1/24/19

1. Data Collection

1.1 Intro of Stats

Statistics is the science of collecting, organizing, summarizing, and analyzing information (data) to draw conclusions or answer questions. Statistics is also about providing a measure of confidence in any conclusions.

Data can be numerical or not, but in both cases, describe characteristics of an individual.

Certainty is not a concept applicable to statistics the way it is for other fields of mathematics.

Language of Statistics

The entire group to be studied is called a **population.**

An **individual** is a person or object that is a member of the population.

A **sample** is a subset of the population (typically more than a single individual).

A **statistic** is a numerical summary of a sample.

A **parameter** is a numerical summary of a population.

Descriptive statistics consist of organizing and summarizing data. (tables, graphs, etc)

Inferential statistics uses methods that take a result from a sample, extend it to a population, and measure the reliability of the result.

Variables are the characteristics of the individuals in the population.

ex If the population is the set of students in 1342.13, then variables would include:

(a) gender

(b) major

(c) age

(d) height

(e) eye color

Qualitative or *categorical* variables allow for classification based on attributes or characteristics.

Quantitative variables provide numerical measures that can be added/subtracted with meaningful results.

Discrete variables are quantitative and have a finite or countable number of positive values.

Continuous variables are quantitative and have an infinite number of possible values that are not countable.

ex Determine whether the variables are discrete or continuous.

(a) the number of heads obtained after flipping a coin five times.

(b) the number of cars that arrive at the fast food drive-thru between 12-1pm.

(c) the distance a minivan can travel on a full tank of gas in city driving conditions.

Be careful to distinguish between variables and the data (values the variables can take).

Levels of Variable Measurement

nominal - name, label or categorize, but do not allow for arrangement by rank or order.

ordinal - allows for arrangement by rank or order.

interval - differences in values have meaning. Addition/subtraction operations possible. $0 \neq$ absence.

ratio - $0 =$ absence of the quantity. Multiplication/division possible.

ex For each variable, determine the level of measurement.

(a) gender – Nominal.

(b) temperature – interval.

(c) number of days this week a college student studied – Ratio.

(d) letter grade earned in your statistics class

1/24/19

1. Data Collection

1.2 Observational Studies vs. Designed Experiments

In studies/experiments researchers investigate a relationship between two variables:

explanatory variable $\longrightarrow$ *response* variable

Observational Study measures the value of the response variable without attempting to influence the value of either the response or explanatory variables.

Designed Experiment is a study where the researcher assigns individuals to a certain group, changes the value of an explanatory variable and records the value of a response variable for each group.

Confounding occurs when the effects of two or more explanatory variables are not separated. Any relation that may exist between explanatory and response variables may be due to some other variable(s) not accounted for.

Lurking variables are explanatory variables not considered in a study, but affecting the response variable value.

Confounding variables are explanatory variables that are considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

* Observational studies do not allow a researcher to claim causation, only association.

Types of Observational Study

(1) Cross-sectional: collects data about individuals at a specific point in time (back in time)

(2) Case-control: collects retrospective data (existing records)

(3) Cohort: group identified, then observations recorded over a long period of time

Existing sources of data: CDC, IRS, US Census

`fedstate.sites.usa.gov`

`https://gssdataexplorer.norc.org`

1. Data Collection

1.3 Simple Random Sampling

Random Sampling is the process of using chance to select individuals from a population to be included in the sample.

Simple Random Sample A sample of size n from a population of size N is obtained through simple random sampling if every possible sample of size n has an equally likely chance of occurring.

Obtaining a Simple Random Sample

(1) Obtain a *frame*: a list of all the individuals within a population.

(2) Use a table of random numbers or random number generator to select from the frame.

Considerations

Are we sampling with or without replacement?

Is it impossible to obtain a frame?

* If convenience is used to obtain a sample, the results of the survey are meaningless!

SRS with a table or random numbers

(1) Collect the frame and assign numbers to the individuals

(2) Determine how many individuals will be selected

(3) Pick a random starting point in the table (point with eyes closed)

(4) Look for numbers in the table which fall within the numbered data range.