Point Estimation

Statistical inference = conclusions about parameters

Parameters == population characteristics

We will use the generic Greek letter θ for the parameter of interest

Process:

- obtain sample data from the population under study
- based on the sample data, estimate θ
- Base conclusions on sample estimates

The objective of point estimation is to estimate θ: compute a single number, based on sample data, that represents a sensible value for θ

Some General Concepts of Point Estimation

A point estimate of a parameter θ is a value (based on a sample) that can be regarded as a sensible guess for θ.

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

Such formulas are called point estimators of θ.

Different samples will generally yield different estimates, even though you use the same estimator.

Example

Sample of 20 observations:

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 is Normal with mean value μ.

Some point estimators of μ:

1) Sample mean
2) Sample median
3) (Max+Min)/2

Estimator quality

Which estimator is the best?

What does “best” mean?

For example, by “best” one might mean:

“which estimator, when used on many samples, will produce estimates closest to the true value, on average?”

and the other might mean:

“which estimator varies the least from sample to sample?”

And yet another

“which estimator is most robust to outliers?”

Estimator quality

An estimator $\hat{\theta}$ is a function of the sample, so it is a random variable.

For some realized samples values , $\hat{\theta}$ (the estimator) will yield a value larger than θ, whereas for other realized samples it will underestimate θ:

$$\hat{\theta} = \theta + \text{error of estimation}$$

It's the distribution of these errors (over all possible 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 the better one.

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

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

Note MSE = Variance for unbiased estimators.

Example: biased and unbiased

Suppose we have two measuring instruments; one instrument accurately calibrated, and the other systematically gives readings smaller than the true value.

Each instrument is used repeatedly on the same object

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

When $X \sim Bin(n, p)$ the sample proportion X / n can be used as an estimator of p.

Note

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

Thus, the sample proportion $\hat{p} = X / n$ is an **unbiased** estimator of p.

No matter what the true value of p is, the distribution of the estimator $\hat{p}$ will be centered at the true value.

Estimators with Minimum Variance

Estimators with Minimum Variance

Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of θ that are both unbiased.

The distribution of each estimator is centered at the true value of θ, but the spreads of the distributions about the true value may be different.

Among all estimators of θ that are unbiased, we will always want to choose the one that has smallest variance.

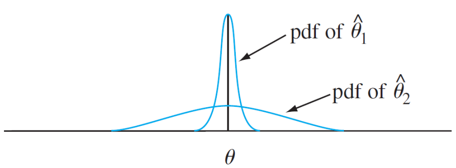
The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

Estimators with Minimum Variance

Here is an example of pdf's of two unbiased estimators

Then $\hat{\theta}_1$ is more likely than $\hat{\theta}_2$ to produce an estimate close to the true θ.

The MVUE is the most likely among all unbiased estimators to produce an estimate close to the true θ.



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.

Example

Sample of 20 observations:

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 is Normal with mean value μ.

Some point estimators of μ :

1) Sample mean
2) Sample median
3) (Max+Min)/2

Example

Assuming normality, the sample mean

$$\hat{\mu} = \overline{X}$$

is the best estimator of μ.

If the value of σ is known to be 1.5, the standard error of X is

$$\sigma_{\overline{X}} = \sigma/\sqrt{n} = 1.5/\sqrt{20} = .335$$

If, as is usually the case, the value of σ is unknown, the estimate $\hat{\sigma} = s$ = 1.462 is substituted $\hat{\sigma}_{\overline{X}}$ to to obtain the estimated standard error

$$\hat{\sigma}_{\overline{X}} = s_{\overline{X}} = s/\sqrt{n} = 1.462/\sqrt{20} = .327$$

General methods for constructing estimators

Setting:

- a sample from a known family of probability distributions
- we don't know the specific parameters of that distribution

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

<u>**Method 1:**</u> Methods of Moments (MoM):

1. set sample statistics (eg. mean, or variance) equal to the corresponding population values
2. solve these equations for unknown parameter values
3. the solution formula is the estimator

What are statistical moments?

For k = 1, 2, 3, . . . , the kth population moment, or kth moment of the distribution $f(x)$, is $E(X^k)$.

The kth sample moment is $(1/n)\sum_{i=1}^{n} X_i^k$.

Eg, the first population moment is $E(X) = \mu$, and the first sample moment is $\Sigma X_i/n = \overline{X}$.

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

Setting $E(X) = (1/n)\,\Sigma X_i$ and $E(X^2) = (1/n)\,\Sigma X_i^2$ 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.

Since there is only one parameter to be estimated, the estimator is obtained by equating $E(X)$ to $\overline{X}$

Since $E(X) = 1/\lambda$ for an exponential distribution, this gives
$1/\lambda = \overline{X}$ or $\lambda = 1/\overline{X}$

The moment estimator of λ is then $\hat{\lambda} = 1/\overline{X}$.

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:

$$L(p \mid sample\ data) = p^3(1-p)^7$$

Likelihood function = function of the parameter only. It is just like the probability of the the sample data, but without any constants

For what value of p is the obtained sample most likely to have occurred?

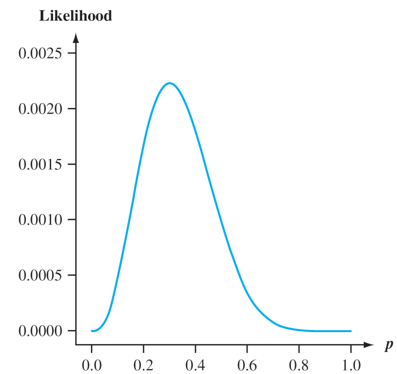
That is, what value of p maximizes the likelihood?

Example MLE

cont'd

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

$$L(p \mid sample\ data) = p^3(1-p)^7$$



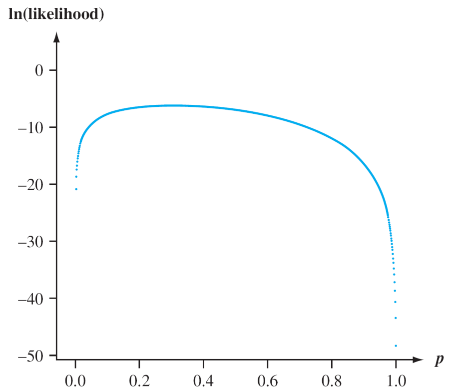
Example MLE

cont'd

The natural logarithm of the likelihood:

log (L (p | sample data)) = *l (p | 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:

the likelihood is typically a product, so its log will be a sum.

Here

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

Example MLE

cont'd

Optimizing the likelihood = optimizing the log-likelihood:

$$\frac{d}{dp}\{3\ln(p) + 7\ln(1-p)\}$$

$$= \frac{3}{p} + \frac{7}{1-p}(-1)$$

$$= \frac{3}{p} - \frac{7}{1-p}$$

Equating this derivative to 0 and solving for p gives 3(1 – p) = 7p, from which 3 = 10p and so p = 3/10 = .30

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?

Answer:

When you take the log, and differentiate wrt parameter, the constants disappear.

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:

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

The natural logarithm of the likelihood function is

$$\ln[\, L(\lambda\, ; x_1, \ldots, x_n)] = n \ln(\lambda) - \lambda \Sigma x_i$$

To find the MLE, we solve $d / d\lambda\, [n \ln(\lambda) - \lambda \Sigma x_i\,] = 0$

$$n / \lambda - \Sigma x_i = 0 \quad \text{i.e.,} \quad \lambda = n / \Sigma x_i$$

Thus the ML estimator is

$$\hat{\lambda} = 1/\overline{X};$$

Example 3 -- MLE

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution. The likelihood function is

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \cdots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)}$$

so

$$\ln[f(x_1, \ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \Sigma(x_i - \mu)^2$$

Example 3, cont.

cont'd

To find the maximizing values of μ and σ^2, we must take the partial derivatives of ln(f) with respect to μ and σ^2:

$$\ln[f(x_1, \ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \Sigma(x_i - \mu)^2$$

equate the partial derivatives to zero, and solve the resulting two equations.

The resulting MLEs are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\Sigma(X_i - \overline{X})^2}{n}$$

Estimating Functions of Parameters

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

Eg, MLE of $\sigma = \sqrt{\sigma^2}$ can be easily derived using the following proposition.

<u>The Invariance Principle</u>

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

$\hat{\sigma}^2 = \Sigma(X_i - \overline{X})^2/n.$

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

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

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

Estimators and Their Distributions

Any estimator, as it is based on a sample, is a random variable. As such, it has its own probability distribution -- how the estimates produced by this estimator vary across all samples (of the same size).

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 rv'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 rv's.

2. Every X_i has the same probability distribution.

We say that 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.

If 20% of all purchasers choose the 2 GB model, 30% choose the 4 GB model, and 50% choose the 8 GB model, then the probability distribution of the cost X of a single randomly selected MP3 player purchase is given by

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

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

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.

Suppose that X_1 and X_2 are independent, each with the probability distribution below.

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

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

Example cont

cont'd

Table below lists all (x_1, x_2) pairs, the probability of each [computed using the assumption of independence], and the resulting $\bar{x}$ and s^2 values.

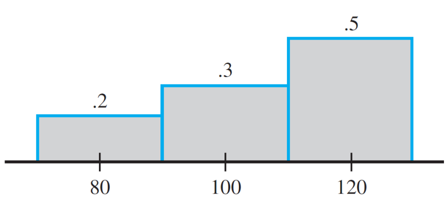
x_1	x_2	$p(x_1, x_2)$	$\bar{x}$	s^2
80	80	.04	80	0
80	100	.06	90	200
80	120	.10	100	800
100	80	.06	90	200
100	100	.09	100	0
100	120	.15	110	200
120	80	.10	100	800
120	100	.15	110	200
120	120	.25	120	0

Example cont

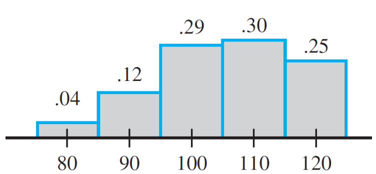
cont'd

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

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



Original distribution:
μ = 106, σ^2 = 244



$\overline{X}$'s distribution

Example cont

cont'd

$$\mu_{\overline{X}} = E(\overline{X}) = \sum \bar{x} p_{\overline{X}}(\bar{x})$$

= (80)(.04) + . . . + (120)(.25) = 106 = μ

It also appears that the $\overline{X}$ distribution has smaller spread (variability) than the original distribution:

$$\sigma^2_{\overline{X}} = V(\overline{X}) = \sum \bar{x}^2 \cdot p_{\overline{X}}(\bar{x}) - \mu^2_{\overline{X}}$$

= $(80^2)(.04)$ + + $(120^2)(.25) - (106)^2$ = 122

= 244/2

$$= \frac{\sigma^2}{2}$$

Example cont

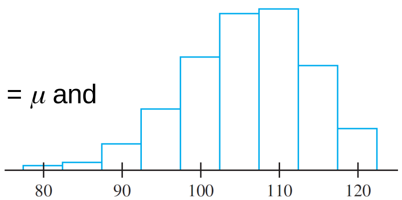
cont'd

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

More calculation eventually yields the pmf of for n = 4 as

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

From this, μ_x = 106 = μ and $\sigma^2_{\bar{X}}$ = 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 $\bar{x}$ 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. $$E(\bar{X}) = \mu_{\bar{X}} = \mu$$
2. $$V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n \text{ and } \sigma_{\bar{X}} = \sigma/\sqrt{n}$$

The standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ 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

```
nj=c(1,10,20,30,40,50,75,100,1000,10000)
par(mfrow=c(2,5))
for (j in 1:10) {
n=nj[j]
x=c(); for (i in 1:100)
     {x = c(x,mean(rnorm(n,1,1) ))}
hist(x,nclass=10,main=NULL)
title(xlab= paste("sample size =", n))
}
```

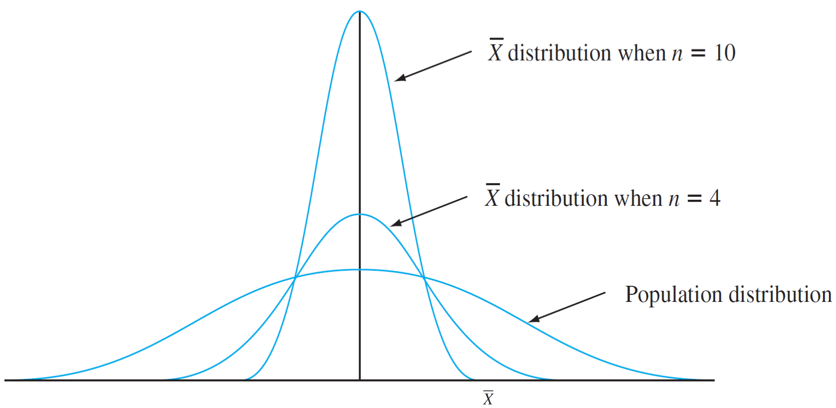
The Case of a Normal Population Distribution

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



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

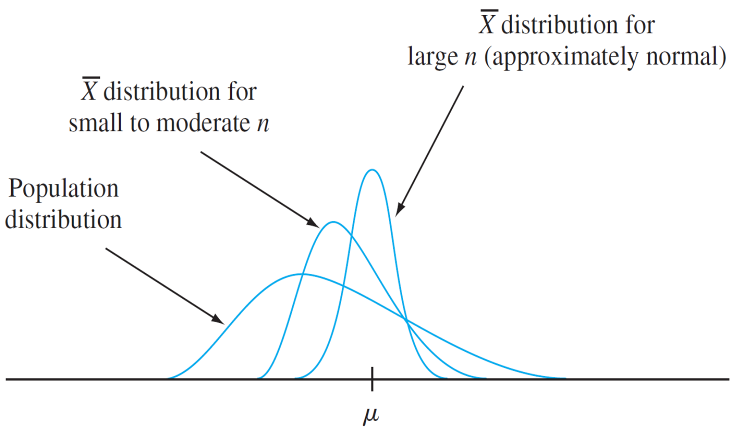
The Central Limit Theorem (CLT)

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

The Case of a Normal Population Distribution

```
nj=c(1,10,20,30,40,50,75,100,1000,10000)
par(mfrow=c(2,5))
for (j in 1:10) {
n=nj[j]
x=c(); for (i in 1:100)
     {x = c(x,mean(rgamma(n,0.1,1) ))}
hist(x,nclass=10,main=NULL)
title(xlab= paste("sample size =", n))
}
```


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?

According to the rule of thumb to be stated shortly, n = 50 is "large enough" for the CLT to be applicable.

Example

cont'd

$\overline{X}$ then has approximately a normal distribution with mean value $\mu_{\overline{X}}$ = 4.0 and

$$\sigma_{\overline{X}} = 1.5/\sqrt{50} = .2121,$$

so

$$P(3.5 \le \overline{X} \le 3.8) \approx P\left(\frac{3.5 - 4.0}{.2121} \le Z \le \frac{3.8 - 4.0}{.2121}\right)$$

$$= \Phi(-.94) - \Phi(-2.36)$$

$$= .1645$$

The Central Limit Theorem

The CLT is why many random variables 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.

There are population distributions for which even an n of 40 or 50 does not suffice, but this is rare.

For others, like in the case of a uniform population distribution, the CLT gives a good approximation for $n \geq 10$.

Normal approximation to Binomial

The CLT can be used to justify the normal approximation to the binomial distribution.

We know that a binomial variable X is the number of successes out of n independent success/failure trials with p = *probability of success* for any particular trial. Define a new rv X_1 as a Bernoulli random variable by:

$$X_1 = \begin{cases} 1 & \text{if the 1st trial results in a success} \\ 0 & \text{if the 1st trial results in a failure} \end{cases}$$

and define $X_2, X_3, \ldots, X_n$ analogously for the other $n - 1$ trials. Each X_i indicates whether or not there is a success on the corresponding trial.

Normal approximation to Binomial

Because the trials are independent and $P(S)$ is constant from trial to trial, the X_i's are iid (a random sample from a Bernoulli distribution).

The CLT then implies that if n is sufficiently large, then the average of all the X_i's have approximately normal distribution.

Normal approximation to Binomial

When the X_i's are summed, a 1 is added for every *success* that occurs and a 0 for every *failure*, so

$$X_1 + \ldots + X_n = X$$

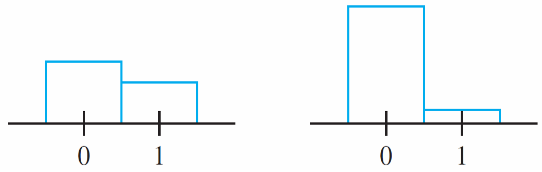
The sample mean of the X_i's is in fact the sample proportion of successes, X / n

That is, CLT assures us that X / n are approximately normal when n is large.

From here, X ($= n X / n$) is also approximately Normally distributed!

Normal approximation to Binomial

The necessary sample size for this approximation depends on the value of p: When p is close to .5, the distribution of each X_i is reasonably symmetric, whereas the distribution is quite skewed when p is near 0 or 1.



Two Bernoulli distributions: (a) p = .4 (reasonably symmetric); (b) p = .1 (very skewed)

Use the Normal approximation only if both $np \geq 10$ and $n(1-p) \geq 10$ – this ensures we'll overcome any skewness in the underlying Bernoulli distribution.

The Case of a Normal Population Distribution

```
nj=c(1,10,20,30,40,50,75,100,1000,10000)
par(mfrow=c(2,5))
for (j in 1:10) {
n=nj[j]
x=c(); for (i in 1:100)
     {x = c(x,mean(rbinom(n,10,0.1) ))}
hist(x,nclass=10,main=NULL)
title(xlab= paste("sample size =", n))
}
```


Other Applications of the Central Limit Theorem

We know that X has a lognormal distribution if $\ln(X)$ has a normal distribution.

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution for which only positive values are possible [$P(X_i > 0) = 1$].

Then if n is sufficiently large, the product

$$Y = X_1 X_2 \ldots\ldots X_n$$

has approximately a lognormal distribution.

Other Applications of the Central Limit Theorem

To verify this, note that

$$\ln(Y) = \ln(X_1) + \ln(X_2) + \cdots + \ln(X_n)$$

Since ln(Y) is a sum of independent and identically distributed rv's [the ln(X_i)s], it is approximately normal when n is large, so Y itself has approximately a lognormal distribution.