Inferences

Problem:[1] A railway company numbers its locomotives 1, 2, $\ldots$, N. One day you see two locomotives and the largest number observed is 60. How many locomotives does the company have?

There are of course many ways to answer this question, no one of which is "right" and yet there are reasonable things to do.

Suppose you imagine this situation repeated many times. Then a method based on a "symmetry principle" will result in you being very close to the true value on average. Imagine you have $N+1$ dots arranged around a circle. Suppose you randomly pick 3 dots: 0-th choice, 1-st choice and 2-nd choice. The remaining dots are scattered in three groups. *The average number of dots per group must be the same by symmetry.* Suppose the 0-th choice dot signifies the place where you break the circle and straighten in out into a line of N dots. Then it must still be true that the average number of dots below the 1-st choice, the average number of dots between the 1-st and 2-nd choice, and the average number of dots above the 2-nd choice is the same. There are $60-2=58$ dots in the first two groups so on average there are $58/2=29$ dots in each group. Therefore, above 60 there are 29 dots on average. Thus on average there are $N=60+29=89$ locomotives.

Suppose you don't get a reward unless you are exactly right — you don't care about being merely close on average, but instead want to be exactly right no matter how faint a hope you have. You want to be right with as high a probability as possible. A good strategy is to pick the highest number, 60, that you saw. The probability that a sample of 2 will contain the maximum is $(N-1)/\binom{N}{2}=2/N$. Thus if there were 60 locomotives and you repeatedly used this kind strategy you would be right one thirtieth of the time. On the other hand, if you always guessed say 5 more than the highest number, you're interested in probability that a sample of 2 will have maximum $N-5$ and this probability is $(N-6)/\binom{N}{2}=2(N-6)/(N(N-1))$. This is lower by a factor of $(N-6)/(N-1)$ and so is an inferior strategy although if there just happened to be 65 locomotives it would be the better answer (but not the better strategy!).

Another possibility is to determine for which N, a sample with maximum 60 is most likely — the method of maximum likelihood. The probability that 60 is the maximum of a size 2 sample is $59/\binom{N}{2}$ =-$118/(N(N-1))$ which gets steadily larger as we let N get smaller. No sample can have a maximum of 60 unless $N\geq 60$ thus the sample of maximum likelihood has $N=60$. So guess 60.

Alternately a confidence interval could be given: we want to be $q\%$ sure that N is in a certain range of values — the closer we take $q\%$ to 100% the larger the range of values will have to be. The probability that the maximum in a sample of two is no more than M, for $M<N$, is $\binom{M}{2}/\binom{N}{2}$ which is about $(M/N)^2$. The probability that the maximum is at least M is about $1-(M/N)^2$. Now take $M/N=1/3$ say. The probability that the maximum is at least $M=N/3$ is about $1-1/9$ which is about 89%. Thus N is no more than $3M$ about 89% of the time. In our specific case, we can be 89% sure that the number N of locomotives is no more than $3\cdot 60=180$.

Comment: During the Second World War serial-numbers from captured German equipment and from seized command post records were used to estimate the German war production. Serial-numbers indicate order of production and these were used as above.

"When the war was over and the official records of the Speer Ministry impounded, it was found that estimates derived from procedures similar to the one just described [like our estimate of 89 using symmetry] were far more accurate than those based on any other source of information. As an example, the serial-number estimate for German tank production in 1942 was 3400, which was very close to the actual figure. The "official" Allied estimate, based on information culled from intelligence reports and espionage activities, was 18,000. Errors of this magnitude were not uncommon. Often the reason for these inflated estimates was the Nazi propaganda machine. Efforts to create the impression that Germany was much stronger than it really was were highly successful. Only the completely objective serial-number procedure remained unaffected!"[2]

[1] Question & some of the discussion from: *Fifty Challenging Problems in Probability with Solutions* by F. Mosteller, Dover, N. Y. — problem #41

[2] Larsen & Marx, "An Introduction to Mathematical Statistics and Its Applications", Prentice-Hall, 1981, pg 203