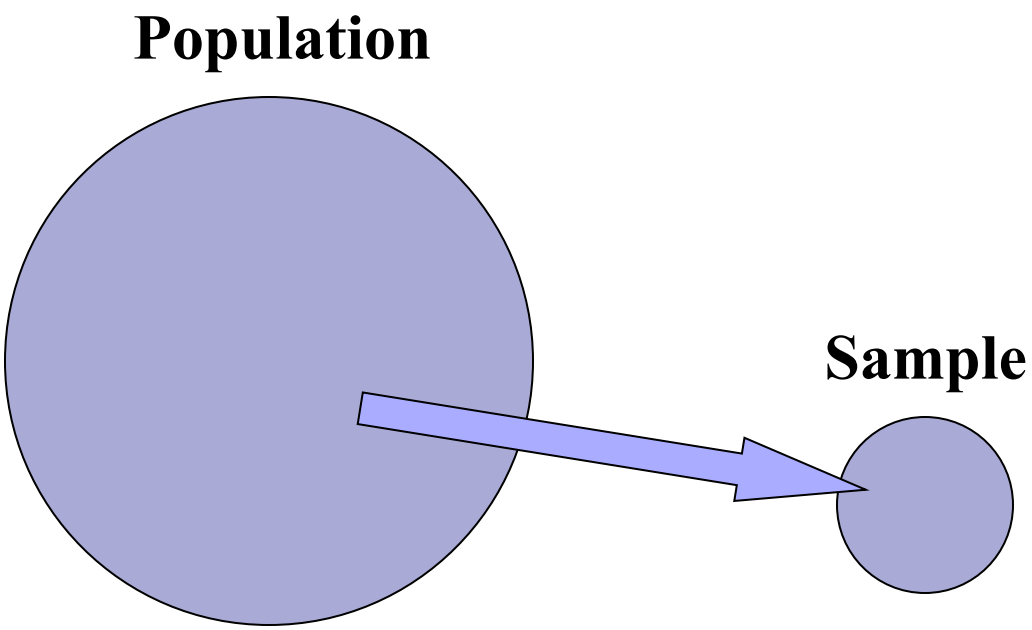
Section 1.2: Sampling

- Idea 1: Examine a part of the whole.



Idea 1: Examine a part of the whole.

Population –

Entire group of individuals that we want to make a statement about.

Sample –

Part of the population we actually examine.

e.g.

Population: My 9am statistics class

Sample: The group defined by all students sitting in a seat with a seat number ending in a ‘2’.

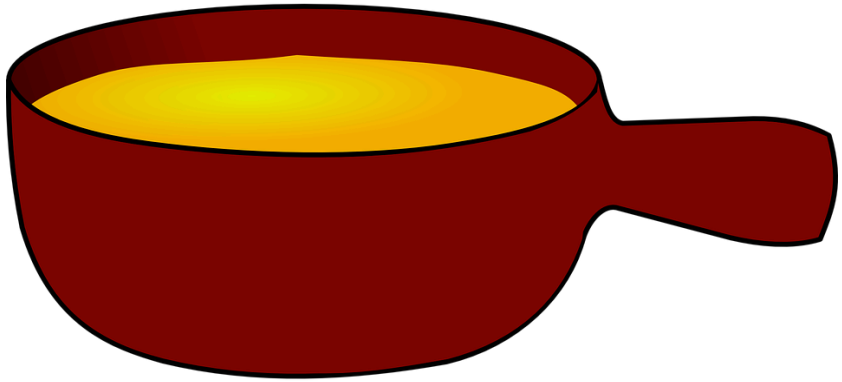
What about a census?

Collect info on everyone

- Would a census of the population be a better way to go?
  - Often difficult to do
    - time, money, resources, non-responders, etc.
  - Populations are often dynamic
    - They're changing as you're collecting the data
  - Can be complex, who gets missed?

Properties of a Sample

- Would like the sample to be <u>representative</u> of the population.



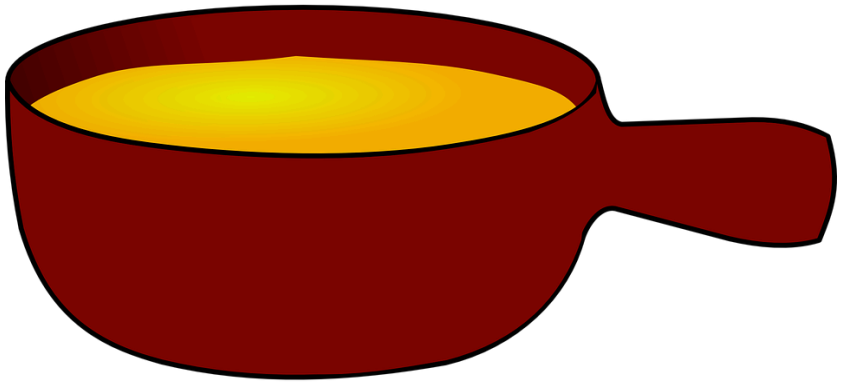
Suppose you want to taste (or sample) your soup.

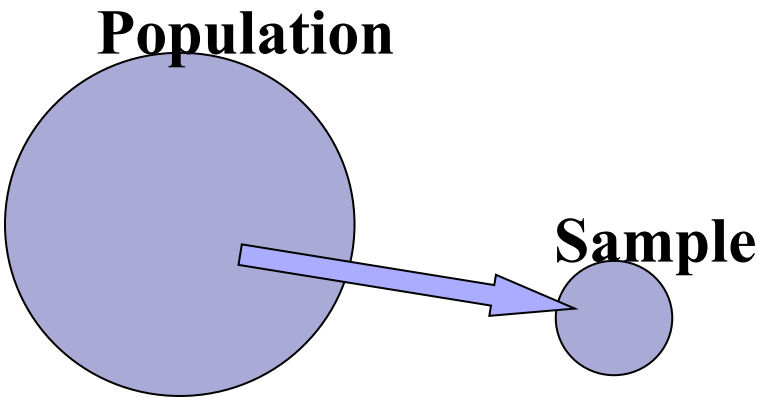
If you leave it sitting for 2 hours and spoon off the top, would that be representative of the soup as a whole? Will you miss some important parts?

If you stir it thoroughly and then take a taste, would that be more representative of the soup as a whole?

Properties of a Sample

- A **representative sample** is a sample in which the relevant characteristics of the sample members are generally the same as the characteristics of the population.





Properties of a Sample

- Getting a perfectly representative sample may not be possible, but we would at least like a sample that is not biased.

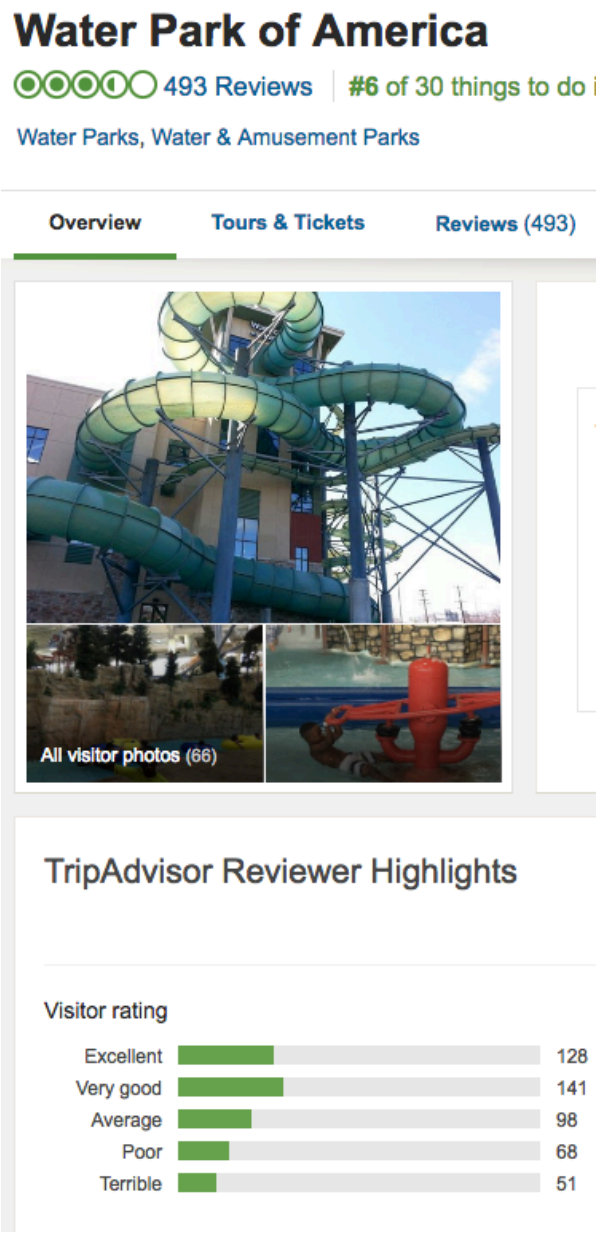
Biased Sample – the sample is 'out of step' with the full population. A biased sample differs in a 'specific way' from the population.

Are we Introducing **bias**? How?

- Response: Grade Point Average (GPA)
  - Population (whole): STAT1010 class
  - Sample (subset): All students in last 3 rows
    - Is it a representative sample?

Are we Introducing **bias**? How?

- Response: Hotel quality
  - Population (whole): All users of the hotel
  - Sample (subset): Users who too the time to upload review on internet
    - Is it a representative sample?



Are we Introducing **bias**? How?

- Response: Defect rate of a product
  - Population (whole): all products produced
  - Sample (subset): products produced on Friday from 3-5pm
    - Is it a representative sample?

Are we Introducing **bias**? How?

- A good statistical study MUST have a representative sample. Otherwise the sample is biased and conclusions from the study are not trustworthy.

- Gallup poll was very 'off' in presidential election prediction in 2012.
  - *Post-election examination determined "that part of the poll's overstatement of Romney support arose from too few phone interviews in the Eastern and Pacific time zones… overstating the white vote..."*
    - (See link to article in USA Today on course website)

Sample Surveys

- Idea 2: Choosing randomly
  - Selecting items for the sample should be done ***at random*** so as to reduce the chance of getting a biased sample.
  - We can't always 'perfectly' use random choice, but we do the best we can for the matter at hand.

Simple Random Sample (SRS)

- Want a representative sample but will settle for one that is not biased.

- SRS of size n=400
  - Give each individual in the population a number, then randomly generate 400 numbers as the ‘chosen’ individuals.

  - Each combination of 400 individuals has the same chance of being selected.

Simple Random Sample

- If one were to do this more than once…
  - Different random numbers will give different samples of 400 students.
  - We have introduced **variability by sampling**!

See web-based GUI applet on sampling words from the Gettysburg Address and observed word length:

http://www.rossmanchance.com/applets/OneSample.html

268 words in the population (whole)

10 chosen… Which were chosen

Paste population data or select from list:

ID	Word	Length	HasE	IsNoun
1	Four	4	No	No
2	score	5	Yes	Yes
3	and	3	No	No
4	seven	5	Yes	No
5	years	5	Yes	Yes
6	ago	3	No	No
7	our	3	No	No
8	fathers	7	Yes	Yes
9	brought	7	No	No

Pop 1
Pop 2
Pop 3
Gettysburg
Pennies
Change

Variable: Length

Use Data | Clear | Top/Bottom

Population size:268

x1 x4 x40

Show Sampling Options:

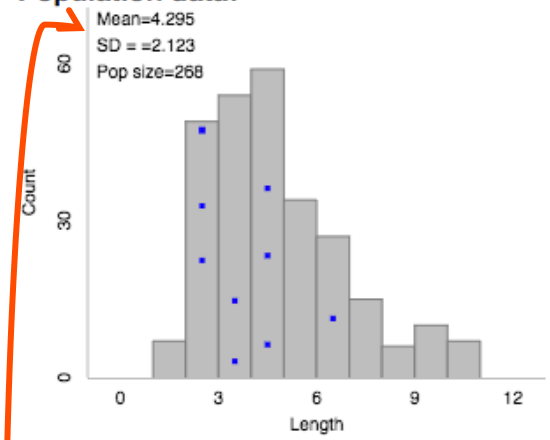
Number of samples: 1

Sample size: 10

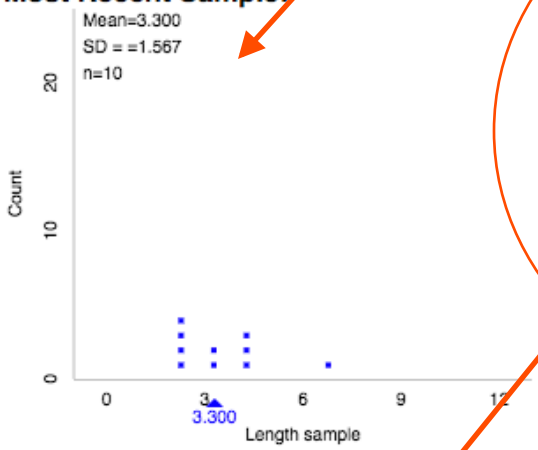
Draw Samples | Reset

28	are	3	Yes	No
237	in	2	No	No
18	in	2	No	No
249	of	2	No	No
121	men	3	Yes	Yes
175	work	4	No	Yes

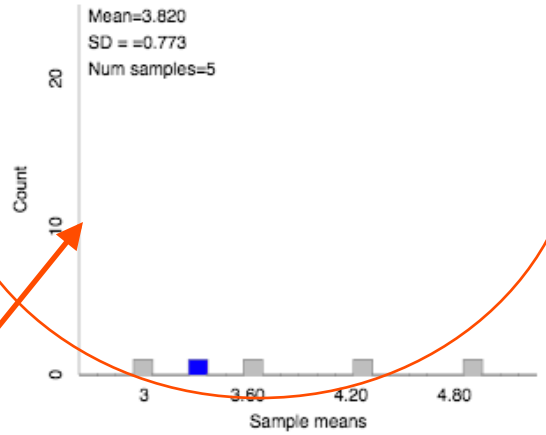
Population data:



Most Recent Sample:



Statistic: Mean Median *t*-statistic



Population scale

Count Samples Greater than ≥ Count

Overlay Normal Distribution:

One sample's information

Population information

Cumulative results over 5 different simulations

Other Sampling Plans

- Systematic Sampling
  - Select in a systematic way from the sampling frame.
    - e.g. Every 60th student (arranged alphabetically) on the list from the Registrar for opinion survey.
    - Use a random start point.
  - Caution- the order must be random...
    - Every Friday on assembly line, not a good idea.
    - Every 15 minutes at museum entry seems fine.

Other Sampling Plans

- Stratified Sampling
  - Divide population into strata (subpopulations) and select a SRS from each strata.
    - e.g. SRS from each county in Iowa.
    - Example strata: race, income, age, sex, etc,
  - Lets you make sure you're getting a certain amount of input from each strata or group.
    - All strata will be represented.

Other Sampling Plans

- Cluster
  - Divide population into clusters, randomly select some of the clusters, <u>choose all members (not SRS) from selected clusters</u> as your sample.
  - Might be more practical than SRS.
  - Note that *ALL* individuals from a chosen cluster are sampled compared to *only some* individuals from each strata in stratified sampling.

Other Sampling Plans

- Convenience
  - Use a sample that is convenient to attain.
    - e.g. Last 3 rows of students to represent class.
    - e.g. Voluntary responses on internet hotel survey.
  - In general, not a good idea.
    - Often gives biased results.
    - Could be justified in some cases, but try to use a different sampling plan if possible.

Other problems

- Question bias/Response bias
- Things that influence the response
  - Question could be worded negatively
    - Would you favor or oppose a law that would take away your constitutional right to own guns?
    - Would you favor or oppose a law that would reduce gun violence in your neighborhood?
  - Respondents don't like the interviewer
  - Respondents are embarrassed to tell truth and give false information

Other problems

- Non response
  - Is there a reason a group doesn't respond?
    - Critical thinking useful here.
  - If it's a health survey, will unhealthy people be less likely to respond?
  - Non response is a *BIG* issue in sample surveys.

Is there an association between breast cancer and abortion?

- Studies include women who have and who have not had breast cancer.
  - An observational study found there was an association.
  - Which group of women is more likely to be TOTALLY honest about their personal health?
- National Cancer Institute (2003)
  - Refuted the reliability of the study.

Variability in Samples

- Results from a **sample** provide <u>estimates</u> of the truth about a **population**.
- 2 different samples will give 2 different estimates (recall word length sampling example).
  - Why? Because we used random chance to select the sample.
  - This allows us to use probability to determine how large of an error we are likely to make – we'll talk more on this later.
- Larger samples give more accurate estimates than smaller samples.

Some main topics from Sections 1.1-1.2

- Parameter (usually a greek letter) vs. Statistic
  - Population vs. Sample
- Choose sample *at random*
  - Helps avoid getting a biased sample
- Sampling methods
  - Simple Random Sample (SRS)
  - Stratified sampling
  - Cluster sampling
  - Convenience sampling (proceed with caution)
  - Systematic sampling