

Computer Lab 3: Expected Values

1. Suppose that 1000 recruits are to be given a blood test to detect the presence of a disease that is known to be present in 5% of all recruits. Instead of testing each recruit separately the recruits are randomly divided into 100 groups of 10 each. In each group of 10 the blood samples are pooled into one sample and that one sample is tested. If the test comes back positive for the disease each of the 10 blood samples from the group is separately tested to determine who among them actually has the disease.

(a) Find the expected number of tests performed using this scheme.

Hint: Label the groups $1, \dots, 100$ and let X_j for $j = 1, \dots, 100$ be the number of tests performed on the j^{th} group. Then X_j takes either the value 1 or 11. Find $P(X_j = 1)$ and $P(X_j = 11)$. The expected value of X_j is just,

$$E(X_j) = 1 \cdot P(X_j = 1) + 11 \cdot P(X_j = 11).$$

The total number of tests, X is,

$$X = X_1 + \dots + X_{100},$$

so the expected value of X is,

$$E(X) = \sum_{j=1}^{100} E(X_j) = 100E(X_1).$$

Use R to compute this.

(b) Now suppose the 1000 recruits are divided up into $\frac{1000}{n}$ groups of size n . Find the expected number of tests performed under this scheme – it is a function $F(n)$. Find the value of n which minimizes the expected number of tests. What is the expected number of tests at the minimum?

Hint: Find the expression for $F(n)$. We can minimize $F(n)$ by using R to graph $F(n)$. One simple way to do this is to define a new function in R . Here is the syntax for doing this,

$$\text{testsplit=function}(n) F(n)$$

The name of the new function is “testsplit”, the “function(n)” tells R to define a new function called “testsplit” which has one argument “n”. The value of testsplit(n) is the expression $F(n)$ you found for the expected value of X , the number of tests performed. To plot this function for n going from 1 to 20 enter,

$$n = 1 : 20$$

$$x = \text{testsplit}(n)$$

$$\text{plot}(x, \text{type} = "l")$$

2. Suppose that an urn has N numbered balls $1, \dots, N$. Suppose that k balls are drawn with replacement from the urn. Let X be the random variable which has value r if the largest number in the sample of k balls is r . So, for example, if you draw 3 balls from the urn and get $\{2, 5, 5\}$ then X is 5 for this outcome. The distribution for X can be found by thinking about the event $(X \leq n)$. In order for the largest of the first k draws to be less than or equal to n all the first k draws must be less than or equal to n . For $n \leq N$ this happens with probability,

$$P(X \leq n) = \left(\frac{n}{N}\right)^k,$$

since sampling with replacement makes the trials independent. The probability density function for X is thus,

$$P(X = n) = P(X \leq n) - P(X \leq n - 1) = \left(\frac{n}{N}\right)^k - \left(\frac{n-1}{N}\right)^k.$$

This makes it possible to find an expression for $E(X)$,

$$E(X) = \sum_{n=1}^N n \left(\left(\frac{n}{N} \right)^k - \left(\frac{n-1}{N} \right)^k \right).$$

Evidently,

$$\begin{aligned} n \left(\frac{n}{N} \right)^k &= \frac{n^{k+1}}{N^k}, \text{ and } n \left(\frac{n-1}{N} \right)^k = ((n-1) + 1) \left(\frac{n-1}{N} \right)^k \\ &= \frac{(n-1)^{k+1}}{N^k} + \frac{(n-1)^k}{N^k}. \end{aligned}$$

Since (write out the first few terms to see how cancellation occurs),

$$\sum_{k=1}^N \left(\frac{n^{k+1}}{N^k} - \frac{(n-1)^{k+1}}{N^k} \right) = \frac{N^{k+1}}{N^k} = N,$$

we see that,

$$E(X) = N - \sum_{n=1}^N \left(\frac{n-1}{N} \right)^k = N \left(1 - \frac{1}{N} \sum_{n=1}^N \left(\frac{n-1}{N} \right)^k \right).$$

Problem. Suppose that the Toyota Prius cars that are shipped to Tucson are numbered $1, 2, \dots, N$. A random selection of 10 such cars gives a highest number 1066. Estimate the total number of such cars in Tucson.

Hint: One way to make such an estimate is to choose N so that

$$E(X) = N \left(1 - \frac{1}{N} \sum_{n=1}^N \left(\frac{n-1}{N} \right)^{10} \right) = 1066.$$

This is pretty complicated looking but R can do the calculations easily. In fact something simplifies dramatically for large N in this problem. Enter the following commands in R ,

```
i = 1 : 1500
x = 1 : 1500
for (N in i) x[N] = 1 - sum(((1 : N - 1)/N)^10)/N
```

This will produce a vector x with 1500 entries the 35th entry of which is (for example),

$$1 - \frac{1}{35} \sum_{n=1}^{35} \left(\frac{n-1}{35} \right)^{10}.$$

To get the expected value of X for $N = 35$ you would just need to multiply this number by 35. Now look at the behavior of entries of x by issuing the command,

```
plot(x, type = "l").
```

You should see that the limit of $x[N]$ as N tends to infinity approaches a constant c . Find c and then estimate N by noting that N will be large and,

$$Nc \simeq 1066$$

so,

$$N \simeq \frac{1066}{c}.$$

In general find the limit,

$$\lim_{N \rightarrow \infty} \left(1 - \frac{1}{N} \sum_{n=1}^N \left(\frac{n-1}{N} \right)^k \right),$$

and use this to give an estimate of N given an observation of X (for general k).

Hint: Think of the sum as a Riemann sum approximation to an integral.