Lecture 20 Random Samples

One of the most important concepts in statistics is that of a “random sample”. The definition of a random sample is rather abstract. However it is critical to understand the idea behind the definition, so we will spend an entire lecture motivating the definition we will do this by giving three motivating examples: polling for elections, testing the lifetime of a Gateway computer, and picking a sequence of random numbers.

First Motivating Example

We recall that a random variable X is a Bernoulli random variable if X takes exactly two values 0 and 1 such that

$$
\begin{aligned}
P(X = 1) &= p \\
P(X = 0) &= q \qquad q = 1 - p
\end{aligned}
$$

In this case we write $X \sim Bin(1, p)$ (the Bernoulli distribution is the special case of the binomial distribution where $n = 1$).

We define a Bernoulli random variable X_{election} as follows.

Choose a random voter in the U.S. Ask him (her) if he (she) intends to vote for Trump in the next election. Record 1 if yes and 0 if no. So X_{election} takes values 0 and 1 with definite (but unknown to us) probabilities q and p.

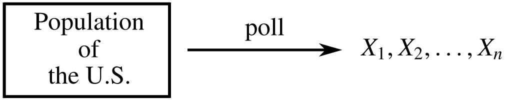
The $ 64,000 question

What is p?

How do you answer this question? Take a poll - in the language of statistics we say one is “taking a sample from a $Bin(1, p)$ distribution where p is unknown.” If we poll n people we arrive at a sequence of 0’s and 1’s

$$x_1, x_2, \ldots, x_n$$

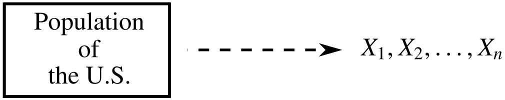
We can represent this schematically by



The X_i’s here should be lower case.

We think of $x_1, x_2, \ldots, x_n$ as the results *after the poll is taken*. We now introduce **random variables** $X_1, X_2, \ldots, X_n$ representing the potential outcomes **before the poll is taken** - we assume we have decided how many people we will talk to and how we are going to choose them.

Thus taking a poll assigns definite values $x_1, x_2, \ldots, x_n$ to the **random variables** $X_1, X_2, \ldots, X_n$. We may schematically represent the situation before the poll is taken by



Population of the U.S. $\dashrightarrow$ $X_1, X_2, \ldots, X_n$

The dotted arrow means we have not yet performed the poll.

It is critical to observe that $X_1, X_2, \ldots, X_n$ are *random* variables ($x_1, x_2, \ldots, x_n$ are ordinary i.e. numerical variables). The X_i's take values 0 and 1 with probabilities q and p respectively. So the $X_1's$ have the same probability distribution as the "underlying" (i.e. the distribution we are sampling from) random variable X_{election}. The random variables $X_1, \ldots, X_n$ will be independent if the poll is constructed properly. Hence, the random variables $X_1, X_2, \ldots, X_n$ are independent and "identically distributed." We say $X_1, X_2, \ldots, X_n$ is a random sample from a $\mathrm{Bin}(1, p)$ distribution.

We conclude this example with a formal mathematical construction of $X_1, X_2, \dots, X_n$. The sample space S of the above poll (“experiment”) is the set of all n-tuples $(x_1, x_2, \dots, x_n)$ of 0’s and 1’s. It is the same as the sample space for n flips of a weighted coin. There is a probability measure P defined on S. For example,

$$P(0, 0, 0) = q^n$$

The random variables $X_1, X_2, \dots, X_n$ are defined to be functions on S defined by

$$X_i(x_1, \dots, x_n) = X_i$$

So they are random variables - a random variable is a function on a probability space, that is a set S with equipped with a probability measure P.

Second Motivating Example

Suppose now we wish to study the expected life of a Gateway (a computer company which I think is no longer in business)computer so in this case we would be studying the random variable X_{Gateway} which is defined as follows:

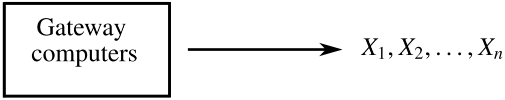
$(X_{\text{Gateway}} = t)$ means that a randomly selected Gateway computer fails at time t.

A good model for the distribution of X_{Gateway} is an exponential distribution with a definite but unknown mean $\mu = \dfrac{1}{\lambda}$.

The new $ 64,000 Question

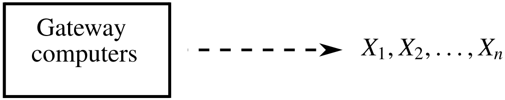
What is μ?

To answer this question,we obtain a number of
Gateway computers and run them until they break down and record these results. We may represent the results schematically by


Gateway computers
$X_1, X_2, \ldots, X_n$

The X_i's should be lower case.

Once again, we introduce random variables $X_1, X_2, \ldots, X_n$, after we have decided how many computers we are going to look at etc, but before we actually test the computers. So schematically we have the "before picture".


Gateway computers
$X_1, X_2, \ldots, X_n$

Mathematically testing the n computers amounts to assigning definite definite numerical values (the failure times)$x_1, x_2, \ldots, x_n$ to the random variables $X_1, X_2, \ldots, X_n$.
Hence, $X_1, X_2, \ldots, X_n$ are random variables with the same probability distribution as the underlying random variable X_{Gateway}.

Assuming that our test is correctly designed, $X_1, X_2, \ldots, X_n$ will be independent so they are identically distributed independent random variables,this will later be the definition of a random sample. So we say $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter.

Once again we have a formal mathematical construction. The sample space S of the experiment is now the set of all n-tuples $(x_1, x_2, \ldots, x_n)$ of positive real numbers, the possible break-down times for the n computers, to be tested. S is a probability space (but not discrete). We define the random variables $X_1, X_2, \ldots, X_n$ is as functions on S as before:

$$X_i(x_1, \ldots, x_n) = x_i, 1 \le i \le n.$$

Third Motivating Example

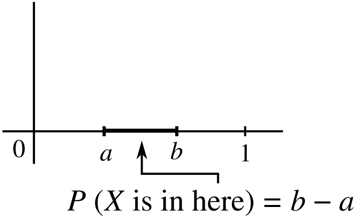
Our third motivating example will be the experiment of "choosing n random numbers from the interval $[0, 1]$". We have seen that a good model for "choosing a random number from $[0, 1]$" is the uniform distribution $U(0, 1)$. Precisely we make $[0, 1]$ into a probability space by defining a probability measure P on $[0, 1]$ by the formula

$$P(a \leq X \leq b) = b - a$$

(assuming $0 \leq a \leq b \leq 1$.

We then define a random variable (function) X on $[0, 1]$ by defining X to be the identity function I. So we think of evaluating I on an element of $[0, 1]$ as selecting a random number.

We may represent the probability space $[0, 1], P$ by



After we choose n random numbers using some procedure for producing random numbers, we obtain n real numbers $x_1, x_2, \ldots, x_n$ in $[0, 1]$.

$$\boxed{[0, 1]} \xrightarrow{\text{Make choices}} X_1, X_2, \ldots, X_n$$

The X_i's should be lower case x_i's.

Before we make the choices we have **random** variables $X_1, X_2, \ldots, X_n$ representing the first, second, ..., n-th choice. Schematically we have

$$\boxed{[0, 1]} \dashrightarrow X_1, X_2, \ldots, X_n$$

The sample space S of all possible choices of n random numbers is given by

$$S = \{(x_1, x_2, \ldots, x_n) : x_i \in [0, 1]\}$$

We have i functions $X_1, \ldots, X_n$ defined by $X_i : S \to [0, 1]$ where $X_i(x_1, x_2, \ldots, x_n) = x_i =$ "the i-th choice" so X_i is a $U(0, 1)$-random variable. We note that $X_1, X_2, \ldots, X_n$ all have $U(0, 1)$-distribution and are all independent.

The definition of a random sample

Hopefully, with the three basic examples we have just discussed we have motivated:

Definition

A random sample of size n is a sequence $X_1, X_2, \ldots, X_n$ of random variables such that

(i) *$X_1, X_2, \ldots, X_n$ are independent*
AND

(ii) *$X_1, X_2, \ldots, X_n$ all have the same probability distribution i.e. are ""identically distributed" often abbreviated to iid.*

The probability distribution common to the X_i's will be called the "underlying distribution"- it is the one we are sampling from.

Now we have a second fundamental definition.

Definition

A statistic is a random variable that is a function $h(X_1, X_2, \ldots, X_n)$ of $X_1, X_2, \ldots, X_n$.

Three very important statistics

The following statistics will be very important to us.

(1) The sample total T_0 defined by

$$T_0 = X_1 + X_2 + \ldots + X_n$$

(2) The sample mean

$$\overline{X} = \frac{1}{n} T_o = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

(3) The sample variance

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{1} (X_i - \overline{X}) \right)^2$$