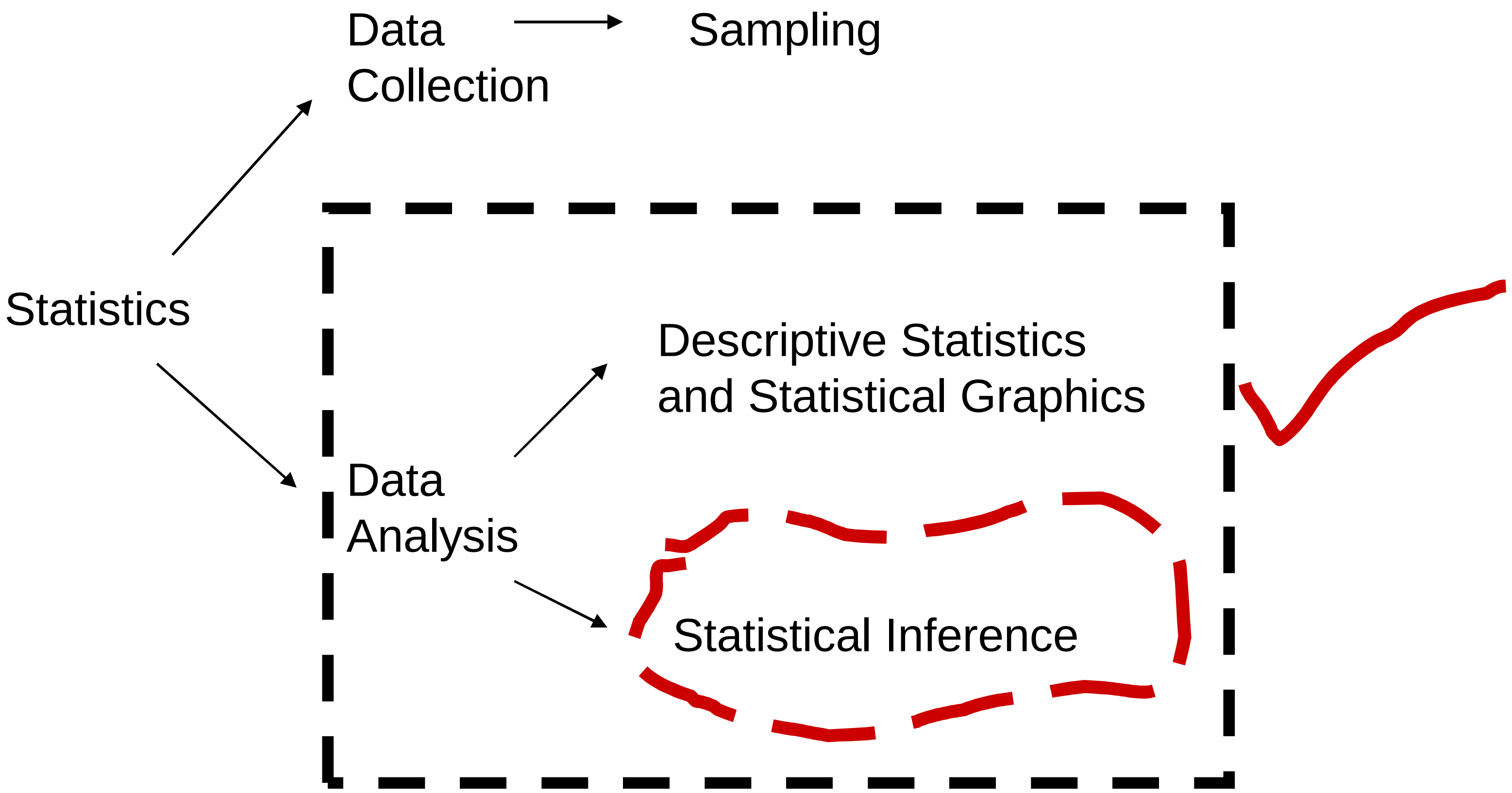
Statistical Inference

What is Statistics?

- It is a science that involves data summarization, experimental design, data collection, etc.
- Recently, it has been considered to encompass the science of basing inferences on observed data and the entire problem of making decisions in the face of uncertainty.

What is Statistics?



Statistics

Data Collection

Sampling

Data Analysis

Descriptive Statistics and Statistical Graphics

Statistical Inference

Difference between probability and statistics

Difference between P and S

	Probability	
1	We have a **fair** coin.	

Difference between P and S

	Probability	
1	We have a **fair** coin.	
2	Flip the fair coin ten times.	

Difference between P and S

	Probability	
1	We have a **fair** coin.	
2	Flip the fair coin ten times.	
3	**P({all are heads}) = ?**	

Difference between P and S

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	
3	**P({all are heads}) = ?**	

Difference between P and S

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	Flip the coin ten times.
3	**P({all are heads}) = ?**	

Difference between P and S

	Probability	Statistics
1	We have a **fair** coin.	We have a coin.
2	Flip the fair coin ten times.	Flip the coin ten times.
3	**P({all are heads}) = ?**	**All heads are obtained**, then **is it a fair coin?**

So, in the same random experiment,

- a probabilitist will only ask the probability of getting a certain event under some probabilistic model assumptions **before** doing the experiment, (kind of mathematics approach), while
- a statistician will make some conclusion about the probability model **after** the experiment (kind of physics approach)

Refer to the above example of tossing a coin.

A probabilitist will tell you that if the coin is fair, then

$$P(\{\text{all are heads}\}) = (0.5)^{10}=0.0009765625.$$

So, in some sense, probability is about **looking forward**.

For a statistician, if all heads are obtained, then s(he) will make a conclusion that the coin is NOT fair; otherwise, it is very unlikely to get ten heads in a row.

So, we can say that statistics is about **looking backward**.

Statistical Inference

What is Statistical Inference

Data =

Use a statistical approach to make an inference about
the **distribution of a sample of data** we collect.

- **What distribution(s) are the data from?**

  Normal distribution? Poisson distribution?

  or other distributions we have not seen before?

Suppose that they are from the normal distribution.

- **What normal distribution(s) are the data from?**

  N(0,1)? N(0,5)? N(-3, 5)? or other normal distributions?

What is Statistical Inference

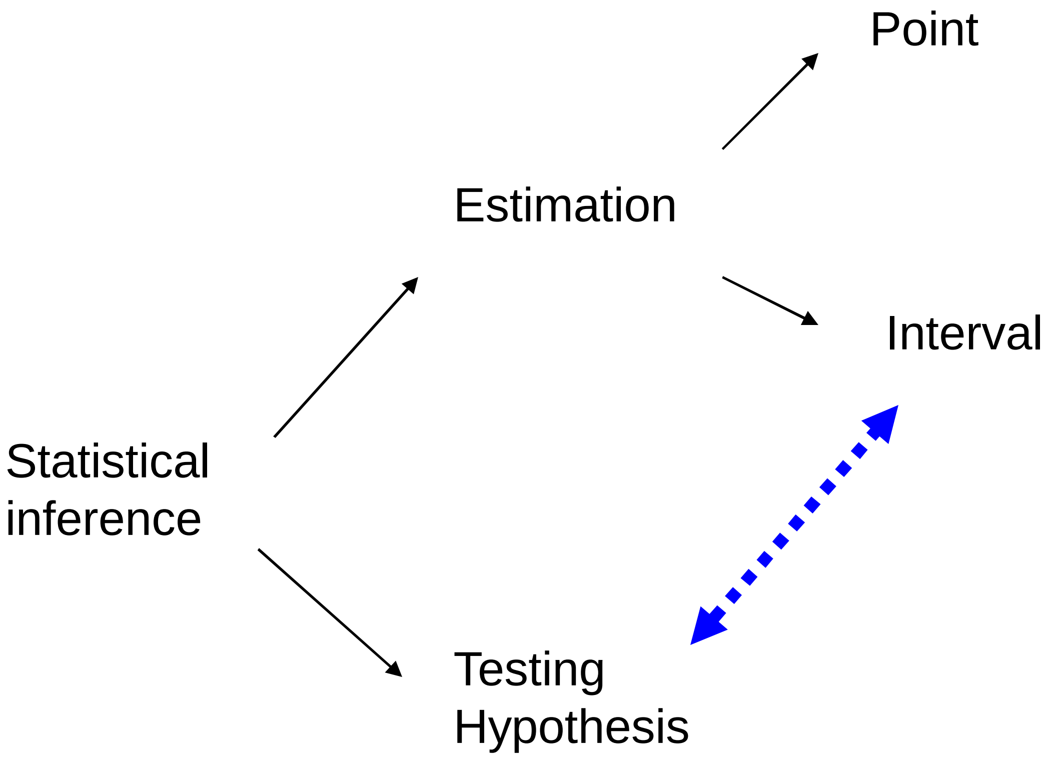
WHY?

Use a statistical approach to **make an inference about**

the distribution of a sample of data we collect.

The population or macroscopic phenomenon is always unknown itself,
because some, but not all, of the data of our interest can be taken.

Statistical Inference



Statistical inference

Estimation

Point

Interval

Testing Hypothesis

Population (macroscopic phenomenon) and Sample

Population is a set of measurements in which we are interested.

If X is the random variable of our interest in a random experiment, then the population is the distribution of X and each observation in a population is just a value of X.

However, it is impossible or impractical to know the underlying distribution of the random variable.

For instance, we are interested in the income of all NY people per month, but it is impractical to collect the data of several million NY people. At least, it is costly and time consuming to do so.

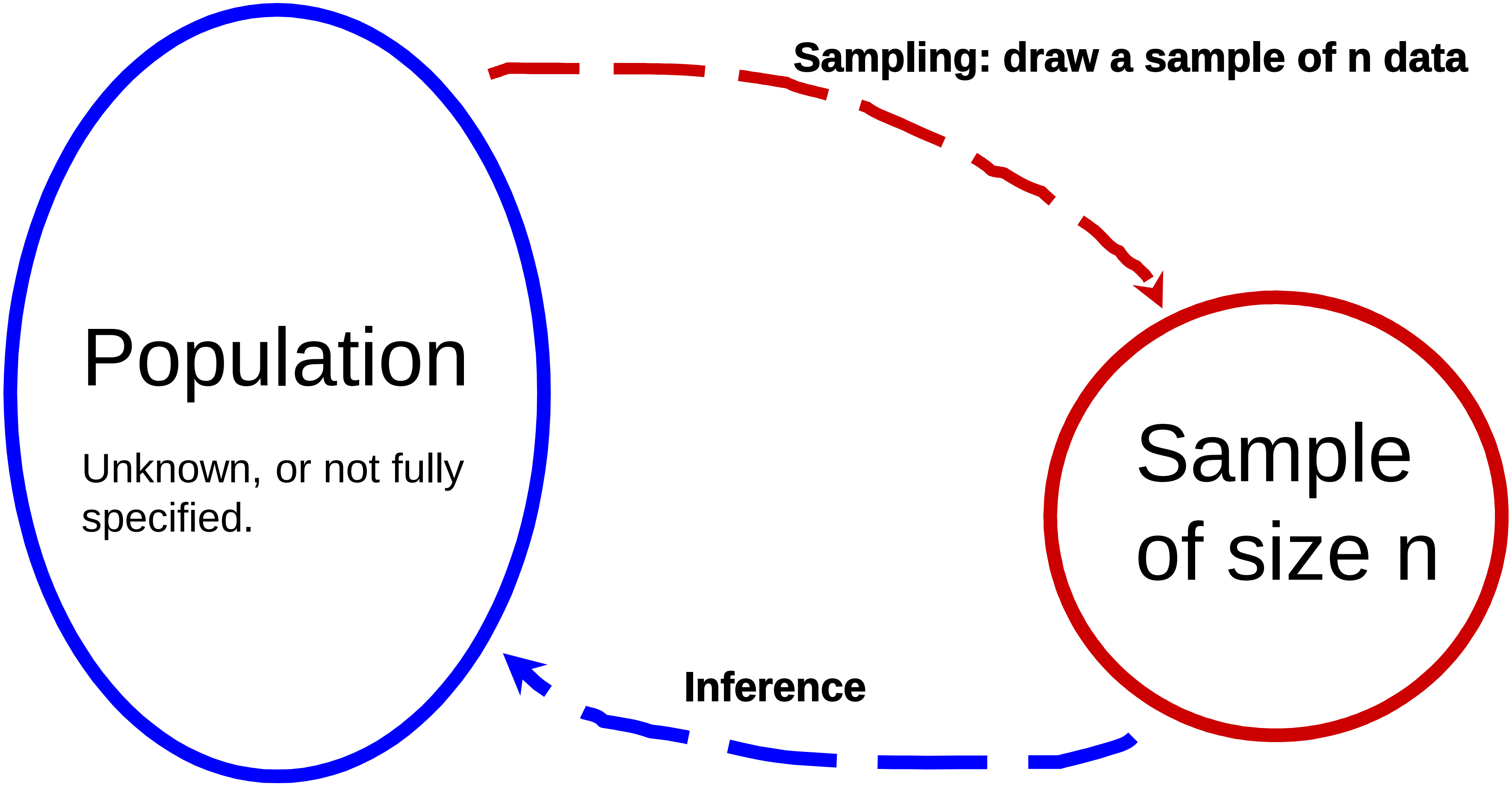
Population and Sample

Thus, what we can do is to use a subset of observations from the population to help us make inferences concerning the population.

This bring us to consider the notion of sampling:
A **sample** is a subset of a population.

The total number of the sample is called a sample size, often denoted by n.

Population and Sample



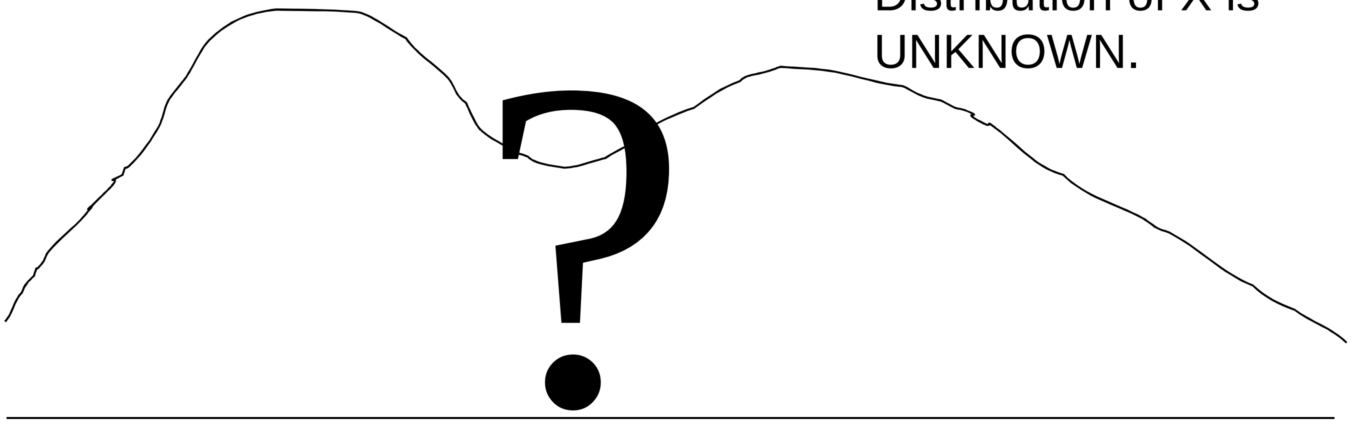
Sampling: draw a sample of n data

Population

Unknown, or not fully specified.

Sample of size n

Inference



Distribution of X is UNKNOWN.

Draw a sample from the population or the distribution of the random variable X.

➔ Obtain the actual/observed values of X.

If we want to draw a sample of size n, then we often denote the sample by $\{X_1, X_2,\ldots,X_n\}$, where X_i , i=1,…, n, represents the i^{th} observations of X.

If we want to draw a sample of size n, then we often denote the sample by $\{X_1, X_2,\ldots,X_n\}$, where X_i , i=1,…, n, represents the i^{th} observations of X.

Before sampling, each X_i is random and have the same distribution as X. In this course, we also assume that all X_i are independent. In statistics, if a set of random variables are **independent and identically distributed** (i.e. same distribution), then it is said to be **a random sample**.

After sampling, we have observed values of $X_1,X_2\ldots,X_n$, and denoted by $x_1,x_2,\ldots,x_n$, where **all x_i is a known real number**.

Referring to the example of the income of NY people per month, we can collect the data/sample by asking some NY people how much they earn.

For instance, at the first observation of the sampling, we may get $10,000, $9,460 at the second, $52,000 at the third, and so on. Thus, we can say that

$$x_1=10000,\ x_2 = 9{,}460,\ x_3=52{,}000,\dots$$

Remark that each observation can provide us some information about the underlying distribution. So, we should collect data/observations as many as possible.

Population Parameter

In most situations, we only want to know some quantities about the population, say the population mean, instead of the population itself.

Such quantities are called **population parameters**, which are often unknown.

For instance, we may have an interest in the average income of NY people only, or we do not care about how the distribution of the mid-term score in our class looks like, and what we really want to know is the mean and the standard deviation of the scores.

Population Parameter

We often use Greek letters to denote population parameters, say μ, σ, θ, λ, and so on.

In this course, we only focus on two population parameters:

Population mean (μ, or E(X))

and population variance (σ^2, or Var(X)).

Mission!!

Use the information from the data we collected to make an inference about the unknown distribution.

Mission!!

Use the information from the data we collected to make

an inference about the unknown distribution.

an inference about the (or the function of) unknown parameter(s) of the specified distribution.

The form of the distribution is known.

For instance, we assume that the data are from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.

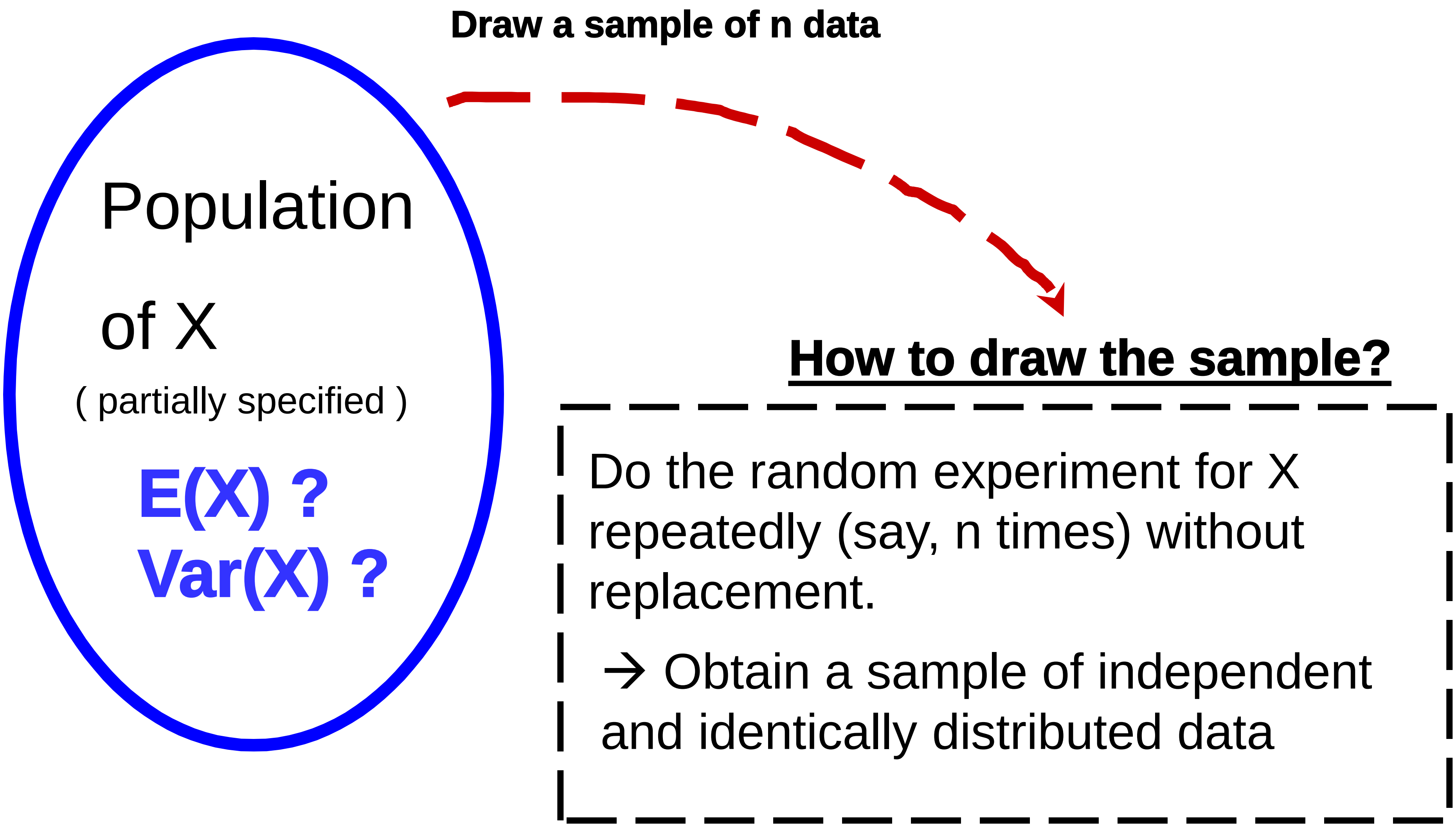
Statistical Inference

Part I: Parameter estimation

How to make a statistical inference about the unknown parameter(s)?

- Parameter estimation
- Hypothesis testing

Data



Draw a sample of n data

Population

of X

(partially specified)

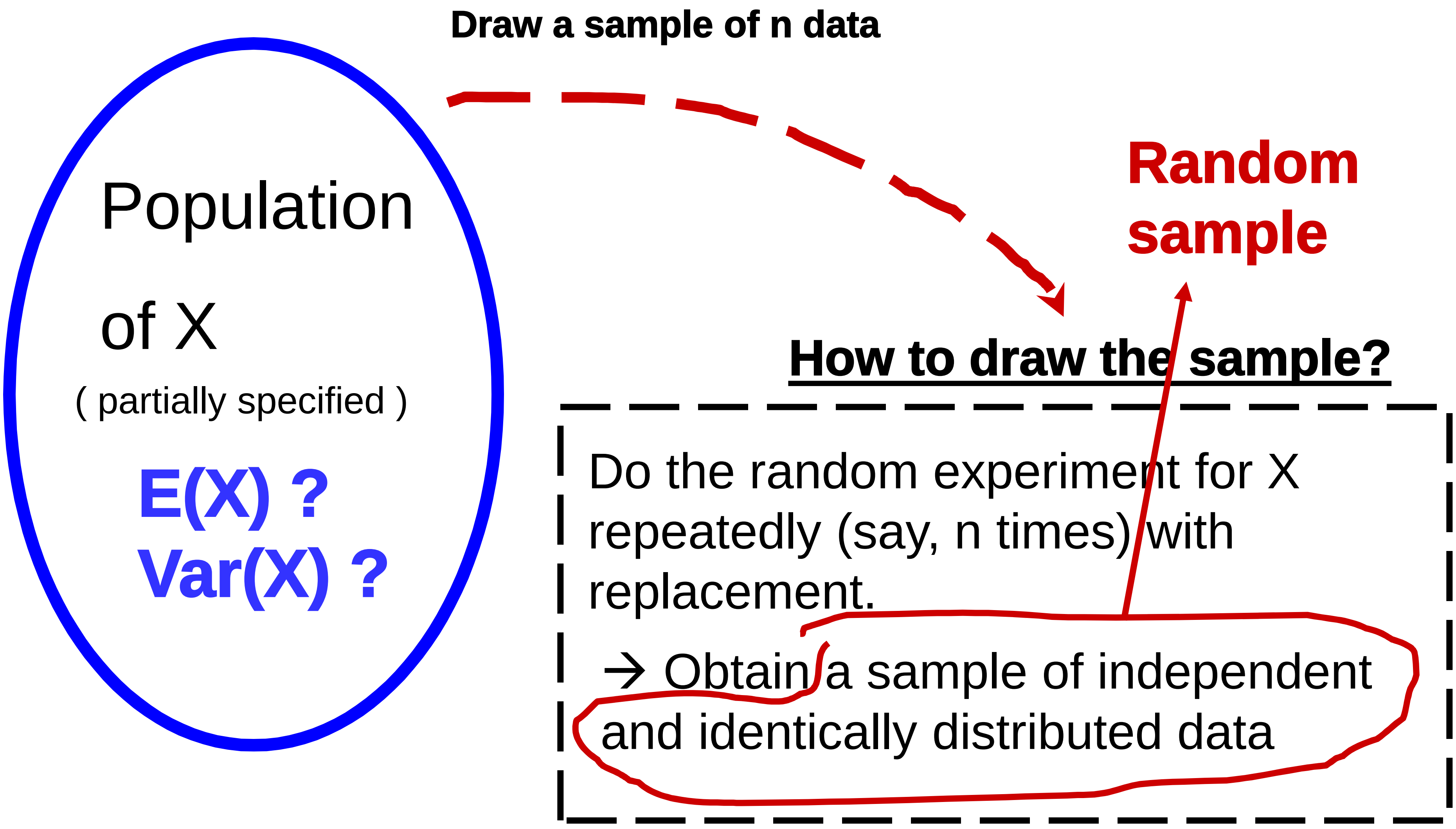
E(X) ?

Var(X) ?

How to draw the sample?

Do the random experiment for X repeatedly (say, n times) without replacement.

→ Obtain a sample of independent and identically distributed data



Draw a sample of n data

Population

of X

(partially specified)

E(X) ?

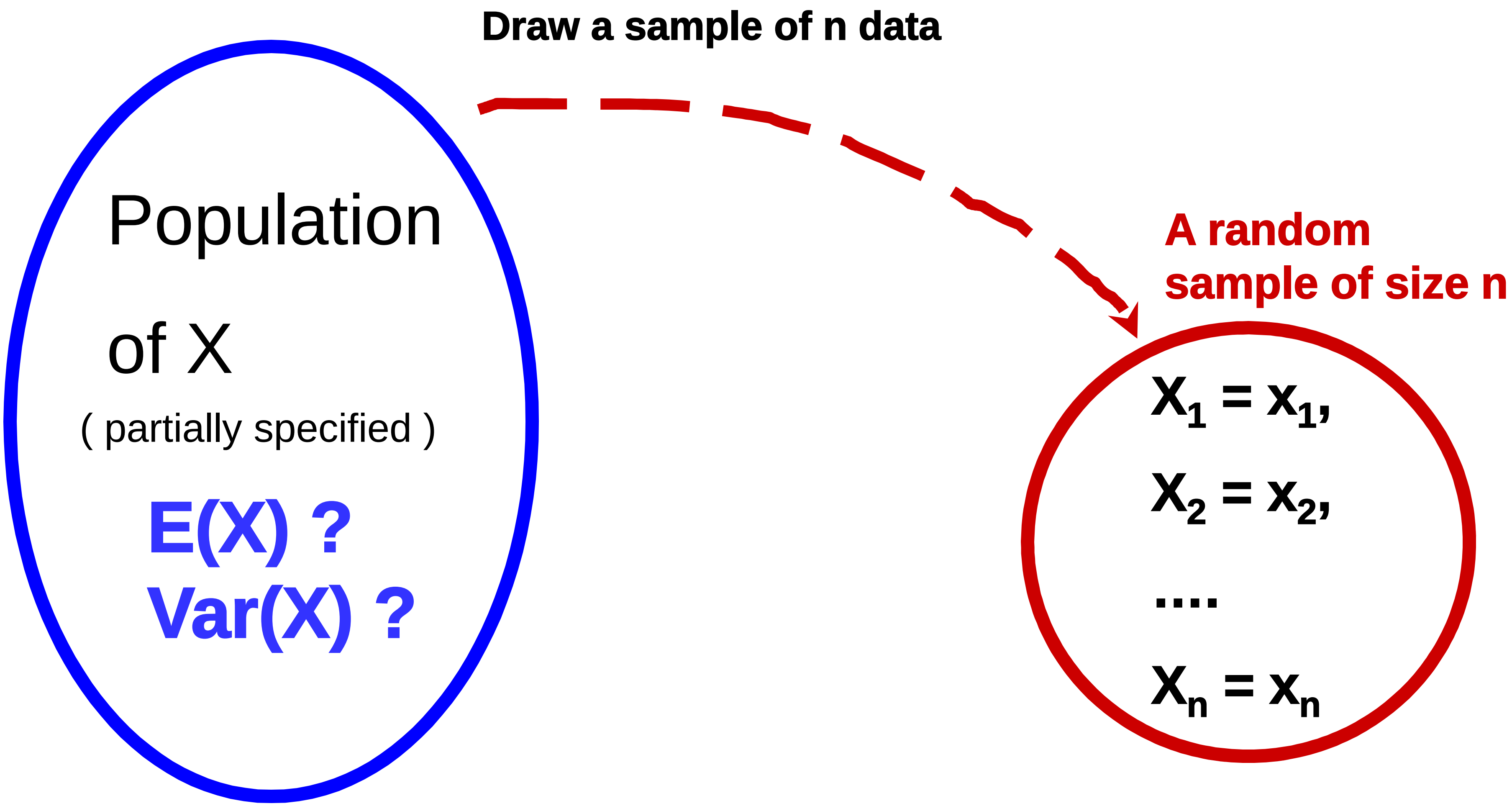
Var(X) ?

Random sample

How to draw the sample?

Do the random experiment for X repeatedly (say, n times) with replacement.

→ Obtain a sample of independent and identically distributed data



Draw a sample of n data

Population

of X

(partially specified)

E(X) ?

Var(X) ?

A random sample of size n

$X_1 = x_1,$

$X_2 = x_2,$

....

$X_n = x_n$

Recall that for $i = 1, \ldots, n$,

X_i represents the i^{th} observation of X **before sampling**, so it is unknown and unpredictable, i.e. X_i is a random variable.

After sampling, the actual value of X_i is known, say x_i ← a fixed number.

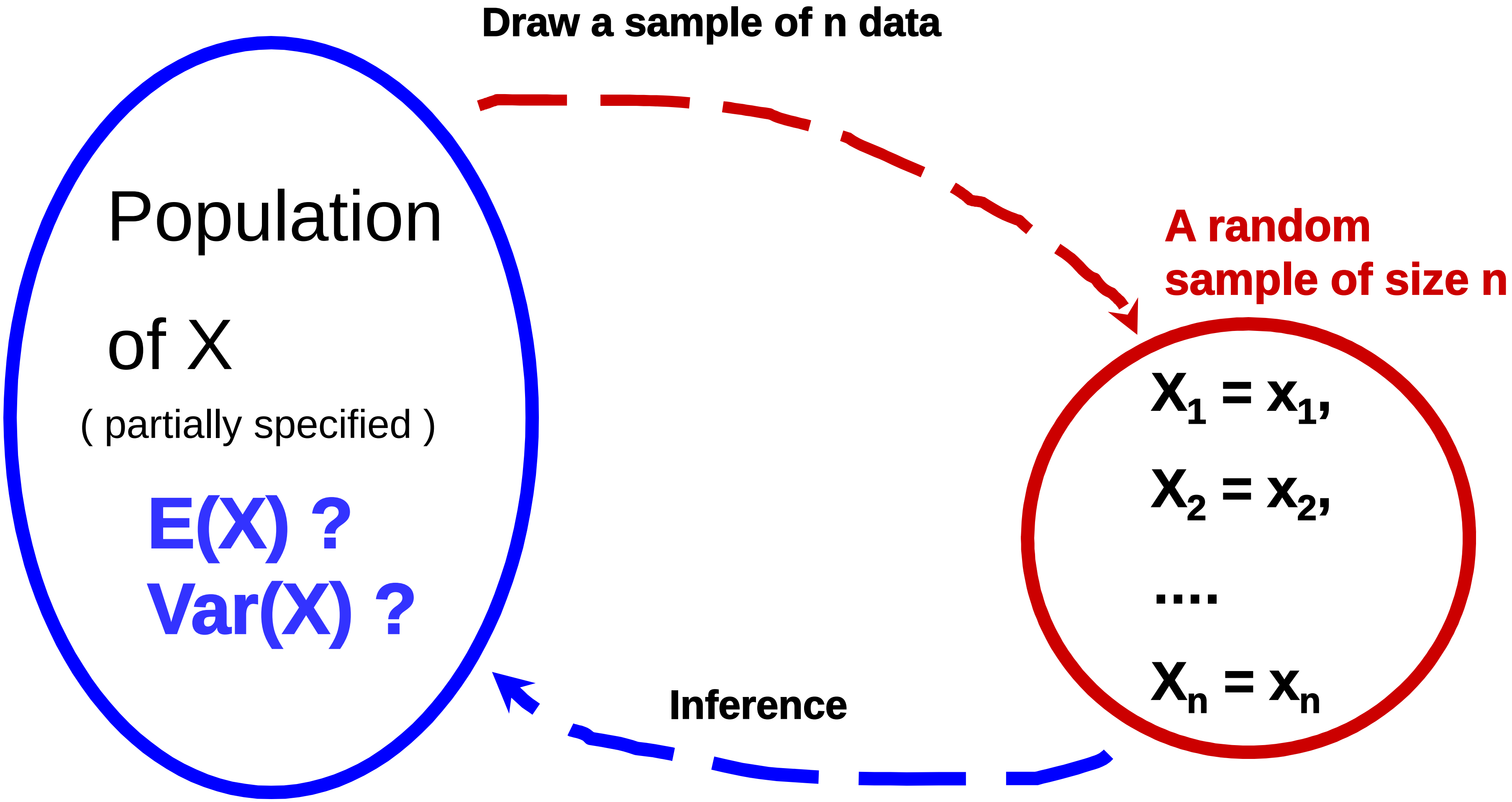
A random sample of size n

$X_1 = x_1$,

$X_2 = x_2$,

….

$X_n = x_n$



Draw a sample of n data

Population

of X

(partially specified)

E(X) ?

Var(X) ?

A random sample of size n

$X_1 = x_1$,

$X_2 = x_2$,

....

$X_n = x_n$

Inference

“A Statistic”

For population parameters, how can we get some information about them from a (random) sample $\{X_1, X_2, \ldots, X_n\}$?

Use **a function of a random sample**, say $T(X_1, X_2, \ldots, X_n)$, called **a statistic**.

For instance,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are statistics.

A Statistic

Caution:

a statistic does not depend on any unknown quantities. So,

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \text{«})^2$$

is NOT a statistic, unless μ is known.

After sampling, we have actual values of a (random) sample {X_1, X_2, …, X_n}, i.e. {x_1, x_2, …, x_n}, so we can also calculate the actual value of the estimators.

For instance,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

are the respective observed values of the sample mean and sample variance.

If we use a statistic to ESTIMATE an unknown parameter(s), then it is called **an (point) estimator of the parameter**.

The typical (point) **estimators** for μ and σ^2 are

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

respectively.

Remark that in statistics, the observed value of the estimator is called an **estimate** of the unknown parameter.

For instance, $\bar{x}$ and s^2 are the respective estimates of μ and σ^2.

Estimator

If θ is the parameter being estimated, then we often denote the estimator of θ by

$$\hat{\theta}(X_1, X_2 \ldots, X_n)$$

or simply

$$\hat{\theta}$$

Remark that since an estimator/ a statistic is a function of random sample, it is also random. Thus, there is a distribution that can be used to describe the behavior of the statistic. Such a probability distribution of a statistic is called a sampling distribution.

Desirable properties of Estimators

As expected, for an unknown parameter, say the population mean μ, we can find a whole bunch of estimators to estimate it, say sample mean, sample median, sample mode, or even a constant like 10.

For selecting a "more reasonable" estimator(s), we require estimators to have some nice properties, say unbiasedness stated below.

Unbiasedness:

An estimator $\hat{\theta}$ of θ is said to be unbiased if $E(\hat{\theta}) = \theta$.
Otherwise, it is biased.

Unbiasedness

Unbiasedness:

An estimator $\hat{\theta}$ of θ is said to be unbiased if $E(\hat{\theta}) = \theta$.
Otherwise, it is biased.

Interpretation:

In the long run, the amounts by which an estimator(s) over- and underestimates the parameter(s) will balance, so that the estimated value will be correct "on the average".

That is, if an estimator is unbiased, then it means that **on average**, the estimator $\hat{\theta}$ is equal to the unknown parameter θ.

The unbiasedness is one of the good properties to evaluate the goodness of estimators.

Example

Consider a discrete random variable X with pmf given by

$$\left.\begin{aligned} &P(X = 2) = p \\ &P(X = 4) = 2p \\ &P(X = 6) = 3p \\ &P(X = 8) = 4p \\ &P(X = 10) = 1- 10p \end{aligned}\right\} \text{Not fully specified}$$

and P(X = i) = 0 otherwise, where p is an **unknown** parameter in (0, 1/10).

After some calculation, we can find that

$$\underbrace{E(X) = 10 – 40\ p \text{ and } Var(X) = 200p – 1600p^2.}_{\text{Unknown}}$$

Now, we want to make an inference about

E(X) = 10 – 40 p and Var(X) = 200p – 1600p^2.

How?: we can draw a random sample of X:

For simplicity, let’s say that a sample of size n=2, X_1 and X_2, are drawn. So, the respective estimators of μ=E(X) and σ2 =Var(X) are

$$\overline{X} = \frac{X_1 + X_2}{2}$$

and

$$nS^2 = \frac{n}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(X_1 - X_2)^2}{2} = \pm^{2}.$$

Suppose that **after sampling**, the actual values of X_1 and X_2 are x_1 = 2 and x_2= 4, respectively. Then we can say that

$$\bar{x} = \frac{x_1 + x_2}{2} = 3$$

and

$$2s^2 = \frac{(x_1 - x_2)^2}{2} = 2 = (\pm\ ^2)$$

are the respective estimates of $\mu = E(X)$ and $\sigma^2 = Var(X)$

based on the observation $(x_1, x_2) = (2, 4)$.

Remark that if we draw another sample of X_1 and X_2, then we would get different actual values, because of the randomness of X_1 and X_2.

Now, based on the setting of X in this example, we can also list all possible values of the sample mean $\overline{X}$ and sample variance S^2, and then find the corresponding probabilities, by the assumption that X_1 and X_2 are the random sample of X.

Here we only focus on the probabilistic statement of $\overline{X}$ and the corresponding result of S^2 can also be obtained in a similar way.

Sampling distribution of $\overline{X}$

Possible value of $\overline{X}$	Probability
2	p^2
3	$4p^2$
4	$10p^2$
5	$20p^2$
6	$2p + 5p^2$
7	$4p – 16p^2$
8	$6p – 44p^2$
9	$8p – 80p^2$
10	$1 – 20p + 100p^2$

Are those results of the sample mean always true?

$$E(\overline{X}) = \sum_{i=2}^{10} iP(X = i) = 10 - 40p.$$

$$= E(X)$$

and

$$Var(\overline{X}) = 100p - 800p^2$$

$$= \frac{200p - 1600p^2}{2} = \frac{Var(X)}{n}$$

Unbiasedness of sample mean

The following theorem shows that the sample mean of a random sample is an unbiased estimator of the population mean.

Theorem:

Consider a random sample of size n, {X_1, X_2, …, X_n} from a common distribution with mean μ and variance $σ^2$. If

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

then

$$E(\overline{X}) = « \quad \text{and} \quad Var(\overline{X}) = \frac{1}{n} \pm^{2}$$

Questions

1. Under what condition is $\sum_{i=1}^{n} c_i X_i$ an unbiased estimator of μ, if $\{X_1,\ldots, X_n\}$ is a random sample from the population with mean μ ?

2. Show that the sample variance S^2 is unbiased for σ^2 , i.e. $E(S^2) = \sigma^2$.

Summary

Let's consider a random experiment of flipping a coin.
Denote by X the random variable of the number of getting a head. So, X can be either 0 or 1.

More specifically, we can use the Bernoulli distribution to describe the randomness of X. That is,

$$X \sim Bernoulli(p),$$

The distribution of X is not fully specified.

Unknown

where $p \in (0, 1)$ is the probability of getting a head.

After studying probability theory and probability distributions, we know that the population mean and variance of X are

$$\mu = E(X) = p \text{ and } \sigma^2 = Var(X) = p(1-p),$$

respectively.

Unknown

However, we cannot draw any further conclusion about the population (the distribution) of X, or equivalently, cannot say anything about the true value of the parameter p.

How does a statistician make an inference about p?

The basic idea is to:
(1) do the experiment repeatedly under the same condition so that we can collect some data, and then
(2) use the information from the data to guess the value of p.

More specifically, by repeating the random experiment $n = 10$ times,
i.e. flipping the same coin n times, we can get n observations of X, **a random sample of size n**.

Let's say we have

$$0, 1, 1, 0, 0, 1, 0, 1, 0, 0$$

If we use x_i to represent the observed value of X at the i^{th} flip, then we will have $x_1 = 0, x_2 = 1, x_3 = 1, \ldots, x_{10} = 0$.

Remark that before sampling, we use X_i to represent the random variable of the outcome at the i^{th} flip, for $i = 1, 2, \ldots, n$.

Now **how do we use the observations to guess p?**

We can consider a function of $x_1, x_2, \ldots, x_n$, that is, **a statistic**. Recall that a statistic is also called **an estimator/estimate** when we use it to estimate an unknown parameter.

So, **what is the reasonable estimator for p?**

Note that p is the probability of getting a head.

So, the natural statistic to estimate p is the proportion of getting a head in the n flips.

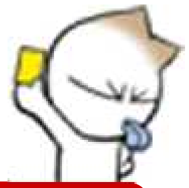
That is, we consider $\frac{1}{n}\sum_{i=1}^{n} x_i$, **the sample mean of the n observations**.

Based on the above observations , our estimate for p is 0.4.

Of course, we can also use other statistics to estimate p,

say sample median, or a constant function (say, your lucky number 15!).

So, **to select a ”more reasonable” estimator**,
or to exclude some ”poor” estimator, say 15,
we consider a criterion, **the unbiasedness of estimator**.



Note that unbiasedness is used for the **estimatOR**, not estimate.

Consider the behavior of the statistic before sampling.

What is the unbiased estimator for p?

Observe that $\{X_1, X_2, \ldots, X_n\}$ is a random sample,
so according to the result we discussed before,

$$\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is unbiased for } p$$

because $E(\hat{p}) = E(\bar{X}) = E(X) = p$.

Let $g(t) = t(1-t)$.
Now we want to find an unbiased estimator for $g(p) = p(1-p)$, i.e. the variance of X.

Remark again that in general $g(\hat{p})$ is NOT unbiased for $g(p)$, even though $\hat{p}$ is unbiased for p.

In fact, we can also see that

$$\begin{aligned}
E(g(\hat{p})) = E\left[\hat{p}(1-\hat{p})\right] = E\left[\bar{X}(1-\bar{X})\right] &= E(\bar{X}) - E(\bar{X}^2) \\
&= E(\bar{X}) - \left[Var(\bar{X}) + \left[E(\bar{X})\right]^2\right] \\
&= p - \left[\frac{p(1-p)}{n} + p^2\right] \\
&= \frac{n-1}{n} p(1-p) = \frac{n-1}{n} g(p).
\end{aligned}$$

Let $g(t) = t(1-t)$.
Now we want to find an unbiased estimator for $g(p) = p(1-p)$, i.e. the variance of X.

Remark again that in general $g(\hat{p})$ is NOT unbiased for $g(p)$, even though $\hat{p}$ is unbiased for p.

According to the theorem we discussed last time,
the sample variance of the n data is the unbiased estimator for the variance of X.
That is, $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$.

Again this result suggests that we should use $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
instead of $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ to estimate the population variance,
because in the sense of unbiasedness, S_{n-1}^2 is better than S_n^2.

Based on the property of unbiasedness, we can only consider the estimator that its mean is equal to the parameter being estimated (i.e. the unbiased estimator).

However, there are still many unbiased estimators of an unknown parameter. Then, **how to compare the performances among unbiased estimators?**

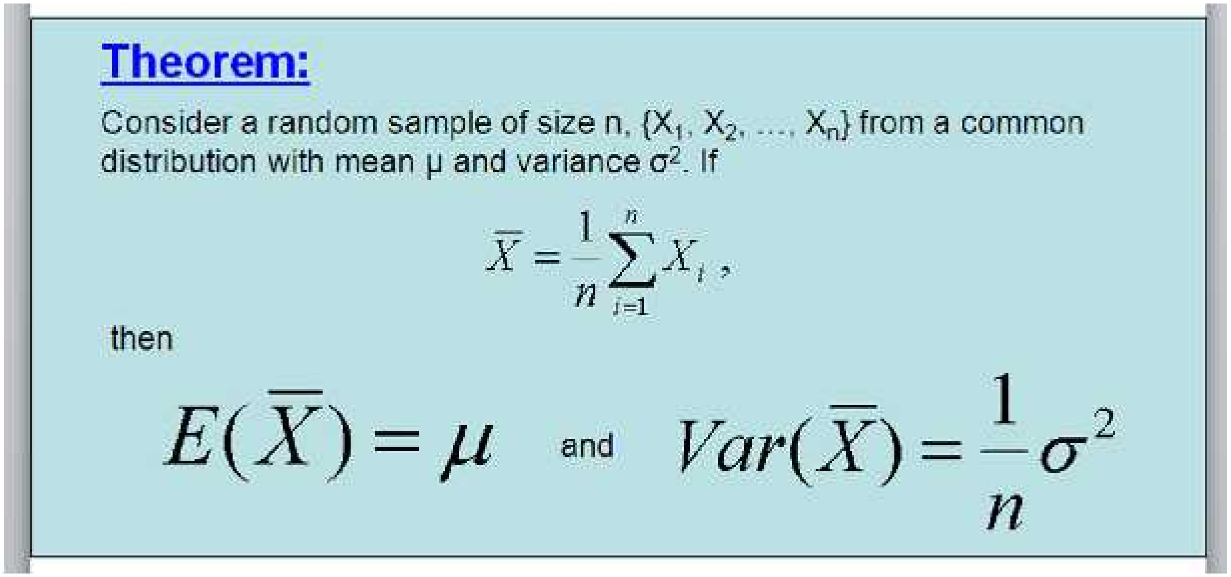
Comparison of unbiased estimators

If there are two unbiased estimators of θ, say $\hat{\theta}_1$ and $\hat{\theta}_2$, then we can **consider their variances**, and prefer the unbiased estimator with smaller variance.

For instance, if $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$, then we prefer $\hat{\theta}_1$, and say that $\hat{\theta}_1$ **is more efficient than** $\hat{\theta}_2$.

Comparison of unbiased estimators

From the previous theorem, we can see that for a random sample of size n, **if n increases, then the variance of the sample mean will decrease**, so the efficiency of the sample mean will increase.



Theorem:

Consider a random sample of size n, $\{X_1, X_2, \ldots, X_n\}$ from a common distribution with mean μ and variance σ^2. If

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

then

$$E(\bar{X}) = \mu \quad \text{and} \quad Var(\bar{X}) = \frac{1}{n}\sigma^2$$

→ A sample mean of 1000 data is more efficient than a sample mean of 100 data.

That's a reason to support that we should draw a sample as many as possible.

Problem with point estimation

The point estimator can only produce a single estimated value of the unknown parameter. Indeed, we can seldom estimate the true value of the unknown parameter correctly. That is, it is almost impossible to have a sample mean exactly equal to the population mean.

Also, as mentioned before, an estimator is a function of a random sample, so we will get a different single value of the unknown parameter when we draw another sample.

If $\hat{\theta}(x_1,\ldots,x_n)=1.2$, then we cannot conclude that the true value of θ is equal or close to 1.2 or not.

For another observed values (x_1’,…, x_n’),

$$\hat{\theta}(x_1',\ldots,x_n')=5.8$$

Then, what is the true value of θ? 1.2 or 5.8? close to 1.2 or 5.8? or in between or far away from these two numbers?

Useless? Why still consider the point estimator?

Problem: Come from the variability of the estimator.

In addition to the (point) estimated value of the estimator, some statisticians suggest that we should also consider the variance of the estimator.

How?

Use the single value and the variance of the estimator to form an interval that has a high probability to cover the true value of the unknown parameter.

This method including the variance of the point estimator is called interval estimation, or "confidence interval".