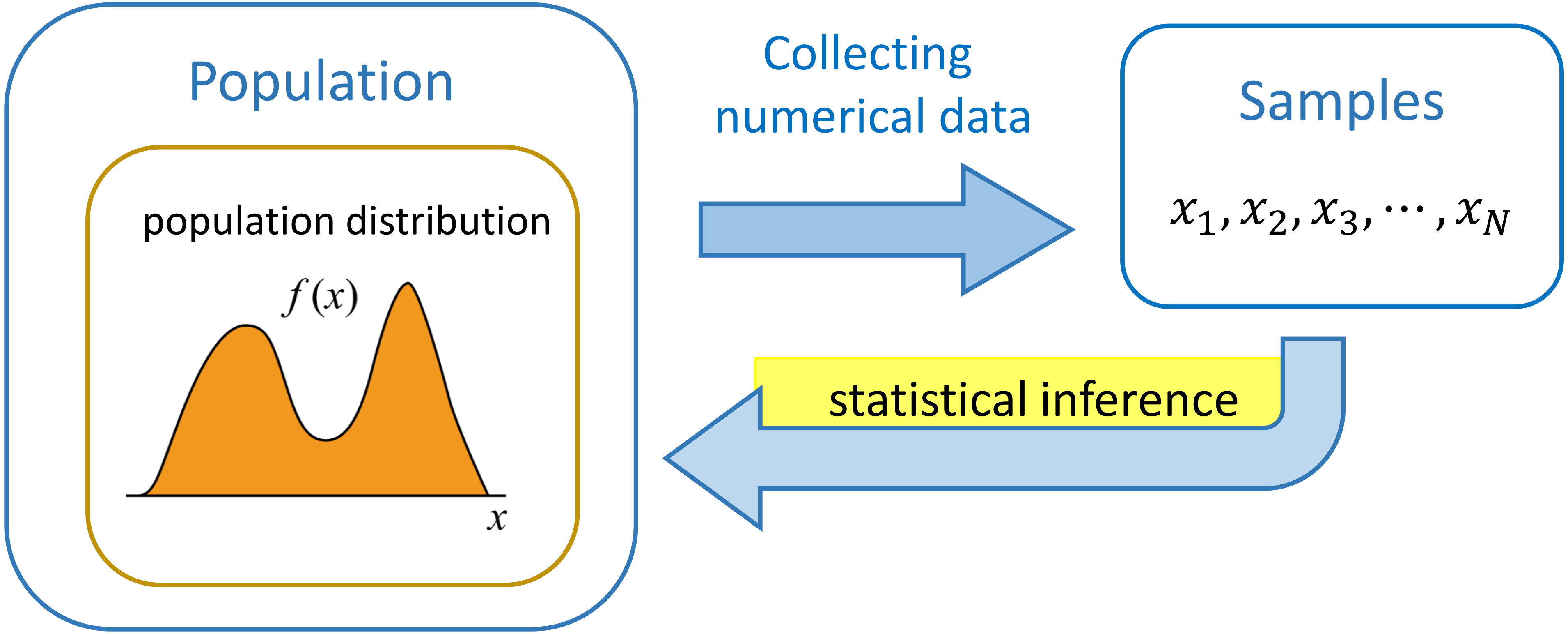
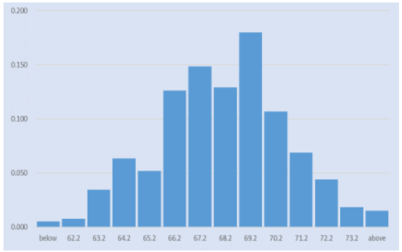
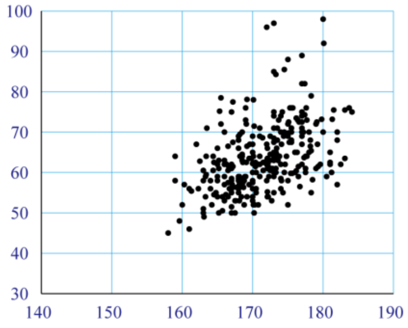
Lecture 2

Statistical Inference

1. Basic problem of statistical inference

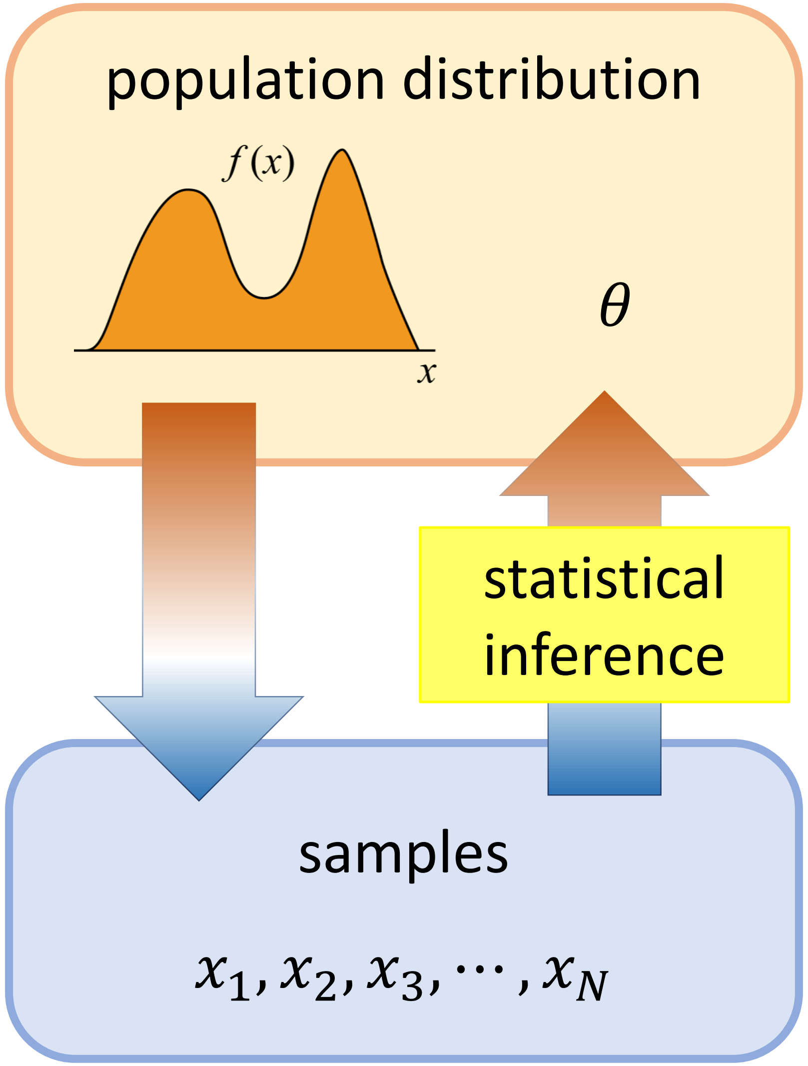






- ✓ The population distribution is our ultimate target.
- ✓ However, it is not reasonable to discuss the full information about it.
- ✓ We will estimate some parameters of the population distribution,
  e.g., mean value, variance, maximal value, correlation coefficient, ...

2. Point estimates of population parameters



θ: a parameter of the population distribution

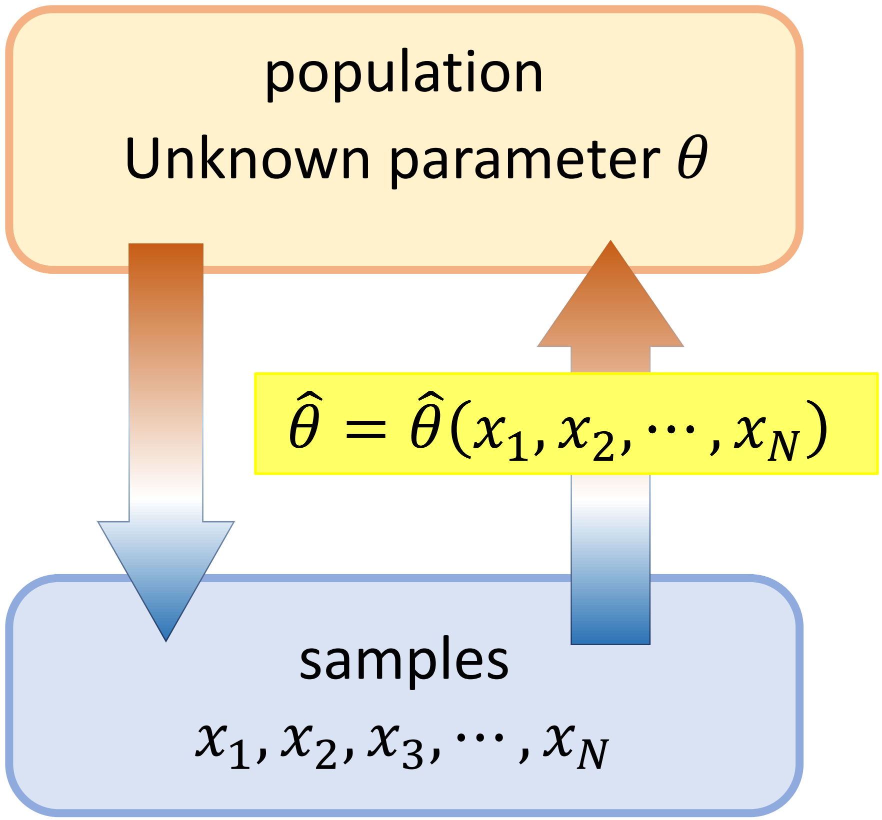
We wish to find a function

$$\hat{\theta} = f(x_1, x_2, x_3, \cdots, x_N)$$

such that $\hat{\theta}$ is a *reasonable estimate* of the unknown parameter θ

We need to measure the difference between θ and $\hat{\theta}$

3. Unbiased estimates



As usual, we adopt
random sampling with replacement

- ✓Sample data $x_1, x_2, x_3, \cdots, x_N$ are definite values once a sampling is performed.
- ✓However, they vary with each sampling though the population remains the same.
- ✓Moreover, because each sample x_i is a result of *random sampling,* it is modeled by a random variable X_i *obeying the population distribution*.
- ✓Thus, the sample data $x_1, x_2, x_3, \cdots, x_N$ are considered as realized values of the *independent random variables* $X_1, X_2, X_3, \cdots, X_N$
- ✓Therefore, $\hat{\theta}$ should be considered as a random variable $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_N)$.

Definition $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_N)$ is called an unbiased estimator of θ if $\mathbf{E}[\hat{\theta}] = \theta$.

4. Sample mean and unbiased variance

For random samples $X_1, X_2, X_3, \cdots, X_N$ the *sample mean* is defined by

$$\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$$

The *unbiased variance* is defined by

$$U^2 = \frac{1}{N-1}\sum_{k=1}^{N}(X_k - \bar{X})^2$$

Theorem The sample mean $\bar{X}$ is an unbiased estimator of the mean value of the population:

$$\mathbf{E}[\bar{X}] = m$$

where m is the mean value of the population.

Theorem The unbiased variance U^2 is an unbiased estimator of the variance of the population:

$$\mathbf{E}[U^2] = \sigma^2$$

where σ^2 is the variance of the population.

5. Better estimates

Let $X_1, X_2, X_3, \cdots, X_N$ be random samples.

1. We have seen that the sample mean

$$\bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$$

is an unbiased estimator: $\mathbf{E}[\bar{X}] = m$.

2. The weighted mean

$$\hat{m} = \sum_{k=1}^{n} a_k X_k, \qquad \sum_{k=1}^{N} a_k = 1$$

is also an unbiased estimator.

In fact,

$$\mathbf{E}[\hat{m}] = \sum_{k=1}^{N} a_k \mathbf{E}[X_k] = m$$

3. Which is better?

Definition Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ, i.e., $\mathbf{E}[\hat{\theta}_1] = \mathbf{E}[\hat{\theta}_2] = \theta$. We say that $\hat{\theta}_1$ is better than $\hat{\theta}_2$ if

$$\mathbf{E}\left[\left(\hat{\theta}_1 - \theta\right)^2\right] \leq \mathbf{E}\left[\left(\hat{\theta}_2 - \theta\right)^2\right].$$

In general, $\mathbf{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbf{V}[\hat{\theta}]$ is called the *mean squared error*.

6. Sample mean is better than weighted mean

For the weighted mean $\widehat{m} = \sum_{k=1}^{n} a_k X_k$
we compute the mean squared error: $\mathbf{E}[(\widehat{m} - m)^2]$.

First note that

$$\widehat{m}^2 = \sum_{j=1}^{n} a_j X_j \sum_{k=1}^{n} a_k X_k$$

$$= \sum_{j \neq k} a_j a_k X_j X_k + \sum_{k=1}^{n} a_k^2 X_k^2$$

and

$$\mathbf{E}[\widehat{m}^2] = \sum_{j \neq k} a_j a_k \, \mathbf{E}[X_j X_k] + \sum_{k=1}^{n} a_k^2 \mathbf{E}[X_k^2]$$

$$= \sum_{j \neq k} a_j a_k \, \mathbf{E}[X_j]\mathbf{E}[X_k] + \sum_{k=1}^{n} a_k^2 \, (\sigma^2 + m^2)$$

$$= \sum_{j \neq k} a_j a_k \, m^2 + \sum_{k=1}^{n} a_k^2 \, (\sigma^2 + m^2)$$

$$= m^2 \sum_{j,k=1}^{n} a_j a_k + \sigma^2 \sum_{k=1}^{n} a_k^2 = m^2 + \sigma^2 \sum_{k=1}^{n} a_k^2$$

Hence

$$\mathbf{E}[(\widehat{m} - m)^2] = \mathbf{E}[\widehat{m}^2] - m^2 = \sigma^2 \sum_{k=1}^{n} a_k^2$$

$$= \sigma^2 \left\{ \sum_{k=1}^{n} \left(a_k - \frac{1}{n} \right)^2 + \frac{2}{n} \sum_{k=1}^{n} a_k - n \times \frac{1}{n^2} \right\}$$

$$= \sigma^2 \sum_{k=1}^{n} \left(a_k - \frac{1}{n} \right)^2 + \frac{\sigma^2}{n}$$

This attains minimum when $a_k = 1/n$ (sample mean).

Exercise 3 (5 min)

A coin is tossed N times and then we obtain samples

$$X_1, X_2, X_3, \cdots, X_N \qquad \text{where } X_k = \begin{cases} 1, & \text{if heads} \\ 0, & \text{if tails} \end{cases}$$

Find an unbiased estimator of the probability that a head occurs.

Exercise 4 (10 min)

A coin is tossed until a head occurs $K\ (\geq 2)$ times. Let N be the total number of tosses.

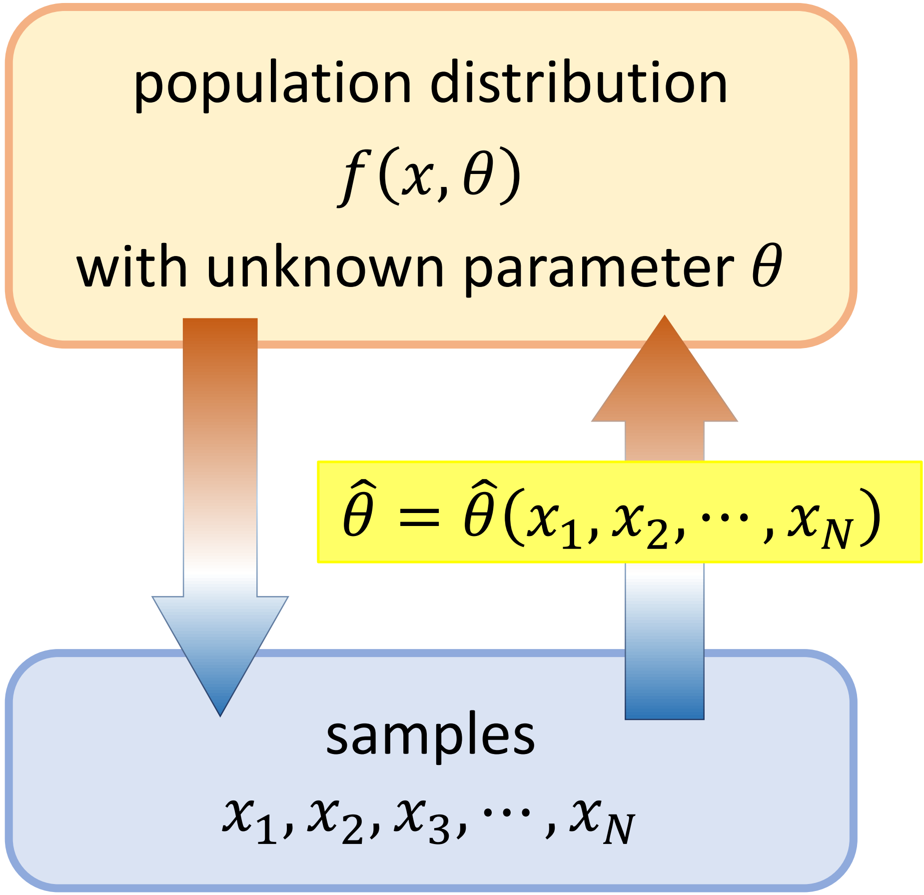
Show that $\hat{p} = \frac{K-1}{N-1}$ is an unbiased estimator of the probability that a head occurs.

[Hint] $$\sum_{n=0}^{\infty} \binom{n+k}{n} x^n = \frac{1}{(1-x)^{k+1}}$$

7. Maximum likelihood estimates

- It is desired to use the best unbiased estimator.
- The best unbiased estimators are known for many concrete cases, however, it is not straightforward to obtain the best unbiased estimator in general.
- Then the maximum likelihood method is employed because
  - ✓ the maximum likelihood estimator is rather easily derived;
  - ✓ and in many cases it is the best unbiased estimator.

8. Formulation



Examples:

(1) Binomial population – opinion poll, etc.

The population distribution is the Bernoulli distribution $B(1, p)$, so the unknown parameter is p.

(2) Poisson population – counting rare events

The population distribution $\mathrm{Po}(\lambda)$ contains just one unknown parameter λ.

(3) Normal population – Population consisting of individuals sharing a lot of small fluctuations

The normal distribution $N(m, \sigma^2)$ contains two unknown parameters m and σ^2.

9. Likelihood functions

Let $f(x, \theta)$ be the population distribution, where θ is unknown.

The *likelihood function* is defined by

$$L(x_1, x_2, \cdots, x_N; \theta) = \prod_{k=1}^{N} f(x_k, \theta)$$

Basic idea:

- Given a set of sample data $x_1, x_2, \cdots, x_N$, we determine the value θ which maximizes $L(x_1, x_2, \cdots, x_N; \theta)$.
- Roughly, we presume that the occurring event (the data $x_1, x_2, \cdots, x_N$) is the one that have the highest probability among many other candidates.

The maximum value of L is found by solving

$$\frac{\partial L}{\partial \theta} = 0.$$

Such a value $\theta = \hat{\theta}$ is called the *maximum likelihood estimator*.

It is often more convenient to consider the *log-likelihood function*:

$$\log L(x_1, x_2, \cdots, x_N; \theta) = \sum_{k=1}^{N} \log f(x_k, \theta)$$

Then

$$\frac{\partial \log L}{\partial \theta} = \sum_{k=1}^{N} \frac{\partial}{\partial \theta} \log f(x_k, \theta).$$

Example: Binomial population

We consider

$$f(x,p) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \end{cases} = p^x(1-p)^{1-x}$$

where p is unknown and to be estimated.

The likelihood function is defined by

$$L(x_1, x_2, \cdots, x_N; p) = \prod_{k=1}^{N} f(x_k, p) = \prod_{k=1}^{N} p^{x_k}(1-p)^{1-x_k}$$

The log-likelihood function is defined by

$$\log L = \sum_{k=1}^{N} \{x_k \log p + (1-x_k)\log(1-p)\}$$

Then we have

$$\frac{\partial}{\partial p}\log L = \sum_{k=1}^{N}\left\{x_k \frac{1}{p} + (1-x_k)\frac{-1}{1-p}\right\}$$

$$= \sum_{k=1}^{N} \frac{x_k - p}{p(1-p)}$$

and solving $\frac{\partial}{\partial p}\log L = 0$, we obtain the maximum likelihood estimator:

$$\hat{p} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

which is known as the sample mean and is an unbiased estimator too.

Problem 5

A number is chosen randomly from the interval $[0, a]$, where $a > 0$ is unknown constant. We obtain N samples

$$X_1, X_2, X_3, \cdots, X_N$$

Set

$$\hat{a} = 2\bar{X} = \frac{2}{N}\sum_{k=1}^{N} X_k \quad \text{and} \quad Y = \frac{N+1}{N}\max\{X_1, X_2, X_3, \cdots, X_N\}$$

(1) Show that both $\hat{a}$ and Y are unbiased estimators of a.

(2) Which estimator is better $\hat{a}$ or Y ?

Problem 6

There are many runners on the street. Each runner is given a number cloth starting with 1 to a certain unknown big number N. Snap shots are taken at random.

(1) In a snap shot there are 3 runners with numbers X_1, X_2, X_3. Set $M = \max\{X_1, X_2, X_3\}$. Prove that

$$P(M = k) = \frac{3(k-1)(k-2)}{N(N-1)(N-2)}$$

(2) Calculating the mean value $\mathbf{E}[M]$, show that $\frac{4}{3}M - 1$ is an unbiased estimator of N.

(3) [Challenge]In a snap shot there are n runners with numbers $X_1, X_2, \cdots, X_n$. Show that $\left(1 + \frac{1}{n}\right)M - 1$ is an unbiased estimator of N, where $M = \max\{X_1, X_2, \cdots, X_n\}$.

Problem 7

There is a deck of N cards and a unique number from 1 to N is assigned to each card as a reward. Three cards are drawn randomly with replacement, say, X_1, X_2, X_3. Consider their mean and median:

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3), \quad M = \text{median}\{X_1, X_2, X_3\}$$

(1) Show that $\bar{X}$ is an unbiased estimator of the mean of the reward.

(2) Show that M is also an unbiased estimator of the mean of the reward.

(3) Calculating the mean squared errors, determine the better estimator.

Problem 8

The exponential distribution with parameter $\lambda > 0$ is defined by the density function:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Then the likelihood function is defined by

$$L(x_1, x_2, \cdots, x_N; \lambda) = \prod_{k=1}^{N} f(x_k, \lambda).$$

Find the maximum likelihood estimator $\hat{\lambda}$ of λ.