Optional Stopping

Malcolm R. Forster
Fifth Draft,
December 20, 1998

Introduction

Suppose you are determined to "prove" that green apples cause cancer. An Optional Stopping strategy (OS) is where you keep looking sampling experimental data until the observed correlation between eating green applies and cancer is significantly different from 0 (where "significantly" means that the null hypothesis is rejected by standard statistical tests). That is, you follow a rule that says "Don't stop until you reject the null hypothesis". This is also the best strategy for confirming the existence of UFOs or establishing the phenomenon of extrasensory perception (ESP) (see Feller, 1940, for the confutation of this tongue-in-cheek assertion). If the data are 'noisy' (and whose data are not?), then this will probably always work in principle, so not always in practice because you won't live long enough to collect enough data.

There are two schools of thought about optional stopping examples of the kind that I consider (see Robbins, 1952, for a more general discussion). The classical hypothesis testers say that it is a bad strategy if the probability of falsely rejecting the null hypothesis when it is true is 1. Some Bayesians, and likelihood theorists, say that it all that matters is how well the hypotheses fit the data, and it makes no difference whether you collect n data by an OS strategy, or if you collect n data with the prior intention of stopping at a sample size of n (strategy FS). About the only thing that has never been said about OS is that it is better than FS (with the same n).

What follows is a series of computer simulation comparing an OS strategy with an FS strategy. The first simulation assumes that the null hypothesis is true, so that rejection of the null hypothesis is always a mistake. When I initially ran the simulation, I found that I had to wait too long for my computer finish doing its experiments (despite its running at 450 MHz). So, I set an upper limit of 2000 data points. For me, 2000 without stopping counts as "no outcome". As you will see from the results below, the experiment had "no outcome" 40% of the time. The next two simulations show how things change when the null hypothesis is false. The results were initially surprising to me. It turns out that OS can be a more reliable method than FS most of time. In the final section, I explain this odd result in terms of an easy-to-understand analogy.

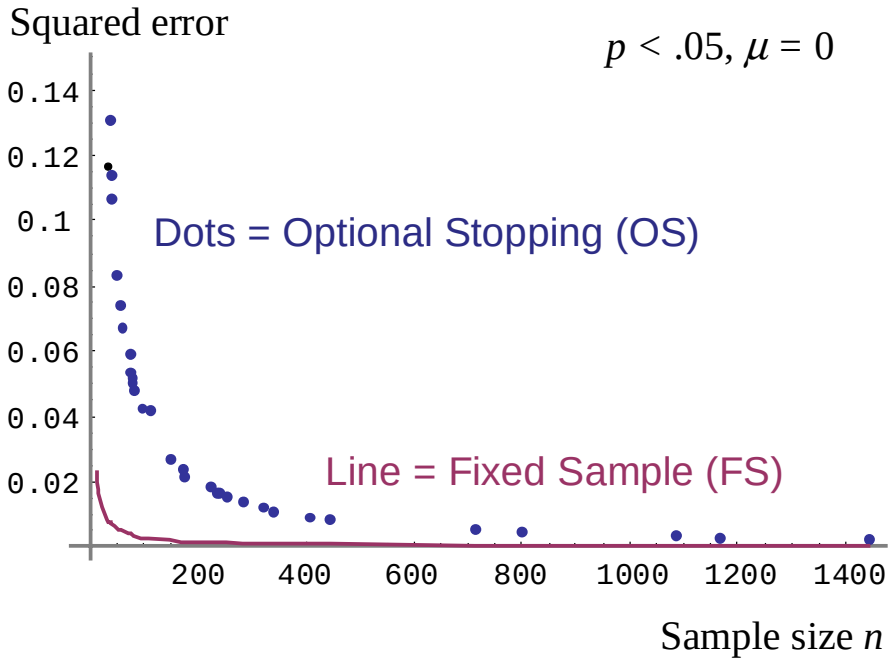
When the Null Hypothesis is True

The first column of the table below is the number of coin tosses, n, it took to produce the required discrepancy. If the required discrepancy was not met, then $n = 2{,}000$. The second column is the p-value achieved. The stopping rule is that $p < .05$, so that you will notice that this is true for the rows in which $n < 2{,}000$, but false for the rows in which $n = 2{,}000$. The third column gives the estimated bias of the coin (that is the relative frequency of heads). Those values will be important when we compare the performance of optional stopping and the fixed sample strategy.

n	p-value	μ-estimate
2	0.002	-2.173
2	0.006	-1.937
2	0.015	1.7172
2	0.029	-1.539
2	0.037	1.4727
2	0.039	1.4532
3	0.008	1.5305
3	0.021	-1.326
3	0.032	-1.236
3	0.033	1.2251
3	0.041	1.1766
3	0.048	-1.140
4	0.020	1.1600
4	0.023	1.1304
5	0.048	0.8841
6	0.036	-0.853
6	0.049	0.8028
7	0.030	-0.816
8	0.040	0.7234
9	0.024	0.7474
12	0.031	-0.619
14	0.036	-0.557
16	0.038	-0.517
17	0.038	0.5011
23	0.039	-0.430
33	0.049	0.3421
36	0.029	-0.362
38	0.043	-0.327
39	0.034	-0.338
46	0.049	-0.289
53	0.046	-0.272
58	0.048	0.2593
74	0.035	0.2440
74	0.046	-0.231
77	0.045	-0.227
77	0.048	0.2252
81	0.047	-0.220
96	0.043	-0.206
110	0.031	0.2052
146	0.046	0.1650
170	0.043	0.1549
175	0.049	-0.148
222	0.041	-0.136
234	0.046	0.1299
236	0.047	0.12877
253	0.048	0.12391
284	0.045	0.11895
319	0.044	-0.1125
340	0.049	-0.1067
407	0.049	-0.0973
445	0.048	0.09335
711	0.049	0.07378
714	0.048	0.07391
797	0.044	-0.0713
1082	0.048	0.060022
1164	0.045	0.058729
1441	0.047	-0.05227
2000	0.101	-0.03658
2000	0.133	-0.03352
2000	0.145	0.032573
2000	0.174	0.030397
2000	0.199	-0.02867
2000	0.208	0.028094
2000	0.272	0.024555
2000	0.337	-0.02143
2000	0.355	-0.02065
2000	0.374	0.01984
2000	0.432	-0.01755
2000	0.437	0.017364
2000	0.475	-0.01596
2000	0.506	-0.01485
2000	0.507	0.014805
2000	0.511	-0.01467
2000	0.540	-0.01367
2000	0.544	-0.01355
2000	0.571	0.012645
2000	0.604	0.011591
2000	0.629	0.010780
2000	0.653	0.010032
2000	0.673	0.009418
2000	0.689	0.008946
2000	0.705	-0.00843
2000	0.715	-0.008164
2000	0.754	-0.006978
2000	0.766	0.0066396
2000	0.773	0.0064232
2000	0.798	0.0056978
2000	0.807	-0.005446
2000	0.807	-0.005445
2000	0.816	-0.005189
2000	0.823	-0.004995
2000	0.846	-0.004335
2000	0.846	0.0043206
2000	0.869	0.0036693
2000	0.920	0.0022416
2000	0.953	-0.001308
2000	0.964	-0.000985
2000	0.984	0.0004443

2000	0.985	0.0004100
2000	0.990	-0.000263

I decided to compare the performance of the optimal stopping strategy (OS) with the fixed sample size strategy (FS) by looking at the expected, or average, squared deviation of the FS inferred value of μ (the relative frequency of heads) from the true value at every value of n at which OS stopped. The results are plotted on the two graphs below. The dots represent the squared deviations for OS, while the curve at the bottom represents the expected squared deviations of FS. I have plotted the lower values of n separately from the higher values of n to make the graphs more readable. The nearer the x-axis, the smaller the error. Needless to say, FS performs much better than OS.

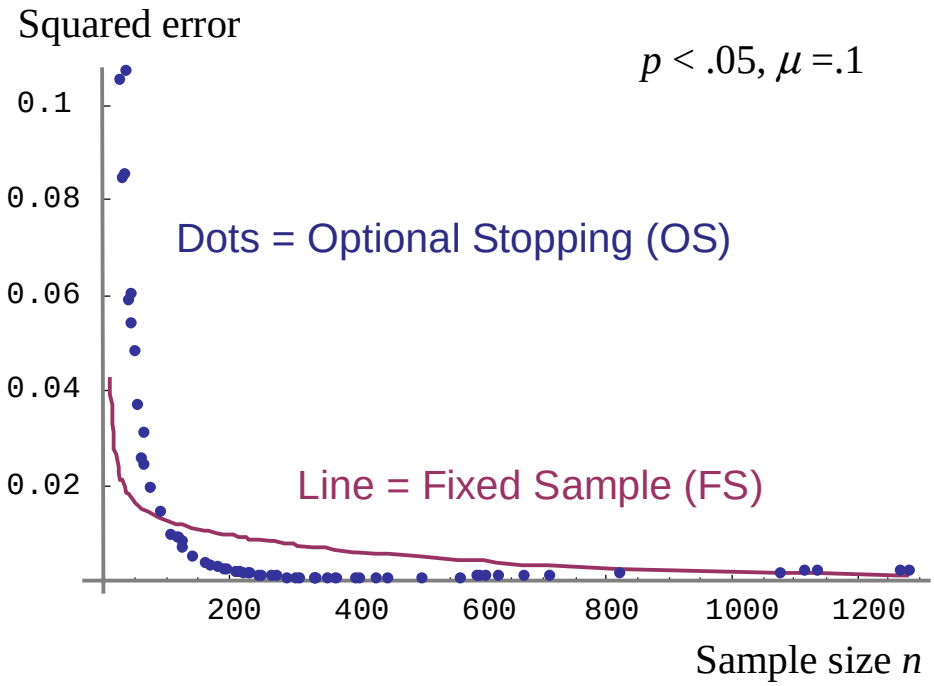


When the Null Hypothesis is False

Part of the reason that FS performs better may be because the null hypothesis (that $\mu = 0$) is actually true. To test the effect of this, I reran the simulations for the cases in which the null hypothesis is false to a varying degrees. The plots below are for $\mu = .1$, and $\mu = .4$. Remember that the true standard deviation is 1 in these simulations, while the standard deviation in a coin tossing example is about 0.5. Therefore, as a fraction of the standard deviation, these values of μ correspond roughly to a coin's propensity to land heads of about .55, and .7. (**Note:** The normal approximation I have used does not work when the coin propensity is near 1; nor will it work so well when n is small. However, I'm sure that one would find that the qualitative results still hold. The normal distribution is really the more interesting and general one in any case.)

The plot for $\mu = .1$ appears below. Even though the null hypothesis is approximately true (since the postulated value of μ differs from the true value by only 10% of a standard deviation), there is a huge difference in these results. First, 100% of the trials stopped before n =1500, as opposed to just 60% for the previous case ($\mu = 0$). Second, for sample sizes of greater than about 100, OS did better than FS! Moreover, the OS strategy is very accurate around the point $n = 400$.

These are surprising results, which demand and explanation. I will attempt one later, but first let me describe the phenomenon some more.



I left out the 20 points for the smallest values of n in order to see more of the detail for the remaining 80 points. Therefore, there are 34 instances on the left (14 shown on the plot) for which OS has a greater error than FS. When one adds the 4 points on the right for which OS a slightly greater error, then about 60% of the points were ones for which OS performed better than FS. When one averages over the errors for all 100 instances, FS comes out to the be more reliable strategy overall.

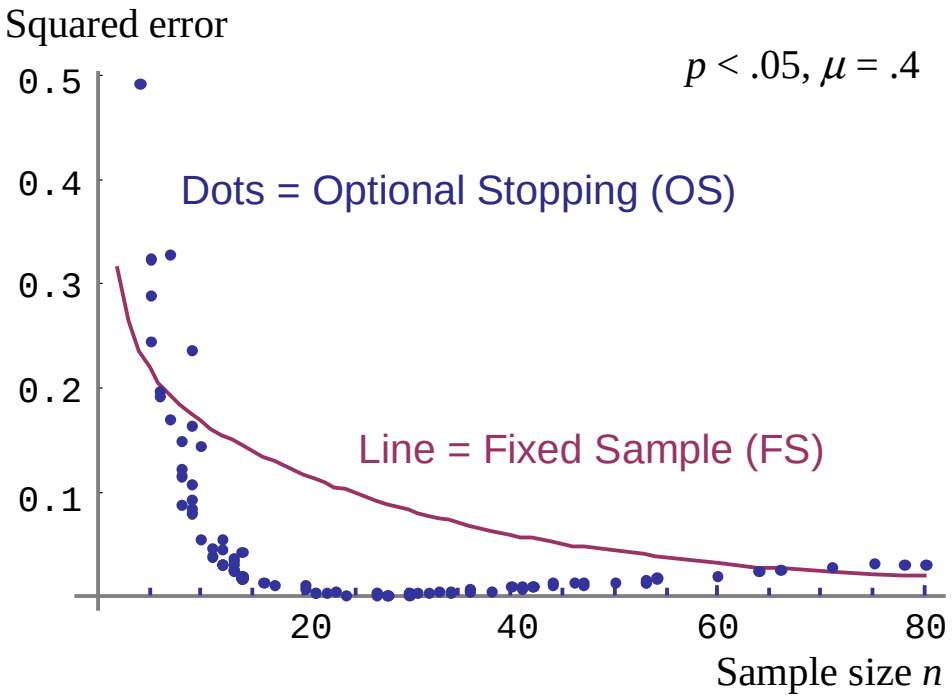
One of the noteworthy features of the plot is the point at the left were OS and FS have the same expected error. This cross-over point is characterized as follows. Call the square of the difference between the null value of μ (which is 0) and true value the *bias* of the null hypothesis. The bias is a measure of the degree to which the null hypothesis is false. In the first run, the bias was 0 because the null hypothesis was true. In this run, the bias is 1/100. Now consider the variance of the sample mean. This is equal to $1/n$. Note that the bias is equal to the variance when $n = 100$, and this is exactly where the null hypothesis and its alternative lead to the same expected error. So *all* methods for choosing between the two alternatives must lead to the same expected error. This phenomenon is exactly as I found in comparing various model selection methods. See "The New Science of Science of Simplicity" and "The Key Concepts of Model Selection: Performance and Generalizability." OS and FS are no exceptions. This also explains why there is no cross-over when $\mu = 0$. For then the bias is zero, in which case there is no n for which the bias is equal to $1/n$.

Let's look at what happens when the true value of μ is 0.4 (below). Again, all the OS trials stopped, this time all at or before $n = 80$. Second, the advantage of OS over FS is quite marked for n greater than about 6, and between 20 and 30, the OS strategy is incredibly accurate. The point $n = 6$ is when the bias is equal to the variance. For values of n less than that FS is more reliable; for values of n greater than that, OS is more reliable, except for large n.

The graph (below) shows all 100 instances of the experiments. So, by inspection, we see that 90% of the cases are ones for which OS performs better than FS. There are some points left

out at the beginning, so it is still not clear that OS performs better than FS on average. But it is clear that the competition is closer, and that OS will do better for higher values of μ. That is, on average, the optional stopping strategy is predictively more accurate in such situations.

Perhaps it is not surprising that the bias of OS against the null hypothesis is to its advantage when the null hypothesis is false by a sufficient degree. However, this does not explain why performs better for intermediate values of *n*. It is this fact that I will explain next.



An Analogy and an Explanation

Is there an intuitive explanation for these results? Here is how I think of optional stopping. Suppose that a blind-folded person starts at one place and is asked to head in the direction of the sun. He cannot see the sun, but he can feel the sun's rays on his face. We expect him to walk roughly in the direction of the sun, but with some random errors in every step. You (the experimenter) do not know the direction of the sun, and your job is to infer it from the behavior of the blind-folded subject.

There are two subjects, whose initials are OS and FS. (OS stands for Optional Stopping, while FS stands for Fixed Sample size.) OS stops when he hits either of the side lines. Then you will record his position, and draw a line from that position to his starting point. That line is your estimate of the direction of the sun.

Now draw a finish line (the vertical line in the figure) that passes through the point at which OS stopped. FS begins walking from the start, and is stopped as soon as he hits this line, and not before. Unlike OS, FS is allowed to cross the side lines any number of times. Now record FS's position at the finish line. If it is between the side lines, you will infer that the sun is on the center line. If he is outside the side lines, then draw a line from that position to his starting point and use this line to estimate the direction of the sun.

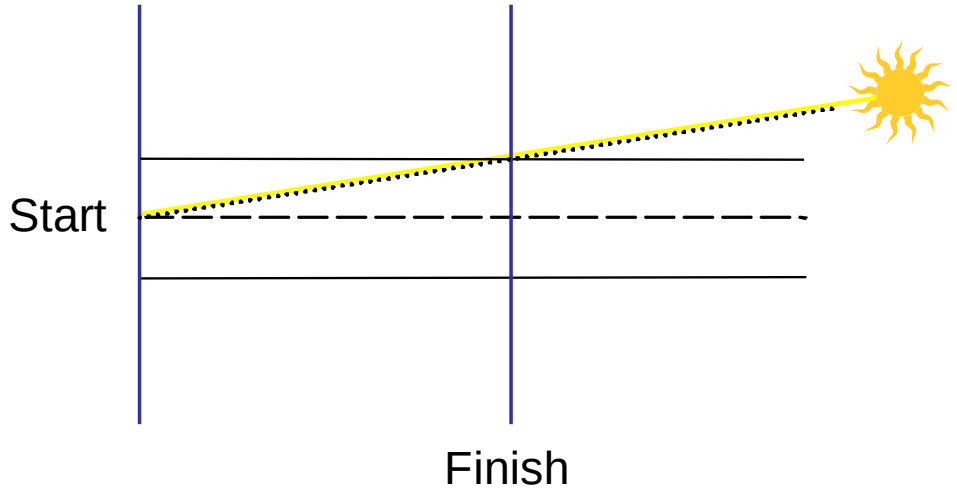
The question is: Which inference is the most reliable? That is, on average, which method of estimating the sun's direction is associated with least error, where error is measured by the square of the discrepancy between the estimated line and the true line. If this example is

analogous to the simulated problem (and I believe it is), then we know that the answer depends on three things. (1) The bias, which is the discrepancy between the true direction of the sun and the center line. The null hypothesis is that. If the direction of the sun is along the center line (the null hypothesis is true), then FS will be the more reliable predictor, on average, because 95% of the time you will be exactly right. The OS line, on the other hand, is always outside the center line, and therefore your inference will always be wrong by at least that amount.

But what if the sun is off the center line (as shown in the diagram)? Now, whenever you reject the null hypothesis, you are making a mistake. Nevertheless, the OS estimate will be wrong most of the time too. So, why should OS have an advantage? It depends on the value of *n*, where *n* is the number of steps taken from the start. I will consider one possible value of *n* that will make the OS estimate extremely accurate. Suppose that the *n* at which OS stopped is such that the finish line passes through the point at which the true line to the sun crosses one of the side lines. (There must be such a point if the sun is not on the center line.) Given that OS stops on this line, and he has moved roughly in the direction of the sun, OS will have stopped very close to the point at which these lines intersect (see the diagram above). Hence, the OS estimate will be fairly accurate, or a least far more accurate than FS, on average, because FS will cross the finish line at a variety of points normally distributed around this point. Remember that if FS's resting place is below this point, then the inferred direction will be the center line, which is inaccurate. If it is above the point (see the diagram), then it is most frequently well above the point at which OS stopped because OS has only just stepped over the side line. FS may have been over the side line for some time. This argument also holds approximately for values of *n* close to this critical value. This explains the middle region of my plots, in which OS provides very accurate results, which are far more accurate than FS.

A second phenomenon is as follows. Suppose, against all odds, OS gets well past the finish line shown in the diagram. Then OS will still finish near the upper side line. But this now provides an inaccurate estimate of the sun's direction. FS, on the other hand, will do better because he is not constrained to stop near the side line. This explains the points on the far right of my plots, for which FS begins to perform better than OS.

In summary, for small values of *n*, FS performs better than OS. When the bias is equal to the variance, FS and OS perform the same (on average). For middle values of *n*, OS performs better than FS. For large values of *n* (if there are any), FS does better than OS once again.



Bibliography

Feller, William (1940): "Statistical Aspects of ESP," *Journal of Parapsychology* **4**: 271-298.

Hacking, Ian (1965). *Logic of Statistical Inference.* Cambridge: Cambridge University Press.

Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago and London, The University of Chicago Press.

Robbins, Herbert E. (1952): "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society* **58**: 527-535.