CS 1173: Populations and samples

The scientific method

In most scientific experiments, scientists measure a small number of items (a **sample**) and then infer characteristics of the entire **population**. The population is usually huge and not completely accessible. For example, the Fisher iris data set contains a sample of 50 observations of sepal length of *Iris setosa*. What population does this sample represent? Statistical analysis assumes that the items of the sample are chosen randomly from the entire population. In reality, biologist Edgar Anderson took these iris measurements in the late 1920's on the Gaspé Peninsula in Quebec, Canada. Does this sample represent all members of *Iris setosa* or only those available at a specific time and place?

We face a similar question for the sleep diary data. What population does our sample represent: All people? All people in the US? All college students? All biology majors at UTSA? All CS 1173 students? CS 1173 students during the specific three weeks of the measurements? None of the above?

For analysis purposes, scientists usually assume that their measurements represent a random sample and use the sample to estimate characteristics such as the population mean and standard deviation. Separately, they analyze sources of bias and then make arguments about how broad a population the sample actually represents.

Two other factors come into play in this type of analysis: sample size and distribution. Generally, the larger the number of items in the sample, the more likely that estimations of the population based on the sample will be valid. Unfortunately, biology is a discipline plagued by small sample size due to expense, availability of samples, and difficulty of the experiments. In designing an experiment, scientists must trade off statistical validity against expense in terms of time and money to make the measurements.

A second factor in making estimations is the statistical distribution of the measured values. Most methods of estimating population characteristics are **parametric**. That is, the methods assume that the data follows a particular statistical distribution and use the characteristics of this distribution as a basis for making estimations of the population. The most widely-widely used methods of estimation assume that the data follows a normal (bell curve) distribution, but methods are available for other types of distributions. Unfortunately, real data seldom exactly follows any distribution represented by a simple formula. The researcher argues the validity of the results based on how closely the data follows the assumed distribution.

Summary of the procedure:

We can summarize the procedure represented by the classical scientific method as follows:

1. **Formulate hypotheses or conjectures about the population.** These hypotheses usually come from previous exploratory work, from theories, from computational models, or from the creative imagination of the scientist. Ideally, the scientist makes testable predictions based on these hypotheses.
2. **Design an experiment to measure characteristics of a sample of this population.** A first step is to make sure that the experiment actually measures quantities that shed light on the truth of the hypothesis. The actual design of the experiment (how many samples and of what type) can be quite complicated with a trade-off of economic factors and feasibility against validity of the results. Often a statistician assists in the design, particularly for experiments that are expensive and/or difficult. At this point, the scientist usually thinks carefully about the population and potential sources of bias.
3. **Perform the experiment.**

4. **Estimate the population parameters** based on measurements of the sample obtained by the experiment.
5. **Test the hypotheses and draw conclusions.**
6. **Verify that the sample was likely to have come from the assumed distribution.** Since real data seldom exactly follows a particular distribution, the scientist usually argues that the data is close enough for the results to be valid.
7. **Argue that the results are unbiased and represent the population as a whole.**

In reality, much scientific work is more exploratory in nature. That is, rather than verifying testable hypotheses, the scientist performs experiments to try to understand how things work. Data analysis and visualization play an important role in exploratory research, because the scientist uses the data as a guide in formulating testable hypotheses. When reading the scientific literature, you should be careful to distinguish between papers reporting exploratory work and more structured verification of hypotheses.

Estimating population characteristics based on samples

As described in the previous section, the classical scientific method involves making measurements on a small number of items (**the sample**) and inferring values representative of the entire collection (**the population**). The two most common population estimates are the population mean (denoted by μ) and the population standard deviation (denoted by σ). Note: If the population actually follows, a normal distribution (a bell curve), the mean and standard deviation completely determine population characteristics.

1. Calculating sample characteristics: Suppose $x = [x_1, x_2, x_3, ..., x_n]$ is a sample representing n measurements of something (e.g., the sepal lengths of a species of iris or the wake-up times of students in CS 1173). One point of confusion here is the word "sample" means a single value in normal English usage. In statistics "sample" means a group of measurements.

We can calculate **the actual mean** of x (denoted by $\bar{x}$) and **the actual standard deviation** of x or the **population standard deviation** (denoted by s) using the formulas:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

and

$$s = \sqrt{\frac{1}{n}\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Example 1: If `Y` is a MATLAB variable representing a sample of 20 values (e.g., a 20 × 1 vector), write a MATLAB expression for the sample mean.

Ans: `mean(Y)`

Example 2: If `Z` is a MATLAB variable representing 3 samples of 20 values each (e.g., a 3 × 20 array), write a MATLAB expression for the means of each of the three samples. How big is the result?

Ans: `mean(Z, 2)` The result is a column vector of size 3 × 1.

Note: Usually we put each sample in a separate column rather than in a separate row.

Example 3: If `W` is a MATLAB variable representing 4 samples of 25 values each (e.g., a 25 × 4 array), write a MATLAB expression to compute the standard deviation of each of the four samples. How big is the result?

Ans: `std(W, 1)` The result is a row vector of size 1 × 4.

Example 4: If `Z` is a MATLAB variable representing 3 samples of 20 values each (e.g., a 3 × 20 array), write a MATLAB expression to compute the standard deviation of each of the three samples. How big is the result?

Ans: `std(Z, 1, 2)` The result is a column vector of size 3 × 1.

2. Estimating the population mean: The most common method of estimating the population mean from a sample is to use the sample mean ($\bar{X}$) to estimate the population mean (μ). On charts or line graphs with error bars, the central point in each error bar usually corresponds to $\bar{X}$, the sample mean. The value of $\bar{X}$ is called a **point estimation** of μ because we are using a single value calculated from a sample to estimate the population parameter.

Example 5: If `Y` is a MATLAB variable representing a sample of 20 values (e.g., a 20 × 1 vector), write a MATLAB expression to estimate the underlying population mean.

Ans: `mean(Y)`

3. Estimating the population standard deviation: The obvious choice for estimating the population standard deviation is the sample standard deviation, s. Unfortunately, empirical studies show that this estimate turns out to be too low for small samples (i.e., it is biased toward low values). A better estimate (Bessel's correction), sometimes called the **unbiased estimator of population standard deviation** or the **sample standard deviation**, is given by $\hat{s}$:

$$\hat{s} = \sqrt{\frac{1}{(n-1)}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

Example 6: If `Y` is a MATLAB variable representing a sample of 20 values (e.g., a 20 × 1 vector), write a MATLAB expression to estimate the underlying population standard deviation.

Ans: `std(Y)`

Note: This is the unbiased estimator $\hat{s}$ of the underlying population mean. The actual standard deviation of the sample is `std(Y, 1).`

Example 7: If `Z` is a MATLAB variable representing 3 samples of 20 values each (e.g., a 3 × 20 array), write a MATLAB expression to estimate the underlying population standard deviations for each of the three samples.

Ans: `std(Z, 0, 2)`

4. How accurate is the sample mean as an estimate of the true population mean? The standard error of the mean (*SEM*) defined by:

$$standard\ error\ of\ the\ mean\ (SEM) = \frac{\hat{\sigma}}{\sqrt{n}}$$

provides a way to characterize the accuracy. Here $\hat{\sigma}$ is an estimate of the population standard deviation (e.g., $\hat{s}$). You can interpret *SEM* as the standard deviation of the sample means from the true population mean. Many figures in biological papers show the sample mean with *SEM* error bars. In these cases, the sample mean estimates the underlying population mean and the *SEM* depicts the likely error.

5. The SEM 95% confidence interval as an interval estimate of the population mean:

An alternative to estimating the population mean by a single point (e.g., the sample mean) is to calculate an interval in which the population is likely to fall. A common method uses *SEM* to produce a 95% confidence interval for the true population mean:

$$[\bar{x} - 1.96\,SEM, \bar{x} + 1.96\,SEM]$$

We would like to say that we are 95% certain that the true mean of the population is in the above interval. However, the actual interpretation is "**we are 95% certain that the interval that we produce in this way contains the actual population mean**". This expression for confidence interval assumes the population is normally distributed. It also doesn't work well for small samples sizes, but that's a story for your statistics courses to address.

Example 8: If `Y` is a MATLAB variable representing a sample of 20 values (e.g., a 20 × 1 vector), write a MATLAB expression for the standard error of the mean (*SEM*).

Ans: `std(Y)./sqrt(length(y))`

Note: This expression uses the unbiased estimator for the standard deviation estimate.

Example 9: `W` is a MATLAB variable representing 4 samples of 25 values each (e.g., a 25 × 4 array). Calculate the mean and SEM for the columns of `W`. Use this result to plot the 95% confidence limits for the estimate of the underlying population mean:

Ans:

```
theMean = mean(Y);
n = size(Y, 1);
theSEM = std(Y) ./ sqrt(n);
confInt =[theMean - 1.96*theSEM, theMean + 1.96*theSEM];
errorbar(theMean, confInt);
```