QUIZ

Fill in the blanks below:

For a large random sample of size n from a distribution with mean μ and standard deviation σ, sample mean has an approximately .... distribution with mean ...... and standard deviation ....

QUIZ

Fill in the blanks below:

For a large random sample of size n from a distribution with mean μ and standard deviation σ, sample mean has an approximately ……normal……… distribution with mean …μ…. and standard deviation …$\sigma/\sqrt{n}$…

(Say σ^2 is variance of population.
Variance of $\bar{x}$ is $\frac{\sigma^2}{n}$.
std. dev. " " " $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$)

Outline

- ➢ We covered sampling distribution of means and CLT (8.1-8.4).
- ➢ Sampling distribution of variance and t-distribution (8.5, 8.6): We will cover them when discussing estimation problems.

Today

- ➢ We will continue with some more examples of CLT.
- ➢ Next we will discuss estimation problems: Chapter 9

Another Example on Central Limit Theorem

Example: X_i: customer spending with $\mu = 80, \sigma = 40$.

Approximate the probability that the average spending of 100 customers is $10\,\%$ or more below average. Use norm.cdf(-2)=0.023.

Solution:

Another Example on Central Limit Theorem

Example: X_i: customer spending with $\mu = 80$, $\sigma = 40$.

Approximate the probability that the average spending of 100 customers is $10\,\%$ or more below average. Use norm.cdf(-2)=0.023.

Solution:

$P(\bar{X} \leq 72) = ?$

Another Example on Central Limit Theorem

Example: X_i: customer spending with $\mu = 80, \sigma = 40.$

Approximate the probability that the average spending of 100 customers is $10\,\%$ or more below average. Use norm.cdf(-2)=0.023.

Solution:

$$P(\bar{X} \leq 72) = ?$$

$$P\left(\frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \leq \frac{72 - 80}{\frac{40}{\sqrt{100}}}\right)$$

z

Another Example on Central Limit Theorem

Example: X_i: customer spending with $\mu = 80, \sigma = 40.$

Approximate the probability that the average spending of 100 customers is $10\,\%$ or more below average. Use norm.cdf(-2)=0.023.

Solution:

$$P(\bar{X} \le 72) = ?$$

$$P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \le \frac{72-80}{\frac{40}{\sqrt{100}}}\right) = P\left(Z \le \frac{-8 \times 10}{40}\right)$$

$\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \rightarrow Z$

$$= P(Z \le -2)$$

$$= 0.023$$

CLT and Difference Between Two Means

Revisit:

Linear combinations/transformations of normal independent variables:

If X_1, X_2 are independent and each is normally distributed

then $Y = a_1X_1 + a_2X_2 + b$ has a normal distribution.

Its mean is $a_1\mu_1 + a_2\mu_2 + b$ and its variance is $a_1^2\sigma_1^2 + a_2^2\sigma_2^2$.

CLT and Difference Between Two Means

Apply it to the difference of two sample means:

If independent samples of size n_1, n_2 are drawn at random from two populations, with means μ_1, μ_2 and variances σ_1^2, σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean $\mu_1 - \mu_2$ and variance $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

Example on CLT and Difference Between Two Means

Case Study 8.2: **Paint Drying Time**: Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$. Use norm.cdf(3)=0.9987.

Solution:

Example on CLT and Difference Between Two Means

Case Study 8.2: **Paint Drying Time**: Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$. Use norm.cdf(3)=0.9987.

Solution:

Let $Y = \bar{X}_A - \bar{X}_B$

$\mu_Y = 0$

$\sigma_Y^2 = \frac{1}{18} + \frac{1}{18} = \frac{1}{9} \Rightarrow \sigma_Y = \frac{1}{3}$

Example on CLT and Difference Between Two Means

Case Study 8.2: **Paint Drying Time**: Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$. Use norm.cdf(3)=0.9987.

Solution:

Let $Y = \bar{X}_A - \bar{X}_B$

$\mu_Y = 0$

$\sigma_Y^2 = \frac{1}{18} + \frac{1}{18} = \frac{1}{9} \Rightarrow \sigma_Y = \frac{1}{3}$

Assuming we can apply CLT on $\bar{X}_A, \bar{X}_B$

$\Rightarrow Y$ has normal distribution.

$P(Y > 1.0) = P\left(\frac{Y - \mu_Y}{\sigma_Y} > \frac{1.0 - 0}{\frac{1}{3}}\right)$ → Z

Example on CLT and Difference Between Two Means

Case Study 8.2: **Paint Drying Time**: Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$. Use norm.cdf(3)=0.9987.

Solution:

Let $Y = \bar{X}_A - \bar{X}_B$

$\mu_Y = 0$

$\sigma_Y^2 = \frac{1}{18} + \frac{1}{18} = \frac{1}{9} \Rightarrow \sigma_Y = \frac{1}{3}$

Assuming we can apply CLT on $\bar{X}_A, \bar{X}_B$

$\Rightarrow Y$ has normal distribution.

$P(Y > 1.0) = P\left(\frac{Y - \mu_Y}{\sigma_Y} > \frac{1.0 - 0}{\frac{1}{3}}\right)$ ($\frac{Y - \mu_Y}{\sigma_Y} \to Z$)

$= P(Z > 3)$

$= 1 - P(Z \leq 3) = 1 - 0.9987 = 0.0013$

Outline

- ➢ We covered sampling distribution of means and CLT (8.1-8.4).
- ➢ Sampling distribution of variance and t-distribution (8.5, 8.6): We will discuss them when relevant in estimation problems.

Today

- ➢ We will continue with some more examples of CLT.
- ➢ Next we will discuss estimation problems: Chapter 9

Statistical Inference and Estimation Problems

Probability:

Distribution $\longrightarrow$ Samples

Given distribution find probabilities of data/events.

Ex: X has distribution $Binomial(X; n, p)$

Probability of $x = 3$ successes in $n = 10$ trials with $p = 0.7$?

Statistical Inference and Estimation Problems

Probability:

Distribution $\longrightarrow$ Samples

Given distribution find probabilities of data/events.

Ex: X has distribution $Binomial(X; n, p)$

Probability of $x = 3$ successes in $n = 10$ trials with $p = 0.7$?

Statistics:

Sample $\longrightarrow$ Distribution

Given data find parameters/properties of distribution.

Ex: We observed $X_1 = 0,\ X_2 = 1,\ \cdots,\ X_{10} = 0$

What is the distribution parameter p , i.e. probability of heads?

Point Estimation

Point Estimate:

Find single “good estimate” of a quantity of interest of a distribution/population using statistics.

Statistics:

Any function of the sample: average, max, max-min, $\cdots$

Point Estimation

Formally, $X_1, \cdots, X_n$ iid data points from some distribution. An estimate of parameter θ of the distribution is:

$$\widehat{\Theta} = r(X_1, \cdots, X_n), \quad \text{for some appropriate function } r.$$

Note 1: A single value of $\widehat{\Theta}$ is denoted with $\hat{\theta}$.

Note 2: θ is considered fixed, unknown quantity. $\widehat{\Theta}$ is a random variable.

Point Estimation

Ex: Estimate $\theta = \mu = \sum_x x\, f(x)$ of an unknown distribution.

Say, true (unknown) value $\theta = 2.5$.

Sample 4 data points X_1, X_2, X_3, X_4, say $3, 6, 5, -2$.

Point Estimation

Ex: Estimate $\theta = \mu = \sum_x x\, f(x)$ of an unknown distribution.

Say, true (unknown) value $\theta = 2.5$.

Sample 4 data points X_1, X_2, X_3, X_4, say $3, 6, 5, -2$.

We can try to estimate θ with any function of $X_1, \cdots, X_4$:

$\widehat{\Theta}$: $\quad \dfrac{X_1 + \cdots + X_n}{n} \quad \dfrac{min(X_1, \cdots, X_n) + max(X_1, \cdots, X_n)}{2} \quad X_1 \cdot X_n$

$\widehat{\theta}$ values:

Point Estimation

Ex: Estimate $\theta = \mu = \sum_x x\, f(x)$ of an unknown distribution.

Say, true (unknown) value $\theta = 2.5$.

Sample 4 data points X_1, X_2, X_3, X_4, say $3, 6, 5, -2$.

We can try to estimate θ with any function of $X_1, \cdots, X_4$:

$\widehat{\Theta}$:	$\dfrac{X_1 + \cdots + X_n}{n}$	$\dfrac{min(X_1, \cdots, X_n) + max(X_1, \cdots, X_n)}{2}$	$X_1 \cdot X_n$
$\widehat{\theta}$ values:	3	2	-6

Unbiased Estimator

A desired property of an estimator:

A statistic $\hat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Unbiased Estimator

A desired property of an estimator:

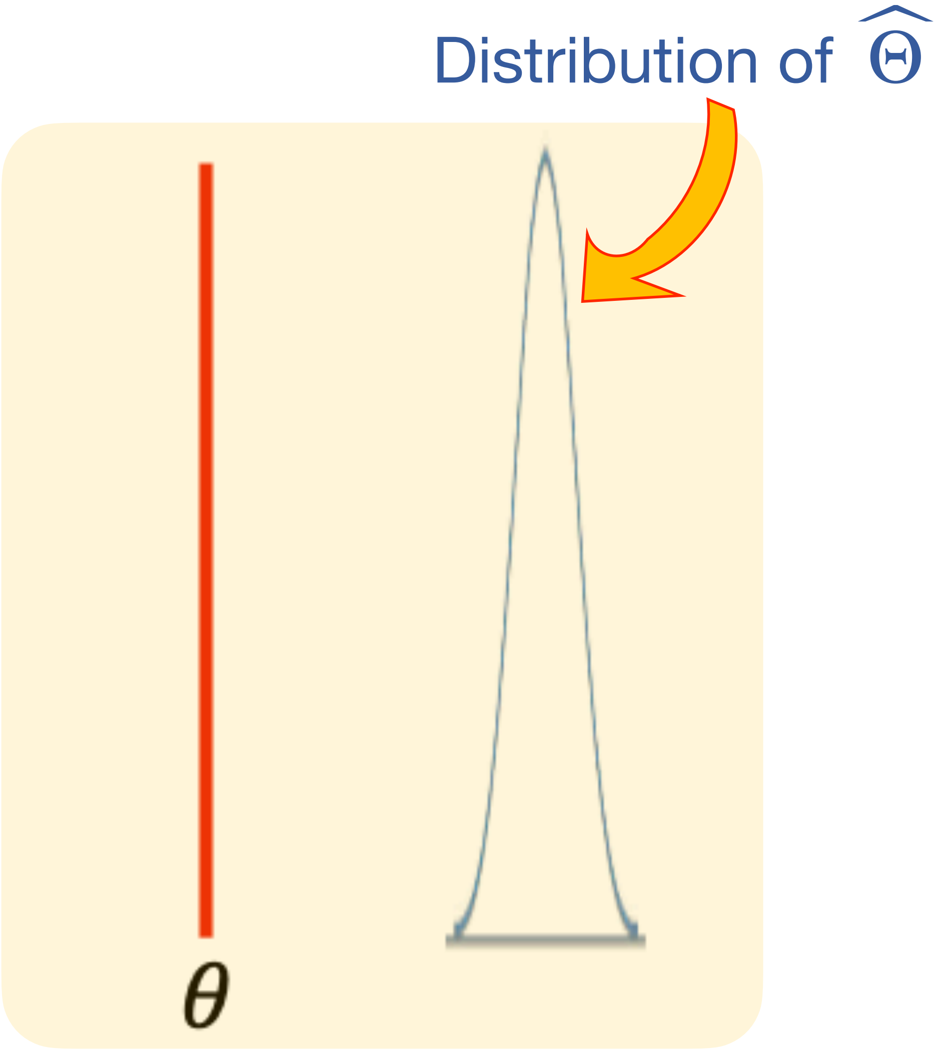
A statistic $\hat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Multiple unbiased estimators:

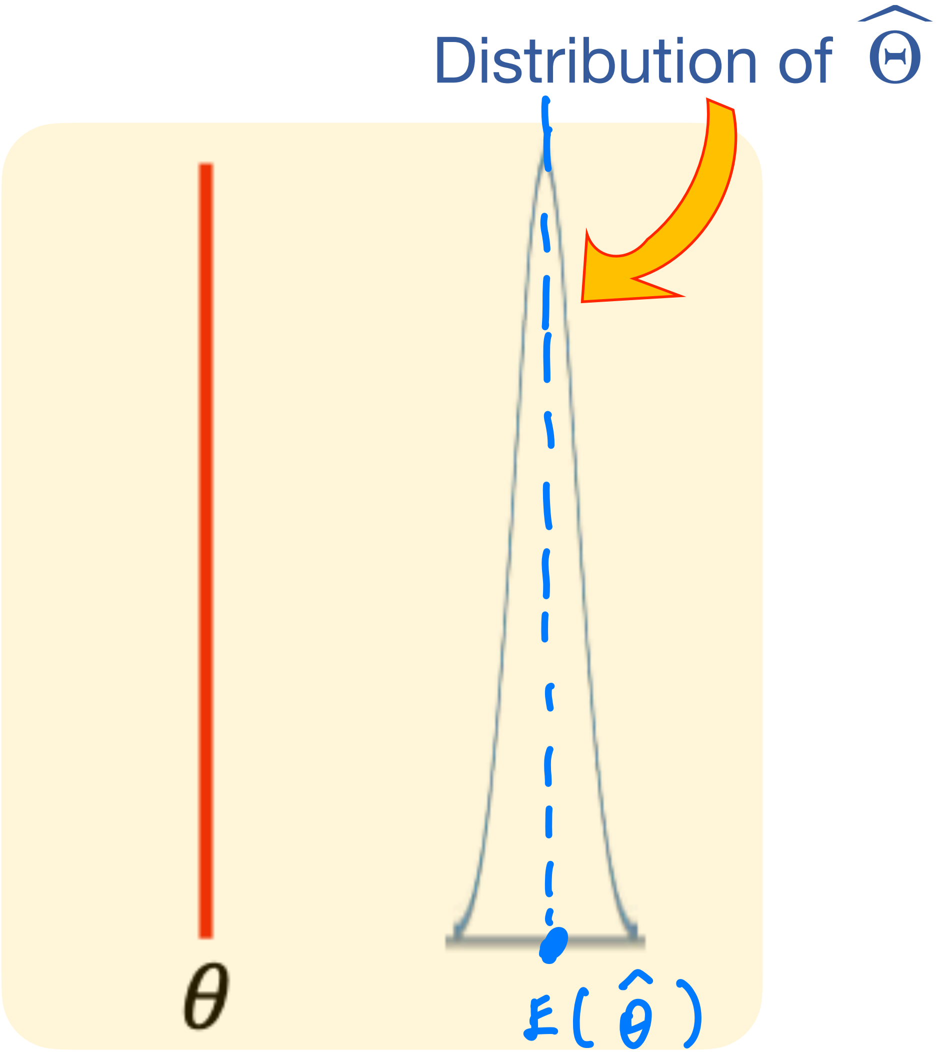
If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the **most efficient estimator** of θ.

Point Estimation



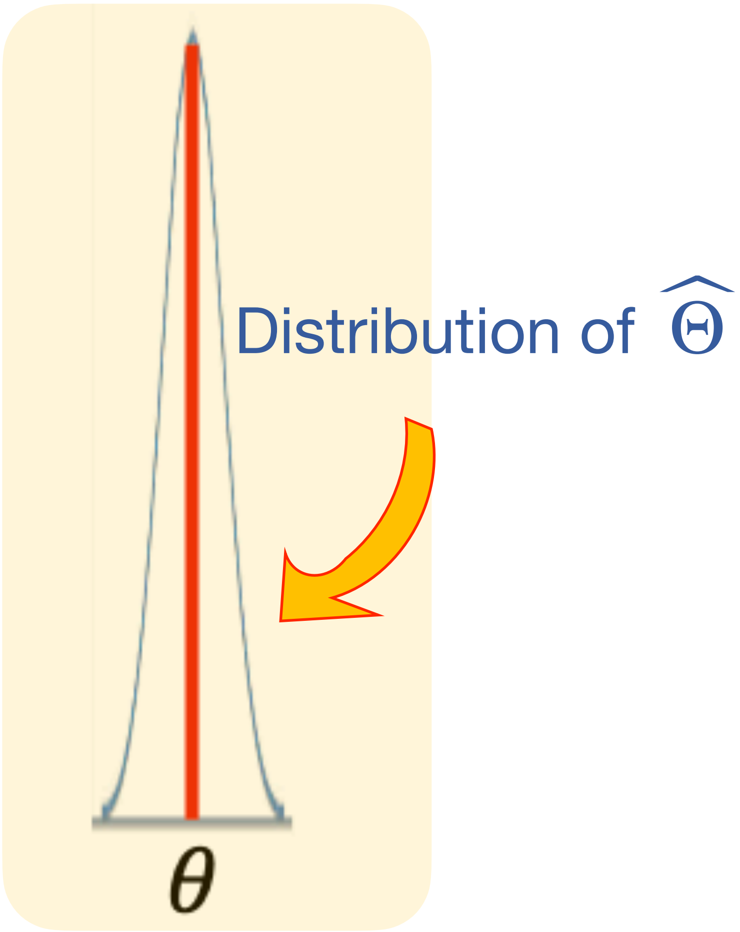
Biased

Point Estimation

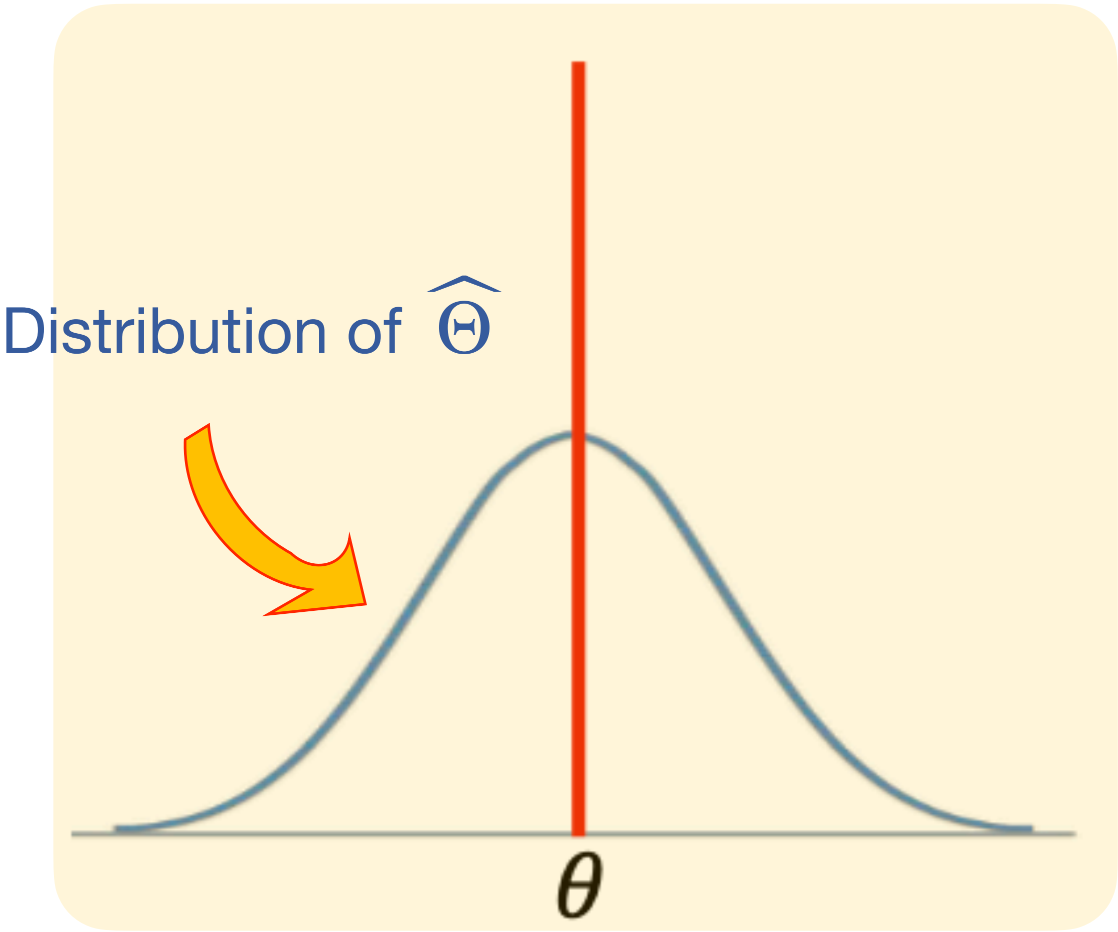


Biased

Point Estimation

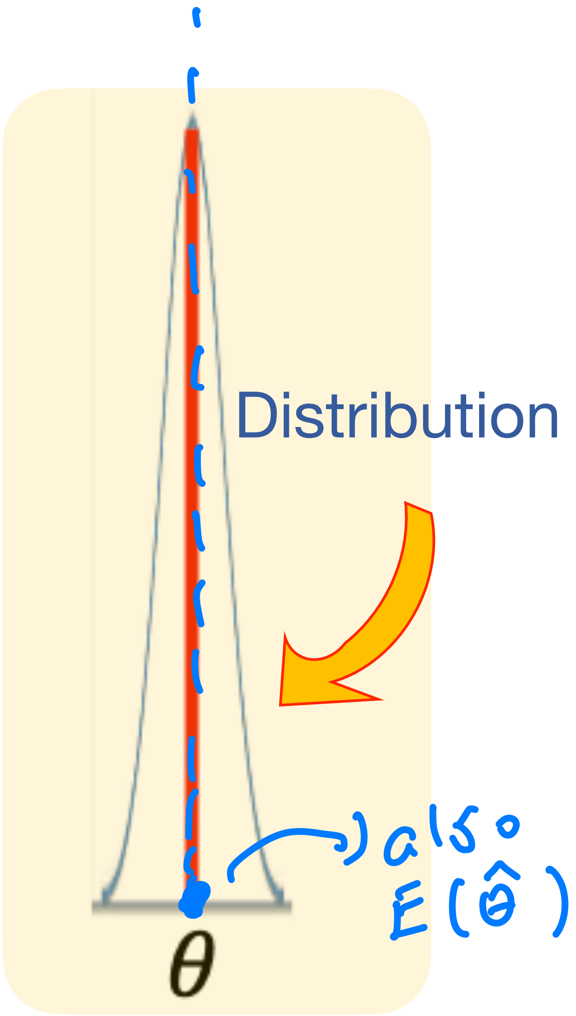


Unbiased
Small variance

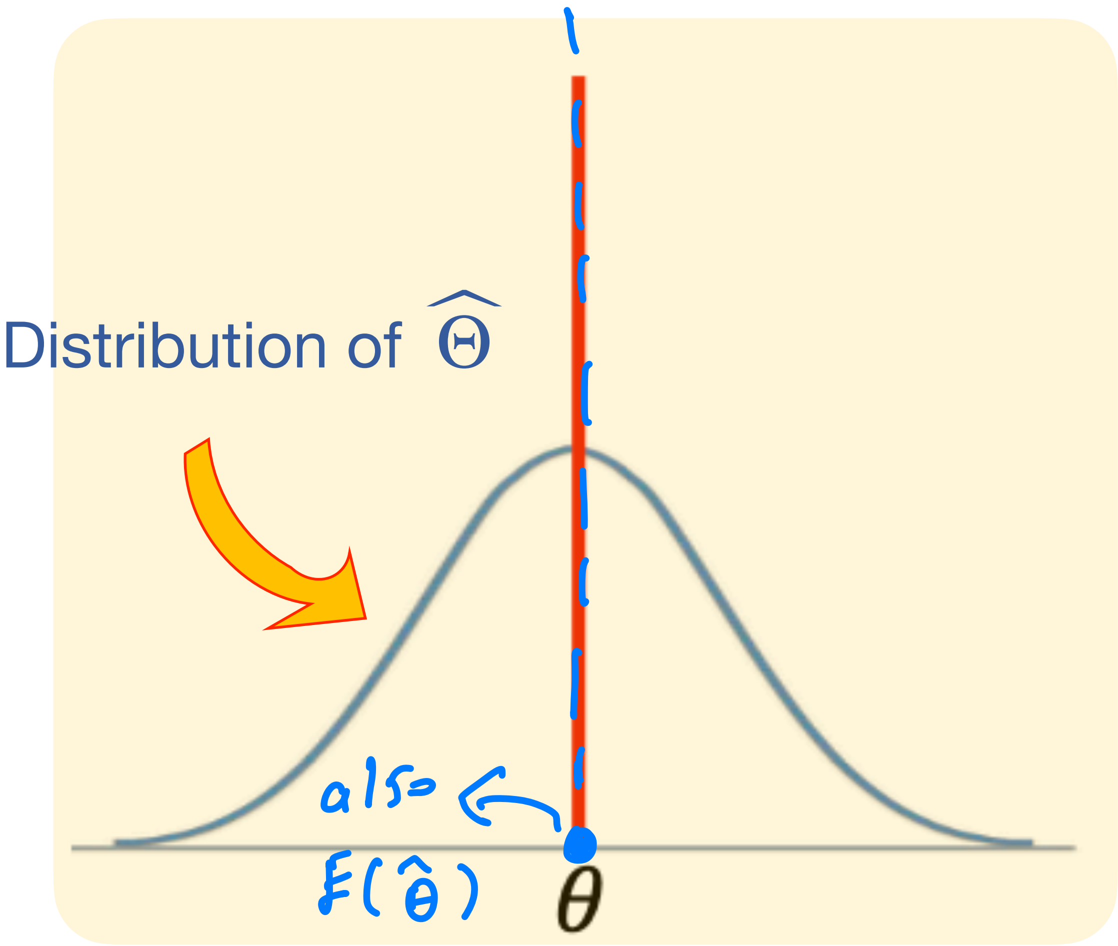


Unbiased
Large variance

Point Estimation

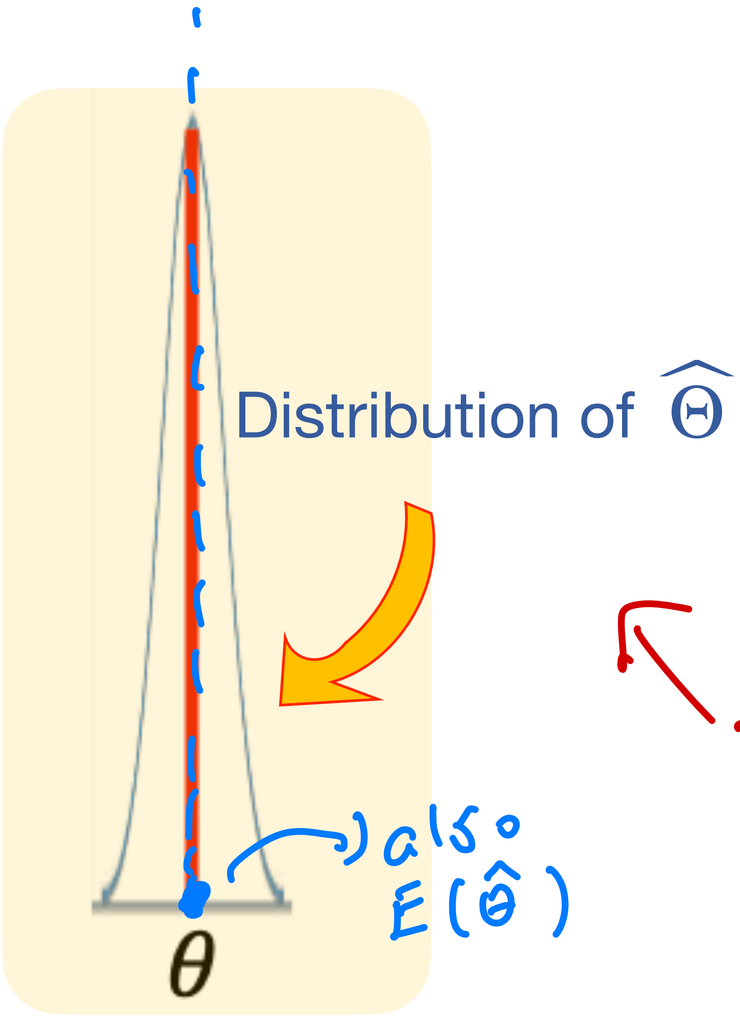


Unbiased
Small variance



Unbiased
Large variance

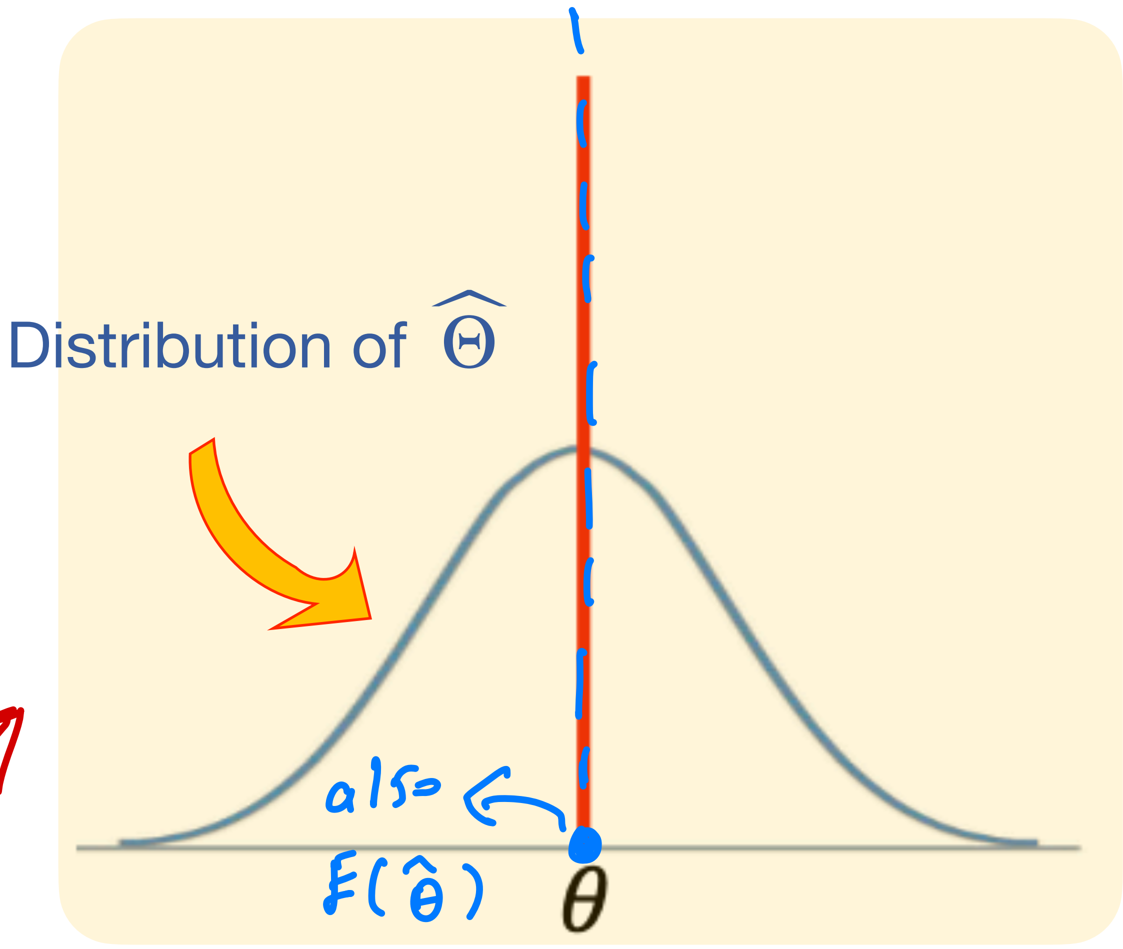
Point Estimation



Unbiased

Small variance

This is more efficient than this



Unbiased

Large variance

Point Estimation

Example: Observe n coin flips $X_1, \cdots, X_n \sim Bernoulli(p)$.

True value of p unknown. We want to estimate it.

Possible estimator: Sample mean $\overline{X} = \frac{1}{n} \sum X_i$

Is it biased? $E(\bar{x}) \stackrel{?}{=} p$

Point Estimation

Example: Observe n coin flips $X_1, \cdots, X_n \sim Bernoulli(p)$.

True value of p unknown. We want to estimate it.

Possible estimator: Sample mean $\overline{X} = \dfrac{1}{n} \sum X_i$

Is it biased? $E(\bar{x}) \overset{?}{=} p$

$$E(\bar{x}) = \frac{1}{n} \sum \underbrace{E(x_i)} = \frac{n \cdot p}{n} = p$$

$\Rightarrow$ Sample mean is an unbiased estimator of p.

Point Estimation

Example: Observe n coin flips $X_1, \cdots, X_n \sim Bernoulli(p)$.

True value of p unknown. We want to estimate it.

Possible estimator: Sample mean $\bar{X} = \frac{1}{n} \sum X_i$

Is it biased? $E(\bar{x}) \stackrel{?}{=} p$

$E(\bar{x}) = \frac{1}{n} \sum \underbrace{E(x_i)} = \frac{n \cdot p}{n} = p$

$\Rightarrow$ Sample mean is an unbiased estimator of p.

(In general, sample mean is an unbiased estimator of μ, since $E(\bar{X}) = \mu$, for any distribution.)