Introduction to Statistics

I. Introduction to Stats

Statistics deals with variation and attempting to draw conclusions from data despite of variation of data.

A. View of Biol data

Biologists view Biological phenomena in two different ways

1. view that Bio Phenom are not deterministic but ***probablistic*** - Response may occur at a characteristic frequency. Thus one makes the assumption that there is inherent variability and randomness. These data can be used to **predict** a **probability** that an event will occur **not determine** if an event will occur.

2. others view that Biol Phenom are Deterministic - inevitable and always occur. However, methods used, environment, organisms all vary and are not identical to one another. Thus will also get variation in results

B. Roles of Stats

Biostatistics has 2 major roles

1. condense variable information into a summary form that conveys information in an intelligibel way (summary stats)

2. Assess whether given variability in your data are consistent with your hypothesis (inferential stats)

e.g.

Parasite data - Statistics helps us to determine what the data are telling us.

C. Definitions

Definitions

Data - information pertinent to answering some question

Population - The entire group of things that are of interest. The group to which you are trying to generalize

1. Observational (wing lengths of House Flies)
2. Experimental (wing lengths of males on standard diet)

Observational Study - investigation of properties of a population.

Experimental Study - Procedure designed to collect observations according to a prearranged plan, under defined conditions, under the control of the investigator.

Samples - the portion of the population that is measured

Sample Unit - the “thing” that is measured (may not be individual)

Random Sample - sample drawn so that all members of a population have an equal and independent chance of being included in the sample

II. Descriptive Stats

A. Location - where on a scale does data fall

1. Mean is simply the average of a sample. It describes the location of data.

Advantage-

simple to compute and interpret

large amount of theory on how to do statistics on mean

Disadvantage

heavily influenced by extremes. If the distribution is skewed then not a good measure of location

e.g.

2. Median - middle value - 50% less than and 50% more than

Rank data smallest to largest - median is value with rank n+1/2

odd

14 17 **18** 20 21

even

14 17 **!** 18 20

This is necessary if data is asymetrical

3. Mode - Most frequent value - commonest

B. Dispersion

Spread of data around central location.

1. Range - difference between max and min

very sensitive to extreme values

2. Standard deviation - a measure of mean deviation of observations from the mean of the distribution (mean distance from the mean). Has same units as original data as does mean.

formula -

3. Variance - quantify how far each observation is from the mean (does not have units associated with variance).

4. Coefficient of variation - the STD expressed as % of the mean. For most biol data if have large mean then have large STD.

III. Inferential Stats

A. T-test

Often one takes 2 independent samples and wishes to compare the sample means.

If the sample means were drawn from the same population with the same means, then the calculated means are estimated of the same population. Any differences between them is a result of chance / sampling error.

The null hypothesis is that the means of the 2 populations are equal.

If you reject the null you are concluding that the two samples were probably drawn from populations with different means.

The level of probability is chosen to be 0.05 (you are willing to take 1 in 20 chances of rejecting the null when it is in fact true). This probability is referred to as alpha.