elements) among the whole population.
- We call this **random selection of the objects/elements** from which we collect data **RANDOM SAMPLING**.

Random sample

A random sample of size n from a population is a sample selected such that the following conditions are satisfied:

1. Each element selected comes from the same population.
2. Each element is selected independently.

Outline

Example 1: Roll a die

- Theoretically, the die is fair, i.e. each outcome occurs equally likely.
- Hence, if $X =$ the number appears in an outcome, X has the expected value (true mean) $\mu = 3$, the true variance is $Var(X) = \sigma^2 =$, and the probability of the outcome $X = 1$ is $p = \frac{1}{6}$.
- We are not sure whether this die is fair, i.e. we are not sure or we do not know the values μ, σ^2, p.
- However, we can conduct the experiment, observe the realized outcomes, collect the data, and then estimate ("guess") the value of μ and σ^2 once we have a data set.

Point Estimation

- Recall that μ is true mean/population mean of a random variable X.
- Also recall that $\bar{X}$ is the sample mean (measure of central location of data).
- Since data is a set of realized results from the underlying experiment of X, data is a good representative of X.
- Thus, it is reasonable to use $\bar{X}$ (the measure of central location in data) to estimate μ (the central location of X.)

Point Estimation

- Also Recall that σ^2 is the measure of variability of a random variable X.
- Again recall that s^2 is the measure of variability of data.
- Since data is a set of results from the underlying experiment of X, data is a good representative of X.
- Thus, it is reasonable to use s^2 to estimate σ^2.

Point Estimation

Experiment	1	2	3	$\cdots$	N
RVs	X_1	X_2	X_3	$\cdots$	X_N

- μ and σ^2 will determine main properties of X.
- These are called the population parameters.
- Point estimation is the process to estimate the population parameter value by using data values.

Point Estimation

the process to estimate the population parameter by using data.

Estimators vs. Estimates

- We call $\bar{X}$ an **estimator** for μ and s^2 an **estimator** for σ^2.
- Suppose we have calculated the sample mean and variance and the results are $\bar{X} = 20$ and $s^2 = 10$.
- These **calculation results** that we get when we "plug" data into estimators are called **estimates**.
- We will conclude that these two numbers (20 and 10) is very close to μ and σ^2.
- Then, we are done with the point estimation for μ and σ^2 of X (the annual income.)

Estimators vs. Estimates. Example: roll a die

Parameter estimated	Estimator	Estimates
μ	$\bar{X} = \frac{1}{N}\sum X_i$	$4, 3.5, 3, 4.2, 3.7, \ldots$
σ^2	$s^2 = \frac{1}{N-1}\sum (X_i - \bar{X})^2$	$2, 3, 2.9, 2.7, 2.4, \ldots$
p	$\bar{p} = \frac{\text{the number of ones}}{N}$	$0.15, 0.2, 0.3, 0.16, \ldots$

- We typically use $\bar{X}$ to estimate μ and s^2 to estimate σ^2, (and sample proportion to estimate population proportion).
- An estimator is a "**formula**" by itself which gives us the way to calculate estimates with data.
- An estimate is the resulted **number** from data.
- As seen, data can change whenever we do sampling again. Thus, estimates can vary. However, estimator is a formula and thus will not "change" whenever we do different sampling.

Why $\bar{X}$ and s^2?

- Intuitively, if data comes from the underlying experiment of X, the data itself is a good representation of the distribution of X.
- Thus, by analyzing data, we can see a part of the whole distribution of X.
- Besides, $\bar{X}$ and s^2 have a very good property: They are **very close** to μ and σ^2.

Law of Large Numbers

The Law of Large Numbers

If X_i are independently and identically distributed ($i = 1, 2, 3, \ldots, N$), As N gets larger (as the sample size increases),

$$\bar{X} = \frac{1}{N} \sum X_i \approx \mu.$$

- It just states that if the sample size N is large enough, the sample mean will be approximately the same as the true mean μ. This property means that the sample mean is a **consistent estimator** for the true mean.
- Let's verify the law of large numbers by simulation.

Example 1: Roll a die

- We roll a die. Whenever we roll a die, we collect 1 data point in our data.
- We roll a die 100 times. So, at the end we will have 100 data points in our data.
- The random variable X in this example is defined by the realized number we get when we roll a die.
- Our purpose is to verify the law of large numbers by simulation.
- Before we do the simulation, let's figure out what is the true mean, $E(X) = \mu$, and variance, $Var(X) = \sigma^2$.

True Mean, True Variance, assuming the die is fair

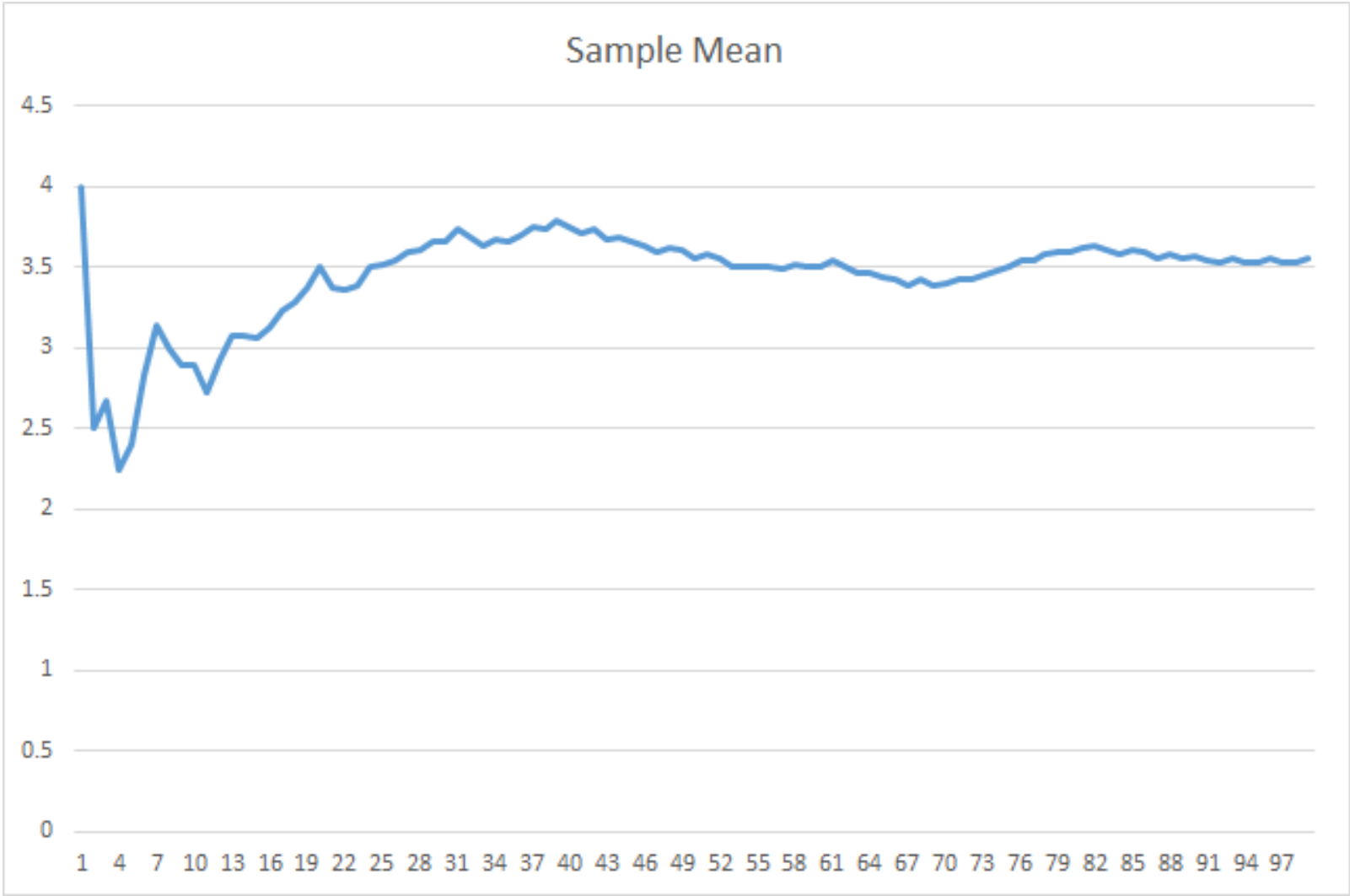
We know our X has $f(x) = \frac{1}{6}$ as the probability function.

$$E(X) = \mu = \sum x \frac{1}{6} = (1/6 + 2/6 + \cdot\cdot + 6/6) = 3.5$$

$$Var(X) = \sigma^2 = \sum (x - 3.5)^2 \frac{1}{6} = 2.91$$

- Then, now let's do the simulation.

Consistency of Sample Mean, $\bar{X}$



Consistency of Sample Mean, $\bar{X}$

- In the previous figure, the numbers on the horizonal axis is the sample size N.
- On the vertical axis, we have the estimates of $\bar{X}$ with a sample size $N = 1, 2, 3, ...100$.
- As one can see, as N gets larger, $\bar{X}$ is really close to $\mu = 3.5$.

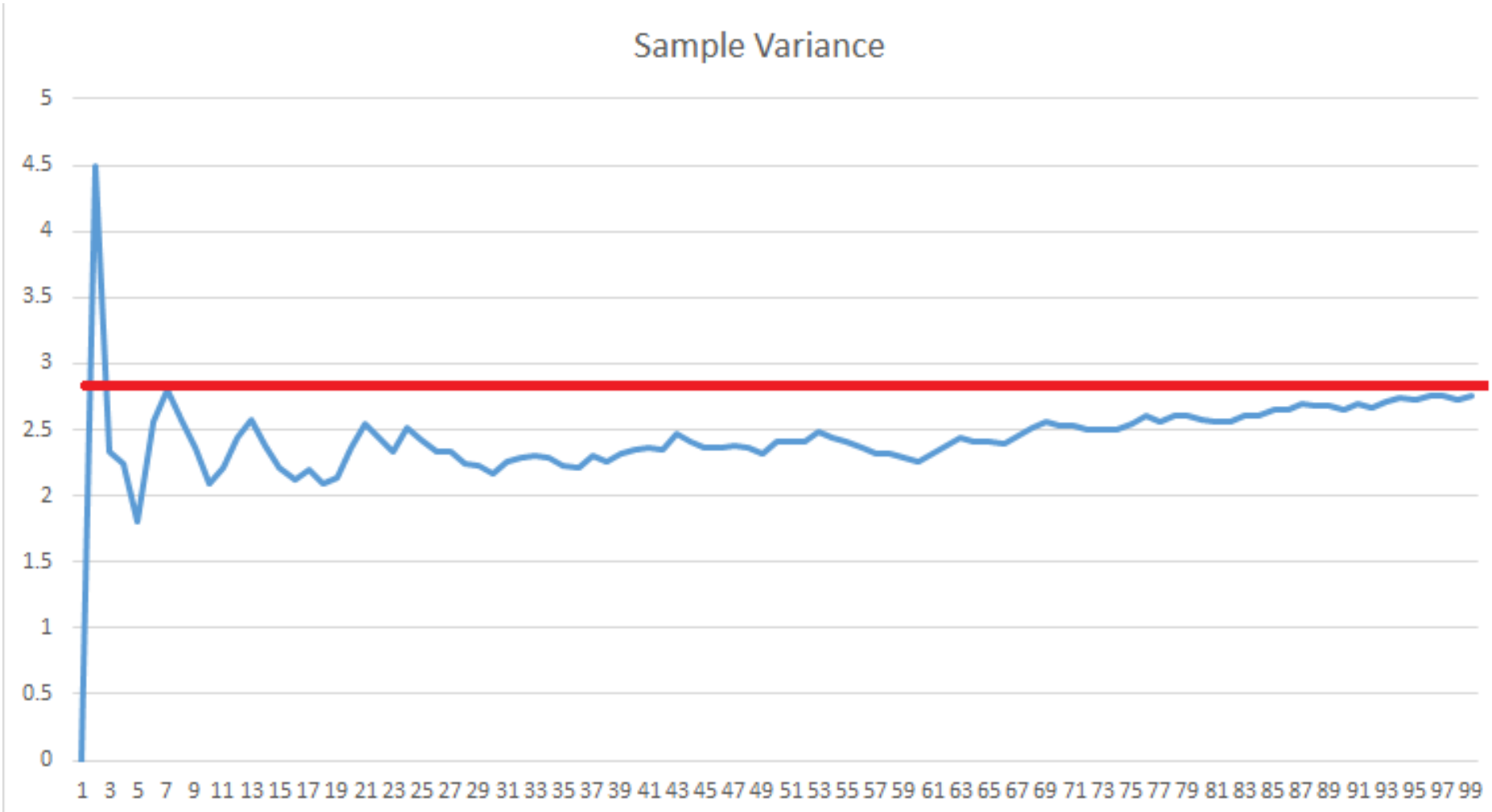
Consistency of Sample Variance, s^2

- Recall that

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- s^2 does not have the exact form of average (not divided by n).
- However, $N \approx N-1$ with a large value of N. Thus, the law of large numbers can be applied (we do not explain in detail here). And, s^2 is a consistent estimator for σ^2.

Consistency of Sample Variance, s^2



Consistency of Sample Variance, s^2

- In the previous figure, the numbers on the horizonal axis is the sample size N.
- On the vertical axis, we have the estimates of s^2 with a sample size $N = 1, 2, 3, ..., 100$.
- As one can see, as N gets larger, s^2 is really close to $\sigma^2 = 2.91$.

Simulation Results

- We've just seen that $\bar{X}$ and s^2 do a good job on estimating μ and σ^2.
- The advantage of point estimation (the advantage of using $\bar{X}$ and s^2 to estimate) is, in fact, we don't need to know a specific distribution that X follows.
- Amazingly, regardless of what kinds of the distribution a random variable X follows (e.g. regardless of whether X follows a normal or uniform), $\bar{X}$ and s^2 give really good approximation of $E(X) = \mu$ and $Var(X) = \sigma^2$.

Summary

- $\bar{X}$ is a very good estimator for μ.
  - Its estimates will be really close to μ if N is large enough.
- s^2 is a very good estimator for σ^2.
  - Its estimates will be really close to σ^2 if N is large enough.
- Thus, we can use $\bar{X}$ and σ^2 to estimate $E(X) = \mu$ and $Var(X) = \sigma^2$.

Example 2: Annual Income

We randomly select 100 people and record their annual incomes in the following table.

X_i	X_1	X_2	X_3	$\cdots$	X_{100}
Data	80	30	15	$\cdots$	50
$\bar{X} = 45$					

- Note that we don't know what kind of distribution that X (the annual income) follows.
- However, we know that there is $\mu = E(X)$ and $Var(X) = \sigma^2$.
- And, we also know that once we calculate $\bar{X}$ and s^2, they will be really close to μ and σ^2.
- In short, we can estimate μ (σ^2) with $\bar{X}$ (s^2).

Example 3: Proportion of married people

Suppose we randomly select 100 people and we record their marital status with the random variable X which takes value 1 only if a person is married and takes 0 otherwise.

X_i	X_1	X_2	X_3	X_4	$\cdots$	X_{100}
Data	0	1	0	1	$\cdots$	0

- Denote the probability that $X = 1$ (i.e. the probability that each person is married) by p.
- At this point, we have no idea what p is.
- p is the value we want to know.

Example 3: Proportion of married people, cont'd

- Assume that every person's marital status is governed by X.
- Then, p is the population proportion of married people.
- It turns out that p is the population mean (or true mean) since

$$E(X) = \sum xf(x) = 1 \times p + 0 \times (1 - p) = p$$

Example 3: Proportion of married people, cont'd

X_i	X_1	X_2	X_3	X_4	$\cdots$	X_{100}
Data	0	1	0	1	$\cdots$	0

- If we calculate $\bar{X}$ in the above data, it will be sample proportion of married people as you can see in the following.

$$\bar{X} = \frac{1}{100}\sum X_i = \frac{\text{\# of married people}}{100}$$

- And, $\bar{X}$ is a good estimator for $E(X) = p$ which represents the proportion of married people in the whole population.
- The sample proportion, $\bar{X}$, is a good estimator for the population proportion.

Example 3: Proportion of married people, cont'd

- Now assume that we have $\bar{X} = 0.3$. Then, we know that $\bar{X} = 0.3$ is really close to $p = \mu = E(X)$.
- Now we are interested in the number of people who are married in our sample.
- Define this new random variable as
  $Y =$ the number of married people in a sample.
- Then, what is the probability that $Y = 3$ when we randomly select 10 people from the population?

Example 4: Application on Poisson Distribution

Define X as the number of aircraft accident occurred in each year. Since X exhibits the property of a Poisson random variable, let's assume that X has a Poisson distribution with the following probability function.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

We don't know the value of μ, but we have data from the history. Suppose we have the following data.

X_i	X_1	X_2	X_3	$\bar{X}$
Data	14	15	16	

Example 4: Application on Poisson Distribution

- Recall that in the question there was a statement saying that An average of 15 aircraft accident occur each year. Previously, we replace the number 15 with μ in $f(x)$.
- The number 15 was sample mean. Basically, we implicitly replace μ with our sample mean which is given in the question.
- The reason why we could do that is because sample mean is a good estimator of μ in the sense that $\bar{X}$ will be approximately close to μ.

Example 4: Application on Poisson Distribution

- Recall that μ governs everything in a Poisson distribution. By finding μ, we can get the probability function that is really close to the true probability function.
- From there, we could calculate various probabilities such as $P[X \leq 2]$, $P[X \geq 3]$ and so on.
- These probabilities give us a predictions such as the probability that there is less than 3 accident in a year or probability that there is at least 1 accident in a month.
- This is why statistics can be useful.

Outline

Estimators are Random Variables

- We've noticed that estimators $(\bar{X}, \bar{p})$ are random variables.
- Recall the following example.

X_i	X_1	X_2	X_3	$\bar{X} = \frac{\sum X_i}{N}$
Data 1	80	30	15	46.6
Data 2	26	39	35	33.3
Data 3	28	12	35	25
.	.	.	.	.
Data 10	33	54	26	37.6

- Also recall that every random variable has a distribution.
- A random variable X distributed by either its probability function or its probability density function.
- Thus, $\bar{X}$ also has a distribution as a random variable.

Central Limit Theorem

Central Limit Theorem (CLT)

If $(X_1, X_2, ..., X_N)$ are independently and identically distributed, then the distribution of $\bar{X}$ (the sampling distribution of the sample mean) can be approximated by a normal distribution as the sample size becomes large:

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

$\mathcal{N}(\cdot, \cdot)$ denotes the normal distribution with mean and variance (respectively) in the bracket.

- As N gets large enough, the mean of $\bar{X}$ is closed to the true population mean μ, and the variance of $\bar{X}$ is closed to $\frac{\sigma^2}{N}$.

Application of CLT, and Sampling of $\bar{X}$

- The Central Limit Theorem provides the sampling distribution of $\bar{X}$.
- The sampling distribution enables us to do interval estimation and hypothesis test.
- Approximately (as the sample size N is large enough), the standard deviation of the sample mean $\bar{X}$ is

$$\sigma_x = \frac{\sigma}{\sqrt{N}}$$

To avoid confusion with the standard deviation σ of the population, we refer to the standard deviation of $\bar{X}$, σ_x, as the **standard error** of the (sample) mean.

Sampling Distribution of Sample Proportion $\bar{p}$

Example: Proportion of Married People:
Suppose we randomly select N people. We're interested in the proportion of married people of the population. So, we record the marital status by the random variable X which takes value 1 if a person is married and takes 0 otherwise.

X_i	X_1	X_2	X_3	X_4	$\cdots$	X_N
Data	0	1	0	1	$\cdots$	0

- We want to estimate the true proportion of married people of the whole population p. So, we use the sample proportion $\bar{p}$ as the estimator for p.

Sampling Distribution of Sample Proportion $\bar{p}$

Sampling Distribution of Sample Proportion $\bar{p}$

The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $Np \geq 5$ and $N(1-p) \geq 5$.
Hence, with a large enough sample size N, we have

$$E(\bar{p}) = p$$
$$Var(\bar{p}) = \frac{p(1-p)}{N}$$

$$\bar{p} \overset{approx}{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{N}\right)$$

$\mathcal{N}(\cdot, \cdot)$ denotes the normal distribution with mean and variance (respectively) in the bracket.
Similarly to the case of sample mean, we refer to the approximated standard deviation of $\bar{p}$, $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$, as the standard error of $\bar{p}$.

Outline

Exercise 1

Explain the following.

a. The Law of Large Numbers

b. Central Limit Theorem

Exercise 2

The Dow Jones Travel Index reported what business travellers pay for hotel rooms per night in major U.S. cities(*The Wall Street Journal*, January 16, 2004). The representative hotel room rates for 18 cities are as follows:

Atlanta	$163	Minneapolis	$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Angeles	192	Washington, D.C.	207

a. Define X as the hotel room rate in U.S.A. Estimate $E(X)$ and $Var(X)$.

b. Explain why you can use the estimator in part (a)

Exercise 3

Assume the population standard deviation is $\sigma = 25$. Compute the standard error (approximate standard deviation) of the sample mean, $\sqrt{Var(\bar{X})}$ for sample size of 50, 100, 150 and 200.

Exercise 4

A population has a mean of 200 and a standard deviation of 50. Suppose a simple random sample of size 100 is selected and $\bar{X}$ is used to estimate μ.

a. What is the approximate probability that the sample mean will be between $\mu - 5$ and $\mu + 5$?

b. What is the approximate probability that the sample mean will be between $\mu - 10$ and $\mu + 10$?

Exercise 5

To study proportion of people who has a college degree, a simple random sample of size 1000 is selected from a population with $p = 0.4$, where p is the probability that a person has college degree. Suppose we use the sample proportion $\bar{p}$ to estimate p.

a. What is $E(\bar{p})$?

b. What is $Var(\bar{p})$?

c. Write the approximate distribution of $\bar{p}$.

d. What is the approximate probability that $\bar{p}$ will be between $p - 0.03$ and $p + 0.03$?

e. What is the approximate probability that $\bar{p}$ will be far from p within the margin ± 0.05?