Lecture 0: Background

Rafael A. Irizarry and Hector Corrada Bravo

January 2010

The purpose of this class is for you to learn machine learning techniques commonly used in data analysis. By the end of the term, you should be able to read papers that used these methods critically and analyze data using them.

When using any of these tools we will be we will be asking ourselves if our findings are "statistically significant". For example, if we make use of a particular classification algorithm and find that we can predict the outcome of 7 out of our 10 cases, how can we determine if this could have happened by chance alone? To be able to answer these questions, we need to understand some basic probabilistic and statistical principles. Today we will review some of these principles.

Probability

If I toss a coin, what is the chance it lands heads?

In this class we will sometimes be using notation like this:

Let X be a random variable that takes values 0 (tails) or 1 (heads) such that

$$Pr(X = 1) = 1/2$$

For die, we would write:

$$Pr(X = k) = 1/6, k = 1, \dots, 6.$$

We will refer to these as **probability distributions**.

More "complicated" distributions can be defined, for example, by considering the random variable

$$Y = \sum_{i=1}^{N} X_i$$

where the X_i's, $i = 1, \dots, N$ are independent tosses of the same die (or coin).

What are possible values of Y? What is the distribution of Y? What does independent mean?

Populations, LLN and CLT

In science, randomness usually comes from either random sampling or randomization.

Side note: What about observational studies?

How does the above relate to populations?

The coin toss can be related to a very large population where each subject is either, say, a democrat (heads) or a republican (tails). If half are democrats and half are republican, then if we pick a person at random, it's just like a coin toss.

If dems are 1, and reps are 0, what is the **population average** $\bar{x}$?

If I take a random sample with replacement (a poll) of $N = 10$ subjects, what is the distribution of the **sample average**?

What happens to the difference between the sample average and the population average as the sample size gets bigger?

Why is the sample average $\bar{X}$ a random variable? What about the distribution? Is the population average a random variable? What does the law of large numbers (LLN) say? What does the central limit theorem (CLT) say?

Inference

How does this all relate to scientific problems? Many times in science we can model the process producing data with a stochastic (probabilistic) model where parameters (such as population averages) are unkowns. We then make **inferences** based on the data.

For example, in the dems and reps problem we may not know the percentages of 1s and 0s. To find out, we take a random sample, and construct estimates (the sample average), confidence intervals and p-values.

How do we construct a confidence interval for the percentage of democrats? What would be an interesting null hypothesis in this case? How would we construct a p-value for this null hypothesis?

For continuous data, this is all pretty much the same. For example, we may want to know if the average weight of Baltimore women is over some recommended ideal weight.

Note: In this case, we could use a t-test if the sample is small.

This inferential approach is used in any situation where a population average is of interest and we can only obtain a random sample. It is also used when randomization is used.