LECTURE 27

UNRELIABLE DATA

It has been my experience, as well as that of many others who have looked, that data is generally much less accurate than it is advertised to be. This is not a trivial point - we depend on initial data for many decisions, as well as for the input data for simulations which result in decisions. Since the errors are of so many kinds, and I have no coherent theory to explain them all, I have therefore to resort to isolated examples and generalities from them.

Let me start with life testing. A good example is my experience with the life testing of the vacuum tubes that were to go into the first voice carrying submarine cable with the hoped for life time of 20 years. (After 22 years we simply removed the cable from service since it was then too expensive to operate - which gives a good measure of technical progress these days.) The tubes for the cable first became available something like 18 months before the cable was to go down. I had a moderate computer facility, including a special IBM 101 statistical sorter, and I made it available to the people who were processing the data, as well as helping them do the more technical aspects of the computing. I was not, however, in any way involved in the direct work of the project. Never-the-less, one day one of the higher ups in the project showed me the test equipment in the attic. Being me, after a time I asked, "Why do you believe that the test equipment is as reliable as what is being tested?" The answer I got convinced me that he had not really thought about it, but seeing that pursuit of the point was fruitless, I let it drop. But I did not forget the question!

Life testing is increasingly important and increasingly difficult as we want more and more reliable components for larger and larger entire systems. One basic principle is accelerated life testing, meaning mainly that if I raise the temperature 17° Centigrade then most, but not all, chemical reactions double their rate. There is also the idea that if I increase the working voltage I will find some of the weaknesses sooner. Finally, for testing some integrated circuits, increasing the frequency of the clock pulses will find some weaknesses sooner. The truth is, all three combined are hardly a firm foundation to work from, but in reply to this criticism the experts say, "What else can we do, given the limitations of time and money?" More and more, the time gap between the scientific creation and the engineering development is so small that there is no time to gain real life testing experience with the new device before it is put into the field for widespread use. If you want to be certain then you are apt to be obsolete.

Of course there are other tests for other things besides those mentioned above. So far as I have seen the basis of life testing is shakey; but there is nothing else available. I had long ago argued at Bell Telephone Laboratories that we should form a life testing department whose job is to prepare for the testing of the next device that is going to be invented, and not just test after the need arises. I got nowhere, though I made a few, fairly weak, suggestions about how to start. There was not time in the area of life testing to do basic research - they were under too much pressure to get the needed results tomorrow. As the saying goes,

"There is never time to do the job right,
but there is always time to fix it later."

especially in computer software!

The question I leave with you is still, "How do you propose to test a device, or a whole piece of equipment, that is to be highly reliable, when all you have is less reliable test equipment, and with very limited time to test, and yet the device is to have a very long lifetime in the field?" That is a problem that will probably haunt you in your future, so you might as well begin to think about it now and watch for clues for rational behavior on your part when your time comes and you are on the receiving end of some life tests.

Let me turn now to some simpler aspects of measurements. For example, a friend of mine at Bell Telephone Laboratories, who was a very good statistician, felt that some data he was analyzing was not accurate. Arguments with the department head that they should be measured again got exactly nowhere since the department head was sure that his people were reliable and furthermore the instruments had brass labels on them saying that they were that accurate. Well, my friend came in one Monday morning and said that he had left his brief case on the railroad train going home the previous Friday and had lost everything. There was nothing else the department head could do but call for remeasurements, whereupon my friend produced the original records and showed how far off they were! It did not make him popular, but did expose the inaccuracy of the measurements which were going to play a vital role at a later stage.

The same statistician friend was once making a study for an outside company on the patterns of phone calling of their headquarters. The data was being recorded by exactly the same central office equipment that was placing the calls and writing the bills for making the calls. One day he chanced to notice that one call was to a nonexistent central office! So he looked more closely, and found that a very large percentage of the calls were being connected for some minutes to nonexistent central offices! The data was being recorded by the same machine that was placing the calls, but there was bad data anyway. You cannot even trust a machine to gather data about itself correctly!

My brother, who worked for many years at the Los Angles Air

Pollution, once said to me that they had found it necessary to take apart, reassemble, and recalibrate every new instrument that they bought! Otherwise they would have endless trouble with accuracy, and never mind the claims made by the seller!

I once did a large inventory study for Western Electric. The raw data they supplied was for 18 months of inventory records on something like 100 different items in inventory. I asked the natural question of why I should believe that the data was consistent - for example, could not the records show a withdrawal when there was nothing in inventory? They claimed they had thought of that and had in fact gone through the data and added a few pseudo transactions so that such things would not occur. Like a fool I believed them, and only late in the project did I realize that there were still residual inconsistencies in the data, and hence I had first to find them, then eliminate them, and then run the data all over again. From that experience I learned never to process any data until I had first examined it carefully for errors. There have been complaints that I would take too long, but almost always I found errors and when I showed the errors to them they had to admit that I was wise in taking the precautions I did. No matter how sacred the data and urgent the answer, I have learned to pretest it for consistency and outliers at a minimum.

I once became involved as an instigator and latter as an advisor to a large AT&T personnel study using a UNIVAC in NYC that was rented for the job. The data was to come from many different places, so I thought that it would be wise to have a pilot study run first to make sure that the various sources understood what was going to happen and just how to prepare the IBM cards with the relevant data. This we did. But when the main study came in some of the sources did not punch the cards as they had been instructed. It took only a little thought on my part to realize that of course the pilot study being small in size went to their local key punch specialty group, but that the main study had to be done by the central group. Unfortunately for me they had not understood the purpose of the pilot study! Once more I was not as smart as I thought I was; I did not appreciate the inner workings of a large organization.

But how about basic scientific data? In an NBS publication on the 10 fundamental constants of physics, the velocity of light, Avagadro's number, the charge on the electron, etc. there were two sets of data with their errors. I promptly noted that if the second set of data were taken as being right (and the point of the table was how much the accuracy had improved in the 24 years between compilations), then the average amount that the new values fell outside the old errors was 5.267 as far, the last column which was added by me, Figure 27-1. Now you would suppose that the values of the physical constants had been carefully computed, yet how wrong they were! The next compilation of physical constants showed an average almost half as large, Figure 27-2. One can only wonder what another 20 or so of years will reveal about the last cited accuracy! Care to bet?

This is not unusual. I very recently saw a table of measurements of Hubble's constant (the slope of the line connecting the red shift with distance) which is fundamental to most of modern cosmology. Most of the values fell outside of the given errors announced for most of the other values.

By direct statistical measurement, therefore, the best physical constants in the tables are not anywhere near as accurate as they claim to be. How can this be? Carelessness and optimism are two major factors. Long meditation also suggests that the present experimental techniques you are taught are also at fault and contribute to the errors in the claimed accuracies. Consider how you, in fact as opposed to theory, do an experiment. You assemble the equipment and turn it on, and of course the equipment does not function properly. So you spend some time, often weeks, getting it to run properly. Now you are ready to gather data, but first you fine tune the equipment. How? By adjusting it so that you get consistent runs! In simple words, you adjust for low variance; what else can you do? But it is this low variance data that you turn over to the statistician and is used to estimate the variability. You do not supply the correct data from the correct adjustments - you do not know how to do that - you supply the low variance data, and you get from the statistician the high reliability you want to claim! That is common laboratory practice! No wonder the data is seldom as accurate as claimed.

I offer you Hamming's rule:

90% of the time the next independent measurement will fall outside the previous 90% confidence limits!

This rule is in fact a bit of an exaggeration, but stated that way it is a memorable rule to recall - most published measurement accuracies are not anywhere near as good as claimed. It is based on a lifetime of experience and represents later disappointments with claimed accuracies. I have never applied for a grant to make a properly massive study, but I have little doubts as to the outcome of such a study.

Another curious phenomenon that you may meet is that in fitting data to a model there are errors in both the data and the model. For example, a normal distribution may be assumed, but the tails may in fact be larger or smaller than the model predicts, and possibly no negative values can occur although the normal distribution allows them. Thus there are two sources of error. As your ability to make more accurate measurements increases the error due to the model becomes an increasing part of the error.

I recall an experience I had while I was on the Board of Directors of a computer company. We were going to a new family of computers and had prepared very careful estimates of costs of all aspects of the new models. Then a salesman estimated that if the selling price were so much then he could get orders for 10,

if another price 15, and another 20 sales. His guesses, and I do not say they were wrong, were combined with the careful engineering data to make the decision on what price to charge for the new model! Much of the reliability of the engineering guesses was transferred to the sum, and the uncertainty of the salesman's guesses was ignored. That is not uncommon in big organizations. Careful estimates are combined with wild guesses, and the reliability of the whole is taken to be the reliability of the engineering part. You may justly ask why bother with making the accurate engineering estimates when they are to be combined with other inaccurate guesses; but that is wide spread practice in many fields!

I have talked first about Science and Engineering so that when I get to economic data you will not sneer at them too much. A book I have read several times is Morgenstern's On the Accuracy of Economic Measurements, Princeton Press, 2nd. ed. He was a highly respected Economist.

My favorite example from his book is the official figures on the gold flow from one country to another, as reported by both sides. The figures can differ at times by more than two to one! If they cannot get the gold flow right what data do you suppose is right? I can see how electrical gear shipped to a third world country might get labeled as medical gear because of different import duties, but gold is gold, and is not easily called anything else.

Morgenstern points out that at one time DuPont Chemical held about 23% of the General Motors stock. How do you suppose this appeared when the Gross National Product, (GNP), figure was computed? Of course it was counted twice!

As an example that I found for myself, there was a time, not too long ago, when the tax rules for reporting inventory holdings were changed, and as a result many companies changed their methods of inventory reporting to take advantage of the new reporting rules, meaning that they now could show smaller inventory and hence get less tax. I watched in vain in the Wall Street Journal to see if this point was ever mentioned. No, it never was that I saw! Yet the inventory holdings are one of the main indices that are used to estimate the expectations of the manufacturers, whether we are headed up or down in the economy. The argument goes that when manufacturers think that sales will go down they decrease inventory, and when they expect sales to go up they increase inventory so that they will not miss some sales. That the legal rules had changed for reporting inventory and was part of what was behind the measurements was never mentioned, so far as I could see.

This is a problem in all time series. The definition of what is being measured is constantly changing. For perhaps the best example, consider poverty. We are constantly upgrading the level of poverty, hence it is a losing game trying to remove it - they will simply change the definition until there are enough of people below the poverty level to continue the projects they

manage! What is now called "poverty" is in many rspects better than what the Kings of England had not too long ago!

In a Navy a yeoman is not the same yeoman over the years, and a ship is not a ship, etc., hence any time series that you study to find the trends of the Navy will have this extra factor to confound you in your interpretations. Not that you should not try to understand the situation using past data, (and while doing it apply some sophisticated signal processing, Lectures 14-17), but that there are still troubles awaiting you due to changing definitions which may never have been spelled out in any official records! Definitions have a habit of changing over time without any formal statement of this fact.

The forms of the various economic indices that you see published regularly, including unemployment (which does not distinguish between the unemployed and the unemployable but should be in my opinion), were made up, usually, long ago. Our society has in recent years changed rapidly from a manufacturing to a service society, but neither Washington, D.C. nor the economic indicators have realized this to any reasonable extent. Their reluctance to change the definitions of the economic indicators is based on the claim that a change, as indicated in the above paragraph, makes the past noncomparable to the present - <u>better to have an irrelevant indicator than an inconsistent one</u>, so they claim. Most of our institutions (and people) are slow to react to changes such as the shift to service from manufacturing, and even slower to ask themselves how what they were doing yesterday should be altered to fit tomorrow. Institutions and people prefer to go along smoothly, and hence lag far behind, than to make the effort to be reasonably abreast of the times. <u>Institutions, like people, tend to move only when forced to.</u>

If you add to the above the simple facts that most economic data is gathered for other purposes and is only incidentally available for the economic study made, and that there are often strong reasons for falsifying the initial data that is reported, then you see why economic data is bad.

As another source for inaccuracy mentioned by Morgenstern, consider that discounts to favored customers is a common practice, and these are jealously guarded secrets. Now it happens that in times of depression the company will grant larger discounts, and decrease them when things are improving, but the government figures of costs must be based on the listed sales prices since the discounts are unknowable. Thus economic down times and up times are systematically biased in different directions in the data gathered.

What can the Government Economists use for their basic data other than much of this inaccurate, systematically biased data? Yes, they may to a lesser or greater extent be aware of the biases, but they have no way of knowing how much the data is in error. So it should not surprise you that many economic predictions are seriously wrong. There is little else they can do, hence you should not put too much faith in their predictions.

In my experience most Economists are simply unwilling to discuss the basic inaccuracy in the economic data that they use, and hence I have little faith in them as Scientists. But who said that Economic Science is a Science? Only the Economists!

If Scientific and Engineering data are not at all as accurate as they are said to be, by factors of 5 or more at times, and economic data can be worse, how do you suppose that Social Science data fares? I have no comparable study of the whole field, but my little, limited experience does suggests that it is not very good. Again, there may be nothing better available, but that does not mean that what data is available is safe to use.

It should be clear that I have given a good deal of attention to this matter of the accuracy of data during most of my career. Due to the attitudes of the experts I do not expect anything more than a slow improvement in the long future.

If the data is usually bad, and you find that you have to gather some data, what can you do to do a better job? First, recognize what I have repeatedly said to you, the human animal was not designed to be reliable; it cannot count accurately, it can do little or nothing repetitive with great accuracy. As an example, consider the game of bowling. All the bowler needs to do is throw the ball down the lane reliably every time. How seldom does the greatest expert roll a perfect game! Drill teams, precision flying, and such things are admired as they require the utmost in careful training and execution, and when examined closely leave a lot to be improved.

Second, you cannot gather a really large amount of data accurately. It is a known fact that is constantly ignored. It is always a matter of limited resources and limited time. The management will usually want a 100% survey when a small one, consisting a good deal less, say 1% or even 1/10%, will yield more accurate results! It is known, I say, but ignored. The telephone companies, in order to distribute the income to the various companies involved in a single long distance phone call, used to take a very small, carefully selected sample, and on the basis of this sample they distributed the money among the partners. The same is now done by the airlines. It took them a long while before they listened, but they finally came to realize the truth of: <u>Small samples carefully taken are better than large samples poorly done.</u> Better, both in lower cost and in greater accuracy.

Third, much social data is obtained via questionaires. But it a well documented fact that the way the questions are phrased, the way they are ordered in sequence, the people who ask them or come along and wait for them to be filled out, all have serious effects on the answers. Of course, in a simple black and white situation this does not apply, but when you make a survey then generally the situation is murky or else you would not have to make it. I regret that I did not keep a survey by the American Mathematical Society that it once made of its members. I was so

indignant at the questions, which were framed to get exactly the answers they wanted, that I sent it back with that accusation. How few mathematicians faced with questions, carefully led up to in each case, such as: is there enough financial support for mathematics, enough for publications, enough for graduate scholarships, etc., would say that there was more than enough money available? The Math Society of course used the results to claim that there was a need for more support for Mathematics in all directions.

I recently filled out a long, important questionaire, (important in the consequence management actions that might follow). I filled it out as honestly as I could, but realized that I was not a typical respondent. Further thought suggested that the class of people being surveyed was not homogeneous at all, but rather was a collection of quite different subclasses, and hence any computed averages will apply to no group. It is much like the famous fact that the average American family has 2 and a fraction children, but of course no family has a fractional child! Averages are meaningful for homogeneous groups (homogeneous with respect to the actions that may later be taken) but for diverse groups averages are often meaningless. As earlier remarked, the average adult has one breast and one testicle, but that does not represent the average person in our society.

If the range of responses is highly skewed we have recently admitted publicly that the median is often preferable to the average (mean) as an indicator. Thus they often now publish the median income and median price of houses, and not the average amounts.

Fourth, there is another aspect that I urge you to pay attention to. I have said repeatedly that the presence of a high ranking officer of an organization will change what is happening in the organization at that place and at that time, so while you are still low enough to have a chance please observe for yourself how questionaires are filled in. I had a clear demonstration of this effect when I was on the Board of Directors of a computer company. I saw that underlings did what they thought would please me, but in fact angered me a good deal, though I could say nothing to them about it. Those under you will often do what they think you want, and often it is not at all what you want! I suggest that, among other things, you will find that when headquarters, in your organization, sends out a questionaire, then those who think that they will rate high will more often than not promptly fill them out, and those who do not feel so will tend to delay, until there is a dead line and then some low level person will fill them out from hunches without making the measurements that were to be taken - it is too late to do it right, so send in what you can! What these "made up" reports do the reliability of the whole is anyone's guess. It may make the results too high, too low, or even not change the results much. But it is from such surveys that the top management must make their decisions - and if the data is bad it is likely that the decisions will be bad.

A favorite pastime of mine, when I read or hear about some data, is to ask myself how people could have gathered it - how their conclusions could be justified? For example, years ago when I was remarking on this point at a dinner party, a lovely widow said that she could not see why data could not be gathered on any topic. After some moments of thought I replied, "How would you measure the amount of adultery per year on the Monterey Peninsula?" Well, how would you? Would you trust a questionaire? Would you try to follow people? It seems difficult, and perhaps impossible, to make any reasonably accurate estimate of the amount of adultery per year. There are many other things like this that seem to be very hard to measure, and this is especially true in social relationships.

There is a clever proposed method whose effectiveness I do not know in practice. Suppose you want to measure the amount of murder that escapes detection. You interview people and tell them to toss a coin without anyone but themselves seeing the outcome, and then if it is heads they should claim that they have committed a murder, while if tails they should tell the truth. In the arrangement there is no way anyone except themselves can know the outcome of the toss, hence no way that they can be accused of murder if they say so. From a large sample the slight excess of murders above one half gives the measure you want. But that supposes that the people asked, and given that protection, will in fact respond accurately. Variations on this method have been discussed widely, but a serious study to find the effectiveness is still missing, so far as I know.

In closing, you may have heard of the famous election where the newspapers announced the victory for President to one man when in fact the other won by a land slide. There is also the famous Literary Digest poll which was conducted via the telephone, and was amazingly wrong afterwards - so far wrong that the Literary Digest folded soon after - some people say because of this faulty poll. It has been claimed that at that time the ownership of a telephone was correlated with wealth and wealth with a political party - hence the error.

Surveys are not a job for an amateur to design, administer and evaluate. You need expert advice on questionaires (not just a run-of-the-mill statistician) when you get involved with a questionaire, but there seems little hope that questionaires can be avoided. More and more we want not mere facts about hard material things, but we want social and other attitudes surveyed - and this is indeed very treacherous ground.

In summary, as you rise in your organization you will need more and more of this kind of information than was needed in the past since we are becoming more socially oriented and subject to law suits for trivial things. You will be forced, again and again, to make surveys of personal attitudes of people, and it is for these reasons I have spent so much time on the topic of unreliable data. You need reliable data to make reliable decisions, but you will seldom have it with any reliability!

MEASUREMENT ACCURACIES

(Parts per million)

		BIRGE 1929		CODATA 1973	FACTOR OF	R
		ESTIMATED ERROR	ACTUAL ERROR	ESTIMATED ERROR	IMPROVEMENT	
Vel. of light	c	13	20	0.004*	5000	1.538
Bohr R	α	800	2000	0.82	2500	2.500
charge	e	1000	6000	2.9	2000	6.000
Planck	h	1200	11,000	5.4	2000	9.167
Avogadro	N_A	1000	7000	1.1*	6400	7.000
mass	me	1500	13,000	5.1	2500	8.667
Rydberg	R_∞	0.6	1.2	0.009*	130	2.000
					Average 3000	= 5.267 Average

*Cohen and Taylor 1975

Figure 7. Measurement accuracies.

Expected "Actual error" is more than 5 times "Estimated error"

Figure 27-1

Slightly different from physical constants available

R
.451
2.552
2.815
3.098
2.980
1.684
2.786
3.000
Av = 2.368

Figure 27-2