The Secret Foundation of Statistical Inference

What you don't know can hurt you

Donald J. Wheeler

When industrial classes in statistical techniques began to be taught by those without degrees in statistics it was inevitable that misunderstandings would abound and mythologies would proliferate. One of the things lost along the way was the secret foundation of statistical inference. This article will illustrate the importance of this overlooked foundation.

A naive approach to interpreting data is based on the idea that "Two numbers that are not the same are different!" With this approach every value is exact and every change in value is interpreted as a signal. We only began to emerge from this stone-age approach to data analysis about 250 years ago as scientists and engineers started measuring things repeatedly. As they did this they discovered the problem of measurement error: repeated measurements of the same thing would not yield the same result. For some, such as the French astronomer Pierre Francois Andre Mechain, this resulted in a nervous breakdown. For others this was the beginning of a new science where two numbers that are not the same may still represent the same thing. While Pierre Simon Laplace twice attempted to develop a theory of errors in the 1770s, it was not until 1810 that he published a theorem that justified Carl Friedrich Gauss's assumption that the appropriate model for measurement error is a normal distribution. Even after this breakthrough, it was another 65 years before Sir Francis Galton laid the groundwork for modern statistical analysis. After Galton's work it took an additional 50 years to fully develop modern techniques of statistical inference that allow us to successfully separate the potential signals from the probable noise.

Statistical inference is the name given to the group of techniques we use to make sense of our data. They work by either filtering out the noise to identify potential signals within our data, or by explicitly showing the uncertainty attached to an estimate of some quantity. This filtering of the noise and the computation of uncertainties are what distinguish statistical inference from naive interpretations of the data where every value is exact and every change in value is a signal. So how does statistical inference work?

ELEMENTS OF STATISTICAL INFERENCE

When we develop a statistical technique we begin with a probability model on the theoretical plane. Probability models are our starting point because they provide a mathematically rigorous description of how some random variable will behave. Using these models we can work out the properties of various functions of the random variables. Once we have a formula that works on the theoretical plane, we move from the theoretical plane to the data-analysis plane and use that formula with our data. In this way we have procedures that are consistent with the laws of probability theory. This allows us to obtain results that are both reasonable and mathematically justifiable. And that is how we avoid the trap of developing *ad hoc* techniques of analysis that

violate the laws of probability theory and confuse noise with signals.

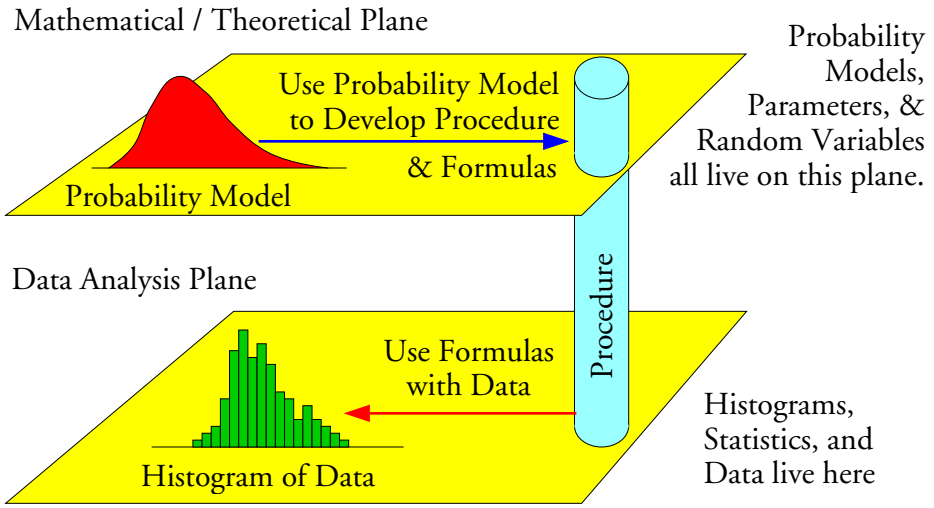

Figure 1: The Theoretical Plane and the Data Analysis Plane

Clear thinking requires that we always make a distinction between the theoretical plane, where we develop our procedures and formulas, and the data analysis plane where we use them. Probability models, parameter values, and random variables all live on the theoretical plane. Histograms, statistics, and data live on the data analysis plane. When we use techniques of statistical inference we frequently have to jump back and forth between these two planes. When we fail to make a distinction between these two planes, confusion is inevitable.

So what are some of the differences between these two planes? While random variables are usually continuous, data always have some level of chunkiness. While probability models often have infinite tails, histograms always have finite tails. While parameters are fixed values for a probability model, the statistics we use to estimate these parameters will vary with different data sets even though these data sets may be collected under the same conditions.

These differences between theory and practice mean that whenever a procedure or formula developed on the theoretical plane is used on the data analysis plane the results will always be approximate rather than exact. This fact of life is one of the better kept secrets of statistical inference. However, if the procedure is sound on the theoretical plane, and if the formula has been proven to be reasonably robust in practice, then we can be confident that our conclusions are reliable in spite of the approximations involved in moving from the theoretical plane to the data analysis plane.

Why do we play this game? Because all statistical inferences are inductive by nature. That is, they begin with the observed data and argue back to the source of those data. Since every inductive inference will involve uncertainty, we need to have a way to make allowance for this uncertainty in our analysis. The use of probability models allows us to make appropriate adjustments when we try to strike a balance between our choice of confidence level and the amount of ambiguity we want to have in our inference. Larger confidence levels (say using 99% instead of 95%) will always result in greater amounts of ambiguity (wider confidence intervals). Since the ambiguity increases faster than the confidence level, this trade-off between confidence

and ambiguity must be made in some rational manner. By working out the details on the theoretical plane, we can be reasonably certain that we end up making the appropriate adjustments in practice.

INTERVAL ESTIMATES OF LOCATION

A confidence interval for location is the first interval estimate most students encounter, so we will use it to illustrate the process shown in Figure 1. In developing the procedure on the theoretical plane the argument proceeds as follows:

1. Assume $\{ X_1, X_2, \ldots, X_n \}$ is a set of n independent and identically distributed normal random variables with unknown mean and variance.

2. To obtain an interval estimate for the parameter *MEAN(X)* we use some function of $\{ X_1, X_2, \ldots, X_n \}$ that is dependent upon the value of *MEAN(X)*. Specifically, in 1908 W. S. Gossett (Student) proved that the formula:

$$T = \frac{\bar{X} - MEAN(X)}{S} \sqrt{n}$$

will have a Student's T distribution with (n-1) degrees of freedom. Thus we know that:

$$\text{Prob}\left\{ -t_{0.05} < T < t_{0.05} \right\} = 0.90$$

$$= \text{Prob}\left\{ -t_{0.05} < \frac{\bar{X} - MEAN(X)}{S} \sqrt{n} < t_{0.05} \right\} = 0.90$$

3. Use the distribution of the random variable T to find a *random interval* that will bracket *MEAN(X)* with some specified probability. With a little work on the inequality within the brackets above we get:

$$\text{Prob}\left\{ \bar{X} - t_{0.05} \frac{S}{\sqrt{n}} < MEAN(X) < \bar{X} + t_{0.05} \frac{S}{\sqrt{n}} \right\} = 0.90$$

The probability that this *random interval* will bracket *MEAN(X)* is 90 percent.

Up to this point the argument has been carried out on the theoretical plane. The data are considered to be observations on random variables that are continuous, independent, and identically normally distributed.

So what happens when we move from the mathematical plane of probability theory down to the data-analysis plane where our data are chunky, our histograms always have finite tails, and our data are never generated by a probability model? We use the theoretical relationships and formulas above as our guide and compute a 90% confidence interval for *MEAN(X)* on the data-analysis plane according to the following:

4. Get n data: $\{ x_1, x_2, \ldots, x \}$

5. Compute the average statistic and the standard deviation statistic for these data:

$$\bar{x} \quad \text{and} \quad s$$

6. Find the Student's T critical value with n-1 degrees of freedom, $t_{.05}$ and compute the endpoints for an observed value of the random interval:

$$\bar{x} \pm t_{.05} \frac{s}{\sqrt{n}}$$

In theory, a 90% confidence interval computed in this manner should bracket *MEAN(X)* exactly 90 percent of the time. However, the approximation that occurs as we move from the theoretical plane to the data analysis plane means that in practice an interval calculated using the formula above *should* bracket *MEAN(X) approximately* 90 percentof the time.

LINE THREE EXAMPLE

Our first example will use the data from Line Three. In order to illustrate how interval estimates work, I used these 200 data to compute a sequence of forty 90% confidence intervals for the mean, each based on five values. While intervals based on such small amounts of data will be fairly wide, the point here is to see *how many* of these 90% confidence intervals bracket *MEAN(X)*.

The Line Three data and the 40 confidence intervals for the mean are given in Figure 8 in the appendix. The histogram for Line Three is found in Figure 4, and the forty confidence intervals are shown in Figure 2. If we consider the grand average of 10.10 to be the best estimate for *MEAN(X)*, then 37 out of 40, or 92.5 percent, of our intervals bracketed the mean. Thus, as expected, about 90 percent of our 90% confidence intervals work in this case.

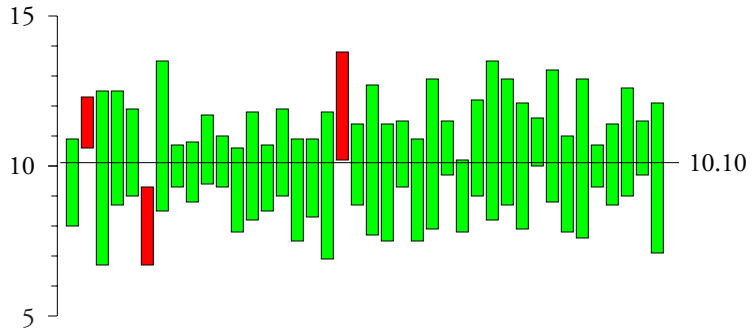

Figure 2: Forty 90% Confidence Intervals for *MEAN(X)* for Line Three

LINE SEVEN EXAMPLE

A second example is provided by the data from Line Seven. Once again, for illustrative purposes these 200 data are subdivided into 40 subgroups of size five and a 90% confidence interval for the mean is computed for each subgroup. The Line Seven data and confidence intervals are given in Figure 9 in the appendix. The grand average for Line Seven is 12.86. The histogram is found in Figure 4 and the forty 90% confidence intervals are shown in Figure 3.

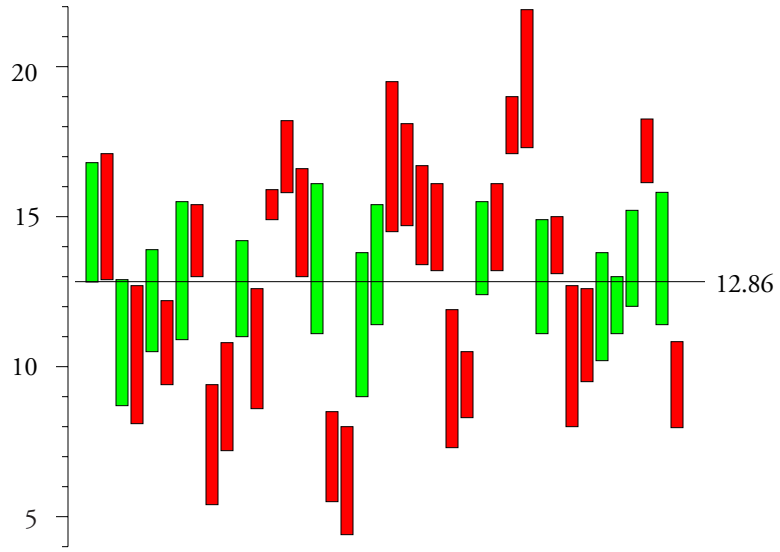

Figure 3: Forty 90% Confidence Intervals for *MEAN(X)* from Line Seven

Only fourteen of the forty 90% confidence intervals in Figure 3 contain the grand average value of 12.86! Thus, rather than working about 90 percent of the time as expected, the 90% confidence interval formula only worked 35 percent of the time with these data! So why did this happen?

"Is this a problem of the small amount of data used for each interval?" No, the 40 intervals of Figure 2 were also based on five values each, and they bracketed the grand average over 90 percent of the time.

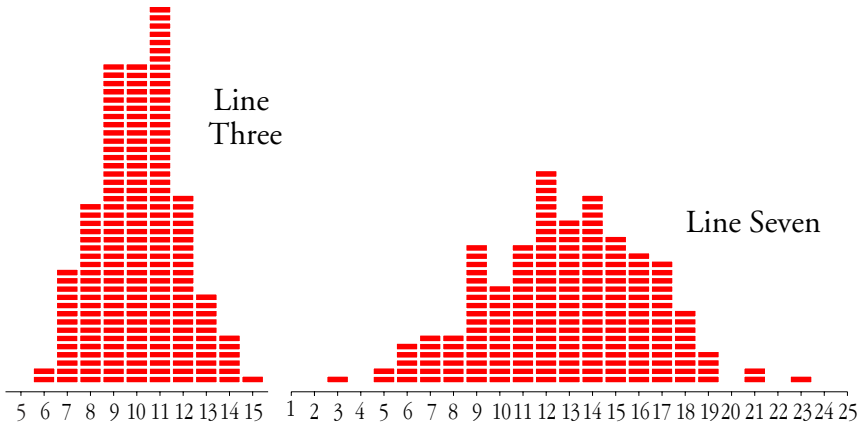

Figure 4: Histograms for Line Three and Line Seven

"Is this a problem with the 'normality' of the data?" No, not only are both data sets reasonably "normal," but the *t*-test and *t*-based confidence intervals have been known for over 60 years to be robust to departures from the normality assumption. The problem in Figure 3 has nothing to do with the shape of the histogram. Instead it has to do with the theoretical assumption that the random variables will be independent and identically distributed.

Virtually all statistical techniques begin with the assumption of independent and identically distributed random variables. (This is so common that it is often abbreviated as i.i.d. in statistical articles.) When this assumption is translated down to the data analysis plane it becomes an assumption that your data are homogenous.

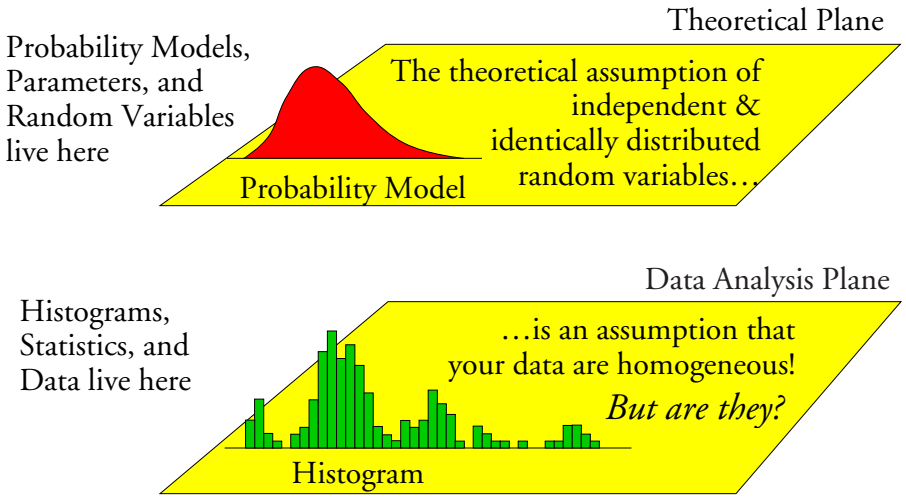

Figure 5: Homogeneity is a Necessary Condition for Statistical Inference

When your data are not homogeneous the techniques of statistical inference that were so carefully constructed on the theoretical plane become a house of cards that is likely to collapse in practice.

"How does a lack of homogeneity undermine the statistical inference?" It does not affect our ability to use the theoretical formulas—we were able to find all 40 confidence intervals in Figure 3 with no difficulty. No, rather than undermining the computations, a lack of homogeneity undermines *our ability to make sense of those computations*. The 90% confidence intervals of Figure 3 do not behave as expected simply because they are not all interval estimates of the same thing. If you assume you have homogeneous data when you do not, it is not your computations that will go astray, but rather your interpretation of the computed values that will be wrong.

WHY WE MISS THIS IN PRACTICE

"Why don't we see this problem when we use the various techniques of statistical inference?"

We miss this for the following reason. While the techniques of statistical inference were developed under the assumption of homogeneity, *they make no attempt to verify that assumption.* The formulas used in statistical inference are almost always symmetric functions of the data. Symmetric functions treat the data without regard to the time order of those data. (A change in the order of the data will not change the value of a symmetric function of those data.) Symmetric functions effectively make a *very strong assumption* of homogeneity. As a result, any lack of homogeneity will undermine the interpretation of the results.

For example, in a typical analysis we would never take the data from Line Three and Line Seven and break them down into subgroups of size five. We would simply dump all 200 data from each line into a computer and let it give us our interval estimates. For Line Three we would

get a 90% confidence interval for the mean of 9.90 to 10.31. For Line Seven we would get a 90% confidence interval for the mean of 12.45 to 13.26. In both cases everything would seem to be okay. There is absolutely nothing in these computations to warn us that the first of these intervals is a reasonable estimate while the second is patent nonsense.

THE QUESTION OF HOMOGENEITY

Virtually every statistical technique is developed using the assumption that, on some level, you are dealing with independent and identically distributed random variables. Because of this, the question of whether or not your data display the appropriate level of homogeneity has always been, and will always be, the primary question of data analysis.

This question trumps all other questions. It trumps questions about which probability model to use. It trumps questions about how to torture the data with transformations. It trumps questions about what alpha level to use. In truth, you cannot define an alpha level, you cannot fit a probability model, and you cannot hope that your statistical inferences will work as advertised if you do not have a homogeneous set of data. If your data are not reasonably homogeneous, it is the height of wishful thinking to imagine that a sophisticated mathematical argument is going to produce anything other than nonsense. Mere computations cannot cure a lack of homogeneity.

The process behavior chart is the premier technique for empirically checking for homogeneity. Unlike other statistical procedures which are gullible about the assumption of homogeneity, process behavior charts are skeptical about this assumption—they explicitly examine the data for evidence of a lack of homogeneity.

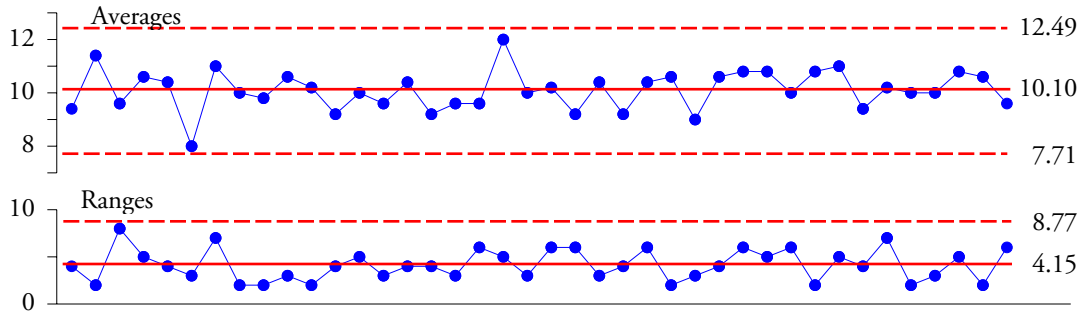

Figure 6: Average and Range Chart for Line Three

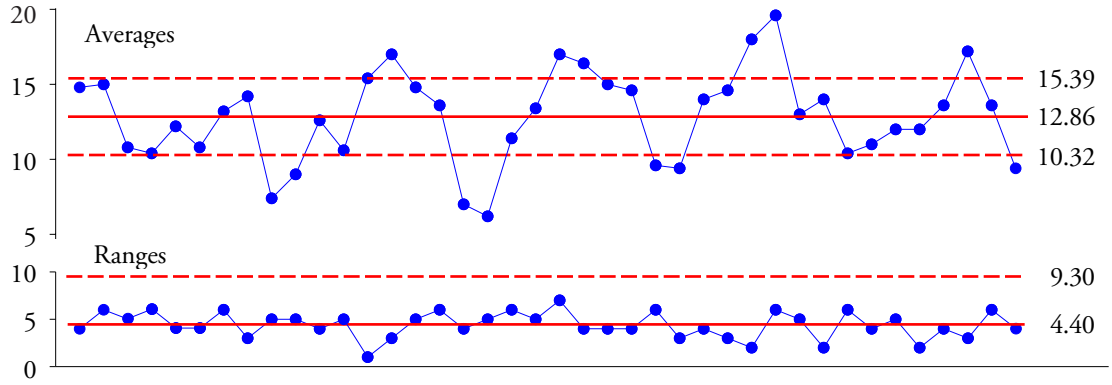

Figure 7: Average and Range Chart for Line Seven

The average and range chart in Figure 6 shows that the data from Line Three are reasonably homogeneous, while that in Figure 7 shows that the data from Line Seven are definitely not homogeneous. Any assumption that the data from Line Seven are identically distributed is inappropriate. There is not one process mean, but many, and the grand average of 12.86 is merely the average of many different things *rather than being an estimate of one underlying property for this process*.

Any analysis is seriously flawed when it does not begin with a consideration of whether or not the data display an appropriate degree of homogeneity.

WHAT ABOUT NORMALITY ?

In statistical inference the assumption of independent and identically distributed random variables is a *necessary condition*. Among other things it justifies the use of symmetric functions of the data, so that we need not be concerned with the time-order sequence of the data. However, as we have seen in Figure 3, if the i.i.d. assumption fails the whole theoretical structure fails, and the notion of underlying parameters vanishes. As noted above, while we may still calculate our statistics, they will no longer represent some underlying parameter.

"Well, if the independent and identically distributed part of the assumption is so important, isn't the normally distributed part equally important?" Not really. The assumption of normally distributed random variables is not a necessary condition, but merely a *worst-case condition* used as a starting point. To illustrate this consider one way we used to compute an estimate of the fraction nonconforming back in the dark ages before computers and capability ratios.

We would convert the specification limits into z-scores by subtracting off the average and dividing by our estimate of the process dispersion. Next we would use these z-scores with a standard normal distribution to obtain the tail areas outside the specifications. When we did this we would obtain *approximate, worst-case values* for the fraction nonconforming. That is, the fractions nonconforming obtained in this way from the normal distribution will either be the worst-case fraction nonconforming possible, or it will provide a reasonably close approximation to the worst-case value.

To understand this, consider the case where the process is centered within the specifications and compare the fractions nonconforming found using both a normal distribution and any chi-square distribution. For capability ratios in the range of 0.2 to 0.7 the normal fractions nonconforming will be greater than or equal to the chi-square fractions. (In some cases these normal fractions will be substantially greater than the chi-square fractions.) Thus, for fractions nonconforming ranging from 55 percent down to 5 percent the normal fractions dominate the corresponding chi-square fractions and are the worst-case values. For all other values of the capability ratio, the chi-square fractions nonconforming never exceed the normal fractions by more than 2 percent nonconforming. Thus, depending upon the capability ratio, using a normal distribution provides fraction nonconforming values that are either the worst-case value or else a close approximation of the worst-case value. You might be better off than what you find using the normal distribution, but you can't be appreciably worse off.

So, the assumption of normally distributed random variables is not a necessary condition, but simply a worst-case condition used as a starting point for the development of statistical techniques. When the techniques we develop under the assumption of a normal distribution turn

out to be robust in practice, we do not need to give any thought to whether or not the data appear to come from a normal distribution. Thus, with robust techniques, the worst-case assumption of normally distributed random variables is used as a starting point, *but it does not become a prerequisite that has to be verified in practice*.

Moreover, attempting to fit a probability model *before* testing for homogeneity is to get everything backwards. Homogeneity is a necessary condition before the notion of a probability model, or pretty much any thing else, makes sense. And the operational definition of homogeneity is a process behavior chart organized according to the principles of rational sampling and rational subgrouping (see my columns for June and July 2015)

And this is why anyone who suggests doing *anything* with your data *prior* to placing them on a process behavior chart is ignoring the secret foundation of statistical inference.

FOOD FOR THOUGHT

A recent release of Apple's OSX 10.10.5 (Yosemite) had 286 reviews posted in the App Store. On a rating scale from one to five stars these 286 reviewers gave the operating system an average rating of 2.96 stars.

The breakdown of these 286 reviews is as follows: 103 reviewers had given the software a rating of five stars; 24 gave it a rating of four stars; 23 gave it a rating of three stars; 31 gave it a rating of two stars; and 105 gave it a rating of one star. Thus, 44 percent of the reviewers loved it, 48 percent hated it, and 8 percent were ambivalent. So which of the two major groups was characterized by the average rating of 2.96 stars?

Without homogeneity, the interpretation of even the simplest of statistics becomes complicated.

POSTSCRIPT

In 1899, T. C. Chamberlin, a geologist, wrote:

> "The fascinating impressions of rigorous mathematical analysis,
> with its atmosphere of precision and elegance,
> should not blind us to the defects of the premises
> that condition the whole process.
> There is, perhaps, no beguilement more insidious and dangerous
> than an elaborate and elegant mathematical process
> built upon unfortified premises."

APPENDIX: LINE THREE DATA

	Line Three Data							*90% Conf. Interval* for *MEAN(X)*	
No.	Individual Values					*Average*	*Std Dev*	lower	upper
-1-	12	9	9	8	9	9.4	1.52	7.95	10.85
-2-	11	13	11	11	11	11.4	0.89	**10.55** *	12.25
-3-	14	6	11	8	9	9.6	3.05	6.69	12.51
-4-	10	12	13	8	10	10.6	1.95	8.74	12.46
-5-	11	10	12	11	8	10.4	1.52	8.95	11.85
-6-	7	10	7	7	9	8.0	1.41	6.65	**9.35** *
-7-	12	12	14	7	10	11.0	2.65	8.48	13.52
-8-	11	10	9	10	10	10.0	0.71	9.33	10.67
-9-	9	11	9	11	9	9.8	1.10	8.76	10.84
-10-	11	10	12	9	11	10.6	1.14	9.51	11.69
-11-	9	11	10	11	10	10.2	0.84	9.40	11.00
-12-	9	10	9	11	7	9.2	1.48	7.79	10.61
-13-	7	10	12	11	10	10.0	1.87	8.22	11.78
-14-	8	10	11	9	10	9.6	1.14	8.51	10.69
-15-	10	8	11	12	11	10.4	1.52	8.95	11.85
-16-	11	8	9	11	7	9.2	1.79	7.49	10.91
-17-	9	9	11	11	8	9.6	1.34	8.32	10.88
-18-	12	9	11	10	6	9.6	2.30	7.40	11.80
-19-	12	11	10	15	12	12.0	1.87	**10.22** *	13.78
-20-	9	12	11	9	9	10.0	1.41	8.65	11.35
-21-	13	11	12	7	8	10.2	2.59	7.73	12.67
-22-	9	9	13	8	7	9.2	2.28	7.03	11.37
-23-	10	12	9	11	10	10.4	1.14	9.31	11.49
-24-	8	8	12	10	8	9.2	1.79	7.49	10.91
-25-	9	13	7	10	13	10.4	2.61	7.91	12.89
-26-	11	10	12	10	10	10.6	0.89	9.75	11.45
-27-	11	9	9	8	8	9.0	1.22	7.83	10.17
-28-	11	12	8	12	10	10.6	1.67	9.00	12.20
-29-	11	14	8	13	8	10.8	2.77	8.15	13.45
-30-	14	12	9	9	10	10.8	2.17	8.73	13.87
-31-	9	11	13	10	7	10.0	2.24	7.87	12.13
-32-	10	11	10	12	11	10.8	0.84	10.00	11.60
-33-	9	10	9	13	14	11.0	2.35	8.76	13.24
-34-	12	10	9	8	8	9.4	1.67	7.80	11.00
-35-	9	7	14	12	9	10.2	2.77	7.55	12.85
-36-	9	10	10	11	10	10.0	0.71	9.33	10.67
-37-	11	8	11	9	11	10.0	1.41	8.65	11.35
-38-	12	11	13	8	10	10.8	1.92	8.97	12.63
-39-	9	11	11	11	11	10.6	0.89	9.75	11.45
-40-	13	11	10	7	7	9.6	2.61	7.11	12.09

Figure 8: 40 Subgroups of Size 5 from Line Three

APPENDIX: LINE SEVEN DATA

	Line Seven Data							*90% Conf. Interval* for *MEAN(X)*	
No.	Individual Values					*Average*	*Std Dev*	lower	upper
-1-	14	17	17	13	13	14.8	2.05	12.85	17.34
-2-	12	14	16	15	18	15.0	2.24	**12.87***	17.13
-3-	9	14	9	12	10	10.8	2.17	8.73	12.87
-4-	11	7	9	12	13	10.4	2.41	8.10	**12.70***
-5-	14	11	10	14	12	12.2	1.79	10.49	13.91
-6-	11	11	10	13	9	10.8	1.48	9.39	**12.21***
-7-	12	14	17	12	11	13.2	2.39	10.92	15.48
-8-	13	15	16	14	13	14.2	1.30	**12.96***	15.44
-9-	5	10	9	7	6	7.4	2.07	5.42	**9.38***
-10-	6	9	9	10	11	9	1.87	7.22	**10.78***
-11-	14	14	12	10	13	12.6	1.67	11.00	14.20
-12-	12	8	9	13	11	10.6	2.07	8.62	**12.58***
-13-	16	15	15	15	16	15.4	0.55	**14.88***	15.92
-14-	18	18	17	15	17	17	1.22	**15.83***	18.17
-15-	15	12	16	17	14	14.8	1.92	**12.97***	16.63
-16-	12	12	12	18	14	13.6	2.61	11.11	16.09
-17-	6	9	8	7	5	7	1.58	5.49	**8.51***
-18-	6	7	3	8	7	6.2	1.92	4.37	**8.03***
-19-	15	13	10	10	9	11.4	2.51	9.01	13.79
-20-	15	14	15	10	13	13.4	2.07	11.42	15.38
-21-	21	16	16	18	14	17	2.65	**14.48***	19.52
-22-	14	17	15	18	18	16.4	1.82	**14.67***	18.13
-23-	16	16	15	12	16	15	1.73	**13.35***	16.65
-24-	17	13	14	15	14	14.6	1.52	**13.15***	16.05
-25-	9	11	8	7	13	9.6	2.41	7.30	**11.90***
-26-	11	10	8	9	9	9.4	1.14	8.31	**10.49***
-27-	14	15	13	16	12	14	1.58	12.49	15.51
-28-	15	13	13	16	16	14.6	1.52	**13.15***	16.05
-29-	17	19	18	17	19	18	1.00	**17.05***	18.95
-30-	17	23	19	21	18	19.6	2.41	**17.30***	21.90
-31-	16	11	12	12	14	13	2.00	11.09	14.91
-32-	13	15	13	14	15	14	1.00	**13.05***	14.95
-33-	6	11	12	12	11	10.4	2.51	8.01	**12.79***
-34-	11	10	12	9	13	11	1.58	9.49	**12.51***
-35-	15	12	10	11	12	12	1.87	10.22	13.78
-36-	13	11	13	11	12	12	1.00	11.05	12.95
-37-	16	14	12	14	12	13.6	1.67	12.00	15.20
-38-	16	17	19	17	17	17.2	1.10	**16.16***	18.24
-39-	11	14	17	14	12	13.6	2.30	11.40	15.80
-40-	9	8	9	9	12	9.4	1.52	7.95	**10.85***

Figure 9: 40 Subgroups of Size 5 from Line Seven