RANDOMIZATION AND SAMPLE SIZE IN EXPERIMENTATION*

Walter T. Federer, Cornell University

236
BU-~~227~~-M

September, 1966

Abstract

This is an expository paper on sample size in experimentation with particular reference to animal experiments. Randomization in animal experiments was discussed first in connection with a consideration of a disturbance of the randomization procedure caused by making the means of all samples equal before starting an experiment. A general discussion of number of animals per cage or pen was then presented. This was followed by a discussion of number of samples allocated to the control relative to other treatments for one statistical criterion. Sample size computations were given for interval estimation using a comparisonwise and an experimentwise error rate base. The required sample size for a two-stage procedure resulting in a confidence interval of a specified length was also presented. The last section dealt with experiments for which the sample size is fixed. A number of situations were presented with a somewhat detailed discussion of one of them.

* Lecture presented at the Food and Drug Administration Statistics Seminar, Washington, D. C., September 19, 1966.

RANDOMIZATION AND SAMPLE SIZE IN EXPERIMENTATION*

BU-~~227~~236-M Walter T. Federer, Cornell University September, 1966

1. Introduction

In line with some suggestions by Jacob N. Eisen, past Chairman of your F.D.A. Statistics Seminar Committee, I shall talk on a number of topics related to experimentation with emphasis on animal experiments. In particular, the following five topics will be discussed:

i) Randomization in animal experiments.

ii) Number of animals per cage or pen.

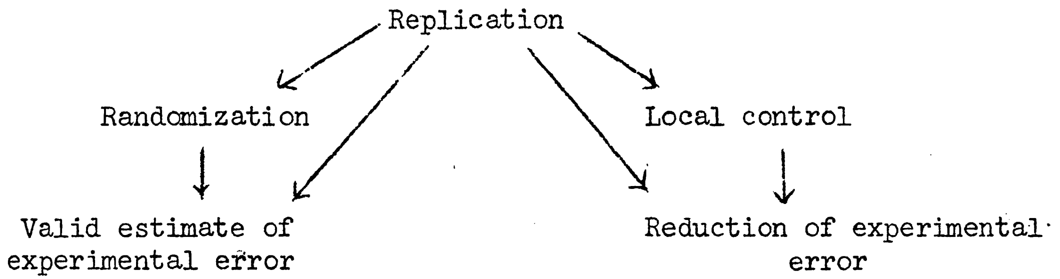
iii) Relative number of experimental units on a control and on a treatment.

iv) Sample size computations for interval estimation.

v) Fixed sample size.

2. Randomization in Animal Experiments

Sir Ronald A. Fisher is said to have had the following diagram, enunciating the three principles of experimental design, hanging on the wall of his office at the Rothamsted Experimental Station:


Replication
Randomization
Local control
Valid estimate of experimental error
Reduction of experimental error

* Lecture presented at the Food and Drug Administration Statistics Seminar, Washington, D. C., September 19, 1966.

He stated that design and analysis are two aspects of the same thing; randomization and replication are necessary to obtain a valid estimate of the error variance of a contrast. If the error variance of a contrast, e.g. the difference between two treatment means, contains all sources of variation inherent in the variation among experimental treatments or entities except that portion due specifically to the treatments themselves, then the error variance is said to be valid. (Fisher, R. A., The Design of Experiments, Section 65, 5th edition, Oliver and Boyd, London, 1949.)

Let us illustrate the effect of a disturbance of the randomization procedure as described in Chapter I, Example I.1, of my text (Experimental Design, Macmillan, New York, 1955). Suppose that fifty pigs are available for experimentation and that their initial weights (randomly selected weights from Table 3.21 of Snedecor, G. W., Statistical Methods, 5th edition, Iowa State University Press, Ames, Iowa, 1956) for 10 random samples of 5 pigs each are as given in Table 1. An analysis of variance on these data result in the one given in the bottom part of Table 1. Since there are no sample differences other than random sampling fluctuations, both mean squares are estimates of the same parameter σ^2.

Suppose now that 10 treatments had been applied and that three of the treatments had a -6 effect, three had a +6 effect, and four had a zero effect. Then, the sum of squares among sample means would be 661.2 + 5(6)(36) = 1741.2, and the mean square would be 1741.2/9 = 193.5. The resulting F statistic would be 193.5/82.87 = 2.33 which is approximately equal to the corresponding tabulated F value at the 3 percent level.

Now, let us disturb the random feature of the experiment and use a "balanced" grouping as utilized in some animal and educational experimentation. The rearrangement to obtain 10 samples of 5 each is made in such a way as to make all

Table 1. Data and analysis of variance on pig weights for 10 randomly selected samples of 5 weights each.

	Sample number									
	1	2	3	4	5	6	7	8	9	10
	30	19	16	17	47	17	41	20	38	42
	29	42	41	30	33	23	26	28	20	47
	39	27	37	24	17	31	19	39	30	41
	17	25	31	28	33	39	32	43	46	31
	12	22	25	35	29	30	27	30	36	29
Mean	25.4	27.0	30.0	26.8	31.8	28.0	29.0	32.0	34.0	38.0

Analysis of variance

Source of variation	d.f.	Sum of squares	Mean square
Among sample means	9	661.2	73.5
Within samples	40	3314.8	82.87

Table 2. Data and analysis of variance on pig weights for 10 balanced samples of 5 weights each.

	Sample number									
	1	2	3	4	5	6	7	8	9	10
	30	19	16	17	47	17	43	20	24	42
	29	42	41	30	25	23	30	28	20	12
	39	27	37	41	17	31	19	39	30	22
	17	25	31	28	33	39	32	33	47	46
	36	38	26	35	29	41	27	31	30	29
Mean	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2

Analysis of variance

Source of variation	d.f.	Sum of squares	Mean square
Among sample means	9	0	0
Within samples	40	3976	99.4

sample means equal. Some such grouping as given in Table 2 might result. The resulting analysis of variance on these "balanced" samples is given in the bottom part of Table 2. Since all means are equal there is no variation among sample means and hence the sum of squares for this category is zero. If this method of balancing were used and if the treatment effects were as above then the among treatment means sum of squares would be $0 + 5(6)(36) = 1080$ with the resulting mean square being 120.0. The F statistic would be $120.0/99.4 = 1.21$ which is considerably lower than the previous F statistic = 2.33. In order to achieve the same F value in the "balanced" array as for the random array almost twice as many animals would have been required, i.e. $2.33 = \frac{n(216/9)}{99.4}$ or $n = \frac{2.33(99.4)(9)}{216} = 9.65$.

The above computations have been performed assuming a within-treatment correlation of unity between initial and final weights in an experiment. If the within-treatment correlation of initial and final weights is zero, "balancing" has no effect and would not affect the validity of the error variance in the analysis of variance. The real-life situation, however, would usually have a correlation between these two extremes. Of course, none of us would let ourselves get into the above situation, but let us suppose that one of our scientific friends used "balancing" and then wanted to compute an error variance and thus salvage this information from his data. Can this be done? The answer is yes. We could use the technique of covariance of final weight on initial weight to eliminate the effect of "balancing" and then obtain valid tests of significance or valid interval estimates (see Yates, F. J., Roy. Stat. Soc. 109: 12-43, 1946).

To sum up, if we utilize a statistical technique requiring randomization, we must either use randomization or devise a scheme which removes the effect of non-randomness from the data.

3. Number of Animals per Cage or Pen

There are many statistical criteria for determining optimum sample size. However, prior to considering any of these, the practical and biological conditions of the experiment must be considered first. If the cages or pens are already available they will have certain limitations on the number of animals that can be accommodated in a cage. Too many animals in a pen or cage may result in the animals contacting a disease or may result in severe competition for food, light, and space. This would result in an additional component of variation within cages or pens which would not enter into the variation among pens. Competition between individuals within a pen or cage is not necessarily bad, but one must know how to handle it. At first thought one could "eliminate" competition by using single animal cages. However, this could induce another kind of competition, e.g. lonesomeness for pigs, antagonisms for pheasants, etc. If the results of an experiment are to have application value, the experiment must be such that the results are applicable to a real-life situation. One way of achieving this is to have the experimental conditions conform with those from real life.

If the cage size can be varied at will and if this has no effect on the response of the animals being used, then we are in a position to invoke statistical considerations. As a general rule, for a fixed total number of animals it is best to use as many cages as possible, resulting in few animals or one animal per cage. The variation among cages is generally greater than within cages and hence the above generalization. If the cost of one animal per cage is $c_a + c_w$ and if the cost of two animals per cage is $c_a + 2c_w$, of three animals per cage is $c_a + 3c_w$, etc., then the optimum (in the sense of minimizing cost for a fixed variance or vice versa) number of animals per cage is given by the formula,

$$k = \sqrt{c_a\sigma_w^2/c_w\sigma_a^2} \quad ,$$

where σ_w^2 is the variance (either an estimate or the actual value) among individuals within cages and σ_a^2 is the variance component among cages or pens. This is the formula well-known to sample surveyors for determining the optimum number of sampling units per stratum. The formula is useful in a variety of contexts and situations, e.g. the optimum number of determinations on a sample relative to the number of samples, the optimum number of readings per determination, etc. In some cases k may turn out to be a fraction, in which instance samples may be bulked prior to making a determination. In other cases the minimum number for k will be one.

Another criterion for determining sample size is a minimax procedure with a specified percentage of selecting a correct hypothesis, the correct binomial ratio, or the correct ranking of treatments. P. NaNagara (M.S. Thesis, Cornell University, 1953) has provided the procedure for binomial ratios and R. Bechhofer (Annals of Mathematical Statistics 25:16-39, 1954) has provided the procedure for ranking means from a normal population.

Many other criteria are available for determining sample size. Three of these will be used in the following sections. One criterion which has no statistical justification is one which might be called a crystal-ball procedure. This involves a purely arbitrary choice of sample size for such various and sundry reasons as:

i) I have always used 3 animals per cage.

ii) My major professor used 5 animals per cage.

iii) A Nobel prize-winner used 4 animals per cage.

iv) A statistician suggests a total sample size of 20 because he knows that the experimenter has only been using one or two.

v) A statistician thought I should use 6 animals per cage.

vi) I like the number 7 so I'll use a total of 7 animals per treatment with one in each cage.

vii) etc.

This procedure brings to mind an amusing incident described by Wilson, E. B., Jr., An Introduction to Scientific Research, McGraw-Hill, New York, 1952, page 46. It appears that chickens were subjected to a specified treatment; it was reported that $33\frac{1}{3}\%$ of the chickens recovered, that $33\frac{1}{3}\%$ died, and that no conclusions could be drawn about the other $33\frac{1}{3}\%$ since that one ran away!

4. Relative Number of Experimental Units on the Control Treatment

Another criterion for determining sample size is that the standard error of a difference between two treatments be a minimum. Let us suppose that we wish to compare a set of v treatments with a control or standard in our experiment and that we do not wish, primarily, to compare the v treatments among themselves. If each treatment is to have r replicates or samples, if the control is to have $r\theta$ replicates, and if $N = r(v+\theta)$ = total number of samples, then the quantity to be minimized is:

$$\sigma\sqrt{\frac{1}{r}+\frac{1}{r\theta}} = \sigma\sqrt{\frac{v+\theta}{N}+\frac{v+\theta}{\theta N}} = \sigma\sqrt{\frac{(v+\theta)}{N}\left(\frac{\theta+1}{\theta}\right)} \ .$$

Minimization results in

$$\theta^2 = v \quad \text{or} \quad \theta = \sqrt{v} \ .$$

For example, suppose that we are interested in comparing four treatments A,B,C, and D with a control or brand X. Then the relative number of replicates or samples on control X should be $\theta = \sqrt{4} = 2$. If $r = 20$ samples for each treatment, then $r\theta = 40$ should be the sample size for the control in order to minimize the standard error of the mean difference between the control and a treatment. The standard error for the above optimum procedure would be

$$\sigma\sqrt{\frac{1}{20} + \frac{1}{40} = \frac{3}{40} = \frac{9}{120}}$$

whereas for equal replication of the control on $N = 4(20) + 40 = 120$ samples, r would be 24 and the standard error would be

$$\sigma\sqrt{\frac{1}{24} + \frac{1}{24} = \frac{10}{120}} \ .$$

Although the difference in the two standard errors isn't world-shaking it does illustrate that equal allocation can be improved upon in certain situations.

5. Sample Size Computations for Interval Estimation

A criterion useful here is to select sample size r such that one has a specified probability, say $1 - \gamma$, that a $1 - \alpha$ percent confidence interval will be less than or equal to a specified length, say 2δ. In other words, we pick a sample size such that $1 - \gamma$ percent of all confidence intervals are less than or equal to 2δ and $\gamma\%$ are greater than 2δ in length. The experimenter is at liberty to choose the error rate base, the confidence coefficient = $1 - \alpha$, the assurance coefficient = $1 - \gamma$, and δ = the difference of interest and importance to him given a specified experimental design and the error variance σ^2. If the value of σ^2 is not known then one may use a previous estimate of σ^2 which has f_1 degrees of freedom. Using the appropriate procedure and the resulting sample size, the experimenter will have a $1 - \gamma$ percent assurance that the $1 - \alpha$ percent confidence interval will be less than or equal to 2δ.

Let us determine the required sample size for two error rate bases, i.e. comparisonwise error rate and experimentwise error rate and for the case where the random errors are normally distributed with mean zero and common variance σ^2. Let us suppose that we plan to use a randomized complete block design with r replicates, that $v = 3$ treatments are to be used, that $\alpha = 10\%$, that $\gamma = 25\%$, that $\delta = 5$, that $s_1^2 = 25$ with $f_1 = 120$ degrees of freedom, that $f_2 = (r-1)(3-1) = 2(r-1)$, and that a comparisonwise error rate is to be used. Then, r may be computed from the following formula:

$$r = 2\left(\frac{s_1}{\delta}\right)^2 t^2_{\alpha,f_2}\, F_\gamma(f_2,f_1)$$

$$= 2\left(\frac{s_1=5}{\delta=5}\right)^2 t^2_{.10,f_2}\, F_{.25}(f_2,120)$$

where t_{α,f_2} is Student's t and F is Snedecor's variance ratio with f_2 degrees in the numerator and f_1 degrees of freedom in the denominator. In order to solve for r we need to select a value of $f_2 = 2(r-1)$. Suppose we try $r = 7$ and then $f_2 = 12$. Substituting in the above formula we obtain

$$2\left(\frac{5}{5}\right)^2 (t_{.10,12} = 1.78)^2(1.26) = 8 \quad .$$

Since 7 was too small let us try r = 8; then

$$2\left(\frac{5}{5}\right) (t_{.10,14} = 1.76)^2(1.24) = 7.7 \quad .$$

Since r = 7 is too small and r = 8 is too large, we would use r = 8 replicates to obtain a 90% confidence interval which would be less than or equal to 10 = 2δ in 75% of all experiments conducted.

Now suppose that instead of a comparisonwise error rate we wish to use an experimentwise error rate of α = 10% for the above experimental situation. What would r have to be? Now r may be obtained from an iterative solution of the following formula:

$$r = \left(\frac{s_1}{\delta}\right)^2 q^2_{\alpha,v,f_2} \, F_\gamma(f_2,f_1)$$

$$= \left(\frac{s_1=5}{\delta=5}\right)^2 q^2_{.10,3,f_2} \, F_{.25}(f_2,120) \quad ,$$

where q_{α,v,f_2} is the tabulated value of the studentized range for v treatments at the α percent level. H. L. Harter et al. have published extensive tables of the studentized range. (E.g. see WADC Technical Report 58-484, volume II, October, 1959, Wright Air Development Center and Annals of Mathematical Statistics 31:1122-1147, 1960.) As before one would need to substitute an arbitrary value for r and then to use $f_2 = 2(r-1)$. Both r = 11 and r = 12 result in values between 11 and 12. For the latter value, $\left(\frac{5}{5}\right)^2(3.06)^2(1.22) = 11.4$.

Hence, one would use $r = 12$ replicates to have a 75% assurance that the 90% confidence interval would be less than or equal to $2\delta = 10$ = two standard deviation units.

Another procedure for obtaining a confidence interval of a specified length has been put forth by Charles Stein (Annals of Mathematical Statistics 16:243-258, 1945). The procedure gives a 100% assurance that the confidence interval will be less than or equal to the specified length and it involves sampling in two stages. Suppose that V is the variance of the linear contrast which will yield a confidence interval of the specified length. For the first stage we decide to take n_1 observations such that n_1 is large enough so that terms of $1/n_1^2$ are small relative to $1/n_1$. An estimate, s_1^2, of the error variance of a single observation is computed from the n_1 observations obtained from the first stage of sampling. The additional number of observations required in the second stage of sampling is computed from the formula:

$$n = \text{maximum} \left\{\frac{s_1^2}{V} + 1,\ n_1 + 1\right\}$$

where n is the smallest integer part of the above maximum. Then, if $\frac{s_1^2}{V} + 1$ is the maximum, $V = s_1^2/(n_1+n_2)$. The value of $\sqrt{V}$ then is substituted in the formula for computing the confidence interval for the contrast μ, say, thus $\mu \pm t_{\alpha,n_1-1}\sqrt{V}$; this interval will always be approximately equal to the specified interval.

F. A. Graybill and T. L. Connell (Annals of Mathematical Statistics 35: 438-440, 1964) present a two-stage procedure for estimating the population variance such that the estimate will be within d units of the true value. This reference illustrates another of the many variations of estimating sample size to achieve a desired goal. The reader is referred to Cochran, W. G., Sampling Techniques, 2nd edition, Wiley, New York, 1963, Chapter 4, for further discussion.

6. Fixed Sample Size

In many situations sample size is fixed by the resources and/or number of sampling units available for an experiment. This fact has been virtually ignored by statisticians, probably feeling that the problem didn't exist or, if it did, that it would go away if they ignored it. Unfortunately, it does exist and it won't go away! Let us get out of the statistical rut of considering that the determination of the sample size r is the item of sole importance and look at the more general problem. Suppose that the experimental design is specified and consider the following class of problems.

Quantities specified	Quantity to be determined
α, γ, σ, δ (or δ/σ)	r
α, γ, σ, r	δ
α, σ, δ (or δ/σ), r	γ
γ, σ, δ (or δ/σ), r	α
α, γ, δ, r	σ
δ/σ, $(1-\alpha)/(1-\gamma) = k$, r	α,γ
etc.	

where the above quantities are as defined previously. All the above except the first consider sample size fixed. Did it ever occur to you to consider the significance level or size of the test as a random variable? The idea that the first problem above was the only one of importance in real-life situations appears to be a figment of the statistician's imagination!

The last problem listed above has been presented to my classes in experimental design for several years. Let me illustrate the procedure for the fixed sample size case where the confidence coefficient relative to the assurance coefficient is a specified value, i.e. $(1\_\alpha)/(1-\gamma) = k$ a constant. Then $\gamma = 1 - (1-\alpha)/k$ and is a function of α. Hence, if we determine α we then have determined γ. Suppose that we use a comparisonwise error rate, then the α satisfying the following formula is one to be used in an experiment:

$$r(\delta/s_1)^2/2 = t^2_{\alpha,f_2}\ F_\gamma(f_2,f_1)$$

For example, let $r = 6$, $v = 5$, $f_2 = 20$, let $s_1^2 = 25$ be estimated from a previous experiment with $f_1 = 60$ degrees of freedom, let $k = (1-\alpha)/(1-\gamma) = 1$ and therefore $\alpha = \gamma$, and let $\delta = s_1 = 5$ units.* Then α is selected to satisfy the following equation which is solved iteratively for α:

$$\frac{6}{2}\left(\frac{5}{5}\right)^2 = 3 = t^2_{\alpha,20}\ F_\gamma(20,60) \quad .$$

First try $\alpha = .10$;

$$(2.97)(1.54) = 4.6 \quad .$$

Next try $\alpha = .20$;

$$(1.76)(1.30) = 2.3 \quad .$$

Next try $\alpha = .15$;

$$(2.5)(1.5) = 3.8 \quad .$$

Next try $\alpha = .17$;

$$1.9(1.5) = 2.85 \quad .$$

* A randomized complete block design is assumed.

(The above values were read from graphs prepared by L. E. Vogler and K. A. Norton, NBS Report 5069, National Bureau of Standards, Boulder, Colorado, 1957.) We would use $\alpha = 17\% = \gamma$ in the above experiment. This might bother the conventional "5-percenters", but it really shouldn't because there is really no halo around the 5% point!

Similar computations are involved using other error rate bases such as experimentwise, per experiment, etc.