# Week 2

Refer to the course map

This week we start with the simplest approach to data: find a way to summarize it, **without pretending to draw any conclusion from it.**. This point needs to be emphasized over and over: observing what a group of people think about an issue, **says absolutely nothing about what the rest of the world thinks, unless several additional steps have been taken**, including the construction of an appropriate mathematical model, and a rigorous verification of what method was used to pick these particular people out of the full population. How to take these steps will be the object of the rest of the course.

As soon as we collect more than a handful of data items, it becomes difficult to grasp them, and there are a number of techniques allowing us to summarize the data to make it more easily understandable. **Almost all these tools will result in information loss, which is something we must also be aware of.** Also, many of these tools require you to make choices in how to "compress" (so to speak) the data, and different choices may sometimes result in different "looks" of the summary. In such cases, it is always a good idea to test a few different options, before committing yourself (for example, this is often an issue when preparing a *histogram*)

## Objectives for this week

* Apply tabular, graphical, and numerical tools for summarizing observations: given a set of observations, display various summarizing methods (histograms, box plots, and so on) as introduced in the text. It's important that choices that go into the method (e.g., size of histogram bins) be explicit, comparing the outcome of different choices. Also, the extent of information loss should be highlighted (e.g., how a histogram might hide the fine structure of the data). Trying various alternatives and comparing the results
* Distinguish the difference between complete observations and sampling. The tools we have at this point are meant to describe *the data that we have, and cannot inform us about questions relating to unobserved items*. Sampling, as discussed in the next couple of weeks, requires some assumed connection between the observations we have and the items we did not observe. The theory that allows us to establish this connection is the purpose of next week's study.

## Reading Material

Refer to the file descriptive.html . This file is a starting point that links to additional files, and to chapters from the Open Stat Book. We discuss the main tools, and how they may be used.

## Assignments

Please, if you haven't done so already, download the (simulated) data files listed in the specific page. These files (choose the format that best suits your software) will be with us all quarter long, as we learn new tools, and apply them to these fictitious data.

This week, you are asked to apply as many of the summarizing tools introduced in the text to these data sets as you can.In particular,

* Set up a summary, in the form of a *frequency table* (numerical and graphic) for the *categorical data set* ("favorite ice cream flavors")
* Set up a similar summary for the *yes-no* data: each data point is one of two values, so adding them up gives the number of "yes". The remaining are, obviously, "no". Additionally, since these are numerical data, you should compute
  + the average (the *sample mean*),
  + and, even if it is not very useful for this type of data, the *population*, and *sample variance*.
* Construct graphical and numerical summaries for the two *normal data* files. That is, try to build histograms (experiment with more than one bin size), box plots, as well as calculate the standard summaries (sample mean, population/sample variance, median, first and third quartile, and more, if you can)
* Repeat, **optionally**, for the last data set, labeled as *exponential*.

**Please, turn in your work on these data sets by Monday, Week 3**

### Quizzes

As discussed in the Syllabus, there is a collection of 80 quizzes available for you to take on the [WAMAP site](http://www.wamap.org). You can take these quizzes any number of times, and every time, you should get a slightly different problem or the same problem, with different numbers (depending on whether you logged out or not in between). While the score you reach will not be used directly towards your grade, **taking these quizzes, and scoring well, possibly after a few attempts, will count as participation, and thus contribute to your grade**

**This week's material is covered in the first 15 quizzes**