Section 5.1. A basic task of statistics is to obtain knowledge of a population from observations of samples. To make the process rigorous, we create a probabilistic model of a random sample. A random sample of size n is modeled by a collection of random variables $X_1, \dots, X_n$ that are mutually independent an all have the same distribution, f. Such a collection is said to be *iid*—independent and identically distributed.

Examples:

1. We take n beads from a large jar and observe whether the color is red or not. Let $X_i = 0$ if the i^{th} bead is not red and let $X_i = 1$ if it is red. In a random sample, we expect the colors of the beads other than the i^{th} have no influence on the color of the i^{th}, and we expect the probability of being red to be the same for all i. (If we drew beads *without* replacement from a *small* jar, then we would violate the assumptions badly.)
2. For each possible height range, there is a definite probability of a random LSU student being in that range. Suppose we took all the LSU student ID numbers, and randomly selected a number from the list. Then repeat 100 times. If X_i represents the height of the i^{th} student selected, then $X_1, \dots, X_{100}$ represents the generic random sample.
3. We select n circuit boards at random from a production line, and then run them until failure. Set X_i to be the number of hours that the i^{th} circuit board lasts. (If the manufacturing process is exactly the same at all times, and there is no relationship of any kind between the particular constitution of one board and the constitution of any other, then this is a random sample.)

By independence, the joint *pmf* or *pdf* of the variables in a sample is:

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i).$$

In many applications, we will know that the distribution of X_i belongs to a family of distributions that are determined by a parameter, θ. Then we write $f(x|\theta)$ for the common distribution of the X_i. In the typical statistical problem, we do not know θ but wish to gain information about it from the sample.

Section 5.2. A *statistic* is a function $Y = g(X_1, \dots, X_n)$ on the set of all possible values of the sample, $(X_1, \dots, X_n)$. The distribution of a statistic Y is called a *sampling distribution*. Some important statistics are:

i) the *sample mean*: $\overline{X} := \frac{1}{n}(X_1 + \cdots + X_n)$;
ii) the *sample variance*: $S^2 = \frac{1}{n-1}\big((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\big)$;
iii) the *sample standard deviation*: $S := \sqrt{S^2}$.

Under assumptions that we will describe later, these statistic provide estimates for the expected value, the variance and the standard deviation of X itself.