Lesson 6
Part 2

Probability Distributions for Discrete Random Variables

Review and Overview

- So far we’ve covered the following probability and probability distribution topics
  - Probability rules
  - Probability tables to describe the distributions of Nominal variables
  - Probability density curves for continuous variables – particularly the Normal Distribution
- Lesson 6 Part 2 covers two probability distributions for discrete random variables with particular focus on
  - Binomial Distribution
  - Poisson Distribution

Review: Probability Distributions

- Any characteristic that can be measured or categorized is called a *variable*.
- If the variable can assume a number of different values such that any particular outcome is determined by chance it is called a *random variable.*
- Every random variable has a corresponding *probability distribution*.
- The probability distribution applies the theory of probability to describe the behavior of the random variable.

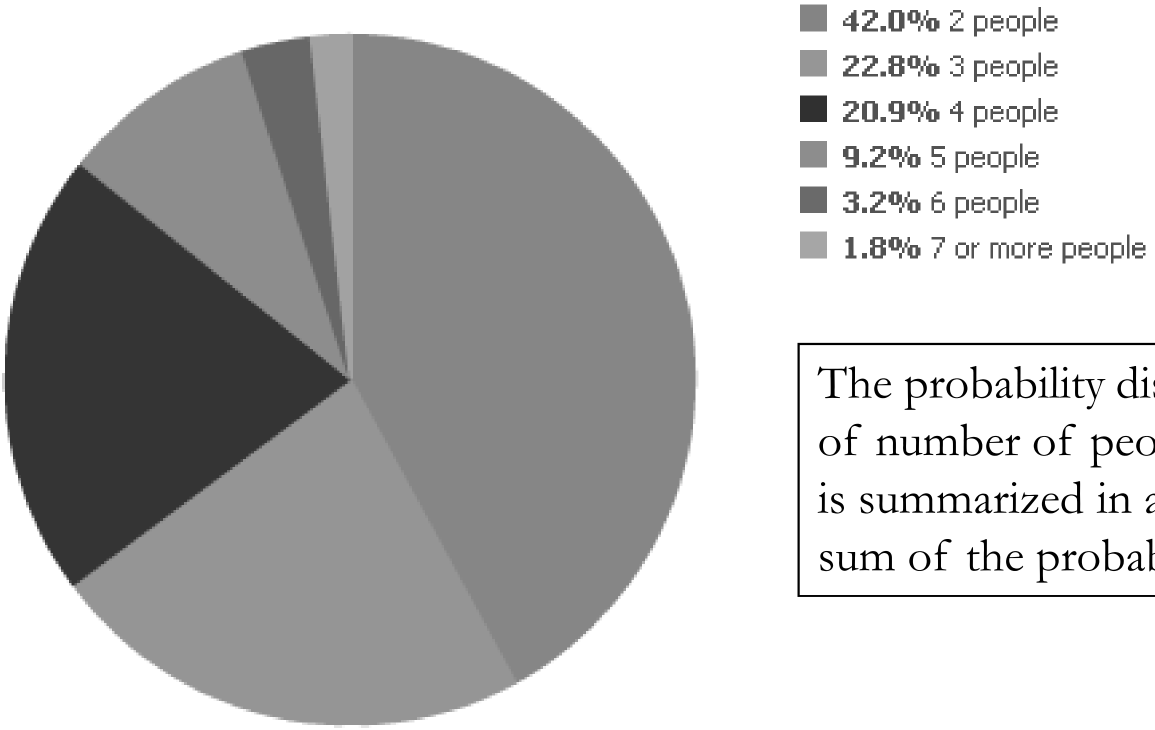
Discrete Random Variable

- A discrete random variable X has a finite number of possible integer values. The probability distribution of X lists the values and their probabilities in a table

Value of X	x_1	x_2	x_3	…	x_k
Probability	p_1	p_2	p_3	…	p_k

1. Every probability p_i is a number between 0 and 1.
2. The sum of the probabilities must be 1.

Distribution of number of people in a family (U.S. 2005 data)



The probability distribution of number of people in a family is summarized in a pie chart. The sum of the probabilities = 1.0

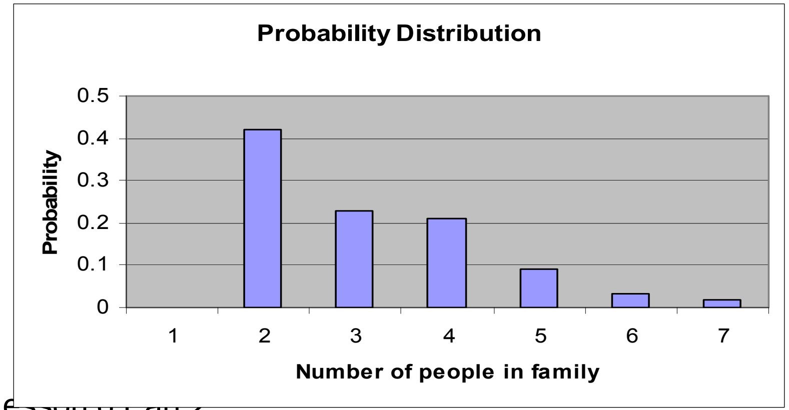
Source: U.S. Census Bureau, *Current Population Survey*, 2005 data.

Probability Distributions for Discrete Random Variables

- With a finite number of possible values the probability distributions for discrete random variables can be summarized in
  - Table of probabilities – use this data to answer Question 1 on the Practice Exercises

People in family	2	3	4	5	6	7+
Probability	0.42	0.228	0.209	0.092	0.032	0.018

  - Pie chart or Bar chart



Binomial Distribution

- The binomial distribution describes the distribution of outcomes from a series of trials where each trial meets certain conditions.
- Recall: a trial is a single event or experiment
  - For example, a single coin toss is a trial
  - For clinical data, a subject in the study can be considered a ‘trial’

Binomial Distribution

IF the following conditions are met

- Each trial has a binary outcome
- One of the two outcomes is labeled a ‘success’
- The probability of ‘success’ for a single trial is known
- The probability of success is constant over all trials
- The number of trials is specified
- The trials are independent. That is, the outcome from one trial doesn’t affect the outcome of successive trials

THEN

- The Binomial distribution describes the probability of X successes in n trials

Examples of Binomial distribution

- A classic example of the binomial distribution is the number of heads (X) in n coin tosses
  - Each trial (toss) has two possible outcomes: heads or tails
  - Label one outcome as a ‘success’ – heads
  - The probability of heads for a single trial = 0.5
  - This probability is constant over all coin tosses
  - Let ‘n’ represent the number of trials
  - Each coin toss is independent
- The number of heads in ‘n’ trials can range from 0 to n. The binomial distribution describes the probability of each possible outcome.

Practice Exercise Q 2a, b

For each coin toss the outcome is either Heads or Tails

- In a series of two coin tosses, what are the 4 possible combinations of outcomes?

- Calculate the probability of 0, 1, or 2 heads in 2 coin tosses

Number of heads	**0**	**1**	**2**
Probability	**0.25**	**0.50**	**0.25**

Bernoulli trial, process, distribution

Sometimes the elements of the binomial distribution are referred to as a ‘Bernouilli’ trial, process or distribution in honor of Jakob Bernoulli, who first described this distribution.

- *Bernoulli* trial: a trial involving a binomial probability
  - The event (or trial) results in only one of two mutually exclusive outcomes
- *Bernoulli* process: a sequence of independent Bernoulli trials
- *Bernoulli* distribution: the probability distribution of a random binary variable

Jakob Bernoulli

Basic principles of the binomial distribution were developed by the Swiss mathematician Jakob Bernoulli (1654-1705)

Following his father's wish, Jakob studied theology and entered the ministry. But contrary to the desires of his parents, he also studied mathematics and astronomy.

Source:http://en.wikipedia.org/wiki/Jakob_Bernoulli

Parameters of Binomial Distribution

- Each Binomial distribution is characterized by two parameters: n and π
  - n: the number of trials must be specified in advance
  - π: the probability of ‘success’ for one trial.
    - The probability of success is a parameter so the Greek letter for ‘p’ is used to represent the probability
- X, the binomial random variable, is the number of successes in ‘n’ trials
  - X can take on values 0 – n
  - The probability of each value of X can be calculated using the Binomial distribution
- The Notation for a binomial distribution is

$$X \sim B(n, \pi)$$

which is read as ‘X is distributed binomial with n trials and probability of success in one trial equal to π ’

Formula for Binomial Distribution

Using this formula, the probability distribution of a binomial random variable X can be calculated if n and π are known

$$P(X) = \frac{n!}{X!(n-X)!}\pi^{X}(1-\pi)^{n-X}$$

n! is called ‘n factorial’ = n(n-1)(n-2) . . .(1)

Example: 5! = 5*4*3*2*1 = 120

Note: 0! = 1

$$\pi^0 = 1$$

Practice Exercise Q 2c.

- Use the binomial formula to calculate the probability of 0, 1 and 2 heads in 2 coin tosses.
  - P(0 heads) =
  - P(1 head) =
  - P(2 heads) =

Another coin toss example

- What is the probability of 'X' heads in 6 coin tosses?
  - Success = 'heads'
  - $n = 6$ trials
  - $\pi = 0.5$
  - X = number of heads in 6 tosses which can range from 0 to 6
  - X has a binomial distribution with $n = 6$ and $\pi = 0.5$

$$X \sim B\,(6,\ 0.5)$$

Binomial Calculation for X ~ B (6, 0.5)

$$P(X=0)=\frac{6!}{0!(6-0)!}0.5^{0}(1-0.5)^{6-0}=0.5^{6}=0.0156$$

$$P(X=1)=\frac{6!}{1!(6-1)!}0.5^{1}(1-0.5)^{6-1}=6*0.5^{6}=0.09375$$

$$P(X=2)=\frac{6!}{2!(6-2)!}0.5^{2}(1-0.5)^{6-2}=15*0.5^{6}=0.234$$

$$P(X=3)=\frac{6!}{3!(6-3)!}0.5^{3}(1-0.5)^{6-3}=20*0.5^{6}=0.3125$$

$$P(X=4)=P(X=2)$$

$$P(X=5)=P(X=1)$$

$$P(X=6)=P(X=0)$$

$X \sim B\ (6, 0.5)$

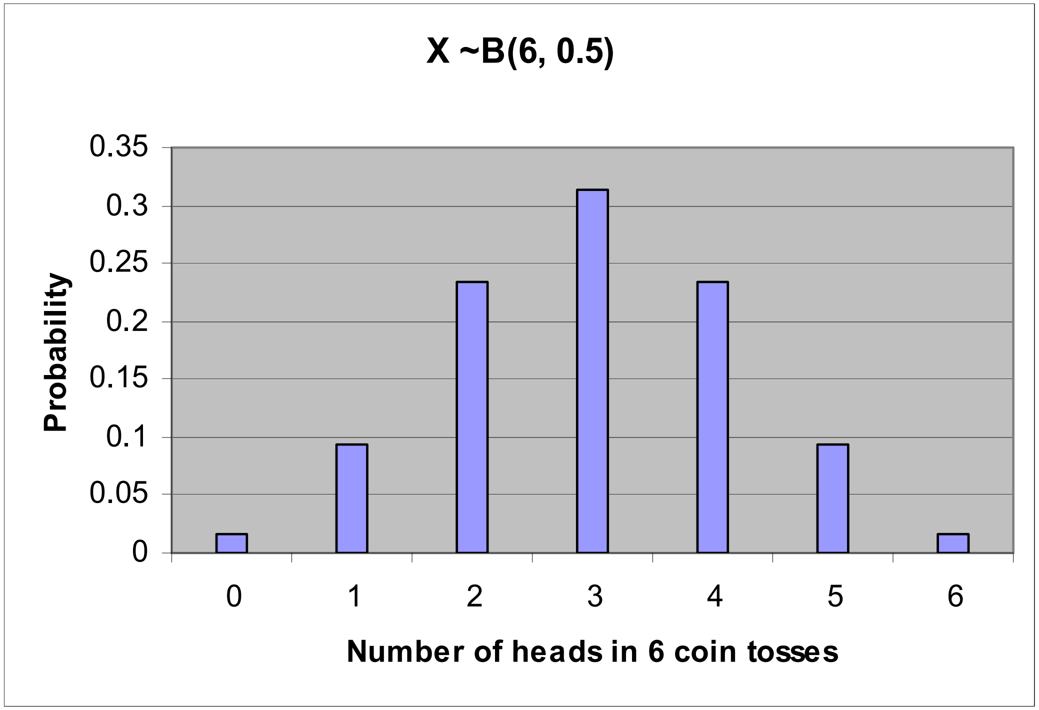
- The binomial distribution for the number of heads in 6 coin tosses is summarized in the table

# Heads	0	1	2	3	4	5	6
Probability	0.0156	0.094	0.234	0.3125	0.234	0.094	0.0156

- This probability distribution is symmetric because the P(Heads) = P(Tails) = 0.5

X ~ B(6, 0.5)

- For a small number of possible values, the binomial distribution can be plotted as a bar chart



A clinical application of the binomial distribution

- If 5 year survival rate for a certain group of prostate cancer patients is known to be 0.80, what is the probability that X of 10 men with prostate cancer survive at least 5 years.
  - 5 year survival is a binary outcome (Yes or No)
  - 'Success' is defined as survival for at least 5 years
  - n = number of 'trials' = 10 subjects
  - $\pi =$ probability of success = 0.80, assumed to be constant for all subjects
  - X is the number of patients surviving at least 5 years. X can range from 0 - 10

$$X \sim B\ (10, 0.8)$$

Binomial Distribution for X~B(10, 0.8)

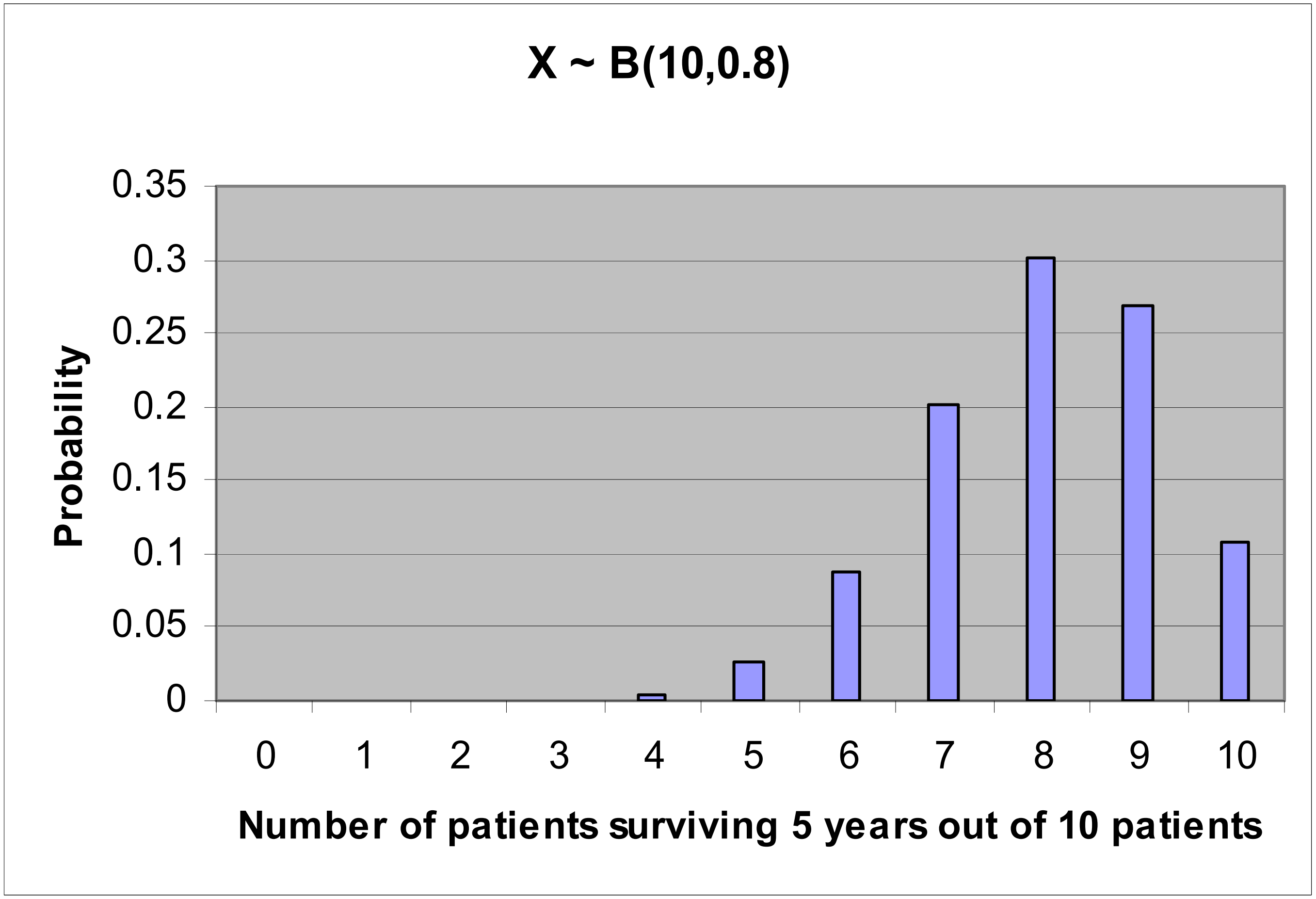
Number of patients surviving 5 years	Probability
0	< 0.0001
1	<0.0001
2	0.0001
3	0.0008
4	0.0055
5	0.0264
6	0.0881
7	0.2013
8	0.3020
9	0.2684
10	0.1074

The binomial distribution formula was used to calculate the probabilities that X of the 10 men would survive at least 5 years.

Notice that this distribution is skewed towards the larger values. This is because the P(success) is > 0.5.

Use this distribution to answer question 3 on the Practice Exercises

Plotting X~B(10,0.8)



Another clinical example

- In a sample of 8 patients with a heart attack, what is the probability that ‘X’ will die if the probability of death from a heart attack = 0.03.

  Assume that the probability of death is the same for all patients.

  - Death from heart attack is a binary variable (Yes or No)
  - ‘Success’ in this case is defined as death from heart attack
  - n = number of ‘trials’ = 8 patients
  - $\pi = 0.03$ = probability of success
  - X = number of deaths. X ranges from 0 – 8

$X \sim B\ (8, 0.03)$

In this example the outcome identified as ‘success’ is not a healthy outcome. ‘Success’ for a binomial distribution is the outcome of **interest** which is not always the clinically desirable outcome.

Binomial Distribution for X~B(8, 0.03)

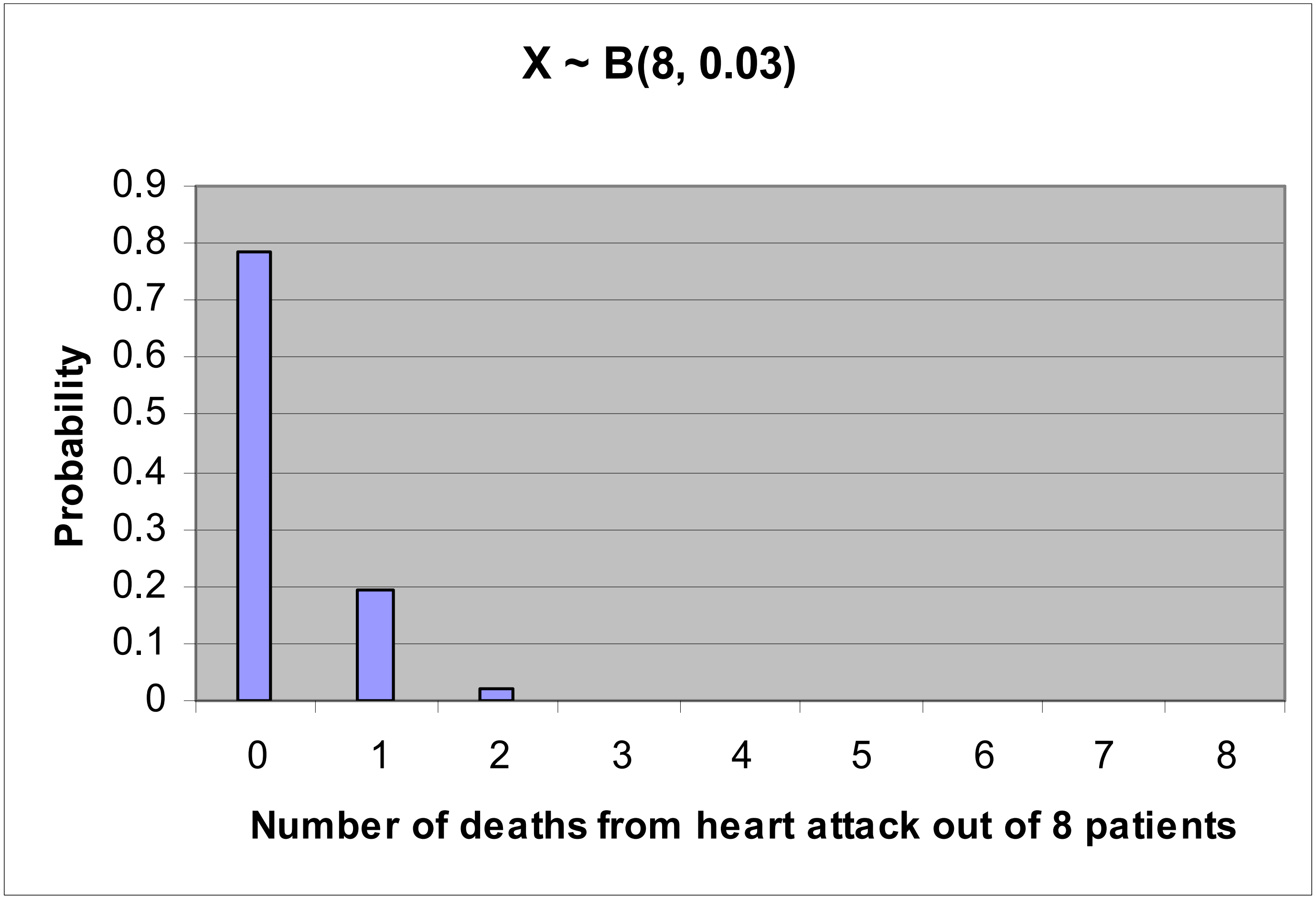
Number of deaths from heart attack	Probability
0	0.784
1	0.194
2	0.021
3	0.0013
4	< 0.001
5	<0.001
6	<0.001
7	<0.001
8	<0.001

The binomial distribution formula was used to calculate the probabilities that X of the 8 patents with a heart attack die

Notice that this distribution is skewed towards the smaller values. This is because the P(death) is very small.

Use this distribution to answer questions on the Practice Exercises.

Plotting X ~ B(8, 0.03)



Online Binomial Calculator

The binomial distribution formula involves time-consuming calculations for larger values of 'n'.

The following websites will calculate the binomial probabilities if you provide n, π, and X

- www.stat.tamu.edu/~west/applets/**binomial**demo.html
- http://www.swogstat.org/stat/public/binomial_calculator.htm
- http://faculty.vassar.edu/lowry/binomialX.html
- http://stattrek.com/Tables/Binomial.aspx

BINOMDIST function in Excel

- The BINOMDIST function in Excel will also calculate binomial probabilities
- BINOMDIST(X,n,π,TRUE) will return the probability of X or fewer events in n trials
- BINOMDIST(X,n, π, FALSE) will return the probability of exactly X events in n trials

Excel Module 6 includes examples of the BINOMDIST function

The risk of MI for 11,037 physicians taking daily aspirin = 0.0126. Is the number of physicians in this group who have MI a binomial distribution?

1. Yes
2. No

The risk of MI for 11,037 physicians taking daily aspirin = 0.0126. What is the probability that a physician taking aspirin will have an MI?

1. 11,037* 0.0126
2. 0.0126
3. 0.0126/11,037

Binomial Distributions

- There are many binomial distributions – each of these are determined by two parameters
  - probability (π) of the outcome of interest
  - number of trials (n).
- Look at this website to see the effect of n and π on the shape of the binomial distribution

  http://www.stat.tamu.edu/~west/applets/binomialdemo1.html
  - You can select the number of trials (n) and the probability of an event (the applet uses 'p' instead of π for the probability) to see how the shape of the binomial distribution changes depending on the number of trials and the probability of the event.

Binomial Demonstration

- Try the following combinations on the website to see how the number of trials (n) and the probability of the event (π) affect the shape of the probability distribution:

n	p	n	p
10	0.1	25	0.1
10	0.9	25	0.9
10	0.5	25	0.5
		60	0.1

Normal Approximation of the Binomial Distribution

- The binomial distribution applet illustrates the following:
  - as the number of trials increases and as the probability of an event is closer to 0.5, the binomial distribution approaches the shape of a normal distribution.
  - The largest probability for each distribution is at $n^*\pi$
- When $n*\pi > 5$ and $n^*(1-\pi) > 5$, the binomial distribution can be approximated by a normal distribution.
  - Most often in health research, the normal approximation to the binomial distribution is used because the sample sizes are large enough. If not, the exact binomial formula can be used

Mean and standard deviation of binomial distribution

- The mean of the binomial distribution = n*π
  - The mean is the ‘expected’ value of X
- The variance of the binomial distribution = n π(1-π)
- The standard deviation of the binomial distribution = $\sqrt{n\pi(1-\pi)}$
- The normal approximation to the binomial has mean = n*π and standard deviation = $\sqrt{n\pi(1-\pi)}$

Binomial Distribution Example

- Osteopenia is a decrease in bone mineral density that can be a precursor condition to osteoporosis
- The risk of osteoporosis for those with osteopenia varies depending on the age of the woman.
- Suppose you have a group of 100 women with osteopenia. The risk of having osteoporosis diagnosed in the following year = 0.20 for these women.
- This is an example of a binomial distribution
  - n = 100
  - ‘Success’ = osteoporosis diagnosed in following year
  - P(success) = risk = 0.20

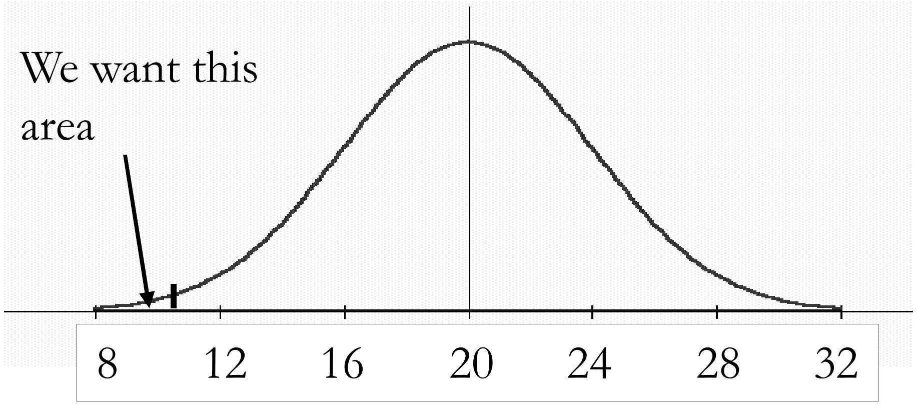
Normal Approximation for X~B(100, 0.20)

- X is distributed binomial with n=100 and π = 0.20
- Is the Normal approximation to this binomial distribution appropriate?
  - 100*0.20 = 20, 100*(1-0.2) = 80
  - 20 and 80 are both > 5 so the normal approximation to binomial is appropriate

- What are the mean and SD of this normal distribution?
  - Mean = n*π = 100 *0.20 = 20
  - Var = n*π (1- π) = 100*0.20*(1-0.20) = 16
  - SD = 4

Normal Approximation to B(100,0.20)

- Use the normal approximation to the binomial to find the probability that 10 or fewer women in this group of 100 are diagnosed with osteoporosis in the next year:



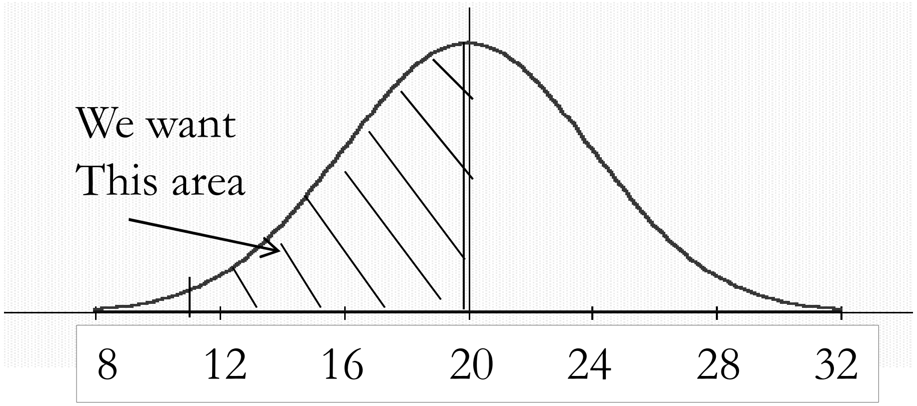
Use the normal distribution with Mean = 20 and SD = 4

- Use Excel to find the probability

=NORMDIST(10, 20, 4, 1) = 0.006

Normal approximation to B(100, 0.20)

- What is the probability that between 10 and 20 women in this group of 100 are diagnosed with osteoporosis in the next year?

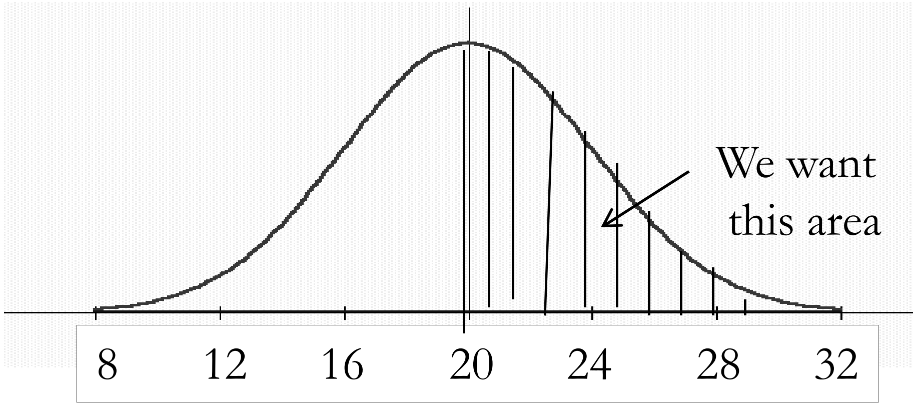


=NORMDIST (20, 20, 4, 1) – NORMDIST (10, 20, 4, 1)

= 0.494

Normal Approximation to B(100, 0.20)

- What is the probability that 20 or more women will be diagnosed with osteoporosis in the next year?



We want this area

= 1 – NORMDIST(20, 20, 4, 1) = 0.50

Practice Exercise Q. 5 works through the normal Approximation to the binomial distribution

The prevalence of overweight among U.S. adolescents = 17%. What is the expected number of overweight adolescents in a school with 500 HS students?

1. 500/17 = 29
2. 500*0.17 = 85
3. 100*0.17 = 17
4. 17/500 x 100 = 34

What is the mean of the normal approximation to the binomial for number of overweight adolescents in a High school of 500?

1. 17
2. 34
3. 85
4. 100

What is the probability that fewer than 34 adolescents in a high school of 200 are overweight?

1. 0.34
2. 0.85
3. 0.50
4. 0.17

In calculating the probability for number of overweight adolescents in a particular school with n enrolled we need to assume

1. Equal probability of being overweight for all students
2. US prevalence of overweight is the same in all schools
3. Whether or not a student is overweight is not influenced by other students’ weight
4. All of the above

Poisson Distribution

- Another probability distribution for discrete variables is the Poisson distribution
- Named for Simeon D. Poisson, 1781 – 1840, French mathematician

- The Poisson distribution is used to determine the probability of the number of events occurring over a specified time or space.
- Examples of events over space or time:

    -number of cells in a specified volume of fluid

    -number of calls/hour to a help line

Poisson Distribution

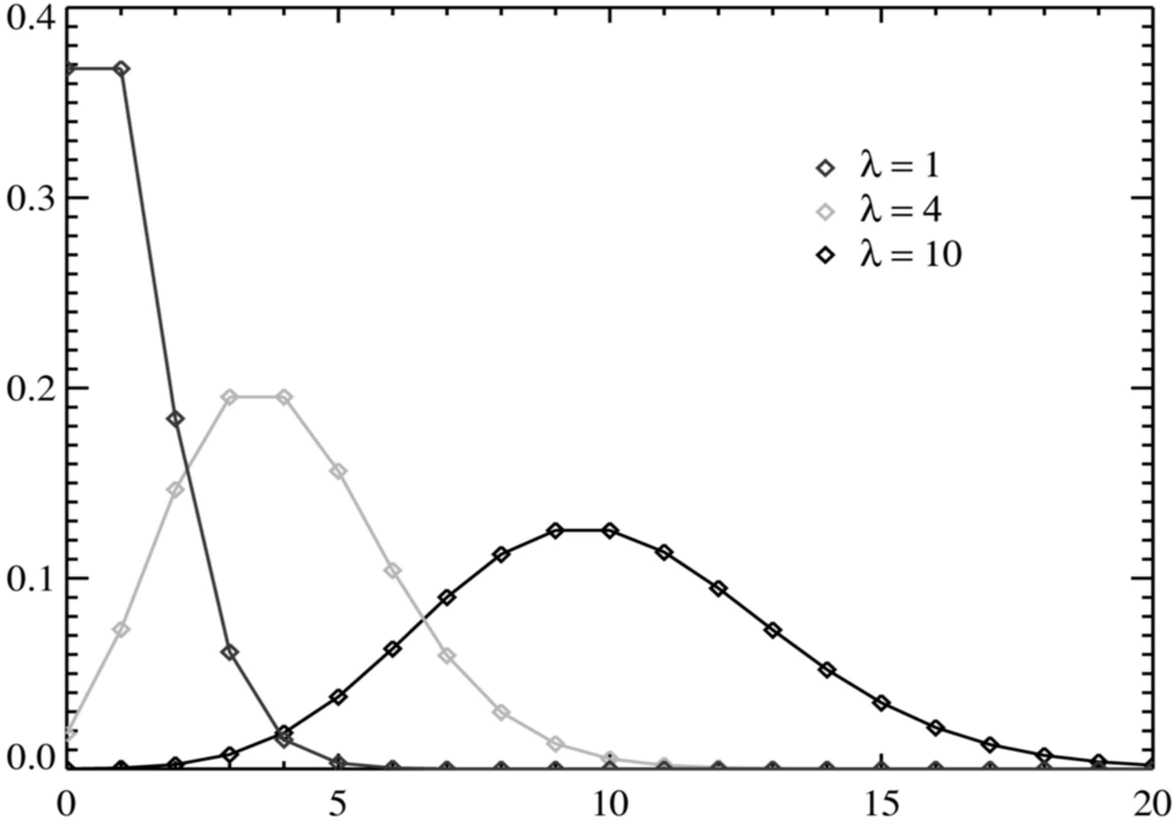
- Like the binomial distribution and the normal distribution, there are many Poisson distributions.
- Each Poisson distribution is specified by the average rate at which the event occurs. The rate is notated with λ
  - $\lambda =$ ‘lambda’, Greek letter ‘L’
  - There is only one parameter for the Poisson distribution

Poisson Distribution Formula

- The probability that there are exactly *X* occurrences in the specified space or time is equal to

$$\boxed{P(X)=\frac{\lambda^{x}e^{-\lambda}}{X!}}$$

What Does it Look Like?



Notice that as λ increases the distribution begins to resemble a normal distribution

- ***The horizontal axis is the index X. The function is defined only at integer values of X. The connecting lines are only guides for the eye and do not indicate continuity.***

- ***Source: http://en.wikipedia.org/wiki/Image:Poisson_distribution_PMF.png***

Normal Approximation of Poisson Distribution

- If λ is 10 or greater, the normal distribution is a reasonable approximation to the Poisson distribution
- The mean and variance for a Poisson distribution are the same and are both equal to λ
- The standard deviation of the Poisson distribution is the square root of λ

Poisson Distribution Example

- A large urban hospital has, on average, 80 emergency department admits every Monday
- What is the probability that there will be more than 100 emergency room admits on a Monday?
  - λ is the rate of admits / day on Monday = 80
  - Use the normal approximation since $\lambda > 10$
- The normal approximation has mean = 80 and

  SD = 8.94 (the square root of 80 = 8.94)
- 1 - NORMDIST (100, 80, 8.94, 1) = 0.0126
- The probability that there are > 100 admits to the ER on a Monday = 0.0126.
  - This information could be used in determining staffing needs for the emergency department.

Practice Exercise Q. 6

- An example of the normal approximation of the Poisson distribution in provided on Practice Exercise Question 6.
- There is also a POISSON function in Excel which will calculate an exact probability for the Poisson distribution

POISSON(X, λ, cumulative)

‘True’ for cumulative returns the probability of X or fewer events

‘False’ for cumulative returns the exact probability of X events

Overview of Probability Distributions for Discrete Data

- For a limited number of discrete outcomes the probability distribution can be presented in a table or chart with the probability for each outcome identified
- The Binomial distribution can be approximated by a normal distribution for large enough number of trials and probability
  - n* π > 5 and n*(1 – π) > 5
- The Poisson distribution can be approximated by a normal distribution for λ > 10.

Readings and Assignments

- Reading: Chapter 4 pgs. 72 – 76
- Lesson 6 Part 2 practice exercises
- Work through Excel Module 6 examples
- Complete Homework 4 by due date