# *Module 3: Models for Data*

3.1 Discrete and Continuous Probability Distributions

## *Probability Distributions*

- Often it is important to know something about the underlying probability distribution that data come from.
- There is a branch of statistics that does not require any assumptions about distributions. These are nonparametric methods and will be discussed later.

## *Probability Distributions*

- Information about the underlying distribution can come from:
  - Knowledge of the data generating process
  - Past data of a similar type
  - Theoretical considerations
  - Plots of current data



## *Discrete Probability Distributions*

- Discrete probability distributions are used when the data can only take on specific values.
- Examples:
  - Contaminated or Non-contaminated(0 or 1)
  - Count data (0, 1, 2, 3, and so on)
  - Value of a rolled die (1, 2, …, 6)

# *Discrete Probability Distributions: Notation*

P(x) = Probability (X=x)

$P(x_1) + P(x_2) + P(x_3) + \dots P(x_n) = 1$

$P(x_i) >= 0$

Expected Value = E(X) = $\sum_{i=1}^{\infty} x_i * P(x_i)$

Variance = Var(X) = $\sum_{i=1}^{\infty} \left(x_i - \mu\right)^2 P(x_i)$

Module 3.1

# *Discrete Probability Distributions: Binomial*

Applies when there are n independent trials

Two outcomes possible: success and failure

p = probability of success is constant

Probability of x successes in n trials is

$$P(x) = \left({}^{n}C_{x}\right) p^{x} (1-p)^{n-x}$$

Expected Value = E(X) = np

Variance = Var(X) = np(1-p)

Module 3.1

## Note on Notation

$$\left({}^{n}C_{x}\right) = \frac{n!}{x!(n-x)!}$$

Module 3.1

## Discrete Probability Distributions: Poisson

- Applies when there are n independent events
- Events occur at a constant rate $\mu$
- Probability of x events in a given period of time or area of space is

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

- Expected Value = E(X) = $\mu$
- Variance = Var(X) = $\mu$

Module 3.1

## *Discrete Probability Distributions: Hypergeometric*

- Applies when a population of N units contains R successes
- Probability that a sample of size n contains x successes is

$$P(x) = \frac{\left({}^{R}C_{x}\right)\left({}^{N-R}C_{n-x}\right)}{\left({}^{N}C_{n}\right)}$$

- Expected Value = E(X) = nR/N
- Variance = Var(X) = R(N-R)n/$N^2$

Module 3.1

## *Continuous Probability Distributions: Normal*

- Bell Shaped Curve
- Occurs:
  - In nature (heights, weights, etc)
  - When the variable is the sum of other variables
  - Distribution of sample mean, total, proportion
- E(X) = $\mu$
- Var(X) = $\sigma^2$
- StDev(X) = $\sigma$

Module 3.1

## *Continuous Probability Distributions: Lognormal*

Right skewed = has long tail to the right

Occurs:

- When the variable is the product of other variables
- Often a good model for environmental contamination

$E(X) = \exp(\mu + \sigma^2/2)$

$Var(X) = \exp(2\mu + \sigma^2)\{\exp(\sigma^2)-1\}$



## *Continuous Probability Distributions: Exponential*

Applies when the time until an event occurs or between events is of interest

$E(X) = \mu$

$Var(X) = \mu^2$