



# ***Module 1: Review of Basic Statistical Concepts***

## **1.1 Understanding Probability Distributions, Parameters and Statistics**



### **Distributions of Data**

- ♦ A variable that can take on any value in a range is called a continuous variable.
  - Example: The concentration of a contaminant in water samples
- ♦ A variable that can take on only certain values is called discrete.
  - Example: The number of animals visiting a contaminated site in a single day



## Distributions of Data

- ◆ A probability distribution describes the values that a variable can take on and the probabilities associated with those values.
- ◆ We use probability density functions (pdfs) to describe these distributions. We can also use cumulative density functions (cdfs)
- ◆ For continuous variables, common distributions are the uniform, the triangular, the normal, and the lognormal

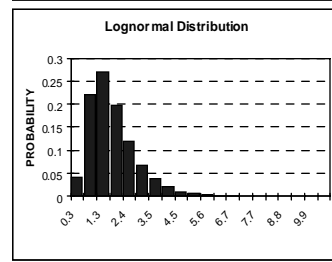
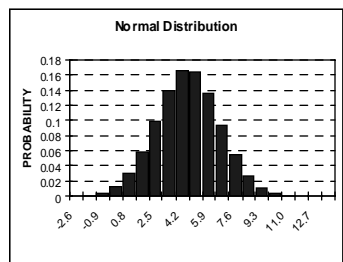
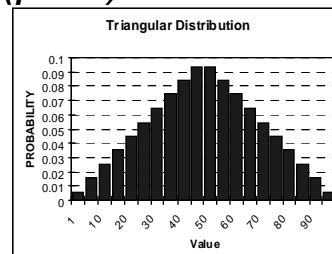
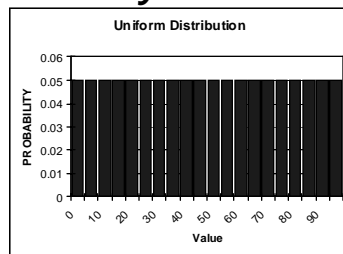
4/18/2002

Module

3



## Examples of Continuous Probability Density Functions (pdfs)



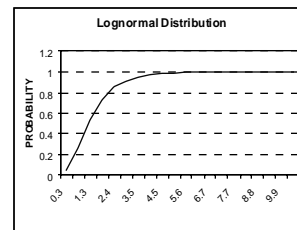
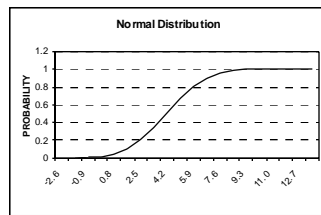
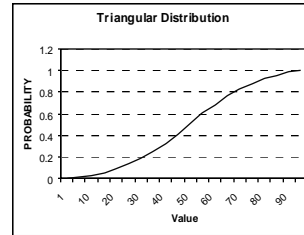
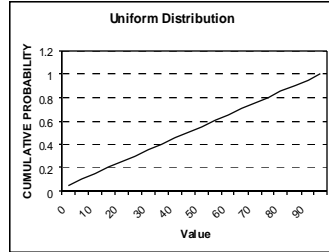
4/18/2002

Module

4



## Examples of Continuous Cumulative Density Functions (cdfs)



4/18/2002

Module

5



## Distributions of Data

- ◆ Values that define key characteristics of probability distributions are called parameters.
- ◆ The uniform distribution means that there is a range of values defined by the parameters minimum and maximum. All of the values in between have an equal probability of occurring.
- ◆ The triangular has a minimum, maximum, and most likely value

4/18/2002

Module

6



## Distributions of Data

- ◆ The normal is the bell shaped curve , its parameters are the mean and standard deviation.
- ◆ The lognormal has smaller values having a higher probability of occurring and larger values having a smaller and smaller probability of occurring.

4/18/2002

Module

7



## Distributions of Data

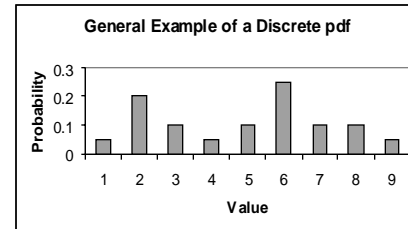
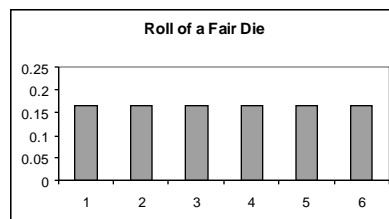
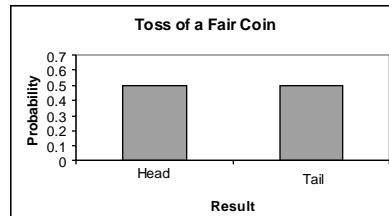
- ◆ Discrete probability distributions simply show the probability of each value occurring.
- ◆ Note that the sum of these probabilities is one. For the continuous pdfs, the area under the curve equals one.

4/18/2002

Module

8

## Examples of Discrete Probability Distribution Functions (pdfs)



4/18/2002

Module

9

## The Normal Distribution

- ◆ The normal distribution is the bell-shaped curve.
- ◆ Many things that occur in nature follow a normal distribution.
- ◆ Some characteristics:
  - It has two parameters: the mean  $\mu$  ( $\mu$ ) and the standard deviation  $\sigma$  ( $\sigma$ )

4/18/2002

Module

10



## The Normal Distribution

- ◆ Some characteristics:
  - The standard normal has  $\mu=0$  and  $\sigma=1$
  - Any normal distribution can be transformed into a standard normal by  $Z=(X-\mu)/\sigma$
  - About 68% of the probability of a normal lies within plus and minus one standard deviation from the mean
  - About 95% lies within plus and minus 2 standard deviations from the mean
  - 99.7% lies within plus and minus 3 standard deviations from the mean

4/18/2002

Module

11



## ***Using the Table of the Normal Distribution***

- ◆ Tables such as Table B1 in Manly relate values of  $Z$  to the probability (area) under the normal pdf from zero to that value.
  - Example: The probability of a value sampled randomly from the standard normal distribution falling between the mean and one standard deviation above it is found by looking up the probability in the table associated with 1.00. It is 0.341. So, the probability of a value falling within 1 standard deviation from the mean is double that or 0.682. Likewise, the probability of a value falling within plus and minus 2 standard deviations is  $2 * 0.477=0.954$ .

4/18/2002

Module

12



## The Binomial Distribution

- Applies in a situation where there are two possible outcomes (success and failure) and the probability of success is constant.
- Example: Failure is defined to be contamination above a regulatory limit. Assume contamination is uniformly dispersed throughout an area such as a lake and  $n$  samples are collected. There will be variability in the amount of measured contamination in the samples due to sampling and measurement errors. There is a probability  $p$  that each of the samples will show contamination above the limit.

4/18/2002

Module

13



## *The Student's t Distribution*

- The Student's  $t$  distribution is similar to the normal but with fatter tails. It is used when the true population standard deviation is not known (most of the time).
- The exact shape of the  $t$  distribution is controlled by the number of data points used to calculate the sample standard deviation. When  $n$  is small, the distribution is wide. When  $n$  gets large, the estimate of  $\sigma$  is good and the  $t$  distribution approaches the shape of a normal distribution.
- The term for this index is called degrees of freedom (df). For use with the  $t$  distribution,  $df = n - 1$ .

4/18/2002

Module

14



## Using the table of the t distribution

- ◆ Table B2 of Manly gives some selected values from t distributions with degrees of freedom ranging from 1 to infinity.
- ◆ Example: If you have 10 data points, you have 9 degrees of freedom. If you want the value along the t scale that has 95% of the probability below it and 5% above, use the first column. That t value is 1.833.

4/18/2002

Module

15



## *Parameters and Statistics*

- ◆ A parameter is a characteristic of a population. It is a value that we would only know if we had perfect information about the entire population. Since we never have this kind of knowledge, parameters can be considered unknown. They are the quantities that we try and estimate from our data.
- ◆ Statistics are quantities calculated from data. For each parameter, there is one or more statistics that estimate it.

4/18/2002

Module

16



## Parameters and Statistics

- Example: The population mean is a parameter denoted by  $\mu$ , the sample mean estimates  $\mu$  and is denoted by  $\bar{X}$  with a bar over it called X bar.
  - Notation:  $N$  = number of units in the population
  - $n$  = number of units in the sample

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

4/18/2002

Module

17



## Parameters and Statistics

- The population standard deviation is a parameter denoted by  $\sigma$  and the sample standard deviation estimates it and is denoted by  $s$ .

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2}$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

4/18/2002

Module

18



## *Parameters and Statistics*

- ◆ Once we have data and an equation to calculate a statistic, it's simple arithmetic to get the estimate. However, the estimate is just that – it's not the actual value of the parameter. The true value of the parameter might be higher or lower than our estimate.
- ◆ If the population was defined by the students registered for this class today, there is a true mean height of that population
- ◆ However, even if I tried to collect data on this population, I couldn't know the true mean. Why?

4/18/2002

Module

19



## *Parameters and Statistics*

- ◆ However, I can collect a sample and calculate a sample mean that would estimate the true mean.
- ◆ I could also calculate a sample standard deviation and create a confidence interval around the true mean. The confidence interval would be a range with a probability attached. It has that probability of including the true mean.

4/18/2002

Module

20