Samples and Populations

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

September 8, 2011

Sex and Older Women

Example

Fertility declines in women as they age until ending at menopause. Younger women may become pregnant relatively easier than older pre-menopausal women. A hypothesis rooted in evolution and psychology states that as women age, they may experience increases in sexual motivation and seek sex more frequently to overcome decreasing fertility.

How can data be collected to examine this hypothesis?

The Scientific Literature

The abstract of a recent (2010) article in the journal *Personality and Individual Differences* titled Reproduction expediting: Sexual motivations, fantasies, and the ticking biological clock begins as follows:

> *Beginning in their late twenties, women face the unique adaptive problem of declining fertility eventually terminating at menopause. We hypothesize women have evolved a reproduction expediting psychological adaptation designed to capitalize on their remaining fertility.*

The Scientific Literature (cont.)

The abstract continues to report the results as follows:

> *The present study tested predictions based on this hypothesis—these women will experience increased sexual motivations and sexual behaviors compared to women not facing a similar fertility decline. Results from college and community samples ($N = 827$) indicated women with declining fertility think more about sex, have more frequent and intense sexual fantasies, are more willing to engage in sexual intercourse, and report actually engaging in sexual intercourse more frequently than women of other age groups. These findings suggest women's "biological clock" may function to shift psychological motivations and actual behaviors to facilitate utilizing remaining fertility.*

The Popular Literature

- *Time* magazine wrote about the scientific publication with an article titled The Science of Cougar Sex: Why Older Women Lust.
- Somewhat surprisingly, the *Time* article is more careful than the article in the primary literature in discussion of the importance in how the data is collected when interpreting the results.
- Much less surprisingly, the *Personality and Individual Differences* article does not use the term *cougar*.

The Big Picture

- Many of the statistical methods we will encounter this semester are based on the premise that the data we have at hand (*the sample*) is representative of some larger group (*the population*).
- We often wish to make *statistical inferences* about one or more populations on the basis of sampled data.
- Statistical methods often assume that samples are *randomly selected* from populations of interest, although in practice, this is frequently not the case.
- We need to understand:
  - how to take random samples; and
  - how non-random sampling may affect inferences.

Samples and Populations

Definition

A *population* is all the individuals or units of interest; typically, there is not available data for almost all individuals in a population.

Definition

A *sample* is a subset of the individuals in a population; there is typically data available for individuals in samples.

Samples and Populations (cont.)

Examples:

- In the cow data set:
  - the sample is the 50 cows;
  - the population is cows of the same breed on dairy farms.
- In the plantation example:
  - the sample is the three sites where data was collected;
  - the population is all plantations in Costa Rica where one might consider restoration to native forest.
- In the older women sex example:
  - the sample is the 827 women included in the study;
  - the population is American women aged 18+.

Properties of Representative Samples

- Estimates calculated from sample data are often used to make inferences about populations.
- If a sample is *representative* of a population, then *statistics* calculated from sample data will be close to corresponding values from the population.
- Samples contain less information than full populations, so estimates from samples about population quantities always involve some uncertainty.
- Random sampling, in which every potential sample of a given size has the same chance of being selected, is the best way to obtain a representative sample.
- However, it often impossible or impractical to obtain a random sample.
- Nevertheless, we often will make calculations for statistical inference *as if a sample was selected at random, even when this is not the case*.
- Thus, it is important to understand both how to conduct a random sample in practice and the properties of random samples.

Random Sampling

Definition

A *simple random sample* is a sample chosen in such a manner that each possible sample of the same size has the same chance of being selected.

- In a simple random sample, all individuals are *equally likely* to be included in the sample.
- The converse, however, is untrue: Consider sampling either all five men or all five women with equal probability from a population with ten people. Each person has a 50% chance of being included, but any sample with a mix of men and women has no probability of being chosen while the two samples of all individuals of the same sex each have probability one half of being selected.

Random Sampling

- Estimates from simple random samples are *unbiased*; there is no systematic discrepancy between sample estimates and corresponding population values.
- For random samples, larger samples are typically more accurate; the chance difference between sample estimates and population values is smaller (on average) for larger samples (but not necessarily for specific samples).
- While it is often impractical to take random samples from a population, it is commonly possible to assign individuals at random to treatment groups.
- It is important to distinguish between *randomness under control of the researcher* and *randomness assumed, but not under control*.

Samples of Convenience

- Researchers often (almost always?) sample individuals that are easily available rather than sampled from a formal random process.
  - Studies of dairy cows are typically performed on cows available in research herds, not from a random sample of the population of cows on farms.
  - Ecological studies are typically performed at sites accessible to a researcher, not from a random sample of all sites of potential interest.
  - Medical studies are typically performed on individuals in a particular region who volunteer to be part of the study.
  - Psychology studies are often performed on volunteers recruited from college campuses.
- It is vital to describe how individuals are sampled so that the potential biases in the sampling process can be considered.

Random Sampling

- Formal simple random sampling requires an accurate and complete list of members of the population.
- Such a list can be numbered from 1 to N.
- In principle, taking a random sample of size n from a population of size N is equivalent to placing the N labels in a hat, mixing, and selecting n labels at random.
- In practice, we use the computer.

Sampling in R

- The function `sample()` is used for random sampling in R.
- The first argument to `sample()` is either an array of the items to be sampled or the number of such items.
- The second argument is the sample size.
- Other optional arguments can allow for sampling with replacement or with nonuniform probabilities.

Example

- The text describes the Prospect Hill Tract of Harvard Forest which in 2001 included 5699 trees.
- Below are three separate samples of size 20, with the IDs sorted for convenience.

```
> sort(sample(5699, 20))

 [1]  135  557  666  806 1944 2018 2208 2682 2794 3034 3090
[12] 3713 3959 3993 4203 4650 5232 5281 5603 5660

> sort(sample(5699, 20))

 [1]  139  292  468  628 1896 1935 2073 2204 2451 2641 2932
[12] 3146 3379 3519 3965 4129 4499 5060 5270 5307

> sort(sample(5699, 20))

 [1]  384  677  770  956 1133 2861 2957 3099 3432 3523 3531
[12] 3548 4015 4789 4920 5228 5429 5478 5529 5554
```


Example

- It would typically be convenient to save the sampled values.

```
> my.sample = sort(sample(5699, 20))
> my.sample

 [1]   25  731 1593 1671 1733 2326 2895 3088 3139 3298 3465
[12] 3530 3696 3761 3892 3945 4466 4610 4677 5223
```

Return to the older women sex study

- The 827 sampled women are not randomly sampled from the population.
- Instead they are recruited form among students at the University of Texas and with an ad on the *Adult Section* from Craigslist.
- (The Adult Section no longer exists, but there is a Personals section for people seeking relationships.)
- A high percentage of the recruited women from the 18–26 year-old group were university students; a high percentage of the 27–45 year-old women were found from the ad on the Adult Section of Craigslist.
- If women on Craigslist differ from women in the population not on Craigslist in terms of sexual behavior, the results of the study would be potentially misleading.
- As the Adult Section of Craigslist was often used by individuals seeking others to engage in sex, it seems likely that the sampling process is biased and is likely to sample individuals who will not be representative of the population as a whole.

Quote from Time

To test this theory, Buss and his students asked 827 women to complete questionnaires about their sexual habits. And, indeed, they found that women who had passed their peak fertility years but not quite reached menopause were the most sexually active. This age group—27 through 45—reported having significantly more sex than the two other age groups in the study, 18 through 26 and 46 and up.

. . .

And yet there are a few flaws with the data in the new paper. Chiefly: some three-quarters of the participants in the study were recruited on Craigslist, a website where many go to seek hookups, meaning there was a self-selection problem with the sample. (The other participants were students at the University of Texas in Austin.)

Discussion about the Sex Study

- What is the difference between this study of sexual behavior of women and a study that examines, say, the response of protein or fat in milk from a study about additives to the diets of cattle?
- Both studies use *samples of convenience*.
- Why might one study be more trustworthy than the other in terms of the accuracy of generalizations to the larger populations?

Cautions

- Many statistical methods assume random sampling; however, it is often impractical to obtain random samples.
- Inferences to populations from nonrandom samples can be justified, but this depends on background information sufficient to determine that a sample is *representative*.
- Biases in sampling procedures can mislead.
- Conclusions from a study should be consistent with how the data was sampled.

What you should know

You should know:

- what a random sample is;
- how to distinguish between random and nonrandom samples;
- what types of bias can result from nonrandom sampling;
- how to use R to take a random sample from a finite population;
- that even when the actual sampling process is not random, calculations of probabilities *as if the sampling process actually was random* can be useful for statistical inference.