

Fall 2001, 1:30 - 2:20, MWF, Harrelson 320.

# INTRO TO MATHEMATICAL STATISTICS I

**Montserrat Fuentes**

*Statistics Department NCSU*

fuentes@stat.ncsu.edu

<http://www.stat.ncsu.edu/~fuentes>

# Introduction

What is Statistics?

- Introduction
- Graphical methods
- Numerical methods
- Inference
- Theory and Reality

Population: The large body of data that is the target of our interest, e.g. collection of all GPA's of US students in 2000.

Sample: subset selected from a population, e.g. a collection of GPA's representing only two scores from each state of US. We are forced to look at samples primarily for the following reasons: *Economy, Timeliness, Large populations, Inaccessibility.*

Statistics; The objective of statistics is to make an inference about a population based on information contained in a sample from that population and to provide an associated measure of goodness for the inference.

Examples:

- sample of voters to estimate the true fraction of all voters who favor a particular candidate.
- A decision is made regarding the relative merits of two manufacturing processes based on examination of samples of products.

Graphical Methods. sort the data into groups defined by possible variable values. Then count how many data items were sorted into each category.

**Frequency:** The count for a particular category.

**Relative frequency:** The decimal value obtained by dividing the frequency by total number of data items.

The graph is constructed by subdividing the axis of measurements into intervals of equal width (histogram).

## NUMERICAL METHODS:

Measures of central tendency:

MEAN: The mean of a sample of  $n$  measured responses  $y_1, y_2, \dots, y_n$  is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

The corresponding population mean is denoted  $\mu$ .

MEDIAN: The value above and below an equal number of observation.

MODE: Most frequently occurring value.

Measures of dispersion:

**VARIANCE:** The variance of a sample is the sum of the square of the differences between the measurements and their mean, divided by  $n - 1$ .

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

The corresponding population variance is denoted by the symbol  $\sigma^2$ .

(the larger the variance the greater the amount of variation within the set of observations.)

**THE STANDARD DEVIATION:** It is the positive square root of the variance.

**RANGE:** Difference between the largest and the smallest values.

For data mound-shaped:

**EMPIRICAL RULE:**

For a distribution of measurements that is approximately normal (bell-shaped), it follows that the interval with endpoints

- $\nu \pm \sigma$  contains approximately 68% of the measurements.
- $\nu \pm 2\sigma$  contains approximately 95% of the measurements.
- $\nu \pm 3\sigma$  contains approximately 99.7% of the measurements.

## INFERENCE STATISTICS:

Studies of the properties of a small sample of data selected from a much larger body of data (*population*) and attempts, using mathematics (probability), to infer the nature of the same properties for the population.

Probability is the mechanism used in making statistical inferences.

Intuitive assessments of probabilities are unsatisfactory, and we need a rigorous theory of probability in order to develop methods of inference.

## THEORY AND REALITY

We will work with theoretical or mathematical models for acquiring and utilizing information in real life. The process of finding a good model is not simple and usually requires several simplifying assumptions (uniform string mass, no air resistance, etc.) The model will not be an *exact* representation of nature.