CSC 223 - Advanced Scientific Programming

Descriptive Statistics

Overview

- *Statistics* is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.
- *Data* consists of information coming from observations, counts, measurements, or responses.
- A *population* is the collection of all outcomes, responses, measurements, or counts that are of interest.
- A *sample* is a subset of the population.
- A *parameter* is a numerical description of a population characteristic.
- A *statistic* is a numerical description of a sample characteristic.

Branches of Statistics

- *Descriptive statistics* is the branch of statistics that involves the organization, summarization, and display of data.
- *Inferential statistics* is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

Data Classification

- Types of data:
  - *Qualitative data* consist of attributes, labels, or nonnumerical entries.
  - *Quantitative data* consist of numerical measurements or counts.
- Levels of measurement:
  - Nominal: categorized using names, labels, or qualities.
  - Ordinal: can be arranged in order or ranked.
  - Interval: can be ordered and meaningful differences between entries can be calculated.
  - Ratio: similar to interval, but there is a zero entry that is an inherent zero (implies none).

Measures of Central Tendency

- The *mean* of a data set is the sum of the data entries divided by the number of entries.
  - Population mean:
    $$\mu = \frac{\sum x}{N}$$
  - Sample mean:
    $$\bar{x} = \frac{\sum x}{n}$$
- The *median* of a data set is the value that lies in the middle of the data when the data is in sorted order.
- The *mode* of a data set is the data entry that occurs with the greatest frequency.

Measures of Central Tendency

- An *outlier* is a data entry that is far removed from the other entries in the data set.
- A *weighted mean* is the mean of a data set whose entries have varying weights. A weighted mean is given by:

  $$\bar{x} = \frac{\sum x \cdot w}{\sum w}$$

  where w is the weight of each entry x.

Measures of Variation

- The *range* of a data set is the difference between the maximum and minimum data entries in the set.
- The *deviation* of an entry x in a population data set is the difference between the entry and the mean μ of the data set.

$$\text{Deviation of x} = x - \mu$$

- The *population variance* of a population data set of N entries is

$$\text{Population variance} = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

where the symbol σ is a lowercase Greek letter Sigma.

Measures of Variation

- The *population standard deviation* of a population data set of N entries is the square root of the population variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Finding Population Variance and Standard Deviation

1. Find the mean of the population data set. $\mu = \frac{\sum x}{N}$
2. Find the devation of each entry. $x - \mu$
3. Square each deviation. $(x - \mu)^2$
4. Add to get the *sum of squares* $SS_x = \sum (x - \mu)^2$
5. Divide by N to get the *population variance*. $\sigma^2 = \frac{\sum (x-\mu)^2}{N}$
6. Find the square root of the variance to get

the *population standard deviation*. $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$

Measures of Variation

- The *sample variance* and *sample standard deviation* of a sample data set of *n* entries are

$$\text{Sample variance} = s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Measures of Variation Symbols

	Population	Sample
Variance	σ^2	s^2
Standard deviation	σ	s
Mean	μ	$\bar{x}$
Number of entries	N	n
Deviation	$x - \mu$	$x - \bar{x}$
Sum of squares	$\sum(x - \mu)^2$	$\sum(x - \bar{x})^2$