

# Design of investigations

## Skeleton notes

D.R.Cox

October 2005

### GENERAL REFERENCES

Breslow, N. E. and Day, N.E.(1980). *Statistical methods in cancer research, vol. 1, The analysis of case-control studies*. Lyon: IARC.

Breslow, N. E. and Day, N.E.(1983). *Statistical methods in cancer research, vol. 2, The analysis of cohort studies*. Lyon: IARC.

Cox, D.R. (1958). *Planning of experiments*. New York: Wiley.

Cox, D. R. and Reid, N. (2000). *The theory of the design of experiments*. London: Chapman and Hall.

Thompson, M. E. (1997). *The theory of sample surveys*. London: Chapman and Hall.

Thompson, S.K. (1992). *Sampling*. New York: Wiley.

### SPECIAL REFERENCES

Baddeley, A.J. and Jensen, E.B. (2004). *Stereology for statisticians*. London: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1979). On sampling and the estimation of rare errors. *Biometrika* **66**, 125-132.

Darby, S.C., McGale, P. and Peto, R. (2003). Mortality rates from cardiovascular disease more than 10 years after radiation for breast cancer: nationwide cohort study of 90, 000 Swedish women. *British Medical Journal* **326**, 256-257.

Doll, R. (2002). Proof of causality. *Perspectives in biology and medicine*. **45**, 499-515.

Doll, R. and Hill, A.B. (1950). Smoking and carcinoma of the lung. Preliminary report. *British Medical Journal* **265**, 739-748.

Farrington, D.P. (2003). British randomized experiments on crime and justice. *Annals of American Academy of Political and Social Science* **589**, 150-167.

Park, A. et al. (editors) (2001). *British Social Attitudes, 18th report..* London: Sage.

Perry, J.E. et al (2003). Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. *Journal of Applied Ecology* **40**, 17-31.

Preece, A.W. et al (1999). Effect of a 915 mHZ simulated mobile phone signal on cognitive function in man. *International Journal of Radiation Biology* **75**, 447-458.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403-411.

Vandenbroucke, J.P. et al (1994). Increased risk of venous thrombosis in oral-contraceptive users who are carriers of factor V Leiden mutation. *Lancet* **344**, 1453-1457.

# 1 Introduction

While the majority of the literature on statistics is concerned with methods for the analysis of data, the design of investigations raises crucial issues. Good design may make analysis simple and convincing; bad design can rarely be overcome by elaboration of analysis.

The role of the statistician in design is, in close to discussion with subject-matter workers, to achieve clarity of interpretation by applying some key principles, taking account of the constraints on investigation, physical, cost-induced or ethical, that are always present.

While most investigations are part of a sequence it is convenient to start with a research question (not necessarily a hypothesis). This may be purely factual; what is the distribution of household income at this time, how many bicycles are there in Oxford today, how many whales are there in a certain area of ocean, and so on. More commonly, however, they concern some notion of dependence. What is the effect of such and such a policy on the distribution of income, do patients having this medication have a better prognosis than patients having some other medication, what is the effect of a GM-crop on various ecological properties of the area as compared with an analogous crop developed out of other plant-breeding programmes, and so on.

Reasonable criteria for any investigation include

- sensible choice of material for study
- sensible choice of features to be measured
- avoidance of systematic error in key comparisons
- reasonable control of random error by choice of design and size of study

- allowing comparison with previous results or relevant theory whenever possible

## 2 Types of investigation

Methods that may be used are

- secondary analysis of already existing data
- cross-sectional study, including descriptive sample survey
- prospective observational study (cohort study)
- retrospective observational study, often a case-control study
- an experiment

There is a quite sharp distinction between an observational study and an experiment. In the latter the investigator has virtual control over the whole system, whereas in the former the investigator's choices are limited to deciding whom or what to observe and what to measure. The types of investigation can often be combined.

*Illustration.* HRT and cardiovascular disease

*Illustration.* Wild life, farm management and bovine TB

*Illustration.* Use of mobile phones in cars and traffic accidents.

Design issues arise in all these and in addition in

- the development of measuring instruments
- the construction of data bases

These last two will not be discussed here.

### 3 Secondary analysis

This aspect will not be discussed in detail. A key aspect here is to know the field well enough to be aware of relevant data. In the social sciences this is aided by the ESRC Data Archive at University of Essex. In the Nordic countries there are various registries concerning population health that are particularly complete and sometimes of long origin.

*Illustration.* Darby et al (2003) investigated the issue of whether radiation treatment of breast cancer enhanced the chance of subsequent cardio-vascular failure.

*Illustration.* The 90-city study of air pollution and total mortality.

## 4 Principles of sampling

### 4.1 Formulation

We suppose that interest focuses on the properties of an existing (i.e. not hypothetical) population of individuals, called units. Let  $Y$  be a property defined for each unit and suppose interest lies in the population mean of  $Y$  which could in principle be measured exactly by complete enumeration.

The first steps are to define clearly the population in question and the notion of a unit.

*Illustration.* In a household survey is a unit a household, a head of household, all members of the household, etc? What constitutes a household?

This leads to the crucial notion of a *sampling frame*, in effect a list of all members of the population. Without a frame reliable sampling is impossible. We suppose for simplicity that the frame is discrete and finite and so can be listed  $1, \dots, N$ , where  $N$  is the finite population size. For each member there is defined a variable  $Y$  and the object is to estimate the population

mean  $\mu = \Sigma Y_j / N$ . A sample  $s$  consists of a subset of the population and the corresponding values  $y_1, \dots, y_n$ , say, and a sampling design  $p(s)$  is a probability distribution defined over the set of possible samples usually considered without regard to the order of selection.

The design problem is to choose  $p(s)$  to achieve the objectives outlined earlier. Freedom from bias in particular requires impersonality of choice.

*Illustration.* Sampling mineral material from a conveyor belt.

*Illustration.* Quota sampling does not satisfy this condition.

Some reduction in random error can be achieved by design; choice of sample size depends on the level of precision required.

## 4.2 Some further ideas

Populations are frequently divided into *strata*, expected to be relatively homogeneous.

*Illustrations.* Male, Female; Urban, Rural.; By Counties; By geographical position across a field.

Populations may be divided into *domains of study*. These are essentially strata of intrinsic interest, i.e. strata such that it may be required to compare at some point or possibly analyse separately.

Populations may be divided into primary units and subunits.

*Illustration.* Households might be a primary unit, individuals in the household subunits. Another example with more levels would be educational districts, schools, classes within schools and individual pupils within classes.

There may be available on every member of the population a vector of variables  $z_1, \dots, z_N$  which can be used for design and analysis.

*Illustration.* In sampling a geographical area modern Geographical Information Systems (GIS) give much information in quite fine detail.

### 4.3 Simple random sampling

A basic scheme, rarely directly employed but a key to more complicated procedures, is to fix a sample size  $n$  and then give all distinct samples of size  $n$  equal probability; this is called *random sampling without replacement*.

Let  $\bar{y} = \Sigma y_t/n$  denote the sample mean. Then with  $E$  denoting an expected value over the sample selection distribution  $p(s)$ , we have

$$E(\bar{y}) = \bar{Y}.$$

To see this note that the left-hand side is a symmetric function of the population values of degree one and hence has the form  $A_{nN}\Sigma Y_j$ , where  $A_{nN}$  is a constant. Consideration of the special population with all values equal leads to the result.

Similarly

$$\text{var}(\bar{y}) = B_{nN}\Sigma(Y_j - \bar{Y})^2,$$

where  $B_{nN}$  is a constant.

Consideration of a special case shows that in fact

$$\text{var}(\bar{y}) = \frac{\Sigma(Y_j - \bar{Y})^2}{(N-1)n}(1-f) = \sigma^2(1-f)/n,$$

say, where  $f = n/N$  is the finite population correction and can often be ignored, and  $\sigma^2$  is a measure of population variability, strictly not quite a variance.

At least for reasonably large  $n$  confidence limits for the population mean follow in the ordinary way.

### 4.4 Stratified random sampling

Suppose that the population is divided into  $k$  strata of sizes  $N_1, \dots, N_k$ . Take a random sample without replacement of size  $n_l$  from stratum  $l$  and estimate

the overall population mean by  $\Sigma N_l \bar{y}_l / N$ , which is clearly unbiased. This estimate has variance

$$\Sigma N_l^2 \sigma_l^2 (1 - f_l) / (N^2 n_l).$$

. If we ignore the finite population correction and treat sample size as a continuous variable, we may now minimize variance for a given total sample size, or minimize sample size for a required variance, and if necessary allow the cost per observation to be different in different strata. There results the optimal allocation formula that when the cost per unit is constant  $n_l$  should be proportional to  $N_l \sigma_l$ .

## 4.5 More general treatment

To specify general sampling schemes it is often helpful to consider the first, second,... order inclusion probabilities,  $\pi_j$ , the probability that population unit  $j$  is in the sample,  $\pi_{j_1 j_2}$ , the probability that both units  $j_1$  and  $j_2$  are included, etc. For fixed sample size  $n$  we have  $\Sigma \pi_j = n$ . Then an unbiased estimate of the population mean is the Horvitz-Thompson estimate  $n^{-1} \Sigma y_j / \pi_{\tilde{j}}$ , where  $\tilde{j}$  is the population unit corresponding to the  $j$  th sample value.

An important special case is probability proportional to size (pps) sampling.

Multistage sampling is dealt with by specifying schemes for selecting primary or first-stage units and then for selecting second,... stage units within the chosen first stage units.

*Illustration.* In the British Social Attitudes survey (Park et al, 2001) a sample of adults aged 18 yr or over was obtained. The sampling frame was the Postal Address File, a list of addressed kept by the Post Office. First postcode sectors were chosen from a list of all sectors. The sectors were

stratified into 37 subregions, and by population density and by the proportion of homes that were owner-occupied. 200 sectors were chosen with probability proportional to the number of households. 31 addresses were then chosen in each of the 200 sectors by systematic sampling of the list of addresses in each sector. At each address eligible individuals were identified and one chosen by a computer-generated random procedure. There were 3 questionnaires, one for all the sample, one for a random two-thirds and one for a random one-third. (Some complications have been omitted in this account.)

If a variable  $z$ , say is available for all members of the population it can be used sometimes assuming proportionality or sometimes a regression relation to adjust  $\bar{y}$  into an improved estimate of  $\bar{Y}$ , in the simplest case by taking  $\bar{y}\bar{Z}/\bar{z}$ , where  $\bar{Z}, \bar{z}$  are respectively population and sample means of  $z$ .

## 4.6 Sample size

Choice of appropriate sample size involves a compromise between the costs of data collection and the dangers of poor conclusions arising, in particular, from large random errors. In principle the choice is an economic one but the cost and other data needed to make such calculations are rarely available. Instead cost is minimized subject to achieving a suitable standard error in the estimation of key parameters, or occasionally to achieve a nominated level of power in testing a relevant hypothesis. Such calculations although very approximate are very important to avoid spending effort on studies highly unlikely to yield interpretable conclusions or unnecessarily expensive. The parameters needed to estimate standard errors can themselves be estimated from pilot studies which are necessary anyway on general grounds if the proposed study enters new territory.

## 4.7 Population model approach

In the above discussion the values attached to study units in the population are *not* regarded as random variables but as fixed and initially unknown constants. Randomness enters only via the method of selection used. For example in simple random sampling done sequentially the first sample value is a random variable with a distribution attaching mass  $1/N$  to the individual population values. There is, however, an alternative approach in which the individual population values are modelled as random variables and probability calculations refer not to the sampling procedure but to the model employed to describe population values. By and large very similar results can be achieved by the two approaches although the assumptions made in the population model route often seem rather contrived.

Systematic sampling, corresponding to taking equally spaced values in the population, is sometimes used as a compromise between random and stratified sampling. It is obviously a bad idea if there is any possibility of a corresponding periodicity in  $Y$ . Estimates of precision are usually based on a time-series or spatial model of variability.

## 4.8 Special topics

Particular fields of application lead to specialized, and often very interesting, problems. Thus stereology deals with the estimation of the properties of systems in, say three dimensions, via data on lower dimensional probes (Baddeley, 1993). Auditing of accounts is sometimes done by an adaptation of probability proportional to size sampling called monetary unit sampling (Cox and Snell, 1979) but there are special problems of analysis.

## 4.9 Nonsampling errors

Especially in sampling human populations nonsampling errors may be serious. These come in particular from systematic reporting errors and from selective non-response. There are various techniques for mitigating the effect of the latter, especially follow-up surveys and the notion of treating the initial non-responders as a stratum to be intensively studied via a sub-sample. Two central questions in considering the results of such surveys are: what was the response rate and what was the sample size?

## 4.10 Key points

Appropriate definitions of variables to measure and of population-definition are crucial. Bias is typically avoided by randomization or occasionally by systematic sampling. Random error is hopefully reduced by stratification and by the use of supplementary variables. Suitable sample size always needs consideration.

# 5 Observational studies of dependence

## 5.1 Formulation

We now consider studies in which the primary purpose is to investigate how one or more variables (outcomes, responses) depend on explanatory features. In principle this is nearly always best achieved by experimentation rather than by observational studies but this is not always feasible, and in some fields virtually impossible. In a sense the object is usually to try and establish causality in one or both of the following senses:

- to understand the physical, biological or social paths that lead from the explanatory variables to the response

- to establish that changes in an explanatory variable lead, other things being equal, to systematic changes in response.

To establish this convincingly is often difficult, especially if the effect involved is relatively modest (Doll, 2002).

Data from cross-sectional studies, for example a sample survey at one time point may establish important dependencies, but rarely achieve the more challenging purpose, in particular because the direction of dependencies cannot be established from the data.

To study processes in time we may proceed prospectively or retrospectively.

*Illustration.* Bradford Hill and Doll's (1950) case-control and a large GP study both investigated smoking and lung cancer.

These illustrate the advantages and disadvantages of the two approaches. The basic desiderata remain those of Section 1.

## 5.2 Prospective study

In a prospective (forward-looking) study, often called a cohort study, we choose a group of individuals, the cohort, determine appropriate explanatory variables initially (at baseline), and then follow the individuals forward in time until a response of interest is observed. Provided there are no important unobserved explanatory variables at baseline study of the relation between the response and the baseline variables establishes which appear to be the most relevant explanatory variables.

In social science applications such studies are called *panel studies* and often have the additional feature that the individuals studied are intended to be representative of a population and thus in principle chosen by one of the sampling schemes of Section 2. Panel members may be questioned every

year, every two years, every election or at whatever frequency is relevant. Panel attrition is an important difficulty overcome to some extent by the technique of sampling with partial replacement of units.

### 5.3 Retrospective study

In the study of rare responses, and problems where the response takes a long time to be observed, prospective studies can be very expensive and, in a sense, inefficient. It is then common to use retrospective methods, starting with the response and looking backwards in time to find relevant explanatory variables. The term case-control study is often used or in econometrics choice-based sampling.

A collection of *cases*, typically individuals showing the rare response, is formed and a group of *controls*, comparable individuals not showing the response and on all individuals explanatory variables are recorded.

Consider a large population with a binary explanatory variable  $X$  and a binary response  $Y$ . Let  $\theta_{ij} = P(X = i, Y = j)$ . One way to describe dependence in the population is via the ratio  $(\theta_{11}\theta_{00})(\theta_{10}\theta_{01})^{-1}$ .

Then in a prospective study we may estimate  $P(Y = j | X = i)$ . One way to describe the effect of  $X$  is by the odds ratio, namely  $P(Y = 1 | X = 1)/P(Y = 0 | X = 1)$  divided by  $P(Y = 1 | X = 0)/P(Y = 0 | X = 0)$  and this is  $(\theta_{11}\theta_{00})(\theta_{10}\theta_{01})^{-1}$ .

Now consider a case-control study on the same population. From the cases, having  $Y = 1$ , we observe the corresponding values of  $X$  and thereby estimate  $P(X = j | Y = 1)$ . The effect of  $X$  can thus be described by the odds ratio, namely  $P(X = 1 | Y = 1)/P(X = 0 | Y = 1)$  divided by  $P(X = 1 | Y = 0)/P(X = 0 | Y = 0)$ . This is again  $(\theta_{11}\theta_{00})(\theta_{10}\theta_{01})^{-1}$ . That is, if we analyse a case-control study as if  $X$  is a response to  $Y$  as

an explanatory variable, then provided we use odds ratio as the parameter of interest we estimate the same quantity as in a prospective study or in a random sample of the whole population. More generally logistic regression is used; for a full justification see Prentice and Pyke (1979).

*Illustration.* Doll and Hill (1950) found strong evidence of a link between smoking and lung cancer from a case-control study. Recently a 50 yr report was given of the analysis of a prospective study of the same issues;

A key question concerns the choice of controls. In a medical context these may be community based, general practitioner based, or hospital based. Often there is one control for each case; since cases are often harder to obtain than controls up to three controls per case may be suitable. Controls may be matched individually to cases or only in some average sense. Selective recall biases are often a concern.

There are various elaborations of the idea of case-control sampling; in some contexts the notion of a case only study is appropriate.

## 6 Design of experiments

### 6.1 Formulation

In outline the specification for an experiment is a set of experimental units  $U_1, \dots, U_n$  and a set of treatments  $T_1, \dots, T_t$ . One treatment *chosen by the investigator* is applied to each unit and a response or outcome variable  $Y$  measured on each unit. In addition there are often intrinsic features  $z_1, \dots, z_n$ , often called *concomitant variables*, measured on each unit before randomization.

*Illustrations.* Agricultural field trials, industrial experiments, clinical trials, psychological experiments and many laboratory experiments in the nat-

ural sciences.

Absolutely key issues are choice of treatments, units and observations but in statistical discussions these are usually regarded as given. The objective is to compare the effect of the different treatments on response and to be confident that any differences found are indeed "caused" by the treatments and not by extraneous features. The formal definition of an experimental unit is that it is the finest division of the experimental material such that any two different units may receive different treatments.

*Illustration.* In an ophthalmological experiment two treatments may be compared. In one design a patient is an experimental unit, in another a single eye is.

*Illustration.* There are difficulties in using experiments to evaluate policy options in a social context but there is some such work, mostly in US, but for criminology in the UK, see Farrington (2003).

A useful formulation, especially for continuous response variables in which additive effects are likely, is to assume that were  $T_s$  to be applied to unit  $U_j$  the resulting response would be

$$\alpha_j + \tau_s,$$

independently of the allocation to other units. In fact for each  $j$  we are allowed to observe only one  $s$ , corresponding to the treatment actually applied to that unit. The other values are so-called counterfactuals; they might have been observed but in fact were not. Note that  $\tau_2 - \tau_1$ , say, is the difference between the observation obtained when  $T_2$  is applied to  $U_s$  and the observation that would have been obtained had  $T_1$  been used instead.

The design allocation problem is to choose how to determine which treatment is applied to each unit. If the allocation is done by an objective randomization scheme that ensures that each unit is equally likely to receive

each treatment an argument exactly parallel to that used in sampling theory shows that if  $\bar{Y}_s$  is the mean response on units actually receiving  $T_s$  then  $\bar{Y}_{s_2} - \bar{Y}_{s_1}$  is an unbiased estimate of  $\tau_{s_2} - \tau_{s_1}$  without further assumption. This argument can be pushed much further (Cox and Reid, 2000, pp 48- ).

The central issue is to aim that all points at which uncontrolled variability might arise systematic errors should be avoided. Critical to this is the elimination of personal choice by the investigator or others and randomization plays a key role in this.

*Illustration.* The role of randomization in clinical trials is primarily to achieve concealment at various stages of the study.

More technical aspects of randomization concern the estimation of standard errors and the validation of tests of significance.

## 6.2 Error control

To reduce the effect of uncontrolled variation we may

- use especially uniform units and high quality measurement devices, the former having some disadvantages
- use the principle of comparing like with like
- use information in the concomitant observations  $z$
- employ repeated use of the same material as distinct units

To compare like with like we arrange the experimental units in blocks, preferably with  $t$  units per block. Then we randomize treatments subject to each treatment occurring just once in each block. The simplest example is the matched pair design,  $t = 2$ . There are important implications for analysis. Sources of variability designed out must be removed in statistical

analysis if appropriate measures of uncertainty are to be obtained. This is typically done by including terms for blocks in the formal model for statistical analysis.

An extension of the idea controls for two sources of variability simultaneously, the key design being the Latin square in which the experiment is built up from  $t$  *timest* arrays in each of which each treatment occurs once in each row and once in each column.

A simple example when  $t = 4$  is

B	C	D	A
D	A	B	C
C	D	A	B
A	B	C	D

Randomization of at least rows and columns would normally be used. There are many elaborations, for example the insertion of a second alphabet orthogonal to the first.

*Illustration.* Preece et al (1999) used 36 subjects in 2 groups of 18 to compare the effect on cognitive function of no signal, sine wave signal, and signal simulating mobile phone message.

### 6.3 Factorial principle

The final broad principle of experimental design is that in suitable contexts it may be much more efficient to investigate several aspects simultaneously in one experiment rather than in a series of separate experiments. There are, however, several practical limitations to this idea.

*Illustration.* A traditional example is of a fertilizer experiment involving the three basic elements, N,P and K. Even if each is restricted to just two levels, an experiment in which all  $2^3$  combinations are investigated may be

much more informative than separate experiments on N, P and K.

*Illustration.* In a different context the factors might be the temperature at which a reaction is run, the type of catalyst and the concentrations of various reactants.

The separate treatments are called *factors* and the different forms *levels*. The levels may be identified qualitatively or by values of a quantitative variable. The simplest factorial experiment is the  $2^p$ , i.e.  $p$  factors each at two levels; it would often be arranged in a number of randomized blocks each of  $2^p$  experimental units.

Note that we can also treat formally as a factor features that describe the experimental units, e.g. gender. The interpretation is different, however. If it is required that the conclusions apply to a specified target population, for example all sufferers from a specific disease in a country, in principle the units should be chosen by an appropriate sampling scheme to represent the population; in practice this is very rarely if ever possible. The ability to extrapolate is then based on some mixture of subject-matter knowledge and observed stability of the conclusions.

The analysis of such designs typically, although not inevitably, is by examining main effects, two-factor interactions, three-factor interactions, . . . and hoping that an interpretation can be based primarily on main effects and a limited number of interactions; note, however, that typically main effects are not a useful base for interpretation in the presence of strong interaction. Error assessment is based on an analysis of variance table which identifies the structure of the design and provides a way of estimating error.

For the  $2^p$  system with factors  $A_1, \dots, A_p$  it is convenient to denote the treatments, i.e. factor combinations, by  $1, a_1, a_2, a_1a_2, \dots, a_1a_2 \dots a_p$ , where for example  $a_1a_2$  means that  $A_1$  and  $A_2$  are at their upper levels and the

remaining factors at their lower level.

We take  $p = 3$  for simplicity of exposition, labelling the factors  $A, B, C$ . Then the main effect of  $A$  is defined by comparing the averages of the two sets of treatments

$$\begin{aligned} &1, b, c, bc; \\ &a, ab, ac, abc. \end{aligned}$$

Similarly the two-factor interaction  $B \times C$  is defined by comparing the averages of the two sets of treatments

$$\begin{aligned} &1, a, bc, abc; \\ &b, ab, c, ac. \end{aligned}$$

It compares, say, the effect of  $A$  with  $B$  at its upper level with the effect of  $A$  with  $B$  at its lower level. There is a detailed theory of such designs. *Illustration.* Designs of this sort are extensively used in some kinds of industrial experimentation and have considerable potential for investigating the sensitivity of elaborate computer model to the choice of input parameters.

There are many developments of these ideas. One hinges on the point that if  $p$ , the number of factors is not small even a single replicate of the  $2^p$  experiment will take a large number of units. The issue arises: can we investigate the  $2^p$  system in  $2^q$  units? There are, of course, similar questions when more than two levels are involved in some or all of the factors.

*Illustration.* Consider a  $2^4$  system with factors  $A, B, C, D$ . The four factor interaction divides the treatments into

$$\begin{aligned} &1, ab, ac, bc, ad, bd, cd, abcd \\ &a, b, c, abc, d, abd, acd, bcd \end{aligned}$$

If one were to observe only, say, the first 8 combinations, a 1/2 replicate of the  $2^4$ , main effects can be estimated separately but two-factor interactions are *aliased*.

For a  $3^3$  experiment with quantitative levels, the simplest reduced design that allows estimation of linear, quadratic and two-way interactive effects is the *central composite design*. This has observations at the points  $(\pm 1, \pm 1, \pm 1), (\pm a, 0, 0), (0, \pm a, 0), (0, 0, \pm a), (0, 0, 0)$ , i.e. 15 treatment combinations as compared with 27 in the full system. There are good arguments for some replication of the central point.

## SUMMARY

Avoidance of bias is achieved by randomization or sometimes by other impersonal device, precision is improved by comparing like with like and by concomitant variables and the factorial principle may lead to a much more informative study. The relative importance of these three aspects varies greatly between fields of application. Choice of units for study, treatments for investigation and observations for inclusion is crucial.

The overriding principles in all studies are that the research question is important and appropriately defined and then that the units of study and the nature of the observations are well chosen. After that it becomes a more technical issue in all modes of study to avoid serious bias, to get good error control and to have a suitable size of investigation. The ways of achieving these objectives are similar in different modes of study but subtly different in detail. The distinction between experiments and observational studies is crucial.

## Exercises

1. Prove that for samples of fixed size  $n$ ,  $\sum \pi_j = n$ . Write down the first and second order inclusion probabilities for simple random sampling with and without replacement from a population of size  $N$ .

2. A region, say a field, of irregular but convex shape is defined on a large-scale map. It is desired to estimate the number of distinct species in the area and their relative abundances. This may be done either by detailed study of a number of distinct small circular areas (quadrats) or by specifying line transects to be surveyed. What steps are desirable in each case to achieve reliable answers?

3. Suppose that  $f(x)$  is a function defined on  $(0, 1)$ . We can think of this as a continuous population with sampling frame the real numbers in  $(0, 1)$  the population value corresponding to a given  $x$  being  $f(x)$ . The population mean is

$$\int_0^1 f(x)dx.$$

A random sample of individuals  $x_1, \dots, x_n$  is chosen by taking independent and identically distributed values from the density function  $p(x)$ ; the corresponding population values  $f(x_i)$  are then observed. Show that the Horvitz-Thompson estimate of the population mean is

$$n^{-1} \sum f(x_i)/p(x_i).$$

Show that this is unbiased and calculate its variance. What guidance does this give over the choice of  $p(x)$ ?

[Note that in a different context this is called importance sampling.]

4. It is required to sample the population of current Oxford undergraduates to find paternal and maternal occupation (and possibly other features) and to study their relation to subject of the undergraduate specialization. How would you proceed?

5. In a country with a number of banks a central agency scores each bank on a five-point scale as to its solidity. To assess the effect of an advertising campaign about the existence and nature of the scale the following is proposed. From a sampling frame of account holders a sample is drawn before the campaign and those selected asked (among other questions) the name of their bank, whether they are aware of the score and, if they are aware, the current score of their bank. (The answer to the last question may be accurate or not.) Then the campaign takes place across the whole country and then on a new sample (which may contain some or all of the earlier sample) is selected and the questions repeated with the addition of a new question about awareness of the recent campaign. Give some more detail about how the samples should be chosen and some brief comments on how the data might be analysed.

6. In a randomized trial to compare two forms  $A, B$  of eye drop for the treatment of glaucoma (raised intraocular pressure) for each patient the left eye is treated with one form the right eye with the other. The drop in pressure is measured after 1 month of treatment. What assumptions are involved in such a design? What is a simple alternative design and what are its advantages and disadvantages? In the above design how do you recommend deciding for each patient which drop to use for each eye?

7. The following are artificial data for the (real) situation of the previous exercise. The rows correspond to 10 patients and the data are pressure drops

in mm Hg.

	<i>A</i>	<i>B</i>
1	5	7
2	0	3
3	-2	3
4	7	7
5	3	6
6	0	1
7	3	5
8	4	5
9	2	3
10	-1	4

An investigator reports that the estimated standard deviation for *A* is 2.8 and that for *B* is 2.6, a pooled value being 2.7. The data means are *A*: 2.1 ; *B*: 4.0. The standard error of the difference of two means of 10 observations each, being  $\sqrt{(2/10)}$  times the standard deviation, is about 1.2 and the difference between the means, namely 1.9, is less than 2 standard errors and thus insignificant. Explain *without using any statistical jargon* why this argument is wrong. How would you analyze these data?

8. What other data might have been used in the example of exercise 7? What would you have recommended for consideration had it been ethical to leave some eyes untreated?

9. Verify in the fractional factorial system given in the notes above that the interactions  $A \times B$  and  $C \times D$  are estimated by the same function of the data and hence are reasonably said to be aliased.

10. Develop a general definition of contrasts (main effects and interactions) in the  $2^p$  system by considering treatments in the form  $a_1^{i_1} \dots a_p^{i_p}$ , where the  $i$ 's take values 0, 1, and examining the linear form  $\alpha_1 i_1 + \dots + \alpha_p i_p \pmod 2$  as dividing the treatments into two sets. [More mathematical question.]