Notes on German Airplanes Problem

Premise

(This problem is also known as the German tank problem.)

The premise of the problem is that, during WWII, the allies wanted to estimate the number of a particular kind of airplane that the Germans had available. Call this number N.

The data they had available were the serial numbers on the sides of the airplanes that they were able to observe in combat. If we assume that

- Each airplane was sequentially assigned a number, starting from 1 and going to N, and
- The airplanes are observed at random (we will consider both with and without replacement),

you could try to construct a guess of the total number of airplanes N.

Disclaimer: The math in this problem does *not* show you how to come up with a good estimate of N. Instead, we will think about how a particular random variable would be distributed if we knew N. Down the line, we will use these calculations to ask questions like "Were the data that I observed consistent with N being greater than 1000?" For now, though, we will stick to probability calculations; statistical inference will come later.

Suppose you observe n serial numbers. One reasonable guess for N based on these serial numbers is the maximum of the n numbers you observed – call this X. Note that X is a random variable because it is a summary of the set of the n randomly sampled serial numbers that you observed.

We will ask here: if N were known, how would X be distributed? For any value $k < N$, what is the probability that X would take a particular value k?

Version With Replacement

Suppose that we are able to observe n airplanes with replacement. For example, we have a spotter with binoculars who is recording the serial numbers of aircraft flying over a particular area every day. Based on this sampling procedure, we could see the same aircraft multiple times, so we can think about this as sampling serial numbers with replacement (granted, the assumption that these serial numbers are sampled at random is a stretch).

Supposing that we know N, what is the probability that X is equal to a particular value k, written $P(X = k)$?

To answer this question, it turns out that it is easier to write down the proability that the maximum is *at most* k. Note that for the maximum to be less than or equal to k, all n serial numbers would have had to be less than or equal to k. Since k of the N possible serial numbers are less than or equal to k,

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

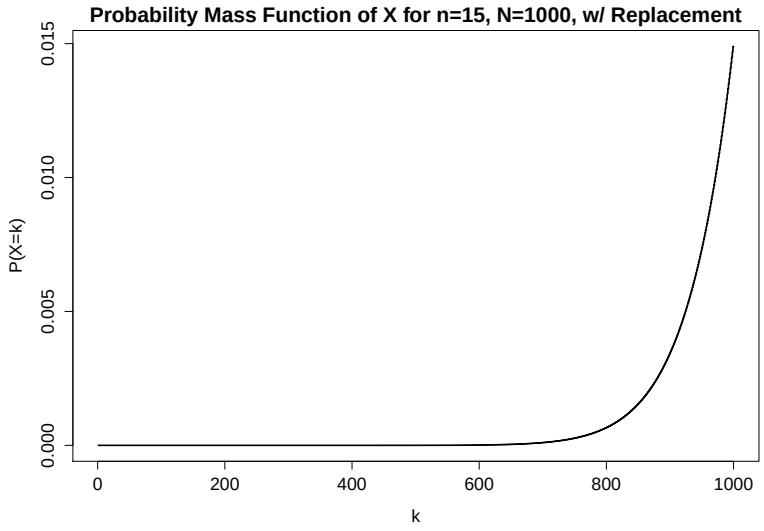
Now we use a trick. Note that because X takes consecutive integer values $\{1, \cdots N\}$, the event $X < k$ is the same as $X \leq (k-1)$ – in other words, to be strictly less than k, X needs to be less than or equal to $k-1$. Using this trick,

$$P(X = k) = P(X \leq k) - P(X < k) = P(X \leq k) - P(X \leq (k-1)).$$

So the probability mass function has a simple form

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

We can use this to visualize the distribution of the maximum observed serial number X for a given value of N. For example, if we suppose $N = 1000$, and we observe $n = 15$ airplanes at random with replacement, the maximum observed value X has a mass function that looks like the following:



Under these circumstances, observing a set of planes with maximum number less than, say, 700 is very rare.

Version Without Replacement

Now suppose that we observe serial numbers without replacement. This may be a scenario, where collect serial numbers from airplans that have been shot down, so once an airplane is observed, it won't be sampled again.

Supposing that we know N, what is the probability that X is equal to a particular value k, $P(X = k)$?

We'll use the same strategy as the "with replacement" version of the problem. For the maximum to be less than or equal to k, all n serial numbers would have had to be less than or equal to k. In this case

there are k serial numbers from which we can sample n values and have the maximum be less than k:

$$\begin{aligned} P(X \leq k) &= \frac{\#\text{ of sets of airplanes whose maximum serial number} \leq k}{\#\text{ of sets of airplains whose maximum serial nubmer} \leq N} \\ &= \frac{\binom{k}{n}}{\binom{N}{n}}. \end{aligned}$$

We'll use the same trick as in the last section:

$$P(X = k) = P(X \leq k) - P(X < k) = P(X \leq k) - P(X \leq (k-1)) = \frac{\binom{k}{n}}{\binom{N}{n}} - \frac{\binom{k-1}{n}}{\binom{N}{n}}.$$

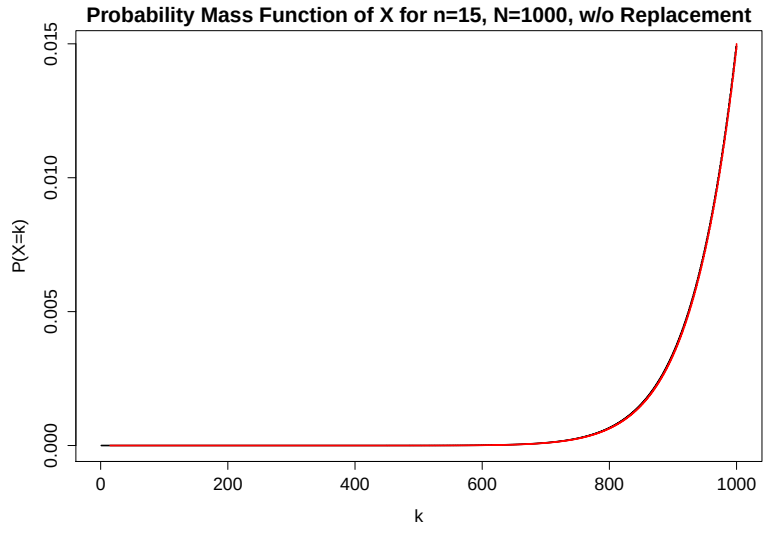
We can simplify this by using the following identity:

$$\binom{k}{n} = \frac{k!}{n!(k-n)!} = \frac{k}{k-n}\frac{(k-1)!}{n!(k-1-n)!} = \frac{k}{k-n}\binom{k-1}{n}.$$

So, factoring out $\binom{k}{n}$ in the numerator,

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

Again, we can use this to visualize the distribution of X. As above, suppose $N = 1000$, and we observe $n = 15$ airplanes at random without replacement, the maximum observed value X has a mass function that looks like the following:



In this figure, the probability mass function for observing aircraft without replacement is plotted in red, on top of the probability mass function for observing aircraft with replacement, which is barely visible because of how close the probability distributions are. Again, the probability of observing a sample of aircraft with maximum serial number less than, say, 700 is very rare.