MASTER

PNL-SA-6902

TRAN - STAT

STATISTICS FOR ENVIRONMENTAL TRANSURANIC STUDIES[†]

April 1978
Number 4

Battelle Memorial Institute
Pacific Northwest Laboratory
Richland, Washington 99352

Coordinator: Richard O. Gilbert, Energy Systems
Department, Statistics Section,
(FTS) 444-7511, Ext. 946-2378

CONTENTS

	Page
Introduction	1
Summary	2
Results from TRAN-STAT Survey	3
Role of the Statistician in Research	5
Discussion	6
Effective Data Summaries and Displays (with examples)	6
Exploratory Versus Confirmatory Data Analyses	9
References	12

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

INTRODUCTION

This issue of TRAN-STAT is concerned with three related topics: (1) results from the questionnaire I included with the 3rd issue (February 1978) of TRAN-STAT, (2) the role of the statistician in transuranic studies, and (3) some effective ways for "exploring" data sets and communicating this information to others via effective data summaries.

Concerning the questionnaires, I feel it is important for you to know how other readers feel about TRAN-STAT and also what they would like future issues to discuss. Their ideas may bring other topics to mind that you would like me to address.

[†] Prepared for the U.S. Department of Energy, Division of Biomedical and Environmental Research, under Contract EY-76-C-06-1830

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

The role of the statistician in transuranic field studies is a topic I wanted to discuss early-on in the publication of TRAN-STAT. I sometimes feel that the contribution a statistician can make to a scientific inquiry is not fully achieved. This can be due to a number of reasons including: (1) incomplete college preparation of statistics students for work on real-world applied problems (2) the critical stance the statistician must take regarding any data set, (possibly irritating the principal investigator), (3) the considerable time and effort it may take for the statistician to become familiar with the subject matter so that he can really communicate with other members of the research team, (4) a belief, perhaps, on the part of some researchers that the statistician is primarily of use after data have been collected, (5) a basic unfamiliarity on the part of researchers of statistical concepts and procedures, and (6) a lack of knowledge of the contributions statistics can make to a study. Hopefully, this issue will contribute to better communication between researcher and statistician and that this will lead to transuranic studies that are well designed and statistically analyzed.

Our discussion of some effective ways for exploring, summarizing, and communicating data begins on page 6. We cannot give a complete discussion of this very broad topic here, but references for further reading are offered. In particular, the discussion here does not specifically address data collected at different locations, a topic for a future issue. Some examples using ^{241}Am soil data are used for illustration purposes.

Enclosed with this issue is a table of contents page for Issue 3 on compositing. You may find this helpful when reading that issue.

SUMMARY

The main points of this issue are:

- (a) Most of you who returned questionnaires are generally pleased with TRAN-STAT thus far, but the organization could be improved by providing better summaries or outlines and providing a table of contents.
- (b) You are interested in and have questions about a broad range of statistical topics.
- (c) The statistician has two major tasks: (1) to help design studies to obtain useful data with a minimum of time and expense, and (2) to assist in the analysis of data to help you reach valid conclusions and to uncover hints about how the study should proceed to the next stage.
- (d) The objective of statistical method is to make the learning process as efficient as possible.
- (e) Effective data handling techniques and display summaries play a vital role in uncovering what our data are trying to tell us and in communicating that information to others so that the learning process can continue to go forward in an efficient manner.

- (f) We should never limit our data analysis to computing the mean and standard deviation. The skewed nature of most transuranic data sets dictates additional summaries.
- (g) Stem-and-leaf displays and "box plots" are useful statistical analysis tools for transuranic data.
- (h) Data analysis can be broken down into two parts; exploratory and confirmatory. The former is in the nature of "detective work," while the latter implies making inferences from collected data to the population from which the data were obtained.
- (i) We should be careful about launching into confirmatory-type data analyses (tests of hypothesis, confidence intervals, etc.) when data are not of sufficient quality and quantity.

RESULTS FROM TRAN-STAT SURVEY

Enclosed with the 3rd issue of TRAN-STAT (February 1978) was a short questionnaire designed to provide me with information to help plan future TRAN-STAT issues. Thus far ten questionnaires have been filled out and returned. I'm very grateful for this response and hope others of you will also respond. The results from the ten returned questionnaires are:

The 3rd issue of TRAN-STAT dealing with compositing

	Yes	No	Don't know
(1) is too long	1	8	1
(2) discusses an important topic	10	0	0
(3) is too theoretical and not very helpful	2	8	0
TRAN-STAT serves a useful purpose and should be continued	10	0	0

Three persons suggested changes in the format of TRAN-STAT in response to the statement: "The format of TRAN-STAT should be changed in the following way." Their comments are listed below. Seven persons indicated no major changes were necessary.

- (i) "Suggest you add a Table of Contents to an issue, or a brief outline at least, for ready reference, especially keyed to those persons not in statistics but using statistical analysis in programs."
- (ii) "If possible, 3-hole punch left margin for loose-leaf binder for ready reference."
- (iii) "Summary or abstract of important points would help (one page)."
- (iv) "Perhaps a list of do's and don'ts for data analysis."
- (v) "Perhaps a glossary to define some of the statistical language for those not that familiar with the field of statistics."

Seven persons listed one or more topics they would like to see discussed in TRAN-STAT. These are (no order of importance intended):

1. Nonparametric statistics as they apply to Pu data.
2. Problem of nondetectable concentrations and how one can treat these statistically when they are mixed with detectable concentrations.
3. More discussion on what the sample is to be used for. To estimate inventory? Why? To account for some release? If it is for some other purpose, e.g., to tell us something about mobility, and transport ratios or processes, biological availability or fate, then sampling on a micro basis needs to be considered. Ten samples taken within a cubic meter or along a root pathway would be much more revealing than 10 samples taken over a hectare.

I would prefer to see the sampling question and design discussed in the same context as the hypothesis and the experimental design.

4. Isopleth mapping of radionuclides.
5. Experimental designs.
6. Effective methods for presenting data sets.
7. Soil sampling methodology for transuranic studies.
8. Sampling for input to modeling (parameter estimation).
9. Efficient designs to test for treatment effects when the CV (coefficient of variation) is large.
10. Appropriate transformations for analyzing transuranic data.
11. Error rates for testing hypothesis (Type I & II) with discussion of appropriate levels of each for research projects.
12. Transformations, whys and rules, etc., plus how to report the results.
13. Data bases and data management; uses, desirability of, problems, etc.
14. How to choose appropriate models.
15. Dose-response models, experiments, and statistics.
16. Concept of power of a statistical test.
17. Mean square estimators.

In general, I think the results of the survey speak for themselves. You will note that some of your suggested changes in format have been initiated with this issue; namely, a Table of Contents, a 3-hole punch left margin, and a summary.

The list of discussion topics will be very helpful in planning future issues of TRAN-STAT. However, I would feel more confident that I am really meeting needs if more of you would return your questionnaires. Only about 20% of the total distributed have been returned thus far. Of course, all of you are encouraged to call or write at any time to express your opinions.

ROLE OF THE STATISTICIAN IN RESEARCH

I noted in the 2nd issue of TRAN-STAT (November 1977) that it might be worthwhile to discuss the role of the statistician in environmental trans-uranic studies. My interest in this topic was renewed when I recently re-discovered an article by George E. P. Box (1974) (a well known and highly respected statistician at the University of Wisconsin) entitled "Statistics and the Environment." To me, as a statistician, this paper is a classic because it describes so well the role that statistics can play in solving our real-world problems. In this section I will attempt to convey his major points to you. This discussion naturally leads into the next section that uses a set of ^{241}Am soil data (collected by the Nevada Applied Ecology Group) to illustrate some methods for uncovering what data are telling us and conveying this information to others.

Dr. Box begins by noting the many urgent problems confronting our nation and that the Scientific Method will, hopefully, contribute to their solution in a timely manner. He states "Scientific Method is a process of controlled learning. The object of statistical method is to make that learning process as efficient as possible." Later he continues "The statistician's job is to advise and assist the investigator in two crucial tasks, so as to allow the investigator to employ his talents most efficiently. These tasks are: (1) deciding what would be appropriate data to get at each stage of the investigation. Broadly we can call this the design problem. (2) deciding what the data entitles us to believe at each stage of the investigation. We call this the analysis problem." Dr. Box considers the design problem of "paramount importance," the analysis of data taking a secondary, but still very important role. He states

"In summary then, we learn through numbers. But what numbers or data should we try to get and what do they mean when we have them? These are the questions that good statisticians are trained to answer. It is very easy to acquire useless or irrelevant data. It is very easy to be misled by data once they are acquired. The design of *each stage* of an enquiry so as to produce useful data with the minimum of time and expense, and the analysis of data of each stage so as to produce, not only valid conclusions, but also valuable hints on how the investigation ought to proceed, these are the two critical tasks in which the statistician plays a key role."

Dr. Box also touches on two other aspects of the statisticians job. First, the critical stance the statistician must take when examining the data relative to the hypotheses or model being investigated. He must become a "doubting Thomas" and be prepared to abandon or modify models that do not agree with data. Second, "another part of the statisticians job is to make sure that statistics and computers do not separate the investigator from his data but, on the contrary, help him to see his data from many different angles."

The remainder of Dr. Box's paper is under the heading of Coping with Natural Variation. He discusses how it is the structure of this variation which determines how we can extract the information contained in data. He notes how the appropriate study of components of variation can lead to more efficient designs; how investigators often error in building models that are too elaborate;

the menacing problem of the "open loop," i.e., of the absence of feedback and communication that results in stagnation; and finally, on the "scarcity of trained statisticians competent to deal with real problems."

Discussion

Dr. Box states that the most crucial task of the statistician is at the design phase of a study. I certainly agree with this view and strongly encourage the active participation of an experienced statistician at the very beginning of any study. If the statistician has had previous experience in environmental transuranic studies he should be able to assist in the definition of study objectives, in determining the type of data to be collected and the choice of measurement instrumentation and laboratory procedures. Furthermore, he can contribute to the writing of field sampling protocols and, of course, to deciding how many samples are required as well as their placement (random, systematic, etc.) in the field.

If a statistician has not had previous transuranic experience, both he and the other investigators will need to invest a certain amount of time and effort to "bring him up to speed." Over the long run, this investment should be worth the effort, and the effort involved may not be large. A note of caution. Don't expect the statistician to work miracles. He is not a magician with a bag of tricks for sale. And he is limited of course, by the quantity and quality of the data and other information known about the area to be studied.

You might enjoy reading the nontechnical articles by Bliss (1969) and Bross (1974) that deal with the relationship between the statistician and his nonstatistical coworkers or clients. The paper by Harshbarger (1974) is also interesting since it is an example of the rather important impact a statistician can have on the conclusions of a study by being involved at the design stage.

EFFECTIVE DATA SUMMARIES AND DISPLAYS

Based on Dr. Box's comments as well as my own experience, it seems important to (1) use statistical analysis methods that are effective in uncovering what the data have to say, and (2) use data display techniques that are effective at communicating information to others via publications and oral presentations. To illustrate these two points we examine here some data collected by the Nevada Applied Ecology Group (NAEG) (Department of Energy, Nevada Operations Office, Las Vegas).

During the summer of 1977 the NAEG collected a set of 24 adjacent surface (5 in. diameter, 5 cm deep) soil samples at Nuclear Site-201 on the Nevada Test Site to examine the question of the effect on concentration level and variability from using different aliquot sizes. A complete account of this study is given by Doctor and Gilbert (1978). Interest in this topic arose from the work of Michels (1977) and Wallace and Romney (1978). Each sample was dried, ball-milled for 5 hours, and screened through a 10-mesh screen. The <10 mesh soil from the 24 samples were composited and mixed by kneading for 10 minutes. The soil was spread out on a sheet in a circular pattern

and divided into 20 equal pie-shaped parts. Each part was placed in a separate bottle. Five aliquots, one each of 1g, 10g, 25g, 50g, and 100g were withdrawn from each bottle in a manner to guard against biasing results with respect to aliquot size. The 100 aliquots were Ge(Li) counted for ^{241}Am for 1000 minutes or until the percent counting error (2σ) was 10% or less (counting errors turned out to be about 7%). Special samples were also prepared to provide radiation counter calibration for each size aliquot. I would be glad to send a copy of the study protocol to those of you interested in more details on the design of this study.

I received the analytical results and eagerly computed the mean concentration and standard deviation for each aliquot size. These results are given in Table 1. My first reaction was one of disappointment because the means were about the same for each aliquot size. I was disappointed because, based on preconceived notions, I had expected to see smaller means for the smaller aliquot sizes. The standard deviations did decrease with increasing aliquot size as I had anticipated. At the time I was too busy doing other things to look more closely at the data, but I showed the results to Pam Doctor with some words to the effect that the data didn't seem to be saying what I had expected with regard to mean levels.

TABLE 1. Means and Standard Deviations of ^{241}Am (Ge(Li)) Concentrations (nCi/g) in Soil Aliquots of Different Sizes from Nuclear Site-201, Nevada Test Site[†]

	Size of Aliquot (grams)				
	1	10	25	50	100
Number of Aliquots	20	20	20	20	20
Mean	1.93	1.82	1.80	1.84	1.92
Standard Deviation	1.52	0.48	0.57	0.20	0.17

[†]Data collected as part of the Nevada Applied Ecology Group Environmental Transuranic Studies Program

A few days later I saw Pam again. She had looked at the data and felt "something was going on." So, I then took the time to jot down the data for each aliquot size in a stem-and-leaf display (Tukey, 1977; Tukey, 1972; Gilbert et al., 1977). The results are given in Table 2. Now we have a much better picture of the data. We see that the distribution of the 20 data points for the 1 gram aliquots is much different than that of the 50g and 100g size. It is more highly skewed and covers a wider range of low and high concentrations. We note that the median rather than the mean concentration increases with aliquot size. The mean does not show this behavior because of the highly skewed nature of the smaller aliquot sizes. It is interesting to note also,

TABLE 2. Stem-and-Leaf Displays of ^{241}Am (Ge(Li)) Concentration (nCi/g) in Soil Aliquots of Different Sizes from Nuclear Site-201, Nevada Test Site. Data collected as part of the Nevada Applied Ecology Group Environmental Transuranics Study Program

Size of Aliquot	1 gram	10 gram	25 gram	50 gram	100 gram
	Stem Leaf	Stem Leaf	Stem Leaf	Stem Leaf	Stem Leaf
	1.0 4 5	1.0	1.0	1.0	1.0
	1.1 0 7	1.1	1.1	1.1	1.1
	1.2 2 2 7	1.2	1.2 3 7	1.2	1.2
	1.3	1.3 6	1.3 3 3 6	1.3	1.3
	1.4 5	1.4 1 6 7 9	1.4 4	1.4	1.4
	1.5 4 7	1.5 0 4	1.5 1 2 2 4 8	1.5 5	1.5
	1.6	1.6 6 7 9	1.6 1	1.6 3 7 8 9	1.6
	1.7 1 1 3	1.7 3 7	1.7 1 5	1.7 0 1 2 3 7	1.7 0 2 4 5 8 9
	1.8 6	1.8 3 6	1.8	1.8 6 7 9	1.8 0 1 3
	1.9	1.9 3 5	1.9	1.9 0 4	1.9 0 1 5 6 9
	2.0	2.0 0 7	2.0	2.0 0 1 2 3	2.0 0 6
	2.1 3	2.1	2.1 9	2.1	2.1 1 3
	2.2	2.2	2.2 0	2.2	2.2 4
	2.3 0	2.3	2.3	2.3	2.3 0
	2.4	2.4	2.4 0 6	2.4 6	2.4
	2.5	2.5	2.5	2.5	2.5
	2.6 9	2.6 3	2.6 8	2.6	2.6
	2.7	2.7	2.7		
	2.8 0	2.8	2.8		
	2.9	2.9	2.9		
	3.0	3.0	3.0		
	3.1	3.1	3.1		
	3.2	3.2	3.2		
	3.3	3.3	3.3		
	3.4	3.4 3	3.4 0		
	3.5	3.5	3.5		
	3.6				
	3.7				
	3.8				
	7.9				
	8.0 0				
	8.1				
n	20	20	20	20	20
Mean ($\bar{X}$)	1.93	1.82	1.80	1.84	1.92
Median	1.56	1.71	1.56	1.82	1.91
Standard Deviation (s)	1.52	0.48	0.57	0.20	0.17
Coefficient of variation (s/ $\bar{X}$)	0.79	0.26	0.32	0.11	0.09

(Example of how to read: the row 1.3|3 3 6 in the 25 gram column represents three concentrations whose leaf digits are 3, 3, and 6; i.e., the three concentrations are 1.33, 1.33, and 1.36.)

that if the single extreme data point (8.0 nCi/g) for the 1 gram aliquot size had, by chance, not been obtained, the mean would have been 1.61 ($n = 19$) rather than 1.93. This illustrates the large variation in $\bar{X}$ due to sampling error for the 1g size.

While the data of this example are themselves very interesting, the important point I want to make here is that we should never limit our data analysis to computing the arithmetic mean and standard deviation. The skewed nature of most transuranic data makes it imperative that we examine the entire data set to determine its structure. The stem-and-leaf display can be jotted down rapidly and is usually a convenient way to learn quickly about the distribution of a data set, if say 20 or more observations are available. The familiar histogram is also useful, but it does not show the actual data values as does the stem-and-leaf.

Figure 1 is a display of "box plots" that might be suitable for use in an oral presentation of the data. The main features of the data are evident pretty much at first glance. Also, the reader can verify that the data are well described by a log-log plot of standard deviation(s) versus aliquot size (grams). The five data points fall quite nicely along a straight line over the range of aliquot sizes considered here (1 to 100 grams). This linear plot could be an effective way to describe and summarize these data. (This linear relationship may not hold, of course, for other aliquot sizes or for other radionuclides at this or other locations where contamination source, spatial distribution, concentration level, or other variables may be different.)

Modifications to Figure 1 could be made to display more information if desired.¹ For example, data values could be plotted as dots along the major axis to show their location. The formats of Table 2 or Figure 1 could be useful for comparing visually concentrations obtained for different times (seasons), locations, or plant and animal species. Tukey (1972) discusses a number of imaginative display devices that could prove useful. McGill et al. (1978) present some interesting variations on box plots that display visually the number of observations in each group and that give an indication of rough significance of differences between medians. The point of this discussion is that the solution of real world problems is likely to progress more rapidly if we fully understand and communicate effectively our present state of knowledge. This will help to facilitate and motivate the kinds of studies that will be most effective in finding solutions to problems.

EXPLORATORY VERSUS CONFIRMATORY DATA ANALYSIS

In the previous section I have been talking primarily about what Tukey (1977) terms "exploratory data analysis," or "detective work." This is to be

¹Note that in Figure 1 the nCi/g scale goes from small values (0.5) at the top of the page to a large value (8.2) at the bottom. This is unconventional and should probably be reversed so that large values are at the top. The present arrangement corresponds with the stem-and-leaf scale in Table 2.

Figure 1. Box Plots of ^{241}Am Concentrations (nCi/g) in Soil Aliquots of Different Sizes from Nuclear Site-201, Nevada Test Site. Data Collected as Part of the Nevada Applied Ecology Group Environmental Transuranics Study Program.

contrasted with "confirmatory data analysis." To help explain the difference we quote Tukey (1977, p. 3) "The processes of criminal justice are clearly divided between the search for the evidence--in Anglo-Saxon lands the responsibility of the police and other investigative forces--and the evaluation of the evidence's strength--a matter for juries and judges. In data analysis a similar distinction is helpful. Exploratory data analysis is detective in character. Confirmatory data analysis is judicial or quasi-judicial in character." He also states, "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step."

Figure 1 and Table 2 come under the heading of exploratory data analysis because our main purpose has been to describe (or explore) what these data are like. Once we understand the data, we can consider whether the data are sufficient for the purpose of making inferences or "confirmations" about the "population" from which the data have been drawn. The data considered here certainly suggest a number of interesting hypotheses about the relation between aliquot size and ^{241}Am or plutonium variability between aliquots, but I would be hesitant to make strong inference statements without further information. I think it is sometimes tempting to put too much faith in our data and to begin making tests of hypotheses, computing "exact" confidence intervals, and performing sophisticated statistical analyses on data sets that are not of a quality commensurate to the task. A safer course would be to rely on good data descriptions that will, hopefully, indicate fruitful areas for more complete studies. My personal viewpoint is to look upon tests of hypotheses and other "inferential procedures" as exploratory rather than confirmatory. Perhaps I'm conservative in that regard, but the great variability we have all observed in transuranic data would seem to suggest caution until we have more experience in designing and analysing field studies.

REFERENCES

- Bliss, C. I. 1969. "Communication Between Biologists and Statisticians, A Case Study." The American Statistician 23, No. 4, pp. 15-20.
- Box, G. E. P. 1974. "Statistics and the Environment." Journal of the Washington Academy of Sciences 64, No. 2, pp. 52-59.
- Bross, I. D. J. 1974. "The Role of the Statistician: Scientist or Shoe Clerk." The American Statistician 28, No. 4, pp. 126-127.
- Doctor, P. G. and R. O. Gilbert. 1978. "Two Studies in Variability for Soil Concentrations: With Aliquot Size and With Distance." Presented at the Nevada Applied Ecology Group Pu Information Conference, San Diego, February 28 - March 1, 1978. To be published in NVO-192, Nevada Operations Office, Las Vegas.
- Gilbert, R. O., L. L. Eberhardt, and D. D. Smith. 1977. "An Initial Synthesis of Area 13 ^{239}Pu Data and Other Statistical Analyses." Environmental Plutonium on the Nevada Test Site and Environs (White, M. G., P. B. Dunaway, and W. A. Howard, eds.). USDOE, Las Vegas, NVO-171, pp. 237-274.
- Harshbarger, B. 1974. "An Example of a How a Designed Experiment Saved Research Workers from False Recommendations." The American Statistician 28, No. 4, pp. 128-129.
- McGill, R., J. W. Tukey, and W. A. Larsen. 1978. "Variations of Box Plots." The American Statistician 32, No. 1, pp. 12-26.
- Michels, D. E. 1977. "Sample Size Effect on Geometric Average Concentrations for Log-Normally Distributed Contaminants." Environmental Science and Technology 11, No. 3, pp. 300-302.
- Tukey, J. W. 1972. "Some Graphic and Semigraphic Displays." Statistical Papers in Honor of George W. Snedecor (Bancroft, T. A. and S. A. Brown, eds.). The Iowa State University Press, Ames, Iowa, pp. 293-316.
- Tukey, J. W. 1977. Exploratory Data Analyses, Addison-Wesley Publishing Company, Reading, MA.
- Wallace, A. and E. M. Romney. 1978. "The Particle Size Problem in the Assay of ^{239}Pu in the Field." Transuranics in Desert Environments (M. G. White and P. B. Dunaway, eds.). USDOE, Las Vegas, NVO-181 (In Press).