

# **STA 291**

# **Summer 2008**

## **Lecture 6**

# Review: Conditional Probabilities

- $P(A \cap B) = P(A, B)$  Joint probability of  $A$  and  $B$   
(of the intersection of  $A$  and  $B$ )
- $P(A|B)$  Conditional probability of  $A$  given  $B$   
“the probability that  $A$  occurs given that  $B$  has occurred.”
- $P(A)$  Marginal probability of  $A$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ provided } P(B) \neq 0$$

# Example: Smoking and Lung Disease I

<b><i>Joint Probabilities</i></b>	Lung Disease	Not Lung Disease	<i>Row Totals</i>
Smoker	.12	.19	.31
Nonsmoker	.03	.66	.69
<i>Column Totals</i>	.15	.85	1.00

# Example: Smoking and Lung Disease II

<b><i>Conditional Row Probabilities</i></b>	Lung Disease	Not Lung Disease	<i>Row Totals</i>
Smoker	.12/.31 =.39	.19/.31 =.61	.31/.31 =1.00
Nonsmoker	.03/.69 =.04	.66/.69 =.96	.69/.69 =1.00
<i>Smokers and Nonsmokers</i>	.15	.85	1.00

# Example: Smoking and Lung Disease III

<b><i>Conditional Column Probabilities</i></b>	Lung Disease	Not Lung Disease	<i>Lung Disease and Not Lung Disease</i>
Smoker	.12/.15 =.80	.19/.85 =.22	.31
Nonsmoker	.03/.15 =.20	.66/.85 =.78	.69
<i>Column Totals</i>	.15/.15 =1.00	.85/.85 =1.00	1.00

# Review: Probability Distributions

- The probability of an event/outcome is the proportion of times that outcome would occur in a (hypothetical) infinitely long run of repeated observations
- It is a number between 0 and 1 (or a percentage between 0 and 100)
- If you select a subject randomly from a population to measure a variable  $Y$ ,  
then the probability distribution for the value of the random variable  $Y$  is the population (relative frequency) distribution of that variable

# Population Distribution: Example

- Population: 228 students in STA 291
- 10% had a score of 93 or higher in the first midterm
- Assume you were to select a STA 291 student randomly and ask whether he/she had 93 or higher in the first exam
- The result is a random variable (because you get different results when you ask different students)
- The probability of getting the answer “yes” is 10%

# Population Distribution vs. Probability Distribution

- If you select a subject randomly from the population,
  - then the **probability distribution** for the value of the random variable  $Y$
  - is the **population distribution** of that variable
  - and the population distribution is just the relative **frequency distribution of the whole population**

# 7. Random Variables

- A variable  $X$  is a **random variable** if the value that  $X$  assumes at the conclusion of an experiment cannot be predicted with certainty in advance.
- There are two types of random variables:
  - **Discrete:** the random variable can only assume a finite or countably infinite number of different values
  - **Continuous:** the random variable can assume all the values in some interval

# *Examples*

Which of the following random variables are discrete and which are continuous?

- a.  $X$  = Number of houses sold by real estate developer per week?
- b.  $X$  = Number of heads in ten tosses of a coin?
- c.  $X$  = Weight of a child at birth?
- d.  $X$  = Time required to run a marathon?

# *Properties of Discrete Probability Distributions*

**Definition:** A Discrete probability distribution is just a list of the possible values of a r.v.  $X$ , say  $(x_i)$  and the probability associated with each  $P(X=x_i)$ .

**Properties:**

1. All probabilities non-negative.
2. Probabilities sum to \_\_\_\_\_ .

$$0 \leq P(x_i) \leq 1$$

$$\sum P(x_i) = 1$$

# Example

The table below gives the # of days of sick leave for 200 employees in a year.

Days	0	1	2	3	4	5	6	7
Number of Employees	20	40	40	30	20	10	10	30

An employee is to be selected at random and let  $X =$  # days of sick leave.

- Construct and graph the probability distribution of  $X$
- Find  $P(X \leq 3)$ .
- Find  $P(X \geq 3)$ .
- Find  $P(3 \leq X \leq 6)$ .

# Population Distribution vs. Probability Distribution

- If you select a subject randomly from the population, then the probability distribution for the value of the random variable  $X$  is the population distribution of that variable
- Example:  
 $X$ =number of sick days/height/grade of a randomly chosen person

# *Cumulative Distribution Function*

**Definition:** The *cumulative distribution function*, or **CDF** is  $F(x) = P(X \leq x)$ .

**Motivation:** Some parts of the previous example would have been easier with this next tool:

**Properties:**

1. For any value  $x$ ,  $0 \leq F(x) \leq 1$ .
2. If  $x_1 < x_2$ , then  $F(x_1) \leq F(x_2)$
3.  $F(-\infty) = 0$  and  $F(\infty) = 1$ .

# Example

Let  $X$  have the following probability distribution

<b>x</b>	<b>2</b>	<b>4</b>	<b>6</b>	<b>8</b>	<b>10</b>
<b>P(x)</b>	.05	.20	.35	.30	.10

- Find  $P(X \leq 6)$ .
- Graph the cumulative probability distribution function.
- Find  $P(X > 6)$ .

# Expected Value and Variance of a Random Variable

The Expected Value, or mean, of a random variable,  $X$ , is

$$\text{Mean} = E(X) = \mathbf{m} = \sum x_i P(X = x_i)$$

- Variance =  $Var(X) = E(X - \mathbf{m})^2 = \mathbf{s}^2$   
$$= \sum (x_i - \mathbf{m})^2 P(X = x_i)$$
  
$$= \sum x_i^2 P(X = x_i) - \mathbf{m}^2$$

## Example

<b>x</b>	<b>2</b>	<b>4</b>	<b>6</b>	<b>8</b>	<b>10</b>
<b>P(x)</b>	.05	.20	.35	.30	.10

**What is  $E(X)$ ? What is  $Var(X)$ ?**

# *Expected Winnings Example*

- In roulette there are 18 red, 18 black, and 2 green pockets.
- If the pockets are all equally likely and a bet on black pays 2-to-1, how much would you expect to win on a \$100 bet?

# Standard Deviation

- Empirical rule and Chebysheff's inequality apply to populations, samples, and probability distribution
- For an approximately bell-shaped distribution, there is ~68% of the probability between  $\mu - s$  and  $\mu + s$
- For **every** distribution, there is at least 75% probability between  $\mu - 2s$  and  $\mu + 2s$

# Bernoulli Distribution

- Suppose we have a single random experiment  $X$  with two outcomes:  
    “success” and “failure.”
- Typically, we denote “success” by the value 1 and “failure” by the value 0.
- It is also customary to label the corresponding probabilities as:  
     $P(\text{success}) = P(1) = p$  and  
     $P(\text{failure}) = P(0) = 1-p=q$
- Note:  $p+q = 1$
- What are expected value and variance of a Bernoulli random variable?

# Binomial Distribution I

- Suppose that instead of just one Bernoulli experiment, we perform several and they are all independent of each other.
- Let's say we do  $n$  of them. The value  $n$  is the **number of trials**.
- We will label these  $n$  Bernoulli random variables in this manner:  $X_1, X_2, \dots, X_n$
- As before, we will assume that the probability of success in a single trial is  $p$ , and that this probability of success doesn't change from trial to trial.

# Binomial Distribution II

- Now, we will build a new random variable  $X$  using all of these Bernoulli random variables:

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^n X_i$$

- What are the possible outcomes of  $X$ ?
- What is  $X$  counting?
- How can we find  $P(X=k)$ ?

# Binomial Distribution III

- We need a quick way to count the number of ways in which  $k$  successes can occur in  $n$  trials.
- Here's the formula to find this value:

$$\binom{n}{k} = {}_n C_k = \frac{n!}{k!(n-k)!}$$

where  $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$  and  $0! = 1$ .

- Note:  ${}_n C_k$  is read as "n choose k."

# Binomial Distribution IV

- Now, we can write the formula for the binomial distribution:
- The probability of observing  $k$  successes in  $n$  independent trials is

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n,$$

- under the assumption that the probability of success in a single trial is  $p$ .

# Using Binomial Probabilities

**Note:** Unlike generic random variables where we would have to be given the probability distribution or calculate it from a frequency distribution, here we can calculate it from a mathematical formula.

Helpful resource (besides your calculator):

- Excel:

=BINOMDIST(4,10,0.2,FALSE)    0.08808

=BINOMDIST(4,10,0.2,TRUE)    0.967207

# Mean, Variance, and Standard Deviation of a Binomial Distribution

$$\mathbf{m} = n \cdot p$$

$$\mathbf{s}^2 = n \cdot p \cdot (1 - p)$$

$$\mathbf{s} = \sqrt{n \cdot p \cdot (1 - p)}$$

Binomial Simulation Tool

# Continuous Probability Distributions

- For continuous distributions, we can not list all possible values with probabilities
- Instead, probabilities are assigned to intervals of numbers
- The probability of an individual number is 0
- Again, the probabilities have to be between 0 and 1
- The probability of the interval containing all possible values equals 1
- Mathematically, a continuous probability distribution corresponds to a (density) function whose integral equals 1

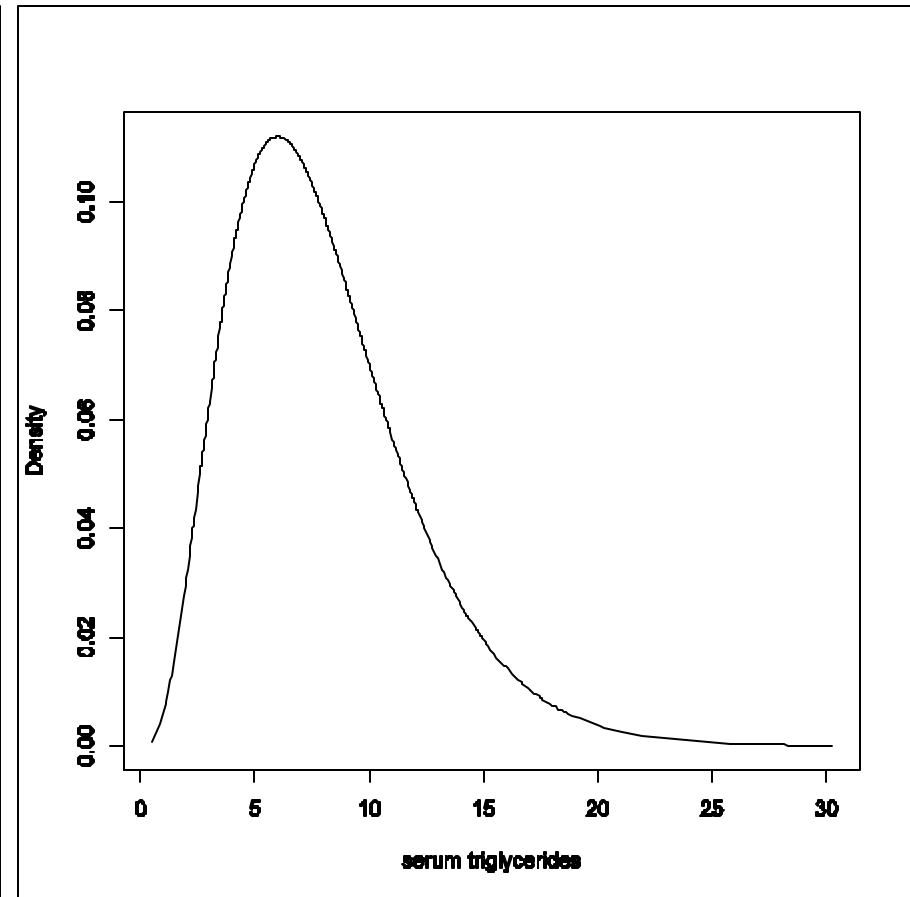
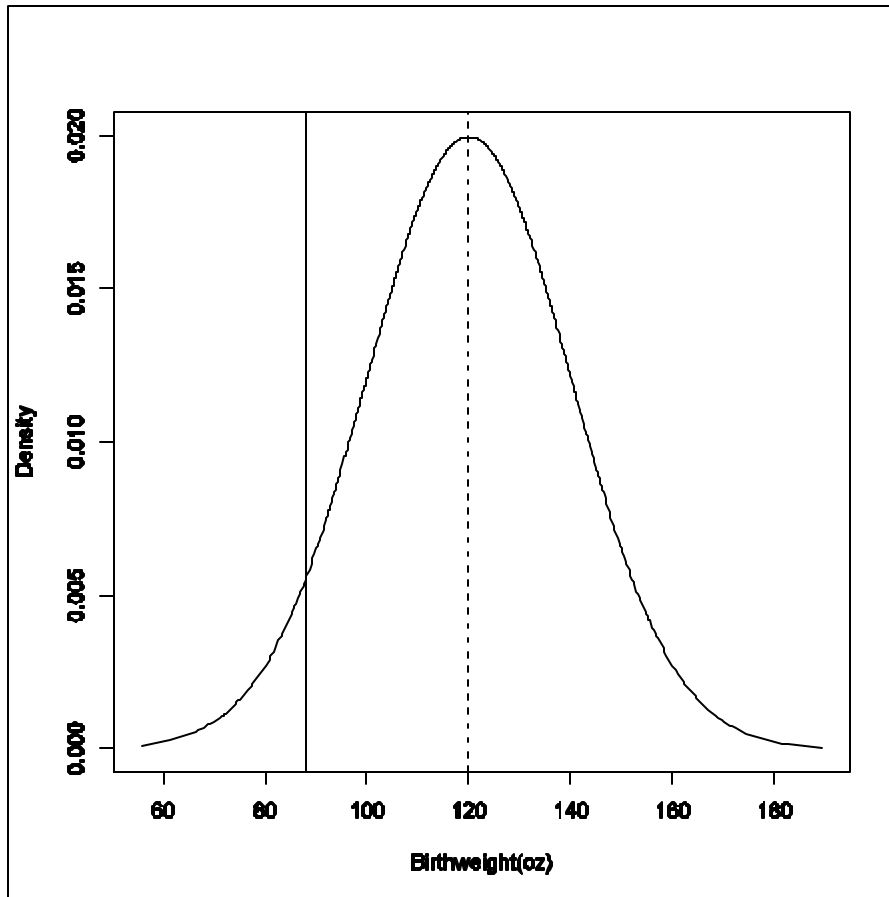
# Continuous Probability Distributions: Example

- Example:  $X$ =Weekly use of gasoline by adults in North America (in gallons)
- $P(16 < X < 21) = 0.34$
- The probability that a randomly chosen adult in North America uses between 16 and 21 gallons of gas per week is 0.34

# Graphs for Probability Distributions

- Discrete Variables:
  - Histogram
  - Height of the bar represents the probability
- Continuous Variables:
  - Smooth, continuous curve
  - Area under the curve for an interval represents the probability of that interval

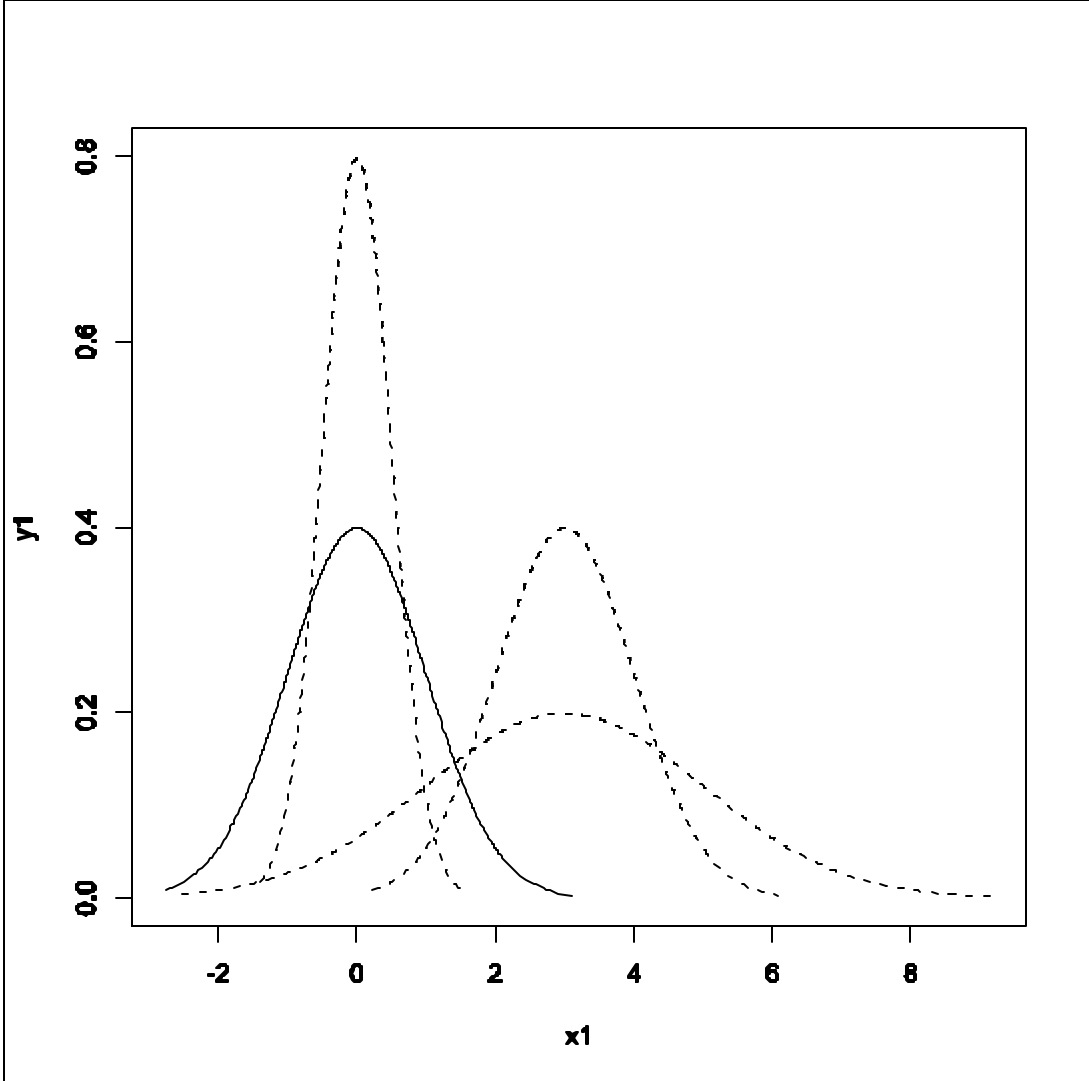
# Continuous Distributions



# The Normal Probability Distribution

- Carl Friedrich Gauß (1777-1855), ***Gaussian Distribution***
- Normal distribution is perfectly ***symmetric*** and ***bell-shaped***
- Characterized by two parameters:  
***mean  $\mu$***  and ***standard deviation  $s$***
- The ***68%-95%-99.7%*** rule applies to the normal distribution
- That is, the probability concentrated within 1 standard deviation of the mean is always 0.68
- The ***IQR = 4/3 s*** rule also applies

# Different Normal Distributions



# Normal Distribution: Example (female height)

- Assume that adult female height has a normal distribution with mean  $\mu=165$  cm and standard deviation  $s=9$  cm
- With probability 0.68, a randomly selected adult female has height between  
$$\mu - s = 156 \text{ cm and } \mu + s = 174 \text{ cm}$$
- With probability 0.95, a randomly selected adult female has height between  
$$\mu - 2s = 147 \text{ cm and } \mu + 2s = 183 \text{ cm}$$
- Only with probability  $1-0.997=0.003$ , a randomly selected adult female has height below  
$$\mu - 3s = 138 \text{ cm or above } \mu + 3s = 192 \text{ cm}$$

# Normal Distribution

- So far, we have looked at the probabilities within one, two, or three standard deviations from the mean

$$(\mu + s, \mu + 2s, \mu + 3s)$$

- How much probability is concentrated within 1.43 standard deviations of the mean?
- More general, how much probability is concentrated within  $z$  standard deviations of the mean?

# Normal Distribution Table

- Table 3 (page B-8) shows for different values of  $z$  the probability between 0 and  $\mu + z\sigma$
- Probability that a normal random variable takes any value between the mean and  $z$  standard deviations above the mean
- For  $z=1.43$ , the tabulated value is .4236
- That is, the probability **between 0 and  $\mu + z\sigma$**  of a normal distribution equals .4236
- Because of the perfect symmetry of the normal distribution, also the probability **between 0 and  $\mu - z\sigma$**  of a normal distribution equals .4236
- So, within 1.43 standard deviations of the mean is how much probability?

# Verifying the Empirical Rule

- The 68%-95%-99.7% rule can be verified using Table 3
- How much probability is *within* one (two, three) standard deviation(s) of the mean?
- Note that the table only answers directly: How much probability is between 0 and one (two, three) standard deviation(s) above the mean?

# Verifying the Empirical Rule ( $z=1$ )

- $z=1.00$ , Table 3:

Between 0 and 1 standard deviations above the mean is probability .3413

- Then, between 0 and 1 standard deviations ***below*** the mean is also probability .3413 (symmetry)

- Therefore, ***within*** one standard deviation from the mean is probability  $.3413 + .3413 = .6828$

# Verifying the Empirical Rule ( $z=2,3$ )

- $z=2.00$ , Table 3:

Between 0 and 2 standard deviations above the mean is probability \_\_\_\_\_

- Then, between 0 and 2 standard deviations **below** the mean is also probability \_\_\_\_\_ (symmetry)
- Therefore, **within** two standard deviations from the mean is probability \_\_\_\_\_
- $z=3.00$ : Table value: \_\_\_\_\_
- Probability within three standard deviations from the mean: \_\_\_\_\_

# Working backwards

- We can also use the table to find z-values for given probabilities
- Find the z-value corresponding to a right-hand tail probability of 0.025
- This corresponds to a probability of 0.475 between 0 and z standard deviations
- Table:  $z = 1.96$
- Probability 0.025 lies above  $\mu + 1.96 s$

# More Examples (Reverse Procedure)

- Verify that the z-value for a right-hand tail probability
  - of 0.1 is  $z=1.28$
  - of 0.05 is  $z=1.65$
  - of 0.01 is  $z=2.33$

# Finding z-Values for Percentiles

- For a normal distribution, how many standard deviations from the mean is the 90<sup>th</sup> percentile?
- Or: What is the value of  $z$  such that 0.90 probability is less than  $\mu + z s$  ?
- If 0.9 probability is less than  $\mu + z s$ , then there is 0.4 probability between 0 and  $\mu + z s$  (because there is 0.5 probability less than 0)
- $z=1.28$
- The 90<sup>th</sup> percentile of a normal distribution is 1.28 standard deviations above the mean

# Application

- SAT scores are approximately normally distributed with mean 500 and standard deviation 100
- The 90<sup>th</sup> percentile of the SAT scores is 1.28 standard deviations above the mean
- $\mu + 1.28 s = 500 + 1.28 \cdot 100 = 628$
- Find the 99<sup>th</sup> and the 5<sup>th</sup> percentile of SAT scores

# Finding z-Values for Two-Tail Probabilities

- What is the z-value such that the probability is 0.1 that a normally distributed random variable falls more than z standard deviations **above or below** the mean
- Symmetry: we need to find the z-value such that the right-tail probability is 0.05 (more than z standard deviations **above** the mean)
- $z=1.65$
- 10% probability for a normally distributed random variable is outside 1.65 standard deviations from the mean, and 90% is within 1.65 standard deviations from the mean
- Find the z-value such that the probability is 0.5 that a normally distributed random variable falls more than z standard deviations **above or below** the mean

# Online Tools

- Normal Density Curve
- Standard Normal Calculator "Surfstat"
- Standard Normal Calculator "Stat Trek"
- Use these to
  - verify graphically the empirical rule,
  - find probabilities,
  - find percentiles
  - calculate z-values for one- and two-tailed probabilities

# Example

- Suppose that the weekly use of gasoline for motor travel by adults in North America is approximately normally distributed with mean 16 gallons and standard deviation 5 gallons.
- What proportion of adults use more than 20 gallons per week?