Introduction to Statistics

Statistics

Statistics are a set of tools that help us to
- Organize data
  - data -- set of systematic measurements or observations
- Summarize data
- Interpret data
  - Determine the relation between two or more variables
  - variable -- a characteristic or condition of an object or human
  - that has different values for different individuals
  - infer if differences exist between different conditions in an experiment

Populations

- The population is the group of all people or objects that we are interested in
- Humans are often the population of interest in psychology studies
- Smaller populations are possible -- all the students in one particular class of introductory statistics
- Usually the entire population cannot be studied. Why?
- Population parameter -- a value that describes a population
Samples

- In most research, a subset of the population called the sample is selected.
- If the sample is selected so that each member of the population has an equal chance of being selected (called a random sample), and
- if the sample is sufficiently large, then
- whatever we learn about the sample will probably also be true of the population.

Sample statistic - a value that describes a sample.

Descriptive Statistics

Descriptive statistics allow us to summarize, organize and simplify data.

- Measures of central tendency tell us about the average value.
  - E.g., the mean, median, and mode.
- Measures of dispersion tell us how similar the data are to the average value.
  - E.g., range, semi-interquartile range, and standard deviation.

Inferential Statistics

Inferential statistics allow us to study samples and then make generalizations about the population from which they were selected.

- E.g., t-test, ANOVA
- Samples are imperfect representations of parameters.
  - Sampling error -- the discrepancy between a sample statistic and a population parameter.
Methods: Correlation

The correlational method involves measuring two (or more) variables to determine whether there is a relationship between them.

Methods: Experiments

Experiments are a special type of research in which all the variables except the independent and dependent variables are held constant.

- Independent variable -- the variable that the researcher systematically manipulates
- Dependent variable -- the variable that the researcher measures or records

Methods: Experiments

- Control condition: the group that does not receive the treatment
- Treatment condition: the group that receives the treatment
- Performance in the treatment condition is compared to performance in the control condition to see if the treatment had an effect
Pick the Variables

A researcher randomly assigned participants to one of two groups -- either the group that consumes 3 beers or the group that consumes no beer. Then both groups view a list of 20 words that they are to remember. Later they recall the words.

What is the independent variable? What is the dependent variable? What is the control condition? What is the treatment condition?

Methods: Experiments

The main advantage of performing experiments is that they allow us, using statistics, to infer if the independent variables causes the change in the dependent variable.

Quasi-Experiments

A quasi-experiment is similar to a real experiment except that the participants have been assigned to the various groups based on some characteristic of the participant. E.g., participants are assigned to one of two groups based on whether they are male or female.
Quasi-Experiments

A researcher assigned students into one of two groups -- whether they over 8 years of age or 8 years or less. They were asked to repeatedly recite the alphabet backwards until they did it perfectly. The number of recitations of the alphabet was recorded.

If there is a difference in the number of recitations, can we say it is because of the difference in age?

Discrete and Continuous Variables

Variables can be either discrete (or discontinuous) or continuous.

A discrete variable consists of separate, indivisible categories. No values can exist between two neighboring categories.

E.g. number of children in a family

A continuous variable can have any value.

E.g. weight

Continuous Variables

Measurements of continuous variables is imperfect.

Real limits are the boundaries of intervals for scores that are represented on a continuous number line.

- Lower limit is the bottom of the interval.
- Upper limit is the top of the interval.

Measure height to nearest cm: 182 cm.

- Lower real limit is 181.5 cm.
- Upper real limit is 182.5 cm.
Levels of Measurement

When we observe and record a variable, it has characteristics that influence the type of statistical analysis that we can perform on it. These characteristics are referred to as the level of measurement of the variable. The first step in any statistical analysis is to determine the level of measurement; it tells us what statistical tests can and cannot be performed.

Nominal Scale

The nominal scale consists of a set of categories that have different names in no particular order. E.g. Hair color, whether a person has schizophrenia or not, five digit ZIP codes. Nominal variables deal with qualitative (and not quantitative) differences.

Nominal Variables

The values of nominal variables can be compared to see if they are equal or not. The values of nominal variables cannot be meaningfully: compared to see if one is larger than another; added or subtracted; multiplied or divided. Cannot calculate the mean (what most people call the average).
Ordinal Scale

*Ordinal scale* consists of a set of categories organized in an ordered sequence. Spacing between categories is not equal.

- E.g. Class rank, order of finishing a horse race, how much you prefer various vegetables

Ordinal Variables

- The values of ordinal variables can be:
  - compared to see if they are equal or not
  - compared to see if one is larger or smaller than another

- The values of ordinal variables cannot be meaningfully:
  - added or subtracted
  - multiplied or divided
  - Cannot calculate the mean

Interval Scale

*Interval scale* consists of ordered categories that are all intervals of exactly the same size. The 0 point is arbitrary.

- E.g. Shoe size, IQ scores, °Fahrenheit
Interval Variables

- The values of interval variables can be:
  - compared to see if they are equal or not
  - compared to see if one is larger or smaller than another
  - added or subtracted

- The values of interval variables cannot be meaningfully:
  - multiplied or divided

  - $70^\circ\text{F}$ is not twice as hot as $35^\circ\text{F}$

Ratio Scale

- *Ratio scale* is an interval scale with an absolute zero

  - The absolute zero implies that a value of zero represents the complete absence of the variable

  - E.g. Weight, reaction time, number correct

Ratio Variables

- The values of ratio variables can be:
  - compared to see if they are equal or not
  - compared to see if one is larger or smaller than another
  - added or subtracted
  - multiplied or divided
Representing Variables

By convention, in statistical formulae variables are represented by a capitalized letter, usually X or Y.

E.g., X might represent how introverted the people in your sample are.

Representing Individual Values

When a variable is subscripted (X_i), the subscript implies that you should deal with a particular observation.

E.g., X_i might represent how introverted the third person in your sample is.

The Summation Operator (∑)

Most statistical procedures involve the summation of the values of variables.

Rather than to write all the values out (X_1 + X_2 + X_3 + ...) a short hand notation is used:

ΣX

N represents the number of observations.

<table>
<thead>
<tr>
<th>i</th>
<th>X_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

ΣX = 12
N = 4