(21) Kennedy, B. A., McQuarrie, D. A., Brubaker, C. H., Jr., *Inorg. Chem.*, **3**, 265 (1964).
(22) Lewis, J., Wilkins, R. G., "Modern Coordination Chemistry," Interscience Publishers, Inc., New York, N. Y., 1960.
(23) Martell, A. E., Calvin, M., "Chemistry of the Metal Chelate Compounds," Prentice-Hall, Inc., New York, N. Y., 1952.
(24) McDonnell, P. M., Pasternack, R. F., *J. Chem. Doc.*, **5**, 56 (1965).
(25) Morris, M. L., Busch, D. H., *J. Am. Chem. Soc.*, **78**, 5178 (1956).
(26) Pasternack, R. F., McDonnell, P. M., *Inorg. Chem.*, **4**, 600 (1965).
(27) Pfeiffer, P., in "Stereochemie," K. Freudenberg, Ed., Franz Deuticke, Leipzig and Vienna, 1932, pp. 1200–1377.
(28) Schwarz, R., "Chemistry of the Inorganic Complex Compounds," translated by L. W. Bass, John Wiley and Sons, Inc., New York, N. Y., 1923, p. 51.
(29) Sievers, R. E., Bailar, J. C., *Inorg. Chem.*, **1**, 174 (1962).
(30) Sutherland, M. M. J., "Textbook of Inorganic Chemistry, Metal-Ammines," Vol. X, J. N. Friend, Ed., Charles Griffin and Co., London, 1928, p. 10.
(31) Swallow, A. G., Truter, M. R., *Proc. Chem. Soc.*, 166 (1961).
(32) Swaminathan, K., Busch, D. H., *J. Inorg. Nucl. Chem.*, **20**, 159 (1961).
(33) Trimble, R. F., Jr., *J. Chem. Educ.*, **31**, 176 (1954).
(34) Weinland, R. F., "Einfuhrung in die Chemie der Komplexverbindungen," 2nd Ed., Enke, Stuttgart, 1924.
(35) Werner, A., *Z. anorg. Chem.*, **3**, 310 (1893).
(36) Werner, A., *ibid.*, **14**, 24 (1897).
(37) Werner, A., "New Ideas on Inorganic Chemistry," translated by E. P. Hedley, Longmans, Green and Co., London, 1911; "Neuere Anschauungen auf dem Gebiete der anorganischem," 3rd Ed., F. Vieweg und Sohn, Braunschweig, 1913.
(38) Wittig, G., "Stereochemie," Akademische Verlagsgesellschaft m.b.H., Leipzig, 1930.

Experimental Designs in Work and Time Studies*

DONALD W. KING
Westat Research Analysts, Inc., Bethesda, Maryland

Received June 22, 1965

I. SUMMARY

This expository paper is addressed to the role of statistics in work and time studies involving input to information storage and retrieval systems. General statistical techniques for gathering data by observation and by experimentation are discussed. Examples are given to illustrate the importance of proper sample and experimental design with regard to reliability and cost.

II. GENERAL COMMENTS

One obtains statistical information in time and work studies to measure characteristics associated with individuals and other system components, and to provide a rational basis for selection among alternative system components. Management must specify system properties of interest such as quality, rate of work, and cost. Furthermore, specific attributes of these properties must be chosen prior to observation or experimentation. For example, quality of indexing may be described by indexing accuracy or consistency (*1*). Analysis should be planned to yield the best information upon which to base decisions. For instance, statements of indexing accuracy are more valuable to management if accompanied by statements concerning the effect of indexing accuracy on missed documents and false drops (*2*).

System characteristics are generally measured by averages, totals, proportions, and measures of variability. A measure of dispersion should accompany each average, total, or proportion to provide a better description of the population studied. For example, average daily room temperature of 70° F. has little meaning for describing work conditions if the temperature ranges from 50 to 100° F. during the day. Average accuracy for an entire group of indexers may be high, but one or more individuals may differ widely from the average. The standard deviation is the most common measure of dispersion. (Standard deviation for a finite population is $\sigma = [\Sigma(X - \mu)^2/N]^{1/2}$, where μ is the population mean, X is an observation from the population, and N is the number found in the population.) The standard error of a mean (average) is measured by the standard deviation divided by the square root of sample size. In a sense, standard error is a description of reliability of statistical estimation.

III. SAMPLE DESIGN

Observational information is obtained by examining a portion of a population in order to make generalizations concerning the entire population. The generalizations may be estimates of population characteristics or tentative hypotheses concerning the population under investigation. Generally, no attempt is made to modify or control any aspect of the system being observed. Statistical sampling theory deals largely with seeking a balance between reliability and cost of sample estimates, while sample design is the mechanism for accomplishing this. Sample design includes the sample plan and estimation procedure.

*Work accomplished under contract to the U. S. Patent Office; presented before the Division of Chemical Literature, Symposium on Work and Time Studies in Technical Information, 149th National Meeting of the American Chemical Society, Detroit, Mich., April 1965.

The sample plan is the procedure used to select sampled elements from the population, including determining sample size and specifying the plan for randomization. The role of randomization is frequently misunderstood. It does not guarantee a "representative sample," as is sometimes supposed, since, by definition, a representative sample must have all characteristics of the population in their correct proportions. The degree of representativeness of a random sample is never known. However, randomization is relied upon to eliminate bias in selection and to provide a valid measure of the reliability of estimates.

One often must compromise with strict rules of randomization. For instance, sampled elements may be systematically chosen from a large list (for example, every 50th item) if there is no reason to suspect periodicity in the sequence of listing. Furthermore, all elements of a population need not be given an equal probability of being included in the sample. However, every element must have a *known* probability of being represented in the sample. Suppose one is sampling from the output of two indexers and the first indexes an average of five documents per day and the other ten documents per day. If documents are chosen randomly from the entire collection, the output of the second indexer has twice the chance of being included in the sample as the output of the first indexer. For some purposes this may be the weighting one desires, but for some other purposes it may not be. Finally, if a small number of elements from a population, say less than four, are chosen for observation, they should be chosen selectively, rather than randomly. For instance, if two indexers are chosen for observation they should be selected by their supervisor as being "representative" of the group. Here the chance that randomization will not yield a typical case outweighs the importance of bias.

Sample size is also an important and sometimes difficult aspect of the sample plan. If the standard deviation is known from previous observations or can be estimated from a preliminary sample, one may choose a sample size to regulate estimation reliability since standard error is inversely proportional to the square root of sample size. Intuitively, one can see that reliable estimates can be obtained with a small sample from a large population if dispersion or variability of data is small. In fact, a sample size of one is all that is necessary if there is *no* variability; *i.e.*, all observations have the same value. Increasingly variable populations require increasingly large samples to obtain equivalent levels of reliability. An example is given to illustrate the relationship between estimation reliability and sample size. Assume that the relative frequency of correct indexing is being measured, that the observed variables have properties of a binomial distribution, and that relative frequency of correct indexing is 0.95; *i.e.*, 95% of the correct terms were in fact indexed. The relationship between standard error and sample size is given in Figure 1. One can see that little is gained in reducing standard error by increasing sample size beyond a certain point. For instance, an increase in sample size from 250 to 500 observations yields quite a different reduction in standard error than an increase from 750 to 1000 observations.

Sometimes data yield unacceptable estimation reliability. In this event, the estimated standard error often can be reduced for a given sample size by implementing an alternative sample plan. For example, the estimated standard error may be reduced by partitioning the population into mutually exclusive and exhaustive strata in which the observations within strata are as homogeneous as possible. Each stratum is then sampled. This form of restricted randomization plan is stratified random sampling. Examples of stratifying factors which can be used to subdivide populations are document length, indexer experience, and days of the week. The stratifying factors are constructed to minimize variability within each stratum and to maximize variability among strata. The estimated standard error can be reduced by the amount attributed to differences among strata.

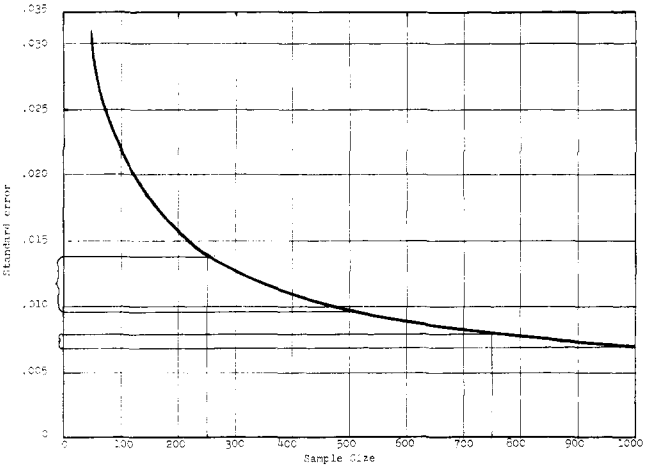

Figure 1. Standard error *vs.* sample size for relative frequency of correct indexing equal to 0.95.

An example is given to illustrate a gain in reliability by stratification. Assume that the total time required to index a large set of documents is estimated from the hypothetical sample observations in Table I. The number of pages in the documents is used as a stratifying factor since indexing time is presumably related to this factor. The total population contains 10,000 documents and the sample 100 documents. The population is stratified into groups of sizes N_i ($i = 1, 2, 3, 4$).

The samples sizes n_i are chosen in proportion to size of the population. The total indexing time in days is estimated to be 1050 days for the entire file of 10,000

Table I. Hypothetical Indexing Time (min.) for Documents Stratified by Number of Pages

Number of pages									
0–10				11–25			26–75		76
45	38	30	21	32	71	53	54	91	92
16	24	25	32	49	56	29	62	69	105
32	16	23	17	39	48	57	95	83	79
27	22	28	29	22	61	63	78	81	112
8	29	14	37	62	48	79	81	68	95
22	12	26	19	45	37	54	67	72	112
18	19	31	25	29	43	41	59	94	116
34	29	26	29	58	47	54	62	67	103
27	23	43	22	52	41	46	131	71	94
21	36	28	15	46	66	58	73	83	85
$n_1 = 40$, $N_1 = 4000$				$n_2 = 30$, $N_2 = 3000$			$n_3 = 20$, $N_3 = 2000$		$n_4 = 10$, $N_4 = 1000$

documents. The estimated standard error, ignoring number of pages, is 118 days and the estimated standard error from the stratified sample is 26 days. Thus, estimates from a stratified random sample yield a considerable gain in reliability for this example.

If strata variability is known, the estimated standard error of total time can be further reduced by a technique of optimum allocation. Furthermore, cost can also be considered in the allocation if it is more costly to obtain information from a sample unit in one stratum than in another.

In some instances one may not be able to identify all elements of a population from which he wishes to draw a sample. For example, suppose one wishes to sample from a population of all compounds in a large collection of documents. Obtaining a complete list of all compounds is virtually impossible. One method of sampling in such cases is to choose a random sample of documents initially and then choose a random sample of compounds within those documents. This sample plan is a two-stage sample with documents identified as primary units and compounds within documents as secondary units. Sample designs have been devised to cope with many problems of this kind, but their characteristics are too technical to discuss here. Cochran's book is a standard reference to sampling theory and practice (*3*).

IV. EXPERIMENTAL DESIGN

The previous section discussed information obtained from observational data. Experimental information is determined by controlling or modifying certain experimental factors in order to measure effect of changes or to compare the effect of different conditions. It is difficult to assign cause and effect by analysis of observational data. The fact that some documents may have been indexed accurately and also indexed rapidly does not mean that indexing faster causes higher accuracy. The relationship might be attributed to the fact that the documents are short. One may observe that college graduates are better indexers than high school graduates, but he cannot say on the basis of this information alone that their better performance is due to college training. Better indexing might be attributed to the possibility that persons who go to college have greater intelligence, or more social contacts, or a combination of factors. The point is that the observational data collected will neither confirm nor deny any of these hypotheses. Thus, when cause and effect relationships are of interest, experimental rather than observational techniques must be utilized.

The general statistical principles stated previously also apply to experimental design. That is, reliability of the information found from experimentation can be measured and controlled by proper experimental design. Furthermore, cost of experimentation can be minimized for specified reliability by constructing experiments intelligently.

Experimental designs are principally concerned with arrangement of experimental factors. Some experimental factors are chosen to establish their effect, and these are termed "experimental treatments." Other experimental factors may be incorporated into the experiment to reduce the experimental variability, thus providing better analysis and reliability of treatment effects. The uncontrolled variation in experimentation is called experimental error. An experimental design might be set up to determine the effect of two difference code sheets on indexing time with two controlled levels of indexer experience and various levels of document sizes. The different code sheets are experimental treatments and indexer experience and document size are factors incorporated to control variability.

When an effect can be attributed to a combination of factors which cannot be separated, the effects are said to be confounded. In the example above, if an indexer indexes one document with the first code sheet and another document with the second code sheet, the code sheet differences are confounded with document differences. That is, the observed differences between the two indexings may be attributed to code sheets, to documents, or to both, and one has no way of knowing which. Statistical techniques for handling confounding are replication, randomization, and partitioning. Replication is the repetition of the experiment using different experimental units. Randomization is the use of a random process to assign experimental units to treatments. Partitioning is the process of separating factors into mutually exclusive groupings. A common technique is to separate two or more factors into an array of cells in which experimental units are assigned to every combination of different levels of the factors. For example, an indexing experiment may be conducted in a new file of documents to determine (1) the effect of learning over time and (2) the effect of indexer experience on indexing accuracy. The arrangement for this experiment might be as given in Table II. The experimental units are documents and at least one document (experimental unit) is assigned randomly to each cell. If more than one document is used the experiment is replicated. Note that in Table II experience is confounded with particular indexers. One could avoid this defect by choosing two or more indexers at each level of experience. Indexer effects should be isolated from experience effects if a large difference between indexers within levels of experience is anticipated.

Table II. Two-Factor Design

		Learning time,[a] weeks					
		L_1	L_2	L_3	L_4	L_5	L_6
Indexing experience, years	E_1	D_1	D_2	D_3	D_4	D_5	D_6
	E_2	D_7	D_8	D_9	D_{10}	D_{11}	D_{12}
	E_3	D_{13}	D_{14}	D_{15}	D_{16}	D_{17}	D_{18}
	E_4	D_{19}	D_{20}	D_{21}	D_{22}	D_{23}	D_{24}

[a] The subscripts represents number of weeks. [b] The subscripts represent number of years.

Frequently, a statistical model is hypothesized for an experiment. In the experiment above the model might be

$$Y_{ijk} = \mu + L_i + E_j + (LE)_{ij} + e_{ijk}$$

where μ is a general constant, L_i is effect due to the ith learning time period, E_j is effect due to the jth level of experience, $(LE)_{ij}$ is effect due to the interaction between experience and learning time in the ijth cell, and e_{ijk} is a random error due to the kth replicate in the ijth cell.

Interaction is the joint effect of two variables. The effect of interaction can be an important part of the

analysis of an experiment. In the example, the effect of interaction between experience and learning time may be that experienced indexers learn faster. Assume that low accuracy of indexing is overcome by reviewing the work of the indexers. The effect of interaction is then important for budgeting since a portion of the indexers require review for a time while others may not.

One problem in work and time studies is that indexers learn from indexing a particular document. Suppose, for instance, an experiment is required to compare two indexing procedures for indexing accuracy. An indexer should repeat his work with the two procedures on the same document to eliminate confounding. However, the indexing accuracy should improve when indexing a document the second time by the second procedure due to the learning effect of having already indexed the same document previously. This is referred to as carryover effect.

Two experimental designs have been used in the Patent Office to cope with carryover effect for work studies involving evaluation of searching with two search systems. The first is a crossover design. The crossover design can be applied to indexing as shown in Table III. In this arrangement $D_1, \ldots, D_8$ are documents and P_1 and P_2 are indexing procedures. In this design the carryover effect is balanced. The numbers of documents and indexers to use in the experiment depend on the variability of the characteristic being observed. There can still be bias in the results if the carryover effects are not equal, that is, if one learns more from one procedure than from another.

Table III. An Experimental Design to Cope with Carryover Effect

Order of indexing	Indexer 1				Indexer 2			
	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
First indexing	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
Second indexing	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1

Another experimental design to cope with carryover effect is given in Table IV. In this arrangement each indexer does one set of documents with each procedure. If variability due to documents is excessive, the indexers should index the same set of documents with a different procedure; *i.e.*, replace $D_9, \ldots, D_{12}$ with $D_5, \ldots, D_8$ and $D_{13}, \ldots, D_{16}$ with $D_1, \ldots, D_4$. The cost of experimentation is higher, however, since fewer completed documents can be added to the file.

The examples in this section are oversimplified to illustrate the basic principles of experimental design. A standard reference to common experimental designs is the book by Cochran and Cox (*4*). A more complex example of an indexing experiment in the Patent Office is given in the next section.

Table IV. An Experimental Design to Cope with Carryover Effect

Indexer 1		Indexer 2	
P_1	P_2	P_1	P_2
D_1	D_5	D_9	D_{13}
D_2	D_6	D_{10}	D_{14}
D_3	D_7	D_{11}	D_{15}
D_4	D_8	D_{12}	D_{16}

V. AN EXAMPLE

The problems discussed previously are typical of those found in indexing chemical compounds found in patent documents. An indexing experiment is being performed currently in the heterocyclics file of the U. S. Patent Office. A document is first analyzed by extracting compounds from the contents of the document and drawing their structures on a card. A second phase involves fragmenting the compounds and indexing (or ciphering) the fragments. Each phase may be performed singly or with a review. That is, there are four alternate indexing procedures: single analysis–single indexing, single analysis–indexing reviewed, analysis reviewed–single indexing, analysis reviewed–indexing reviewed. Data are accumulated for frequency of indexing errors. The principal analysis consists of determining the effect of the four indexing procedures on accuracy and time. Three types of compounds are also considered.

The basic experimental arrangement is given in Table V. Each experimental replication consists of four indexers who index six compounds each, one of each type that was analyzed by a single analyst and one of each type that was analyzed by an analyst and reviewed.

Table V. Split–Split Plot Experimental Design Applied to an Indexing Experiment of the U. S. Patent Office Heterocyclics File

Design I[a]

		Rep 1		
		T_1	T_2	T_3
	A	$I_1\ I_2\ I_3\ I_4$		
		c_1	c_2	c_3
	AR	c_4	c_5	c_6

Design II[a]

	Rep 1′							
	I_1		I_2		I_3		I_4	
	A	AR	A	AR	A	AR	A	AR
T_1	U R_1 R_2							
	c_1	c_4	c_7	c_{10}	c_{13}	c_{16}	c_{19}	c_{22}
T_2	c_2	c_5	c_8	c_{11}	c_{14}	c_{17}	c_{20}	c_{23}
T_3	c_3	c_6	c_9	c_{12}	c_{15}	c_{18}	c_{21}	c_{24}

[a] Rep, replications, T, types of compound, A, analysis, AR, analysis reviewed, I, indexer, c, compound, U, unreviewed indexing mode, R_1, R_2, reviewer 1 and reviewer 2 indexing mode.

An indexer is not allowed to index a compound more than once, *i.e.*, a compound (1) analyzed and (2) analyzed and reviewed is not indexed twice by the same indexer. Thus, the indexer carryover effect is eliminated at the cost of confounding analysis procedures and chemical compound effects. In this design best estimates can be made for indexer effects, interactions between indexers and analysis procedures, and interactions between indexers and types of compounds.

This experimental design is called a split–split plot design. The name is a carryover from early agricultural experiments. If replications and types of compounds are considered alone, the experiment becomes a randomized block design with replications corresponding to blocks and types of compounds corresponding to treatments.

The design does not yield information about the effect of reviewing the indexing. An alternative arrangement of factors which yields this information is constructed by constructing four replicates from the following: (1) the first indexer's six indexing from replicate 1, (2) the second indexer's six indexings from replicate 2, (3) the third indexer's six indexings from replicate 3, and (4) the fourth indexer's six indexings from replicate 4. Each indexed compound is then reviewed by two independent reviewers. Thus, the alternative arrangement splits the indexer plots into reviewed and unreviewed plots. The results of the two reviewers can then be compared with the unreviewed indexing. This design provides the best estimates of the effect of the two indexing procedures. A number of interactions between the various main effects can also be evaluated.

Analysis of the experiment yields information concerning the following general questions:

(1) What is the trade off between accuracy and time for four alternative indexing procedures?
(2) Should different types of compounds be indexed by different indexing procedures?
(3) Are some types of compounds more difficult for different indexers?
(4) Are there other significant interactions between the different factors?

The cost of the experiment is important. The principal cost is derived from repeating the indexing procedure four times. However, the cost is not excessive since the basic experimental unit is a single compound and the time required to index a compound is short.

Accuracy of indexing is measured by the relative frequency of omitted and committed fragments in this experiment. Assessment of errors of indexed compounds is relatively accurate. Thus, only one person is used to assess errors. Estimates of indexing accuracy will be used as parameters of a retrieval model to determine the effect of these errors on the ultimate retrieval in terms of missed documents, false drops, and total retrieved. This procedure (1) provides a much better basis for selecting among alternative indexing procedures and (2) provides a means for setting standards for an indexing quality control program.

An entire patent document is used as an experimental unit for assessing indexing time for four indexing procedures. In this case, the time actually spent indexing a document is recorded. This yields a biased estimate for determining the total indexer time for budgeting purposes. However, the amount of bias can be measured by comparing indexing time when indexers are aware of being observed and when indexers are not aware of being observed. If the indexing process is easily identified, among other functions performed by the indexers, an alternative method is to spot check the proportion of indexers that are actually indexing at random time periods during the day. Proper sample planning and analysis can yield sufficiently precise estimates for this purpose.

VI. RÉSUMÉ

Statistics are used in work and time studies for description, for analysis, and for prediction. In the first instance, a gross amount of data is reduced to a few meaningful descriptors such as mean and variance. By analysis is meant that statistics provide an objective means of interpreting statistical information once obtained. Furthermore, one has a means of analyzing sampled observations in view of reliability of sample estimates. This is important since one can control reliability of estimates as a function of sample size (or cost). Also, effect of change or stimuli introduced to a given situation can be predicted by statistical models. Models also provide a framework for determining what information to obtain in research.

One of the principal advantages in statistical experimental and sample designs is that one can utilize prior information to increase reliability of estimation or testing of hypotheses which in turn reduces experimental costs and time in some instances. Economy is also gained in that one can answer many questions from a single experiment or sample. For example, analysis of an experiment yields a description in terms of mean and variance as well as determining the effect of various factors on these population estimates. Finally, proper statistical design ensures valid results and forces one to consider beforehand what information shall be found and how this information will be utilized once obtained.

LITERATURE CITED

(1) Bryant, E. C., "Evaluation of Information Retrieval Systems in Patent Office Environments. Part I. Statistical Concepts," Westat Research Analysts, Inc., PB 168,000, U. S. Department of Commerce, U. S. Patent Office, Washington, D. C., Feb. 1965.
(2) King, D. W., *J. Chem. Doc.*, **5**, 96 (1965).
(3) Cochran, W. G., "Sampling Techniques," John Wiley and Sons, Inc., New York, N. Y., 1953.
(4) Cochran, W. G., Cox, G. M., "Experimental Designs," John Wiley and Sons, Inc., New York, N. Y., 1957.