

# THE BASICS

## What are statistics?

Statistics can summarize and simplify large amounts of numerical data.

Using statistics one can draw conclusions about data.

Statistics is a discipline that examines data and can calculate numerical estimates of "true" values.

Statistics **can not** prove anything- estimates are normally presented in probabilistic terms (e.g. we are 95% sure ...)

Statistics **can not** make bad data better - "garbage in, garbage out"

## Why use statistics?

Want to characterize something (species, community composition, stratigraphic range, average grain size, etc...) for which we have only a limited sample- we must therefore **estimate** the "true" parameters by employing statistical methods.

Statistics may reveal underlying patterns in data not normally observable (especially true in multivariate analyses).

If used correctly, statistics can separate the probable from the possible

### Types of Data

- **ratio-scale data.** Measurements along a continuous scale whose scale begins at 0 (e.g. lengths or widths in mm).
- **interval-scale data.** Same as ratio, but data do not have 0 as low end of scale (e.g. temperature).
- **ordinal-scale data.** Generally used for irregular scaled data converted to ranks or relative position (e.g. position of stratigraphic stages).
- **discrete data.** Not continuous, usually counts (e.g. number of individuals per sample).
- **nominal or categorical data.** Includes binary data (e.g. presence/absence) or group data (e.g. sandstone/siltstone/mudstone).

## Some Basic Definitions

- **variable:** Anything that varies and can be measured (e.g. measurement, property, quantity, and attribute). Determining the relationships between variables is the realm of R-mode analysis.
- **object:** Unit of study *on which* variables can be measured (e.g. case, individual, specimen). Determining the relationships between objects is the realm of Q-mode analysis.
- **population:** The total set of measurements. The limits of the population should be designated before any analysis. (e.g. the size of all specimens of brachiopod species x.)

Usually the population is *unknowable* and must be *estimated* by a **sample**.

- **sample:** Collection of objects which are a subset of the population of interest and are taken as *representative* of the population. (e.g. the size of 20 specimens of brachiopod species x from a particular outcrop.)

Note: if the sample is not representative of the population, it is said to be **biased**. e.g. collecting only large specimens of brachiopod species x only within arm's reach of an outcrop would lead to a biased sample.

- **sample size:** How big must it be for the sample to represent the population? No real answer as it depends upon the variability of the population and the degree of precision one wants to achieve in answering the question.

more on how to determine sample size later ....

- **parametric statistics:** Statistical procedures used on interval or ratio data. Usually many assumptions must be made.
- **nonparametric statistics:** Statistical procedures used on ordinal data based on ranks. Not so many assumptions are necessary.
- **precision:** *Reliability* of a measurement. Usually determined by taking repeated measurements.

- **accuracy:** The *closeness* of a measurement to the true value. Usually unknown in biology, geology and paleontology, but can sometimes be determined from a known standard.

What statistics to use? answer depends on . . .

- what questions you want to ask
- types and quality of data available (e.g. parametric or nonparametric)
- can be descriptive, comparative, or classificatory
- can involve one or more samples (one-way or two-way analyses) and one or more variables (e.g. univariate , bivariate, or multivariate)
- topic of course ! ! !

## SIGNIFICANT FIGURES

- Should always maintain significant digits through all calculations
- The last digit should imply precision *and is an estimate*
- For example:  
  
45.346 implies any number between 45.3455 and 45.3465
- The last digit, including 0 (zero) to the right of decimal points, is always significant!
- Should use enough significant figures to have at least 30-300 divisions between lowest and highest measurements
- Because an error of 1 digit in samples with less than 30 divisions will have an error of more than is acceptable for most statistical tests

## ROUNDING NUMBERS

- Should round numbers to get desired significant number of digits
- The Rules:
  - Not changed if it is followed by a number equal to or less than 5
  - Add 1 to the number if it is followed by a number greater than 5

---

Number	Significant Figures	Answer
26.58	2	27
133.7137	5	133.71
0.03725	3	0.0372
0.037152	3	0.0372

---