Statistics

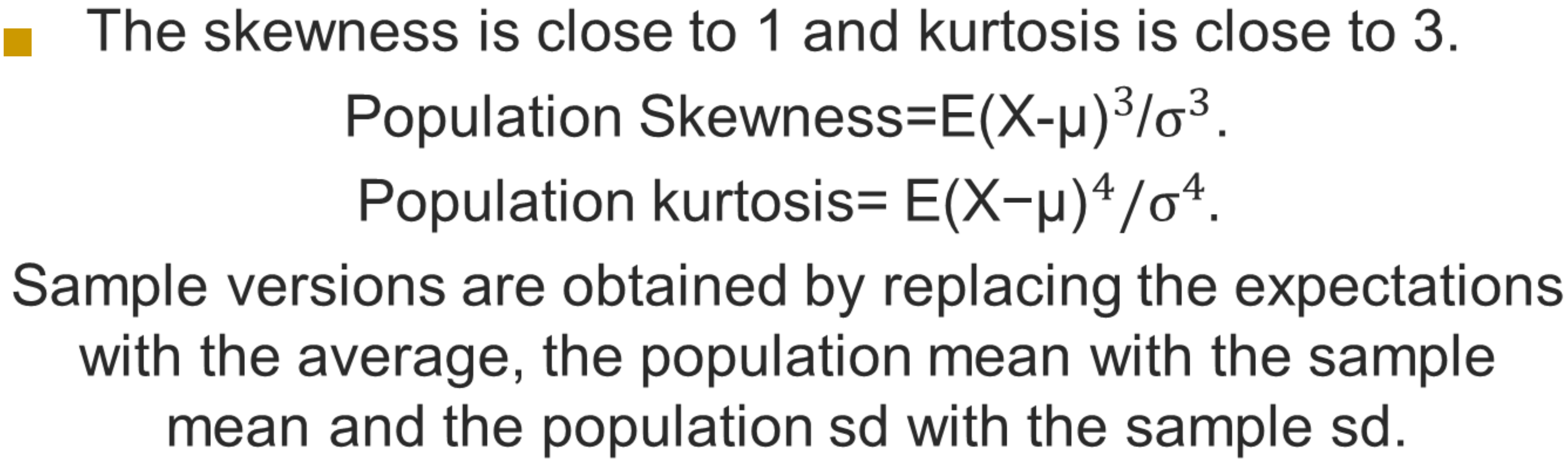
Chapter 5: Distributions with R

5.1: Descriptive Methods for Assessing Normality

If the data are normal

- A histogram or stem-and-leaf display will look like the normal curve
- The mean ± s, 2s and 3s will approximate the empirical rule percentages
- The ratio of the interquartile range to the standard deviation will be about 1.3
- The skewness is close to 0 and kurtosis is close to 3.
- A *normal probability plot* , a scatterplot with the ranked data on one axis and the expected *z*-scores from a standard normal distribution on the other axis, will produce close to a straight line

5.1: Descriptive Methods for Assessing Normality



- The skewness is close to 1 and kurtosis is close to 3.

Population Skewness=$E(X-\mu)^3/\sigma^3$.

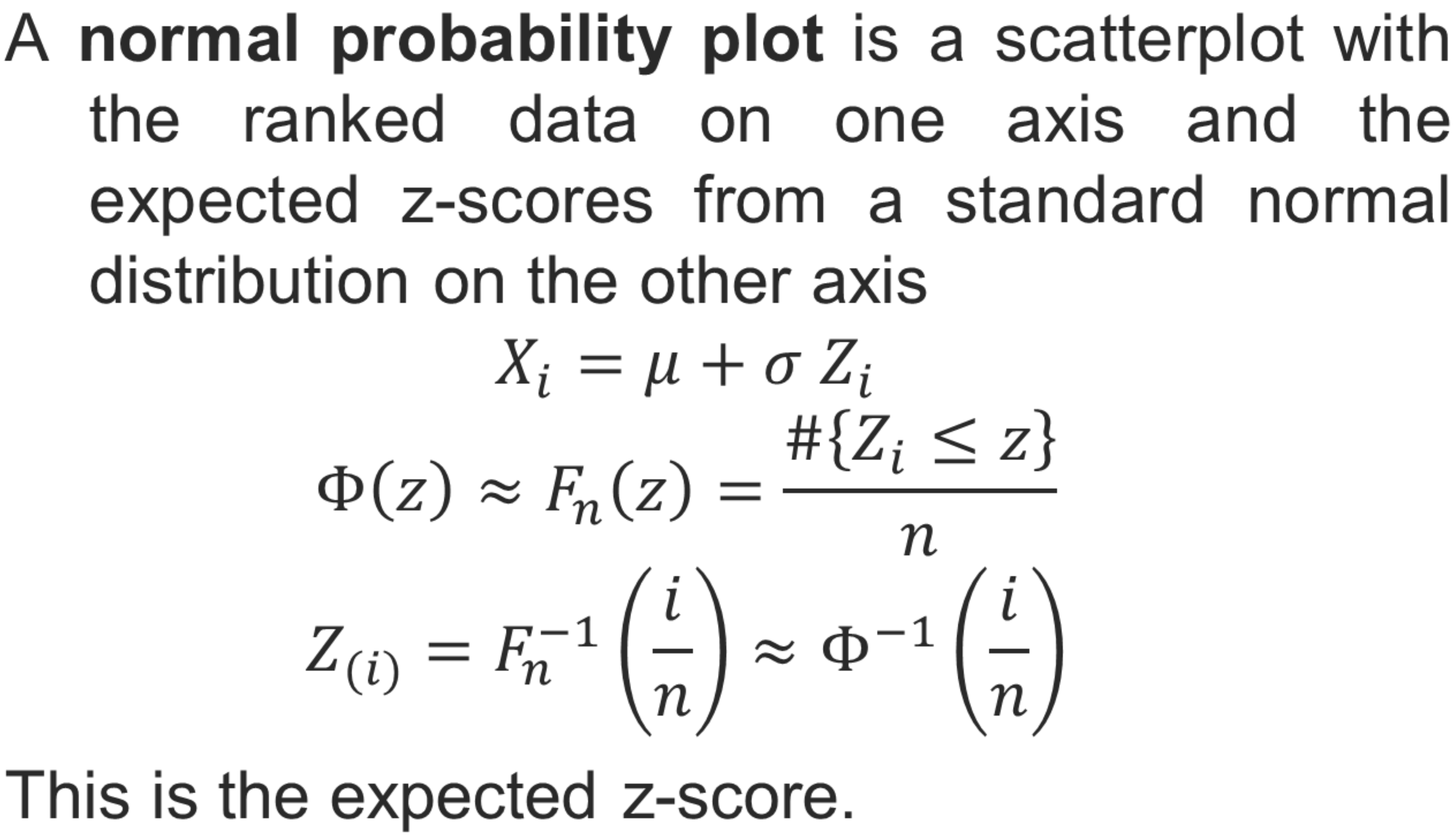
Population kurtosis= $E(X-\mu)^4/\sigma^4$.

Sample versions are obtained by replacing the expectations with the average, the population mean with the sample mean and the population sd with the sample sd.

5.1: Descriptive Methods for Assessing Normality

A **normal probability plot** is a scatterplot with the ranked data on one axis and the expected z-scores from a standard normal distribution on the other axis

5.1: Descriptive Methods for Assessing Normality



A **normal probability plot** is a scatterplot with the ranked data on one axis and the expected z-scores from a standard normal distribution on the other axis

$$X_i = \mu + \sigma Z_i$$

$$\Phi(z) \approx F_n(z) = \frac{\#\{Z_i \leq z\}}{n}$$

$$Z_{(i)} = F_n^{-1}\left(\frac{i}{n}\right) \approx \Phi^{-1}\left(\frac{i}{n}\right)$$

This is the expected z-score.

5.1: Descriptive Methods for Assessing Normality

- Beginning in 2017, public companies will be required to disclose the ratio of CEO pay to median worker pay. The Glassdoor Economic Research Blog has published the data for 2014. The data includes CEO identities, companies, CEO compensation, median worker compensation (compiled by Glassdoor), and the ratio of CEO to worker compensation.
- You can download the data from here: https://dasl.datadescription.com/download/data/3105

5.1: Descriptive Methods for Assessing Normality

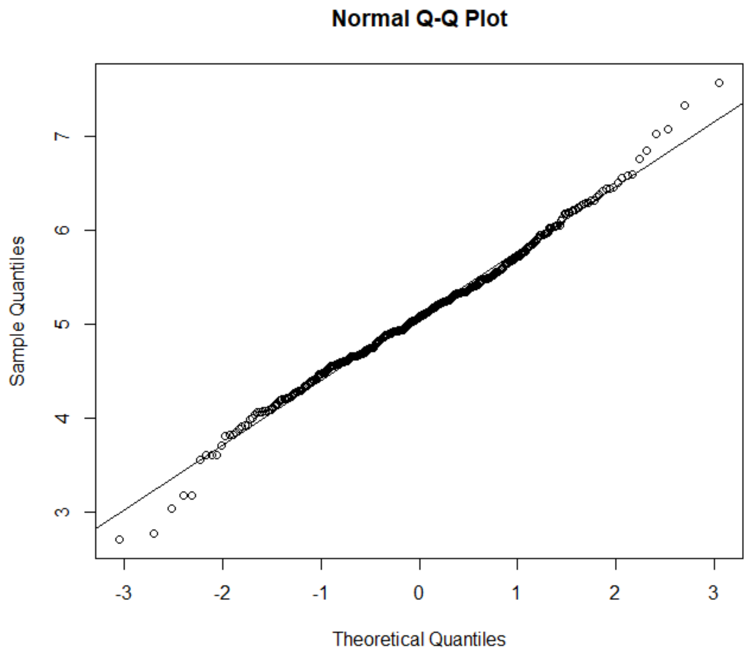
```
attach(ceo.compensation.2014)
x<-log(Ratio)
library(timeDate)
skewness(x)  # 0.08324829
kurtosis(x,method="moment") #3.873771
IQR(x)/sd(x)   #1.201651
```

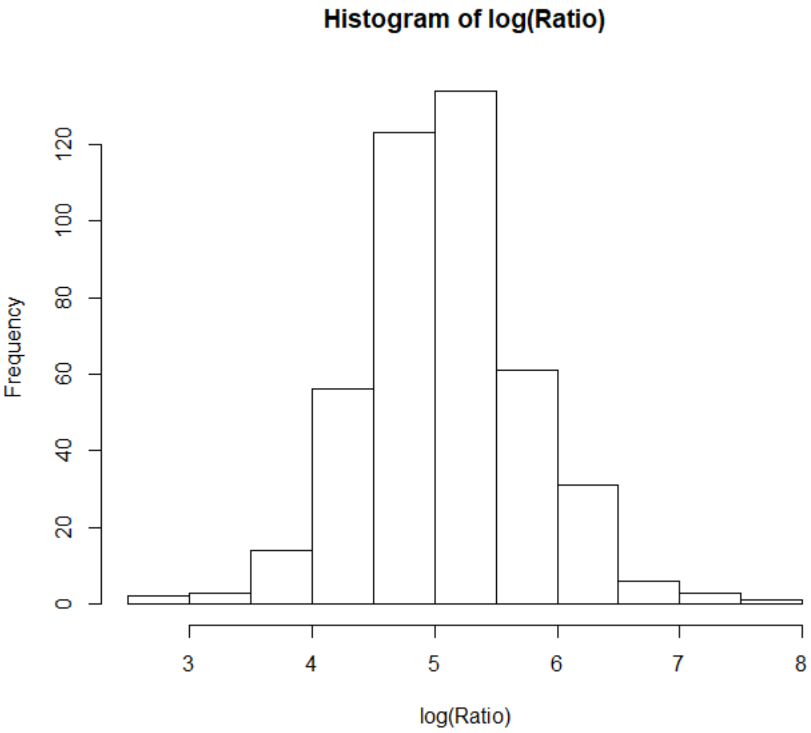
5.1: Descriptive Methods for Assessing Normality

```
hist(x)
length(which(x<mean(x)+sd(x)&x>mean(x)-sd(x)))/length(x)
length(which(x<mean(x)+2*sd(x)&x>mean(x)-2*sd(x)))/length(x)
length(which(x<mean(x)+3*sd(x)&x>mean(x)-3*sd(x)))/length(x)
pnorm(1)-pnorm(-1)
pnorm(2)-pnorm(-2)
pnorm(3)-pnorm(-3)
qqnorm(x)
abline(mean(x),sd(x))
```

0.7211982	0.6826895
0.9562212	0.9544997
0.9907834	0.9973002

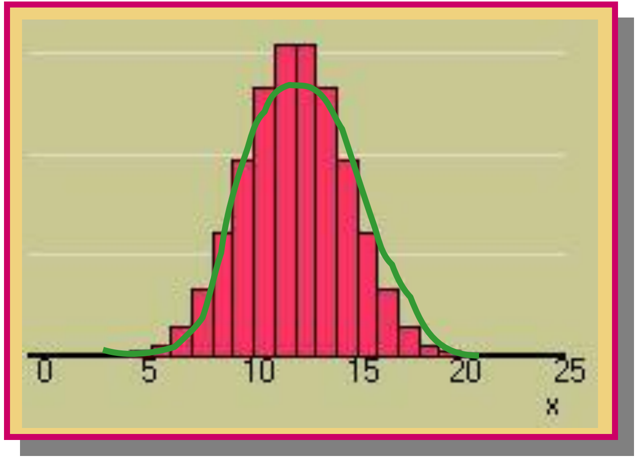
5.1: Descriptive Methods for Assessing Normality





5.2: Approximating a Binomial Distribution with the Normal Distribution

- We can calculate binomial probabilities using
  - The binomial formula
  - The cumulative binomial table
- When *n* is large, and *p* is not too close to zero or one, areas under the normal curve with mean *np* and variance *npq* can be used to approximate binomial probabilities.



5.2: Approximating a Binomial Distribution with the Normal Distribution

- Discrete calculations may become very cumbersome
- The normal distribution may be used to approximate discrete distributions
  - The larger *n* is, and the closer *p* is to .5, the better the approximation
- Since we need a range, not a value, the **correction for continuity** must be used
  - A number *r* becomes *r+.5*

5.2: Approximating a Binomial Distribution with the Normal Distribution

Calculate the mean plus/minus 3 standard deviations

$$\mu \pm 3\sigma = np \pm \sqrt{npq}$$

If this interval is in the range 0 to *n*,
the approximation will be reasonably close

Express the binomial probability as a range of values

$$P(x \leq a)$$

$$P(x \leq b) - P(x \leq a)$$

Find the z-values for each binomial value

$$z = \frac{(a + .5) - \mu}{\sigma}$$

Use the standard normal distribution to find
the probability for the range of values you calculated

5.2: Approximating a Binomial Distribution with the Normal Distribution

Flip a coin 100 times and compare the binomial and normal results

Binomial: $P(x=50)=\binom{100}{50}.5^{50}.5^{50}=.0796$

Normal:

$$\mu=100\cdot .5=50$$

$$\sigma=\sqrt{100\cdot .5\cdot .5}=5$$

$$P(49.5\le x\le 50.5)=P\left(\frac{49.5-50}{5}\le z\le\frac{50.5-50}{5}\right)=$$

$$P(-0.10\le z\le= 0.10)=.0796$$

5.2: Approximating a Binomial Distribution with the Normal Distribution

Flip a weighted coin *[P(H)=.4]* 10 times and compare the results

Binomial: $$P(x=5)=\binom{10}{5}.4^5.6^5=.1204$$

Normal: $$\mu=10\cdot .4=4$$

$$\sigma=\sqrt{10\cdot .4\cdot .6}=1.55$$

$$P(4.5\le x\le 5.5)=P\left(\frac{4.5-4}{1.55}\le z\le \frac{5.5-4}{1.55}\right)=$$

$$P(-0.32\le z\le= 0.32)=.1255$$

5.2: Approximating a Binomial Distribution with the Normal Distribution

Flip a weighted coin *[P(H)=.4]* 10 times and compare the results

Binomial: $P(x=5)=\binom{10}{5}.4^5.6^5=.1204$

Normal: $\mu=10\cdot.4=4$

$\sigma=\sqrt{10\cdot.4\cdot.6}=1.55$

$P(4.5\le x\le 5.5)=P\left(\frac{4.5-4}{1.55}\le z\le\frac{5.5-4}{1.55}\right)=$

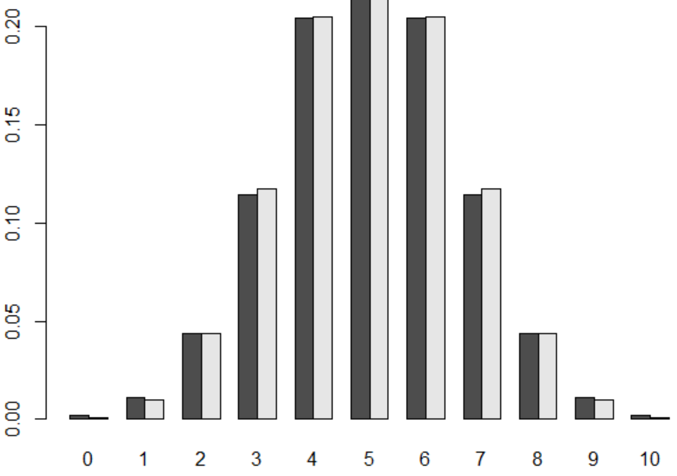
$P(-0.32\le z\le=0.32)=.1255$

The more *p* differs from .5, and the smaller n is, the less precise the approximation will be

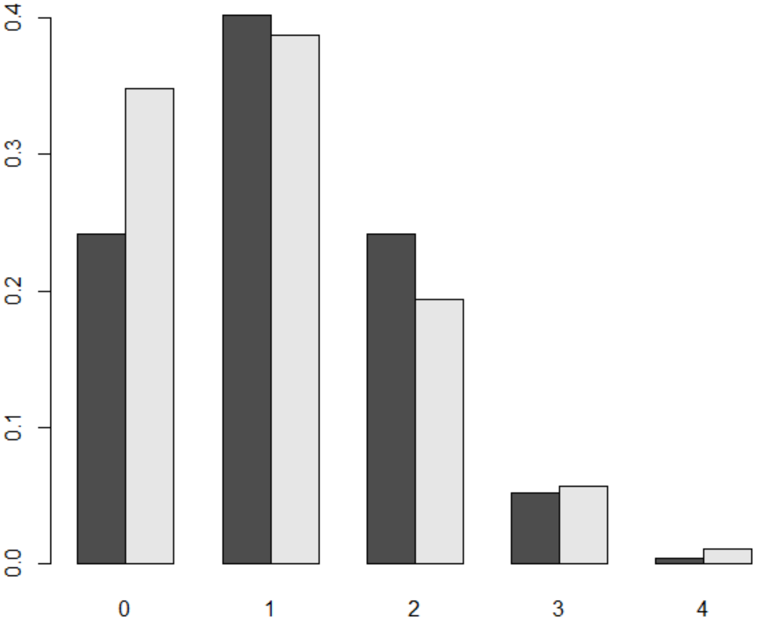
5.2: Approximating a Binomial Distribution with the Normal Distribution

```
p<-0.1
b<-rep(0,11)
n<-b
for (i in 1:11) b[i]<-dbinom(i-1,10,p)
m<-10*p
s<-sqrt(10*p*(1-p))
for (i in 1:11)n[i]<-pnorm((i-.5-m)/s)-pnorm((i-1.5-m)/s)
max(abs(b-n))
barplot(rbind(n[1:5],b[1:5]),beside=T,names.arg=c(0:4))
```

5.2: Approximating a Binomial Distribution with the Normal Distribution



p=0.5, n=10
MaxDiff=0.0027



p=0.1, n=10
MaxDiff=0.1065

5.3: Sampling Distributions

- In practice, sample statistics are used to estimate population parameters.
  - A **parameter** is a numerical descriptive measure of a population. Its value is almost always unknown.
  - A **sample statistic** is a numerical descriptive measure of a sample. It can be calculated from the observations.

5.3: Sampling Distributions

	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	$\boldsymbol{\sigma^2}$	s^2
Standard Deviation	$\boldsymbol{\sigma}$	s
Binomial proportion	p	$\hat{p}$

5.3: Sampling Distributions

- Numerical descriptive measures calculated from the sample are called **statistics**.
- Since we could draw many different samples from a population, the sample statistic used to estimate the population parameter is itself a **random variable**.

5.3: Sampling Distributions

- The **sampling distribution** of a sample statistic calculated from a sample of *n* measurements is the probability distribution of the statistic.
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs.

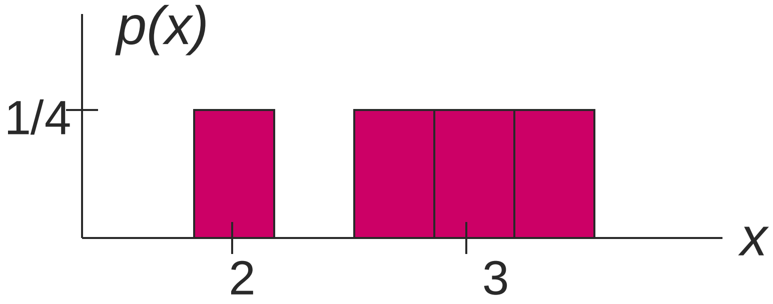
5.3: Sampling Distributions

Population: 3, 5, 2, 1

Draw samples of size *n* = 3 without replacement

Possible samples	$\bar{x}$
3, 5, 2	$10/3 = 3.33$
3, 5, 1	$9/3 = 3$
3, 2, 1	$6/3 = 2$
5, 2, 1	$8/3 = 2.67$

Each value of x-bar is equally likely, with probability 1/4



5.3: Sampling Distributions

Imagine a very small population consisting of the elements (N=population size=3) 1, 2 and 3. Below are the possible samples (without replacement) of sample size n that could be drawn, along with the means of the samples.

n = 1	$\bar{x}$
1	1
2	2
3	3

n = 2	$\bar{x}$
1, 2	1.5
1, 3	2
2, 3	2.5

n = 3 (= N)	$\bar{x}$
1, 2, 3	2

$$\boxed{\frac{\sum \bar{x}}{3} = 2}$$

$$\boxed{\frac{\sum \bar{x}}{3} = 2}$$

$$\boxed{\frac{\sum \bar{x}}{1} = 2}$$

5.4: Strong Law of Large Numbers

- Suppose $X_1, X_2, \ldots X_n$ are iid random variables with mean μ, then

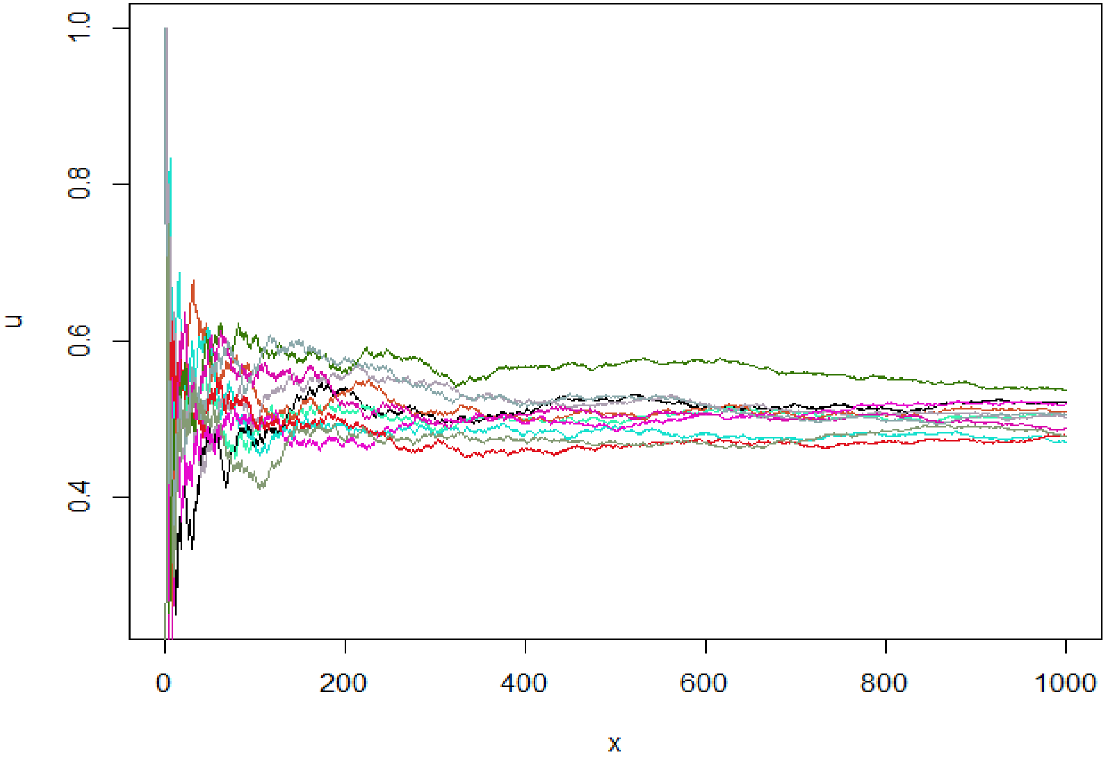
$$P(\bar{X} \rightarrow \mu \; as \; n \rightarrow \infty) = 1$$

- Sample mean is consistent estimator of true mean.

5.4: Strong Law of Large Numbers

```
runningmean = function (x,N){
 y = sample(x,N, replace=TRUE)
 c = cumsum(y)
 n = 1:N
 c/n
 }
u = runningmean(c(0,1), 1000)
x=1:1000; plot(u~x, type="l");
replicate(10, lines(runningmean(c(0,1), 1000)~x,
type="l", col=rgb(runif(3),runif(3),runif(3))))
```

5.4: Strong Law of Large Numbers



Running average of x Bernoulli random variables

5.5: The Central Limit Theorem

Properties of the Sampling Distribution of $\bar{x}$

The mean of the sampling distribution equals the mean of the population

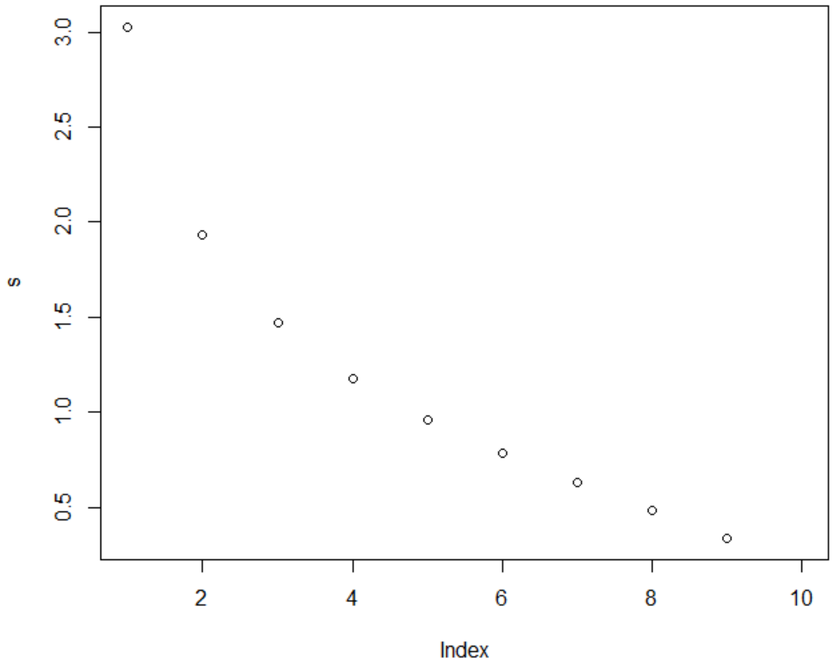
$$\mu_{\bar{x}} = (E\bar{x}) = \mu$$

The standard deviation of the sampling distribution [the **standard error (of the mean**)] equals the population standard deviation divided by the square root of *n*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

5.5: The Central Limit Theorem

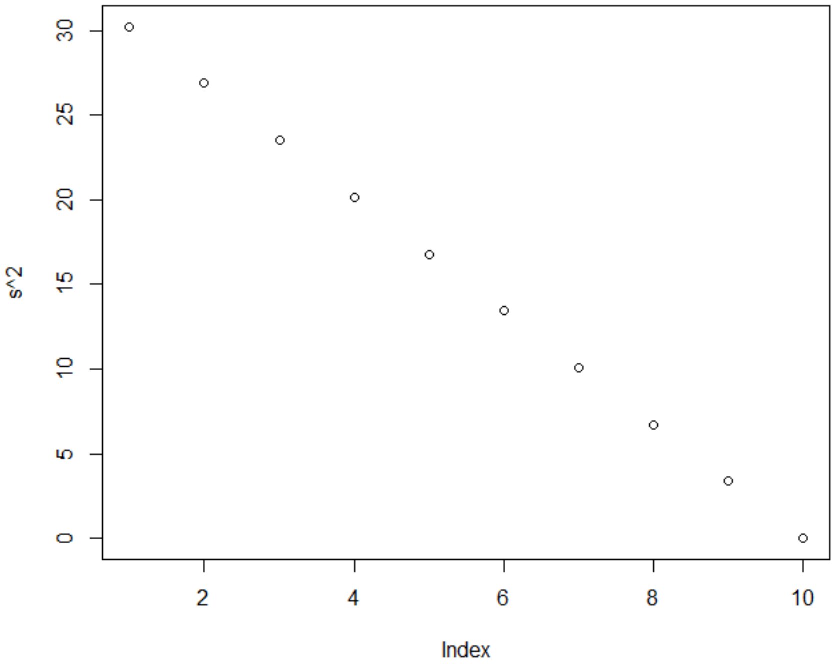
```
S<-rep(0,9)
for(i in 1:9){
m <- combn(10, i,mean)
s[i]<-sd(m)}
plot(s)
```



This is sampling without replacement. The result on the previous page is for sampling with replacement.

5.5: The Central Limit Theorem

```
S<-rep(0,9)
for(i in 1:9){
m <- combn(10, i,mean)
s[i]<-sqrt(10-i)*sd(m)/3}
plot(s^2)
```



Finite population correction (N-n)/(N-1) for variance.

5.5: The Central Limit Theorem

Here's our small population again, this time with the standard deviations of the sample means. Notice the mean of the sample means in each case equals the population mean and the standard error falls as *n* increases.

n = 1	$\bar{x}$
1	1
2	2
3	3

n = 2	$\bar{x}$
1, 2	1.5
1, 3	2
2, 3	2.5

n = 3 (= N)	$\bar{x}$
1, 2, 3	2

$$\frac{\sum \bar{x}}{3} = 2$$

$$\sigma_{\bar{x}} = .82$$

$$\frac{\sum \bar{x}}{3} = 2$$

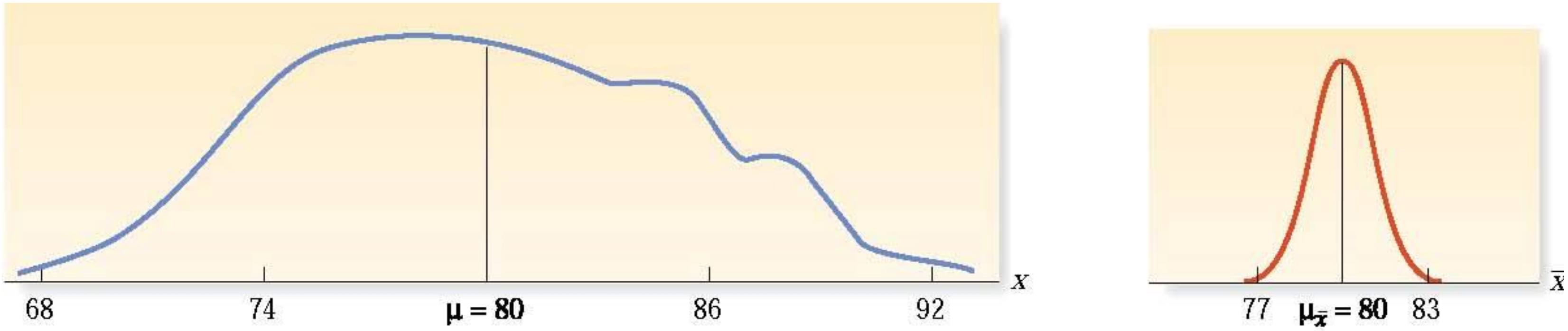
$$\sigma_{\bar{x}} = .41$$

$$\frac{\sum \bar{x}}{1} = 2$$

$$\sigma_{\bar{x}} = 0$$

5.5: The Central Limit Theorem

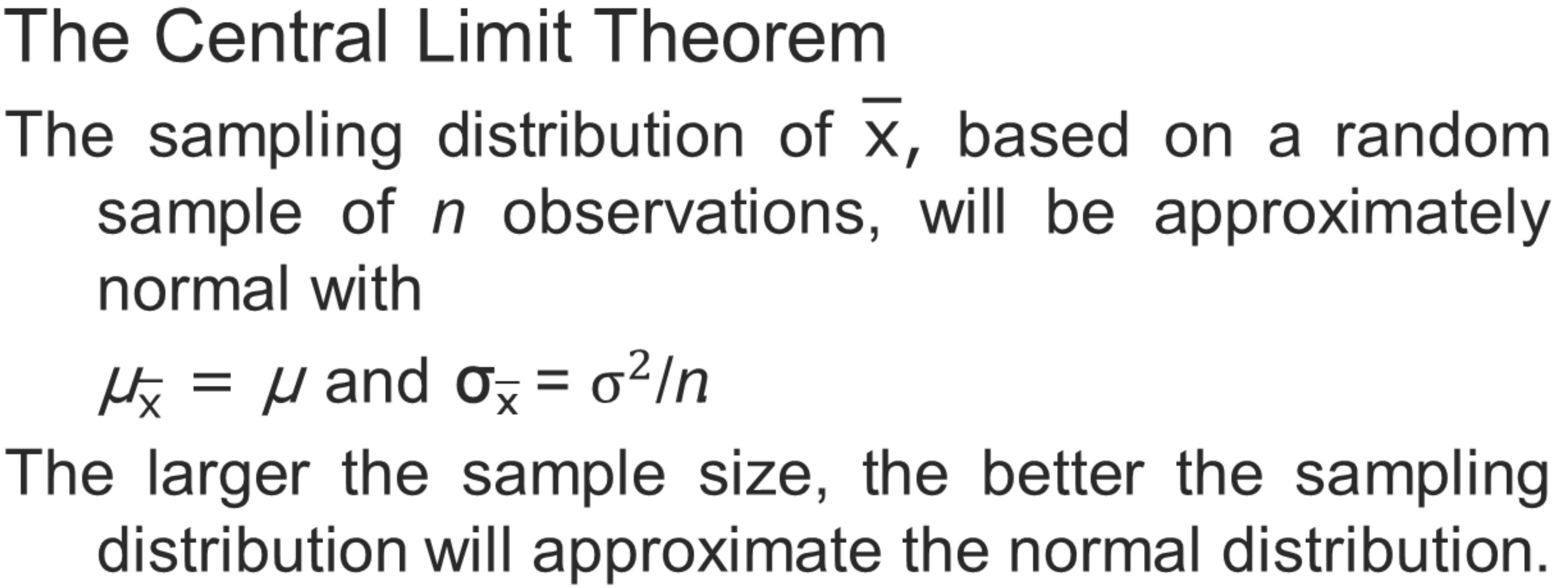
- If a random sample of *n* observations is drawn from a normally distributed population, the sampling distribution of $\bar{x}$ will be normally distributed



a. Population relative frequency distribution

b. Sampling distribution of $\bar{x}$

5.5: The Central Limit Theorem



- The Central Limit Theorem

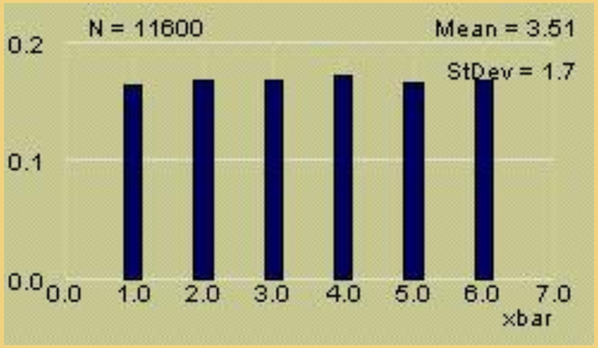
  The sampling distribution of $\overline{x}$, based on a random sample of *n* observations, will be approximately normal with

  $\mu_{\overline{x}} = \mu$ and $\boldsymbol{\sigma}_{\overline{x}} = \sigma^2/n$

  The larger the sample size, the better the sampling distribution will approximate the normal distribution.

5.5: The Central Limit Theorem

Roll a fair die *n* = 1 time. The distribution of *x* the number on the upper face is flat or uniform.

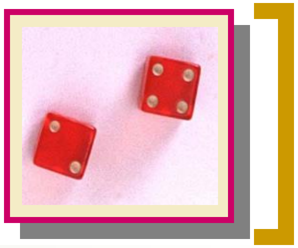


$$\mu = \sum xp(x)$$

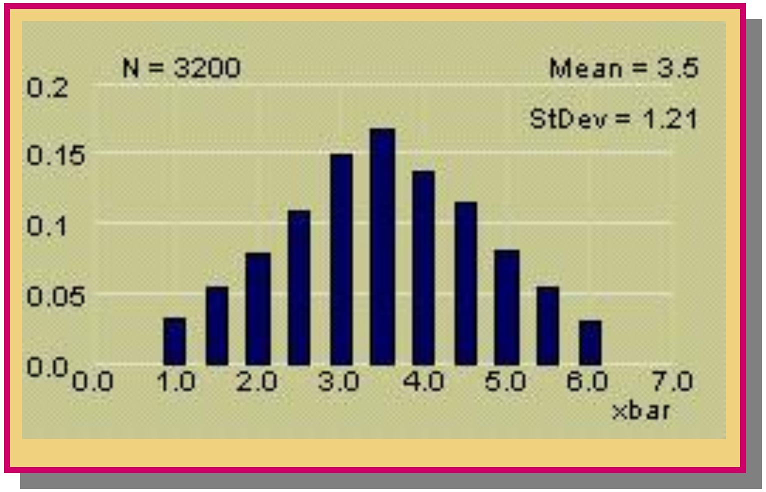
$$= 1(\frac{1}{6}) + 2(\frac{1}{6}) + \ldots + 6(\frac{1}{6}) = 3.5$$

$$\sigma = \sqrt{\sum (x-\mu)^2 p(x)} = 1.71$$

5.5: The Central Limit Theorem



Roll a fair die *n* = 2 time. The distribution of *x* the average number on the two upper faces is mound-shaped.



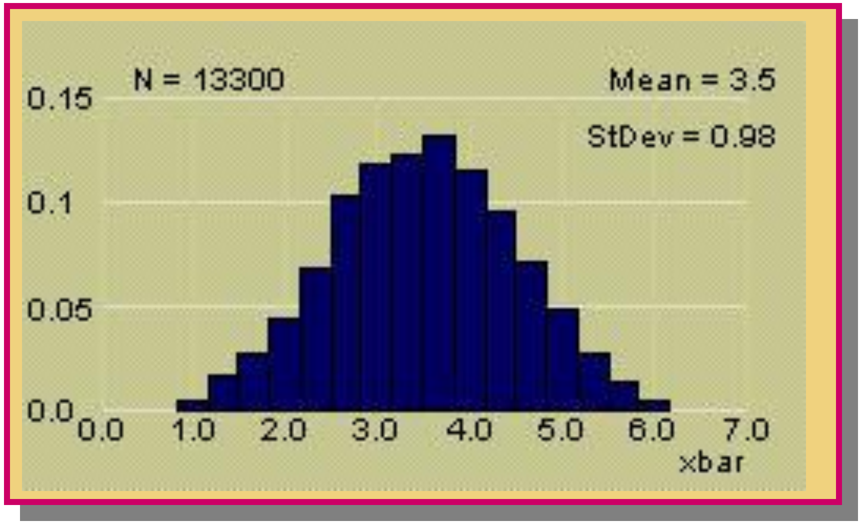
$$\text{Mean}: \mu = 3.5$$

$$\text{Std Dev}:$$

$$\sigma/\sqrt{2} = 1.71/\sqrt{2} = 1.21$$

5.5: The Central Limit Theorem

Roll a fair die *n* = 3 time. The distribution of *x* the average number on the two upper faces is approximately normal.

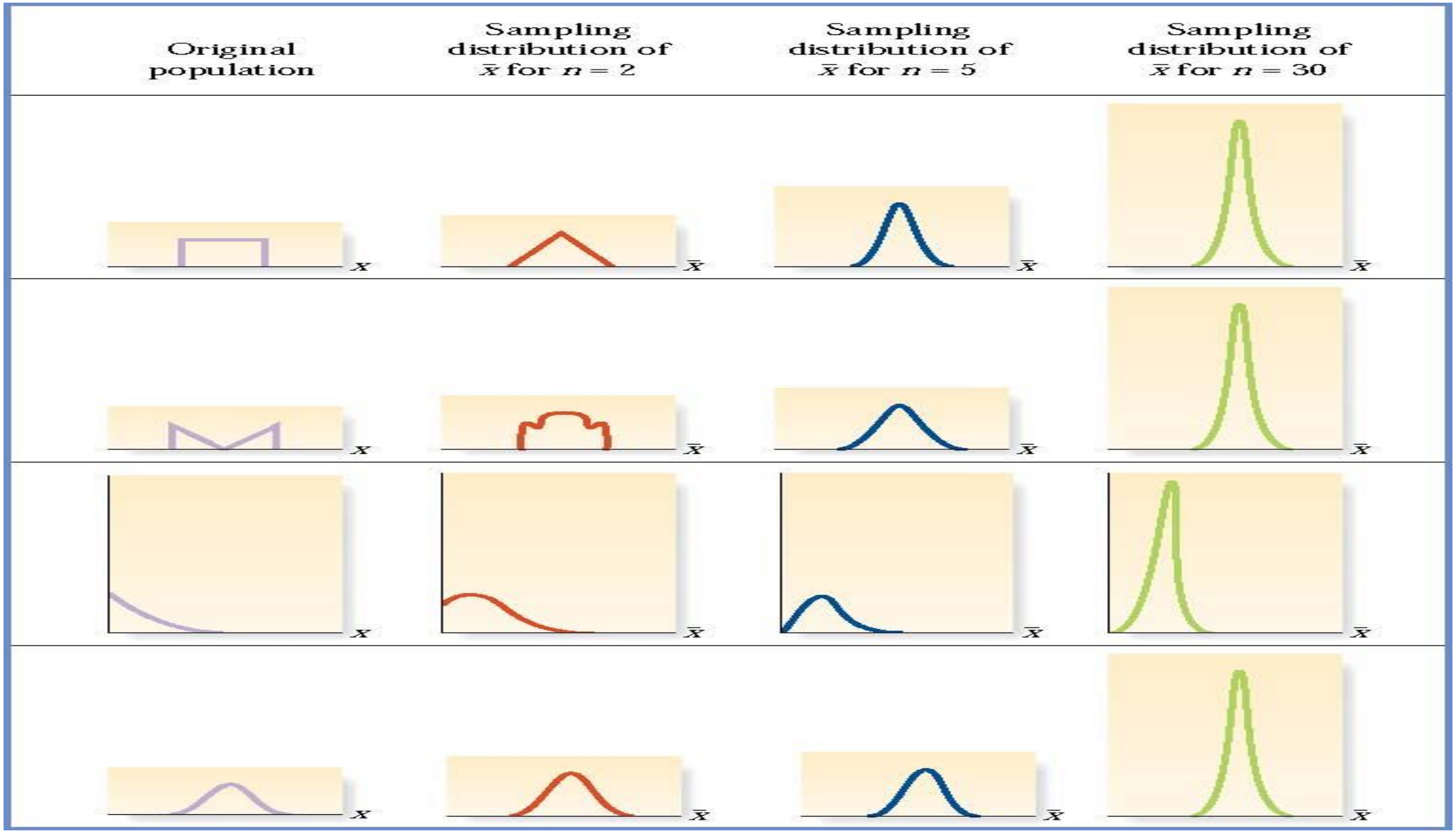


$$\text{Mean}: \mu = 3.5$$

$$\text{Std Dev}:$$

$$\sigma/\sqrt{3} = 1.71/\sqrt{3} = .987$$

5.5: The Central Limit Theorem


Original population
Sampling distribution of $\bar{x}$ for $n = 2$
Sampling distribution of $\bar{x}$ for $n = 5$
Sampling distribution of $\bar{x}$ for $n = 30$
x
$\bar{x}$

5.5: The Central Limit Theorem

- ✓The Central Limit Theorem also implies that the sum of *n* measurements is approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$.

- ✓Many statistics that are used for statistical inference are sums or averages of sample measurements.

- ✓This will allow us to describe their behavior and evaluate the reliability of our inferences using the normal distribution, even if we do not know the original distribution.

How large is large?

If the sample is normal, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size.

When the sample population is approximately symmetric, the distribution becomes approximately normal for relatively small values of *n*.

When the sample population is skewed, the sample size must be $\bar{x}$ at least 30 before the sampling distribution of becomes approximately normal.

5.5: The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected house will have at least 2300 ft^2 ?

5.5: The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected house will have at least 2300 ft^2 ?

$$P(x \geq 2300) =$$

$$P\left(z \geq \frac{2300-2200}{250}\right) =$$

$$P(z \geq 0.40) = .3446$$

5.5: The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected sample of 16 houses will average at least 2300 ft^2 ?

5.5: The Central Limit Theorem

Suppose existing houses for sale average 2200 square feet in size, with a standard deviation of 250 ft^2.

What is the probability that a randomly selected sample of 16 houses will average at least 2300 ft^2 ?

$$P(\bar{x} \geq 2300) =$$

$$P\left(z \geq \frac{2300 - 2200}{250 / \sqrt{16}} \right) =$$

$$P(z \geq 1.60) = .0548$$

5.5: The Central Limit Theorem

✓The Central Limit Theorem can be used to conclude that the binomial random variable *x* is approximately normal when *n* is large, with mean *np* and standard deviation npq .

✓The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by *n*.

✓From the Central Limit Theorem, the sampling distribution of $\hat{p}$ will also be approximately normal, with a *rescaled* mean and standard deviation.

5.5: The Central Limit Theorem

✓A random sample of size *n* is selected from a binomial population with parameter *p*.

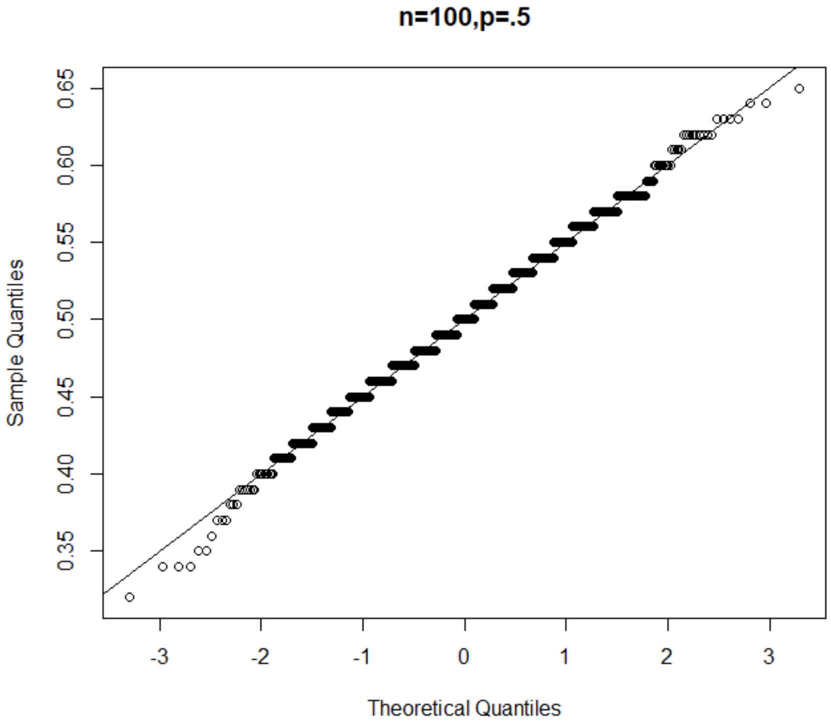
✓The sampling distribution of the sample proportion, $\hat{p} = \frac{x}{n}$

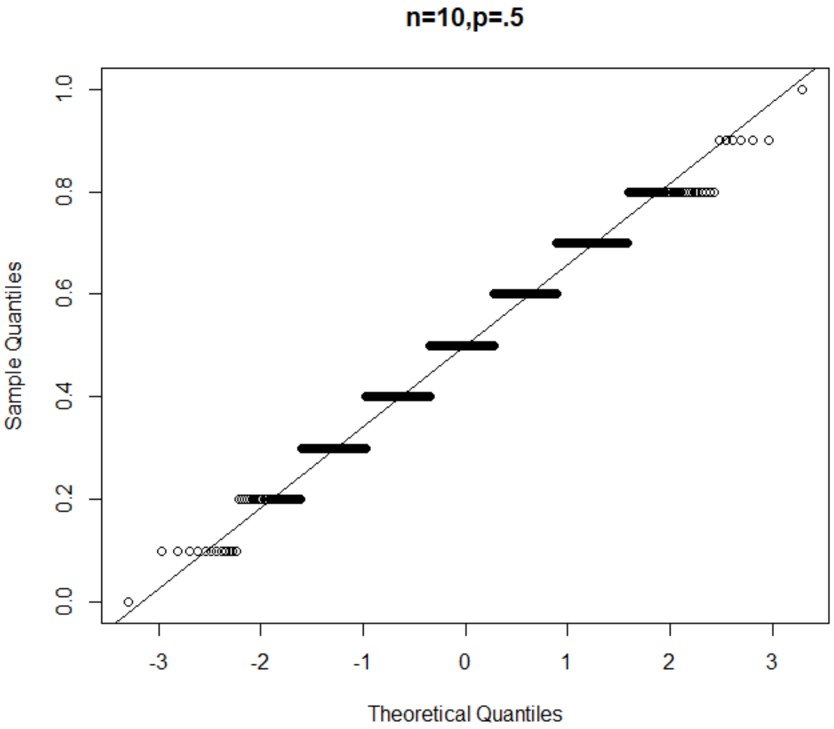
will have mean *p* and standard deviation $\sqrt{\frac{pq}{n}}$

✓If *n* is large, and *p* is not too close to zero or one, the sampling distribution of $\hat{p}$ will be approximately normal.

5.5: The Central Limit Theorem

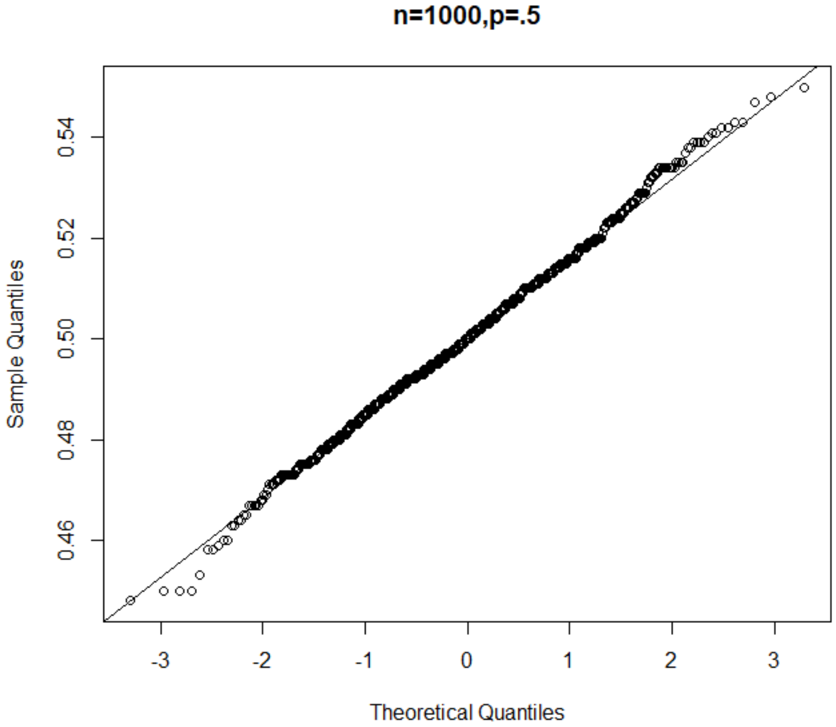
```
x<-rbinom(1000,n,p)/n; qqnorm(x); abline(p,sqrt(p*(1-p)/n))
```

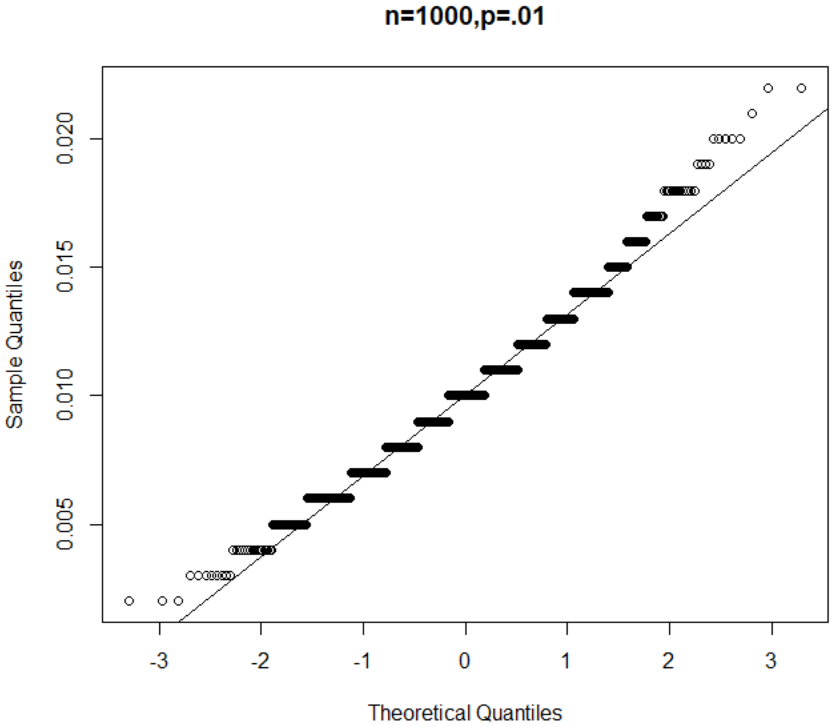




5.5: The Central Limit Theorem

```
x<-rbinom(1000,n,p)/n; qqnorm(x); abline(p,sqrt(p*(1-p)/n))
```


n=1000,p=.5
Sample Quantiles
Theoretical Quantiles


n=1000,p=.01
Sample Quantiles
Theoretical Quantiles