

Data in Biology

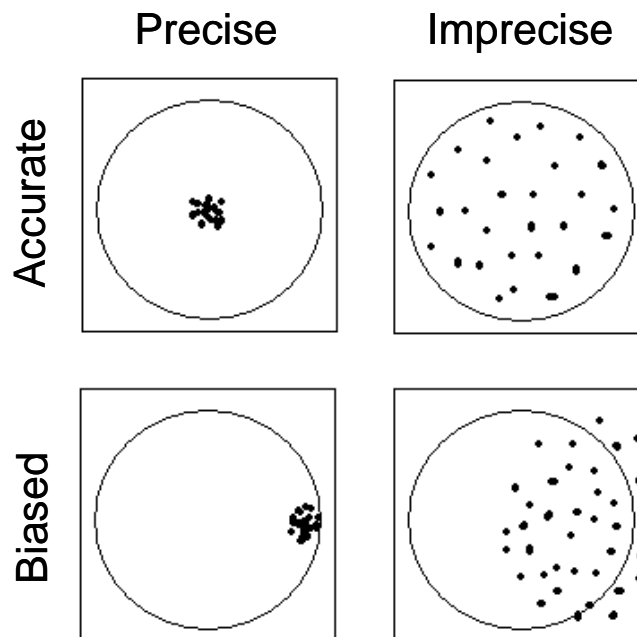
- I. Samples and Populations
 - a. **Individual observations** (observations) – observations or measurements taken on the smallest sampling unit
 - i. The standard length of a roundnose minnow
 - b. **Variable** (character – evolutionary and systematic biology) – the actual property measured by the individual observations
 - i. Univariate – one variable
 - ii. Bivariate – two variables
 - iii. Multivariate – three or more variables
 - c. **Sample of observations** (sample) – a collection of individual observations selected by a specified procedure; a subset of the population
 - i. a portion of the population
 - ii. The 124 lengths of individual minnows
 - d. **Population** – the totality of individual observations about which inferences are to be made, existing anywhere in the world or at least within a specified sampling area limited in space and time
 - i. universe of potential observations
 - ii. The lengths of all 124 minnows
 - e. **Statistic** – any one of many calculated or estimated statistical quantities that characterize a sample
 - i. E.g., mean, standard deviation, correlation coefficient
 - ii. Symbolized by Arabic letters
 - f. **Parameter** - any one of many calculated or estimated statistical quantities that characterize a population
 - i. E.g., mean, standard deviation, correlation coefficient
 - ii. Symbolized by Greek letters

- II. Variables in Biology
 - a. Quantitative vs. qualitative variables
 - i. **Quantitative data** consist of numbers representing counts or measurements.
 1. The weights of manatees.
 - ii. **Qualitative** (or **categorical** or **attribute**) **data** can be separated into different categories that are distinguished by some non-numeric characteristic.
 1. The genders of bears.
 - b. Discrete vs. continuous variables
 - i. **Discrete data** result when the number of possible values is either a finite number or a “countable” number.
 1. The number of eggs that hens lay.
 - ii. **Continuous data** result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps.
 1. The amount of milk from cows.

- c. Levels of measurement
- i. The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only; the data cannot be arranged in an ordering scheme.
 1. e.g., colors
 - ii. Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences between data values either cannot be determined or are meaningless.
 1. e.g., course grades
 - iii. The **interval level of measurement** is like the ordinal level, with the additional property that the difference between any two data values is meaningful; however, data at this level do not have a *natural zero* starting point.
 1. e.g., temperature (°F)
 - iv. The **ratio level of measurement** is the interval level with the additional property that there is also a natural zero starting point.
 1. e.g., distance

III. Precision and Accuracy

- a. **Precision** is the closeness of repeated measurements to the same quantity
 - i. Assessed by repeated measurements
 - ii. Deviation in precision is impression or insensitivity
- b. **Accuracy** is the closeness of a measured or estimate value to its true value
 - i. Assessed by some reference to theory
 - ii. Deviation in accuracy is a bias



Descriptive Statistics

- I. Basic Definitions
 - a. **Descriptive statistics** – methods used to summarize or describe the important characteristics of a data set.
 - b. **Inferential statistics** – methods used to make inferences or generalizations about a population that goes beyond the data.

- II. Important Characteristics of Data
 - a. **Center** - a representative or average value that indicates where the middle of the data set is located
 - b. **Variation** – a measure of the amount that the data values vary among themselves
 - c. **Distribution** – the nature or shape of the distribution of the data (such as bell-shaped, uniform, or skewed)
 - d. **Outliers** – sample values that lie very far away from the vast majority of other sample values

- III. Measures of center
 - a. The **arithmetic mean** of a set of values is the measure of center found by adding the values and dividing the total by the number of values.
 - i. Sample: $\bar{x} = \frac{\sum x}{n}$
 - ii. Population: $\mu = \frac{\sum x}{N}$
 - b. The **median** of a data set is the measure of center that is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude.
 - c. The **mode** of a data set is the value that occurs most frequently.
 - i. When two values occur with the same greatest frequency, each one is a mode and the data set is **bimodal**.
 - ii. When more than two values occur with the same greatest frequency, each is a mode and the data set is said to be **multimodal**.
 - iii. When no value is repeated, we say that there is no mode.
 - d. The **midrange** is the measure of center that is the value midway between the highest and lowest values in the original data set. It is found by adding the highest value to the lowest value and then dividing the sum by two.
 - e. A distribution of data is **skewed** if it is not symmetric and extends more to one side than the other.

- IV. Measures of variation
 - a. The **range** of a set of data is the difference between the maximum value and the minimum value.
 - i. Range = (Maximum value) – (Minimum value)
 - b. The **standard deviation** of a set of sample values is a measure of variation of values about the mean.

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

$$\text{ii. Population: } \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

- c. The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.

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

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

- d. The **coefficient of variation** for a set of non-negative sample or population data, expressed as a percent, describes the standard deviation relative to the mean and is given by

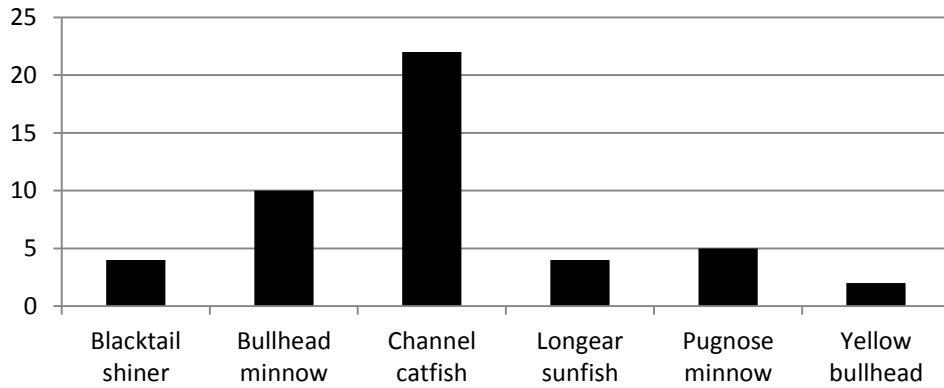
$$\text{i. Sample: } CV = \frac{s}{\bar{x}} * 100\%$$

$$\text{ii. Population: } CV = \frac{\sigma}{\bar{x}\mu} * 100\%$$

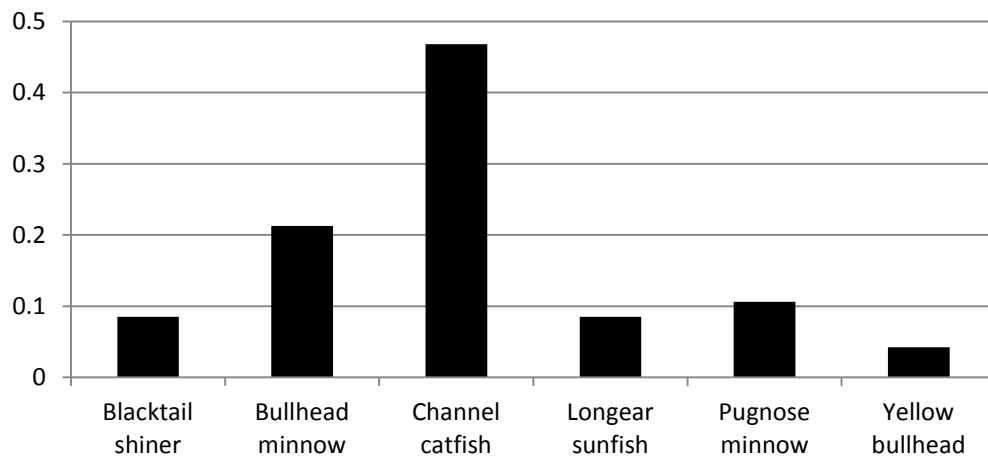
V. Frequency Distributions

- a. A **frequency distribution** lists data values (either individually or by groups or intervals), along with their corresponding frequencies (or counts).
- b. A **relative frequency distribution** includes the same class limits as a frequency distribution, but relative frequencies are used instead of actual frequencies.
- $$\text{i. relative frequencies} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$
- c. A **cumulative frequency distribution** includes the same class limits as a frequency distribution, but cumulative frequencies (i.e., the sum of the frequencies for a class and all previous classes) are used instead of actual frequencies.

Absolute Frequencies



Relative Frequencies



Cummulative Frequencies

