Evaluation: since CSC239 is primarily a statistics course, almost every task in the class is dedicated to the statistical analysis of data.

Automation: because CSC239 relies on MicroSoft Excel for computing results, most of the course is also dedicated to mapping statistical concepts to specific formulae and worksheets.

Recollection: much of statistics relies on summarizing data in a form that can be used for future calculations. In particular, encoding a data set as its mean and standard deviation provides a foundation for all the more advanced statistics of CSC239.

Computation: many of the analyses in CSC239 are amenable to "mini-algorithms" in which the students start with a mean and standard deviation (as well as one or two other input parameters) and proceed through a number of steps to derive a final analysis.

Constructing a histogram

The construction of a histogram is an important first step in any statistical analysis. The visual appearance of the histogram can be used to determine if there are outliers and what the appropriate statistics are for reporting center and spread. For more complex statistical analysis, the shape of the histogram can be used to determine what analyses are appropriate and which are not. However, the creation of a histogram is not a simple process and benefits greatly from *computation*.

In CSC239, students are shown the algorithm for constructing a histogram:

1) Determine the minimum and maximum of the data
2) Create a number of bins for aggregating data
    a. Bin range must equal or exceed the data range
    b. Bins must not be too few or too many
3) Create a distribution table for the data based on the bins
4) Graph the distribution table
    a. If the resulting histogram is too spiky, return to step 2 and create fewer bins
    b. If the resulting histogram is too shapeless, return to step 2 and create more bins

There are computer programs that completely automate the construction of a histogram. However, since the histogram is a visual statistic, it is best constructed using human automation; in particular the choice of the number of bins is critical in generating a histogram that best provides visual information about the distribution of the underlying data.

The actual construction of a histogram may also be considered a *recollection* task: the encoding of a set of numbers into a visual representation that can be used to organize and recommend further analysis. In CSC239, students are shown that the construction of a histogram is critical in determining whether the mean and standard deviation or the mode and quartiles are better for summarizing the center and spread of the data.

Computing measures of center and spread

One of the fundamental problems of dealing with data is in extracting meaning from the data. At a very basic level, it can be extremely useful to calculate the center and spread of a data set. This is an example of *recollection*, in which an entire set of numbers is encoded in two (mean and standard deviation) or three (first quartile, median and third quartile) numbers. In particular, the mean and standard deviation provide efficient summaries for the calculation of further statistical values, including correlations, confidence intervals, and tests of statistical significance. Computing the center and spread is also an example of *evaluation*.

Using the normal approximation

The normal approximation is a simple and easy way to model data, requiring knowledge only of the mean and standard deviation of the data, both easily computed using simple arithmetic. In CSC239, students discuss a set of questions often asked about data, and generate a worksheet designed to answer those questions. The questions include

1) What percentage of the data is above a certain value *x*?
2) What percentage of the data is below a certain value *x*?
3) What percentage of the data is between two values *a* and *b*?
4) What value is the smallest value in the top *p* percent of the data?
5) What value is the largest value in the bottom *p* percent of the data?

Construction of the worksheet involves both *evaluation* and *automation*, as students create a worksheet that uses formulae to generate answers to these questions based on a normal approximation of the underlying data.

Computing confidence intervals

Two main tasks of statistics are summary and prediction. The computation of a confidence interval for some unknown parameter of a population based on a sample is one of the most basic predictive tasks in statistics. The task of creating confidence intervals is one of *evaluation*, *automation* and *computation*.

In particular, students are shown that given the sample mean, standard deviation and size of a sample set, they can derive a prediction for the unknown population mean given a certain confidence level. The task is broken down into simple steps to arrive at the solution:

1) Use the standard deviation and sample size to calculate the standard error
2) Use the confidence level and sample size to calculate a critical value from the t-distribution
3) Calculate the lower level of the confidence interval
4) Calculate the upper level of the confidence interval

Students create a worksheet using Excel formulae to accept the confidence level as a parameter and compute all the necessary intermediate values to arrive at a final prediction of the population mean in the form of a confidence interval.

Testing hypotheses

Hypothesis testing can be used to verify a claim about an unknown population parameter in light of known sample values. In CSC239, students are asked to verify or refute claims about a population average based on sample statistics. The task of conducting a hypothesis test is one of *evaluation*, *automation* and *computation*.

In particular, students are given a problem statement and an associated question about the problem, and asked to generate an answer to the question:

1) Generate a null hypothesis
2) Generate an alternative hypothesis
3) Set a significance level α
4) Use the standard deviation and sample size to calculate the standard error
5) Use the mean, null value and standard deviation, calculate sample test statistic, *t**
6) Calculate the *p*-value from the t-distribution using *t**
7) Compare the *p*-value and α to reject or not reject the null hypothesis

Students create a worksheet using Excel formulae to accept a null value as a parameter and compute all the necessary intermediate values to arrive at a decision regarding the hypotheses.