APPENDIX: Elementary Medical Statistics: An overview of content

In Chapter 1, Mainland outlines the purpose of his book. He indicates it is designed primarily for students, for whom the four chief objectives are addressed in Chapters 1 to 7, of 8. These are:

1. To help develop an ability to evaluate what is read in medical journals and heard at medical meetings.
2. To provide basic methods of observation, analysis and reasoning necessary in learning from clinical experience.
3. To give instruction in some of the simpler statistical techniques.
4. To provide some knowledge of other techniques and terms that are met more and more frequently in the medical literature.

The final chapter gives further help for investigators, especially for those early in their careers.

Comments on each chapter are given in the subsequent sub-sections, with topics of particular interest highlighted.

Chapter 1: The Place of Statistics in Medicine

In this first chapter, Mainland argues that many aspects of medicine, including diagnosis, are essentially quantitative. In spite of this, proper methods of handling quantitative data have been neglected and the reasons for the neglect relate to medical teaching, which is often about 'facts' or relate to differences which are so large as not to need statistical investigation. In addition, the link between methods in public health which are often statistical and clinical work is often not made. Mainland then praises the work of Gossett and particularly Fisher in establishing the principles of 'experimenter's statistics', both for analysis and design, and Snedecor in promoting them in the US. Notably on page 5, Mainland makes the observation that there is no absolute division between observation with experiment and observation without experiment.

Mainland then returns to medical attitudes to statistics, stressing the transformative aspect of using statistical methods in laboratory and clinical work but acknowledging the advances in medicine which can occur without statistical methods and the 'moral problem' some medical researchers feel about the term 'experimentation' in medicine. Mainland also addresses mathematics in statistics concluding "the validity of a statistical method is not to be found in mathematical proofs, but in the accumulated experience of statisticians and in special explorations devised by them".

The chapter finishes with a section on the purpose of the book as outlined previously and brief comments on the relationship of statistical methods to clinical practice which must take place even if sound statistical knowledge is not available. However, Mainland writes

“the general effect of statistics is to make an observer more critical and more aware when, for want of reliable information, he is acting on impression or opinion".

Chapter 2: On Looking at Evidence

This long 34-page chapter should be read by students, according to Mainland’s introduction to the chapter, quickly once and then again a second time when definitions can be learned. However, he argues, extensive memorisation of the material is not needed because the “ideas and methods are useless unless they can be applied, and ability to apply them can be achieved only by practice".

After a short section discussing the concept of variation, nine questions useful in looking at evidence are given: Who, Why (purpose of investigation), What, Where, When, How, How much, How many and Why (causal interpretation of results). These questions provide the starting point for a wide ranging discussion, with four primary examples of investigations for illustration.

Some of the issues discussed are:

- [Who?] The reputation of an investigator does not establish an authority that is better than their evidence.
- [Why?] Statistics is the “the science of arguing from samples". The populations from which the samples are taken is critical to arguments made from samples and the purpose of the sampling “should be clear, and should reveal an open-minded inquiry". The latter characteristic is because "a scientist sets up a hypothesis and then tries to knock it down, whereas the unscientific worker sets up a hypothesis and then tries to keep it up".
- [What?] Under this question, Mainland first argues that it is often important to classify samples by subdivision to reduce variation and/or bias. Failure to allow for such classification can make a sample “different from what it seems to be or purports to be". While the term is not used, much of this discussion relates to bias due to confounding. Mainland says, however, that while classes can be made narrower and narrower, “it is wise to stop the division at some point, which is determined largely by the use we wish to make of the information". After division, what remains is individual or ‘residual’ variation. A specific section deals with classification of vital statistics and introduces the term ‘rate’ which may be, for example, race-, age- or cause-specific.
- [Where?] In evaluating data from an investigation, Mainland argues that the ‘environment’ in which the work was done should be considered in its widest sense. The implication, unstated, is that this will influence the interpretation of any findings.
- [When?] This question relates to being aware of variation associated with time which might be periodic, such as seasonal, or non-periodic, such as age of disease classifications over time. Also mentioned is "The ‘Previous Series’ Method" which

essentially addresses the issue of historical controls which, Mainland says, “can never give indisputable conclusions". A final comment is made on allowance for ‘duration of exposure’.

- [How?] The importance of objectivity of measurement is stressed with reference to the MRC trial of streptomycin. Under this question, Mainland also addresses the distinction between two types of data, measurement and enumeration data. These are addressed separately by the next two questions.
- [How much?] In the discussion of measurement data, such as continuous data, under this question, Mainland defines errors of bias and random errors. He stresses that error does not necessarily suggest a fault but simply a “departure, difference or variation from some true value". After discussion of systematic errors leading to bias and the random errors that might be seen, for example, in repeat measurements, Mainland returns to the term ‘true value’ indicating the difficulties in defining ‘true’. He essentially argues that the definition which will allow investigators to know what is meant by the term is to regard true values as population averages that would arise from an ever increasing number of measurements of the same thing. He recommends that this term be used rather than true values. A final point is made that increased precision arises from replicate measurement of data to be compared in an experiment, for example comparing average adult heights in ten cities and ten rural districts.
- [How many?] In considering enumeration, or count data, Mainland focusses on two common faults. The first is spurious replication and the issue of identifying the number of independent observations or individuals in a study. The point made is that, in modern terms, clustering of data should be recognized in assessing “the true (or effective) size" of an experiment. The second fault is the use of incorrect numerators or denominators and a number of examples of this are given.
- [Why?] The second ‘why’ section addresses the issue of causal interpretation and this is discussed in the main text of this article.

This chapter continues with some summary remarks on the questions indicating that essentially all of them insist on the proper naming or labelling of things and that they can be used in criticism, investigation and even in activities such as writing patient case histories.

A final section outlines some principles of investigation recognising that the chapter might well prompt the question “How can we ever hope to draw any reliable conclusions?" In answer, Mainland recognises the need for some assumptions that cannot be tested in any investigation and also highlights that while experimental observations are best, nonexperimental observations are often all that is available, and also can be needed to provide the basis for properly planned experiments. He finishes with ‘Five Rules for Investigation’ which relate to consistency of observation, the definition of subpopulations and the use of ‘systematic sampling’ as required, random sample selection if the

investigation is experimental, the use of statistical tests to show how chance could account for any results and to not go beyond the evidence in linking association with causation.

This chapter ends with some questions which help the reader to think through the application of the material in a variety of contexts. Answers to these questions are given at the end of the book along with answers to questions found at the ends of various other chapters.

Chapter 3: Estimating the Error in Enumeration Data

This chapter begins, as do a number of others, with instructions to the 'student' on how the chapter might be read. The technical content of the chapter relates primarily to the determination of confidence intervals for observed binomial samples. Pages 7 to 10 of the chapter establish the concepts of a binomial classification, using a classification of success and failure for illustration, and a standard of confidence, using 95% and 99% for illustration and later tabulation, and then illustrate how to find a binomial confidence interval from tables for a sample of 20 with 2 failures. Pages 33 to 38 of the chapter then present further examples of binomial confidence intervals. These two sections are the material that Mainland recommends the student pay the most attention to first. However, Mainland then says that second and later readings are essential to go beyond these technical explanations and, indeed, to lay down the foundation material for all subsequent chapters.

The early pages of this chapter lay out the example of a surgeon with 18 successes in 20 operations. In examining this, Mainland stresses that some knowledge of error is essential to give the numbers meaning. He revisits errors of bias and random errors from the previous chapter and introduces the concept of chance by considering the innumerable factors that might affect each patient and which cannot be allowed for by systematic sampling or otherwise. He then defines an experiment with an easily recognised and pure chance outcome. This involves sampling from 1000 well mixed circular disks with 300 marked with "F' for failure and the rest with "S" for success. A sample of 20 disks is taken from a box containing the 1000 disks, the results are recorded and the 20 disks are returned to the box. This process can then be repeated. Mainland refers to the selection of the 20 disks as being due to chance which "is to be defined, not as something mysterious or metaphysical, but as the action of a multiplicity of independent causes". This example subsequently reappears throughout the book. The argument is made that with 1000 disks, this essentially represents sampling from an infinite population and the conception of sampling from an infinite population is typically relevant in medicine and adds that is "necessary for the making of certain tests and estimates". This again, it can be surmised, represents the influence of Fisher's views as outlined by Cox (2016). Some further discussion of and a definition of independent random samples is then given.

Between the material defining a binomial confidence interval and that giving examples, the material is, as Mainland suggests in his opening remarks, foundational. The first section of this material is titled "Statistical Significance" and this is where the concept of a significance test is introduced. The term significance level is not used but "The probability, P" is used instead, and the possible use of different levels of significance is discussed.

The second section addresses “Misconceptions Regarding Significance". Here Mainland points out that a significant results does not imply the certainty of a finding, that significance does not imply a difference is important, and that non-significance does not imply no difference. The section concludes with the suggestion that conventional significance levels of 5% and 1% have been found useful, that other levels can be used if the implications of their use is understood and that it must be remembered that rare events can occur by chance.

Mainland introduces the rest of the material leading up to the section of examples writing “The following sections are designed to show more fully the meaning of confidence limits and significance tests. …they introduce much that is basic in statistics: the testing of hypotheses; tabular and graphical presentation of data; examples of sampling variation; definitions (for example, probabilities, mode and skewness); the binomial distribution; and the normal curve". This material does indeed discuss “more fully" some topics already introduced and also introduces some more technical material, notably the binomial expansion and the normal curve. The latter is introduced for use as an approximation to the binomial but a brief section on the “Importance of the Normal Curve" indicates the "curve is of fundamental importance" and that it will be seen subsequently “how it underlies almost the whole treatment of measurement data".

Following the material on examples of binomial confidence intervals, this chapter concludes with a discussion of how the methods of the chapter can be used to illustrate the influence of sample sizes, the calculation of binomial confidence intervals when no observations of one type are observed and that small samples are particularly uninformative if no significant effects are observed. Brief mention is made of multinomial samples but no technical material is presented although a reference to such material in a later chapter is made.

Chapter 4: Comparison of Samples of Enumeration Data

In this chapter, Mainland again signposts for the student the order in which material is to be read. However, for purposes of summarisation, there seems no reason not to discuss the material in the order presented as, unlike Chapter 3, the sections signposted for a first reading are not just the details and examples of the procedure.

The chapter begins by setting out the problem of comparing success rates in two different surgical samples. The details of calculating a chi-square test statistic for a 2×2 (referred to as a fourfold) table and the determination of a significance level through comparison with some critical values from the chi-squared distribution. Following immediately is the section “On Planning a Simple Experiment" which has been discussed in the main text and deals with the use of systematic and random sampling and causal interpretations enabled by randomisation. There is then a brief discussion of the generalisation of findings and eight recommendations, for ‘beginners’, on how to design an investigation.

Intercurrent Events

There is a brief subsection in the section on experimental planning on what Mainland calls intercurrent events. This would appear to be the first use of this now more commonly used

term, often in the context of discussions of intention-to-treat (Chalmers et al. 2023a, 2023b), to refer to the possibility of events taking place after treatment which may influence the patient's outcome. The given examples of these events are treatment supplementation, change of treatment, accidents or diseases which may or may not be associated with the condition under study, the suspension of treatment for the patient's business or domestic affairs and loss of follow-up of a patient for a variety of possible causes including death. Mainland summarises what should be done as follows: "In deciding what should be done with data from any such patients the criterion must always be whether their inclusion or omission would introduce bias. Unless the appropriate decision is obvious, the best plan is to analyze all the data together, then to analyze the special cases and the main series separately". The inclusion of such a section in 1952 seems particularly remarkable.

The next section is on "Some Common Research Designs" and includes material on intrasubject comparisons beginning with the cross-over design but also discussing the options within such a design for subsequent changes in treatment in some cases. The usefulness of twin studies is also highlighted as is the potential of losing information when artificial pairing is introduced. Finally there is a very good discussion of the issues raised when treatment must be applied to groups of patients, *i.e,* cluster designs. This is followed by a section giving a set of examples of unsatisfactory sampling, stressing that the aim was not to criticise the investigators but to show how they were handicapped because knowledge of proper sampling methods had not been disseminated.

A section on "Implications of Levels of Significance" discusses errors of the first and second kind, what are now referred to as Type I and Type II errors. Interestingly, Mainland writes that "except in special cases …there is no way of telling how often errors of the second kind occur". While true for any specific study, this broad statement may again reflect Mainland's dependence of the writings of Fisher who largely deals with so-called "pure" significant tests and makes little of no mention of the "power" of tests which was a feature of the Neyman-Pearson approach to significance testing.

Following a brief discussion of some situations when the customary 5% and 1% levels of significance may not be appropriate, for example if one treatment has advantages other than the outcome of interest or when a one-sided test may be appropriate (which is discussed in a later chapter), and a few comments on the criticism that significance tests give a false sense of exactness to conclusions, there is a section entitled "Justification of Chi-square with Fourfold Tables". This section begins by outlining Fisher's exact test for 2×2 tables, although not referring to the approach as such, while indicating that it will be further discussed in Chapter 8. A comparison is then made, for a particular example of comparing two samples with failure rates of 1/10 and 5/10, of the results obtained using the exact calculation of a significance level and the chi-squared approximation. The results are similar and reference is made to a larger study of the approximation. Some precautions when using the approximation are then given, for example that all observed cells must be greater than 1. The guideline of expected values in a contingency table being greater than 5 when chi-squared is used is not discussed until Chapter 8 when tables larger than 2×2 are discussed. Seven examples of this methodology are given in the next section with broad

ranging discussion of each, including topics such as importance versus significance and the nature of the sampling.

The importance of estimation and not just significance testing is highlighted in the next section on estimating the difference in failure rates between samples but only the simple method of comparing the rates using binomial variances is presented which is not surprising as later alternatives based on odds ratios were not in widespread use at the time.

A section on “Sizes of Samples" makes a few points about sample size beyond the obvious observation that the amount of evidence is most influenced by sample size. These include the advantages of equal sample sizes, the presence of bias even in large samples, the confusion of the term “representative" with random and the recognition that sample size is reflected in significance tests if understood correctly and should not influence testing levels. Some guidelines and methods for sample size estimation are then given,

A brief discussion of tables larger than 2×2 follows with a forward reference to more discussion in Chapter 8. The example used for this is discussed broadly and notably Mainland points out that it is not appropriate to combine categories after examining the data, concluding the section with the comment that “individual comparisons are permissible only if …a complete analysis has revealed significant heterogeneity".

Chapter 5: Variation Between Measurements

In this chapter, Mainland turns to the analysis of measurement, or continuous, data. The early sections of the chapter relate, first, to significance testing regarding a sample mean and, secondly, the comparison of (two) sample means, both using the *t*-distribution as the basis of statistical inference. Necessary definitions, such as of the standard deviation, the standard error of the mean and the pooled estimate of an (assumed) constant variance for the two sample problem, are given and, for both problems, calculations for the relevant confidence intervals are given. The examples used for testing a sample mean are all based on differences between observations on the same subject, differing in time or with respect to another characteristic, so that a test of the mean equal to zero is of interest. A brief discussion is given to suggest that a better experiment would involve control subjects where the intra-subject measurements do not differ with respect to the characteristic of interest.

The normal distribution is mentioned in passing (p.149) in the discussion of t-tests but there then follows a section, “Normal Curve Methods in Testing Means", that justifies the assumption, and use, of normal distributions that lie behind t-tests. It is mentioned that measurement data may take different forms and need not necessarily be continuous as long as their distribution approximates the bell shape of a normal distribution. Some plots illustrate this. Also, a subsection points out that distributions of means are likely to be normal even if the data distribution is not,

Another subsection, “Criteria of Suitability of Data", argues that even without formal testing there is broad experience showing that t-test methodology based on normal assumptions is generally safe, with a reference to Fisher’s book *Design of Experiments*. The recognition that

not all data can be handled by these methods is made however and guidelines given for thinking about this for particular studies.

Two brief sections follow, on “Averages" and “Standard Deviations", to further discuss distributional shape. In the former, modes are introduced along with the concept of multi-modal distributions, as are medians. It is noted that the mean may not necessarily be of most interest but that for normal distributions it is of greatest usefulness. In the latter section, standard deviations of a population or large sample, standard errors of means and the standard error of a difference between means are discussed. The interpretation of multiples such as 0.6745 (“probable error"), 1.96 and 2.776, of standard deviations in terms of distributional shape is presented.

After a brief section on significance tests with known standard deviations, it is explained that Student’s t-test is the solution to the problem that the standard deviation for most samples of interest are not known. This is followed by an expanded discussion of the t-distribution and its use in significance testing and in determining confidence intervals, noting also that it approaches the normal distribution as the degrees of freedom becomes large, although this term is not used until Chapter 8. This section is followed by a brief section on sample size which reiterates the general discussion of this for enumeration data in Chapter 4.

Situations when the use of the *t*-distribution may be questioned are addressed in the next section on “Non-Normal Distributions". Hidden heterogeneity is illustrated when a single sample of treated individuals is examined when, in fact, the treatment only is effective on a subgroup of the sample. It is stressed that there is nothing wrong with the t-test, *per se*, but the interpretation of the results is the problem. The use of logarithms is suggested in order to apply the t-test to positively skewed data such as durations with grouping into classes and the use of contingency table methods advocated as often useful with such data to avoid the need to use logarithms or more complex distributional fitting. There is advice to avoid the unsuitable application of t-tests to coarsely grouped data and, for percentages and ratios, it is noted that regression methods to be discussed later may be useful and that special methods do exist, with reference again to Fisher’s work.

This long chapter continues with a discussion of the Analysis of Variance and the use of the F-test to compare more than two means. The “most obvious" advantage of the method is discussed to be that for all comparisons the information on variation can be derived from the full sample size. The brief mention of the method, and the experimental designs introduced by Fisher through it, is justified as it is unlikely to be used by medical students. However, the importance of being aware of the method is said to be because no one who is unaware of it knows what “modern statistics" means, the method will increasingly be seen in medical research and it is needed to see how defective “old-fashioned" methods still in use are. The possibilities with more complex experimental designs are illustrated with discussion of a simple factorial design. Neuhauser, Provost, and Provost (2020) provide an interesting historical perspective on factorial designs in medical research and this issue is also discussed by Matthews (2026b).

A very brief section of this chapter then discusses the comparison of standard deviations with only a forward reference to F-tests found in Chapter 8 and the use of coefficients of variation to compare different types of measurement although no formal significance tests are available. Two sections then give some examples of how enumeration tests may be applied to measurement data and how procedures for measurement data can sometimes by applied to enumeration data but noting that t-tests should not be used with ratios.

Mixture models

A short subsection of interest, titled “Heterogeneous Samples", makes the general point about being aware of known heterogeneity but focusses on the example of dental caries where patients with zero-caries may have a high or low frequency independent of the frequencies of non-zero categories of caries. This can be seen to be an early example of the possible value of mixture models for zero-heavy count data and other similar data (Farewell et al. 2017).

The material in this chapter finishes with some, now dated, comments on reading of scales with advice on the inadvisability of too much precision being assumed, a few remarks on the choice of the number of decimal places in data recording and the reporting of statistical tests, and some, still very relevant, comments on outliers and decisions to include or exclude them.

Chapter 6: Relationships Between Measurements - Concomitant Variation and Trends

This chapter deals with methods to examine how measurement variables vary together, although it is noted that the methods can also be used for enumeration data in some cases. The primary focus is on methods for two variables. In the chapter introduction, Mainland acknowledges that the chapter is short, although whole books have been written on the topic, and that the focus is on showing the value of the methods but also showing the dangers of misinterpretation of the methods.

After giving a few examples of concomitant variables, such as height and weight, Mainland first examines the use, and misuse, of graphs. He starts by giving two cautions, one from Fisher and one from Hill, that relate to the use of graphs to illustrate possible relationships but not to be regarded as evidence of relationships. Mainland then also points out some common problems with graphs such as different scales on the two axes, non-zero origins, confusion of relative and absolute changes, and the use of logarithms to examine rates of change. He also highlights that it should not be necessary to refer to the text of a presentation to know what the graph represents. A few comments are also made on scatter plots before indicating the need to use other methods, notably regression lines, to examine the significance of relationships.

Although not giving the origin of the term (heights of sons regressing towards the mean compared to fathers’ heights), the natural sense of regression as “going back" is said to be unhelpful. Mainland suggests the terms “mean trend line" or “mean relation line" and then

presents an ad-hoc justification of least squares estimation of a regression line by showing an indirect method of calculating a simple mean as “the value such that the sum of squares of deviations would be less than from any other value".

The general form of a linear regression line is then given as $Y = \bar{y} + b(x - \bar{x})$, defining Y as the dependent variable and X as the independent variable, noting however that no causal relationship is implied. The form without standardising around $\bar{x}$ is also mentioned after a simple example is discussed. A brief argument is then given that a t-test can be used to test the significance of a regression by defining a t-statistic that compares the estimated value of b to its estimated standard error which involves the sum of squares of deviations from the observed regression line. The precise formula, which also depends on the sum of square deviations of the x variable, is not given. Additional, more technical, comments on regression line are given in Chapter 8 (p.294-296).

In the next section, “Coefficient of Correlation", a number of topics are addressed. The relationship between the regression coefficient and the correlation coefficient is made and the use of the correlation coefficient to indicate the strength of a relationship is illustrated. The idea of thinking of regression as explaining variation is then introduced along with the partitioning of total variation into that due to regression and that due to deviations from the regression line. The use of a table, from Fisher and Yates, to test the significance of a correlation coefficient is illustrated. The point is made that correlation, unlike regression, does not entail a distinction between the variables, *i.e.* dependent and independent, but simple measures a mutual relationship and this contrasts therefore with regression which is also more generally applicable because there is no need for a distributional assumption about X. Comments follow on the misinterpretation of correlation coefficients, and the inadvisability of using the term as a synonym for association, and some particular cautions about over interpreting high values.

The chapter continues by discussing that care should be taken when relating variables when one is a component of the other or ratios are related to one of their components in a section titled “Complex Variates". The possibility of quadratic regression is then mentioned under the section heading “Curvilinear Regression", with comments on cautions regarding the use of the method and the need to see if departures from linear relationships are significant. A brief section on “Multiple Regression" introduces the model and the terms partial regression and a coefficient of partial correlation. In the same section, perhaps surprisingly, discriminant functions are mentioned as being ‘akin" to multiple regression.

Finally, in a section titled “Practitioners and Regression Methods", Mainland says that while few practitioners might calculate a regression coefficient, the chapter should help to recognise when they ought to be used and that apparent trends and relationships need to be properly assessed through formal statistical tests. In addition, Mainland indicates that if data are to be used to examine relationships, the chapter’s purpose will have been achieved if “the practitioner knows that in such a case he should consult someone who can show him what to do".

Chapter 7: Statistical Ideas in Clinical Practice

In this chapter, Mainland deals with three topics which will be relevant to a clinical practitioner. While some technical details are given, this is described as “mostly …a chapter to read and think about".

The first topic is titled “Assessment of an Individual" but is primarily focused only on the setting of ‘normal’ values of measurements or observations from a patient. A discussion of the terminology leads to a definition of a normal feature to be one that “falls within a certain range of variation derived from the examination of healthy people" with a follow-up discussion on possible limitations of the term healthy. The importance of using proper sampling methods to acquire data for the setting of normal limits is stressed as well as the need to recognise the range of variation in a healthy population. Recognising both the skewness of some distributions and that viewing a measurement as suspicious when it is from a healthy individual is less serious than missing an abnormal reading, it is suggested that a percentile method of choosing the 10th and 90th percentiles to define a normal range is sensible. A brief section deals with applying the concept of normal values for qualitative data as well.

The second topic is covered in a section titled “Errors in Routine Measurement and Counts". This section has the aim of highlighting the caution that Mainland felt was appropriate for a clinician to have when using laboratory findings, measurement or counts. With general comments, and with application to some specific clinical settings, the need to recognise the accuracy of such findings, the sources of error such as observer and random, and the nature of errors, bias or random, is highlighted. A couple of brief but more technical subsections deal with measurements, such as blood cell counts, where error can be proportional to the count itself and with the need to correctly allow for the increased error, relative to a single measurement, when a difference in two measurements is being examined.

The final section of this chapter, titled “Routine Statistics - Public Health and Clinical", deals first with population based mortality rates. The potential problems that arise through cause of death classifications, either through inaccurate information on the certificates or changes over time are presented. As well, the tradeoff between the usefulness and precision of a population specific rate because of its dependence on the size of the population is highlighted. More criticism is given of rates based on hospital data and here Mainland chooses to discuss in detail an example used by Berkson to illustrate confounding, *i.e.* Berkson’s bias. Nevertheless, the usefulness of such rates, in spite of their potential problems, is stressed with reference to Greenwood, “the perfect is the enemy of the good" and their use by Florence Nightingale in her investigations. In the closing paragraph of this section, Mainland also presciently outlines the value of longitudinal data, carefully collected by practitioners, in understanding the natural history of disease.

The remainder of the chapter presents, with a reference to pioneering texts such as that of Pearl (1941), the essentials of how life tables are calculated and interpreted and the need for and usefulness of standardised rates in comparisons of a set of, say, age, race and sex

specific rates across populations. For the latter, a particular discussion is given of the use of subject-years as a denominator, its value and its limitations.

Chapter 8: Some Further Hints for Investigators

According to Mainland's introduction, this chapter "embodies a number of elementary items …useful in helping those with little statistical experience", with "no claim to completeness or to judicious balance". The topics are varied, some being technical and some being general comments and advice. The technical work is perhaps elementary in terms of statistical theory although more technical and complex than the rest of the book.

Under the title "Sources of Help", personal help is discussed with Mainland pointing out that the best people to help the medical researcher may not be a mathematician, or even a public health researcher who does not undertake experimental research, but rather a worker in applied science, such as agricultural science. Books that go beyond the methods discussed in this book are highlighted along with books of statistical tables. A brief discussion is given of computational aids.

What follows in this miscellaneous collection of topics are two sections which present various algebraic developments related to binomial probabilities and a discussion of the use of $\sqrt{Npq}$ the standard deviation of a binomial distribution which can be used for significance tests and confidence intervals when a normal approximation to the binomial is reasonably made. Comparative comments are made with respect to methdodology based on chi-squared distributions.

A brief section is then presented describing the Poisson distribution as a limiting case of the binomial as p becomes small. with a few examples of applications. Following this, Mainland returns to the topic of "Random Sampling", particularly noting that the randomisation plan for an experiment should be done in advance and that samples that do not 'appear' random should not be rejected. Note however that Fisher (see conversation reported in Morgan and Rubin (2012)) and Yates (1948), in considering restricted randomisation, were both a little more liberal in this regard.

As discussed in the main text of this paper, this section presents a critique of alternate treatment allocation in clinical trials. Comments on random number tables and some examples of randomisation in different contexts are also given.

Lost to follow-up studies

In a single paragraph of the random sampling section, labelled "Difficult Problems", there is an interesting reference to the problem of 'no response' (to letters) with the suggestion that a random sample of non-responses be followed up with various other methods so that comparisons can be made between the responding and non-responding subjects.

The next section presents a general discussion of one-sided tests under the assumption that the investigator knows that one treatment may be better but cannot be any worse than an alternative. This argues for a doubling of the usual chi-squared significance levels and is an example of "the right of an experimenter to choose his own level of significance". No

discussion is given of one-sided tests in the situation when departures in only one direction are 'of interest'.

A technical section follows on how to calculate the hypergeometric probabilities needed for Fisher's exact test for 2×2 tables although the term hypergeometric is not used. A brief section is also given on confidence intervals for the difference between population percentages from a 2×2 table. This presents a possible alternative to the normal based methods presented in Chapter 4.

The ninth section in the chapter, "Chi-Square Tests with Multinomial Samples", outlines chi-squared tests for goodness-of-fit (to specified class probabilities) and for two-way contingency tables larger than 2×2. The reliability of the chi-square approximation when expected values are greater than 5 is discussed, with reference to recent results suggesting lower expectations can be allowed in some cases.

While degrees of freedom are mentioned in the previous section, the next section presents a specific discussion of this topic explaining that these are related to the number of independent pieces of information contributing to an estimate. This is illustrated with reference to variance estimation and the single degree of freedom relevant for tests of significance with 2×2 tables.

Section 11 discusses some suggested methods to combine information from multiple 2×2 tables, pointing out particularly the need for a test of heterogeneity before pooling of data can be used. A reference is made to Snedecor (1946) for the proper handling of $2 \times 2 \times 2$ tables with a note that a previous suggestion of Mainland (1948) may mislead if samples are small.

The next section picks up the topic of Analysis of Variance, first mentioned in Chapter 5. Computational and technical detail is provided with reference to the various sums of squares, degrees of freedom, significance tests *etc.* Interactions are discussed, and the necessary calculations given, for a design with two classifications, followed by brief mention of more complex designs, including incomplete block designs, and the analysis of covariance.

A brief section follows presenting the details of the F-test for homogeneity of variance, the issue of which has been alluded to twice in earlier chapters. A range of transformations are explained and discussed in the next section including the square root transformation of counts, the inverse sine for percentages, the logarithm for skewed distributions and the probit for use in dose response studies. The following brief section indicates that skewness that may not be dealt with through a logarithmic transformation can be addressed through a categorisation of the response and enumeration methods, perhaps with some loss of sensitivity.

The numerical details of a regression analysis are given in the last section with reference to other books for more information before explaining the structure of regression coefficient in terms of comparing a covariance, of Y and X, to a variance, of X.