Statistics lets us make inferences about a population by studying a sample chosen from it.

1.1 Sampling

e.g. We'll grill brats for a school picnic and want to know how many of 42,000 students will attend.

- If we don't know statistics, $\cdots$

- If we know statistics, $\cdots$

  Estimate that population proportion is ____________, so about ____________ will attend.

  Our estimate is unlikely to be correct. Questions include, e.g.:

  - Given a sample proportion of 40%, what surrounding interval would give us 95% confidence that it contains the population proportion?
    e.g. $(.40 \pm .00001)$? ____________
    e.g. $(.40 \pm .6)$? ____________
  - We're ____________ certain that 42,000 brats will be enough.
    We're ____________ certain that 19 will not be enough.
    Can we be 95% certain that 17000 brats will be enough?

Simple Random Samples

- A *population* is the set of individuals (objects or outcomes) about which we seek information.
- A *sample* is a ____________ of the population containing the individuals we actually observe.
- A *simple random sample (SRS) of size* n is a sample chosen so that each subset of n individuals is ________________________

  To draw a simple random sample of size n from a population of size N,

  - number individuals in population with 1 through N
  - generate n random integers in ____________, and use the corresponding individuals

- *Sampling variation* is the variation that occurs between ______________________________ from the same population.

How to Sample Badly

- A *sample of convenience* consists of individuals in the population that are ____________________

  e.g.

- A sampling design is *biased* if $\cdots$

- A *voluntary response sample* consists of people who ____________________ by responding to a broad appeal. It's biased because people with strong opinions are most likely to respond.

  e.g.

Determining Whether a Sample Is a Simple Random Sample

Independence

Items in a sample are *independent* if knowing values of some doesn't help predict values of others.

e.g. Put ten balls labeled 0 through 9 in a bucket $\cdots$

P(draw 3) =

Suppose we draw a 3; then P(draw 3) =

To *sample with replacement*, replace an item after selecting it.

e.g. Then P(draw 3) = ____________, even after drawing 3.

e.g. For a large population, this effect is negligible: with 10000 each of the ten balls in a bucket, drawing a 3 changes P(draw 3) from ____________ to ____________. We treat items in a sample with $n < (5\%)N$ as independent (even when sampling without replacement).

Other Sampling Methods

In *weighted sampling*, some items are given more weight than others.

e.g. Put ten balls labeled 0 through 9 in bucket, then add ten 3s.

P(3) =

P(i) = ____________ for each $i \neq 3$.

In *stratified random sampling*, the population is divided into subpopulations called "strata" (layers), and a SRS is taken from each stratum.

e.g. To get a sample of 200 from 42000 students and 2000 teachers at UW-Madison, $\cdots$

In *cluster sampling*, individuals are grouped into clusters, a sample of clusters is chosen, and individuals in those clusters are studied.

Types of Data

- With *quantitative or numerical* data, each item is assigned $\cdots$

  e.g.

- With *categorical or qualitative* data, each item is assigned $\cdots$

  e.g.

Controlled Experiments and Observational Studies

- A *controlled experiment* ____________________________________ individuals in order to observe their responses. Its purpose is to study whether treatment causes a change in the response. It can lead to a claim of ____.
- An *observational study* ____________________________________ and measures variables of interest, but doesn't attempt to influence responses. Its purpose is to describe some group.

e.g.