6 Point Estimation

Stat 4570/5570

Material from Devore's book (Ed 8), and Cengage

Point Estimation

Statistical inference: directed toward conclusions about one or more parameters. We will use the generic Greek letter θ for the parameter of interest.

Process:

- Obtain sample data from each population under study
- Based on the sample data, estimate θ
- Conclusions based on sample estimates.

The objective of **point estimation** = estimate θ

Some General Concepts of Point Estimation

A **point estimate** of a parameter θ is a value (based on a sample) that is a sensible guess for θ.

A point estimate is obtained by a formula ("estimator") which takes the sample data and produces an point estimate.

Such formulas are called **point estimators** of θ.

Different samples produce different estimates, even though you use the same estimator.

Example

20 observations on breakdown voltage for some material:

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

Assume that after looking at the histogram, we think that the distribution of breakdown voltage is normal with mean value μ. What are some point estimators for μ?

Estimator “quality”

“Which estimator is the best?”

What does “best” mean?

Estimator "quality"

In the best of all possible worlds, we could find an estimator $\hat{\theta}$ for which $\hat{\theta} = \theta$ always, in all samples. Why doesn't this estimator exist?

For some samples, $\hat{\theta}$ will sometimes be too big, and other times too small.

If we write $\hat{\theta} = \theta$ + error of estimation

then an accurate estimator would be one resulting in small estimation errors, so that estimated values will be near the true value. It's the distribution of these errors (over all samples) that actually matters for the quality of estimators.

Measures of estimator quality

A sensible way to quantify the idea of $\hat{\theta}$ being close to θ is to consider the squared error $(\hat{\theta} - \theta)^2$

and **the *mean squared error*** MSE = $E[(\hat{\theta} - \theta)^2]$.

If among two estimators, one has a smaller MSE than the other, the first estimator is usually the better one.

Another good quality is ***unbiasedness***: $E[(\hat{\theta})] = \theta$

Another good quality is **small variance**, $Var[(\hat{\theta})]$

Unbiased Estimators

- Suppose we have two measuring instruments; one instrument is accurately calibrated, and the other systematically gives readings smaller than the true value.
- When each instrument is used repeatedly on the same object, because of measurement error, the observed measurements will not be identical.
- The measurements produced by the first instrument will be distributed about the true value *symmetrically*, so it is called an unbiased instrument.
- The second one has a systematic bias, and the measurements are centered around the wrong value.

Example: unbiased estimator of proportion

If X denotes the number of sample successes, and has a binomial distribution with parameters n and p, then the sample proportion X / n can be used as an estimator of p.

Can we show that this is an unbiased estimator?

Estimators with Minimum Variance

Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of θ that are both unbiased. Then, although the distribution of each estimator is centered at the true value of θ, the spreads of the distributions about the true value may be different.

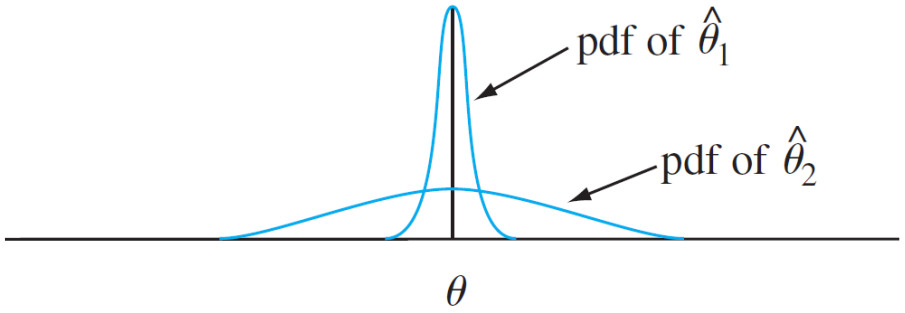
Among all estimators of θ that are unbiased, we will always choose the one that has minimum variance. WHY?

The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

Estimators with Minimum Variance

Figure below pictures the pdf's of two unbiased estimators, with $\hat{\theta}_1$ having smaller variance than $\hat{\theta}_2$.

Then $\hat{\theta}_1$ is more likely than $\hat{\theta}_2$ to produce an estimate close to the true θ. The MVUE is, in a certain sense, the most likely among all unbiased estimators to produce an estimate close to the true θ.



Graphs of the pdf's of two different unbiased estimators

Reporting a Point Estimate: The Standard Error

Besides reporting the value of a point estimate, some indication of its precision should be given.

The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. It is the magnitude of a typical or representative deviation between an estimate and the true value θ.

Basically, the standard error tells us roughly within what distance of true value θ the estimator is likely to be.

The Mean is unbiased

Note that the following result shows that the arithmetic average is unbiased:

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Then

1. $E(\overline{X}) = \mu$
2. $V(\overline{X}) = \sigma^2/n \quad \text{and} \quad \sigma_{\overline{X}} = \sigma/\sqrt{n}$

Thus we see that the arithmetic average is an unbiased estimator for the mean for any random sample of any size from any distribution.

General methods for constructing estimators

We have:

- a sample from a probability distribution ("the model")
- we don't know the parameters of that distribution

How do we find the parameters to best match our sample data?

Method 1: Methods of Moments (MoM):

1. equate sample characteristics (eg. mean, or variance), to the corresponding population values
2. solve these equations for unknown parameter values
3. the solution formula is the estimator (need to check bias).

Method 2: Maximum Likelihood Estimation (MLE)

Statistical Moments

For k = 1, 2, 3, . . . , define the ***k*-th population moment**, or ***k*-th moment of the distribution *f*(*x*),** to be $E(X^k)$.

and the ***k*-th sample moment** is $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$.

Thus the first population moment is $E(X) = \mu$, and the first sample moment is

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

The second population and sample moments are $E(X^2)$ and $M_2 = \Sigma X_i^2/n$, respectively.

The Method of Moments

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are parameters whose values are unknown.

Then the **moment estimators** $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$.

If, for example, $m = 2$, $E(X)$ and $E(X^2)$ will be functions of θ_1 and θ_2.

Setting $E(X) = M_1$ and $E(X^2) = M_2$ gives two equations in θ_1 and θ_2. The solution then defines the estimators.

Example for MoM

Let $X_1, X_2, \ldots, X_n$ represent a random sample of service times of n customers at a certain facility, where the underlying distribution is assumed exponential with parameter λ.

What is the MOM estimate for λ?

Example 2 for MoM

Let $X_1, X_2, \ldots, X_n$ represent a random sample from a Gamma distribution with parameters *a* and *b*.

How do we use MoM to estimate a and b?

MLE

Method 2: Maximum likelihood estimation (MLE)

The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s.

Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have many desirable mathematical properties.

Example for MLE

A sample of ten independent bike helmets just made in the factory A was up for testing. 3 helmets are flawed.

Let p = P(flawed helmet). The probability of X=3 is:

$$P(X=3) = C(10,3)\, p^3(1-p)^7$$

But the **likelihood function** is given as*:*

$$L\,(p \mid \textit{sample data}) = p^3(1-p)^7$$

Likelihood function = function of the parameter only.

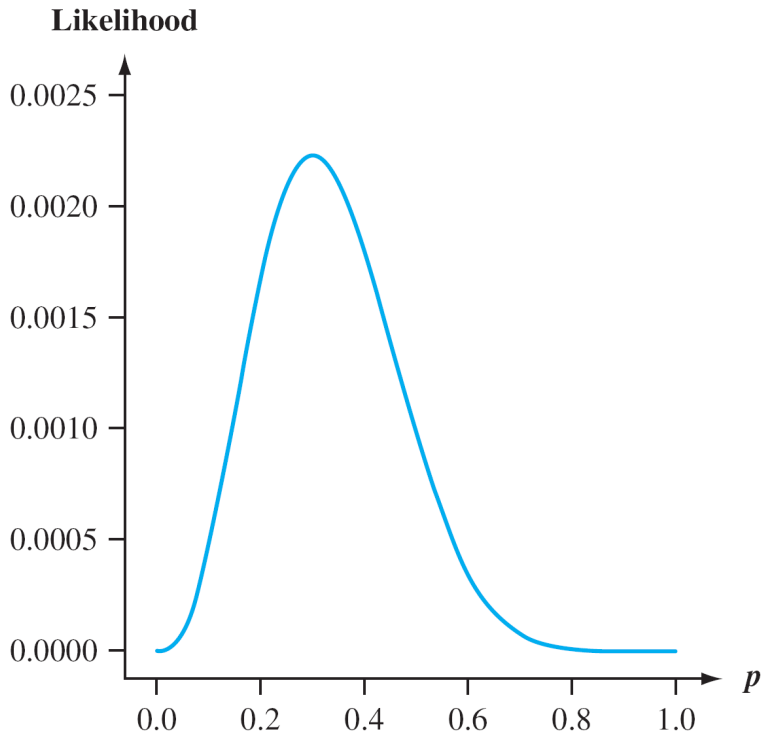
For what value of p is the obtained sample most likely to have occurred? bi.e., what value of p **maximizes** the likelihood?

Example MLE

cont' d

Graph of the *likelihood* function as a function of p:

$$L\,(p \mid \textit{sample data}) = p^3(1-p)^7$$



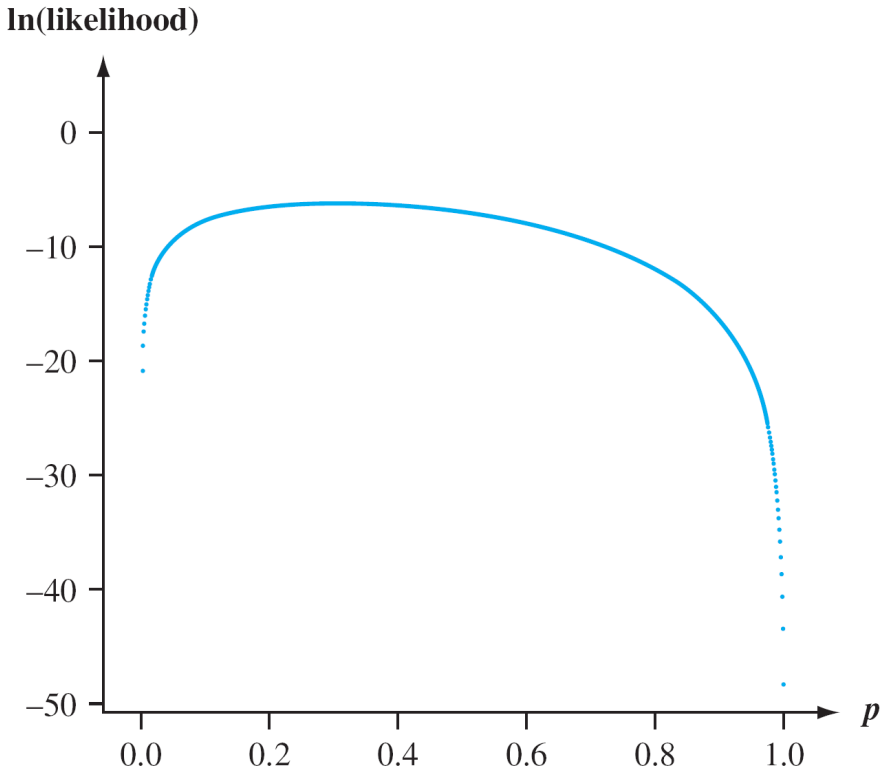
Example MLE

cont' d

The natural logarithm of the likelihood:

$\log (L (p \mid \text{sample data})) = l (p \mid \text{sample data}))$

$= 3 \log(p) + 7 \log(1 - p)$



Example MLE

cont' d

We can verify our visual guess by using calculus to find the actual value of *p* that maximizes the likelihood.

Working with the natural log of the likelihood is often easier than working with the likelihood itself. WHY?

How do you find the maximum of a function?

Example MLE

cont' d

That is, our MLE estimate that the estimator $\hat{p}$ produced is 0.30. It is called the *maximum likelihood estimate* because it is the value that maximizes the likelihood of the observed sample.

It is the most likely value of the parameter that is supported by the data in the sample.

Question:

Why doesn't the likelihood care about constants in the pdf?

Example 2 - MLE (in book's notation)

Suppose $X_1, \ldots, X_n$ is a random sample (iid) from Exp(λ). Because of independence, the joint probability of the data = likelihood function is the product of pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \ldots \cdot (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \Sigma x_i}$$

How do we find the MLE?

What if our data is normally distributed?

Estimating Functions of Parameters

We've now learned how to obtain the MLE formulas for several estimators. Now we look at functions of them.

The Invariance Principle

Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ be the mle's of the parameters $\theta_1, \theta_2 \ldots \theta_m$.

Then the **mle** of any function $h(\theta_1, \theta_2, \ldots, \theta_m)$ of these parameters is the function $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ of the mle's.

Example

In the normal case, the mle's of μ and σ^2 are $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = \sum(X_i - \overline{X})^2/n$.

To obtain the mle of the function $h(\mu, \sigma^2) = \sqrt{\sigma^2} = \sigma$, substitute the mle's into the function:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[\frac{1}{n}\sum(X_i - \overline{X})^2\right]^{1/2}$$

The mle of σ is not the sample standard deviation *S*, though they are close unless *n* is quite small.

The Central Limit Theorem

Estimators and Their Distributions

Any estimator, as it is based on a sample, is a random variable that has its own probability distribution.

This probability distribution is often referred to as the **sampling distribution of the estimator.**

This sampling distribution of any particular estimator depends:

1) the population distribution (normal, uniform, etc.)
2) the sample size *n*
3) the method of sampling

The standard deviation of this distribution is called **the standard error of the estimator.**

Random Samples

The r.v.'s $X_1, X_2, \ldots, X_n$ are said to form a (simple) **random sample** of size n if

1. The X_i's are independent r.v.'s.

2. Every X_i has the same probability distribution.

We say that these X_i's are *independent and identically distributed* (***iid***).

Example

A certain brand of MP3 player comes in three models:

- 2 GB model, priced $80,
- 4 GB model priced at $100,
- 8 GB model priced $120.

Suppose the probability distribution of the cost X of a single randomly selected MP3 player purchase is given by

x	80	100	120
$p(x)$	.2	.3	.5

From here, $\mu = 106$, $\sigma^2 = 244$

Example, cont

cont' d

Suppose on a particular day only two MP3 players are sold. Let X_1 = the revenue from the first sale and X_2 the revenue from the second. X_1 and X_2 are independent, and have the previously shown probability distribution.

In other words, X_1 and X_2 constitute a random sample from that distribution.

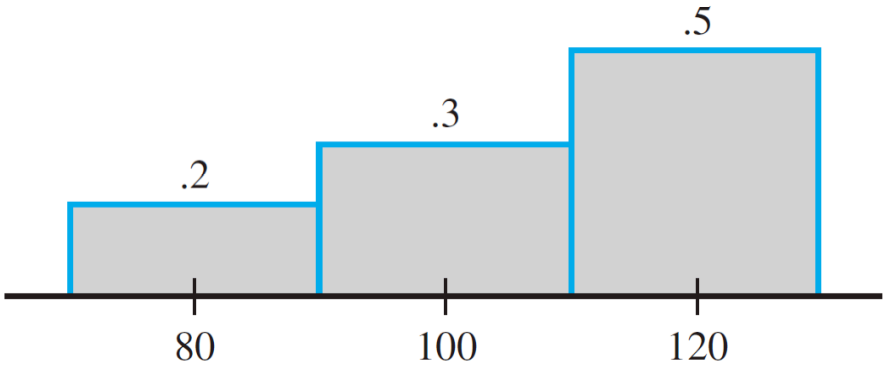
How do we find the mean and variance of this random sample?

Example cont

cont' d

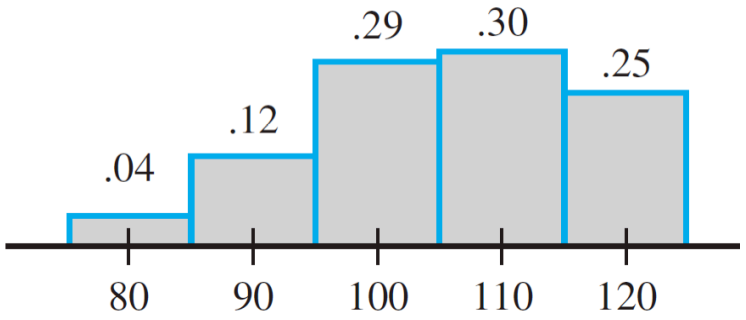
The complete sampling distributions of $\overline{X}$ is :

$\overline{x}$	80	90	100	110	120
$p_{\overline{X}}(\overline{x})$	.04	.12	.29	.30	.25



Original distribution:

μ = 106, σ^2 = 244



$\overline{X}$'s distribution

Example cont

cont' d

What are the mean and variance of this estimator?

What do you think the mean and variance would be if we had four samples instead of 2?

Example cont

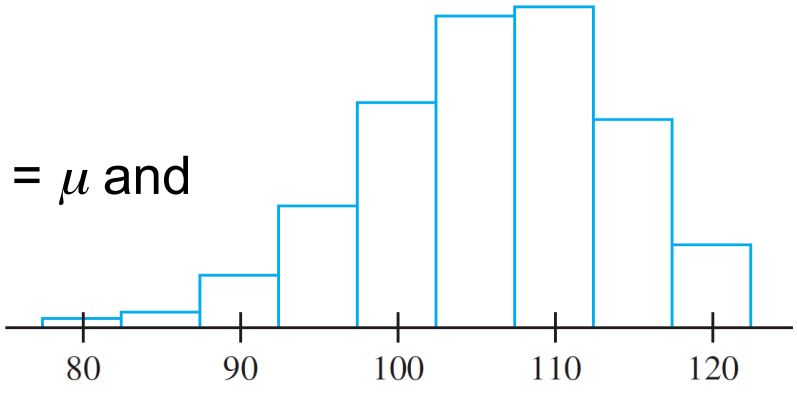
cont' d

If there had been four purchases on the day of interest, the sample average revenue $\overline{X}$ would be based on a random sample of four X_i's, each having the same distribution.

More calculation eventually yields the pmf of $\overline{X}$ for $n = 4$ as

$\overline{x}$	80	85	90	95	100	105	110	115	120
$p_{\overline{X}}(\overline{x})$	.0016	.0096	.0376	.0936	.1761	.2340	.2350	.1500	.0625

From this, $\mu_{\overline{X}} = 106 = \mu$ and $\sigma^2_{\overline{X}} = 61 = \sigma^2/4$.



Simulation Experiments

With a larger sample size, any unusual *x* values, when averaged in with the other sample values, still tend to yield $\bar{x}$ an value close to μ.

Combining these insights yields a result:

$\bar{X}$ based on a large *n* tends to be closer to μ than does $\bar{X}$ based on a small *n*.

The Distribution of the Sample Mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$

2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$

The standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{n}$ is also called the *standard error of the mean*

Great, but what is the *distribution* of the sample mean?

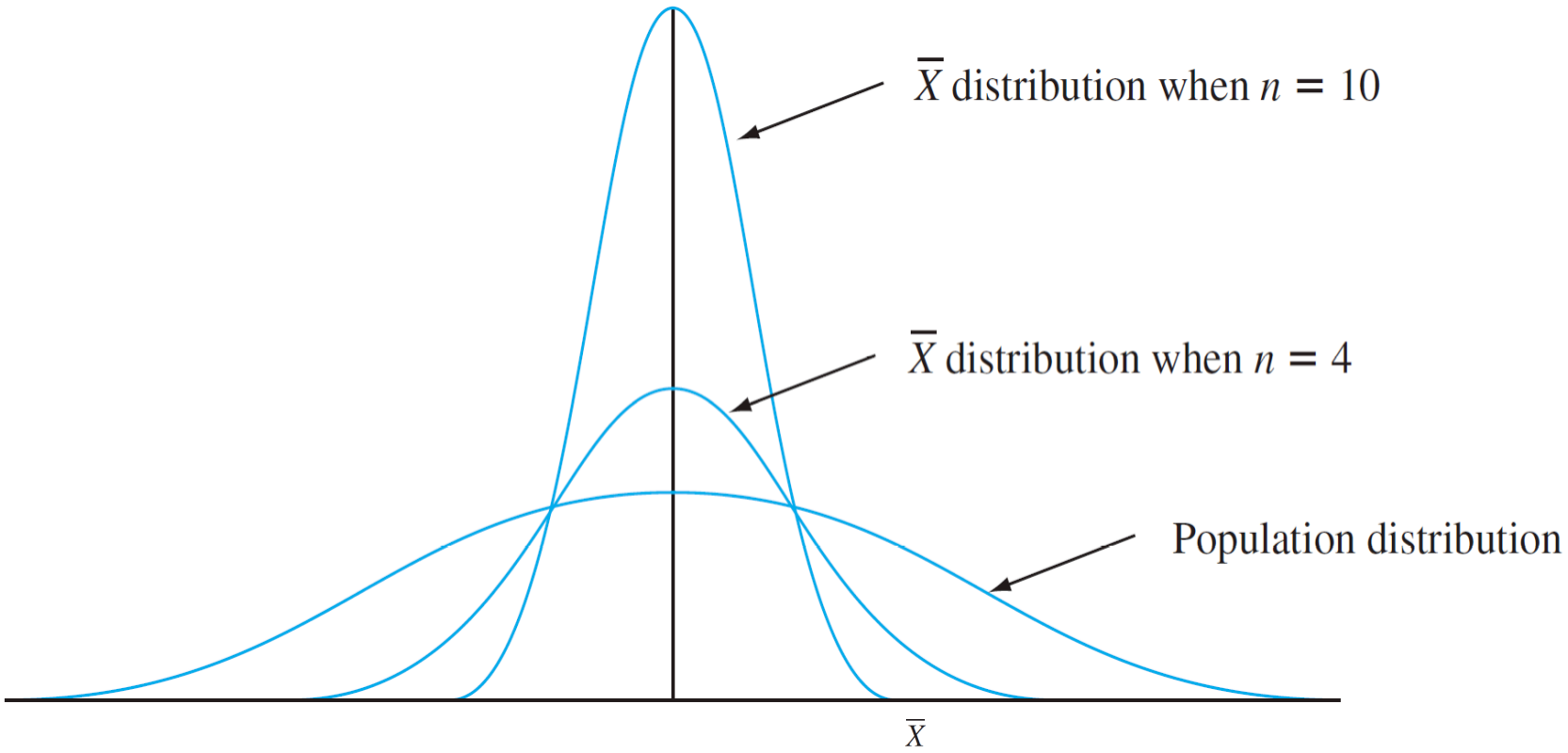
The Case of a Normal Population Distribution

Proposition:
Let $X_1, X_2, \ldots, X_n$ be a random sample from a *Normal* distribution with mean μ and standard deviation σ. Then for *any n*, $\overline{X}$ is normally distributed (with mean μ and standard deviation $\sigma/\sqrt{n}$)

We know everything there is to know about the $\overline{X}$ distribution when the population distribution is Normal.

In particular, probabilities such as $P(a \le \overline{X} \le b)$ can be obtained simply by standardizing.

The Case of a Normal Population Distribution



But what if the underlying distribution of X_i's is not Normal?

The Central Limit Theorem

The Central Limit Theorem (CLT)

When the X_i's are normally distributed, so is $\overline{X}$ for every sample size n.

Even when the population distribution is highly nonnormal, averaging produces a distribution more bell-shaped than the one being sampled.

A reasonable conjecture is that if n is large, a suitable normal curve will approximate the actual distribution of $\overline{X}$.

The formal statement of this result is one of the most important theorems in probability: CLT

The Central Limit Theorem

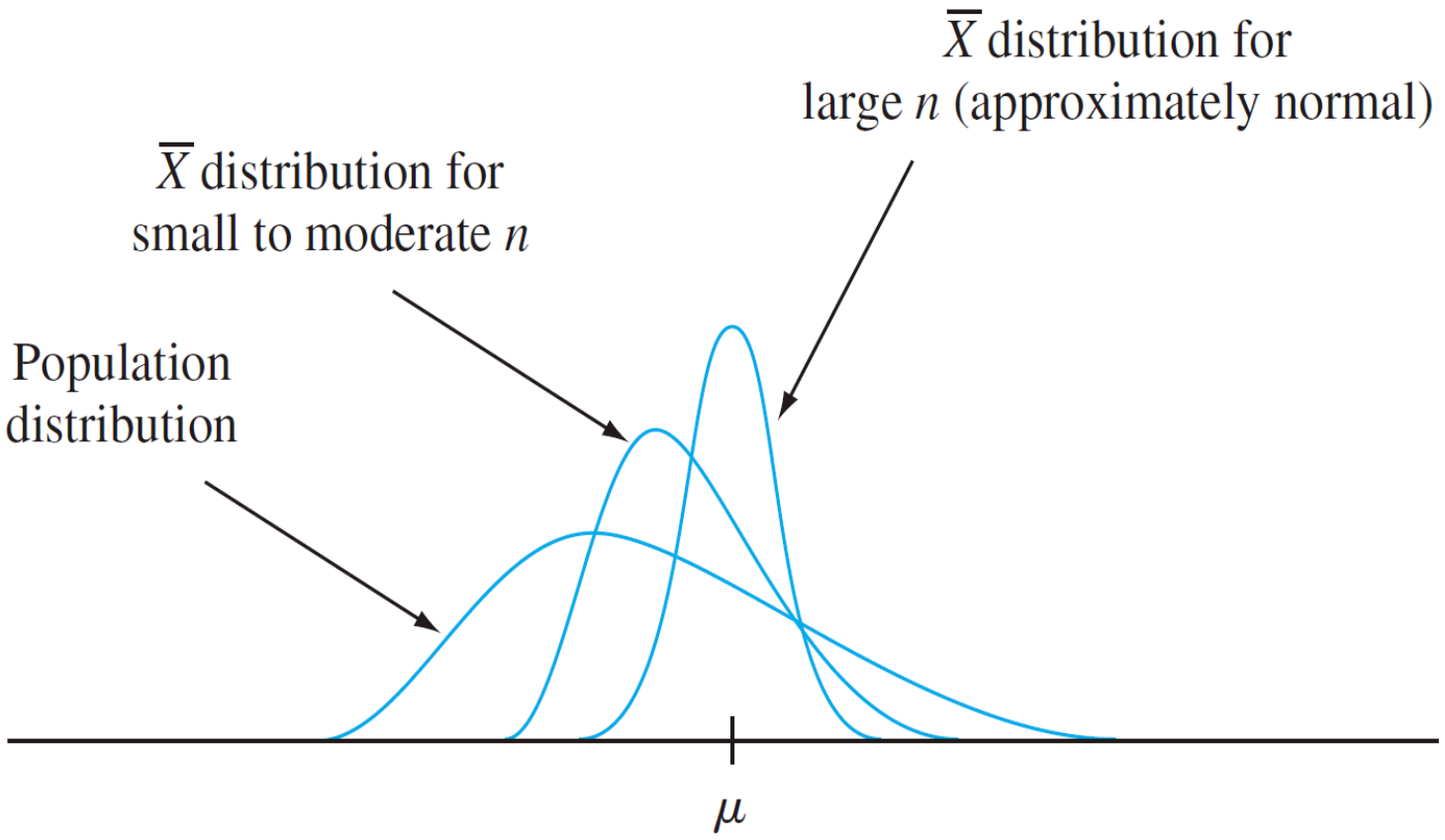
Theorem
The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2.

Then if n is sufficiently large, $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma^2_{\overline{X}} = \sigma^2/n$,

The larger the value of n, the better the approximation.

The Central Limit Theorem



The Central Limit Theorem illustrated

Example

The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

If 50 batches are independently prepared, what is the (approximate) probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

Side note: according to the rule of thumb to be stated shortly, $n = 50$ is "large enough" for the CLT to be applicable.

The Central Limit Theorem

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when *n* is sufficiently large. The problem is that the accuracy of the approximation for a particular *n* depends on the shape of the original underlying distribution being sampled.

The Central Limit Theorem

If the underlying distribution is close to a normal density curve, then the approximation will be good even for a small *n*, whereas if it is far from being normal, then a large *n* will be required.

Rule of Thumb
If $n > 30$, the Central Limit Theorem can be used.

R CODE