My wife and I are long-time Volvo owners – myself, since 1968. In 2001 we ordered a new car, our first in 17 years, a V70 station wagon (fits 4’x8’ plywood!) with manual transmission (nice for mountain driving) and fabric seats (no burns or frostbite on the bum). We had to wait awhile for it to be built. I began to notice that Volvos comprised a fair fraction of the heavy tractors (of semis) on the Interstates – in fact, about 1 in 8. The fraction I observed in any one run varied a lot, however. I might see 10 in 58 or 7 in 100! What is the distribution of probabilities in these runs? Think of the probability of seeing *m* Volvos in a series of *n* consecutive trucks, if the real probability in the appearance in any one truck is 1/8.

Let the real fraction in the vast population of trucks be α. For a single element in the series, the probability of this next truck being a Volvo is just α, and the probability of it being another make is just 1- α. Now, assuming truckers don’t match speeds deliberately with trucks of similar make, the probability of a series of *n* trucks with *m* of them being Volvos seems to be

$$P = \alpha^m (1-\alpha)^{n-m}$$

Oops. This might apply to the sequence with the first *m* being Volvos and the next *n-m* being other makes. However, the Volvos can be scattered in any of the *n* positions. We have to multiply by the number of different sequences in which *m* objects, considered indistinguishable, fit into *n* boxes. For *m* = 1, there are *n* different places in the sequence where the object can go. We should multiply the probability by *n*. For *m* = 2, there are *n-1* places to put the 2nd object, so we should multiply the simple probability by *n*(n-1)*. Wait; the objects are indistinguishable; swapping the two gives the same result, so we should then divide by 2. For *m* = 3, we want to multiply by *n(n-1)(n-2)* to account for the alternative places....and divide by the number of ways to swap 3 identical objects, which is 6 – consider changing the order 1,2,3 to 1,3,2, then to 2,1,3, 2,3,1, 3,1,2, and 3,2,1. There are *m!* (“em factorial”) ways to do this, 1*2*....*m. So, the final multiplier for arbitrary *m* and *n* is

$$\begin{aligned} k &= \frac{n(n-1)(n-2)...(n-m+1)}{1*2*3*...m} \\ &= \frac{n(n-1)(n-2)....2*1}{(n-m)(n-m-1)....1}\frac{1}{1*2*3...m} \\ &= \frac{n!}{(n-m)!m!} \end{aligned}$$

Let’s write the final formula,

$$P = \alpha^m (1-\alpha)^{n-m} \frac{n!}{(n-m)!m!}$$

Now let’s look at an example. Suppose we observe 16 trucks. What are the probabilities that there are 0 Volvos, 1 Volvo, 2 Volvos, ...? For 0 Volvos, the probability is just (1- α)^{16} = (7/8)^{16} = 0.118.

For 1 Volvo, replace a (7/8) by (1/8) (same as multiplying by 1/7) and multiply by 16: P = 0.270.

For 2 Volvos, multiply the last result by 1/7 again and then by 15/2: P = 0.289.

For 3 Volvos, multiply this last results by 1/7 and then by 14/3: P = 0.193.

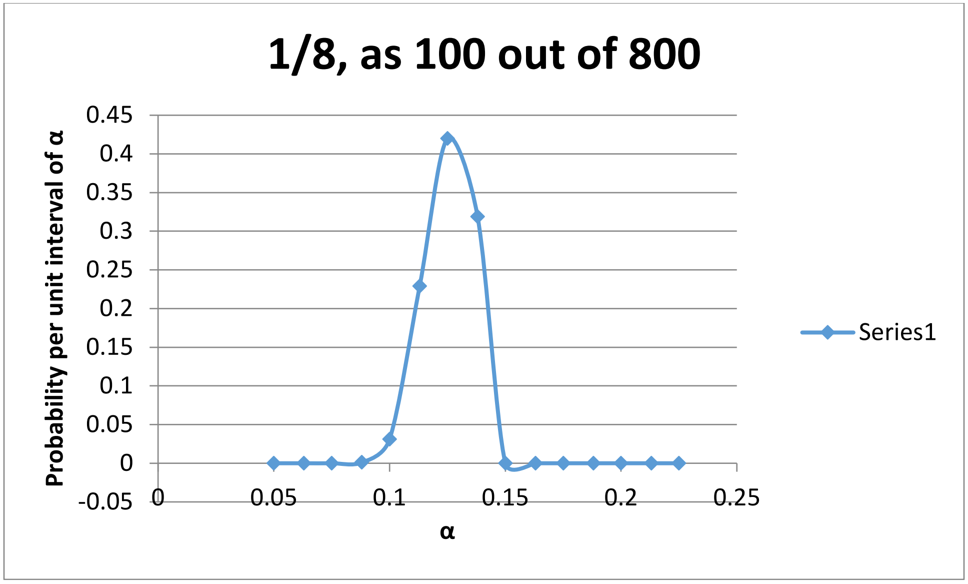
We get a table:

m	P
0	0.118

1	0.270
2	0.289
3	0.193
4	0.090
5	0.031
6	0.008
7	0.002
8, 9, ….16	Essentially 0.000
Total	1.001 (we have some rounding errors)

Hmm. The probability of getting “exactly the right fraction” is only a bit more than ¼. Consider a bigger number of observations, n = 32, for which we expect 4 Volvos. Our formula gives P = 0.209. For 112 observations, the probability of exactly 14 Volvos, or n/8, is only 0.113. So, P by itself doesn’t directly give us confidence that we know the right fraction, α.

In that case, we should consider as a comparison not P=P(n,m) itself, but its ratio to P(n,m) with a different assumed value of α, or a variety of assumed values of α. Let’s try this, for, say, 800 observations (a lot! This might take a day). Suppose we did get exactly 100 Volvos in this set. What is the probability that α is 1/8 (0.125) vs. some other values? I ran simulations with various values of α. Here is the plot:



Sure, the most likely value of α is 1/8 = 0.125, but others are possible. The plot considers α as a continuous variable, so we can compare the probability that α is 0.125 to the probability for any other α, say, 1/10. Well, the ratio of those probabilities is 0.42/0.031 > 13. That is, we’re much more confident that α is near 1/8 than it is near 1/10. We can even use the curve to get the interval that comprises, say, 95% of the likely values. This is about the range 0.11 = 1/9 to 0.14 = 1/7. That’s a surprisingly large range for so many observations. It should not be surprising, then, that one can have long runs with a fraction notably different from 1/8 even when α is really 1/8.

One quick note: since we’re considering truck arrivals as random events, it may be also surprising that it doesn’t matter if you accidentally fail to note any number of trucks in the series. Say, your vision was blocked but you know that trucks 13, 21, and 37 out of 40 trucks *were* semis but you couldn’t identify them. Missing observations, if not deliberately tied to previous observations (e.g., “I won’t count the next truck after I see 3 volvos), don’t affect the statistics. Any random sample from a series of random events gives the same answer.

P vs. other alpha
P vs. other m
Skipping?