Statistics 301: Probability and Statistics

Introduction to Statistics

Module 1

2018

Introduction to Statistics

Statistics is a *science*, not a branch of mathematics, but uses mathematical models as essential tools.[1]

Like other branches of science, statistics uses math, but is not *inherently* a math class. Just like physics uses math [for everything!], but is *not* a math class.[2]

Defined, statistics is the science of data. Sample information is obtained and inferences about the population are made from the sample information. We use *descriptive statistics* (graphs, numerical summaries, etc.) and *inferential statistics* (formal methods using probabilities)

Definitions I

Population
the entire group of people or things that we are interested in researching

Sample
the subset of the population; the group from which data is obtained

Parameter
measurement/property of the population

Statistic
measurement/property of the sample; estimate of the parameter

Statistical Inference
educated estimations about populations with sample information

Probability
the likelihood (chance) of an event occurring

Examples of Definitions I

Suppose we are interested in average spending on on-campus students for food per month.

population: all students living on campus

sample: a subset, say 50 randomly selected students

parameter: true average spending of all students living on-campus of food per month

sample: average spending of the sample of students living on-campus of food per month

[1] John Tukey

[2] M. Londe

Definitions II

Variable
characteristic of interest for each person/thing
- numeric (quantitative)
- categorical (qualitative)

Data
the actual values of the variable (numeric or categorical)
- datum is singular, data are plural
- the number of observations in a sample, the sample size, is denoted as n

Mean
mathematical average, arithmetic mean, etc.; is a center location

Proportion
is a part of a whole; commonly expressed as a percentage in study results
- Example: proportion of people who like Brand X is 0.55, or 55%

Data Types

Qualitative data happens when information is categorized (descriptions of attributes). Examples include hair color, blood type, type of car, etc.

Quantitative data are always numbers, resulting form counting or measuring attributes of a population. Examples include weight, height, age, time to complete a task, etc. Quantitative data can be either discrete or continuous.

(1) *discrete*: results of counting; they take on ony whole number values, usually specified. Example: number of phone calls you receive each day of the week (it will not be an infinite number... you hope)

(2) *continuous*: results of measuring; they take on any value within an interval. Example: measuring height, someone can be 56.259997861024 inches tall

Data types

Levels of Measurement:
(1) Nominal scale: qualitative data (favorite foods cannot be analyzed)
(2) Ordinal scale (Like a Likert scale)
(3) Interval scale (like temperature)

Sampling I

Researching populations is costly and/or impossible, which is why we sample

Observational study
no treatment and no real active researcher interference involved (surveys, etc.)

Representative sample
a sample that contains the characteristics of the population of interest

Random sampling
samples are selected using a random sampling method; random sampling best mirrors the population of interest

Sampling II

Sampling without Replacement (swor)
when an individual is randomly chosen for the sample, they are 'removed' from the population, so that there is one less individual in the population to choose from every time (and they have already been chosen)

Sampling with Replacement (swr)
when an individual is randomly chosen for the sample, they are placed back into the population, so that the total number of individuals in the population to choose from does not change; an individual could be chosen more than once

Sampling designs

Simple Random Sample (SRS): any group of n individuals have an equal chance of being chose for the sample

Stratified Random Sample (StRS): separate the population into groups called strata and then take SRS of each stratum.

1-stage Cluster Design: separate the population into groups called clusters and then take SRS of n clusters (using *all* individuals from chosen cluster(s)).

2-stage Cluster Design: separate the population into groups called clusters and then take SRS of n clusters, then an SRS is taken of the individuals from chosen clusters in first stage.

Systematic sample with a Random Start (SyRS): randomly select a starting individual then observe every k^{th} individual, usually from a list.

Sampling designs example data

The data: a class roster with year and major information per student supplied for 30 students

survey examples

Errors

Sampling bias: created when samples are collected from populations where some individuals are more likely to be chosen. A sample is never perfectly representative but a large enough sample can reduce the variation.

Choosing a sample using easy to access data (no random design) leads to bias is called *convenience sampling*. A sample that "chooses itself" (website polls, call-in polls, etc.) leads to bias is called *volunteer sampling* (ie, usually people with really strong opinions on either side are more apt to call in)

Errors in sampling:
(1) *sampling errors*: the actual process of sampling can cause errors (because you are not looking at the whole population). One reason may include a sample size that is not large enough.
(2) *nonsampling errors*: factors not related to the sampling process; a defective measurement tool can cause nonsampling errors.

Variation

Variation : is present is any dataset; measuring the same thing with the same tool can sometimes produced variation in the measurements. Using random assignment for samples means every sample will have slightly different results, since they are not “capturing” the same individuals each time so the measurements will vary from sample to sample. This is called *sampling variability.*

Once data is collected, the next step is to organize it so that analyzing it is easier.

Frequencies

frequency : the number of times a value of the data occurs.

relative frequency (rf) : is the proportion of the number of times a value of the data occurs. To find rf, divide each frequency by the total number of observations in the study.

cumulative relative frequency : the accumulation of the previous relative frequencies.

Frequencies example

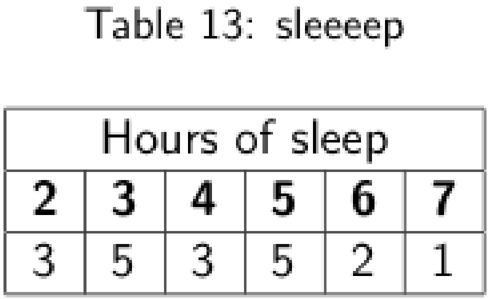

Table 13: sleeeep

Hours of sleep					
2	**3**	**4**	**5**	**6**	**7**
3	5	3	5	2	1

Figure 1: sleeeep

Frequencies example

```
         2      3      4      5      6       7
rf  0.1579 0.2632 0.1579 0.2632 0.1053 0.05263
crf 0.1579 0.4211 0.5789 0.8421 0.9474 1.00000
```

Experimental Design I

Experiment
researchers actively impose a treatment (an experimental condition) to individuals (people, animals, things, etc.) and record the results. Main purpose is to establish cause and effect between 2 or more variables

Unit
an individual in the experiment; also referred to as experimental unit

Subject
human or animal experimental units

Experimental Design II

Blinding
when the subject does not have any knowledge of the treatment they are receiving

Double-blinding: *when neither subjects nor researchers directly interacting with the subjects know how the treatments are assigned*

Placebo
a fake treatment; when subjects know what treatment they are receiving, results can be skewed (placebo effect)

Control
the group receiving either a placebo or no treatment; used as a basline to detect significant changes/differences by comparing treatment groups to the control group(s)

Experimental Design III

Response variable
the variable that is studied for changes in response to another variable (explanatory)

Explanatory variable
the variable that causes the change in the response variable

Lurking variable
variables that are not accounted for in the experiment but can interfere with results

Confounding
when effects of 2 or more variables cannot be distinguished from one another

Experimental Designs: CRD, RCBD, Matched pairs

Completely Randomized Design (CRD): one which each unit has an equal chance to be chosen for the treatment group; all units are as homogeneous (similar) as possible.

Randomized Complete Block Design (RCBD): when the experimental units are not homogeneous (similar) enough, detecting differences among treatment groups can be difficult. When we think the entire group of units is not similar enough, we create groups, called blocks, for comparing t treatments in b blocks. The units within the blocks are similar to each other; all blocks see all treatments in random order.

Matched Pairs : when we take measurements on the same unit, usually once before and once after a treatments (examples include weight loss programs, Coke vs. Pepsi,...)

Diagram of CRD

Want to test new and old fertilizer on tomato plants, and will include a control with no fertilizer, just water.

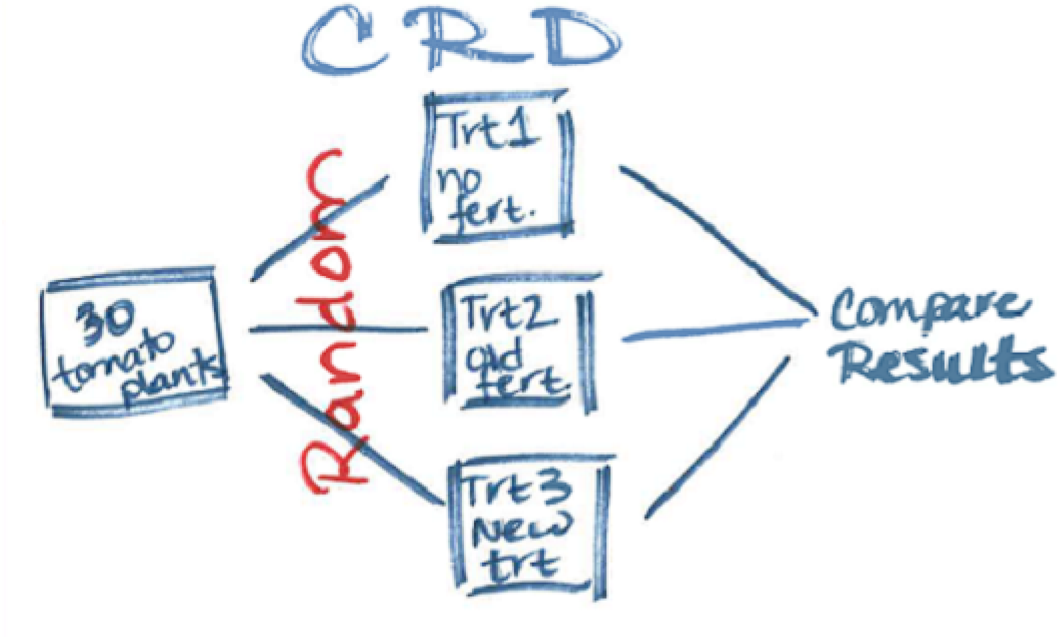
CRD
Random
30 tomato plants
Trt1 no fert.
Trt2 old fert.
Trt3 New trt
Compare Results

Figure 2: CRD