Data as Evidence: Likelihood

by
Kevin T. Kilty

Formalities often dictate the course of an experiment.

For example, if the experimenter is a person who adheres to the Neyman-Pearson style of hypothesis testing, this person will follow a proscribed procedure as follows. Decide on a test statistic. Set a significance level for the experiment. Run the experiment, and only after the data are collected, calculate a test statistic and decide to reject a nominal hypothesis or provisionally accept its alternative.

Alternatively, experimenters who subscribe to the Fisherian philosophy will not set a significance level, but rather at the end of the experiment, will decide on the significance of the results through a calculated probability level. A small $p-$value indicates a significant experimental result. People familiar with the Analysis of Variance table that most spreadsheets produce for regression and curve fitting will recognize this as the "p" values provided for significance of model coefficients and F ratios.

In either case, there is no "peeking" at data in advance, and certainly no stopping the experiment before the data are all collected because the size of the sample is dictated by a design resolution, or in other words by a power of test. The experimenter will decide, on the basis of past experience and a calculation of standard error, the size of the sample (number of experiment runs and replications) required to achieve a design goal. The goal is usually to resolve a minimal difference in effect between treatments. This is only an estimate, of course, and the actual experiment may fail or succeed to achieve the goal. There are two disadvantages to such approaches to experimentation. These have to do with interpretation of the results–specifically with what evidence the results have with regard to one hypothesis or another–and with a sort of "mindlessness" that comes along with following hide-bound procedures.

Seeing significance as a measure of evidence

The first disadvantage is that people view the significance level, which means nothing more than how likely it is that such experimental results could arise by pure chance, as a measure of strength of evidence about the hypothesis.

To see why this is not necessarily a valid interpretation, consider the two graphs in Figure 1.

The curve labelled "nominal" pertains to the probability density of the nominal hypothesis. If the nominal hypothesis is correct then one expects the test statistic value to lie close to the mean of this distribution–in other words close to the peak of the curve near a value of 18. An experimental result that plots out on one of the tails of the nominal distribution, such as what the label "result" points to in Figure 1 at a value of 19, is "improbable" and a classical statistician would consider such a result as strong evidence against the nominal hypothesis and in favor of its alternative. But what is the alternative hypothesis? Often the alternative isn't specified directly. Possibly curve 2 describes the probability density of an alternative. In this case the experimental results are also very improbable if the alternative hypothesis is true. In this case the experimental data favors neither the nominal or alternative hypotheses, and in no case could the data be construed as strong evidence on way or the other.

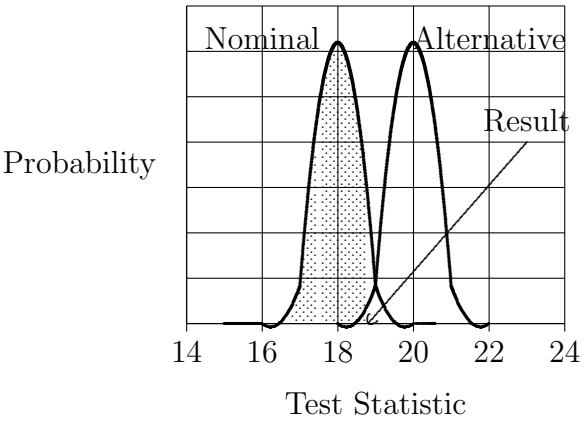

Figure 1: Diagram explaining why statistical significance does not necessarily represent strength of evidence.

Tests of significance classify errors as being Type I[1] or Type II[2]. But these two classifications focus on the hypothesis itself, which we do not know the truth of, whereas the logical focus of attention should be on the data as

[1]Rejecting the nominal hypothesis when in fact it is better than the alternative.

[2]Accepting the nominal hypothesis when its alternative is true.

evidence about the hypothesis. It seems more clear to think of evidence as being strong, weak, or misleading.

Finally, it is possible that the model a person used to design the experiment is mis-specified, and therefore the hypothesis is neither true nor false in a clear sense. Also it is easy to find a highly significant, but very incorrect model for any set of data. One should not place faith in the significance test as a measure of evidence strength.

Likelihood ratio

Some people see the likelihood ratio as the proper measure of evidence. Likelihood is nothing more than the probability of having acquired the observed data *a priori*. If, for example, the probability density of a random variable, X, is $P(X = x) = p(x)$, then the likelihood of a certain sequence of independent observations, $x_i \ \ for \ \ i = 0, \cdots, n$, is the product of the individual probabilities $L(data) = p(x_0)p(x_1)\cdots p(x_n)$. The likelihood function depends on parameters typically, so I should write likelihood as $L(data|\lambda_0, \lambda_1, \cdots \lambda_k)$, which means the likelihood of observing specific data *given* the k parameter values–a conditional probability. Because people perform experiments to detect differences in these parameters generally, it makes more sense to make the parameters the explicit focus of attention by writing $L(\lambda_0, \lambda_1, \cdots \lambda_k|data)$, which stands for the likelihood of these being the correct parameter values *given* the observed data.

Likelihood is pretty meaningless without comparing it to something else. Naturally the comparison is of one set of parameters against another, and so the likelihood ratio $L(\lambda_i|data)/L(\lambda_j|data)$ is the evidence in the current data for parameter value λ_i against λ_j. A large value of the ratio, or a very small one represent strong evidence[3], a small ratio value suggests weak evidence, and the probability of misleading evidence one can adjust by varying the size of the experiment just as one would in a test of significance.

Before I give the impression that likelihood is a panacea, I should mention that the common problem of testing by any means is that models and hypotheses are mis-specified, and likelihood ratio is no more help in this instance than is significance level. However, likelihood has appeal because it puts the horse before the cart so to speak. In order to use it a person has to

[3]It is customary to use values of $\frac{1}{8}$ and 8 or $\frac{1}{32}$ and 32 as boundaries separating weak from strong evidence.

have a true, or at least a defensible probability model to begin with rather than use one by default, and a person also has to specify the numerical difference between hypotheses being compared to one another. Using a model of normality just because one cannot think of anything better will lead to something resembling tests of significance once again. Unfortunately this is a common situation and so likelihood ratio and tests of significance often arrive at the same conclusion. Obviously if there were a huge difference between them in common usage, then people would have abandoned one for the other long ago. Never the less, likelihood ratio provides us with a tool to examine pieces of data as evidence, and to evaluate evidence as it arrives piece by piece.

Before embarking on an example involving real data, I'd like to discuss the second disadvantage of the usual approaches to tests of significance.

Running experiments to completion

The second disadvantage of Neyman-Pearson or Fisherian methodology involves this business about not peeking at data in advance. By not peeking at data I mean that experiments are run to completion according to the original design. The main reason for this is that the experimenter originally decided on a probability of Type II error[4] he or she was willing to tolerate, and sample size, or equivalently the number of runs and replications, partially determines this.

In addition, to terminate an experiment early is to leave oneself open to the accusation of choosing one's data–terminating the experiment early if the data are favorable, or letting the experiment run if they are not. However, in instances where the experimental results are actually very clear-cut after a few runs, this hide-bound insistence on running the experiment to completion also results in:

- wasted resources of time, money, and experimental materials; and,
- the possibility of ethical dilemnas regarding the experimental units if these happen to be animals or people.

How can one organize a means to look at data as it becomes available, and not stand accused of fudging? There are two distinct approaches to this that

[4]In statistics as generally practiced a Type II error is accepting a nominal hypothesis when it is not true. It is the same as saying "I made a mistake because of lack of resolution."

I know of. The first is the approach that manufacturing engineers take in what is called Statistical Process Control (SPC). SPC calls for operators of a process or machine to occasionally take small samples of product, calculate a statistic from this sample, and plot the result on a control chart. Examining the behavior of successive samples on this control chart allows the operator or engineer to detect improbable trends in the data that suggest the process is going awry. Improbable behavior might mean a trend in the data, or too many successive values within a single chart zone, or a few successive values beyond control limits[5].

A nurse or doctor does this same thing by keeping a chart of a patient's vital statistics. However, there is one extremely important difference between SPC and the typical medical control chart. The SPC engineer has fully characterized the process under control[6], whereas the doctor probably has not. By having a well understood baseline of operation the engineer knows of quirks in process operation that really do not constitute improbable behavior; whereas, the doctor may have to depend on population means as a baseline for monitoring a patient.

The other approach makes use of Bayes' theorem to update a prior probability about the hypothesis as each new data value become available from an experiment run. What amounts to the same thing is to constantly update a likelihood ratio.

An example using real data

Two climatologists argue about the frequency of major hurricanes in the Altantic Ocean. They argee that the Poisson distribution describes the variation in hurricane numbers from year to year, but one, influenced perhaps by recent history, insists the rate is at least 4 per year while the other feels it is only about 2.5. What evidence do observations of hurricane numbers provide? In Table 1 I provide the numbers of major hurricanes per year since 1944.

[5]Control limits are centered on process target and set anywhere from $\pm 2.57\sigma$ to $\pm 6\sigma$ depending on industry tradition.

[6]By fully characterize I mean that the engineer knows how the process behaves when running properly. There is an established baseline with a target value and a known variance to the product it produces. Furthermore the engineer will know about cycles that are part of the normal process.

Year	count	Year	count	Year	count	Year	count
1944	3	1958	4	1973	1	1987	1
1945	2	1959	2	1974	2	1988	3
1946	1	1960	2	1975	3	1989	2
1947	2	1961	6	1976	2	1990	1
1948	4	1962	0	1977	1	1991	2
1949	3	1963	2	1978	2	1992	1
1950	7	1964	5	1979	2	1993	1
1951	2	1965	1	1980	2	1994	0
1952	3	1966	3	1981	3	1995	5
1953	3	1967	1	1982	1	1996	6
1954	2	1968	0	1983	1	1997	1
1955	5	1969	3	1984	1	1998	3
1956	2	1970	3	1985	3	1999	5
1957	2	1971	3	1986	0	2000	3
—-	-	1972	3	—-	-	—-	-

Table 1: Major Altantic Hurricanes 1944-2000 Source: NOAA

The Poisson density, which our hypothetical scientists agree describe these variations[7], depends only upon the single parameter $rate = \lambda$,

$$P(count = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

Therefore the likelihood function for $n = 57$ observations of k_i hurricanes in year $1943 + i$ becomes,

$$L(\lambda | k_1, \cdots k_n) = \Pi_i (\frac{\lambda^{k_i}}{k_i!}) e^{-\lambda} \tag{2}$$

Where the symbol Π_i stands for the product of n terms. Because the product of exponential terms equals the exponential of the sum of such terms,

[7]Comparing these data to what one expects of a Poisson process shows that the generation of hurricanes in this region is surprisingly Poisson-like. I say *surprisingly* because the Poisson process presumes independent events, and hurricanes are not independent of one another in that each one alters the ocean surface behind it, and a very busy year may alter the ocean sufficiently to produce a slow year following.

$$L(\lambda|k_1,\cdots k_n) = e^{-n\lambda}\Pi_i(\frac{\lambda^{k_i}}{k_i!}) \quad (3)$$

The likelihood ratio of these data, using two different values for the rate is,

$$L(\lambda_1 = 4.0|k_1 = 3,\cdots k_n = 3)/L(\lambda_2 = 2.5|k_1 = 3,\cdots k_n = 3)$$
$$= e^{57(2.5-4.0)}\Pi_i(\frac{4.0}{2.5})^{k_i} \quad (4)$$

Therefore the numerical value of the likelihood ratio is 7×10^{-10}, which is overwhelmingly strong evidence in favor of the rate being 2.5 hurricanes per year rather than 4 or any rate greater[8]. In fact, the likelihood ratio drops below the *strong* evidence threshold of 1/8 at observation 19 which occurs in 1963, and never rises back into the *weak* evidence category[9].

When not completing an experiment is a virtue

Royall[10] summarized a particularly galling example of how strict adherence to a hide-bound procedure caused uneeded harm to experimental subjects. The example deals with clinical trials of a medical procedure[11]. It is highly unlikely an amateur scientist would ever become involved in testing on human subjects, but a situation similar to this one could arise in using animal subjects, or it may arise in regard to expensive or rare experimental materials.

This particular procedure dealt with respiratory problem of premature births. A particular procedure was known to have a success rate of only 20%. Some people associated with a new experimental procedure thought it might have a success rate as high as 80%. The design of the clinical trial was that of "winner continues." There are only two treatments involved–the

[8]Peak likelihood for these data occurs at a value of $\lambda = 2.3$.

[9]Comparing values as similar as 3.0 and 2.5 still provides very strong evidence ($LR < 1/32$) for the true value being 2.5, but values of 2.3 and 2.5 are almost indistinguishable from one another, the LR being 0.97.

[10]Royall, Statistical Evidence, Chapman-Hall, 1998. However, the story unfolds completely in the following papers: Bartlett, et al., Pediatrics, 76, 479-487, 1985. Ware, Statistical Science, 4, 298-340, 1989. UK EMCO Trial, The Lancet, 348, 75-82, 1996.

[11]Extracorporeal membrane oxygenation (ECMO).

current treatment (control) and the experimental one (cases). In this design a random choice leads to the assignment of one or the other treatment to the first subject. If the procedure is successful the trial continues to use it, otherwise a failure causes the trial to switch to the alternative treatment on the next subject. I mentioned in my book[12] that this design (protocol) can lead to a trial that gets stuck applying only one treatment repeatedly, and in the normal course of trials, run as they are according to the demands of the Neyman-Pearson or Fisherian philosophies, this presents a problem. Comparison of two treatments by typical statistics is most efficient when the controls and cases, the two *arms* of the trial, are of equal size[13].

In this particular trial the new treatment was exceptionally successful. The first 11 subjects survived, and the trial continued to apply the experimental treatment. Finally, the experimenters forced a switch to the alternative treatment on the 12^{th} subject, who did not survive. This resulted in lopsided arms to the trial, which resulted in biostatisticians arguing about the effectiveness of the new procedure, and leading eventually to not one additional trial, but several. If one examines the evidence of 11 straight successes with the new procedure in comparison with the established success rate of the old procedure, one finds a likelihood ratio (LR), using the binomial density, of

$$LR = (0.8)^{11}/(0.2)^{11} = 4,194,304 \tag{5}$$

Overwhelmingly in favor of the success rate of the new procedure being 0.8. In fact the evidence showed strongly (LR above 32) after the third run, and even if one believed the new procedure was only 50% better than the original one (0.3 success rate), the LR is 86.5, which still constitutes strong evidence.

The survival of 11 straight subjects with the new procedure, even though it was thought to be 80% effective[14] should have caused a reasonable person to hesitate to switch from it. Certainly an engineer or operator looking at a control chart for this sequence of subjects would have concluded the new treatment was highly successful. This run of successes combined with the extensive experience regarding effectiveness of the alternative treatment

[12] *Experiments and Studies*, BookSurge, 2003

[13] Also statisticians would prefer to have both trial arms represented within the current experiment, rather than rely on one arm that resulted from a separate study or experiment.

[14] Some people thought it might be less so

would cause such a reasonable person to say the trial is over. However, adherence to an inflexible, objective trial process causes people to become not reasonable but overly rational instead.

Data as Evidence: Missing and Censored Data

At this point I'd like to discuss a couple of situations that arise commonly with experimental data.

Let's begin by examining an experiment to gather data about hydraulic pump durability. The goal of this actual industrial experiment was to substantiate claims about a line of high quality hydraulic pumps. The current pumps installed on trucks were failing in a short time, the mean time between failure (MTBF) appeared to be about 6 months, and the company was looking for a source of more durable pumps. An alternative manufacturer advertizes that the MTBF of their pumps is one year.

In order to evaluate these new pumps, we organized an experiment that involved replacing each original pump that fails with a pump from the new manufacturer and keeping track of subsequent repair records. The experiment began on 12 March, 1997, and continued until 13 October, 2001. Data in Table 2 shows the performance of 26 pumps that failed during this time period. A cursory inspection of failure times suggests a MTBF much shorter than the manufacturer claims, perhaps only 0.6 years. However, there are two problems with the data in Table 2.

Missing data

First, by examining the table, one will see that there is missing data. Note that unit number 19 appears to be missing data from 2 May, 1997 to 24 November, 1997. What happened during this 7 month period is not known, but the pertinent question is "what should I do about the missing data?" My options are

- Ignore it.
- Replace missing data with a best guess of what its value probably was.

unit	Date		Time	Percentile
no.	installed	replaced	Years	Rank
28	7/5/2000	7/20/2000	0.042	0.019
18	2/14/2001	3/8/2001	0.067	0.058
15	3/9/2000	4/4/2000	0.069	0.096
16	3/28/2001	5/3/2001	0.097	0.135
8	8/2/2000	9/19/2000	0.131	0.173
19	3/12/1997	5/2/1997	0.139	0.211
16	5/3/2001	6/27/2001	0.15	0.25
5	6/22/2001	9/4/2001	0.2	0.29
28	2/9/2001	5/1/2001	0.228	0.33
30	4/4/2001	8/10/2001	0.35	0.36
28	2/14/2000	7/5/2000	0.39	0.40
28	9/9/1999	2/14/2000	0.43	0.44
28	5/1/2001	10/31/2001	0.5	0.48
28	7/20/2000	2/9/2001	0.55	0.52
19	11/24/1997	6/15/1998	0.558	0.55
19	1/11/1999	8/3/1999	0.56	0.59
19	6/15/1998	1/11/1999	0.57	0.63
29	9/19/2000	4/25/2001	0.6	0.67
30	8/1/2000	4/4/2001	0.675	0.71
8	11/2/1999	8/2/2000	0.75	0.75
5	5/3/2000	6/22/2001	1.136	0.78
28	7/2/1998	9/9/1999	1.186	0.82
5	2/17/1999	5/3/2000	1.211	0.86
19	1/16/2000	6/14/2001	1.411	0.90
19	8/3/1999	1/16/2001	1.453	0.94
18	4/14/1999	2/14/2001	1.833	0.981

Table 2: Pumps sorted by time in service

In the case of our pump survival data, we can just ignore the missing observation since nothing in the method of analysis, maximum likelihood estimation of MTBF in this case, depends on having this missing piece of data. This is also true of regression analysis, and many other analysis methods. However, when a person uses Analysis of Variance (ANOVA) to analyze multiway data, uses Latin Squares, or uses factorial designs, then missing

data upset the balance among and within treatments, which is an important element of the experiment design. Take as an example Table 3, below which is a randomized block design in which each block is meant to test all four treatments (A through D).

Treatment	Block					Totals
	1	2	3	4	5	
A	32.3	34.0	34.3	35.0	36.5	172.1
B	33.3	33.0	36.3	36.8	34.5	173.9
C	30.8	34.3	35.3	32.3	35.8	168.5
D	-NA-	26.0	29.8	28.0	28.8	112.6
Totals	96.4	127.3	135.7	132.1	135.6	627.1

Table 3: Randomized block experiment with one missing data value.

For whatever reason, the experimental run for treatment 'D' in the first block is missing. Perhaps someone was careless and forgot to write down the result; perhaps there was an accident. The point is that the experiment itself is threatened over this missing data. It is obvious that the issue of missing data is not solved so simply as just replacing it with an average value derived from the block, or an average from the treatment because there is an obvious trend to both the blocks and treatments. The best estimate of the missing value is,

$$X = \frac{tT + bB - S}{(t-1)(b-1)} \tag{6}$$

and one must adjust the treatments mean squares for bias by subtracting,

$$Bias = \frac{(B - (t-1)X)^2}{t(t-2)^2} \tag{7}$$

where,

$$t = \text{number of treatments} \tag{8}$$
$$b = \text{number of blocks} \tag{9}$$
$$T = \text{sum of results with the same treatment} \tag{10}$$
$$B = \text{sum of results within the same block} \tag{11}$$
$$S = \text{sum of all results} \tag{12}$$

The analysis of variance is done as I explained in *Design your experiments*[15] except that the total number of degrees of freedom is reduced by 1 to account for the missing data item. This replacement value will produce the correct treatment means, and correct mean square of experimental error in the resulting ANOVA table[16].

A Latin Square considers two categories of nuisance factors, call them row and column factors for now, and a number of treatments, n, to compare to one another. The design of a Latin Square requires as many rows and columns factors as there are treatments. Thus each category has n factors. One should replace a missing data item in a Latin Square with the value,

$$X = \frac{n(R + C + T) - 2S}{(n-1)(n-2)} \tag{13}$$

and adjust the mean squares for treatments by subtracting a bias of,

$$Bias = \frac{(S - R - C - (n-1)T)^2}{(n-1)^3(n-2)^2} \tag{14}$$

where,

$$n = \text{number of treatments, row factors, and column factors} \tag{15}$$
$$T = \text{sum of results with the same treatment} \tag{16}$$
$$R = \text{sum of results within the same row} \tag{17}$$
$$C = \text{sum of results within the same column} \tag{18}$$
$$S = \text{sum of all results} \tag{19}$$

[15] And in my book *Experiments and Studies*.

[16] See Snedecor and Cochran, *Statistical Methods*, U. Iowa Press, 1965, for further deatils.

Keep in mind one important rule about replacing missing data. The data truly have to be missing. If the experiment failed completely because of the treatment, then the data are not missing, the data are actually zero.

Censored data

Returning to Table 2 again, the second obvious problem with the data is that it implies many pumps which were installed on trucks, but which have not yet failed by the time we ended the experiment. In effect these are experiment runs that are not yet completed. Each of these pumps would eventually return information about MTBF, but we cannot wait for this to happen. This situation is so common that statisticians refer to the data as *censored*. In fact, in this case where the effect is to eliminate experiment runs with possibly the largest values, they refer to the data as *suspended*. I think the most insightful way to think about the experimental runs that complete as being samples that offered themselves to the experimenter. They offered themselves in exactly the way that respondents to the *Cosmopolitan* sex surveys volunteer information. In other words, they probably constitute a biased sample.

Before I speak about suspended data, let me make a few remarks about data censored for the reason that the values are too small to measure. These are truly called *censored* data. An example occurs in the running of chemical tests where many concentrations are too small for the test or instrument to measure. The experimenter will report these values as *below detection limit* or BDL. How should one go about calculating the mean and standard deviation (uncertainty) of data that contains censored values? There are several procedures in common use,

- Ignore the censored values. In effect, throw them out.
- Replace each censored value with a zero (0) value.
- Replace each censored value with the detection limit of the test in use.
- Replace each with $\frac{1}{2}$ of the detection limit.

What method a person decides to employ may depend on a traditional protocol, but I always suggest making the most conservative choice, which could be any of the above choices depending on the consequences of what

decisions may result. For example do large or small sample means cause a person to take some risky option? If so, then use replacement values that bias against this decision. In general, a person can always calculate descriptive statistics using all four of these choices, and make a table of the results to show the sort of uncertainty lurking within the data.

Suspended data are another matter. I've indicated that just throwing them out leaves one with a biased sample, and because the unknown values might be very large, the bias is possibly very large as well. At the web site "www.weibull.com" there is advice about how to handle such data through ranking the results. However, ranking methods sometimes result in a very large uncertainty about statistical measures of the data. The example at the weibull.com web site shows this explicitly.

An alternative is to predict what value(s) these data would have obtained had the experiment continued indefinitely[17]. Let me provide a simple explanation of how to do this with the pump failure data, even though what I propose is not valid in this particular instance.

In reliability experiments of this sort, one normally uses the Weibull density to model probability of failure. The Weibull density is,

$$P(TBF = t) = \lambda\beta(\lambda t)^{\beta-1}e^{-(\lambda t)^{\beta}} \tag{20}$$

where the parameter λ is a scale factor and β is a parameter to describe how the MTBF changes as parts age during use. An important characteristic of the Weibull density is that it contains enough parameters to describe how parts age during use. A much simpler probability density to use for analyzing MTBF is the exponential density, which contains a single parameter. Let's use λ for this parameter again–λ, then, is the inverse of MTBF. The failure density model is, therefore,

$$P(TBF = t) = \lambda e^{-\lambda t} \tag{21}$$

For the suspended data what I really need is a *conditional* density. It is,

[17]For example, in the third section of this paper I mention small world experiments. Dodds, et al., Science, 301, 827-829, 2003, tried to adjust the 98% of their trials that failed according to survival rates indicated in the mere 2% that succeeded, but I can't see how anyone could be convinced of the correctness of what resulted. In effect they are treating the undelivered items as suspended data–messages that would eventually reach the target with enough time. There is no reason to believe that these are suspended trials, rather than complete failures.

$$P'(TBF = t|t > s) = \lambda e^{-\lambda t} \tag{22}$$

The way to interpret this P' density is that it describes the probability of the time between failures (TBF) lying between t and $t + dt$ *given* the condition that the part has already survived for a time $t > s$ where s is the time the part accumulated in use before we suspended the experiment.

What is wonderful about the exponential density is that it is very simple to calculate P'. By just calculating the integral on the right hand side of Equation 17, and using the fact that the cumulative probability from the time of suspended usage, s, out to $t = +\infty$ equals 1, I find the stunning result that a part having an expected life equal to the MTBF before it is put in service, has an additional expected life equal to MTBF after it has been running for a time of s. It is just as though the part has never been in service at all–this hypothetical part does not wear out. In other words, the exponential density describes a process with *no memory*, which is why it is not a good model to use for reliability of real parts which do wear out in fact.

Nevertheless, using Table 2, I can make a summary of the censored (suspended) data, and list the minimum time between failures for each unfinished experiment run. In order to turn these suspended data into pseudo-data to use in my analysis, I'll simply add an expected MTBF to each suspended time. I have two possible values of MTBF. I can use the one the manufacturer claims, one year, or I can use what the finished experiments indicate.

In order to calculate a *best* value for MTBF, I'll maximize the likelihood function for n observations of time between failures, t_i. This is,

$$L(\lambda|t_1, \cdots t_n) = \lambda^n e^{-\lambda t_1} \cdots e^{-\lambda t_n} \tag{23}$$

Or, because the product of exponential terms equals the exponential of the sum of such terms,

$$L(\lambda|t_0, \cdots t_n) = \lambda^n e^{-\lambda \Sigma t_i} \tag{24}$$

From Table 2 $\Sigma_i t_i = 15.286$. This is probably biased, but using it tentatively I maximize this likelihood function in Equation 19 by simply trying various values of λ. This leads me to $\lambda = 1.7$, or an $MTBF = 0.59\, years$. Then, if I add this estimated MTBF to the usage times in Table 4, and treat these as observed times to failure, $\Sigma_i t_i$ is now 23.11, and the maximum of

unit	Date		Time
no.	installed	running as of	Years
28	-NA-	-NA-	-NA-
18	3/8/2001	10/13/2001	0.598
15	4/4/2000	10/13/2001	1.525
16	6/27/2001	10/13/2001	0.294
8	9/19/2000	10/13/2001	1.07
19	6/14/2001	10/13/2001	0.331
5	9/4/2001	10/13/2001	0.11
30	8/10/2001	10/13/2001	0.18
29	4/25/2001	10/13/2001	0.47

Table 4: Pumps still in service at study conclusion and their minimum service times

the likelihood function occurs at $\lambda = 1.4$, or in other words at a MTBF of 0.71 $years$. A second application of this scheme in which I use an MTBF of 0.71 produces a MTBF of 0.74 $years$, which is practically no change at all, so I will stop the process at this point. Indeed, these pumps do not appear to be quite as durable as their manufacturer suggests.

Conclusion

Accidents, instruments, measurement methods, and record keeping often result in missing, censored and suspended data. Sometimes such things do not affect an experimentat all, but at other times the experimenter really needs to address these occurrences and repair the data accordingly.

Data as Evidence: Circular Logic

We in the sciences like to think of our work as an objective adjudication of truth, but now and then, perhaps more than even I suspect, scientists engage in completely invalid thinking that infects their hypotheses and experimental results. One example of such invalid reasoning is *circular reasoning*. Circular reasoning is assuming beforehand what it is you intend to demonstrate through an experiment, study or proof. In other words, the hypothesis has become an axiom, or something close to an axiom. One may wonder how such a thing could possibly happen other than by shear fraud, because it seems so easy to avoid. The most common way this happens is that scientists fail to carefully exam the axiomatic elements of their experiment design–that is, the instruments and procedures they depend on provide circularity. The five examples I provide below illustrate how this happens.

The Fudging of Isaac Newton

My regular reading of *Science* during graduate school occasionally paid off handsomely. Among the articles that made a lasting impression on me, and one that encouraged my sense of skepticism, was Westfall's summary of the deliberate deceptions of Isaac Newton[18].

In his *Principia* Issac Newton presented examples of his mechanical system of the world. These displayed the power of his system as much as they illustrated its uses. Through many successive editions Newton worked with his editor to revise applications and maintain conformance with experimental findings. One body of experimental findings involved the speed of sound in air, a topic that required continual alteration in successive editions.

Newton understood very well the mechanical principles involved in sound propagation, but the specific details eluded him because of an incomplete understanding of heat transfer in such a fast process. As a result Newton used the value of isothermal compressibility of air rather than isentropic compressibility, which would have rendered his estimates of sound speed wrong by 20% or so. However, in order to maintain the illusion that all was right with his mechanical system he engaged in a pattern of fudging the theory with ingenious, but unfounded and indefensible"corrections" to his calculations. According to Westfall

[18] *Newton and the fudge factor*, Science, 179, 751-8, 1973.

> In examining the alterations, let us start with the velocity of sound since the deception in this case was patent enough that no one beyond Newton's most devoted followers was taken in. Any number of things were wrong with the demonstration. It calculated a velocity of sound in exact agreement with Derham's figure, whereas Derham himself had presented the conclusion merely as the average of a large number of measurements. Newton's assumptions that air contains vapor in the quantity of 10 parts to 1 and that vapor does not participate in the sound vibrations were wholly arbitrary, resting on no empirical foundation whatever. And his use of the "crassitude" of the air particles to raise the calculated velocity by more than 10 percent was nothing short of deliberate fraud.

Newton always knew what value for the speed of sound his theory needed to reproduce, which allowed him to fudge exact correction factors. This was circular reasoning, perhaps even outright dishonesty. But more important is this observation. Through circular reasoning Newton managed to justify mechanical corrections that were non-existent, but which nearly all people failed to appreciate. A well made circular argument can prove nearly anything.

While this example came from the earliest 18^{th} Century, the following ones are all very recent. I have chosen one each from physical sciences, sociology, forensics, and medical science.

Climate Obtained from Boreholes

There is no way for me to explain the nature of this circular argument without referring to two separate studies. The first one illustrates a method of fitting data to a non-linear model of heat conduction; while the second study merely summarizes what particular borehole temperature records have to say about climate over the past five centuries.

The inverse method

The objective of this method is to use temperature measured in a borehole to figure what the ground surface temperature (GST) was like over the past 1000 years. To do this the authors organize a model of temperatures in the subsurface based on heat transport by conduction, combined with a computer program that uses the observed temperatures to find the parameters

of the model. At each iteration this computer model will provide a set of parameters[19], m, and also the likely borehole temperature observations that result from it, d. The central idea is to begin with an initial set of parameters for the model, which I denote here symbolically as m_o, a set of temperature observations, d_o, and systematically adjust the parameters to minimize an objective function like the following...

$$S = \text{misfit to observations} + \text{change from initial model} \tag{25}$$

Specifically the objective function in matrix form is

$$S = (d - d_o)^t C_d^{-1}(d - d_o) + (m - m_o)^t C_m^{-1}(m - m_o). \tag{26}$$

The reasoning behind this objective is as follows. The penalty for misfit to observed data ($(d - d_o)^t C_d^{-1}(d - d_o)$) helps insure that whatever ground surface temperature history results from the analysis conforms to the observed borehole temperatures somewhat. The penalty for deviating from the initial model ($(m - m_o)^t C_m^{-1}(m - m_o)$) prevents the resulting surface temperature history from obtaining "unrealistic" oscillations. Values in the matrices C_m and C_d provide an optimum balance between honoring the observed data (d_o) and adhering to an *a priori* model (m_o) of GST.

However, a common problem in this instance is that heat conduction destroys information regarding long past temperature quite completely, and, therefore, many extremely different different temperature histories explain borehole data equally well. Where ever thermal diffusion has destroyed information, there is very little resulting variation in $(d - d_o)$ to tradeoff against $(m - m_o)$, resulting in a final model unduly influenced by *a priori* assumptions. Furthermore, the investigators set C_m and C_d to allow four times more recent GST variation than they allow freedom to vary 1000 years ago.

Crux of the circular argument

There is no invalid argument implied in anything the investigators have done in this first study. However, in a subsequent study the same authors use this inverse method and choose as their initial model one of zero (none) past temperature variation. They claim that using an initial model of zero temperature variation is neutral, but it is surely not. It is merely one choice

[19]Including the ground surface temperatures.

among an infinite number of possible choices. The conclusion they found was that while temperature increased quite rapidly over the past 150 years, temperature 500 to 1000 years ago was very stable in comparison. Therefore, they reason, the current rate of temperature increase is unnatural, and must be man-caused.

But given the characteristics of their method, and how they claim to set the matrices C_m and C_d, this finding is entirely expected no matter how surface temperature was behaving 500 to 1000 years ago. The authors interpret the zeroed GST output of their method as being characteristic of climate rather than being characteristic of their assumptions and method. Their conclusion may turn out to be perfectly correct for unrelated reasons, but their argument certainly appears circular and therefore invalid.

The Small World Studies

In 1967 a psychologist described an experiment to demonstrate what is now called *six degrees of separation*[20]. This has become a new age notion that all people have unrecognized connection to one another through family, friends, and acquaintances, and that this chain of connection between arbitrary total strangers is surprisingly short. The typical claim is that it is about 6 links in length.

There are a number of objections to the original experiment carried out in 1969[21], and others which attempted to replicate the results[22], but one flaw involves a circular logic.

The original experiment had people in Kansas (subjects) attempt to contact people whom they did not know personally in Massachusetts (targets) through the mail. A recent experiment uses the internet. In both cases the investigators are using experimental apparatus (post office or internet) that makes communications possible when there are no social ties at all. Unless there is a careful control for this factor, then some messages advance irrespective of social networks[23].

[20]Milgram. Psychology Today, 1, 61-67, 1967.

[21]See Kleinfeld's thoughts mentioned in Six Degrees of Uncertainty. Science. 294, 777, 26 Oct 2001. More details regarding her thoughts are found at Kleinfeld, Society, 39, 61, 2002.

[22]Dodds, et al., Science, 301, 827-829, 2003.

[23]The 1969 researchers noted a huge discrepancy in success depending on how official the letters appeared, but never made an allowance for it. The 2003 researchers produced a

Let me make a clarifying analogy. Suppose that I claimed I had measured something, and I had found its value to be 2.718. Suppose I admitted that my apparatus would often produce the value 2.718 even when I did not have any signal applied to it. Surely, you would doubt that I could claim the result of my measurements was 2.718. Small world experiments contain an element of circular logic of just this sort. The experimental apparatus on its own provides the sort of thing the experiment intends to measure. I don't know how much the results depend on this mistake, perhaps only a little, but a better experiment would have used something other than the post office or internet.

The Baltimore Fiasco

Teresa Imanishi-Kari was alleged to have falsified experiments, and data, in her laboratory notebooks. She was exonerated of all charges in 1996, but only after a Kafka-esque decade in which the charges against her continually metamorphosed, during which she could not actually examine the specific charges or evidence against her, and during which she was characterized as an example of corrupt science by government oversight committees. One of the principal charges against her was that she had fabricated a laboratory notebook to bolster her claim of having performed experiments on immune response – experiments which lead to an acclaimed paper in the journal *Cell*.

Imanishi-Kari's notebooks were apparently very messy. They contained paper strips of instrument readings that were taped to pages, and which were in varying colors and degrees of fadedness, printed with varying ribbons on a variety of printers. The pages themselves appeared out of order, dates were scratched through and corrected, there was white-out correction fluid used back-to-back on both sides of some sheets, and mechanical impressions suggested that pages were, in fact, dated out of the order they were written. The U. S. Secret Service was asked to perform forensic analysis on the notebooks to determine whether this messiness was the product of a hurried attempt to deceive. Part of the forensic analysis compared Imanishi-Kari's notebooks against a sample of the notebooks of other scientists.

---

study design that appears to contain most of the same flaws. It uses self-reporting of non-random volunteers. The success rate is a dismal 3%, but the researchers make adjustments for the unsuccessful 97% as though these were merely censored data, and so forth. They used self-reporting to try to correct for spurious connection due soley to the internet.

Secret Service investigators were unable to examine more than a tiny fraction (an estimate was perhaps 1%) of the notebooks produced by researchers in the same laboratory during the same time period. In the forum of the hearing, Imanishi-Kari's counsel cross-examined one of the Secret Service investigators about why he had not included for examiniation the laboratory notebook of someone who was a research collaborator of Imanishi-Kari at an earlier time. The agent testified that they excluded this notebook because it looked a lot like that of Imanishi-Kari so they were

> ... unwilling to use that as an example of what a normal or a usual notebook would be.

Since an explicit goal of the comparison was to establish the normalcy of Imanishi-Kari's notebook, the decision to exclude other notebooks becomes neat circular logic.

- Collect a sample of notebooks to establish a norm.
- Exclude from this sample notebooks that appear similar to that of the subject.
- Conclude that the subject notebook is unusual and abnormal.

Hormone Replacement Therapy

Many women suffer real physical symptoms of menopause and afterward including night sweats, hot flashes, irritability, and clumsiness. The lives of these women are modestly to greatly improved through hormone replacement therapy (HRT). However, a recent study concluded that HRT increases a woman's risk of breast cancer, heart disease, and strokes, and *that women show very little difference in their quality of life by going on HRT* compared to women on a placebo.[24] This is in striking contrast to what mountains of anecdotal evidence suggest.[25]

What the study, the Women's Health Initiative, found before it was halted in July 2002, was that of every 10,000 women taking HRT there were:

[24] See an article on this topic, *Look beyond scary news on HRT*, Fairbanks News-Miner Heartland, Sunday, September 7, 2003.

[25] I'm not a fan of having anecdotal evidence drive research and experimentation, but in this case there is so much anecdotal evidence that I consider it statistically significant.

- Seven more heart disease events.
- Eight more strokes.
- Eight more invasive breast cancers.
- Six fewer colorectal cancers, and,
- Five fewer hip fractures.

The last two items on this list received far less publicity than the first three scary items. It is true that HRT did not provide benefits with regard to heart disease and cancer, which was the point of the WHI study. More to the point of this paper, however, the subconclusion of the study that HRT provides only limited improvement of life quality seems founded on circular reasoning. The researchers excluded from their study women who had ever experienced severe postmenopausal symptoms. I am told[26] that the criterion for excluding a participant in the study was whether or not she had ever had a *hot flash*. In other words, the study screening excluded those women most likely to benefit from HRT, and those most likely to see their quality of life improve. By excluding such women the study should have remained silent on the issue of quality of life.

I have several other examples of circular reasoning in medical studies, including the misdiagnosis of Japanese Encephalitis, and the example of the near misdiagnosis of Toxic Oil Syndrome. These examples are fully explained in my book *Experiments and Studies*.

Conclusions

If we are to see data as evidence, then this data has to be gathered faithfully without bias. Circular reasoning strikes directly at the heart of these fair characteristics; and, though it seems easy enough to avoid, circular reason is, in fact, insidious.

Yes, Issac Newton was a deceiver. He also remains, in my mind, among the true giants of physical science. I hesitate to call what he did fraud in the current sense, prefering to think that he acted out of an absolute, blinding belief that his mechanical system was completely correct, without recognition that it was incomplete. I'd call this a cautionary tale about abandoning one's

[26]Kleinfeld, personal communication, October 4, 2003

skepticism. It also occured well before anyone had formulated the idea of a *Fair Test*[27] We know today that a fair test cannot contain biases and invalid reasoning.

The borehole temperature example shows how circular reasoning derives from excessive exhuberance combined with indirect methods of measurement that depend on long chains of mathematical correction and reasoning. I maintain that it is always best to employ simple, direct measurements.

Small world problems are difficult to study without using a communications network. Yet the network itself makes the world small. Removing this circularity is really a challenge for clever experiment design. I've not seen one yet.

The prosecution of Teresa Imanishi-Kari illustrates a sort of circular logic common to criminal and civil courts. Get the defendant! Luckily our adversarial legal system manages to provide some needed balance. In particular the "beyond a reasonable doubt" standard of proof is essential.

The study on HRT therapy shows how circular reasoning can result from trying to push results of a study far beyond its valid inference space. How much current research is a victim of something similar? What is perhaps worst about the HRT example is that the general public are now left with an impression that HRT is not useful, but is, rather, dangerous. This is a medical disservice.

Finally the examples of Japanese Encephalitis and Toxic Oil Syndrome both suffered from scientists involved assuming that a test with an entirely predictable outcome would accurately affirm a causative agent. Nothing that is entirely predictable provides useful information unless the entirely predictable outcome fails to materialize.

[27]Refer to the James Lind Library at www.jameslindlibrary.org for an explanation of what constitutes a fair test.