The Imaginary Theorem of Large Samples

How Many Data Do You Need?

Donald J. Wheeler

Courses in statistics generally emphasize the problem of inference. In my December column I defined this problem in the following manner: Given a single unknown universe, and a sample drawn from that universe, how can we describe the properties of that universe? In general, we attempt to answer this question by estimating characteristics of the universe using statistics computed from our sample.

One of the lessons that most students of statistics manage to learn is that, in questions of inference, the uncertainty of an estimator is inversely related to the amount of data used. To illustrate this relationship I will use drawings from my Bead Box. A paddle with holes on one side is used to obtain a sample of 50 beads. The number of yellow beads in each sample of 50 beads is recorded. The beads are replaced in the Bead Box, the beads are stirred up, and the whole sequence begins again. After ten such drawings I have drawn 500 beads and have found a total of 65 yellow beads.

My point estimate for the proportion of yellow beads in the Bead Box is thus:

$$p = 65/500 = 0.1300 \text{ or } 13\%$$

and the usual 90% Interval Estimate for the proportion of yellow beads is:

$$p \pm 1.645\sqrt{\frac{p\,(1-p)}{n}} = 0.1300 \pm 0.0247 = 0.1052 \text{ to } 0.1547$$

I repeat this experiment of drawing of 10 samples and find 43 yellow beads out of 500 beads sampled. Combining the results from both experiments we have a point estimate for the proportion of yellow beads in the Bead Box of:

$$p = 108/1000 = 0.1080 \text{ or } 10.8\%$$

and the usual 90% Interval Estimate for the proportion of yellow beads is:

$$p \pm 1.645\sqrt{\frac{p\,(1-p)}{n}} = 0.1080 \pm 0.0161 = 0.0919 \text{ to } 0.1241$$

While the point estimate changed, the element of interest here is how the uncertainty decreased from 0.0247 to 0.0161 as we went from using 10 samples to using 20 samples in our estimate. With increasing amounts of data our estimates come to have lower levels of uncertainty. Table 1 shows the results of 20 repetitions of this experiment of drawing 10 samples of 50 beads each from my Bead Box. The first column gives the cumulative number of yellow beads. The second column gives the cumulative number of beads sampled. The third column lists the cumulative point estimates of the proportion of yellow beads, while the last two columns list the end points for the 90% interval estimates for the proportion of yellow beads. Figure 1 shows these last three columns plotted against the number of the experiment.

Table One: Twenty Bead Box Experiments

Experiment	Number Yellow	Number Drawn	p	90% Interval Estimate Lower	Upper
1	65	500	0.130	0.105	0.155
2	108	1000	0.108	0.092	0.124
3	164	1500	0.109	0.096	0.123
4	219	2000	0.110	0.098	0.121
5	276	2500	0.110	0.100	0.121
6	336	3000	0.112	0.103	0.121
7	389	3500	0.111	0.102	0.120
8	448	4000	0.112	0.104	0.120
9	506	4500	0.112	0.105	0.120
10	553	5000	0.111	0.103	0.118
11	609	5500	0.111	0.104	0.118
12	655	6000	0.109	0.103	0.116
13	714	6500	0.110	0.103	0.116
14	770	7000	0.110	0.104	0.116
15	817	7500	0.109	0.103	0.115
16	874	8000	0.109	0.104	0.115
17	927	8500	0.109	0.103	0.115
18	981	9000	0.109	0.104	0.114
19	1040	9500	0.109	0.104	0.115
20	1105	10000	0.111	0.105	0.116

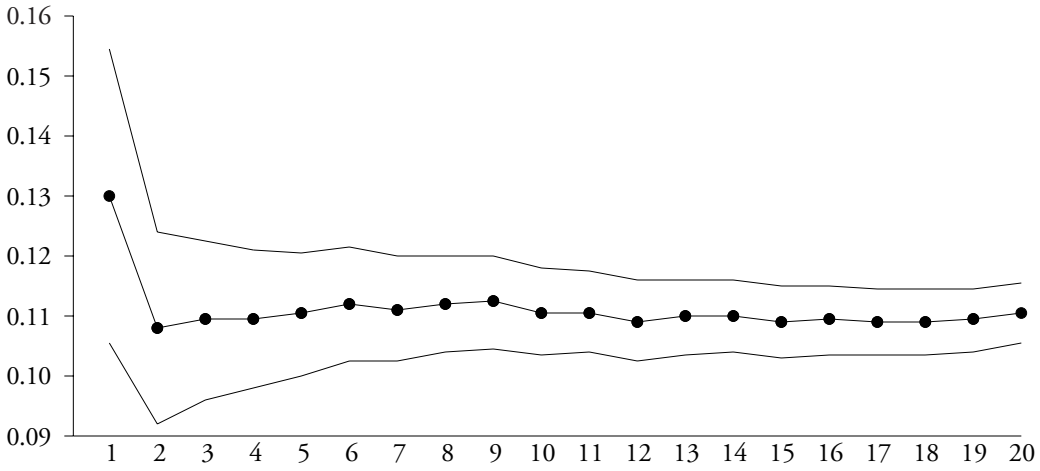

Figure One: Cumulative Proportion Yellow and 90% Interval Estimates

As we look at Figure 1 we see the point estimate converge on a value near 0.11 and stabilize there while the uncertainty keeps dropping and the interval estimate gets tighter and tighter. This is the picture shown in textbook after textbook, and the source of the theorem of large samples. Based on this graph, it would appear that this experiment will yield an average of 11% yellow beads.

But is this a reasonable estimate of the proportion of yellow beads in my Bead Box?

Since there are only 4800 beads in the box, the 20 repetitions of our experiment effectively looked at every bead in the box twice. Yet, by actual count, the box contains only 10% yellow beads, a value that was outside the interval estimate from Experiment 5 on. As we collected more and more data our point estimate did converge, but it did not converge to the "true value."

So here we come to the first problem with the theorem of large samples. The whole body of computations involved with estimation are built on certain assumptions. One of these is the assumption that we have drawn random samples from the universe. However, random samples are nothing more than a concept. There is no rigorous mathematical definition of random. In practice we always have to use some sort of sampling system or sampling device. Here we used mechanical sampling. And regardless of how careful we may be, mechanical sampling is not the same as the assumption of random sampling. Here we ended up with an excess number of yellow beads in our samples. Nothing in our computations can compensate for this bias. Moreover, in practice, where we cannot stop the experiment and find the "true value" by counting all the beads, there will be no way to even detect this bias. Thus, in practice, the first problem with the theorem of large samples is that, because of the way we obtain our data, our estimates may not converge to the values that we expect them to converge to.

And if this problem is not enough to give you pause, there is an even bigger problem with the theorem of large samples.

Table Two: The Batch Weight Data

Batch No.	Batch Weights (kilograms of product exiting blender)									
1-10	905	930	865	895	905	885	890	930	915	910
11-20	920	915	925	860	905	925	925	905	915	930
21-30	890	940	860	875	985	970	940	975	1000	1035
31-40	1020	985	960	945	965	940	900	920	980	950
41-50	955	970	970	1035	1040	1000	1000	990	1000	950
51-60	940	965	920	920	925	900	905	900	925	885
61-70	1005	1005	950	920	875	865	880	960	925	925
71-80	875	900	905	990	970	910	980	900	970	900
81-90	895	885	925	870	875	910	915	900	950	880
91-100	910	965	910	880	900	920	940	985	965	925
101-110	925	975	905	890	950	975	935	940	900	915
111-120	980	880	905	915	960	900	915	920	865	980
121-130	935	840	900	965	890	875	1020	780	900	900
131-140	800	960	845	820	910	885	940	930	925	850
141-150	965	1010	1030	980	1010	950	940	1005	880	930
151-160	845	935	905	965	975	985	975	950	905	965
161-170	905	950	905	995	900	840	1050	935	940	920
171-180	985	970	915	935	950	1030	875	880	955	910
181-190	1050	890	1005	915	1070	970	1040	770	940	950
191-200	1040	1035	1110	845	900	905	910	860	1045	820
201-210	900	860	875	1005	880	750	900	835	930	860
211-220	960	950	1020	975	950	960	950	880	1000	1005
221-230	990	1020	980	1020	920	960	1000	1000	860	1130
231-240	830	965	930	950	945	900	990	865	945	970
241-250	915	975	940	870	890	915	935	1060	1015	1100
251-259	810	1010	1140	805	1020	1110	975	970	1090	

To illustrate this second problem I shall use the Batch Weight Data shown in Table 2. There you will find the weights, in kilograms, of 259 successive batches produced during one week at a plant in Scotland. For purposes of this example assume that the specifications are 850 kg to 990 kg. After every tenth batch the Capability Ratio is computed using all of the data for that week. Thus, just like the proportion of yellow beads, we would expect to see these Capability Ratios converge to some value as the uncertainty drops with the increasing amounts of data used. Table 3 shows the number of batches used for each computation, the Capability Ratios found, and the

90% Interval Estimates for the process capability. These values are plotted in sequence in Figure Two.

Table Three: Cumulative Estimates of Capability Ratio for Batch Weight Data

Batches	C_p	Lower	Upper	Batches	C_p	Lower	Upper
10	1.102	0.575	1.596	140	0.677	0.590	0.762
20	1.205	0.795	1.594	150	0.656	0.575	0.736
30	0.898	0.648	1.137	160	0.654	0.576	0.731
40	0.904	0.687	1.114	170	0.623	0.550	0.693
50	0.966	0.758	1.167	180	0.614	0.545	0.682
60	1.022	0.821	1.216	190	0.555	0.495	0.615
70	0.951	0.779	1.118	200	0.528	0.472	0.584
80	0.85	0.706	0.991	210	0.509	0.456	0.561
90	0.856	0.720	0.990	220	0.511	0.459	0.562
100	0.849	0.720	0.974	230	0.502	0.452	0.552
110	0.846	0.724	0.966	240	0.489	0.442	0.536
120	0.806	0.695	0.915	250	0.489	0.442	0.535
130	0.716	0.620	0.809	259	0.455	0.412	0.497

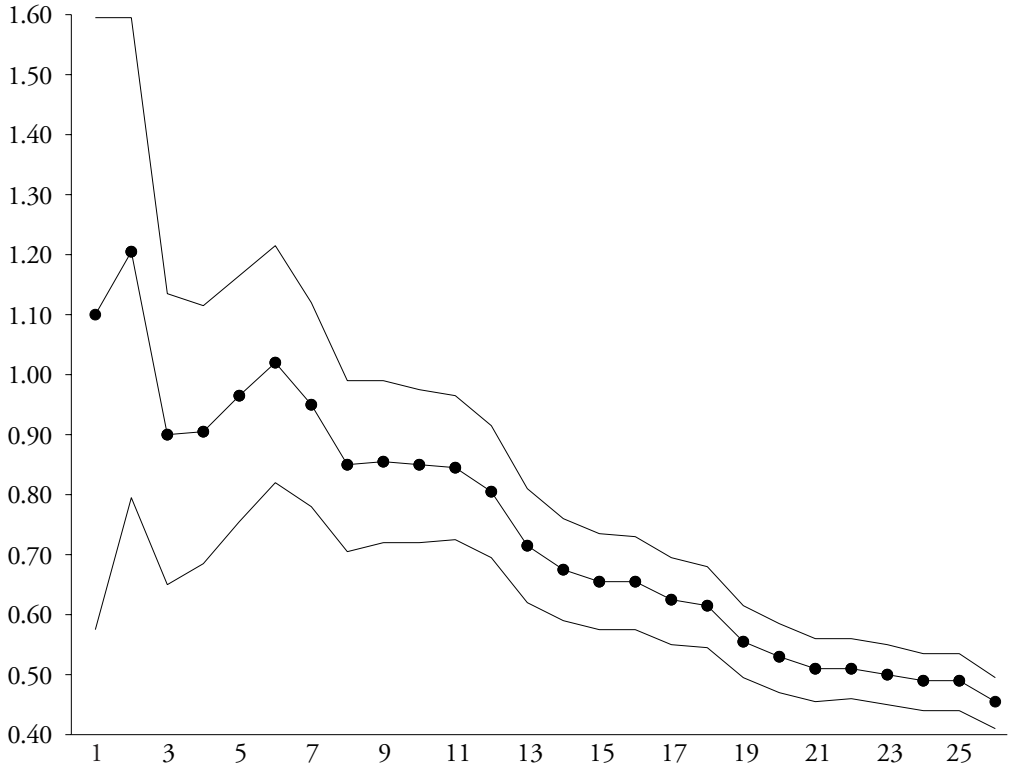

Figure 2: Cumulative Estimates of Capability Ratio for the Batch Weight Data

There we see that these Capability Ratios do not converge to any one value. Instead they rather meander around over time. While the uncertainties decrease with increasing amounts of data, this reduction in uncertainty is meaningless when the target itself is uncertain. We get better and better estimates of something, but that something may have already changed by the time we have the estimate.

To understand the behavior of these Capability Ratios we need to look at the *XmR* Chart for these data in Figure 3. The limits shown are based on the first 60 values. This baseline was chosen in order to obtain a reasonable characterization of the process potential. Here we see a process that is not only unpredictable, but one that gets worse as the week wears on!

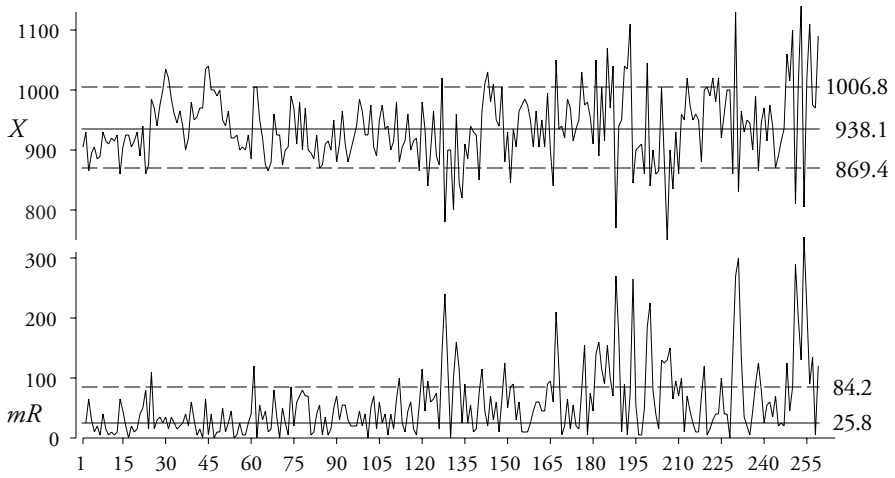

Figure 3: *XmR* Chart for the Batch Weight Data

The Theorem of Large Samples implicitly assumes that there is one universe. When there are multiple universes, and these universes are changing around without warning, no amount of data will ever be sufficient to provide a good estimate of any process characteristic!

This takes us back to the question of homogeneity, which is the fundamental question of data analysis. Did these data come from a process or system that appears to be operating in the same way over time? Or do these data show evidence that the underlying process has changed in some manner while the data were collected? And the only statistical technique that can answer this question of homogeneity is the process behavior chart.

If the process behavoir chart does not show any evidence of a lack of homogeneity, then our process may be predictable, and the Theorem of Large Samples may only suffer from the problem of estimating the wrong thing. (Drawings out of a Bead Box are a classic example of a predictable process.)

But if we find evidence of a lack of homogeneity within our data, then we know that our process has multiple personality disorder, and any attempt to use the Theorem of Large Samples is merely wishful thinking—no amount of data will ever be sufficient to provide a reliable estimate of a process parameter.