No. 230
July 1980

CONFRONTATIONS WITH DATA:
SOME EXAMPLES FOR THE ENGINEER

John B. Woodward, III
Professor

For Presentation to
The Society of Naval Architects
and Marine Engineers
Great Lakes and Great Rivers Section
January 22, 1981

Department of Naval Architecture
and Marine Engineering
College of Engineering
The University of Michigan
Ann Arbor, Michigan 48109

ABSTRACT

The paper consists of discussions of several sets of contrived experimental data. In each example there is a "what to do with it" choice, for example, whether there is a relationship with an independent variable to be found by regression, or whether simple statistics (such as mean, variance) should be found. Common equations of statistics and regression analysis are used throughout, but with the purpose of illustrating choice of use rather than of introducing the details of computation.

INTRODUCTION

Data is (are?) a collection of numbers that can be reduced to a much smaller collection -- perhaps just a single number -- that is comprehensible to a potential user. An example: a listing of the age of every individual in the USA is clearly an immense set of data, incomprehensible as it stands; reduced to a median (or to a mean or other parameter) it is comprehensible, and immediately becomes the grist for every grinding sociologist, politician, insurance actuary, etc.

Data reduction may be quite simple. The mean or median, for example, can surely be found readily by an analyst who knows little more than elementary arithmetic and the definitions of these words. Complication abound, however. The data is likely to be scattered, i.e. all points neither lie precisely along some recognizable curve, nor do they fall precisely into some well-recognized distribution about their mean, say. Often the data covers only a small sample of a large population, leaving doubt over how well the reduced data (the "statistics") represents the true value(s) (the "parameters"). Resolving these complications calls forth the many techniques of the probability-statistics field, and a reader who has training therein doubtless is aware of these techniques.

The probability-statistics field is so important in so many disciplines that the technical literature is swamped with textbooks, and some of them are quite suitable for the self-education of anyone having a reasonable acquaintance with mathematics. It therefore would be redundant for this paper to explain formulas and techniques. I use a lot of formulas here, but my purpose is not to introduce them; they are standard stuff, and the references can tell you about them. What the extensive literature and most college courses in probability and statistics don't cover is the choice of technique to use when confronted with a jumble of numbers -- material handed you in the expectation that you will distill that comprehensible essence from it. Judgement is required, for firm rules are lacking. To reinforce your judgement capability a bit in this area, I'm going to demonstrate via examples what might be done with several data sets. The purpose here is therefore to illustrate choice of technique, with only passing attention being given to details of those techniques.

For an introductory example, look at Tables 1 and 2. They contain data distributed in the 1970s by the Alken-Murray Corporation, showing the changes in fuel consumption occurring when its "Alken Even-Flo" fuel treatment is used. Its purpose is doubtless to convince potential customers of the product's merit. An organized jumble of numbers, all different -- what do they mean to you? Perhaps you operate a ship, or ships, and must decide from this data what benefit you would receive. The key

ALKEN-MURRAY CORPORATION
ESTABLISHED 1894
EXECUTIVE OFFICES 111 FIFTH AVENUE • NEW YORK, N.Y. 10003 • TEL. (212) 777 6580

SUMMARY OF ENGINE LOG ABSTRACTS OF STEAM TANKERS USING ALKEN EVEN-FLO FUEL TREATMENTS
COMPARING BEFORE TREATMENT WITH TREATMENT PERIODS (Assuming 300 Days Per Year At Sea)

(Seabuoy to Seabuoy)

Tankers		Total Engine Miles	Total Fuel Consumed Bbl or Ton	Fuel per Eng. Mile	Fuel % Inc. or Dec.	Avg. Eng. Speed	** Speed % Inc. or Dec.	Annual Value Inc. Speed	Annual Value Fuel Saving	Annual Total Amount Saved	Annual Cost of ALKEN Treatment
A 26554 DWT	No Treat.	23878	38902 B	1.63 B		18.94		(13.1 days)			
18600 shp.	Alken	59013	93689 B	1.588 B	-2.58%	19.77	+4.38%	$65500	$11976	$77476	$5016
	No Treat.	15323	24622 B	1.607 B	+1.20%	19.21	-2.83%	$44000*	$ 5409*	$49409*	
B 29000 DWT	No Treat.	52183	9276 T	.1777 T		15.24		(2.8 days)			
12000 shp.	Alken	28312	4830 T	.1706 T	-4.0%	15.38	+0.92%	$7000	$10953	$17953	$4090
C 20198 DWT	No Treat.	41962	39588 B	.940 B		14.48		(8.9 days)			
6000 shp.	Alken	155988	141654 B	.908 B	-3.39%	14.91	+2.97%	$35600	$ 7697	$43297	$3936
D 19165 DWT	No Treat.	49211	37456 B	.7611 B		13.331		(9 days)			
5000 shp.	Alken	60042	44110 B	.7346 B	-3.48%	13.746	+3.11%	$24750	$ 5506	$30256	$2510
E 19165 DWT	No Treat.	41272	33174 B	.804 B		12.65		(8.3 days)			
5000 shp.	Alken	35641	30048 B	.843 B	+4.85%	13.19	+4.27%	$33200	-$ 6622	$26578	$2275
F 18000 DWT	No Treat.	77153	67999 B	.8814 B		15.87		(3.4 days)			
6600 shp.	Alken	98062	84332 B	.8600 B	-2.43%	16.05	+1.13%	$11900	$ 5569	$17469	$4142
(T-2)	Alken	31913	26425 B	.8280 B	-6.06%	15.98	+0.69%	$ 7350 (2.1 days)	$13826	$21176	$4124
G 34000 DWT	No Treat.	39084	8751 T	.2239 T		13.98		(14.6 days)			
14000 shp.	Alken	35697	7913 T	.2217 T	-0.98%	14.66	+4.86%	$43710	$ 2970	$46680	$4655
	No Treat.	36734	8561 T	.2331 T	+5.14%	14.04	-4.23%	$40200*	$15500*	$55700*	
H 10535 DWT C-2	No Treat.	78095	45807	.5866		16.74		(13.8 days)			
6000 SHP	Alken	56261	34103	.6062	+3.34%	17.51	+4.60%	$41400	-$7108	$37292	$3060
	Net Improved Boiler Efficiency-7.86% (Allowing 2.0% for Deeper Draft)										
I 16900 DWT T-2	No Treat.	29274	20466	.6991		15.77		(1.7 days)			
6600 SHP	Alken	15871	10621	.6692	-4.28%	15.86	+0.57%	$5100	$6542	$11642	$3057
	Net Improved Boiler Efficiency-5.42%										
J 22672 DWT Bulk	No Treat.	14545	16105	1.1073		15.85		(5.3 days)			
9350 SHP	Alken	28956	31473	1.0869	-1.84%	16.13	+1.77%	$15900	$5091	$20991	$4918
	Net Improved Boiler Efficiency-5.38%										
K 47164 DWT Tkr.	No Treat.	132042	236695	1.7926		18.06		(6.3 days)			
21500 SHP	Alken	63045	113731	1.8040	+0.64%	18.44	+2.10%	$100800	$3351	$97449	$10881
	Net Improved Boiler Efficiency-5.56% (Allowing 2.0% for Deeper Draft by 2'1½")										
	% Over Performance Curve**										
L 11330 DWT C-2	No Treat.	28257	16840	.5960	10.24%	16.11		(7 days)			
6000 SHP	Alken	32212	17722	.5502	4.66%	16.49	+2.36%	$21000	-	$21000	$2375
	Net +5.58%**										
M 66000 DWT Tkr.	No Treat.	122317	30351	.2481		17.16		(4.5 days)			
19000 SHP	Alken	103234	24787	.2401	-3.24%	17.42	+1.51%	$33525	$16380	$49905	$7020
	Net Improved Boiler Efficiency-6.26%										

*Double the % of increase or decrease in speed and add (or subtract) that figure to the % of increase or decrease in fuel consumption to get the approximate % of fuel saved if speed had remained the same.

KENNETH F. YARRINGTON
Sales Manager - Marine Division

ALKEN-MURRAY CORPORATION
ESTABLISHED 1934
EXECUTIVE OFFICES 111 FIFTH AVENUE • NEW YORK, N.Y. 10003 • TEL. (212) 777-6660

SUMMARIES OF ENGINE LOG ABSTRACTS OF MOTOR VESSELS USING ALKEN EVEN-FLO FUEL TREATMENT
COMPARING BEFORE TREATMENT WITH TREATMENT PERIODS (Assuming 300 Days Per Year At Sea)
(Seabuoy to Seabuoy)

		Total Engine Miles	Total Tons Fuel Consumed	Fuel Cons. per Mile	Fuel % Inc. or Dec.	Avg. Eng. Speed	** Speed % Inc or Dec.	Annual Value Inc. Speed	Annual Value Fuel Saving	Annual Total Amount Saved	Annual Cost of ALKEN Treatment
A 24000 DWT Bulk Car.	No Treat.	64918	5821	.0896		15.89		(9 days)			
10500 BHP Sulzer	Alken	36674	3316	.0940	+0.89%	16.37	+3.02%	$27000	-$1488	$25572	$2365
B 16585 DWT Bulk Car.	No Treat.	28353	12733	.0449		11.67					
5680 BHP M.A.N.	Alken	12586	5424	.0431	-4.01%	11.66	Same		$2416	$ 2416	$ 856
C 16585 DWT Bulk Car.	No Treat.	55878	2538.5	.0454		11.53		(8.44 days)			
5680 BHP M.A.N.	Alken	17821	804.9	.0452	-0.44%	11.85	+2.83%	$25320	$ 283	$25603	$ 825
D 23690 DWT Bulk Car.	No Treat.										
9100 BHP Sulzer	to 6/11/65	58324	5122.3	.0878		15.01		(2.6 days)			
	Alken	78421	6753.3	.0861	-1.93%	15.14	+0.86%	$ 5200	$2592	$ 7792	$2040
	No Treat.	31659	2846.6	.0899	+4.41%	14.42	-4.76%	-$29600	-$5522	-$35122	
								(14.8 days)			
E 27625 DWT Bulk Car.	No Treat.	47335	4044.3	.0854		13.84		(7.4 days)			
9100 BHP Sulzer	Alken	15903	1333.0	.0838	-1.87%	14.19	+2.53%	$18500	$2285	$20785	$1860
F 12700 DWT Dry Cargo	No Treat.	96707	8248.8	.0853		18.30		(3.5 days)			
10500 BHP M.A.N.	Alken	79503	6619.0	.0833	-2.34%	18.51	+1.15%	$10500	$3890	$14390	$2890
G 25500 DWT Bulk Car.	No Treat.	38910	3905.0	.1004		15.83					
10500 BHP Sulzer	Alken	33931	3224.4	.0950	-5.38%	15.87	+0.25%	None	$15370	$15370	$2796
		(Obs.)		(Obs.)							
H 3640 DWT Dry Cargo	No. Treat.	52128	1919 T.	.0368		13.59		(15.9days)			
3200 BHP Fiat	Alken	19260	684 T.	.0355	-3.53	14.31	+5.30	$15900	$ 2903	$18803	$ 912
200 Sec. Redwood		10830	349 T.	.0322	-12.50	13.74	+1.10				
I 3640 DWT Dry Cargo	No Treat.	39277	1517.6 T.	.0386		13.09		(4.1 days)			
3200 DHP Fiat	Alken	41586	1560.4 T.	.0375	-2.85	13.27	+1.38	$ 4100	$ 2380	$ 6480	$ 906
200 Sec. Redwood											
		(Eng.)		(Eng.)							
J 30000 DWT Tanker	No Treat.	38538	36307 B.	.9421		14.12		(12.1days)			
20160 BHP Pielstik	Alken	12835	12599 B.	.9816	+4.19	14.69	+4.04	$24200	-$13050	$11150	$4158
Light Diesel											
K 20115 DWT Tanker	No Treat.	27242	1560 T.	.0650		14.64		(- 2 days)			
8250 BHP B&W	Alken	25774	1513 T.	.0608	-6.46	14.54	-0.68	-$ 3000	$ 6017	$ 3017	$1656
1500 Sec. Redwood											
L 20861 DWT Tanker	No Treat.	43409	3777 T.	.0870		14.46		(3.5 days)			
9230 BHP B&W	Alken	37090	3191 T.	.0860	-1.15	14.63	+1.18	$ 7000	$ 1782	$ 8882	$2227
1500 Sec. Redwood											
M 6075 DWT Reefer	No Treat.	42303	2630 T.	.0622		17.96					
8400 BHP Sulzer	Alken	89273	5358 T.	.0600	-3.54	17.96	none	-	$ 4625	$ 4625	$2044
1500 Sec. Redwood											
N 44500 DWT Bulker	No Treat.	102695	11023	.1073		16.69		(8.1 days)			
13800 BHP B&W	Alken	38842	4099	.1055	-1.68	17.14	+2.70	$29970	$ 4320	$34290	$3138
1000 Sec. Redwood											

*Double the % of increase or decrease in speed and add (or subtract) that figure to the % of increase or decrease in fuel consumption to get the approximate % of fuel saved if speed had remained the same.

KENNETH F. YARRINGTON
Sales Manager - Marine Division

information is the change in specific fuel consumption, but what will it be for a ship that isn't part of the sample, yet belongs to the population sampled? I would advise you first to read the tables carefully in order to understand just what is being presented, then to check for obvious errors (there are a few). Then what? The fundamental fact to be uncovered is whether the data represents a single number (fractional improvement in specific fuel consumption) with unidentified chance factors giving the data its scatter, or whether there is an underlying relationship that is the key to your answer. For example, the fractional improvement may be a function of propulsive power, or of viscosity of the fuel, of some other factor, or of a combination of factors. If there is a relationship, you obviously want to discover what the relationship is. In either case, you will want to decide how much margin to allow for uncertainties, based on how confident you must be in the final answer.

Now, it is not appropriate here to reach any conclusions regarding the merits of the Alken product. Tables 1 and 2 serve only as examples of the kind of data that a marine engineer may encounter. Examples that appear henceforth are based on artifical data contrived to illustrate the several points to be made. None of it derives in any way from those first two tables.

All formulas to be encountered following come from references 1, 2, and 3. Reference 4 is included as an additional source, particularly in non-linear regression.

EXAMPLE 1

Look at Figure 1, a plot of 100 data points, represented simply as X and Y. Since these are pairs of data points, the plotter must believe that there is a relationship between X and Y. The figure does suggest that such a relationship exists (i.e. that there is a "correlation" here); the points, though scattered, seem to lie along a straight line. The correlation is a good one, since it is clear where the line lies — to close approximation — without any formal curve fitting. You could doubtless draw the line by eye.

However, best results come from a formal process of regression. Linear regression is a clear choice here, hence the process is expected to produce an equation of the form

$$Y = a + b\,X \qquad (1)$$

"Best" is defined as values of Y that will minimize the sum of the square errors, i.e. if Y' represents the values of Y given by equation (1), then the sum of all 100 $(Y - Y')$ shall be a

minimum. (You may then recognize the process as "least squares curve fitting.") the resulting formula for the linear regression coefficient b is

$$b = \frac{\Sigma(x\ y) - (\Sigma x)(\Sigma y)/n}{\Sigma x^2 - (\Sigma x)^2/n} \qquad (2)$$

where the summation is taken over all n data points,

and $a = \bar{y} - b\ \bar{x}$ (3)

where the bar designates a mean value of the variable.

When the formulas are applied to the data pictured in Figure 1, the resulting linear regression equation is

$$Y = 0.329 + 0.146\ X \qquad (4)$$

The data is repeated in Figure 2, and equation (4) is plotted there.

Finding the best linear regression line was the obvious thing to do with this data, and the process was fairly easy (I used a computer routine). But finding the line is just the beginning. No more than two or three of the data points actually fall on the regression line, hence you should question just how good the prediction of the line is, is spite of its being "best." A significant fact is the ability of the regression formulas to produce a line that is best no matter how scattered the data is; best may be merely least bad. Although clearly (by inspection) not the case here, the line can be so bad as to be worthless. A common test of this possibility is to calculate the correlation coefficient. This parameter ranges in magnitude from 0 to 1.0, with 0 indicating absolutely no relationship between X and Y, and 1.0 indicating perfection. (A negative value simply indicates a negative slope of the regression line.)

One of the several formulas for correlation coefficient is

$$r = \frac{\Sigma(x\ y) - (\Sigma x)(\Sigma y)/n}{[\Sigma x^2-(\Sigma x)^2/n]\,[\Sigma y^2-(\Sigma y)^2/n]} \qquad (5)$$

Equation (5) processes Figure 1 data into a value of r = 0.98, indicating an excellent correlation by the line in Figure 2. A low value might indicate that the assumed linear relationship is a poor choice, or indicate such extreme scatter that any relationship is doubtful. Visual inspection of the plotted data would probably reveal the culprit.

Here the correlation is good, but scatter is nonetheless present, and should not be ignored. One doubt raised by scatter is that X is not the only independent variable; something

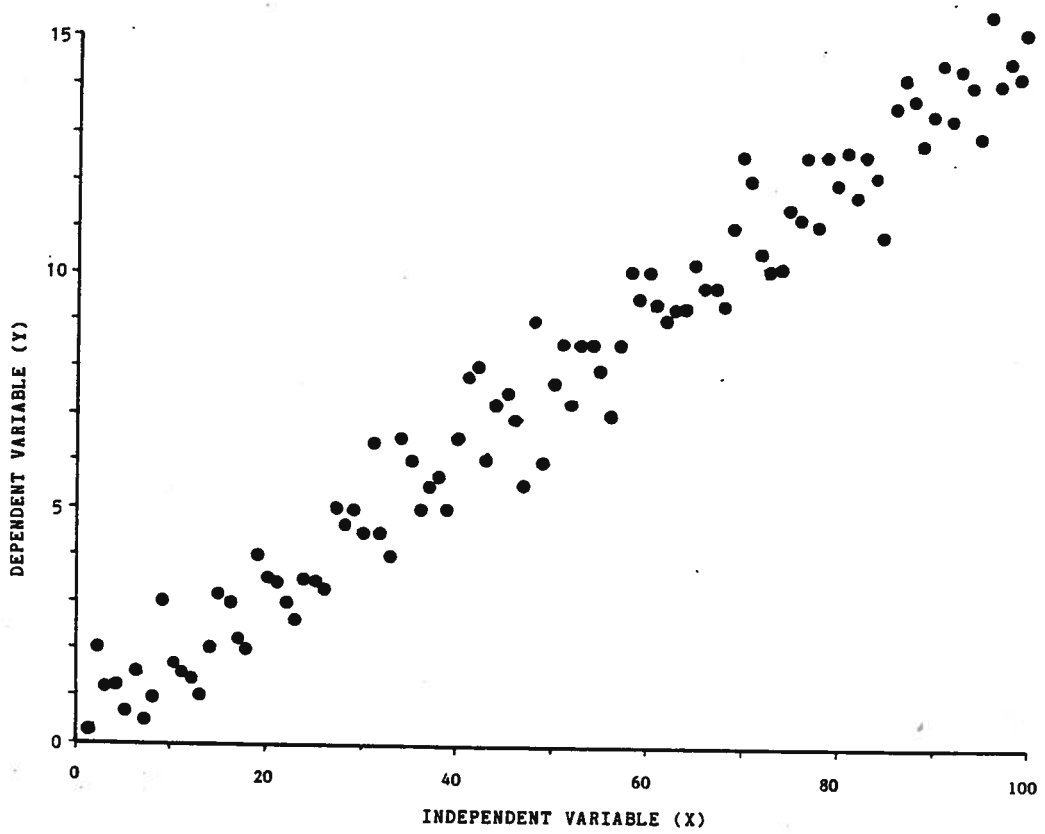

FIGURE 1

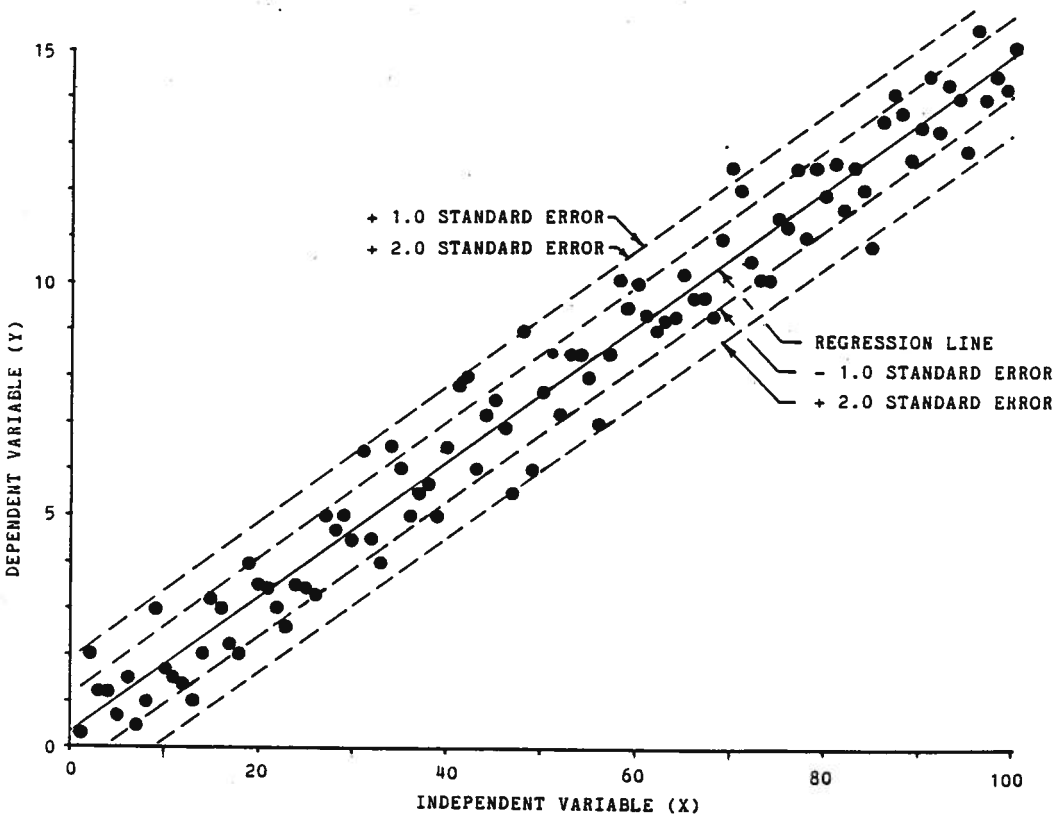

FIGURE 2

unknown may be producing the variation of points from the line. You should look for it, but of course it cannot be found here because no further information is given. The scatter may otherwise be due to errors in making the measurements, or to the operation of chance in the X-Y relationship. Whatever, it is irreducible, but should not be ignored in producing the answer you expect to extract from the data. Let's say you are asked to state the value of Y when X is 60 (perhaps the percent fuel saving when shp is 60,000). You would read Y = 9.09 from the line, or from equation (4), but how about the data point at X = 60 that has a value of 10? You will use the value read from the line, but must give some respect to that scatter.

A first step in that direction is to calculate the standard error. A formula is

$$S = [RSS/(n-2)]^{1/2} \tag{6}$$

where RSS = the residual sum of squares

$$= \Sigma x_D^2 - (\Sigma x_D \Sigma y_D)^2 / \Sigma x_D^2 \tag{7}$$

$$x_D = x - \bar{x}$$

$$y_D = y - \bar{y}$$

S is in effect the standard deviation of the data points about the regression line (and S^2 is the variance). Here S has the value 0.813.

Since the "least squares curve fitting" concept is based on an assumed normal distribution of the data about the regression line, you should reflect that when data is normally distributed, about 68 percent of it lies within 1.0 standard deviation of its mean, about 95 percent lies within 2.0 standard deviations of its mean, etc. In Figure 2 auxilliary lines are plotted at 1.0S and 2.0S on either side of the regression line. On counting the data points, you will find that approximately the predicted numbers lie within these distances of that line.

To complete the answer, you must depart from calculation into the realm of judgement based on the use of the answer. If, for instance, it were the input to an expected value decision process, then Y = 9.09, being an expected (mean) value, would be the appropriate answer. On the other hand, suppose that only a single event will follow from the a decision based on X = 60. If you want to be 95 percent confident that the value chosen for Y will turn out to be correct, then say that Y will lie in the range 7.46 to 10.7, i.e. plus or minus two standard errors. If the Boss doesn't accept that much waffling, then tell her (!) that Y will be 9.09 with a possible error of plus or minus 1.62 — really the same thing, but may sound more definite. If that

doesn't go over, then investigate a little further the end use of your answer. If it goes into estimating a cost, and Boss is a known conservative on cost estimates, then say that Y will be no greater than 10.7 (agains assuming tha 95 percent confidence), or no greater than 9.9 (gambling on 68 percent confidence). If Boss still doesn't understand why you are being so slippery, then include the estimates of a new employer in your calculations.

EXAMPLE 2

Look at Figure 3 and Table 3. These show contrived data for fractional fuel rate improvement produced by an imaginary fuel additive. In Figure 3 the data is plotted as a function of shp, and the regression line is drawn. Its equation, found as before, is

$$Y = 0.03147 + 0.000424\ X \tag{8}$$

where X is shp/1000.

The correlation coefficent is 0.846, which is not bad, though the scatter suggests that you should not be satisfied. In this case there is a possible second independent variable given, the viscosity of the fuel. Perhaps fuel rate change is more a function of this than of shp. If X is taken to be viscosity, then

$$Y = 0.0368 + 8.52\text{x}10\ X \tag{9}$$

The correlation coefficient is 0.428. This is certainly better than zero, but does suggest a rather poor correlation. You might then reasonably ask "is this really a correlation, or does this seeming correlation occur by chance?" Let's take a look at a possible answer to the question.

A procedure that may illuminate the situation is to establish a null hypothesis, i.e. declare that there is no correlation, and hence if all possible X-Y values were known, then the value of r would turn out to be zero; it must be chance operating among those 20 data points that gives r = 0.428. The null hypothesis must then be put to a test, and if it withstands attack, then you reasonably conclude it to be true and abandon correlation. The "t" test is appropriate to the testing of the hypothesis, whence

$$t = r\sqrt{-2}/\sqrt{1-r^2} \tag{10}$$

$$= 2.01 \text{ in this case}$$

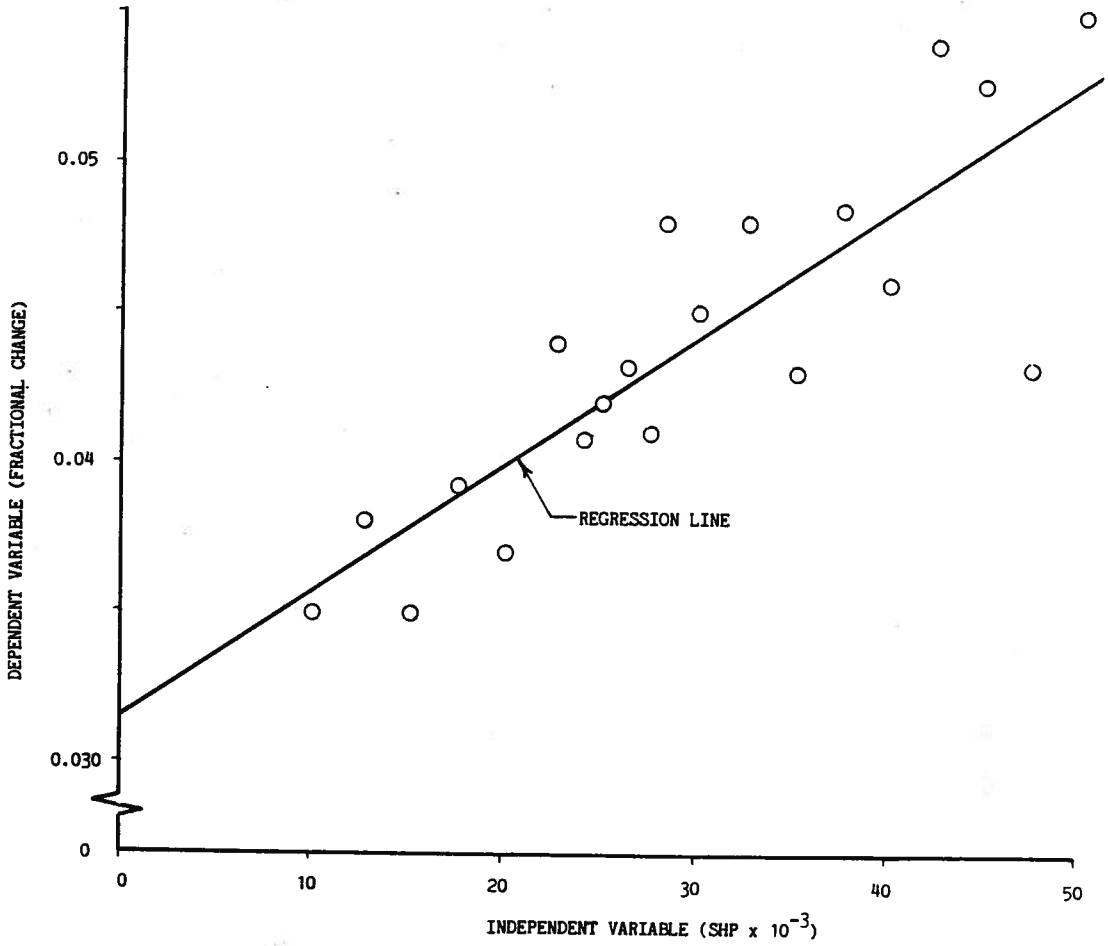

FIGURE 3

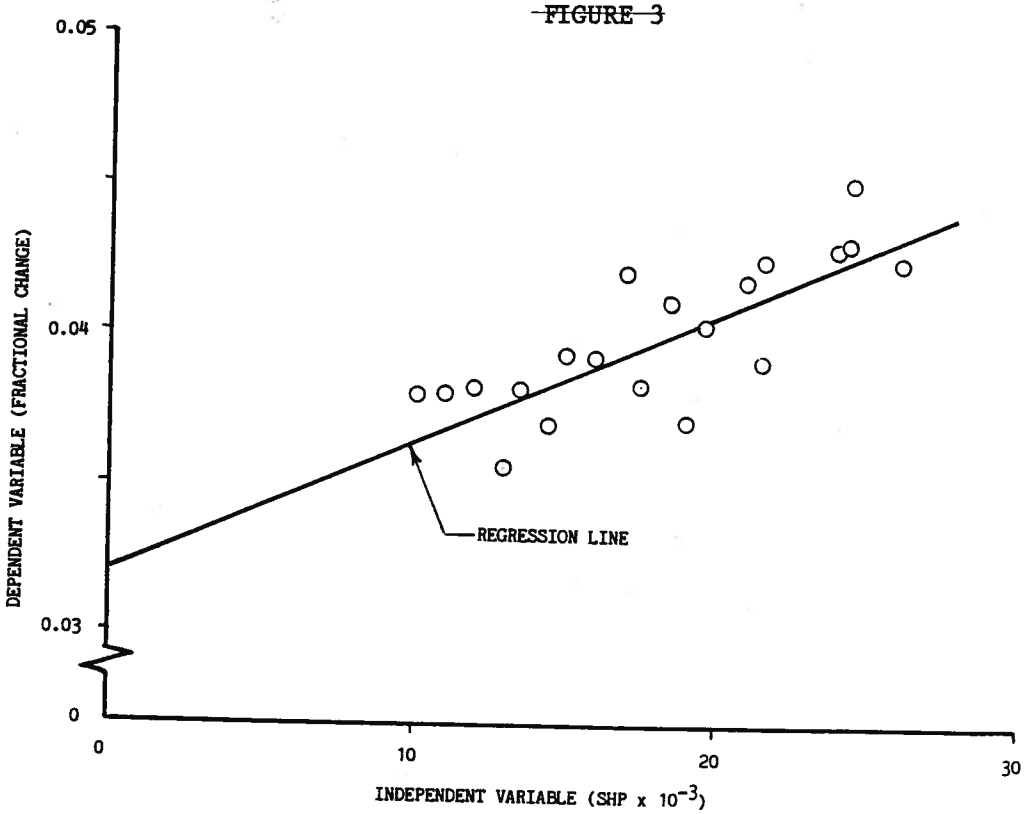

FIGURE 4

TABLE 3

Example Fuel Rate Improvement Data - Steam Propulsion

	POWER (SHP/1000)	VISCOSITY (Centistokes)	IMPROVEMENT (Fractional)
1.	10.0	900	0.0350
2.	37.5	950	0.0486
3.	26.0	900	0.0432
4.	27.5	980	0.0410
5.	12.5	900	0.0380
6.	25.0	880	0.0420
7.	42.5	1200	0.0540
8.	30.0	870	0.0450
9.	28.0	1180	0.0480
10.	15.0	360	0.0350
11.	35.0	400	0.0430
12.	24.0	880	0.0408
13.	17.5	920	0.0392
14.	20.0	350	0.0370
15.	47.5	400	0.0431
16.	50.0	900	0.0550
17.	22.5	1200	0.0440
18.	32.5	1240	0.0480
19.	40.0	400	0.0460
20.	45.0	900	0.0528

TABLE 4

Example Fuel Rate Improvement Data - Steam and Diesel Propulsion

	STEAM		DIESEL	
	POWER (SHP/1000)	IMPROVEMENT (Fractional)	POWER (SHP/1000)	IMPROVEMENT (Fractional)
1.	10.0	0.0350	16.0	0.0392
2.	37.5	0.0486	21.5	0.0390
3.	26.0	0.0432	17.5	0.0382
4.	27.5	0.0410	19.0	0.0370
5.	12.5	0.0380	21.6	0.0425
6.	25.0	0.0420	13.0	0.0355
7.	42.5	0.0540	13.5	0.0381
8.	30.0	0.0450	17.0	0.0420
9.	28.0	0.0480	21.0	0.0417
10.	15.0	0.0350	26.0	0.0434
11.	35.0	0.0430	15.0	0.0392
12.	24.0	0.0408	10.0	0.0380
13.	17.5	0.0392	11.0	0.0380
14.	20.0	0.0370	12.0	0.0381
15.	47.5	0.0431	14.5	0.0370
16.	50.0	0.0550	18.5	0.0410
17.	22.5	0.0440	24.5	0.0450
18.	32.5	0.0480	24.2	0.0429
19.	40.0	0.0460	19.5	0.0401
20.	45.0	0.0528	24.0	0.0428

This statistic is a measure of the probability that r = 0.428 could occur by chance when its true value is really 0.0. If t could then be as large as 0.428 by chance on only 5 percent of the occasions when such samples were taken (or 1 percent, or 0.1 percent, whatever your judgement is) you would probably say that the null hypothesis is destroyed, i.e. that the probability is just too small for you to believe it true. How to know? The arbitrary choice of 5 percent (or 1 percent, or 0.1 percent, etc) sets the level of significance at 5 percent (etc, etc); the sample size of 20 sets the degrees of freedom at n - 2 = 18. A glance at a table of the t distribution shows that for these two arguments, t = 2.10. Here then is a borderline case in which calculated t and tabulated t are nearly the same. The null hypothesis is upheld, though shakily. If calculated t were much larger, then the the probability of a chance occurrence of r = 0.428 would be very small, and you would be forced to admit that a correlation must exist to produce it.

But borderline? We've struggled to no decision (c'est le vie, eh?), but doubt is surely cast on the use of viscosity to predict values of the fuel rate improvement. If it were essential to know whether the correlation exists, then the remedy would be to go back for more data. However, one might pursue the alternative path of investigation the possibility that fuel rate saving is a function of both shp and viscosity. If so, then the scattering evident when either independent variable is taken separately may be caused by the omission of the other. We try

$$Y = a + b_1 X_1 + b_2 X_2 \tag{11}$$

Formulas are

$$b_1 = \frac{(\Sigma x_{D1} y_D)(\Sigma x_{D2}^2) - (\Sigma x_{D2} y_D)(\Sigma x_{D1} x_{D2})}{\Sigma x_{D1}^2 \Sigma x_{D2}^2 - (\Sigma x_{D1} x_{D2})^2} \tag{12}$$

$$b_2 = \frac{(\Sigma x_{D2} y_D)(\Sigma x_{D1}^2) - (\Sigma x_{D1} y_D)(\Sigma x_{D1} x_{D2})}{\Sigma x_{D1}^2 \Sigma x_{D2}^2 - (\Sigma x_{D1} x_{D2})^2} \tag{13}$$

$$a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \tag{14}$$

With appropriate numbers

$$Y = 0.0239 + 0.000429\ X_1 + 8.91 \times 10^{-6} X_2 \tag{15}$$

where X_1 is shp/1000, and X_2 is viscosity.

The joint correlation coefficient is

$$r = \frac{b_1 \Sigma x_{D1} y_D + b_2 \Sigma x_{D2} y_D}{\Sigma y_D^2} \tag{16}$$

where $x_{D1} = x_1 - \bar{x}_1$

$x_{D2} = x_2 - \bar{x}_2$

$y_D = y - \bar{y}$

With appropriate numbers, r = 0.960. Since this is higher than the coefficients of either Y vs X_1 (0.846) or Y vs X_2 (0.428), the correlation is much better, as is indeed confirmed by a few trials of equation (15) to see how it reproduces the data.

EXAMPLE 3

Look at Table 4, Figure 3, and Figure 4. Table 4 repeats the data of Table 3, omitting the viscosity column, and includes similar data for 20 diesel ships. The diesel ship data is plotted in Figure 4.

For the diesel data, the linear regression process produces

$$Y = 0.0322 + 0.000429\ X \qquad (17)$$

and a correlation coefficient of r = 0.802. These are much like the steam result (e.g. equation (8)), but do show obvious differences. Nonetheless, the similarity is strong enough that one should ask if the two data sets do not come from the same population, i.e. if "D" and "S" are not the values of a false independent variable that should be discarded. If so, then we have a combined data set of 40 pairs instead of 20. Since large samples typically furnish more accurate statistics, such a conclusion will be beneficial.

One should pose two null hypotheses to test the conclusion, namely that the two regression coefficients are really the same, and the the two correlation coefficients are really the same also (and hence differ only because of chance). If the hypotheses can be upheld, then the data can be combined.

Take the regression coefficient first. The t test is to be used again. In this instance

$$t = \frac{b_1 - b_2}{S_{b1-b2}} \qquad (18)$$

$$S_{b1-b2} = S^2_{y,x}\left[\frac{1}{\Sigma x^2_{D1}} + \frac{1}{\Sigma x^2_{D2}}\right]^{1/2} + \qquad (19)$$

$$S^2_{y,x} = \frac{RSS_1 + RSS_2}{n_1 + n_2 - 4} \qquad (20)$$

where RSS is the residual sum of squares as given by equation (7).

The resulting t is approximately 1×10^{-4}. With 36 degrees of freedom, and a level of significance of 0.05, one finds in a table that $t_{.} = 2.03$. Since the calculated value is so much less than the tabulated one, the difference in b can occur by chance with a much higher frequency than the one-in-twenty indicated by the chosen level of significance.

Now the correlation coefficient. Here a common test procedure is provided by way of the Z parameter, where

$$Z_r = 0.5[\ln(1 + r) - \ln(1 - r)] \tag{21}$$

(calculated for each of the 2 data sets)

The difference in the two Z_rs, divided by a standard deviation, produces Z, which is to be compared with a tabulated value.

$$Z = \frac{Z_{r1} - Z_{r2}}{\sigma_{Z1-Z2}} \tag{22}$$

$$\sigma_{Z1-Z2} = [\sigma^2_{Zr1} + \sigma^2_{Zr2}]^{1/2} \tag{23}$$

$$\sigma_{Zr} = \frac{1}{\sqrt{n-3}} \tag{24}$$

With the values of r that have been stated, z = 0.403. Using that same level of significance (0.05) again, one finds a tabulated Z = 1.96. Again, the null hypothesis is upheld.

Since both tests indicate the same population, you are safe in saying that steam and diesel data should be combined. The resulting regression line is

$$Y = 0.0324 + 4.02\ X \tag{25}$$

and the combined correlation coefficient is 0.866.

Now, a reader may snarl "Combustions in a boiler and in a diesel cylinder are different phenomena. You gotta expect different results, and damnit, they are different." Okay, but this assertion is not proven by the data given here. You are unlikely to prove it (or disprove it) with statistics, but if "judgment" says that there must be a difference that the data doesn't show, the way to get support from statistics is to take more data. You can see this by checking the role that n has played in producing the numbers given above.

EXAMPLE 4

The previous discussion deals exclusively with linear regresssion. Data may represent a non-linear phenomenon, of course, and regression may be applied to such data. For example, a polynomial relationship may be assumed, least-squares curve fitting applied to it, and from the process will come the best polynomial regression line.

Trouble is, the regression techniques do not themselves choose the form of the curve. The linear technique will produce that linear regression line from any collection of points. For example, you might carefully plot points on the circle $(X - 5)^2 + (Y - 5)^2 = 25$. Equations (2) and (3) cheerfully produce the line $Y = 5$ as the linear fit. However, equation (5) produces a correlation coefficient of $r = 0.0$, indicating most forcefully that that line is a poor fit. But it doesn't say why, i.e. whether the data points are a hopeless jumble of numbers, or whether they lie precisely on an obvious curve. That's for you to judge by examination of the data, and in this extreme case you would surely say "It's a BLEEPin' circle!" And proceed from there.

The example we're pushing toward is not this obvious. Look at Figure 5. Here are 20 data points, and from them comes the linear regression line

$$Y = -1.847 + 1.358\ X \tag{26}$$

with the correlation coefficient $r = 0.970$. Not bad as correlations go, and the figure does confirm a good fit. Nonetheless, the points show a distinct "curviness;" there's the suggestion of a parabolic variation. Perhaps the regression line should have the form

$$Y = a\ X^b \tag{27}$$

If so, linear techniques can be still be used, since the data can first be plotted on log-log paper; note that equation (27) becomes linear when the logarithm of both sides is taken. The equivalent procedure is to plot log Y vs log X (or ln Y vs ln X) on linear paper. The latter is done in Figure 6, and by equations (2) and (3)

$$\ln Y = -1.57 + 1.977 \ln X \tag{28}$$

Equation (5) produces a correlation coefficient $r = 0.991$, indicating a better fit to a straight line of the log (or ln) data. Inspection of Figure 6 confirms the conclusion.

Keep in mind that it was an arbitrary choice by the analyst to try equation (27) for a better fit. She could have been

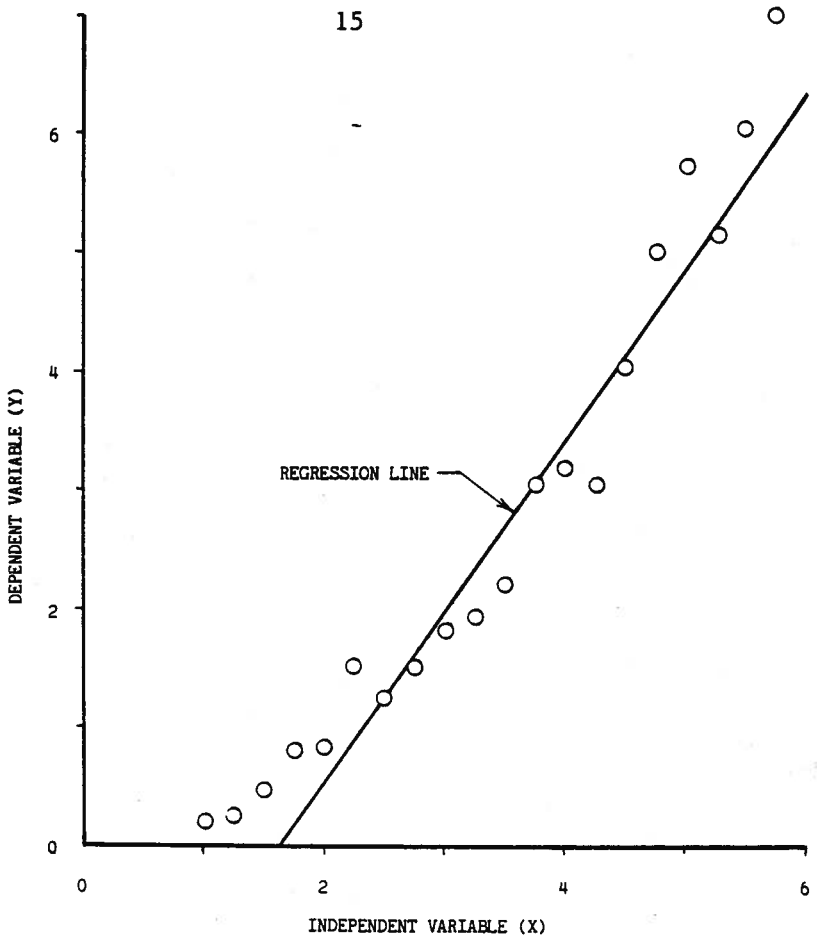

FIGURE 5

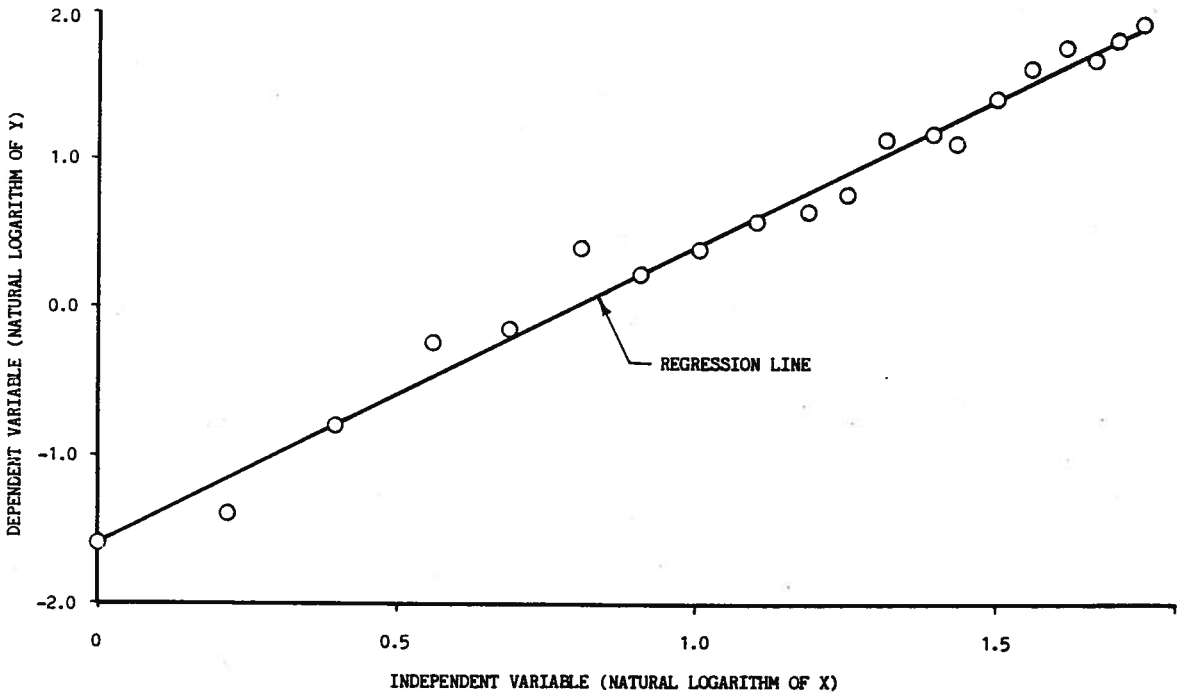

FIGURE 6

satisfied with equation (26), since nothing in the technique points to non-linearity, and the correlation coefficient is already pleasingly high. (However, there are computer programs that will try a number of different curves to fit, and select the best on the basis of highest r.)

EXAMPLE 5

Look at Figure 7 and Table 5. Here is some rather widely scattered data, but as always it is possible to find the linear regression equation, and its plot appears in the figure. Correlation coefficient is r = 0.357. This value, plus examination of the figure, casts strong doubt on the merit of the regression line. You can get further support for doubt by testing the null hypothesis that correlation coefficient is really 0.0 with the 0.357 value occurring by chance. The t test is appropriate, as it was in example 2, and equation (10) is used again to produce t. Its value here (18 degrees of freedom) is 1.624. If you once more elect the 0.05 significance level, you find the tabulated t = 2.101. The calculated t could therefore occur by chance with a higher frequency than one-in-twenty. The null hypothesis is therefore not destroyed, adding further evidence that linear regression is just not the answer here.

Some alternative should be explored, but the figure suggests no reasonable curve to try a fit on (I, at least, sure can't see anything among those points). Possibly there is no X-Y relationship, or possibly it's so messed up by undetected independent variables that a X-Y line just can't be constructed.

Let's abandon regression and see if the data (the Y values, that is) does not fall into a distribution that will reveal something of what is going on. Look at Figure 8, a distribution of the frequency of occurrence of Y. The heights of the bars represent the number of Ys that fall between 80 and 85, 85 and 90, etc. Obviously there is some arbitrariness here; the figure depends somewhat on how the boundaries are treated (where did I count Y = 85, for instance?), and inevitably is somewhat lumpy because of the small number of data points. Nonethelss, it suggests a normal (or Gaussian) distribution, a distribution whose density function is also plotted in Figure 8. The fit is far from perfect, but a quantitative measure of "goodness of fit" can form the basis for a rational decision.

Judging the fit of data to a suspected distribution, or indeed to a curve of any kind, is a common problem (we met it in example 4), hence there are several tools in widespread use. Here I use the Kolmogorov-Smirnov test.

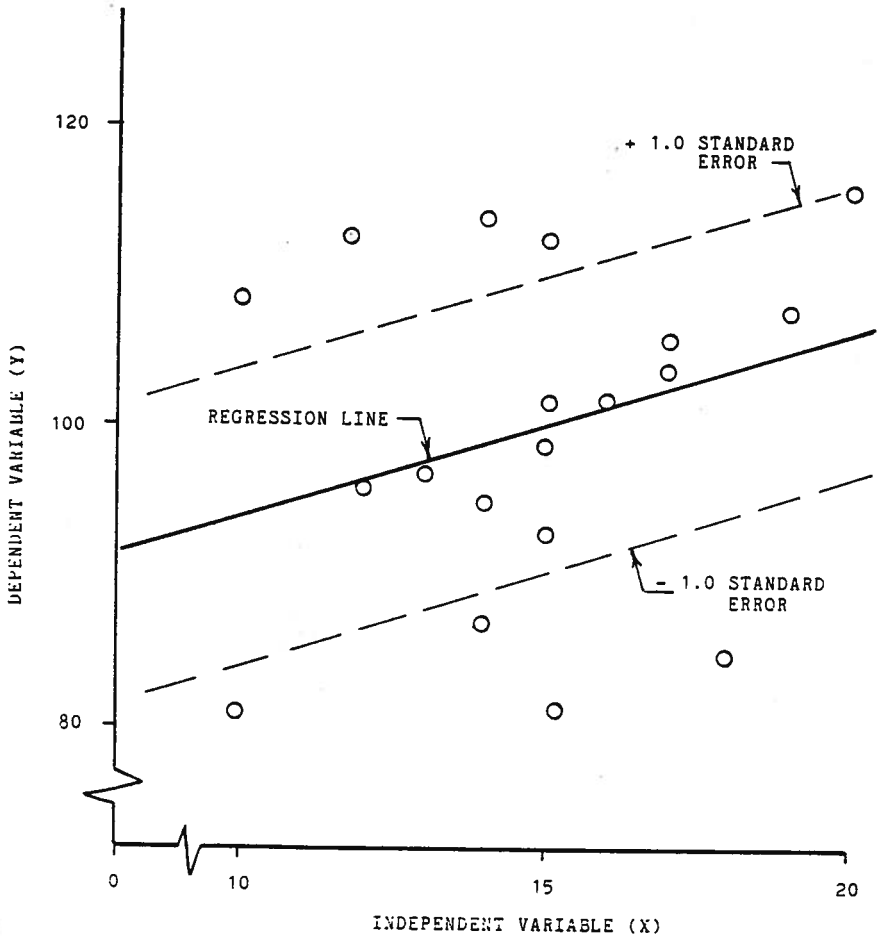

FIGURE 7

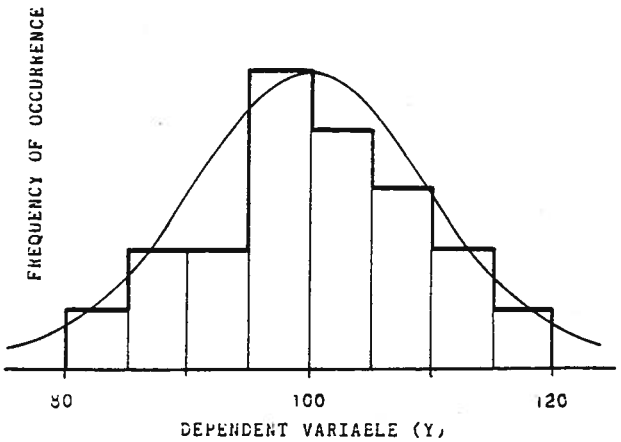

FIGURE 8

TABLE 5

Example Data (X-Y Pairs)

	X	Y
1.	10.0	81.0
2.	14.0	87.0
3.	13.0	94.0
4.	12.0	96.0
5.	13.0	97.0
6.	10.0	100.0
7.	16.0	102.0
8.	17.0	106.0
9.	10.0	109.0
10.	14.0	114.0
11.	18.0	85.0
12.	15.0	93.0
13.	14.0	95.0
14.	14.0	97.0
15.	15.0	99.0
16.	15.0	102.0
17.	17.0	104.0
18.	19.0	108.0
19.	15.0	113.0
20.	20.0	118.0

TABLE 6

Goodness of Fit Test for Figure 8

	m	θ	Cumulative	Theoretical	Cumulative	Difference
$\bar{Y}$-3s	1	0.05	0.05	0.0228	0.0228	0.0272
$\bar{Y}$-2s	2	0.10	0.15	0.1360	0.1587	0.0087
$\bar{Y}$-s	8	0.40	0.55	0.3413	0.5000	0.0500
$\bar{Y}$+s	6	0.30	0.85	0.3413	0.8413	0.0087
$\bar{Y}$+2s	3	0.15	1.00	0.1360	0.9773	0.0227

The question to be tested is whether the data of Table 5 approximates a normal distribution sufficiently well to justify use of that distribution in any subsequent analysis. The first step in the test is to calculate the mean and standard deviation of the data, namely Y = 100.0 and S = 9.47, and assume that they have the same values as the corresponding parameters of the distribution. The data is then broken into arbitrary groups bracketted by Y - 3S, Y - 2S, ... Y + 3S, as seen in Table 6. The second column is the number of points occurring in each group, the third is their fractional occurrence, and the fourth is the cumulative frequency. Columns 5 and 6 repeat 3 and 4 for the normal distribution having the parameter values stated just above; the numbers you see are read from a handbook table of the normal distribution. For example, 0.3413 is the tabulated area lying between the mean and the mean plus or minus one standard deviation.

The last column is the difference in each row between column 4 and column 7. The null hypothesis at this point is that there is no difference between the data and the exact distribution, save that which could occur by chance. Only if the difference could occur by chance fewer than one-in-twenty -- as before several times here -- will we abandon the supposed fit. At the 0.05 level just announced, and with 20 degrees of freedom, a Kolmogorov-Smirnov table shows a difference of 0.294. The largest difference in Table 6 is 0.05, indicating that the deviation from perfect fit shown by Figure 8 could occur with a frequency much greater than one-in-twenty.

Since the null hypothesis is upheld, it now seems better to discard the hoped-for Y vs X relationship, and to say that Y = 100 + e no matter what X is, where e is your choice for a measure of uncertainty, just as in Example 1. For instance, if Y were to be the basis for an expected value decision, then e = 0 is appropriate; if you wish to be 95 percent confident of predicting Y in a single trial, then e = 2S = 18.94, etc.

Look again at Figure 7, observing the plus-or-minus 1.0 STANDARD ERROR lines. If I stubbornly persist in using the regression line, and am asked the value of Y when X = 10, I would say Y = 94.4 + 18.6 for that 95 percent confidence. From the discussion just above, however, Y = 100.0 + 18.94 appears to be the preferred answer.

EXAMPLE 6

Here is an example something like examples 4 and 5, but approached in a slightly different way. Data (not printed here) on failure times of vacuum tubes is given on page 147 of reference 5. Over a period of 2064 hours, or 74,208 total tube

operating hours, 48 tubes fail. The fractional number of tubes surviving at any time is the reliability to that time. From knowledge of elementary reliability theory, one expects that the data should fit the line

$$R = e^{-\frac{48}{74,208}t} \tag{29}$$

and consequently that the data should fit a straight line on a semi-log plane. See Figure 9; in it equation (29) appears as the dashed straight line ("theoretical function"), and the cumulative data (fraction surviving) is the broken line ("observed function"). Is the straight line a good fit for the data?

The Kolmogorov-Smirnov test used in example 5 is again applied. Level of significance is once again 0.05, and degrees of freedom are 48. A handbook table of the function gives a difference value of 0.196. If we were to follow the procedure of example 5, we would compare this value to the largest difference between cumulative theoretical and cumulative data. The same information is shown in an alternative manner in Figure 9 by the two boundary lines plotted at $\pm$ 0.196 on either side of the theoretical line. One then notes that the observed line doesn't reach either boundary line, and concludes that no difference is as great as 0.196. The null hypothesis of no-difference-except-by-chance is thus upheld, and equation (29) survives as an adequate regression line for the data.

Adequate, yes, but is it the best? The question naturally arises just from looking at the figure. It appears that a straight line more closely approximating the data line could be put in even by eye. Working on a slightly more elegant level, I go back to data in the reference (pairs of times and fractions surviving at those times), take the natural log of the dependent data, then apply equations (2) and (3) to get the best linear regression line. The result on the semi-log plane is

$$Y = 0.1889 - 0.0009725\ X \tag{30}$$

which, upon putting back onto the linear plane, becomes

$$R = 1.208\ e^{-\frac{t}{1028}} \tag{31}$$

Figure 10 is a replot of Figure 9 with my regression line used. The boundary lines are also plotted. Since the data line does not approach them as closely as in the previous figure, the fit can be said to be better.[1]

[1] If you look closely, you will see that the boundary lines lie at a distance of 0.24 from the regression line. The reason is that the data actually has only 32 pairs of times and reliabilities. Simultaneous failures account for the 48 points used in the reference.

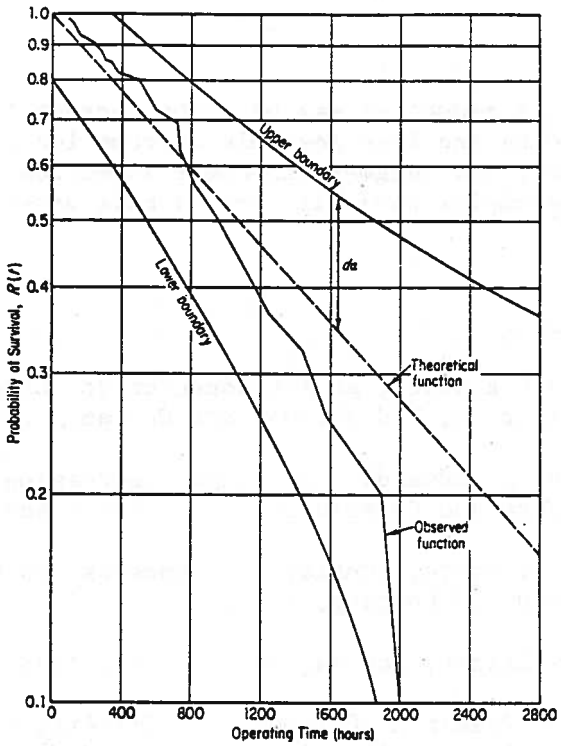

FIGURE 9

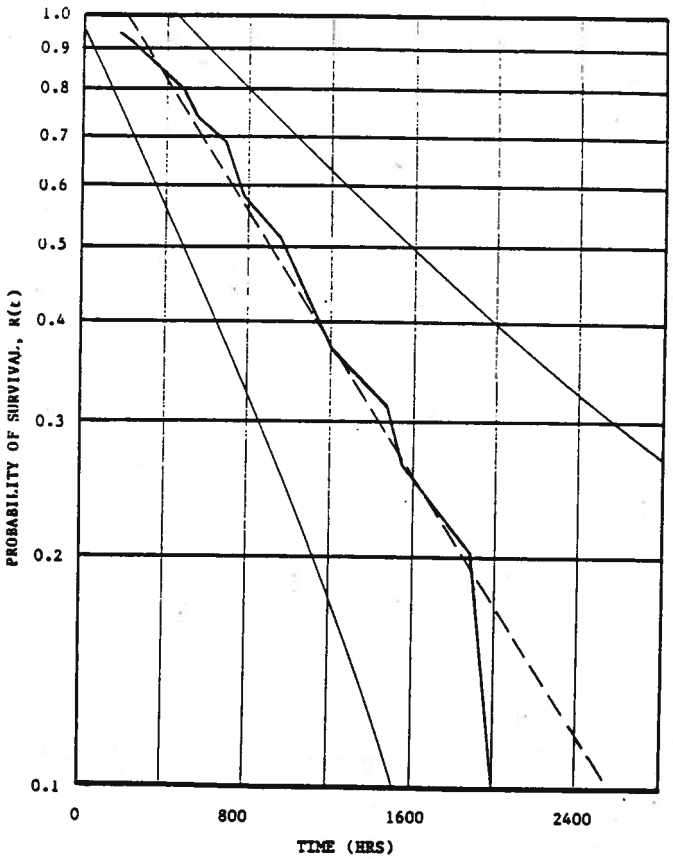

FIGURE 10

SUMMARY

A data reduction may be quite correct in its technique, yet not produce the best possible information. Knowledge of several techniques, and judgment in their selection, are both important. Several examples have illustrated this assertion.

REFERENCES

1. Allen L Edwards, An Introduction to Linear Regression and Correlation, W H Freeman and Company, 1976.

2. Allen L Edwards, Multiple Regression and the Analysis of Variance and Covariance, W H Freeman and Company, 1979.

3. David S Moore, Statistics: Concepts and Controversies, W H Freeman and Company, 1979.

4. E J Williams, Regression Analysis, John Wiley and Son, 1959.

5. ARINC Research Corporation (William H Von Alven, editor), Reliability Engineering, Prentice-Hall, 1964.