The definition of a Random Sample

August 29, 2005

1 The Introduction

One of the most important concepts in statistics is that of a "random sammple". The definition of a random sample is rather abstract. However it is important to understand the idea behind the definition so we will spend an entire lecture motivating this definition. We will do this by giving three motivating examples, polling for an election, testing Dell computers and picking a sequence of random numbers. We conclude this introduction by giving the formal mathematical definition which we will try to motivate in the following subsections

Definition 1 (Definition of a Random Sample). *A random sample of size n is an n-tuple of identically-distributed independent random variables.*

2 First Motivating Example

We assume that in the 2008 presidential election Hillary Clinton will be running against Condoleeza Rice. We define a Bernoulli random variable $X_{election}$ as follows. Choose a random voter in the US. Ask him (her) if he (she) intends to vote for Hillary Clinton in the next presidential election. Record 1 if yes and 0 if no. So $X_{election}$ takes values 1 and 0 with some definite (but unknown to us) probabilities p and $q = 1 - p$.

The $ 64, 000 dollar question

What is p?

How do you answer this question - take a poll. In the language of statistics we say one " takes a sample from a Bernoulli distribution with unknown success probability p". If we poll n people we arrive at a sequence $x_1, x_2, \cdots, x_n$ of 0's (Condoleeza) and 1's (Hillary). It is important to note that the above sequence records the results *after the poll is taken.* We now introduce random variables $X_1, X_2, \cdots, X_n$ representing the potential outcomes *before the poll is taken* - we assume that we have decided how many people (namely n) we are going to poll, we just haven't talked to them yet. Thus taking a poll assigns numerical values $x_1, x_2, \cdots, x_n$ to the random variables $X_1, X_2, \cdots, X_n$.

It is critical to observe that $X_1, X_2, \cdots, X_n$ are *random* variables (and $x_1, x_2, \cdots, x_n$ are *ordinary* variables). The random variables $X_i, 1 \leq i \leq n$ take values 0 and 1 with probabilities p and q respectively (because they represent results that will come from talking "random chosen American voters") or put more formally, they are drawn from the sample space of $X_{election}$. So all the X_i's *have the same probability distribution.* Note also if the poll is constructed properly that the random variables $X_i, 1 \leq i \leq n$ are *independent.* So $X_1, X_2, \cdots, X_n$ is an n-tuple of identically-distributed (with Bernoulli distribution with success probability p) independent random variables. So we have arrived at the formal definition of the Introduction.

3 Second Motivating Example

Suppose now that we work for Consumer Reports and we wish to estimate the expected lifetime of a line of computers made by Dell. So in this case we would be sampling from the sample space of a random variable X_{Dell} which is defined as follows

$(X_{Dell} = t)$ means a randomly selected Dell computer breaks down at time t.

A good model for X_{Dell} is an exponential random variable with mean $\mu = 1/\lambda$.

The new \$ 64, 000 dollar question

What is μ, or, equivalently, what is λ?

How does Consumer Reports answer this question. They obtain n Dell computers (of the kind they are studying) and run them until they break down. They record the breakdown times as $x_1, x_2, \cdots, x_n$. This sequence is now a sequence of positive real numbers (not just zeroes and ones) and represents a sample from an exponential distribution (namely X_{Dell}) with parameter λ. Once again we introduce random variables $X_1, X_2, \cdots, X_n$ which represent potential outcomes *after we take the sample.* Testing n computers assigns numerical values $x_1, x_2, \cdots, x_n$ to the random variables $X_1, X_2, \cdots, X_n$. Again, $X_1, X_2, \cdots, X_n$ are random variables with the same probability distribution (exponential with parameter λ) as the underlying population random variable X_{Dell}. Assuming that our test has been designed correctly the X_i's will be independent. So again we arrive at the formal mathematical definition.

4 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 let

$$X = \text{ a random number in } [0, 1].$$

Then $P(a \leq X \leq b) = b - a$ (assuming $0 \leq a \leq b \leq 1$). We can repeat the experiment n times. *After* we actually perform the experiment n times we obtain n definite numbers $x_1, x_2, \cdots, x_n$ in $[0, 1]$. *Before* we make the choices we have *random* variables $X_1, X_2, \cdots, X_n$ representing the first, second, ..., n-th choice. We note that

$X_1, X_2, \cdots, X_n$ all have $U(0,1)$-distribution and are all independent. We say that $X_1, X_2, \cdots, X_n$ is a random sample from $U(0,1)$.

Once again we have arrived at a n-tuple of identically-distributed, independent random variables.