COMP 441/COMP 552 — Large Scale Machine Learning Jan 9, 202		Jan 9, 2017
Lecture 1: Mark and Recapture Estimation		
Lecturer: Anshumali Shrivastava	Scribe By: Anshume	ali Shrivastava

This scribe may contain errors, please do not cite. Please email if you find any errors.

## 1 Random Variable

Informally, a *random variable* is a variable whose value depends on the outcome of some stochastic (or random) process (or experiments). If we repeat the same experiment the value of the variable can change. For example, if we toss a fair coin and define a variable X such that

$$X = \begin{cases} 1, & \text{if the coin shows heads} \\ 0, & \text{otherwise,} \end{cases}$$
(1)

then X is a random variable.

For discrete random variable X, every possible value of X is associated with a chance or a probability value. In the above case, X = 1 has associated probability of 0.5 because the coin is fair. Similarly X = 0 also has the associated probability value of 0.5. For continuous random variables, we have a density function  $\phi$  associated with the random variable.  $\phi(x)$  can be thought of as the probability value associated with interval  $[x - \Delta, x + \Delta]$  for infinitesimal  $\Delta$ . The definition for the continuous case is not formal in the strictest sense, as we need familiarity with measure theory which we will skip. Since random variables, such as X, do not have a fixed value one important practical quantity of interest is the expectation E(X). The expectation of X is defined as the weighted summation of all possible values of the given random variable X, weighted by the probability associated with the value.

## 2 Mark and Capture Estimation: Counting Turtles

**Problem:** You are hired to estimate the number of turtles in a giant pond.

A Reasonable Strategy: Capture  $k_1$  turtles, mark them and release them in the pond. Then after some days, assuming marked turtles have mixed nicely with other turtles, catch  $k_2$  turtles. Count the number of marked turtles, call the number of marked turtles seen as M. (M is a random variable)

**Intuition:** If there are *n* turtles, then after marking  $k_1$  turtles.  $\frac{k_1}{n}$  is the fraction of turtles that are marked. If the second  $k_2$  sample is random, then the fraction of turtles marked in this sample should be same as in the original population (assuming sample is representative of the population). Therefore, the formula

$$\frac{k_1}{n} \simeq \frac{M}{k_2}; \qquad n \simeq \frac{k_1 k_2}{M}, \tag{2}$$

gives a good estimator for n

From The First Principles: Doing the Math Name the turtles in the  $k_2$  samples as  $\{T_1, T_2, ..., T_{k_2}\}$ . A turtle can be marked or unmarked. Define  $k_2$  random variables  $\{X_1, X_2, ..., X_{k_2}\}$  as

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ turtle, i.e. } T_i, \text{ is marked} \\ 0, & \text{otherwise.} \end{cases}$$
(3)

1: Mark and Recapture Estimation-1

We can write our random variable of interest and its expectation as

$$M = \sum_{i=1}^{k_2} X_i; \quad E(M) = E(\sum_{i=1}^{k_2} X_i) = \sum_{i=1}^{k_2} E(X_i)$$

Even though  $X_i$ 's are correlated, for expectation all we need is the value of  $E(X_i)$  which is  $\frac{k_1}{n} \forall i$ . This is because if any  $T_i$  is randomly selected (and turtles have mixed well) then  $Pr(X_i = 1) = \frac{k_1}{n}$  Thus we get,

$$E(M) = \frac{k_1 k_2}{n}; \quad n = \frac{k_1 k_2}{E(M)}$$

therefore  $\hat{n} = \frac{k_1 k_2}{M}$  is a reasonable estimator for *n*. Note: We only see *M*, *E*(*M*) is a property of random variable *E* which we don't observe.

In probability, it is recommended to do the Math from first principles, unless you are in a puzzle solving competition running out of time. Intuitions can be MISLEADING in probability.

## 3 In Hindsight: A very high-level recipe of Random Estimation

We want to estimate a target quantity T. We design a (randomized) experiment and get, say k, observed quantities  $O_i$ s which are related to the target t. These quantities are usually random variables which, under the given the assumptions, have some known properties. Example, we know mean-variance, distribution, etc. of  $O_i$ s

The name of the game is to figure out the estimation formula which is a function of the observed  $O_i$ s. In other words, we figure out the relationship between observables and target, under some realistic assumptions. Typically, we want the estimator,  $\hat{n} = f(O_1, O_2, ..., O_k)$ , such that  $E(\hat{n})$  is close to n. The concentration of estimate is determined by the variance of  $\hat{n}$ .

In the estimation problem, discussed in the last section, the target was n, we had only 1 observable M with  $f(M) = \frac{k_1 k_2}{M}$ . The assumption was that the second sample  $k_2$  is uniformly generated, i.e., we catch any turtle with equal probability.

The art is to design experiments, which are easy/cheap to conduct and leads to observables  $O_i$  with sharp concentration and which correlates nicely with the target. For the design part, there is almost always room for improvements. Assumptions are also part of the design.