Topic 5
(Chapter 6)

Probability Distributions

1. How can we summarize possible outcomes and their probabilities

2. How can we find probabilities when each observation has two possible outcomes?

3. How can we find probabilities for bell-shaped distributions?

Review

- Inferential Statistics

  Use *sample* data to make decisions and predictions about a *population*

- Shape of distribution

  symmetric or skewed (right or left)

- Empirical rule and z-score formula

- Sample space of random experiments

  List all the outcomes, specify the events & find probabilities using basic rules

Randomness

Often, the randomness results from

- selecting a random sample for a population
- performing a randomized experiment

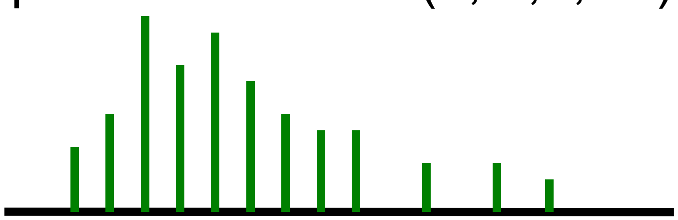
Random Variable

- A *random variable* is a numerical measurement of the outcome of a random phenomenon.
- The *numerical values* that a *variable* assumes are the *result of some random phenomenon*:

Random Variables

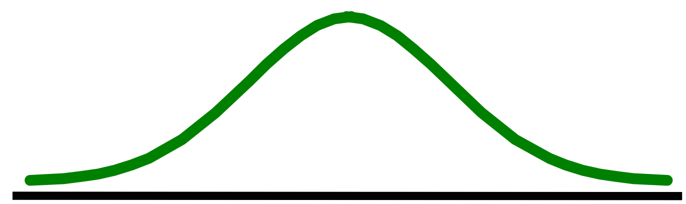
Discrete

The possible outcomes are a set of separate numbers: (0, 1,2, …).



Continuous

A continuous random variable has an infinite possible values in an interval.



- Capital letter, such as X, refers to the random variable itself.
- A small letter, such as x, refers to a particular value of the variable.

Probability Distribution

The *probability distribution* of a random variable specifies its possible values and their probabilities.

- In the form of a table
- In the form of a graph
- In the form of a mathematical formula

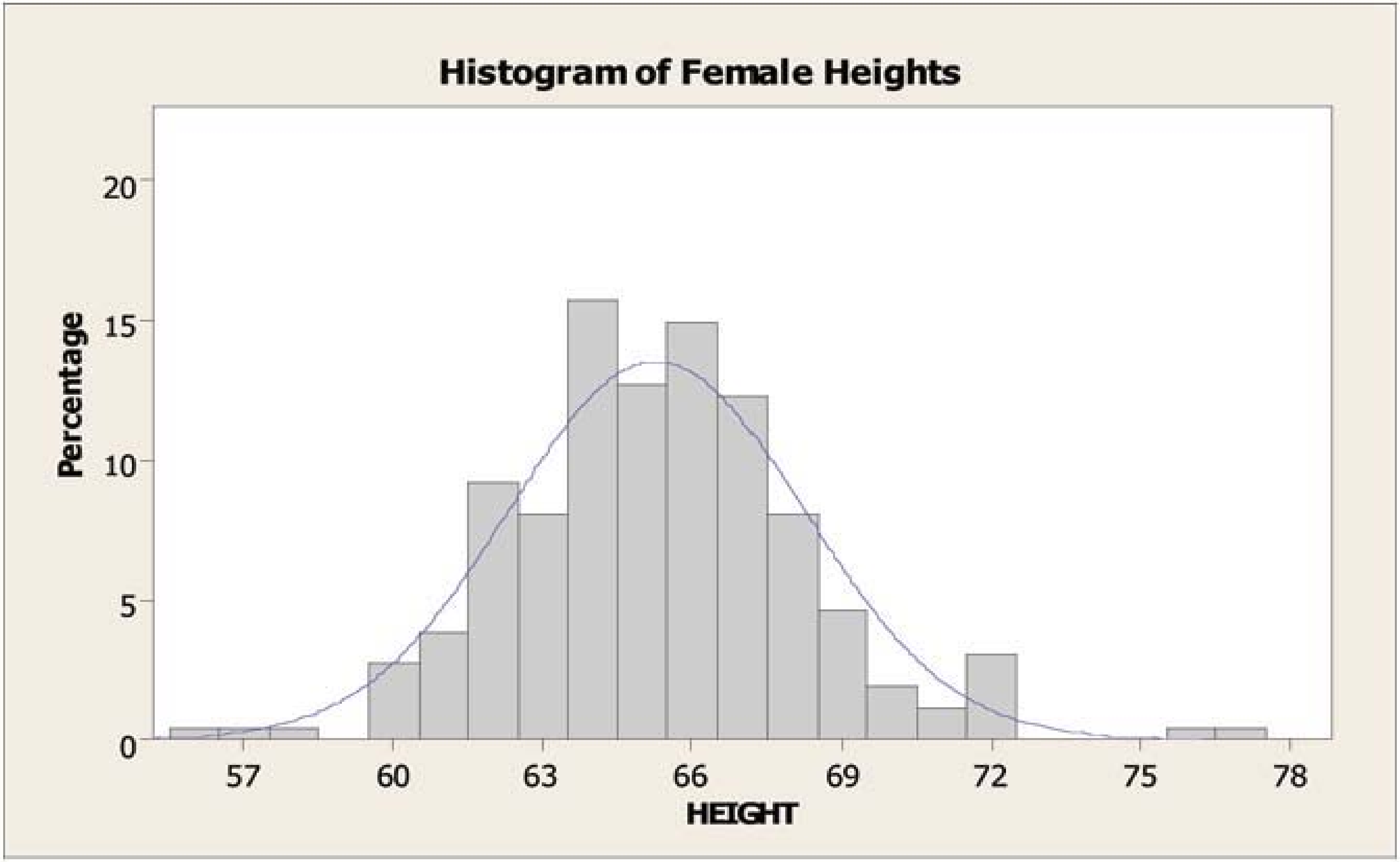
Note: It is the *randomness* of the variable that allows us to specify probabilities for the outcomes.

Probability Distribution of a Discrete Random Variable

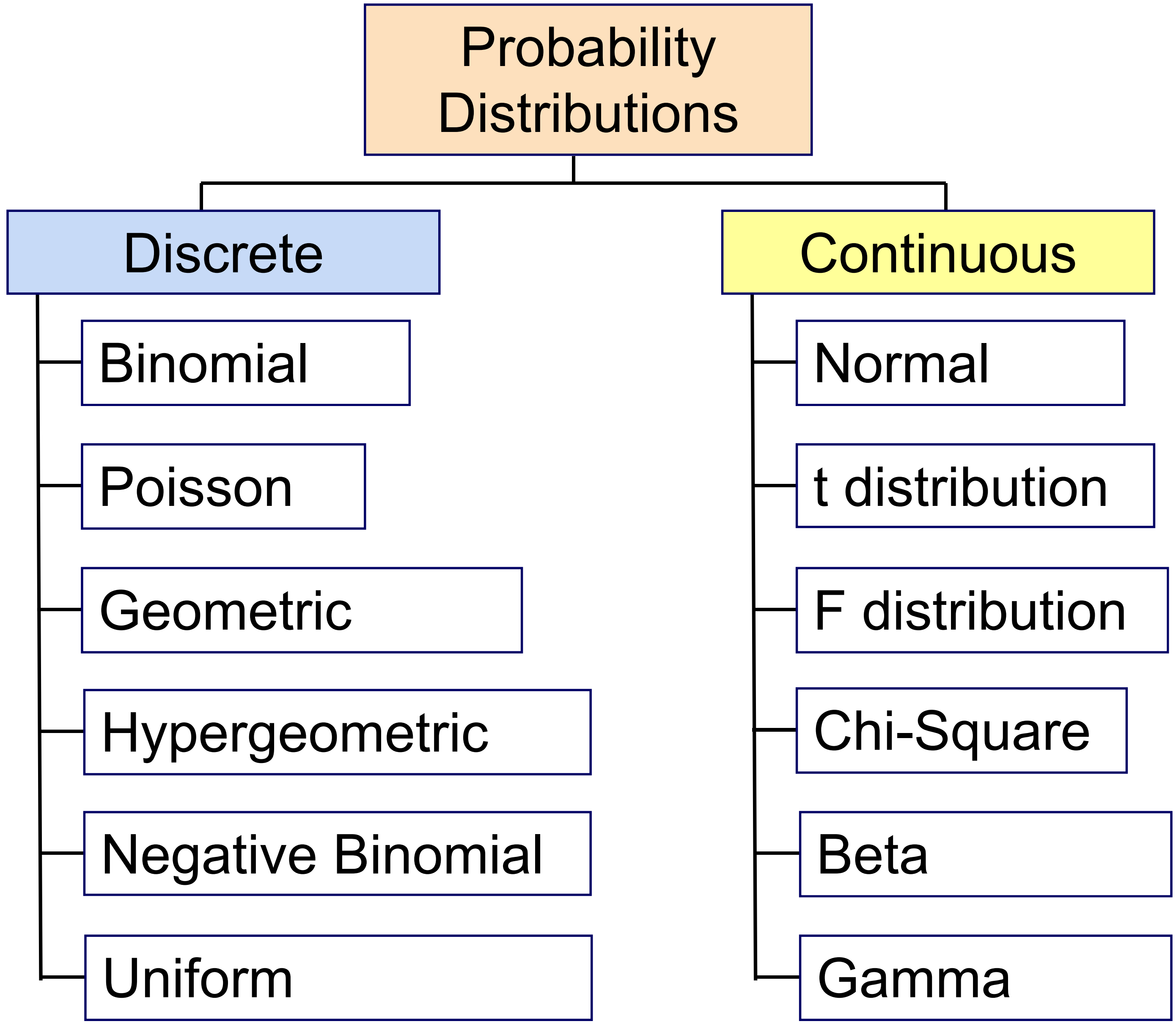
- Its probability distribution assigns a probability P*(x)* to each possible value *x*:
- For each *x*, the probability P*(x)* falls between 0 and 1
- The sum of the probabilities for all the possible *x* values equals 1

Probability Distribution of a Continuous Random Variable

- A continuous random variable has possible values that form an interval.
- Its probability distribution is specified by a curve.
- Each interval has probability between 0 and 1.
- The interval containing all possible values has probability equal to 1.



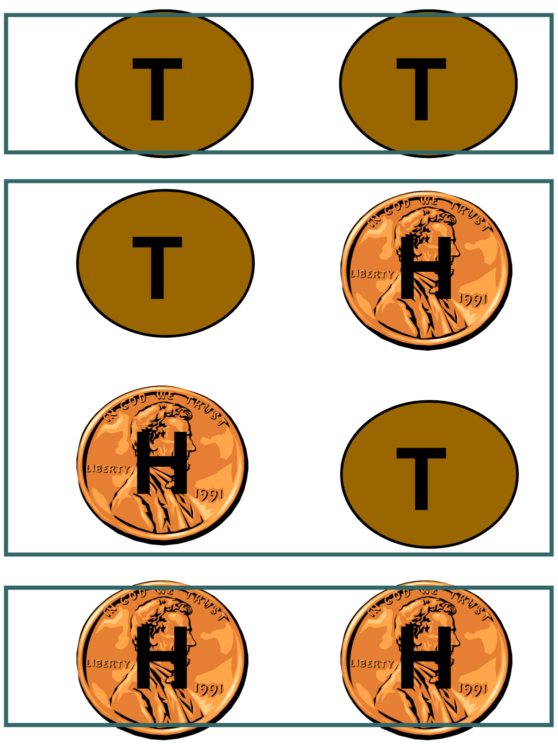
Many continuous variables are measured in a discrete manner because of rounding.


Probability Distributions
Discrete
Binomial
Poisson
Geometric
Hypergeometric
Negative Binomial
Uniform
Continuous
Normal
t distribution
F distribution
Chi-Square
Beta
Gamma

Discrete Probability Distribution

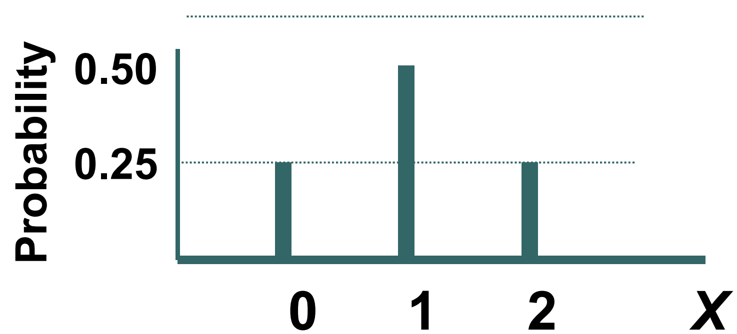
Experiment: Toss 2 Coins. Let X = # heads.

4 possible outcomes



Probability Distribution

x Value	Probability
0	1/4 = 0.25
1	2/4 = 0.50
2	1/4 = 0.25



Parameters of a Probability Distribution

The Mean

- The *mean* of a probability distribution is denoted by the parameter, μ.

The Standard Deviation

- The *standard deviation of a probability distribution*, denoted by the parameter, σ.

The Mean of a Discrete Probability Distribution

- The *mean of a probability distribution* for a discrete random variable is $$\boxed{\mu = \sum x \cdot P(x)}$$ where the sum is taken over all possible values of x.

- The mean of a probability distribution of a random variable X is also called the *expected value of X*.

- The expected value reflects not what we'll observe in a *single* observation, but rather that we expect for the *average in a long run* of observations.

Example:

Find the mean of this probability distribution.

x	$P(x)$	$x \cdot P(x)$	$(x-\mu)^2$	$(x-\mu)^2 \cdot P(x)$
0	0.23	0	1.9044	
1	0.38			
2	0.22			
3	0.13			
4	0.03			
5	0.01			
Total	1.00			

$$\mu = \sum x \cdot P(x)$$

The Standard Deviation of a Probability Distribution

- The *standard deviation of a probability distribution*, denoted by the parameter, σ, measures its spread.

$$\boxed{\sigma = \sqrt{\sum (x-\mu)^2 \cdot P(x)}}$$

$$\boxed{\sigma = \sqrt{\sum x^2 \cdot P(x) - \mu^2}}$$

- Larger values of σ correspond to greater spread.

Example:

- The mean of the probability distribution, $\boldsymbol{\mu}$

  $=$ 1.38

- The standard deviation of the probability distribution, $\boldsymbol{\sigma}$

  $=$ 1.1205

Question: Discrete Probability Distribution

Does the following table describe a discrete probability distribution?

x	–1	0	1	2
$P(x)$	0.2	0.3	0.1	0.4

The Binomial Distribution

The observation of a binomial distribution is binary: it has one of two possible outcomes.

Conditions for the Binomial Distribution

- Each of n trials has two possible outcomes: "success" and "failure".
- Each trial has the same probability of success, denoted by p.
- The n trials are independent.
- The binomial random variable X is the number of successes in the n trials. P(n~X)

Example :
Finding Binomial Probabilities for An ESP Experiment

- John Doe claims to possess ESP.
- An experiment is conducted:
  - A person in one room picks one of the integers 1, 2, 3, 4, 5 at random.
  - In another room, John Doe identifies the number he believes was picked.
  - The experiment is done with three trials.
  - Doe got the correct answer twice.
- If John Doe does not actually have ESP and is actually guessing the number, what is the probability that he'd make a correct guess on two of the three trials?

Example :
Finding Binomial Probabilities for An ESP Experiment

The probability of a correct guess is 0.2 on each of the three trials, if John Doe does not have ESP.

Outcome	Probability	Outcome	Probability
SSS	$0.2 \times 0.2 \times 0.2 = (0.2)^3$	SFF	$0.2 \times 0.8 \times 0.8 = (0.2)^1(0.8)^2$
SSF	$0.2 \times 0.2 \times 0.8 = (0.2)^2(0.8)^1$	FSF	$0.8 \times 0.2 \times 0.8 = (0.2)^1(0.8)^2$
SFS	$0.2 \times 0.8 \times 0.2 = (0.2)^2(0.8)^1$	FFS	$0.8 \times 0.8 \times 0.2 = (0.2)^1(0.8)^2$
FSS	$0.8 \times 0.2 \times 0.2 = (0.2)^2(0.8)^1$	FFF	$0.8 \times 0.8 \times 0.8 = (0.8)^3$

- The three ways John Doe could make two correct guesses in three trials are: SSF, SFS, and FSS.
- Each of these has probability: $(0.2)^2(0.8)=0.032$.
- The total probability of two correct guesses is $3(0.2)^2(0.8)=0.096$.

Probabilities for a Binomial Distribution

- Denote the probability of success on a trial by *p*.
- For *n* independent trials, the probability of *x* successes equals:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

$$x = 0, 1, 2, \ldots, n$$

Example :
Using the Binomial Formula in ESP Experiment

- The probability of exactly 2 correct guesses is the binomial probability with

  n = 3 trials,

  p = 0.2 probability of a correct guess.

  x = 2 correct guesses and

Example:
Testing for Gender Bias in Promotions

- A group of female employees has claimed that female employees are less likely than male employees of similar qualifications to be promoted.
- Of 1000 employees, 50% are female.
- None of the 10 employees chosen for management training were female.
- If the employees are selected randomly in terms of gender, <u>about half</u> of the employees picked should be females and about half should be males.
- How can we investigate statistically the women's assertion of gender bias?

Example:
Testing for Gender Bias in Promotions

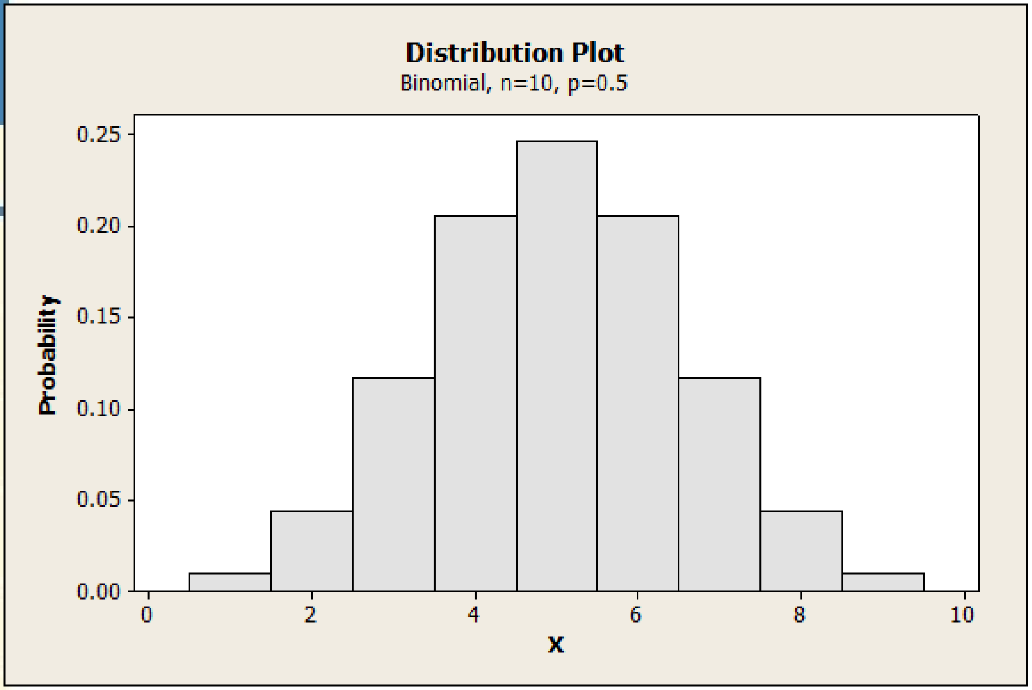
- Due to ordinary sampling variation, it need not happen that *exactly* 50 % of those selected are females.
- If employees were actually selected at random for the training, the probability that no females are chosen is:

$$P(0) = \frac{10!}{0!10!}(0.50)^0(0.50)^{10} = 0.001$$

It is very unlikely (one chance in a thousand) that *none* of the 10 selected for management training would be female if the employees were chosen randomly.

Binomial Probability Distribution for $n = 10$ and $p = 0.5$

x	P(x)	x	P(x)
0	0.001	6	0.205
1	0.010	7	0.117
2	0.044	8	0.044
3	0.117	9	0.010
4	0.205	10	0.001
5	0.246		



Do the Binomial Conditions Apply?

Before you use the binomial distribution, check that its three conditions apply:

- Binary data (success or failure).
- The same probability of success for each trial (denoted by p).
- Independent trials.

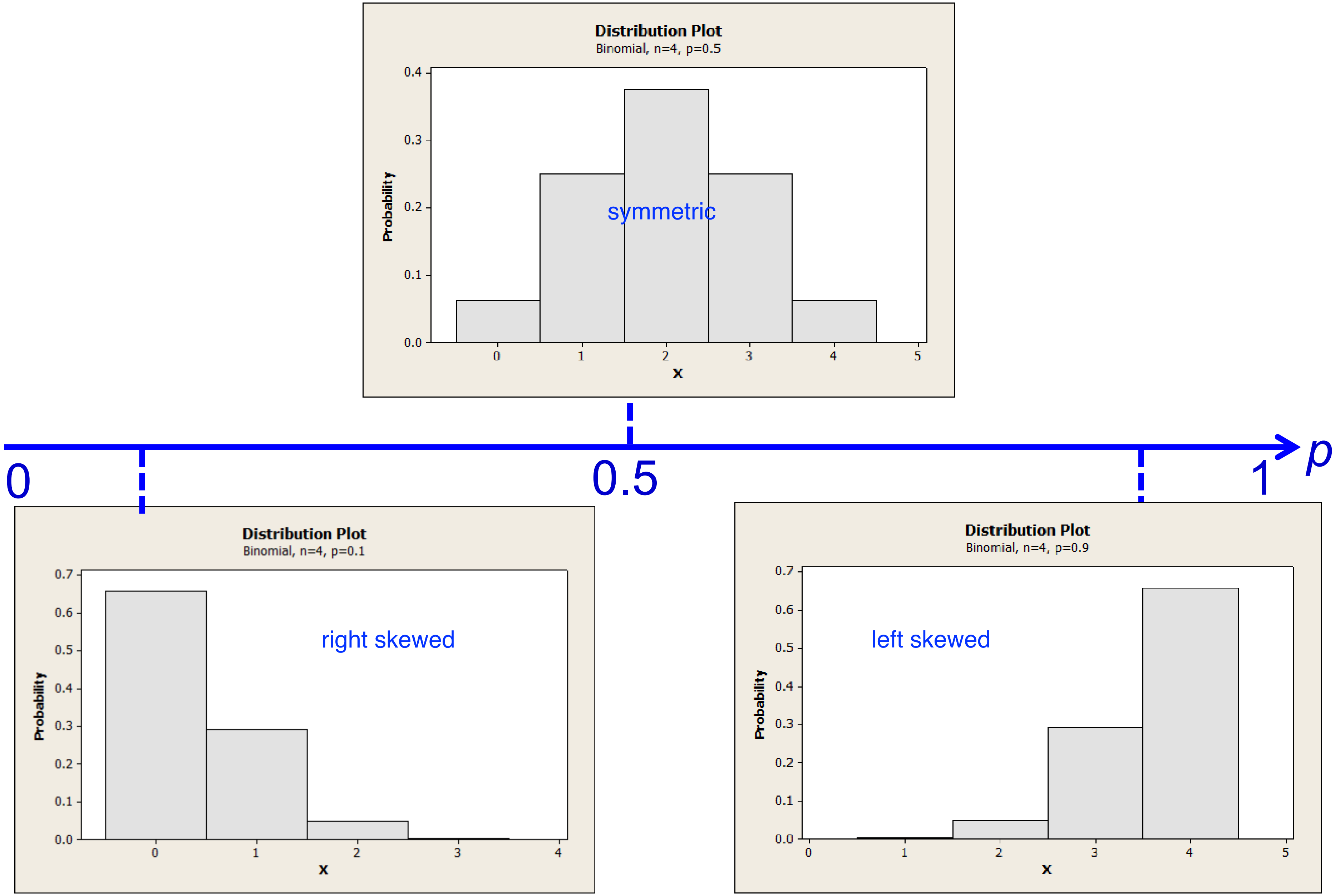
Mean and Standard Deviation of the Binomial Distribution

- The binomial probability distribution for *n* trials with probability *p* of success on each trial has mean μ and standard deviation σ given by:

$$\boxed{\mu = np, \quad \sigma = \sqrt{np(1\text{-}p)}}$$

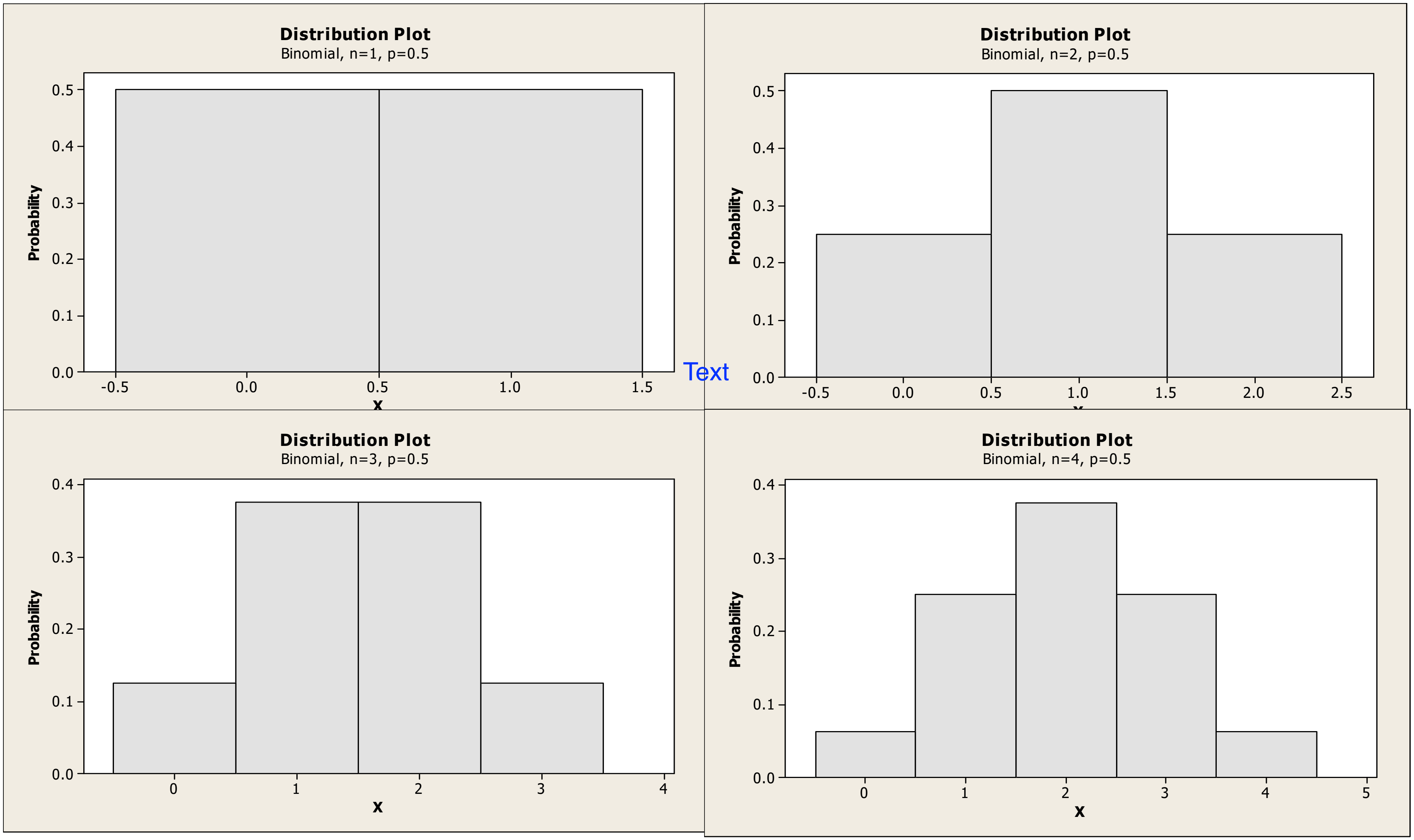
- The shape of the distribution depends on the values of *p* and *n.*

Effect of p

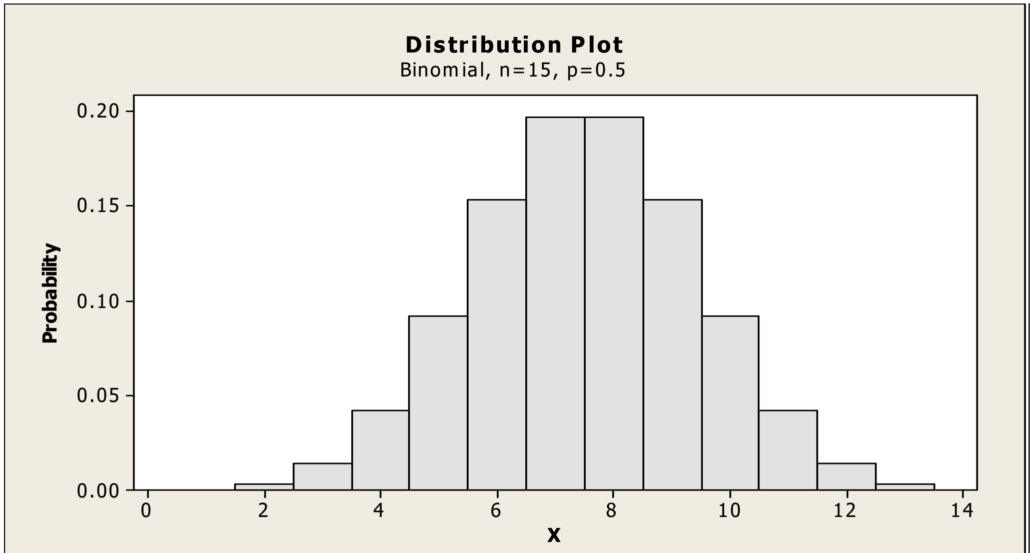


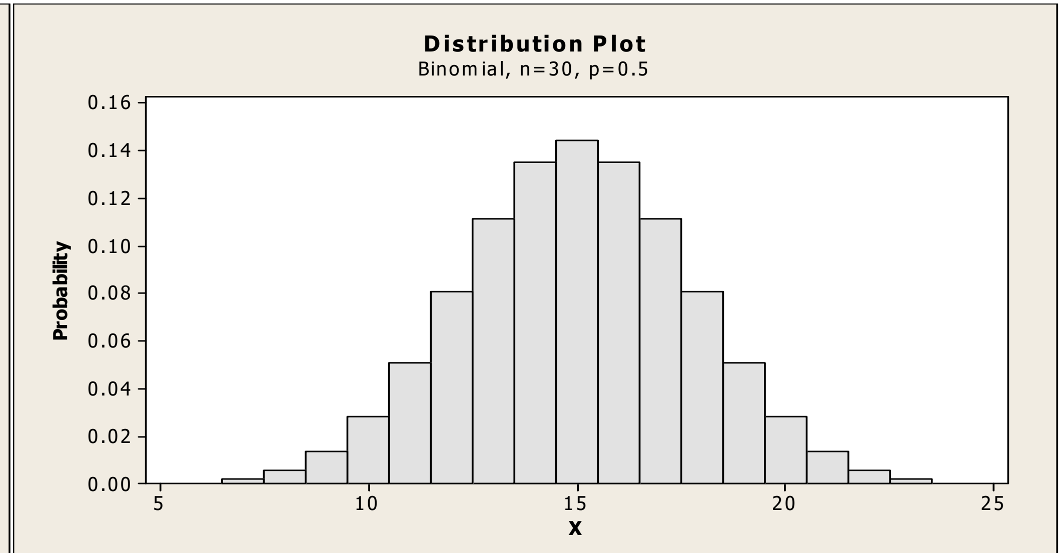
Distribution Plot
Binomial, n=4, p=0.5
symmetric
Probability
x
0
0.5
1
p
Distribution Plot
Binomial, n=4, p=0.1
right skewed
Probability
x
Distribution Plot
Binomial, n=4, p=0.9
left skewed
Probability
x

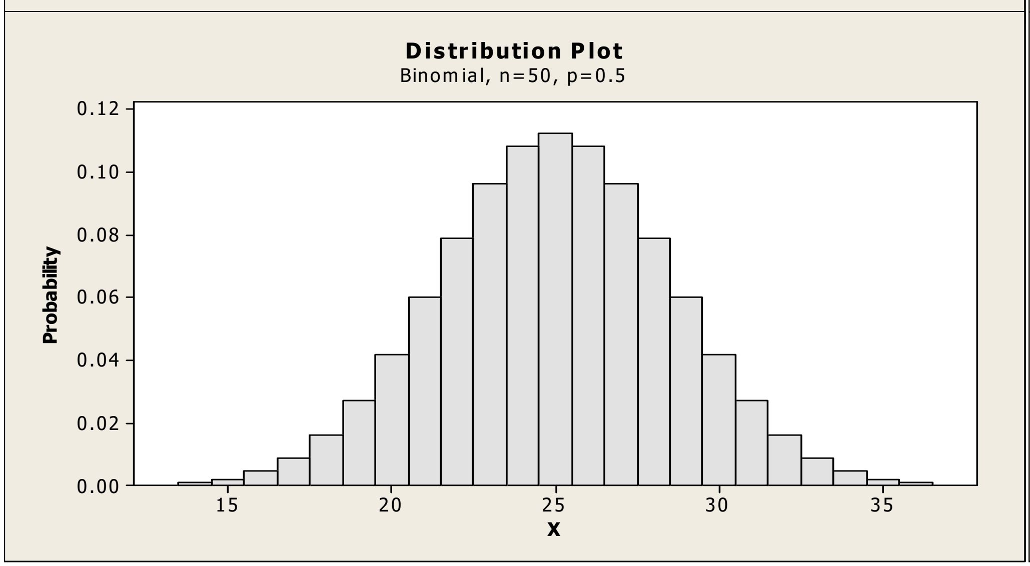
Effect of *n* when *p* = 0.50

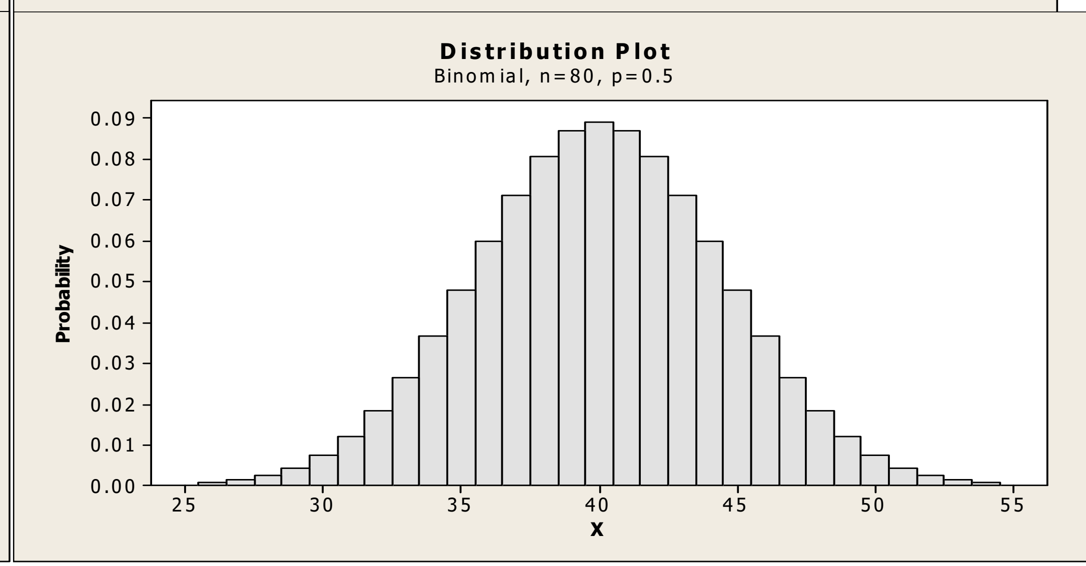


Effect of *n* when *p* = 0.50



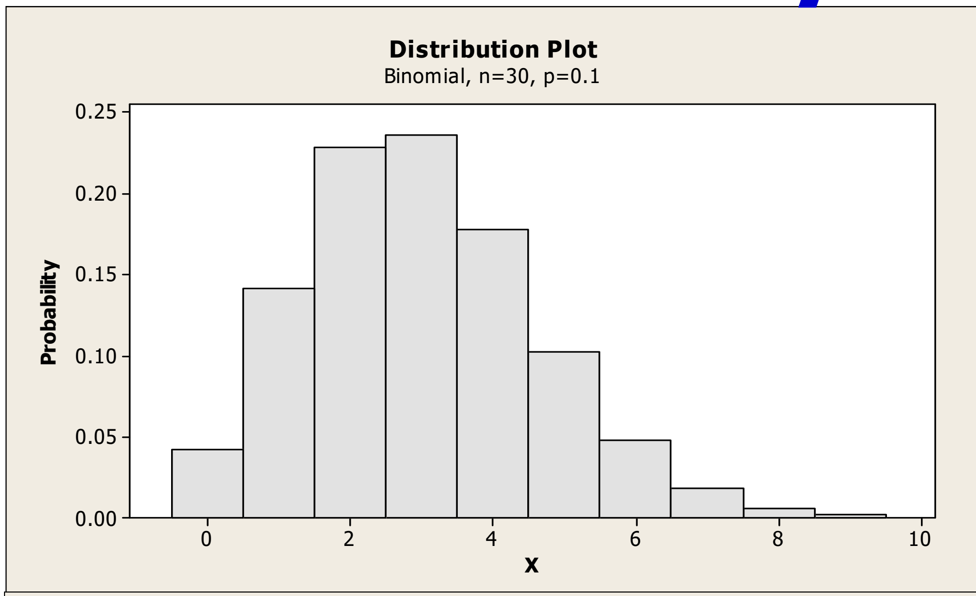


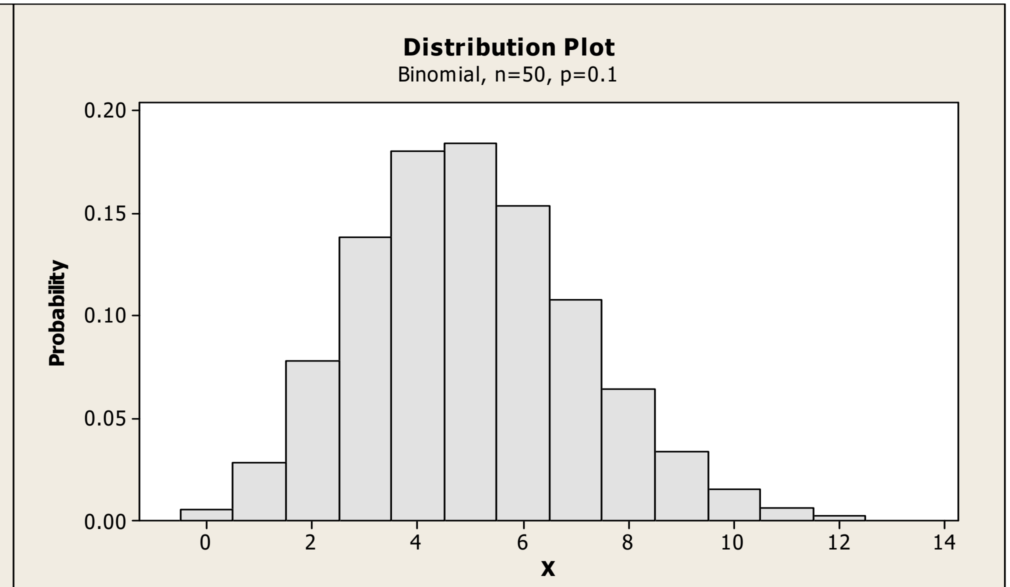


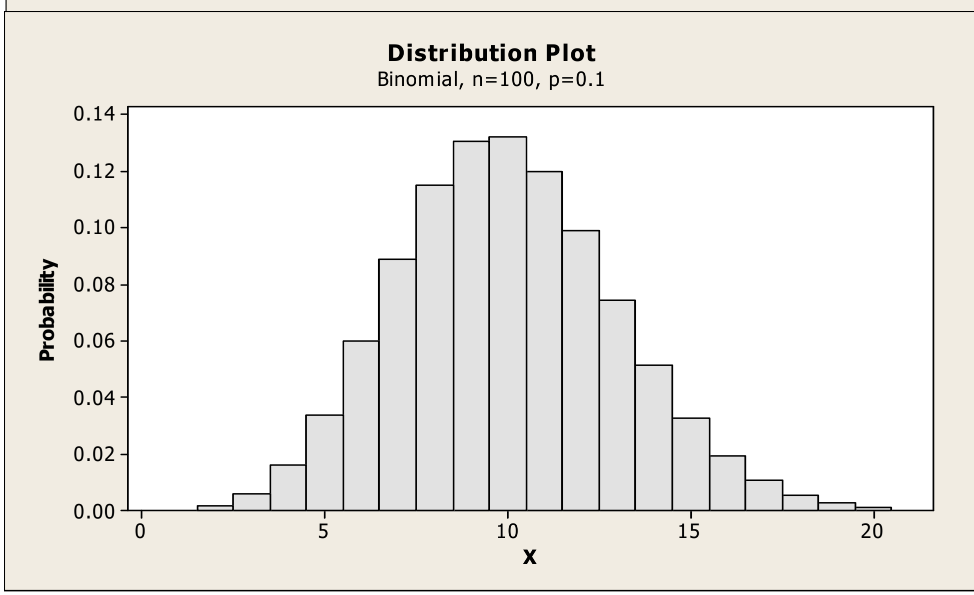


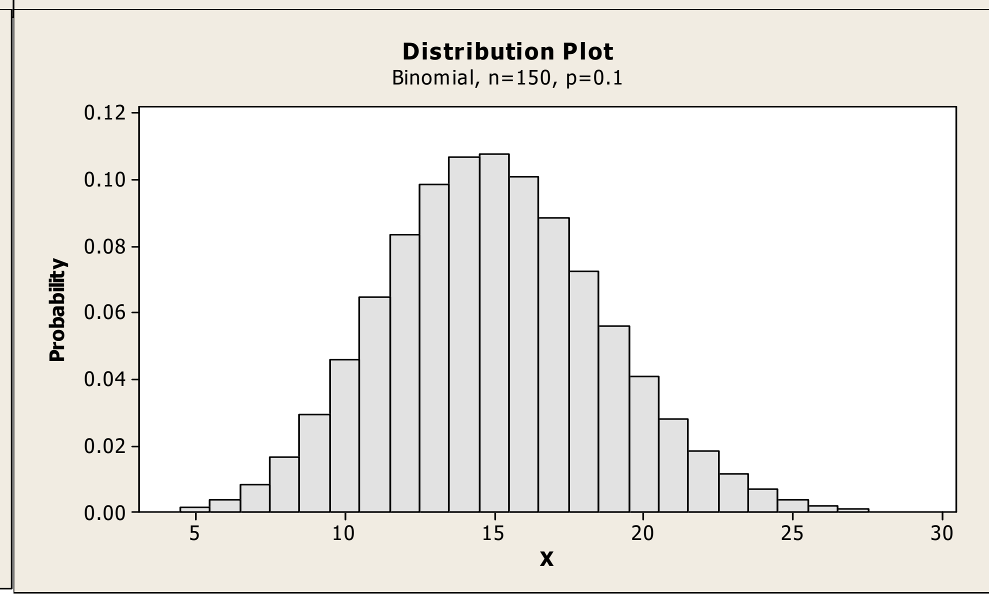
Effect of n when $p \neq 0.50$

$p = 0.10$ RIGHT SKEWED

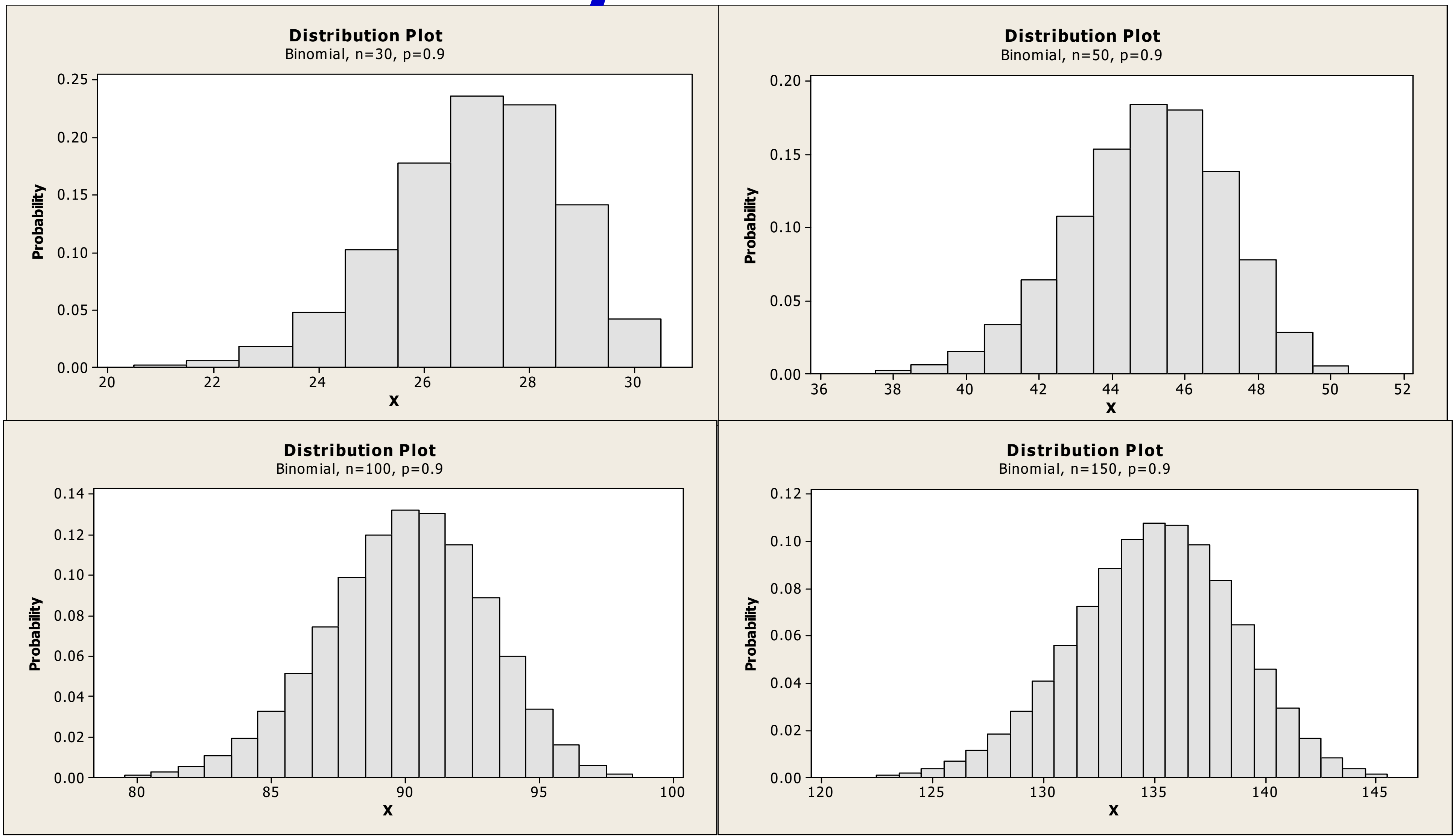








Effect of n when $p \neq 0.50$
$p = 0.90$



When Is the Binomial Distribution Bell Shaped?

The binomial distribution will be close to a bell-shaped distribution when:

The expected number of successes, *np*, & the expected number of failures, *n(1-p)*

are both at least 15.

Example: Checking for racial profiling

- Study conducted by the American Civil Liberties Union.
- Study analyzed whether African-American drivers were more likely than other in the population to be targeted by police for traffic stops.
- Data:
  - 262 police car stops in Philadelphia in 1997.
  - 207 of the drivers stopped were African-American.
  - In 1997, Philadelphia's population was 42.2% African-American.
- Does the number of African-Americans stopped suggest possible bias, being higher than we would expect (other things being equal, such as the rate of violating traffic laws)?

Example: Checking for racial profiling

- Assume:
  - 262 car stops represent $n = 262$ trials.
  - Successive police car stops are independent.
  - P(driver is African-American) is $p = 0.422$.

- Calculate the mean and standard deviation of this binomial distribution:

Example: Checking for racial profiling

- Recall: Empirical Rule

  When a distribution is bell-shaped, about all (99.7%) of it falls within 3 standard deviations of the mean.

- If no racial profiling is happening, we would not be surprised if between about ____87____ and ____135____ of the 262 people stopped were African-American.
- The actual number stopped (207) is well above these values. The number of African-American stopped is too high, even taking into account random variation.

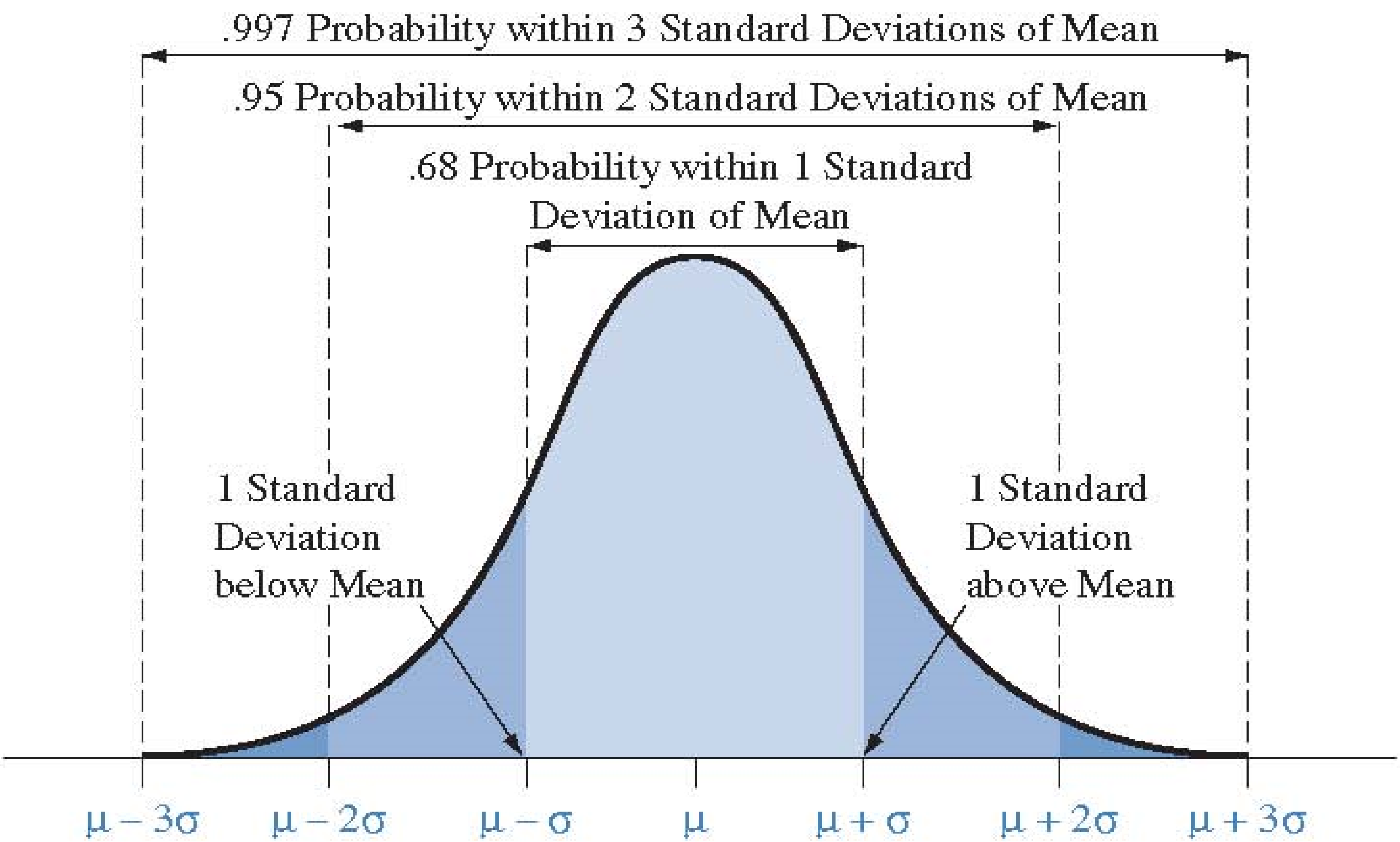
Question: Binomial Probability Distribution

X has a binomial probability distribution with $n = 20$. If the expected number of successes is 14, then the expected number of failures is

a. 6
b. 10
c. 14
d. 20

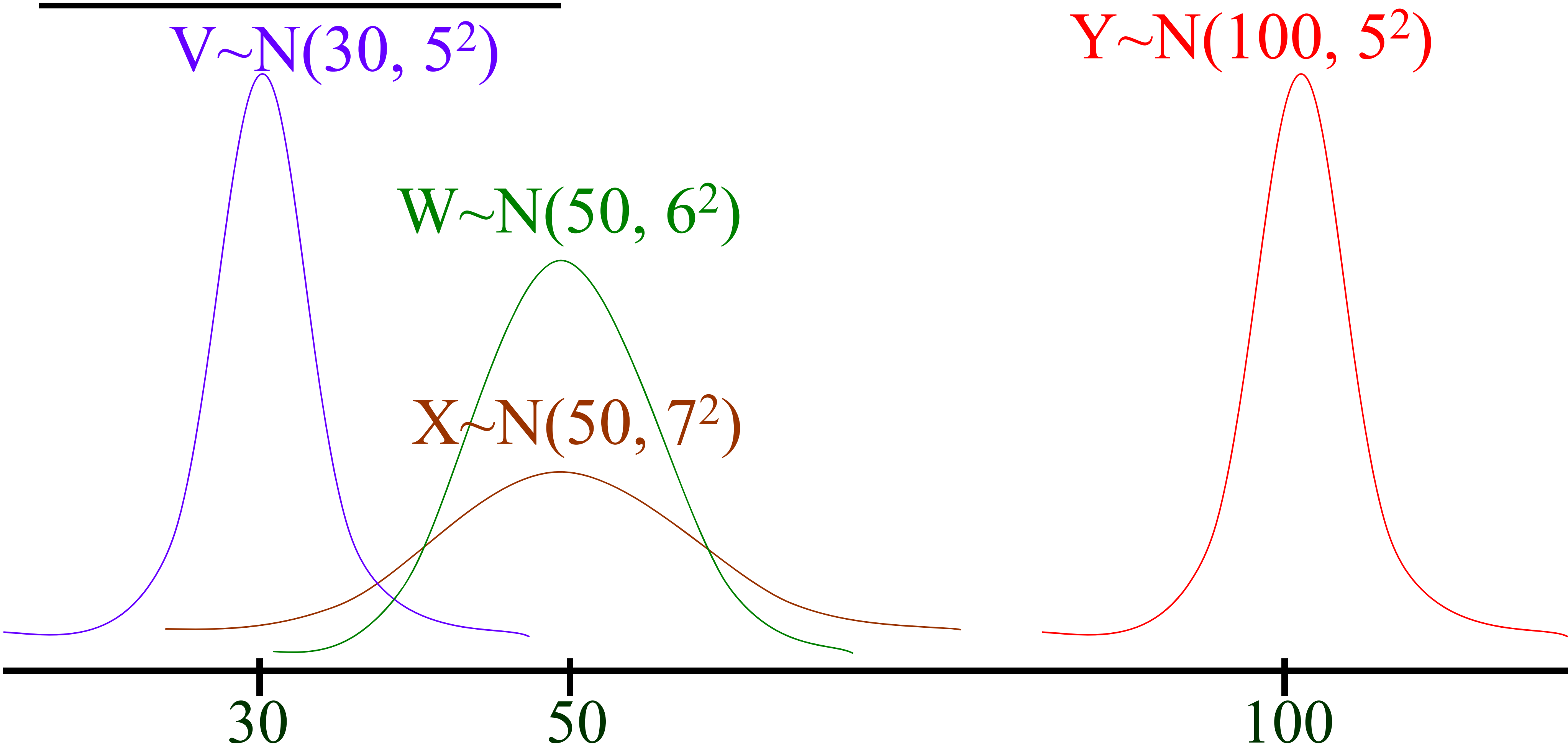
Normal Distributions $X \sim \text{Normal}(\mu, \sigma^2)$

The normal distribution is bell-shaped (symmetric & unimodal).
It is fully characterized by mean μ and standard deviation σ.



Family of Normal Distributions

The probability of falling within any particular number of standard deviations, z of the mean, μ is the <u>same for all normal distributions</u>.



Z-Score & The Empirical Rule

- The *z-score* for an observation is the number of standard deviations that it falls from the mean.

$$z = \frac{x - \mu}{\sigma}$$

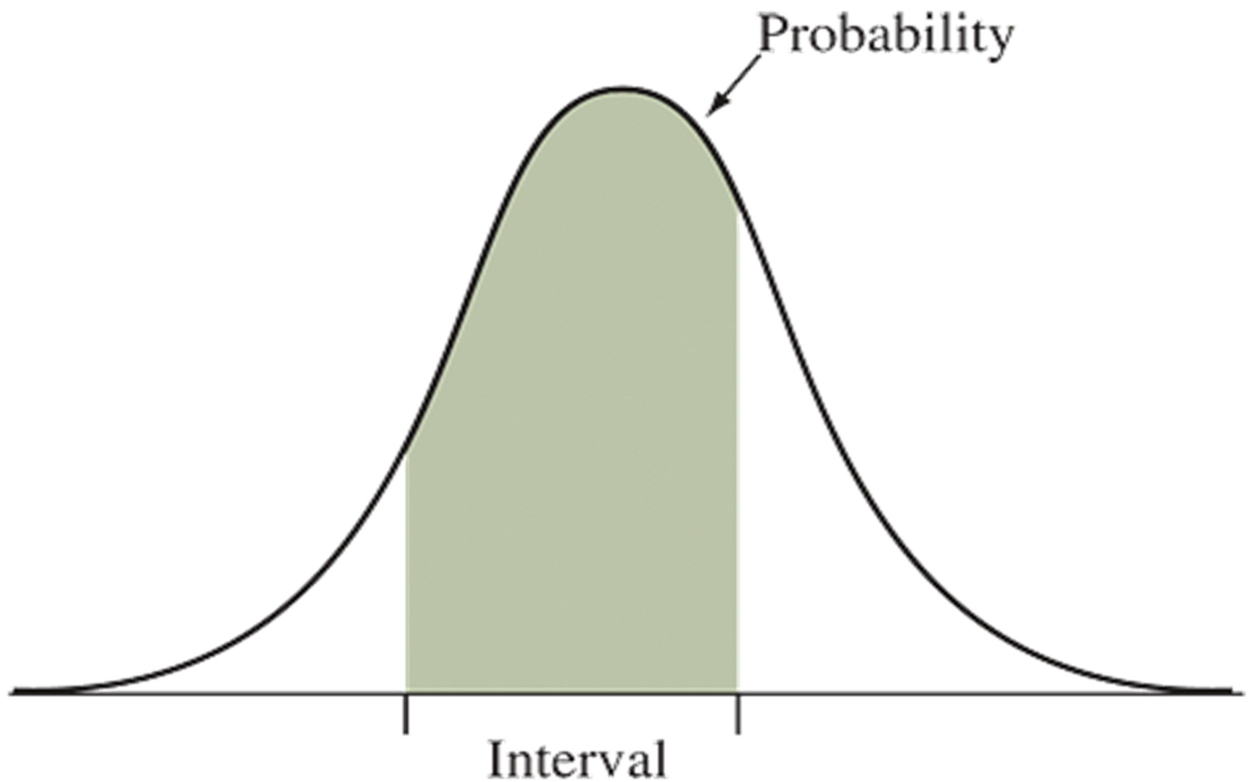
- For z = 1:
  68% of the area (probability) of a normal distribution falls between: $\mu - 1\sigma$ and $\mu + 1\sigma$

- For z = 2:
  95% of the area (probability) of a normal distribution falls between: $\mu - 2\sigma$ and $\mu + 2\sigma$

- For z = 3:
  99.7% of the area (probability) of a normal distribution falls between: $\mu - 3\sigma$ and $\mu + 3\sigma$

Probability of Normal Distributions

The area under the curve within an interval is the probability of the variable in that interval.

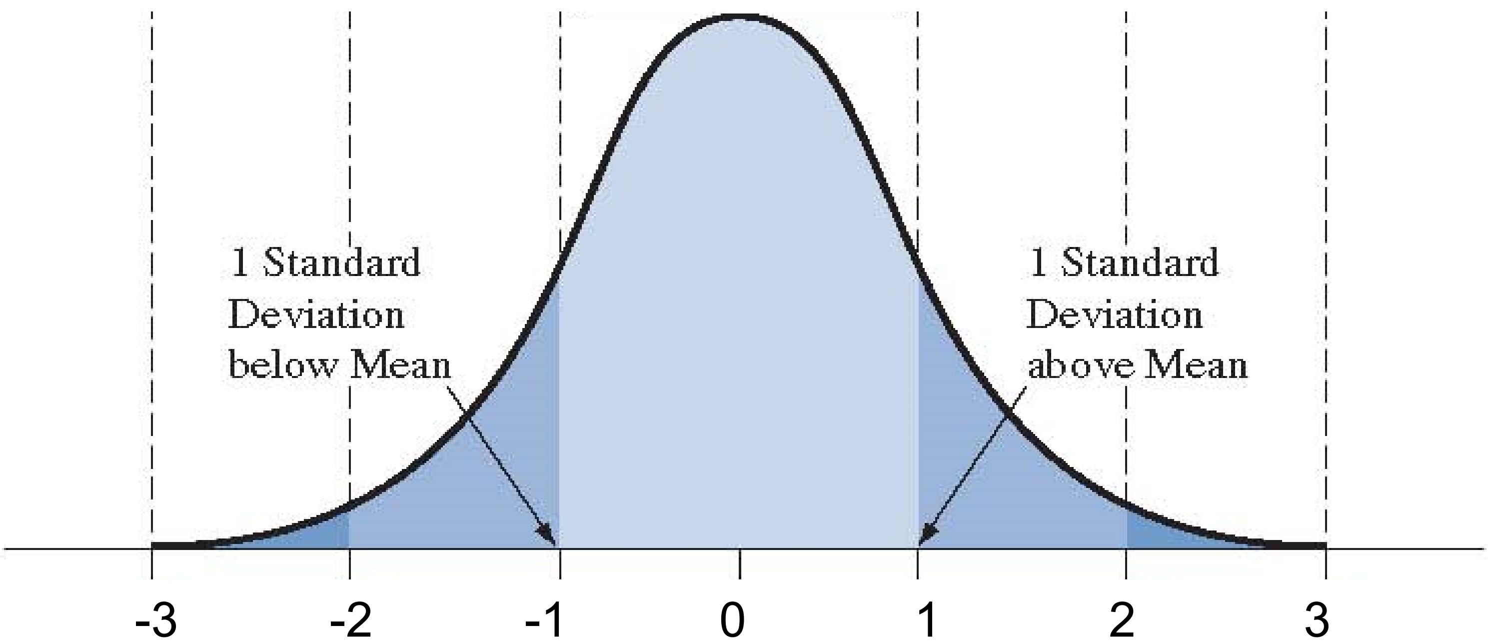


The Standard Normal Distribution

When a random variable has a normal distribution and its values are converted to z-scores by subtracting the mean and dividing by the standard deviation,

The z-scores have the standard normal distribution.

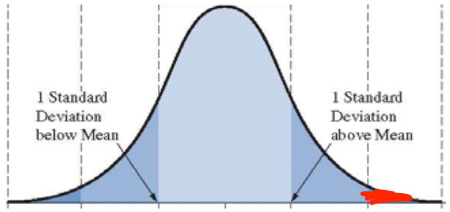
$$Z \sim N\ (\mu = 0, \sigma^2 = 1\ )$$



Example: Finding probabilities for various Z-values

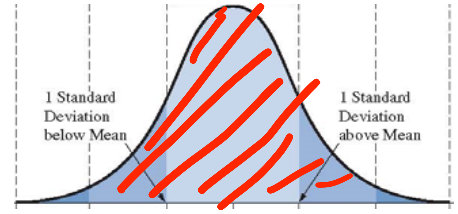
P(Z>3.00)

= 0.00135



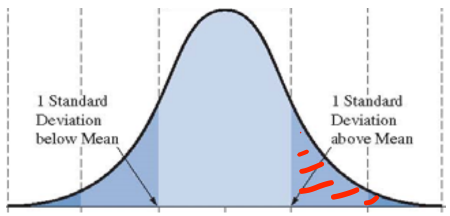
P(Z<3.00)

= 1-0.00135



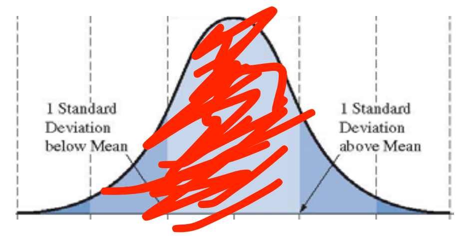
P(Z> 1.23)

= 0.1093



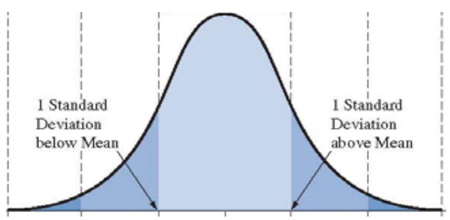
P(Z<1.23)

= 1- 0.1093



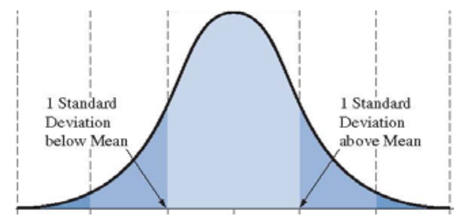
P(Z> -1.23)

= 1- 0.1093



P(Z<-1.23)

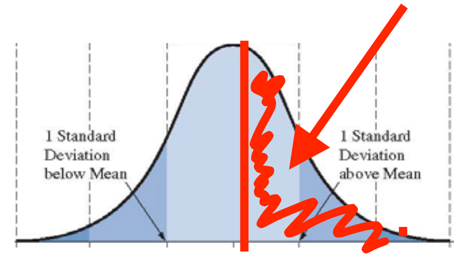
= 0.1093



Example: Finding probabilities for various Z-values

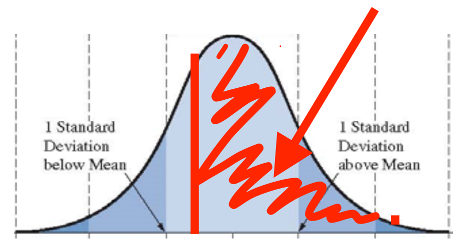
P(0.12 <Z< 2.34)

= 0.4522-0.00964



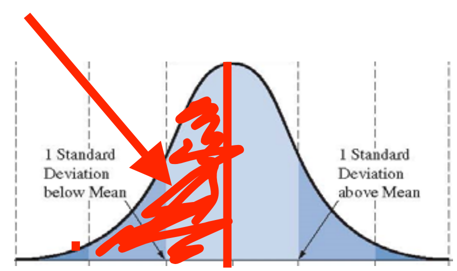
P(-0.12 <Z< 2.34)

= 1-0.4522-0.00964



P(-2.34 <Z<- 0.12)

= 0.4522-0.00964



Normal distributions

$$\begin{tikzcd}
X \sim N\,(\mu,\, \sigma^2) \arrow[d, "\text{z-score formula}"'] & \\
\boxed{Z \sim N\,(0,\, 1)} \arrow[r, shift left] & \text{Probabilities} \arrow[l, shift left, "\text{std normal table}"]
\end{tikzcd}$$

Example: Find z^* with the given probability

$P(Z > z^*) = 0.500$

$z^* =$ 0

$P(Z < z^*) = 0.500$

$z^* =$

$P(Z > z^*) = 0.025$

$z^* =$

$P(Z < z^*) = 0.025$

2.81 $z^* =$

$P(Z > z^*) = 0.975$

$z^* =$

$P(Z < z^*) = 0.975$

$z^* =$

Example:
What proportion of students get a grade of B?

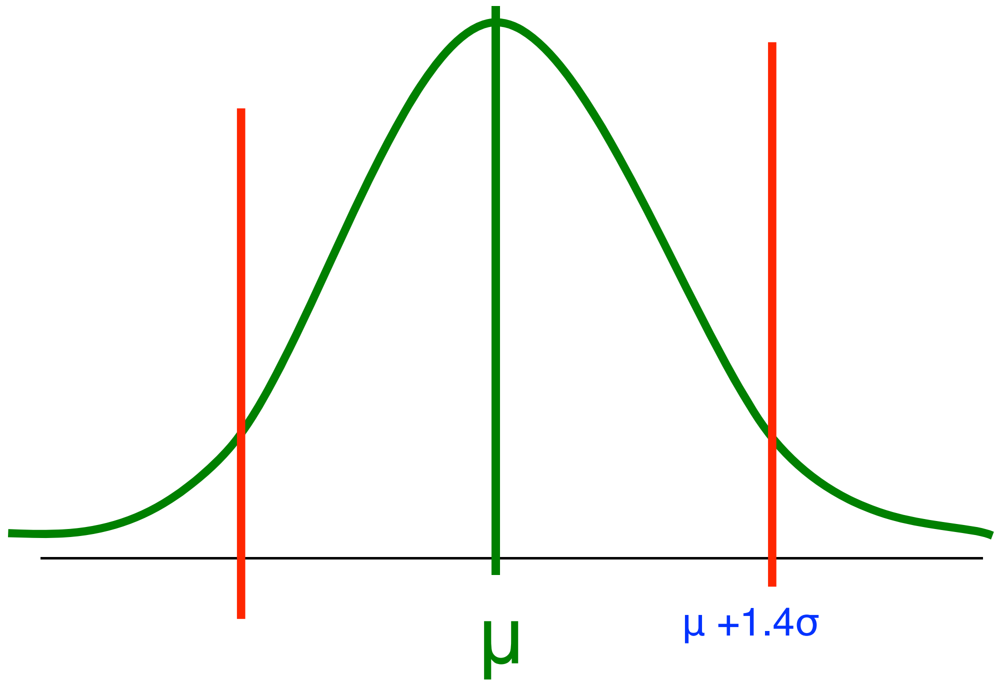
- On the midterm exam in introductory statistics, an instructor always give a grade of B to students who score between 70 and 80.
- One year, the scores on the exam have approximately a normal distribution with mean 73 and standard deviation 5.
- About what proportion of students get a B?

Let X = Exam score

X~ Normal ($\mathbf{\mu = 73, \sigma^2 = 5^2}$)

P(70< X < 80)

(70-73)/5 <χ<(80-73/5)

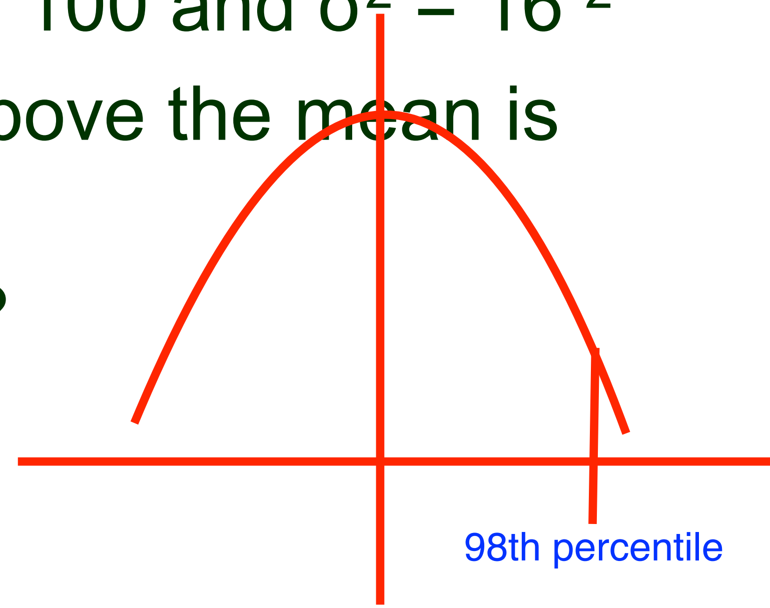


P(-0.6 < z < 1.40) = 1 – 0.2743 – 0.0808 = 0.6449

About 64% of the exam scores were in the ‘B’ range.

Example:
What IQ do you need to get into Mensa?

- Mensa is a society of high-IQ people whose members have a score on an IQ test at the 98th percentile or higher.
- IQ is normally distributed with $\mu = 100$ and $\sigma^2 = 16^2$
- How many standard deviations above the mean is the 98th percentile?
- What is the IQ for that percentile?



Let X = IQ of people

$X \sim$ Normal ($\boldsymbol{\mu = 100, \sigma^2 = 16^2}$)

$x = \mu + 2.05\sigma$

Example:
What IQ do you need to get into Mensa?

- How many standard deviations above the mean is the 98th percentile?

  $P(Z<z^*) = 0.98 \rightarrow z^* =$ 2.05

- What is the IQ for that percentile?

  Since $\mu = 100$ and $\sigma = 16$,

  let 98th percentile of IQ = x^*

  $x^* = \mu + z^* \sigma$

  133

$$z = \frac{x - \mu}{\sigma}$$

Topic 5
(Chapter 7)

Sampling Distributions

probability distribution for statistics

Learn

1. How likely are the possible values of a sample proportion?

2. How close are sample means to population means?

3. How can we make inferences about a population?

Statistics and Parameters

Recall:

A *statistic* is a numerical summary of sample data, such as a *sample proportion* or a *sample mean*.

A *parameter* is a numerical summary of a population, such as a *population proportion* or a *population mean.*

How do we know that a sample statistic is a good estimate of a population parameter?

To answer this, we need to look at a probability distribution called the sampling distribution.

Sampling Distribution

The *sampling distribution* of a **statistic** is the probability distribution that specifies probabilities for the possible values the statistic can take.

Random variable → **Probability distribution (Standard deviation)**

Sample statistic → **Sampling distribution (Standard error)**

To distinguish the standard deviation of a *sampling distribution* from the standard deviation of an ordinary probability distribution, we refer to it as a ***standard error***.

The Sampling Distribution of the
Sample Proportion

- Look at each possible sample.
- Find the sample proportion for each sample.
- Construct the frequency distribution of the sample proportion values.
- This frequency distribution is the sampling distribution of the sample proportion.

Example:
Sampling Distribution of The Sample Proportion

- Which Brand of Pizza Do You Prefer?
  - Two Choices: A or D.
    2 outcomes
  - Assume that half of the population prefers Brand A and half prefers Brand D.
    $p = 0.50$
  - Take a random sample of $n = 3$ tasters.

2 x 2 x 2 = 8 possible outcomes

Example:
Sampling Distribution of The Sample Proportion

$x = \{0, 1, 2, 3\}$ random variable, X

$p = \{0, 1/3, 2/3, 1\}$ Statistics, p

Sample	# Prefer Pizza A, X	Sample Proportion, $\hat{p}$
(A,A,A)	3	1
(A,A,D)	2	2/3
(A,D,A)	2	2/3
(D,A,A)	2	2/3
(A,D,D)	1	1/3
(D,A,D)	1	1/3
(D,D,A)	1	1/3
(D,D,D)	0	0

Example:
Sampling Distribution of The Sample Proportion

x~B(n=3, p= 0.5)

Probability Distribution

X	Probability	$X \cdot P(X)$	$X^2 \cdot P(X)$
0	1/8	0	0
1	3/8	3/8	3/8
2	3/8	6/8	12/8
3	1/8	3/8	9/8
	Total	12/8	24/8

Mean = 1.5 = np

SD = sqrt(0.75) = npq

Sampling Distribution

$\hat{p}$	Probability	$\hat{p} \cdot P(\hat{p})$	$\hat{p}^2 \cdot P(\hat{p})$
0	1/8	0	0
1/3	3/8	1/8	1/24
2/3	3/8	2/8	4/24
1	1/8	1/8	3/24
	Total	4/8	8/24

Mean = 0.5 = p

SE = sqrt(1/12) = sqrt(p(1-p)/n)

Mean and Standard Deviation of the Sampling Distribution of a Proportion

For a random sample of size *n* from a population with proportion *p*, the sampling distribution of the sample proportion has

$$\text{Mean} = p \;\; ; \;\; \text{standard error} = \sqrt{\frac{p(1-p)}{n}}$$

If *n* is sufficiently large such that the expected numbers of outcomes of the two types, *np* and *n(1-p),* are both at least 15, then this sampling distribution has a bell-shape.

same as binomial distribution

Statistics	Sample proportion, $\hat{p} = \dfrac{X}{n}$	Sample mean, $\bar{x} = \dfrac{\sum X}{n}$
Mean	μ = p	
Standard Error	σ = sqrt(p(1-p)/n)	

Example: Election

- Prior to counting the votes, the proportion in favor of president T was an unknown parameter.
- An exit poll of 3160 voters reported that the sample proportion in favor of president T was **0.54**. $= \hat{p}$
- If a different random sample of about 3160 voters were selected, a different sample proportion would occur. Imagine all the distinct samples of 3160 voters you could possibly get. Each such sample has a value for the sample proportion.

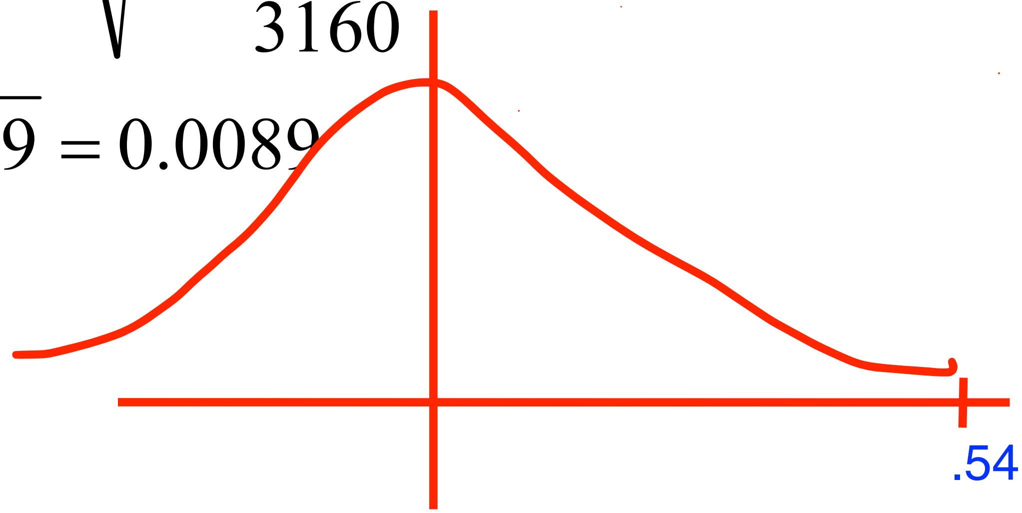
Example: Election

- If the population proportion supporting president T was 0.50, how likely to observe the exit-poll sample proportion of 0.54?

$$\text{Mean} = p$$
$$= 0.50$$

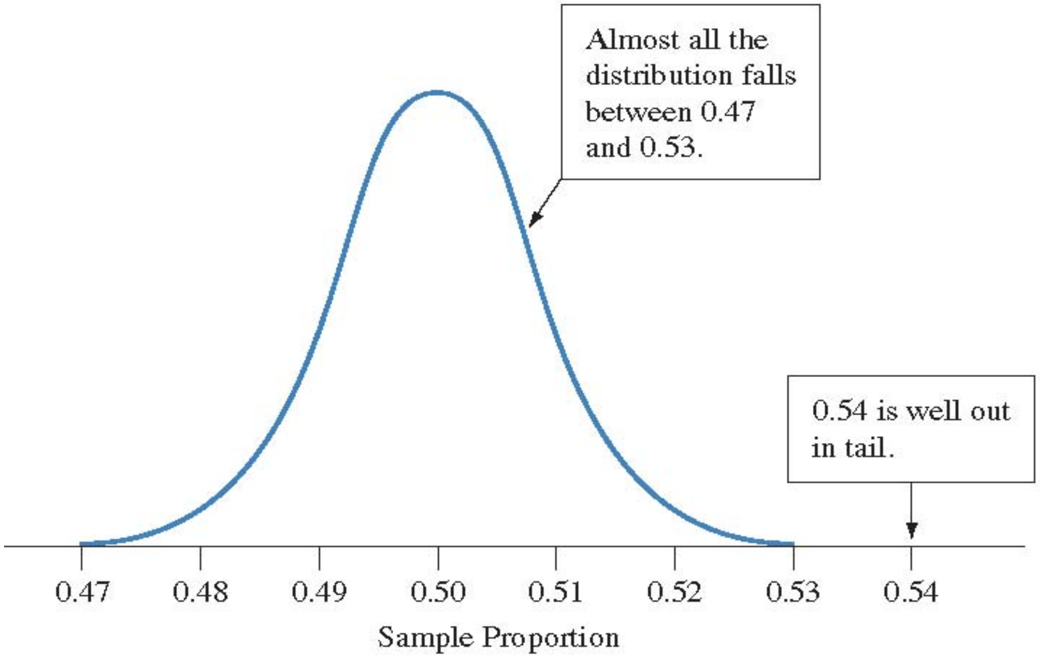
$$\text{Standard Error} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.50)(0.50)}{3160}}$$
$$= \sqrt{0.000079} = 0.0089$$

$$z = \frac{(0.54 - 0.50)}{0.0089} = 4.5$$



Example: Election

- The sample proportion of 0.54 is more than four standard errors from the expected value of 0.50.



- The sample proportion of 0.54 voting for him would be very unlikely if the population support were p = 0.50.

- We then have strong evidence that the population support was **larger than 0.50**.

- The exit poll gives strong evidence that president T would win in the election.

Mean and Standard Error of the Sampling Distribution of the Sample Mean

- For a random sample of size *n* from a population having mean μ and standard deviation σ, the sampling distribution of the sample mean has:

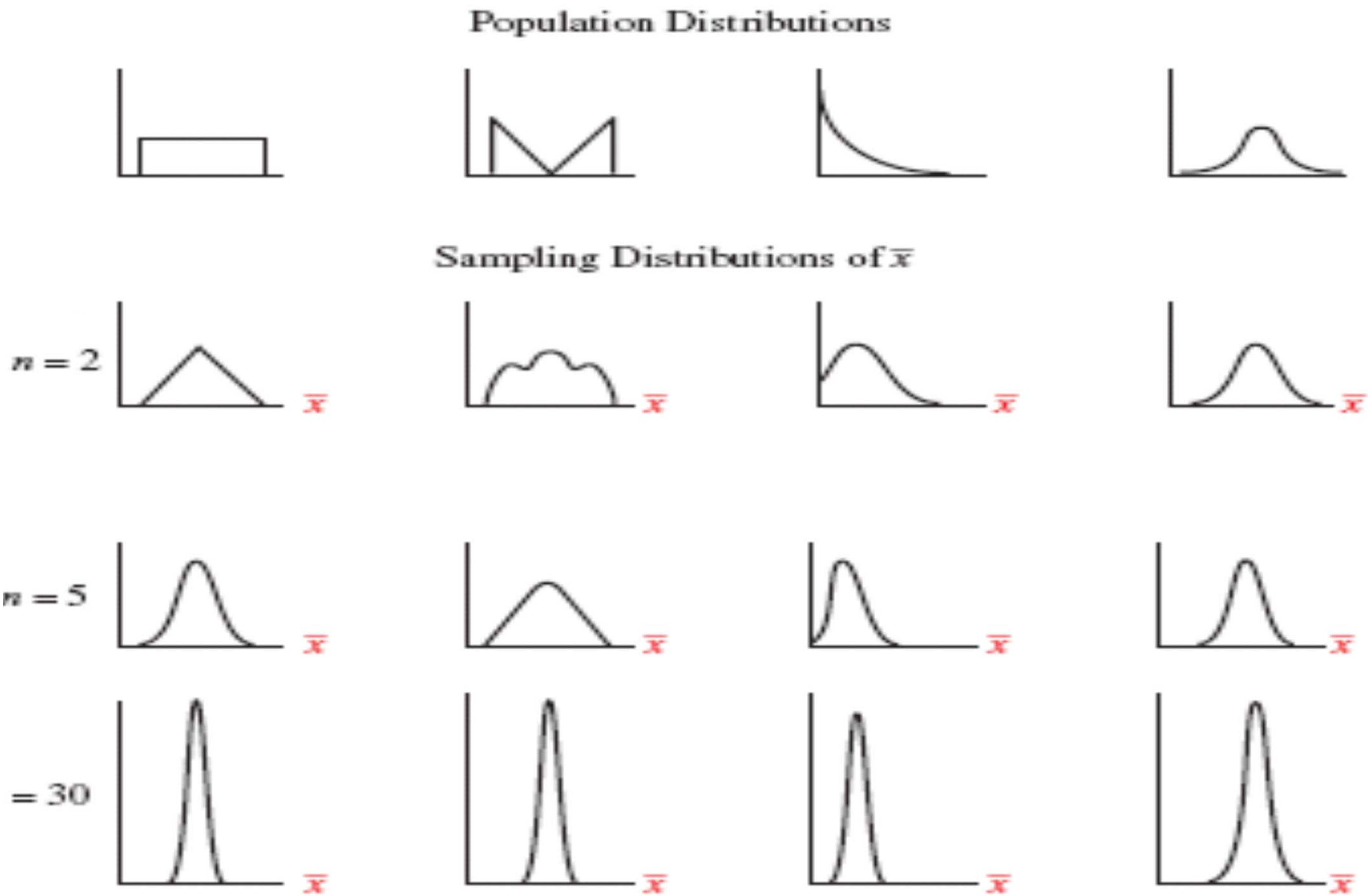
$$\text{Mean} = \mu \ ; \ \text{standard error} = \frac{\sigma}{\sqrt{n}}$$

- If *n* is sufficiently large ($n \geq 30$), then this sampling distribution has a bell-shape.

Central Limit Theorem (CLT)

- For random sampling with a large sample size *n*, the sampling distribution of the sample mean is approximately a normal distribution.
- This result applies *no matter what the shape* of the probability distribution from which the samples are taken.
- The sampling distribution of the sample mean takes more of a bell shape as the random sample size *n* increases.
- The more skewed the population distribution, the larger *n* must be before the shape of the sampling distribution is close to normal.
- In practice, the sampling distribution is usually close to normal when the sample size *n* is at least about 30.

CLT: Impact of increasing n



Example:
Sampling Distribution of The Sample Mean

Consider a small population: 2, 4, 6, 8

- The mean $\mu = \dfrac{\sum x}{N} = \dfrac{2+4+6+8}{4} = 5.0$

- The standard deviation

$$\sigma = \sqrt{\frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4}}$$

$$= \sqrt{5} \quad = 2.2$$

Example:
Sampling Distribution of The Sample Mean

Population: **2, 4, 6, 8**

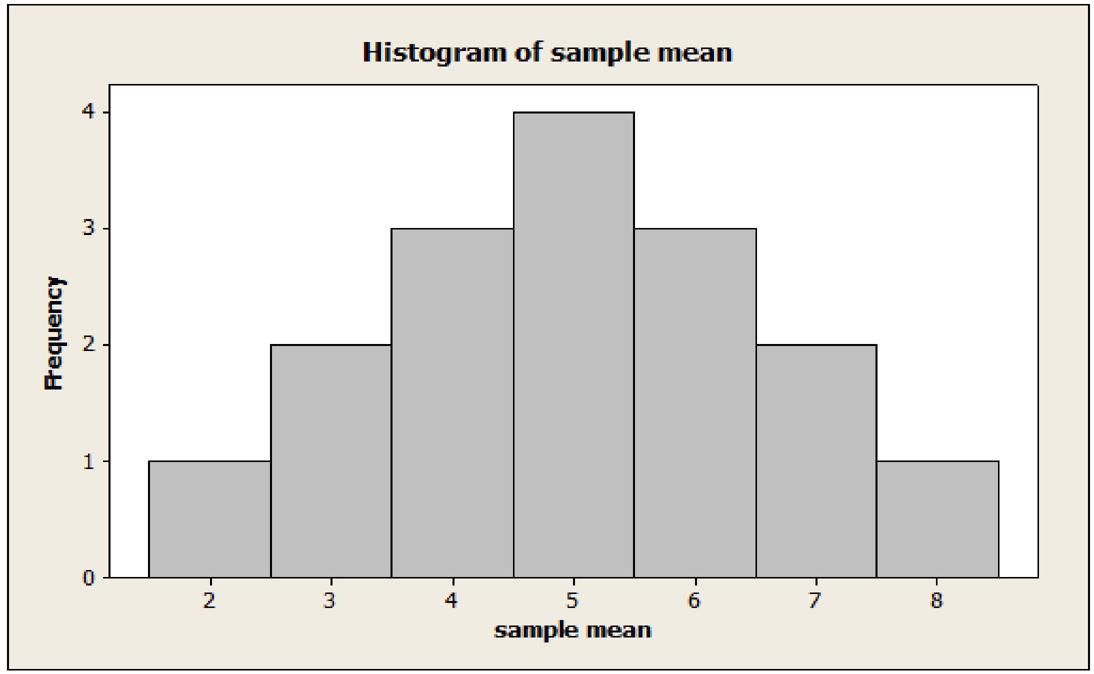
All samples of size 2 are taken with replacement and the mean of each sample is computed:

Sample	Mean	Sample	Mean
2, 2	2	6, 2	4
2, 4	3	6, 4	5
2, 6	4	6, 6	6
2, 8	5	6, 8	7
4, 2	3	8, 2	5
4, 4	4	8, 4	6
4, 6	5	8, 6	7
4, 8	6	8, 8	8

Example:
Sampling Distribution of The Sample Mean

Frequency distribution of the sample means:

Sample Mean	Frequency	Relative Frequency
2	1	1/16=0.0625
3	2	2/16=0.1250
4	3	3/16=0.1875
5	4	4/16=0.2500
6	3	3/16=0.1875
7	2	2/16=0.1250
8	1	1/16=0.0625
Total	**16**	**1**



Example:
Sampling Distribution of The Sample Mean

The mean of all the sample means,

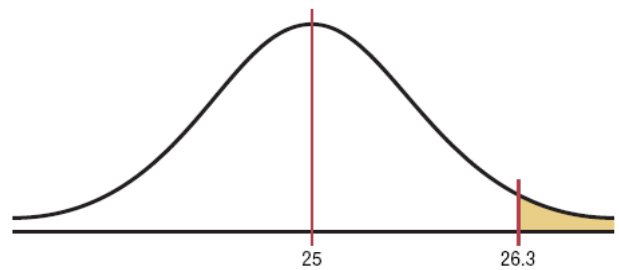
$$\mu_{\bar{x}} = \frac{1(2)+2(3)+3(4)+4(5)+3(6)+2(7)+1(8)}{16}$$

The standard deviation of all the sample means,

$$\sigma_{\bar{x}} = \sqrt{\frac{(2-5)^2+(3-5)^2+(3-5)^2+\Lambda\ +(7-5)^2+(8-5)^2}{16}}$$

Example: Hours of Television

A. C. Neilsen reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours.



Let X = amount of time spent watching TV.

$X \sim$ normal ($\mu = 25$, $\sigma^2 = 3^2$)

For n = 20, $\bar{x} \sim$ normal ($\mu = 25$, $\sigma^2 = (3/\sqrt{20})^2$)

$P(\bar{x} > 26.3)$

Three Distinct Types of Distributions

1. Population Distribution

- described by parameters (usually unknown)

2. (Sample) Data Distribution

- described by statistics

3. Sampling Distribution

- For large *n*, the sampling distribution is approximately a normal distribution.
- provides probabilities for all the possible values of the statistic.