7.1 Statistical Inference

Models and Parameters and Statistics

Data comes in many forms. Often, it consists of a list of numeric or character values. But it can also be organized more rigidly, as a matrix or array. Or it can present itself non-numerically, as a function, graph, or image.

A model is a structure for data production:

Def: We are given a random experiment with sample space S and a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ whose components take values in a space D (often $D = \mathbb{R}$). For an outcome $\omega \in$, we refer to $\mathbf{X}(\omega)$ as the **observations** or **data**.

Ex 1: Suppose a class of 25 students contains a mix of 1st, 2nd, 3rd, and 4th year students. We are interested in sampling students in the class to learn more about the distribution of years. To do so, we might select two students uniformly and independently at random (there is a chance that both individuals selected are the same student) and inquire about each student's year in school. Let X_1 denote the year of the first student, and let X_2 denote the year of the second student. Since student selection was random, then X_1 and X_2 are random variables. Unlike many variables we have seen in the past, the support set D of these variables is not a subset of the real numbers, but rather the list $D = \{1\text{st}, 2\text{nd}, 3\text{rd}, 4\text{th}\}$.

What is the sample space for this experiment? Many different sample spaces are possible. But for us, we need a sample space upon which we can simultaneously define both X_1 and X_2. That is, since X_1 and X_2 are by definition functions from S to D, then for any specific $\omega \in S$, we need $X_1(\omega)$ and $X_2(\omega)$ to represent the year in school for independent students.

As a first (but incorrect) attempt, we might choose S to be the list of students in the class. And then say that $X_1(\omega)$ and $X_2(\omega)$ represent the years in school of student ω. But this is a problem, since for every $\omega \in S$ (say, $\omega =$ Jonathan), then $X_1(\omega) = X_2(\omega)$, since both variables record the year in school of student ω, which wouldn't represent the desired phenomon (we want to allow X_1 and X_2 to possible take different values from each other).

So a second (and valid) choice might be to say that S consists of ordered pairs of students; each element $\omega \in S$ is a vector of length two, $\omega = (\omega_1, \omega_2)$. The first coordinate ω_1 of the vector is the name of one student, while the second coordinate ω_2 is the name of a second (but possibly the same) student. We can define the variables X_1 and X_2 on each $\omega = (\omega_1, \omega_2) \in S$ by

$$X_1(\omega) = \omega_1 \qquad X_2(\omega) = \omega_2$$

That is, for any pair of students $\omega = (\omega_1, \omega_2)$, then $X_1(\omega)$ is the year in school of the first student, and $X_2(\omega)$ is the year in school of the second student.

This isn't the only sample space possible. We could instead of have said that S represented triplets of students, and then said that X_1 is the year of the first student in the triplet, that X_2 is the year of the second triplet, and then ignore entirely the year of the third student in the triplet.

Because we are primarily interested in the distribution of random variables, and because many different sample spaces can produce the same distribution for the random variables, we will not often focus explicitly definining the sample space. But sometimes, it will be important to be explicit, and so it's important to investigate how we might construct appropriate sample spaces.

Def: A **model** $\mathcal{P}$ is a proposed list of joint distributions for which the vector of samples $\mathbf{X}$ could belong. The set Ω of all possible values of the parameter is called teh **parameter space**.

Ex 2: A bag contains 100 red and blue tickets. Let p denote the proportion of red tickets in the bag. Suppose we draw a number of tickets from the bag, with replacement. Let X_1 represent the number of red tickets in the first draw, let X_2 represent the number of red tickets in the first two draws, let X_3 represent the number of red tickets in the first three draws, etc. Note that the X_i are not independent (if $X_1 = 1$, then $X_2 \geq 1$.)

The implied model for the distribution of X_i for this experiment is $\mathcal{P} = \{\text{Bin}(i, p)\}$; if we draw from the bag i times, and each draw independently has probability p of being a red ticket, then the number of red tickets in i draws is $\text{Bin}(i, p)$. Note that this model consists of many different distributions (one for each different value of p). In this case, since p is the proportion

of red tickets in the bag, and since this proportion must be a fraction out of 100, then the parameter space is

$$\Omega = \left\{ \frac{0}{100}, \frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}, \frac{100}{100} \right\}$$

Statisticians create models and parameterizations. But the values of parameters will almost always be unknowable. The aim of Statistics is to use data inductively to narrow down the value of the parameter.

Def: A **statistic** T is a function from a random vector $\mathbf{X}$ on a sample space to one (or possible more than one) real number.

Ex 3: In the ticket box example, the number of red tickets X_i in the first i draws is a statistic. But so is the random variable $T(X_i)$ which takes the value 1 if the first i tickets contains at least 1 red ticket, and 0 otherwise.