Stat 400, section 5.3-5.4 Introduction: Populations and Samples

notes by Tim Pilachowski

1. Background

In the real world, people often want (or need) to know information about some large group of people or items: a **population**. Educational researchers want to know about all of the students in a district, county, state, or nation. Political scientists want to know about all of the voters in an election. A manufacturer wants to know about the reliability of the machines coming off a production line. Advertisers want to know whether their wares appeal to potential customers. A government needs to know the rate of inflation and the percent of unemployment to predict expenditures and future economic trends.

The numbers associated with the desired information are called the **population parameters**. The two most common parameters measured are population mean [symbol μ] and population proportion (percent, relative frequency) [symbol p]. For each of these, there exists the population variance [σ^2] and standard deviation [σ].

Researchers usually don't have enough time and resources to determine actual population means and proportions. Instead, a **sample** will be taken and numbers (called **sample statistics**) calculated. Think of a sample as being a subset of the population.

The items included in the analysis and the people actually answering a survey are the sample, while the whole group of things or people we want information about is the population.

Example A: "National NAEP reports statistical information about student performance and factors related to educational performance for the nation and for specific student groups in the population (e.g. race/ethnicity, gender). It includes students drawn from both public and nonpublic (private) schools and reports results for student achievement at grades 4, 8, and 12." Source: http://nces.ed.gov/nationsreportcard/about/national.asp
Identify the population and the sample in this scenario.

Important note: Be careful to distinguish between
a) population/sample (in this case, students) and
b) data (in this case, scores on the NAEP).

The questions are always, "How well does a sample reflect the status or opinions of the population?" "How reliable are the sample statistics as indicators of the population parameters?" The answers depend on many variables, some easier to control than others.

Sources of Bias. One source of bias is carelessly prepared questions. For example, if a pollster asks, "Do you think our overworked, underpaid teachers deserve a raise?" the answers received are likely to reflect the slanted way in which the question was asked! Another danger is a poorly selected sample. If the person conducting the survey stands outside a shopping center at 2 pm on a weekday, the sample may not be representative of the entire population because people at work or people who shop elsewhere have no chance of being selected. Questions may be asked, either deliberately or unintentionally, in a manner that leads to particular conclusion. (In a courtroom this type of questioning is called "leading the witness".) Another possible source of error comes from respondents who misinterpret a question, either because the question was too complex or because the respondent is not knowledgeable (but maybe wants to appear to be so).

Random Sampling. To eliminate bias due to poor sample selection, statisticians insist on **random sampling**, in which *every member of the population is equally likely to be included in the sample*. For a small population, simple random sampling can be achieved by putting everyone's name on a slip of paper, placing all the slips in a large bin, mixing well, and selecting the sample.

Technically, each slip should be replaced before the next slip is drawn, so the probability doesn't change for subsequent draws. That is, the events should be independent, and the probability not conditional. Of course, the slips should be mixed before each draw, as well. In practice, most populations being tested are large enough that the conditional probabilities don't significantly alter the results.

Representative Samples. It is important to note that saying the sample is free of bias does not guarantee that it will perfectly represent the population. A random sample may, by chance variation, be very different from the population. However, the mathematics underlying statistics will allow us to specify how much a random sample is likely to vary and how often it will be extremely non-representative. Bias, on the other hand, affects the results in unknown ways that cannot be quantified or specified in any such manner.

2. Theory and Practice

When designing a survey/study/experiment, researchers will choose a sample size, n. Random variables $X_1, X_2, \ldots, X_n$, will be defined representing the potential outcomes of the experiment. Random variables must fulfill two (important) requirements:

1. The X_i's are independent random variables.
2. Every X_i has the same probability distribution.

Data is collected from the sample, i.e., the random variables $X_1, X_2, \ldots, X_n$ each receive values $x_1, x_2, \ldots, x_n$. These values are used to calculate sample statistics. The sample statistics we'll be most interested in are:

1. The sample total $T_0 = X_1 + X_2 + \ldots + X_n$.
2. The sample mean $\overline{X} = \frac{1}{n} * T_0$. (The calculated sample mean is symbolized by $\bar{x}$.)
3. The sample variance $S^2 = \frac{1}{n-1} * \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$. (The calculated sample variance is symbolized by s^2.)

3. Examples

Example A. NAEP defines a scale score as, "A score, derived from student responses to assessment items, that summarizes the overall level of performance attained by that student. While NAEP does not produce scale scores for individual students, NAEP does produce summary statistics describing scale scores for groups of students. NAEP subject area scales typically range from 0 to 500 (reading, mathematics, U.S. history, and geography) or from 0 to 300 (science, writing, and civics)."
http://nces.ed.gov/nationsreportcard/glossary.asp#scale_score

Suppose that a sample of 10 students has scale scores of 290, 340, 350, 400, 410, 440, 460, 460, 460, and 490 in the mathematics portion of the NAEP. Calculate the sample mean and sample standard deviation.

Example B. A news organization polls n voters and asks, "Do you intend to vote for incumbent Senator Phillip E. Buster in the upcoming election?" Pollsters record a value of 1 for a "Yes" answer and 0 for a "No" answer.

Example C. A factory needs to evaluate the lifetime of its production machinery. The lifetime of similar machinery has an exponential probability distribution.

Example D. McCormick Consumer testing asks participants to rate various characteristics of food items on a scale from 1 to 10. There is a baseline assumption that, if participants simply make random choices, the values have a uniform probability distribution.

4. Should you believe statistics?

In his 1924 *Autobiography*, Mark Twain writes, "Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies and statistics.' " This quote is often taken out of context and used in support of an assertion that statistics can never be relied upon or trusted.

However, the sardonic and satirical journalist, author and speaker was recognizing the persuasive power of authoritatively made numeric presentations to a largely innumerate public. He did not mean that statistics are worthless, but that they can be interpreted and manipulated to support very different arguments. Critical thinking skills are necessary to evaluate claims and decisions made based on statistical analysis.

One purpose of statistics is to go beyond simple statements of "right and wrong" and "agree or disagree" to analysis (using probabilities) which will either support or disprove particular statements. For example, parents may believe school A is better than school B—information about students performance, teacher reputation, and experience of other parents would be helpful in either validating or challenging their belief.

When it becomes time to interpret a statistical study, several things can go wrong. A researcher or reader might give a conclusion more weight than is warranted by the scope of the study. For example, a study of primary-age children is likely not applicable to middle school students. A statistical study taken out of context may lead to conclusions far afield from the original intent. It is also true that journals and news reports often present summarized results and conclusions without a full presentation of the data. A reader might never know how well data was collected and whether or not the author was seeking to justify an already-made conclusion.

Does all this mean that statistics should never be trusted? No, only that we, as citizens, voters and consumers, should exercise thinking skills in evaluating the merits of statistical analysis. A good measure of critical thinking goes a long way toward weeding out the "damned lies" from useful information.

The purpose of statistical analysis is to take an otherwise overwhelming amount of information and work with it to find patterns and relationships which can inform us about our world. Controlled experiments lessen the effects of unknown or non-measurable variables. The goal is to move beyond limited personal experience and guesswork to intelligent and justified analysis and conclusions. Peer review helps to ensure that rigorous standards of data collection and analysis are in place. The possibility of bias and error is thus lessened.

By the way, Mark Twin attributed the quote cited at the beginning of this section to Benjamin Disraeli (1804–1881), British statesman and author. However, the words have never been found among Disraeli's works; alternative attributions include the radical journalist and politician Henry Labouchère (1831-1912).

5. Homework exercise. In an effort to identify the status of salaries of women working full time in the United States, a political action group undertook several studies, one of which collected data seeking to correlate the age of a woman at the birth of her first child with her current annual salary. Data collected from one sample were as follows:

Age at birth of first child (years)	18	17	20	22	23	26	27	19
Current annual income ($*1000)	16	17	17	18	17	21	20	17

1a. Describe the *population* in this scenario.

1b. Describe the *sample* in this scenario.

2. What is the *frequency* of women who were age 24 at the birth of their first child?

3. What is the *relative frequency* of women who earned $20,000 or more?

4a. Find the (sample) mean of the *ages*.

4b. Find the (sample) variance of the *ages*.

5a. Find the (sample) mean of the *incomes*.

5b. Find the (sample) variance of the *incomes*.