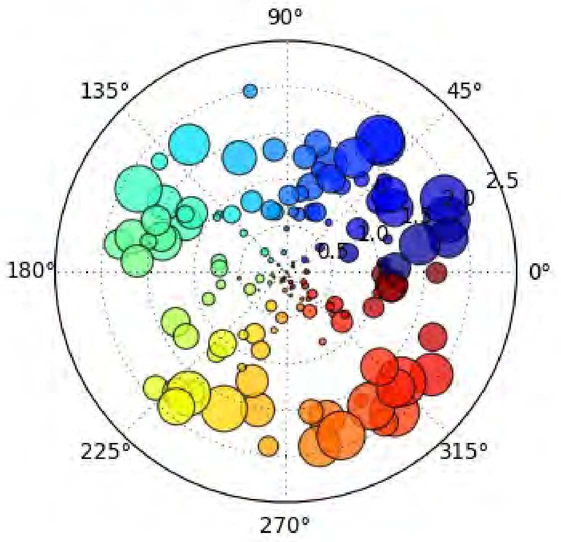
Statistical Analysis Handbook

A comprehensive handbook of statistical concepts, techniques and software tools

Dr M J de Smith



Statistical Analysis Handbook

A comprehensive handbook of statistical concepts, techniques and software tools

by Dr M J de Smith

Statistical Analysis Handbook

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

Front cover image: Polar bubble plot (source: MatPlotLib library, Python)

Rear cover image: Florence Nightingale's polar diagram of causes of mortality, by date (source: Wikipedia)

Citation: de Smith M J (2015) STATSREF: Statistical Analysis Handbook - a web-based statistics resource. The Winchelsea Press, Winchelsea, UK

Table of Contents

Part I

Introduction

Introduction

The definition of what is meant by *statistics* and *statistical analysis* has changed considerably over the last few decades. Here are two contrasting definitions of what statistics is, from eminent professors in the field, some 60+ years apart:

"Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition 'natural phenomena' includes all the happenings of the external world, whether human or not." Professor Maurice Kendall, 1943, p2 [MK1]

"Statistics is: the fun of finding patterns in data; the pleasure of making discoveries; the import of deep philosophical questions; the power to shed light on important decisions, and the ability to guide decisions..... in business, science, government, medicine, industry..." Professor David Hand [DH1]

As these two definitions indicate, the discipline of statistics has moved from being grounded firmly in the world of measurement and scientific analysis into the world of exploration, comprehension and decision-making. At the same time its usage has grown enormously, expanding from a relatively small set of specific application areas (such as design of experiments and computation of life insurance premiums) to almost every walk of life. A particular feature of this change is the massive expansion in information (and misinformation) available to all sectors and age-groups in society. Understanding this information, and making well-informed decisions on the basis of such understanding, is the primary function of modern statistical methods.

Our objective in producing this Handbook is to be comprehensive in terms of concepts and techniques (but not necessarily exhaustive), representative and independent in terms of software tools, and above all practical in terms of application and implementation. However, we believe that it is no longer appropriate to think of a standard, discipline-specific textbook as capable of satisfying every kind of new user need. Accordingly, an innovative feature of our approach here is the range of formats and channels through which we disseminate the material - web, ebook and in due course, print. A major advantage of the electronic formats is that the text can be embedded with internal and external hyperlinks. In this Handbook we utilize both forms of link, with external links often referring to a small number of well-established sources, notably MacTutor for bibliographic information and a number of other web resources, such as Eric Weisstein's Mathworld and the statistics portal of Wikipedia, for providing additional material on selected topics.

The treatment of topics in this Handbook is relatively informal, in that we do not provide mathematical proofs for much of the material discussed. However, where it is felt particularly useful to clarify how an expression arises, we do provide simple derivations. More generally we adopt the approach of using descriptive explanations and worked examples in order to clarify the usage of different measures and procedures. Frequently convenient software tools are used for this purpose, notably SPSS/PASW, The R Project, MATLab and a number of more specialized software tools where appropriate.

Just as all datasets and software packages contain errors, known and unknown, so too do all books and websites, and we expect that there will be errors despite our best efforts to remove these! Some may be genuine errors or misprints, whilst others may reflect our use of specific versions of software packages and their documentation. Inevitably with respect to the latter, new versions of the packages that we have used to illustrate this Handbook will have appeared even before publication, so specific examples, illustrations and comments on scope or restrictions may have been superseded. In all cases the user should review the documentation provided with the software version they plan to use, check release notes for changes and known bugs, and look at any relevant online services (e.g. user/ developer forums and blogs on the web) for additional materials and insights.

The interactive web version of this Handbook may be accessed via the associated Internet site: www.statsref.com. The contents and sample sections of the PDF version may also be accessed from this site. In both cases the information is regularly updated. The Internet is now well established as society's principal mode of information exchange, and most aspiring users of statistical methods are accustomed to searching for material that can easily be customized to specific needs. Our objective for such users is to provide an independent, reliable and authoritative first port of call for conceptual, technical, software and applications material that addresses the panoply of new user requirements.

Readers wishing to obtain a more in-depth understanding of the background to many of the topics covered in this Handbook should review the Suggested Reading topic. Those seeking examples of software tools that might be used for statistical analysis should refer to the Software section.

References

[DH1] D Hand (2009) President of the Royal Statistical Society (RSS), RSS Conference Presentation, November 2009

[MK1] Kendall M G, Stuart A (1943) The Advanced Theory of Statistics: Volume 1, Distribution Theory. Charles Griffin & Company, London. First published in 1943, revised in 1958 with Stuart

How to use this Handbook

This Handbook is designed to provide a wide-ranging and comprehensive, though not exhaustive, coverage of statistical concepts and methods. Unlike a Wiki the Handbook has a more linear flow structure, and in principle can be read from start to finish. In practice many of the topics, particularly some of those described in later parts of the document, will be of interest only to specific users at particular times, but are provided for completeness. Users are recommended to read the initial four topics - Introduction, Statistical Concepts, Statistical Data and Descriptive Statistics, and then select subsequent sections as required.

Navigating around the PDF or web versions of this Handbook is straightforward, but to assist this process a number of special facilities have been built into the design to make the process even easier. These facilities include:

- Tests Index - this is a form of 'how to' index, i.e. it does not assume that the reader knows the name of the test they may need to use, but can navigate to the correct item by the index description
- Reference links and bibliography - within the text all books and articles referenced are linked to the full reference at the end of the topic section (in the References subsection) in the format [XXXn] and in the complete bibliography at the end of the Handbook
- Hyperlinks - within the document there are two types of hyperlink: (i) internal hyperlinks - when clicking on these links you will be directed to the linked topic within this Handbook; (ii) external hyperlinks - these provide access to external resources for which you need an active internet connection. When the external links are clicked the appropriate topic is opened on an external website such as Wikipedia
- Search facilities - the web and PDF versions of this Handbook facilitate free text search, so as long as you know roughly what you are looking for, you should be able to find it using this facility

Intended audience and scope

Ian Diamond, Statistician and at the time Chief Executive of the UK's Economic and Social Research Council (ESRC), gave the following anecdote (which I paraphrase) during a meeting in 2009 at the Royal Statistical Society in London: "Some time ago I received a brief email from a former student. In it he said '*your statistics course was the one I hated most at University and was more than glad when it was over.... but in my working career it has been the most valuable of any of the courses I took... !*'" So, despite its challenges and controversies, taking time to get to grips with statistical concepts and techniques is well worth the effort.

With this perspective in mind, this Handbook has been designed to be accessible to a wide range of readers - from undergraduates and postgraduates studying statistics and statistical analysis as a component of their specific discipline (e.g. social sciences, earth sciences, life sciences, engineers), to practitioners and professional research scientists. However, it is not intended to be a guide for mathematicians, advanced students studying statistics or for professional statisticians. For students studying for academic or professional qualifications in statistics, the level and content adopted is that of the Ordinary and Higher Level Certificates of the Royal Statistical Society (RSS). Much of the material included in this Handbook is also appropriate for the Graduate Diploma level also, although we have not sought to be rigorous or excessively formal in our treatment of individual statistical topics, preferring to provide less formal explanations and examples that are more approachable by the non-mathematician with links and references to detailed source materials for those interested in derivation of the expressions provided.

The Handbook is much more than a cookbook of formulas, algorithms and techniques. Its aim is to provide an explanation of the key techniques and formulas of statistical analysis, often using examples from widely available software packages. It stops well short, however, of attempting a systematic evaluation of competing software products. A substantial range of application examples is provided, but any specific selection inevitably illustrates only a small subset of the huge range of facilities available. Wherever possible, examples have been drawn from non-academic and readily reproducible sources, highlighting the widespread understanding and importance of statistics in every part of society, including the commercial and government sectors.

References

Royal Statistical Society: Examinations section, Documents: http://www.rss.org.uk/main.asp?page=1802

Suggested reading

There are a vast number of books on statistics - Amazon alone lists 10,000 "professional and technical" works with *statistics* in their title. There is no single book or website on statistics that meets the need of all levels and requirements of readers, so the answer for many people starting out will be to acquire the main 'set books' recommended by their course tutors and then to supplement these with works that are specific to their application area. Every topic and subtopic in this Handbook almost certainly has at least one entire book devoted to it, so of necessity the material we cover can only provide the essential details and a starting point for deeper understanding of each topic. As far as possible we provide links to articles, web sites, books and software resources to enable the reader to pursue such questions as and when they wish.

Most statistics texts do not make for easy or enjoyable reading! In general they address difficult technical and philosophical issues, and many are demanding in terms of their mathematics. Others are much more approachable - these books include 'classic' undergraduate text books such as Feller (1950, [FEL1]), Mood and Graybill (1950, [MOO1]), Hoel (1947, [HOE1]), Adler and Roessler (1960, [ADL1]), Brunk (1960, [BRU1]), Snedecor and Cochrane (1937, [SNE1]) and Yule and Kendall (1950, [YUL1]) - the dates cited in each case are when the books were originally published; in most cases these works then ran into many subsequent editions and though most are now out-of-print some are still available. A more recent work, available from the American Mathematical Society and also as a free PDF, is Grinstead and Snell's (1997) An Introduction to Probability [GRI1]. Still in print, and of continuing relevance today, is Huff (1954, [HUF1]) "How to Lie with Statistics" which must be the top selling statistics book of all time. A much more recent book, with a similar focus, is Blastland and Dilnot's "The Tiger that Isn't" [BLA1], which is full of examples of modern-day use and misuse of statistics. Another delightful, lighter weight book that remains very popular, is Gonik and Smith's "Cartoon Guide to Statistics" (one of a series of such cartoon guides by Gonik and co-authors, [GON1]). A very useful quick guide is the foldable free PDF format leaflet "Probability & Statistics, Facts and Formulae" published by the UK Maths, Stats and OR Network [UKM1].

Essential reading for anyone planning to use the free and remarkable "R Project" statistical resource is Crawley's "The R Book" (2007, [CRA1]) and associated data files; and for students undertaking an initial course in statistics using SPSS, Andy Field's "Discovering Statistics Using SPSS" provides a gentle introduction with many worked examples and illustrations [FIE1]. Both Field and Crawley's books are very large - around 900 pages in each case. Data obtained in the social and behavioral sciences do not generally conform to the strict requirements of traditional (parametric) inferential statistics and often require the use of methods that relax these requirements. These so-called nonparametric methods are described in detail in Siegel and Castellan's widely used text "Nonparametric Statistics for the Behavioral Sciences" (1998, [SIE1]) and Conover's "Practical Nonparametric Statistics" (1999, [CON1]).

A key aspect of any statistical investigation is the use of graphics and visualization tools, and although technology is changing this field Tufte's "The Visual Display of Quantitative Information" [TUF1] should be considered as essential reading, despite its origins in the 1980s and the dramatic changes to visualization possibilities since its publication.

With a more practical, applications focus, readers might wish to look at classics such as Box *et al*. (1978, 2005, [BOX1]) "Statistics for Experimenters" (highly recommended, particularly for those involved in industrial processes), Sokal and Rohlf (1995, [SOK1]) on Biometrics, and the now rather dated book on Industrial Production edited by Davies (1961, [DAV1]) and partly written by the extraordinary George Box whilst a postgraduate student at University College London. Box went on to a highly distinguished career in statistics, particular in industrial applications, and is the originator of many statistical techniques and author of several groundbreaking books. He not only met and worked with R A Fisher but later married one of Fisher's daughters! Crow *et al.* (1960, reprinted in 2003, [CRO1]) published a concise but exceptionally clear "Statistics Manual" designed for use by the US Navy, with most of its examples relating to ordnance - it provides a very useful and compact guide for non-statisticians working in a broad range of scientific and engineering fields.

Taking a further step towards more demanding texts, appropriate for mathematics and statistics graduates and post-graduates, we would recommend Kendall's Library of Statistics [KEN1], a multi-volume authoritative series each volume of which goes into great detail on the area of statistics it focuses upon. For information on statistical distributions we have drawn on a variety of sources, notably the excellent series of books by Johnson and Kotz [JON1], [JON2] originally published in 1969/70. The latter authors are also responsible for the comprehensive but extremely expensive nine volume "Encyclopedia of Statistical Sciences" (1998, 2006, [KOT1]). A much more compact book of this type, with very brief but clear descriptions of around 500 topics, is the "Concise Encyclopedia of Statistics" by Dodge (2002, [DOD1]).

With the rise of the Internet, web resources on statistical matters abound. However, it was the lack of a single, coherent and comprehensive Internet resource that was a major stimulus to the current project. The present author's book/ebook/website www.spatialanalysisonline.com has been extremely successful in providing information on Geospatial Analysis to a global audience, but its focus on 2- and 3-dimensional spatial problems limits its coverage of statistical topics. However, a significant percentage of Internet search requests that lead users to this site involve queries about statistical concepts and techniques, suggesting a broader need for such information in a suitable range of formats, which is what this Handbook attempts to provide.

A number of notable web-based resources providing information on statistical methods and formulas should be mentioned. The first is Eric Weisstein's excellent Mathworld site, which has a large technical section on probability and statistics. Secondly there is Wikipedia (Statistics section) - this is a fantastic resource, but is almost by definition not always consistent or entirely independent. This is particularly noticeable for topics whose principal or original authorship reflects the individual's area of specialism: social science, physics, biological sciences, mathematics, economics etc, and in some instances their commercial background (e.g. for specific software packages). Both Mathworld and Wikipedia provide a topic-by-topic structure, with little or no overall guide or flow to direct users through the maze of topics, techniques and tools, although Wikipedia's core structure is very well defined. This contrasts with the last two of our recommended websites: the NIST/SEMATECH online Engineering Statistics e-Handbook, and the UCLA Statistics Online Computational Resource (SOCR). These latter resources are much closer to our Handbook concept, providing information and guidance on a broad range of topics in a lucid, structured and discursive manner. These sites have a further commonality with our project - their use of particular software tools to illustrate many of the techniques and visualizations discussed. In the case of NIST/SEMATECH e-Handbook a single software tool is used, Dataplot, which is a fairly basic, free, cross-platform tool developed and maintained by the NIST. The UCLA Statistics Online Computational Resource project makes extensive use of interactive Java applets to deliver web-enabled statistical tools. The present Handbook references a wider range of software tools to illustrate its materials, including Dataplot, R, SPSS, Excel and XLStat, MATLab, Minitab, SAS/STAT and many others. This enables us to provide a broader ranging commentary on the toolsets available, and to compare the facilities and algorithms applied by the different implementations. Throughout this Handbook we make extensive reference to functions and examples available in R, MATLab and SPSS in particular.

References

[ADL1] Adler H L, Roessler E B (1960) Introduction to Probability and Statistics. W H Freeman & Co, San Francisco

[BLA1] Blastland M, Dilnot A (2008) The Tiger That Isn't. Profile Books, London

[BOX1] Box G E P, Hunter J S, Hunter W G (1978) Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. J Wiley & Sons, New York. The second, extended edition was published in 2005

[BRU1] Brunk H D (1960) An Introduction to Mathematical Statistics. Blaisdell Publishing, Waltham, Mass.

[CHA1] Chatfield C (1975) The Analysis of Times Series: Theory and Practice. Chapman and Hall, London, UK (see also, extended 6th ed.)

[CON1] Conover W J (1999) Practical Nonparametric Statistics. 3rd ed., J Wiley & Sons, New York

[CRA1] Crawley M J (2007) The R Book. J Wiley & Son, Chichester, UK

[CRO1] Crow E L, Davis F A, Maxfield M W (1960) Statistics Manual. Dover Publications. Reprinted in 2003 and still available

[DAV1] Davies O L ed. (1961) Statistical Methods in Research and Production. 3rd ed., Oliver & Boyd, London

[DOD1] Dodge Y (2002) The Concise Encyclopedia of Statistics. Springer, New York

[FEL1] Feller W (1950) An Introduction to Probability Theory and Its Applications. Vols 1 and 2. J Wiley & Sons

[FIE1] Field A (2009) Discovering Statistics Using SPSS. 3rd ed., Sage Publications

[GON1] Gonik L, Smith W (1993) Cartoon Guide to Statistics. Harper Collins, New York

[GRI1] Grinstead C M, Snell J L (1997) Introduction to Probability, 2nd ed. AMS, 1997

[HOE1] Hoel P G (1947) An Introduction to Mathematical Statistics. J Wiley & Sons, New York

[HUF1] Huff D (1954) How to Lie with Statistics. W.W. Norton & Co, New York

[JON1] Johnson N L, Kotz S (1969) Discrete distributions. J Wiley & Sons, New York. Note that a 3rd edition of this work, with revisions and extensions, is published by J Wiley & Sons (2005) with the additional authorship of Adrienne Kemp of the University of St Andrews.

[JON2] Johnson N L, Kotz S (1970) Continuous Univariate Distributions - 1 & 2. Houghton-Mifflin, Boston

[KEN1] Kendall M G, Stuart A (1943) The Advanced Theory of Statistics: Volume 1, Distribution Theory. Charles Griffin & Company, London. First published in 1943, revised in 1958 with Stuart

[KOT1] Kotz S, Johnson L (eds.) (1988) Encyclopedia of Statistical Sciences. Vols 1-9, J Wiley & Sons, New York. A 2nd edition with almost 10,000 pages was published with Kotz as the Editor-in-Chief, in 2006

[MAK1] Mackay R J, Oldford R W (2002) Scientific method, Statistical method and the Speed of Light, Working Paper 2002-02, Dept of Statistics and Actuarial Science, University of Waterloo, Canada. An excellent paper that provides an insight into Michelson's 1879 experiment and explanation of the role and method of statistics in the larger context of science

[MOO1] Mood A M, Graybill F A (1950) Introduction to the Theory of Statistics. McGraw-Hill, New York

[SIE1] Siegel S, Castellan N J (1998) Nonparametric Statistics for the Behavioral Sciences. 2nd ed., McGraw Hill, New York

[SNE1] Snedecor G W, Cochran W G (1937) Statistical Methods. Iowa State University Press. Many editions

[SOK1] Sokal R R, Rohlf F J (1995) Biometry: The Principles and Practice of Statistics in Biological Research. 2nd ed., W H Freeman & Co, New York

[TUF1] Tufte E R (2001) The Visual Display of Quantitative Information. 2nd edition. Graphics Press, Cheshire, Conn.

[UKM1] UK Maths, Stats & OR Network. Guides to Statistical Information: Probability and statistics Facts and Formulae. www.mathstore.ac.uk

[YUL1] Yule G U, Kendall M G (1950) An Introduction to the Theory of Statistics. Griffin, London, 14th edition (first edition was published in 1911 under the sole authorship of Yule)

Web sites:

Mathworld: http://mathworld.wolfram.com/

NIST/SEMATECH e-Handbook of Statistical Methods: http://www.itl.nist.gov/div898/handbook/

UCLA Statistics Online Computational Resource (SOCR) : http://socr.ucla.edu/SOCR.html

Wikipedia: http://en.wikipedia.org/wiki/Statistics

Notation and symbology

In order to clarify the expressions used here and elsewhere in the text, we use the notation shown in the table below. Italics are used within the text and formulas to denote variables and parameters. Typically in statistical literature, the Roman alphabet is used to denote sample variables and sample statistics, whilst Greek letters are used to denote population measures and parameters. An excellent and more broad-ranging set of mathematical and statistical notation is provided on the Wikipedia site.

Notation used in this Handbook

Item	Description
$[a,b]$	A closed interval of the Real line, for example [0,1] means the infinite set of all values between 0 and 1, including 0 and 1
(a,b)	An open interval of the Real line, for example (0,1) means the infinite set of all values between 0 and 1, NOT including 0 and 1. This should not be confused with the notation for coordinate pairs, (x,y), or its use within bivariate functions such as $f(x,y)$ - the meaning should be clear from the context
$\{x_i\}$	A set of n values x_1, x_2, x_3, … x_n, typically continuous interval- or ratio-scaled variables in the range $(-\infty,\infty)$ or $[0,\infty)$. The values may represent measurements or attributes of distinct objects, or values that represent a collection of objects (for example the population of a census tract)
$\{X_i\}$	An ordered set of n values X_1, X_2, X_3, … X_n, such that $X_i \leq X_i+1$ for all i
$\mathbf{X}$,$\mathbf{x}$	The use of bold symbols in expressions indicates matrices (upper case) and vectors (lower case)
$\{f_i\}$	A set of k frequencies ($k \leq n$), derived from a dataset $\{x_i\}$. If $\{x_i\}$ contains discrete values, some of which occur multiple times, then $\{f_i\}$ represents the number of occurrences or the count of each distinct value. $\{f_i\}$ may also represent the number of occurrences of values that lie in a range or set of ranges, $\{r_i\}$. If a dataset contains n values, then the sum $\sum f_i=n$. The set $\{f_i\}$ can also be written $f(x_i)$. If $\{f_i\}$ is regarded as a set of weights (for example attribute values) associated with the $\{x_i\}$, it may be written as the set $\{w_i\}$ or $\mathrm{w}(x_i)$. If a set of frequencies, $\{f_i\}$, have been standardized by dividing each value f_i by their sum, $\sum f_i$ then $\{f_i\}$ may be regarded as a set of estimated probabilities and $\sum f_i=1$
Σ	Summation symbol, e.g. $x_1+x_2+x_3+\ldots+x_n$. If no limits are shown the sum is assumed to apply to all subsequent elements, otherwise upper and/or lower limits for summation are provided
$\cap$	Set intersection. The notation $P(A \cap B)$ is used to indicate the probability of A *and* B

Item	Description
∪	Set union. The notation P(A∪B) is used to indicate the probability of A *or* B
Δ	Set symmetric difference. The set of objects in A that are not in B plus the set of objects in B that are not in A
∫	Integration symbol. If no limits are shown the sum is assumed to apply to all elements, otherwise upper and/or lower limits for integration are provided
∏	Product symbol, e.g. $x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n$. If no limits are shown the product is assumed to apply to all subsequent elements, otherwise upper and/or lower limits for multiplication are provided
^	Hat or carat symbol: used in conjunction with Greek symbols (directly above) to indicate a value is an estimate of a parameter or the true population value
→	Tends to, typically applied to indicate the limit as a variable tends to 0 or ∞
¯	Solidus or overbar symbol: used directly above a variable to indicate a value is the mean of a set of sample values
~	Two meanings apply, depending on the context: (i) "is distributed as", for example *y*~N(0,1) means the variable *y* has a distribution that is Normal with a mean of 0 and standard deviation of 1; (ii) negation, as in ~A means NOT A, or sometimes referred to as the complement of A. Note that the R language uses this symbol when defining regression models
!	Factorial symbol. $z=n!$ means $z=n(n-1)(n-2)\ldots1$. $n>=0$. Note that 0! is defined as 1. Usually applied to integer values of *n*. May be defined for fractional values of *n* using the Gamma function. Note that for large *n* Stirling's approximation may be used. R: factorial(*n*) – computes *n*!; if a range is specified, for example 1:5 then all the factorials from 1 to 5 are computed
$\binom{n}{r}$	Binomial expansion coefficients, also written as ^nC_r, or similar, and shorthand for $n!/[(n-r)!r!]$.
≡	'Equivalent to' symbol
≈	'Approximately equal to' symbol
∝	Proportional to
∈	'Belongs to' symbol, e.g. $x \in [0,2]$ means that *x* belongs to/is drawn from the set of all values in the closed interval [0,2]; $x \in \{0,1\}$ means that *x* can take the values 0 and 1
≤	Less than or equal to, represented in the text where necessary by <= (provided in this form to support display by some web browsers)

Item	Description
≥	Greater than or equal to, represented in the text where necessary by >= (provided in this form to support display by some web browsers)
$\lfloor x \rfloor$	Floor function. Interpreted as the largest integer value not greater than *x*. Sometimes, but not always, implemented in software as int(*x*), where int() is the integer part of a real valued variable
$\lceil x \rceil$	Ceiling function. Interpreted as the smallest integer value not less than *x*. Sometimes, but not always, implemented in software as int(*x* +1), where int() is the integer part of a real valued variable
A|B	"given", as in P(A|B) is the probability of A given B or A *conditional upon* B

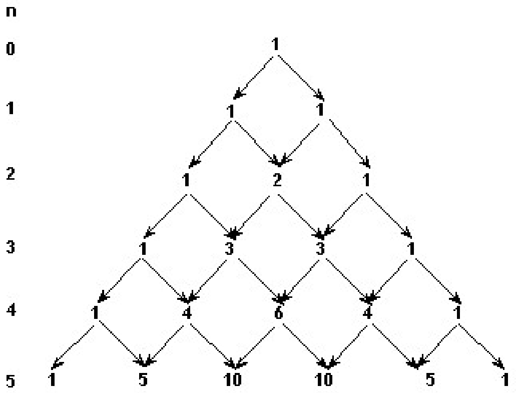
References

Wikipedia: Table of mathematical symbols: http://en.wikipedia.org/wiki/Table_of_mathematical_symbols

Historical context

Statistics is a relatively young discipline - for discussions on the history of statistics see Stigler (1986, [STI1]) and Newman (1960,[NEW1]). Much of the foundation work for the subject has been developed in the last 150 years, although its beginnings date back to the 13th century involving the expansion of the series $(p+q)^n$, for n=0,1,2.... The coefficients of this 'binomial' expansion were found to exhibit a well defined pattern (illustrated below) known as Pascal's triangle. Each coefficient can be obtained as the sum of the two immediately above in the diagram, as indicated.

Coefficients of the Binomial expansion



Pascal used this observation to produce a formula for the coefficients, which he noted was the same as the formula for the number of different combinations (or arrangements) of r events from a set of n (r=0,1,...n). , usually denoted:

$$^{n}C_{r} \text{ or } \binom{n}{r}$$

This formula is typically expanded as:

$$^{n}C_{r}=\frac{n!}{(n-r)!r!}$$

Hence with n=5, and noting that 0! is defined as 1, we have for r=[0,1,2,3,4,5] the values [1,5,10,10,5,1] as per Pascal's triangle, above. What this formula for the coefficients says, for example, is that are 5 different ways of arranging one p and four q. These arrangements, or possible different combinations, are:

pqqqq, *qpqqq*, *qqpqq*, *qqqpq*, and *qqqqp*

and exactly the same is true if we took one q and four p's. There is only one possible arrangement of all p's or all q's, but there are 10 possible combinations or sequences if there are 2 of one and 3 of the other. The possible different combinations are:

ppqqq, *qppqq*, *qqppq*, *qqqpp*, *pqpqq*, *pqqpq*, *pqqqp*, *qpqpq*, *qpqqp*, *qqpqp*

In these examples the order of arrangement is important, and we are interested in all possible *combinations*. If the order is not important the number of arrangements would be greater and the formula simplifies to counting the number of *permutations*:

$$^{n}P_{r}=\frac{n!}{(n-r)!}$$

Assuming (p+q)=1 then clearly (p+q)^n=1. Jakob Bernoulli's theorem (published in 1713, after his death) states that if p is the probability of a single event occurring (e.g. a 2 being the result when a six-sided die is thrown), and q =1-p is the probability of it not occurring (e.g. the die showing any other value but 2) then the probability of the event occurring *at least m times* in n trials is the sum of all the terms of (p+q)^n starting from the term with elements including p^r where r≥m, i.e.

$$\sum_{r=m}^{n}\frac{n!}{r!(n-r)!}p^{r}q^{n-r}$$

So, if a die is thrown 5 times, the expected number of occasions a 2 will occur will be determined by the terms of the binomial expansion for which p =1/6, and q =1-p = 5/6):

$$p^{0}q^{5},5p^{1}q^{4},10p^{2}q^{3},10p^{3}q^{2},5p^{4}q^{1},p^{5}q^{0}$$

which in this case give us the set of probabilities (to 3dp): 0.402,0.402,0.161,0.032,0.003,0.000. So the chance of throwing *at least one* "2" from 5 throws of an unbiased die is the sum of all the terms from m=1 to 5, i.e. roughly 60% (59.8%), and the chances of all 5 throws turning up as a 2 is almost zero. Notice that we could also have computed this result more directly as 1 minus the probability of no twos, which is 1-(1/6)^0(5/6)^5=1-0.402, the same result as above.

This kind of computation, which is based on an *a priori* understanding of a problem in which the various outcomes are equally likely, works well in certain fields, such as games of chance - roulette, card games, dice games - but is not readily generalized to more complex and familiar problems. In most cases we do not know the exact chance of a particular event occurring, but we can obtain an estimate of this assuming we have a fairly large and *representative* sample of data. For example, if we collate data over a number of years on the age at which males and females die in a particular city, then one might use this information to provide an estimate of the probability that a woman of age 45 resident in that location will die within the next 12 months. This information, which is a form of *a posteriori* calculation of probability, is exactly the kind of approach that forms the basis for what are known as mortality tables, and these are used by the life insurance industry to guide the setting of insurance premiums. Statisticians involved in this particular field are called actuaries, and their principal task is to analyze collected data on all manner of events in order to produce probability estimates for a range of outcomes on which insurance premiums are then based. The collected data are typically called *statistics*, here being the plural form. The term *statistics* in the singular, refers to the science of how best to collect and analyze such data.

Returning to the games of chance examples above, we could approach the problem of determining the probability that at least one 2 is thrown from 5 separate throws of the die by conducting an experiment or *trial*. First, we could simply throw a die 5 times and count the number of times (if any) a 2 was the uppermost face. However, this would be a very small trial of just one set of throws. If we conducted many more trials, perhaps 1000 or more, we would get a better picture of the pattern of events. More specifically we could make a chart of the observed *frequency* of each type of event, where the possible events are: zero 2s, one 2, two 2s and so on up to five 2s. In practice, throwing a

6-sided die a very large number of times and counting the frequency with which each value appears is very time-consuming and difficult. Errors in the process will inevitably creep in: the physical die used is unlikely to be perfect, in the sense that differences in the shape of its corners and surfaces may lead some faces to be very slightly more likely to appear uppermost than others; as time goes on the die will wear, and this could affect the results; the process of throwing a die and the surface onto which the die is thrown may affect the results; over time we may make errors in the counting process, especially if the process continues for a very long time... in fact there are very many reasons for arguing that a physical approach is unlikely to work well.

As an alternative we can use a simple computer program with a random number generator, to simulate the throwing of a six-sided die. Although modern random number generators are extremely good, in that their randomness has been the subject of an enormous amount of testing and research, there will be a very slight bias using this approach, but it is safe to ignore this at present. In the table below we have run a simple simulation by generating a random integer number between the values of 1 and 6 a total of 100,000 times. Given that we expect each value to occur with a probability of 1/6, we would expect each value to appear approximately 16667 times. We can see that in this trial, the largest absolute difference between the simulated or observed frequency, f_o, and the *a priori* or expected frequency, f_e, is 203, which is around 1.2%.

Face	Frequency	|Observed-Expected|
1	16741	74
2	16870	203
3	16617	50
4	16635	32
5	16547	120
6	16589	78

This difference is either simply a matter of chance, or perhaps imperfections in the random number algorithm, or maybe in the simulation program. Some of this uncertainty can be removed by repeating the trial many times or using a larger number of tests in a single trial, and by checking the process using different software on different computers with different architectures. In our case we increased the trial run size to 1 million, and found that the largest percentage difference was 0.35%, suggesting that the random number generator and algorithm being used were indeed broadly unbiased, and also illustrating the so-called "Law of large numbers" or "Golden theorem", also due to Bernoulli. Essentially this law states that as the sample size is increased (towards infinity), the sample average tends to the true 'population' average. In the example of rolling a die, the possible values are 1,2,...6, the average of which is 3.5, so the long term average from a large number of trials should approach 3.5 arbitrarily closely. There are actually two variants of this law commonly recognized, the so-called Weak Law and the Strong Law, although the differences between these variants are quite subtle. Essentially the Weak Law allows for a larger (possibly infinite) number of very small differences between the true average and the long term sampled average, whilst the Strong Law allows just for a finite number of such cases.

This example has not directly told us how likely we are to see one or more 2s when the die is thrown five times. In this case we have to simulate batches of 5 throws at a time, and count the proportion of these batches that have one or more 2s thrown. In this case we again compute 100,000 trials, each of which involves 5 throws (so 0.5 million iterations in total) and we find the following results from a sequence of such trials: 59753, 59767,59806,... each of which is very close to the expected value based on the percentage we derived earlier, more precisely 59812 (59.812%). In general it is unnecessary to manually or programmatically compute such probabilities for well-known distributions such as the Binomial, since almost all statistical software packages will perform the computation for you. For example, the Excel function BINOMDIST() could be used. Until relatively recently statistical tables, laboriously calculated by hand or with the aid of mechanical calculators, were the principal means of comparing observed results with standard distributions. Although this is no longer necessary the use of tables can be a quick and simple procedure, and we have therefore included a number of these in the resources topic, Distribution tables section, of this Handbook.

A number of observations are worth making about the above example. First, although we are conducting a series of trials, and using the observed data to produce our probability estimates, the values we obtain vary. So there is a *distribution* of results, most of which are very close to our expected (true) value, but in a smaller number of cases the results we obtain might, by chance, be rather more divergent from the expected frequency. This pattern of divergence could be studied, and the proportion of trials that diverged from the expected value by more than 1%, 2% etc. could be plotted. We could then compare an observed result, say one that diverged by 7% from that expected, and ask "how likely is it that this difference is due to chance?". For example, if there was less than one chance in 20 (5%) of such a large divergence, we might decide the observed value was probably not a simple result of chance but more likely that some other factor was causing the observed variation. From the Law of Large Numbers we now know that the size of our sample or trial is important - smaller samples diverge more (in relative, not absolute, terms) than larger samples, so this kind of analysis must take into account sample size. Many real-world situations involve modest sized samples and trials, which may or may not be truly representative of the populations from which they are drawn. The subject of statistics provides specific techniques for addressing such questions, by drawing upon experiments and mathematical analyses that have examined a large range of commonly occurring questions and datasets.

A second observation about this example is that we have been able to compare our trials with a well-defined and known 'true value', which is not generally the situation encountered. In most cases we have to rely more heavily on the data and an understanding of similar experiments, in order to obtain some idea of the level of uncertainty or error associated with our findings.

A third, and less obvious observation, is that if our trial, experiments and/or computer simulations are in some way biased or incorrectly specified or incomplete, our results will also be of dubious value. In general it is quite difficult to be certain that such factors have not affected the observed results and therefore great care is needed when designing experiments or producing simulations.

Finally, it is important to recognize that a high proportion of datasets are not obtained from well-defined and controlled experiments, but are observations made and/or collections of data obtained, by third parties, often government agencies, with a whole host of known and unknown issues relating to their quality and how representative they are. Similarly, much data is collected on human populations and their behavior, whether this be medical research data, social surveys, analysis of purchasing behavior or voting intentions. Such datasets are, almost by definition, simply observations on samples from a population taken at a particular point in time, in which the sampling units (individual people) are not fully understood or 'controlled' and can only loosely be regarded as members of a well-defined 'population'.

With the explosion in the availability of scientific data during the latter part of the 18th century and early 19th century, notably in the fields of navigation, geodesy and astronomy, efforts were made to identify associations and patterns that could be used to simplify the datasets. The aim was to minimize the error associated with large numbers of observations by examining the degree to which they fitted a simple model, such as a straight line or simple curve, and then to predict the behavior of the variables or system under examination based on this approximation. One of the first and perhaps most notable of these efforts was the discovery of the method of Least Squares, which Gauss reputedly devised at the age of 18. This method was independently discovered and developed by a number of other scientists, notably Legendre, and applied in a variety of different fields. In the case of statistical analysis, least squares is most commonly encountered in connection with linear and non-linear regression, but it was originally devised simply as the 'best' means of fitting an analytic curve (or straight line) to a set of data, in particular measurements of astronomical orbits.

During the course of the late 1900s and the first half of the 20th century major developments were made in many areas of statistics. A number of these are discussed in greater detail in the sections which follow, but of particular note is the work of a series of scientists and mathematicians working at University College London (UCL). This commenced in the 1860s with the research of the scientist Sir Francis Galton (a relation of Charles Darwin), who was investigating whether characteristics of the human population appeared to be acquired or inherited, and if inherited, whether humankind could be altered (improved) by selective breeding (a highly controversial scientific discipline, known as Eugenics). The complexity of this task led Galton to develop the concepts of correlation and regression, which were subsequently developed by Karl Pearson and refined by his student, G Udny Yule, who delivered an influential series of annual lectures on statistics at UCL which became the foundation of his famous book, An Introduction to the Theory of Statistics [YUL1], first published in 1911. Another student of Pearson at UCL was a young chemist, William Gosset, who worked for the brewing business, Guinness. He is best known for his work on testing data that have been obtained from relatively small samples. Owing to restrictions imposed by his employers on publishing his work under his own name, he used the pseudonym "Student", from which the well-known "Students t-test" and the t-distribution arise. Also joining UCL for 10 years as Professor of Eugenics, was R A Fisher, perhaps the most important and influential statistician of the 20th century. Fisher's contributions were many, but he is perhaps most famous for his work on the Design of Experiments [FIS1], a field which is central to the conduct of controlled experiments such as agricultural and medical trials. Also at UCL, but working in a different field, psychology, Charles Spearman was responsible for the introduction of a number of statistical techniques including Rank Correlation and Factor Analysis. And lastly, but not least, two eminent statisticians: Austin Bradford Hill, whose work we discuss in the section on statistics in medical research, attended Pearson's lectures at UCL and drew on many of the ideas presented in developing his formative work on the application of statistics to medical research; and George Box, developer of much of the subject we now refer to as industrial statistics. Aspects of his work are included in our discussion of the Design of Experiments, especially factorial designs.

Substantial changes to the conduct of statistical analysis have come with the rise of computers and the Internet. The computer has removed the need for statistical tables and, to a large extent, the need to be able to recall and compute many of the complex expressions used in statistical analysis. They have also enabled very large volumes of data to be stored and analyzed, which itself presents a whole new set of challenges and opportunities. To meet some of these, scientists such as John Tukey and others developed the concept of Exploratory Data Analysis, or "EDA", which can be described as a set of visualization tools and exploratory methods designed to help researchers understand large and complex datasets, picking out significant features and feature combinations for further study. This field has become one of the most active areas of research and development in recent years, spreading well beyond the confines of the statistical fraternity, with new techniques such as Data Mining, 3D visualizations, Exploratory Spatio-Temporal Data Analysis (ESTDA) and a whole host of other procedures becoming widely used. A further, equally important impact of computational power, we have already glimpsed in our discussion on games of chance - it is possible to use computers to undertake large-scale simulations for a range of purposes, amongst the most important of which is the generation of pseudo-probability distributions for problems for which closed mathematical solutions are not possible or where the complexity of the constraints or environmental factors make simulation and/or randomization approaches the only viable option.

References

[FIS1] Fisher R A (1935) The Design of Experiments. Oliver & Boyd, London

[NEW1] Newman J R (1960) The World of Mathematics. Vol 3, Part VIII Statistics and the Design of Experiments. Oliver & Boyd, London

[STI1] Stigler S M (1986) The History of Statistics. Harvard University Press, Harvard, Mass.

[YUL1] Yule G U, Kendall M G (1950) Introduction to the Theory of Statistics. 14th edition, Charles Griffin & Co, London

MacTutor: The MacTutor History of Mathematics Archive. University of St Andrews, UK: http://www-history.mcs.st-and.ac.uk/

Mathworld: Weisstein E W "Weak Law of Large Numbers" and "Strong Law of Large Numbers": http://mathworld.wolfram.com/WeakLawofLargeNumbers.html

Wikipedia: History of statistics: http://en.wikipedia.org/wiki/History_of_statistics

An applications-led discipline

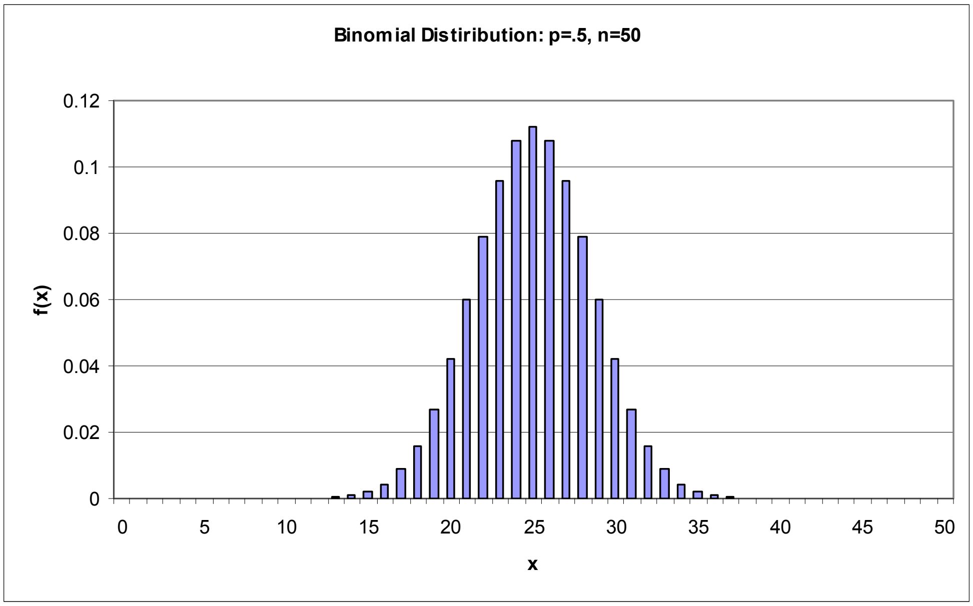
As mentioned in the previous section, the discipline that we now know as Statistics, developed from early work in a number of applied fields. It was, and is, very much an applied science. Gambling was undoubtedly one of the most important early drivers of research into probability and statistical methods and Abraham De Moivre's book, published in 1718, "The Doctrine of Chance: A method of calculating the probabilities of events in play" [DEM1] was essential reading for any serious gambler at the time. The book contained an explanation of the basic ideas of probability, including permutations and combinations, together with detailed analysis of a variety of games of chance, including card games with delightful names such as Basette and Pharaon (Faro), games of dice, roulette, lotteries etc. A typical entry in De Moivre's book is as follows:

"Suppose there is a heap of 13 cards of one color [suit], and another heap of 13 cards of another color; what is the Probability, that taking one Card at a venture [random] out of each heap, I shall take out the two Aces?" He then goes on to explain that since there is only one Ace in each heap, the separate probabilities are 1/13 and 1/13, so the combined probability (since the cards are independently drawn) is simply:

$$\frac{1}{13}\times\frac{1}{13}=\frac{1}{169}$$

hence the chance of not drawing two Aces is 168/169, or put another way, the *odds* against drawing two Aces are 168:1 - for gambling, whether the gambler or the gambling house, setting and estimating such odds is vitally important! De Moivre's book ran into many editions, and it was in the revised 1738 and 1756 editions that De Moivre introduced a series approximation to the Binomial for large *n* with *p* and *q* not small (e.g. not less than 0.3). These conditions lead to an approximation that is generally known as the Normal distribution. His motivation for so developing this approximation was that computation of the terms of the Binomial for large values of *n* (e.g. >50, as illustrated below) was extremely tedious and unrealistic to contemplate at that time. Furthermore, as *n* increases the individual events have very small probabilities (with *n*=500 the maximum probability for an individual event with *p*=0.5 is 0.036 - i.e. there is just under 4% chance of seeing exactly 250 heads, say, when 500 tosses of an unbiased coin are made). For this reason one tends to be interested in the probability of seeing a group or range of values (e.g. 400 or more heads from 500 tosses), rather than any specific value. Looking at the chart below the vertical bars should really be just vertical lines, and as the number of such lines becomes very large and the interval between events becomes relatively smaller, a continuous smooth bell-like curve approximation (which is what the Normal distribution provides) starts to make sense (see further, the Normal distribution).

Binomial distribution, mean = 25



De Moivre also worked extensively on another topic, mentioned in the previous section, mortality tables. This work developed following the publication by John Graunt in 1662 of figures on births and deaths in London, and similar research by Edmund Halley (the astronomer) of birth and deaths data for the City of Breslau (modern day Wrocław in Poland) between 1687 and 1691 [HAL1]. Halley was interested in using this data in order to "ascertain the price of annuities upon lives", i.e. to determine the level at which life insurance premiums (or *annuities*) might be set. As an illustration, Halley observed that (based on his data) there was only 100:1 chance that a man in Breslau aged 20 would die in the following 12 months (i.e. before reaching 21), but 38:1 if the man was 50 years old. De Moivre included Halley's data and sample annuity problems and solutions in the 1756 edition of his "Doctrine of Chance" book, cited above.

A very different application of statistics arose during the 19th century with the development of new forms of communication, especially the development of telephony and the introduction of manual and then mechanical exchange equipment. A Danish mathematician, Agner Erlang, working for the Copenhagen Telephone Authority (KTAS), addressed the important questions of queuing and congestion. Answers were needed to questions such as "how many operators are needed to service telephone calls for 1000 customers?" and "how many lines are required to ensure that 95% of our customers can call other major towns in the country without finding that the line is busy". Questions such as these are closely related to problems of queuing and queue management, such as "how many checkouts do I need in a supermarket to ensure customers on a busy Saturday do not have to wait in line more than a certain amount of time?", or "how long should we have a stop sign on red before we allow the traffic to cross an intersection?". Erlang investigated these questions by assuming that there are a large number of customers but only a small chance that any particular customer would be trying to make a call at any one time. This is rather like the Binomial with *n* large and *p* very small, which had been shown by the French mathematician, Siméon Poisson (in a work of 1837) to have a simple approximation, and is now given the name Poisson Distribution. Erlang also assumed that when a call was made, the call lengths followed an Exponential Distribution, so short calls were much more common than very long calls. In fact, this assumption is unnecessary - all that really matters is that the calls are made independently and have a known average duration over an interval of time, e.g. during the peak hour in the morning. The number of calls per hour made to the system times their average length gives the total *traffic*, in dimensionless units that are now called Erlangs and usually denoted by the letter *A* or *E*. Erlang derived a variety of statistical measures based on these assumptions, one of the most important being the so-called Grade of Service (GoS). This states the probability that a call will be rejected because the service is busy, where the traffic offered is *E* and the number of lines or operators etc available is *m*. The formula he derived, generally known as the Erlang B formula, is:

$$GoS = \frac{E^m / m!}{\sum_{k=0}^{m} E^k / k!}$$

Hence, if we have 2 units of traffic per hour (*E*=2) and *m*=5 channels to serve the traffic, the probability of congestion is expected to be just under 4%. Put another way, if you are designing facilities to serve a known peak traffic *E* and a target GoS of 5%, you can apply the formula incrementally (increasing *m* by 1 progressively) until you reach your target. Note that this very simple example assumes that there is no facility for putting calls into a queuing system, or re-routing them elsewhere, and critically assumes that calls arrive independently. In practice these assumptions worked very well for many years while telephone call traffic levels were quite low and stable over periods of 0.5-1.0 hours. However, with sudden increases in call rates people started to find lines busy and then called back immediately, with the result that call arrival rates were no long Poisson-like. This leads to a very rapidly degrading service levels and/or growing queuing patterns (familiar problems in physical examples such as supermarket checkouts and busy motorways, but also applicable to telephone and data communications networks). Erlang, and subsequently others, developed statistical formulas for addressing many questions of this type that are still used today. However, as with some other areas of statistical methods previously described, the rise of computational power has enabled entire systems to be simulated, allowing a range of complex conditions to be modeled and stress-tested, such as varying call arrival rates, allowing buffering (limited or unlimited), handling device failure and similar factors, which would have been previously impossible to model analytically.

The final area of application we shall discuss is that of experimental design. Research into the best way to examine the effectiveness of different treatments applied to crops led R A Fisher to devise a whole family of scientific methods for addressing such problems. In 1919 Fisher joined the Rothamsted Agricultural Experiment Station and commenced work on the formal methods of randomization and the analysis of variance, which now form the basis for the design of 'controlled' experiments throughout the world. Examples of the kind of problem his procedures address are: "does a new fertilizer treatment X, under a range of different conditions/soils etc, improve the yield of crop Y?" or "a sample of women aged 50-60 are prescribed one of three treatments: hormone replacement therapy (HRT); or a placebo; or no HRT for *x* years - does the use of HRT significantly increase the risk of breast cancer?".

As can be seen from these varied example, statistics is a science that has developed from the need to address very specific and practical problems. The methods and measures developed over the last 150-200 years form the basis for the many of the standard procedures applied today, and are implemented in the numerous software packages and libraries utilized by researchers on a daily basis. What has perhaps changed in recent years is the growing use of computational methods to enable a broader range of problems, with more variables and much larger datasets to be analyzed. The range of applications now embraced by statistics is immense. As an indication of this spread, the following is a list of areas of specialism for consultants, as listed by the websites of the UK Royal Statistical Society (RSS): and the US American Statistical Association (ASA):

Statistical Consultancy - Areas of specialism - RSS

Applied operational research	Epidemiology	Neural networks and genetic algorithms	Sampling
Bayesian methods	Expert systems	Non-parametric statistics	Simulation
Bioassay	Exploratory data analysis	Numerical analysis and optimization	Spatial statistics
Calibration	Forecasting	Pattern recognition and image analysis	Statistical computing
Censuses and surveys	GLMs and other non-linear models	Quality methodology	Statistical inference
Clinical trials	Graphics	Probability	Survival analysis
Design & analysis of experiments	Multivariate analysis	Reliability	Time Series

Statistical Consultancy - Areas of specialism - ASA

Bayesian Methods	General Advanced Methodological Techniques	Quality Management, 6-Sigma	Statistical Software - SAS
Biometrics & Biostatistics	Graphics	Risk Assessment & Analysis	Statistical Software - SPSS
Construction of Tests & Measurements	Market Research	Sampling & Sample Design	Statistical Training
Data Collection Procedures	Modeling & Forecasting	Segmentation	Survey Design & Analysis
Decision Theory	Non Parametric Statistics	Statistical Organization & Administration	Systems Analysis & Programming
Experimental Design	Operations research	Statistical Process Control	Technical Writing & Editing
Expert Witness	Probability	Statistical Software - other	Temporal & Spatial Statistics

References

[DEM1] De Moivre A (1713) The Doctrine of Chance: A method of calculating the probabilities of events in play; Available as a freely downloadable PDF from http://books.google.com/books?id=3EPac6QpbuMC

[HAL1] Halley E (1693) An Estimate of the Degrees of Mortality of Mankind. Phil. Trans. of the Royal Society, January 1692/3, p.596-610; Available online at http://www.pierre-marteau.com/editions/1693-mortality.html . Also available in Newman J R (1960) The World of Mathematics. Vol 3, Part VIII Statistics and the Design of Experiments, pp1436-1447. Oliver & Boyd, London

Part II

Statistical data

Statistical data

Statistics (plural) is the field of science that involves the collection, analysis and reporting of information that has been sampled from the world around us. The term *sampled* is important here. In most instances the data we analyze is a sample (a carefully selected representative subset) from a much larger *population*. In a production process, for example, the population might be the set of integrated circuit devices produced by a specific production line on a given day (perhaps 10,000 devices) and a sample would be a selection of a much smaller number of devices from this population (e.g. a sample of 100, to be tested for reliability). In general this sample should be arranged in such a way as to ensure that every chip from the population has an equal chance of being selected. Typically this is achieved by deciding on the number of items to sample, and then using equi-probable random numbers to choose the particular devices to be tested from the labeled population members. The details of this sampling process, and the sample size required, is discussed in the section Sampling and sample size.

The term *statistic* (singular) refers to a value or quantity, such as the mean value, maximum or total, calculated from a sample. Such values may be used to estimate the (presumed) *population* value of that statistic. Such population values, particular key values such as the mean and variance, are known as *parameters* of the pattern or *distribution* population values.

In many instances the question of what constitutes the population is not as clear as suggested above. When undertaking surveys of householders, the total population is rarely known, although an estimate of the population size may be available. Likewise, when undertaking field research, taking measurements of soil contaminants, or air pollutants or using remote sensing data, the population being investigated is often not so well-defined and may be infinite. When examining a particular natural or man made *process*, the set of outcomes of the process may be considered as the population, so the process outcomes are effectively the population.

Since statistics involves the analysis of data, and the process of obtaining data involves some kind of measurement process, a good understanding of measurement is important. In the subsections that follow, we discuss the question of measurement and measurement scales, and how measured data can be grouped into simple classes to be produce data distributions. Finally we introduce two issues that serve to disguise or alter the results of measurement in somewhat unexpected ways. The first of these is the so-called statistical grouping affect, whereby grouped data produce results that differ from ungrouped data in a non-obvious manner. The second of these is a spatial effect, whereby selection of particular arrangement of spatial groupings (such as census districts) can radically alter the results one obtains.

Measurement

In principle the process of measurement should seek to ensure that results obtained are consistent, accurate (a term that requires separate discussion), representative, and if necessary independently reproducible. Some factors of particular importance include:

- **framework** - the process of producing measurements is both a technical and, to an extent, philosophical exercise. The technical framework involves the set of tools and procedures used to obtain and store numerical data regarding the entities being measured. Different technical frameworks may produce different data of varying quality from the same set of entities. In many instances measurement is made relative to some internationally agreed standard, such as the meter (for length) or the kilogram (for mass). The philosophical framework involves the notion that a meaningful numerical value or set of values can be assigned (using some technical framework) to attributes of the entities. This is a model or representation of these entity attributes in the form of numerical data - a person's height is an attribute that we can observe visually, describe in words, or assign a number to based on an agreed procedure relative to a standard (in meters, which in turn is based on the agreed measurement of the speed of light in a vacuum)

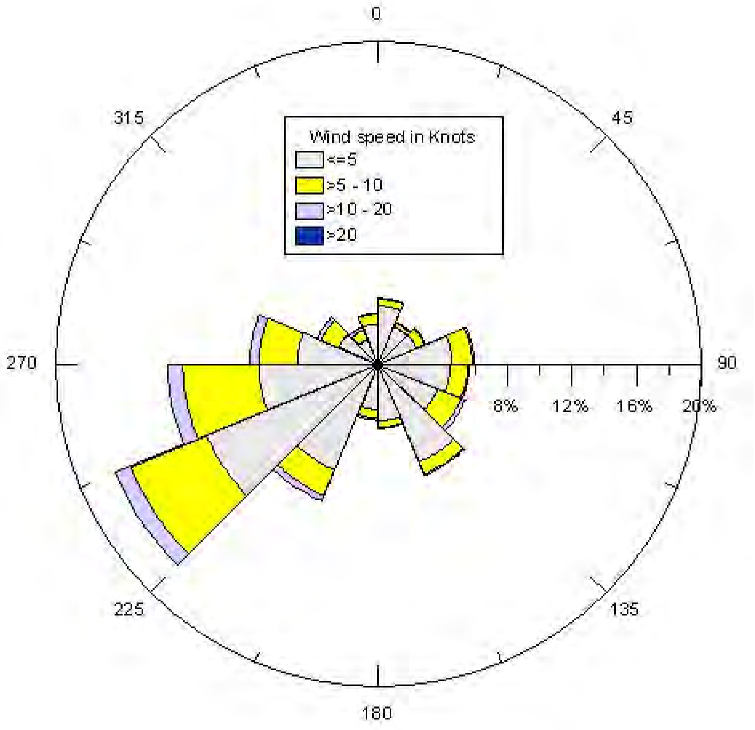
- **metrics** - when measuring distance in the plane using Euclidean measure the results are invariant under translation, reflection and rotation. So if we use Euclidean measure we can safely make measurements of distances over relatively small areas and not worry about the location or orientation at which we took the measurements and made the calculation. However, over larger areas and/or using a different metric, measurements may fail the invariance test. In the case of measurements that seek to compute distances, measurements made using the so-called City block or Manhattan distance are not invariant under rotation. Likewise, Euclidean distance measurements give incorrect results over larger distances on the Earth's surface (e.g. >20 kilometers). When making other forms of measurement similar issues apply (e.g. the effect of the local gravitational field on weight, the local magnetic field on magnetic flux, etc.)
- **temporal effects** - measurement made at different times of the day, days of the year and in different years will inevitably differ. If the differences are simply random fluctuations in a broadly constant process (results are unaffected by temporal translation of the data) the process is described as being *stationary*. If a trend exists (which could be linear, cyclical or some other pattern) the process is said to be *non-stationary*. All too often consideration of the temporal aspect of measurement is omitted, e.g. a person's height will be measured as shorter in the evening as compared with the morning, a persons academic or sporting achievement can be significantly affected by when they were born (see Gladwell, 2008, for an extensive discussion of this issue, [GLA1]) - the issue is always present even if it is not of direct concern. Frequently the sequence of event measurement is important, especially where humans are doing the measurements or recordings, since issues such as concentration become important over time; event sequences may also be explicitly monitored, as in control charts, time series analysis and neural network learning
- **spatial effects** - measurements made at different locations will typically exhibit spatial variation. If all locations provided identical data the results would be a spatially uniform distribution. If the results are similar in all directions at all locations, then the process is described as *isotropic* (i.e. results are rotationally invariant). If the results are similar at all locations (i.e. the results are translationally invariant) then the process can be described as stationary. In practice most spatial datasets are non-stationary
- **observer effects** - in both social and pure science research, observer effects can be significant. As a simple example, if we are interested in measuring the temperature and air quality in a process clean room, the presence of a person taking such measurements would inevitably have some affect on the readings. Similarly, in social research many programmes can display the so-called Hawthorne Effect in which changes (often improvements) in performance are partially or wholly the result of behavioral changes in the presence of the observer (reflecting greater interest in the individuals being observed)

Measurement scales

Measurement gives rise to values, such as counts, sets of decimal values, binary responses (yes/no, presence/absence) etc., which may be of different types (scales). The principal scales encountered are:

- **Nominal** (or Categorical): data is really just assignment to named classes, such as Red, Blue, Green or Utah, Nevada, New York...An attribute is nominal if it successfully distinguishes between groups, but without any implied ranking or potential for arithmetic. For example, a telephone number can be a useful attribute of a place, but the number itself generally has no numeric meaning. It would make no sense to add or divide telephone numbers, and there is no sense in which the number 9680244 is more or better than the number 8938049. Likewise, assigning arbitrary numerical values to classes of land type, e.g. 1=arable, 2=woodland, 3=marsh, 4=other is simply a convenient form of naming (the values are still nominal)
- **Ordinal**: this term refers to data values that involves a concept of order, from least to greatest and may include negative numbers and 0. A set of apparently ordered categories such as: 1=low, 2=medium, 3=high, 4=don't know does not form an ordinal scale. An attribute is ordinal if it implies a ranking, in the sense that Class 1 may be better than Class 2, but as with nominal attributes arithmetic operations do not make sense, and there is no implication that Class 3 is worse than Class 2 by the precise amount by which Class 2 is worse than Class 1. An example of an ordinal scale might be preferred locations for residences - an individual may prefer some areas of a city to others, but such differences between areas may be barely noticeable or quite profound. Analysis of nominal and ordinal data is often qualitative, or uses visualization techniques to highlight interesting patterns, and may use non-parametric statistical methods especially when count data are available
- **Interval**: numeric data that exhibits order, plus the ability to measure the interval (distance) between any pair of objects on the scale (e.g. $2x-x\equiv 3x-2x$). Data are interval if differences make sense, as they do for example with measurements of temperature on the Celsius or Fahrenheit scales, or for measurements of elevation above sea level
- **Ratio**: interval plus a natural origin, e.g. temperature in degrees Kelvin, weights of people (i.e. so $x=2y$ is meaningful); Interval or ratio scales are required for most forms of (parametric) statistical analysis. Data are ratio scaled if it makes sense to divide one measurement by another. For example, it makes sense to say that one person weighs twice as much as another person, but it makes no sense to say that a temperature of 20 Celsius is twice as warm as a temperature of 10 Celsius, because while weight has an absolute zero Celsius temperature does not (but on an absolute scale of temperature, such as the Kelvin scale, 200 degrees can indeed be said to be twice as warm as 100 degrees). It follows that negative values cannot exist on a ratio scale.
- **Cyclic**: modulo data - like angles and clock time. Measurements of attributes that represent directions or cyclic phenomena have the awkward property that two distinct points on the scale can be equal - for example, 0 and 360 degrees. Directional data are cyclic (see the sample *wind rose* diagram below) as are calendar dates. Arithmetic operations are problematic with cyclic data, and special techniques are needed to handle them. For example, it makes no sense to average 1° and 359° to get 180°, since the average of two directions close to north clearly is not south. Mardia and Jupp (1999, [MAR1]) provide a comprehensive review of the analysis of directional or cyclic data

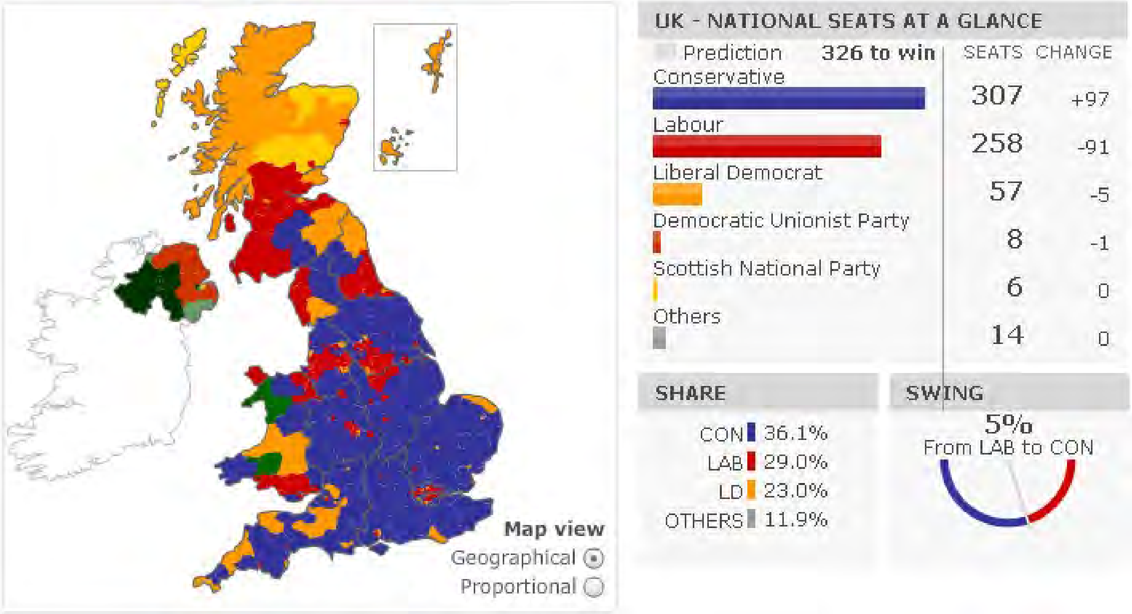
Cyclic data — Wind direction and speed, single location



Bar charts, Histograms and Frequency distributions

- **Bar chart:** The process of measurement may produce data that are recorded as counts and assigned to purely nominal classes, for example counts of different bird species in a woodland. In this instance a simple bar chart may be produced to illustrate the different relative frequencies of each species. Each class is assigned an individual vertical or horizontal bar and typically each bar being the same width (so height indicates relative frequency). Bars are separated by distinct gaps and the order in which the bars are placed on the horizontal or vertical axis is of no importance. The example below (upper diagram) shows the results of the UK parliamentary election in May 2010. The bar chart indicates the seats one in the "first past the post" system used currently in the UK, with a geographic map of the spread of these results.

BBC UK Election 2010 results



UK - NATIONAL SEATS AT A GLANCE

Prediction 326 to win	SEATS	CHANGE
Conservative	307	+97
Labour	258	-91
Liberal Democrat	57	-5
Democratic Unionist Party	8	-1
Scottish National Party	6	0
Others	14	0

SHARE

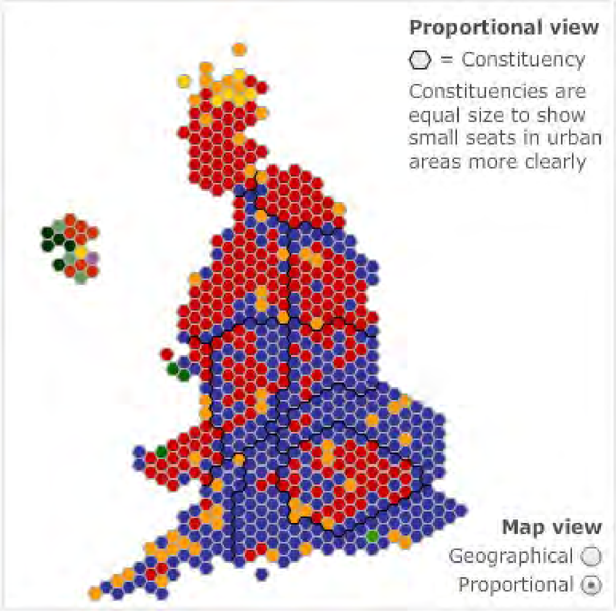
CON	36.1%
LAB	29.0%
LD	23.0%
OTHERS	11.9%

SWING

5%
From LAB to CON

Map view
Geographical ●
Proportional ○

source: *http://news.bbc.co.uk/1/shared/election2010/results/*



Proportional view

⬡ = Constituency

Constituencies are equal size to show small seats in urban areas more clearly

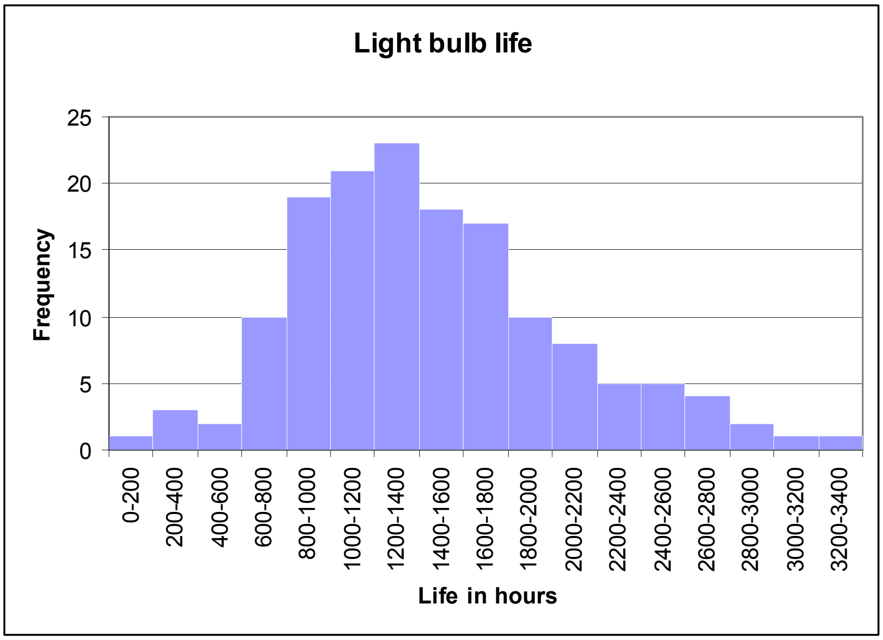
Map view
Geographical ○
Proportional ●

source: http://news.bbc.co.uk/1/shared/election2010/results/. Note that the first diagram is misleading as it does not reflect the density of voters, suggesting the seats obtained by parties in large rural areas are somehow more significant than densely populated urban areas. This view can be corrected in various ways, most simply by adjusting the areas to reflect the populations in each. The result is a proportional map, with areas distorted but still recognizable, as illustrated in the second diagram.

- **Histogram**: If measurements yield numerical values on an interval or ratio scale, these can be grouped into classes and the counts (or frequencies) in each class plotted as a bar chart in which the order on the horizontal axis (or *x*-axis) is important. A bar chart of this type is called a *histogram* and should be plotted without spaces between the vertical bars reflecting the continuous nature of the scale (see example of light bulb life data, below). The term histogram was introduced by Karl Pearson in the late 19th century to describe any chart of this type, especially charts in which the horizontal axis represented time. He liked the idea that the Greek word *histos*, which means anything placed vertically, like a ship's mast, is similar to the word *historical*, giving the idea of a frequency chart with a time-based *x*-axis..

- **Frequency distribution**: A frequency distribution is a tabulated set of sample data, showing the number of occurrences of events or observations that fall into distinct classes or that have particular values. As such, it can be seen as a convenient way of avoiding the need to list every data item observed separately. However, frequency distributions can often provide greater insight into the pattern of sample values, and enables these patterns to be compared with well-understood standard distributions, such as the Binomial (discrete) and Normal (continuous) distribution. A simple example is shown in the table below together with the chart (or histogram) of the data. In this table there are 17 equal interval classes, for each of which the number of light bulbs in a sample of *N*=150 that fail after a certain time are listed.

Length of life of electric light bulbs - tabulated and histogram

Life (hours)	Frequency	Histogram of frequencies
0-200	1	
200-400	3	
400-600	2	
600-800	10	
800-1000	19	
1000-1200	21	
1200-1400	23	
1400-1600	18	
1600-1800	17	
1800-2000	10	
2000-2200	8	
2200-2400	5	
2400-2600	5	
2600-2800	4	
2800-3000	2	
3000-3200	1	
3200-3400	1	
Total	150	

after Pearson E S (1933, [PEA1])

Several observations should be made about this particular frequency distribution:

(i) it has a single category or *class* containing the most frequent bulb life (1200-1400hrs) - this category is called the mode, and because there is a single mode, the distribution is said to be *unimodal*

(ii) the set of classes in the tabulated list are not really correctly defined - the boundaries are indeterminate, and should be specified as [0,199.9],[200-399.9], etc (or similar) or better still [0,<200], [200,<400] etc (in Pearson's paper, which was primarily concerned with production control and sampling, he actually only supplied the frequency diagram, not the tabulated data) - the precise definition of the boundaries of classes avoids the problem of deciding how to assign values that lie on the boundary (e.g. a bulb with measured lifespan of exactly 200 hours)

(iii) each class is the same width (duration) and every data value is allocated to a unique class; however, when performing certain calculations, such as computing the mean value, a decision has to be made as to whether to use the recorded frequencies in the various classes or *bins*, or the source data (if available). If the frequencies have to be used, it is necessary to define a representative value for each interval, which is usually taken to be the mid-interval value. Note that this assumption hides the within-class variation in values which may create some errors in computations, especially if the class widths are large. The question of bin selection is discussed later in this section

(iv) the width (duration) of each class is somewhat arbitrary and this choice significantly affects the form of the frequency distribution. If the class width was very small (1 hour say) most classes would contain the frequency 0, and a few would contain just 1 failure. At the opposite extreme, if the class width was 3400 hours all the results would be in just the one class. In both these examples very little information would be gained from inspecting the pattern of frequencies. Selecting the class boundaries and number of classes is an important operation - it should ensure that the minimum of information is lost, whilst also ensuring that the distribution communicates useful and relevant information. Many authors recommend the use of an odd number of classes, and there are a myriad of rules-of-thumb for choosing the number of classes and class boundaries (see Class Intervals, below)

(v) all the data fits into the classes (in this example). This is often not possible to achieve with equal interval classes, especially at the upper and lower ends of the distribution. Indeed, frequency distributions with very long tails are common, and often the final category is taken as 3000+ for example

(vi) the data being analyzed in this example can be regarded as a continuous variable (lifespan of the bulb) and is a single variable (i.e. univariate data)

There are several extensions and variations that can be applied to the above model. The first is to rescale the vertical axis by dividing each class value by the total sample size (N=150), in which case the data are described as *relative frequencies*, and in examples such as this, the values can be considered as *estimated probabilities*.

A second important variant is the extension of the frequency table and chart to multivariate and multi-dimensional cases. In the bivariate case the data may simply be separate measures applied to the same classes, or they may be joint measures. For example, suppose that our classes show the heights of individuals in a large representative sample. The first column of a bivariate frequency tabulation might show the frequency distribution for men over 18 years, whilst the second column shows the same data but for women. However, if the mix of those sampled included fathers and sons, one could construct a two-way or *joint* frequency distribution (or *cross-tabulation*) of the men with classes "Tall" and "Short", where Tall is taken as over some agreed height. The table below illustrates such a cross-tabulation, based on a study of families carried out by Karl Pearson and Dr Alice Lee from 1893 onwards:

Cross-tabulation of father-son height data

	Father short	Father tall	Total fathers
Son short	250	89	339
Son tall	215	446	661
Total sons	465	535	1000

simplified, after K Pearson and A Lee (1903, Table XXII [PEA2]; the overall sample size of 1000 families and the cell entries are simply a proportional reduction from the 1078 cases in the original data).

In this example each part of the frequency distribution is divided into just 2 classes, but each could readily have been separated into 3 or more height bands. Indeed, the original table is divided into 20 rows and 17 columns (illustrated in full in the Probability section of this Handbook), but inevitably many of the table entries are blank. Row and column totals have been provided, and these are sometimes referred to as *marginal frequencies* or *marginal distributions*. They are essentially the univariate frequency distributions for the rows and columns taken separately.

As with the univariate frequency data, this table could be converted to relative frequencies by dividing through by 1000, but it also affords another perspective on the data; we can consider questions such as: "what is the probability that a tall son has a tall father?" If the data are truly representative of the population of fathers and sons, then the estimated probability is 446/1000 or 44.6%. But when we examine the table, we find that there are far more tall fathers and tall sons than short fathers and short sons. We could then ask "does this estimate of probability suggest that tall fathers have tall sons, i.e. some genetic or other relationship factor?". Overall we can see from the *totals* entries that 53.5% of our sample fathers are tall and 66.1% of the sons are tall, and if these two groups were completely independent we might reasonably expect 53.5% x 66.1% of the father-son combinations to be tall (applying the rule of multiplication for independent probabilities). But this combination is actually only 35.4%, so the 44.6% finding does suggest a relationship, but whether it is significant (i.e. highly unlikely to be a chance result) requires more careful analysis using a particular statistical technique, contingency table analysis. Cross-classifications of this kind do not require numeric classes or classes derived from numeric values as in this example - in many instances the rows contain classes such as "Success, Failure" or "Survived, Died" and the columns might contain "Treatment A, Treatment B, Placebo, No treatment", with the table entries providing a count of the number of plants, patients etc. recorded in that combination of classes. In general such multivariate classification tables are restricted to 2-way, and occasionally 3-way analysis, and rarely are the number of classes in each dimension of the classification large if analyzed in this manner - often they are 5 or less.

Frequency distributions can also by multi-dimensional. For example, the distribution of cases of a particular disease around a point source of contamination might be measured in distance bands and radial sectors around this location. This pattern might then be compared with a known bivariate frequency distribution, such as the bivariate Normal distribution. In three dimensions one could be looking at the distribution of bacteria in cheese, or the distribution of stars in a region of space.

Class intervals, bins and univariate classification

If sampled data are measurements of a continuous variable, *x*, such as the light bulb lifespans described above, then the standard procedure in frequency chart (or histogram) production is to create a set of equal width class intervals (or *bins*) and count the frequencies occurring in each interval. The values at which the bins are separated are often referred to as *cut-points*. The number of intervals to be used is a matter for the researcher to determine, depending on the problem requirements. It is often helped, in interactive software packages, by viewing a display of the resulting histogram as different options are selected. For visualization purposes it is desirable to limit the number of classes to between 5 or 9, as using large numbers of classes (20+) can be difficult to display and interpret with clarity, and an odd number of intervals will ensure there is a central class. On the other hand, with a large set of observations that exhibit considerable spread across the range, a larger number of classes may be more helpful and will avoid the problem of having much of the real variation hidden by large class intervals.

There are several rules of thumb for determining the ideal number of bins and/or the width for fixed-width bins for real-valued continuous data. These include the following (*n* is the number of observations or data items to be grouped, *k* is the number of classes, *h* is the bin width, *s* is the standardized spread or standard deviation of the sample data):

$$k=\left\lceil \frac{\max-\min}{h}\right\rceil,\ \text{or}\ h=\left\lceil \frac{\max-\min}{k}\right\rceil$$

These options use the range and a pre-selected bin width to define the number of bins, *k*, or alternatively the number of bins is specified and the range used to determine the bin width, *h*. Note that if the distribution has a very long tail, e.g. a few data items that are very much larger or smaller than all the others, these formulas will produce excessively wide bins.

The next formula is due to Scott (1979, [SCO1]) and uses the standard deviation of the dataset, *s*, rather than the range to determine bin width:

$$h=\left\lceil 3.5s/n^{1/3}\right\rceil$$

Thus for 1000 data items with a standard deviation of 25, *h*=9. The number of bins still remains to be chosen, and this will be a matter of choice again, but could safely use the range calculation for *k*, above, in most cases. Scott's model is built on an analysis of the optimal properties of a binning arrangement with constant bin widths and an examination of the ideas of so-called kernel density estimation (KDE) techniques. The latter use all the data points to create a smooth estimated probability distribution (or probability density function, which has been shown to produce excellent results but may require a considerable amount of data processing.

As mentioned earlier, if the frequencies are to be used in computations it is necessary to define a representative value for each interval, which is usually taken to be the mid-interval value. Thus if the bin width is *h*, and the mid-interval value is x_i, the interval has a range from x_i-*h*/2 to x_i+*h*/2. This assumption hides the within-interval variation in values which may create some errors in computations, especially if the class width are large. The so-called Sheppard's correction, named after its author William Sheppard (1897), is an adjustment to estimates of the variance when (Normally distributed) fixed width bins are used. Without correction the computations tend to over-estimate the variance since they effectively treat all values in a range as the same as the mid-value. Sheppard's correction to the variance is -h^2/12, an amount that is the variance of the Uniform distribution defined over an interval of width, *h*.

The table below provides details of a number of univariate classification schemes together with comments on their use. Such schemes are essentially a generalization of fixed-width binning. Many statistical software packages provide classification options of the types listed, although some (such as the box, Jenks and percentile methods) are only available in a limited number of software tools.

The scheme described in the table as Natural breaks or Jenks' method is an automated procedure utilizing the following algorithm:

Step 1: The user selects the attribute, *x*, to be classified and specifies the number of classes required, *k*

Step 2: A set of *k*-1 random or uniform values are generated in the range [min{*x*},max{*x*}]. These are used as initial class boundaries or 'cut points'

Step 3: The mean values for each initial class are computed and the sum of squared deviations of class members from the mean values is computed. The total sum of squared deviations (TSSD) is recorded

Step 4: Individual values in each class are then systematically assigned to adjacent classes by adjusting the class boundaries to see if the TSSD can be reduced. This is an iterative process, which ends when improvement in TSSD falls below a threshold level, i.e. when the within class variance is as small as possible and between class variance is as large as possible. True optimization is not assured. The entire process can be optionally repeated from Step 1 or 2 and TSSD values compared

Univariate binning/classification schemes

Classification scheme	Description/application
Unique values	Each value is treated separately - this is effectively a nominal data classification model
Manual classification	The analyst specifies the boundaries between classes/bins as a list, or specifies a lower bound and interval or lower and upper bound plus number of intervals required. This approach is widely used in statistical software packages
Equal interval	The attribute values are divided into *n* classes with each interval having the same width=range/*n*
Exponential interval	Intervals are selected so that the number of observations in each successive interval increases (or decreases) exponentially
Equal count or quantile	Intervals are selected so that the number of observations in each interval is the same. If each interval contains 25% of the observations the result is known as a quartile classification. Ideally the procedure should indicate the exact numbers assigned to each class, since they will rarely be exactly equal
Percentile	In the standard version equal percentages (percentiles) are included in each class, e.g. 20% in each class. In some implementation of percentile plots (specifically designed for exploratory data analysis, EDA) unequal numbers are assigned to provide classes that, for example, contain 6 intervals: <=1%, >1% to <10%, 10% to <50%, 50% to <90%, 90% to <99% and >=99%
Natural breaks/Jenks	Used within some software packages, these are forms of variance-minimization classification. Breaks are typically uneven, and are selected to separate values where large changes in value occur. May be significantly affected by the number of classes selected and tends to have unusual class boundaries. Typically the method applied is due to Jenks, as described in Jenks and Caspall (1971, [JEN1]), which in turn follows Fisher (1958, [FIS1]). Very useful for visualization work, but unsuitable for comparisons
Standard deviation (SD)	The mean and standard deviation of the data are calculated, and values classified according to their deviation from the mean (z-transform). The transformed values are then grouped into classes, usually at intervals of 1.0 or 0.5 standard deviations. Note that this often results in no central class, only classes either side of the mean and the number of classes is then even. SD classifications in which there is a central class (defined as the mean value +/- 0.5SD) with additional classes at +/- 1SD intervals beyond this central class, are also used

Classification scheme	Description/application
Box	A variant of quartile classification designed to highlight outliers, due to Tukey (1977, Section 2C, [TUK1]). Typically six classes are defined, these being the 4 quartiles, plus two further classifications based on outliers. These outliers are defined as being data items (if any) that are more than 1.5 times the inter-quartile range (IQR) from the median. An even more restrictive set is defined by 3.0 times the IQR. A slightly different formulation is sometimes used to determine these box ends or hinge values

Supervised binning and classification

Some statistical software packages differentiate between *unsupervised* and *supervised* schemes. These terms have different meanings within different packages and application areas, which can be confusing. In broad terms an unsupervised method utilizes the data directly, whereas a supervised method cross-refers the sample data to some other dataset that is already divided into a number of distinct classes or categories. It then uses this other dataset to guide (or supervise) the classification process.

In SPSS, for example, supervised (or *optimal*) binning refers to a procedure in which the source data is divided into bins using cut-points that seek to minimize the mix of a separate, but linked, nominal variable in each bin. For example, the variable to be binned might be household income in $000s p.a., and the supervisor or control variable might be the level of education achieved by the main earner of the household. The principal technique used, known as MDLP, starts by placing every (sorted) data item (observation) into a single large bin. The bin is then divided using cut-points, and the mix of the linked nominal variable in each bin is examined (using an Entropy or Diversity statistic). If every entry in the bin has the same linked nominal category then the Entropy measure will be 0, which is regarded as optimal with respect to the nominal variable. On the other hand if there is a large mix of nominal variables represented, of roughly equal numbers, the bin will have a higher Entropy score. The algorithm adjusts the cut points and increases the number of cut pints (and hence bins) to achieve an improvement in the total Entropy of the binning process.

In remote-sensing applications (for example, multi-spectral satellite imagery) the task is to classify individual image pixels into groups, which may be pre-defined (e.g. land use categories, such as Forest, Grasslands, Buildings, Water etc) or derived from the data. Unsupervised classification in this instance refers to the use of wholly automated procedures, such as K-means clustering, in order to group similar pixels. Supervised classification refers to a multi-stage process, in which the dataset is compared to a reference dataset that has already been classified, and the similarity between pixels in the dataset to be classified and the reference set is used as a means for achieving the 'best' classification. Clearly procedures such as this, which arise in a number of disciplines, essentially belong in the realm of multivariate data classification, which may or may not use statistical techniques and measures as part of that process.

Scale and arrangement

In the preceding subsections we have seen that determining the number and size of bins can be a quite complicated exercise. It was noted that with too many bins only frequencies of 1 and 0 would be recorded, whereas with very few bins, almost all the variation in the data would be hidden within the bin, or class, with little or no variation detectable between classes. This is often the exact opposite of the ideal classification or grouping schemes, where the aim is generally to minimize within-class variance as compared to between class variance - making sure that classes or groupings are as homogeneous as possible. Two additional, and somewhat unexpected factors, come into play when such groupings are made. These are known as the *statistical effect* and the *arrangement effect*.

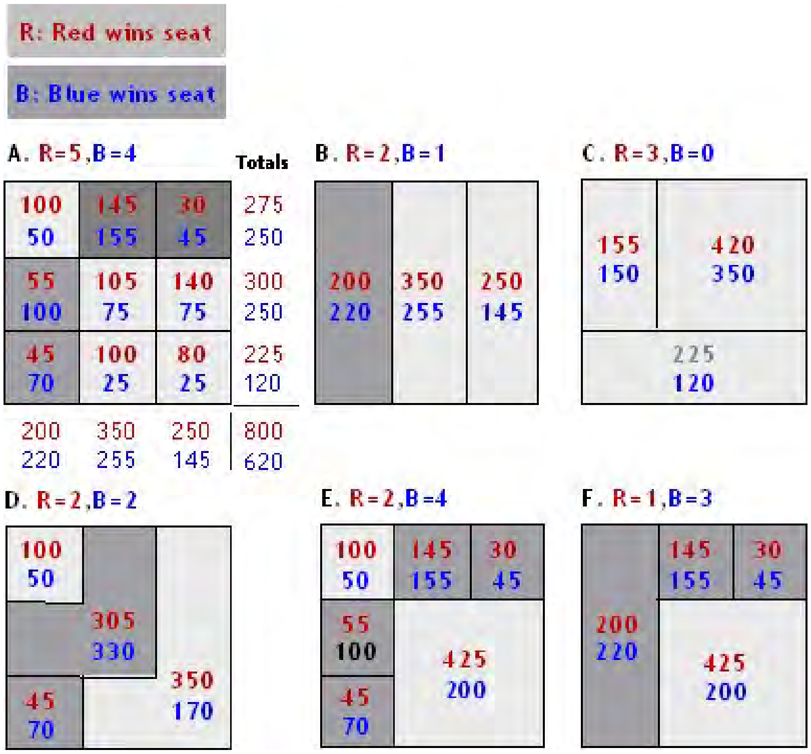
To understand the statistical effect (which is a scale or grouping effect) look at the regional employment statistics shown in the Table below (after de Smith *et al*. (2009, [DES1]). Areas A and B both contain a total of 100,000 people who are classified as either employed or not. In area A 10% of both Europeans and Asians are unemployed (i.e. equal proportions), and likewise in Area B we have equal proportions (this time 20% unemployed). So we expect that combining areas A and B will give us 200,000 people, with an equal proportion of Europeans and Asians unemployed (we would guess this to be 15%), but it is not the case - 13.6% of Europeans and 18.3% of Asians are seen to be unemployed! The reason for this unexpected result is that in Area A there are many more Europeans than Asians, so we are working from different total populations.

Regional employment data - grouping effects

	Employed (000s)	Unemployed (000s)	Total (000s) (Unemployed %)
Area A			
European	81	9	90 (10%)
Asian	9	1	10 (10%)
Total	90	10	100 (10%)
Area B			
European	40	10	50 (20%)
Asian	40	10	50 (20%)
Total	80	20	100 (20%)
Areas A and B			
European	121	19	140 (13.6%)
Asian	49	11	60 (18.3%)
Total	170	30	200 (15%)

There is a further, less well known problem, which has particular importance in the process of elections and census data collection but also has much wider implications. This is due to the way in which voting and census areas are defined. Their shape, and the way in which they are aggregated, affects the results and can even change which party is elected. The Grouping Data diagram below illustrates this issue for an idealized region consisting of 9 small voting districts. The individual zone, row, column and overall total number of voters are shown in diagram A, with a total of 1420 voters of whom roughly 56% (800) will vote for the Red party (R) and 44% (620) for the Blue party (B). With 9 voting districts we expect roughly 5 to be won by the Reds and 4 by the Blues, as is indeed the case in this example. However, if these zones are actually not the voting districts themselves, but combinations of the zones are used to define the voting areas, then the results may be quite different. As diagrams B to F show, with a voting system of “first past the post” (majority in a voting district wins the district) then we could have a result in which every district was won by the Reds, to one in which 75% of the districts were won by the Blues. So it is not just the process of grouping that generates confusing results, but also the pattern of grouping. We are rarely informed of the latter problem, although it is one that is of great interest to those responsible for defining and revising electoral and census district boundaries.

Grouping Data - Zone arrangement effects on voting results



This is not just a problem confined to voting patterns and census data. For example, suppose the information being gathered relates to the average levels of lead and zinc in the soil within each field. Samples based on different field boundaries would show that in some arrangements the average proportion of lead in the soil exceeded that of zinc, whilst other arrangements would show the opposite results. .

References

[DES1] de Smith M J, Goodchild M F, Longley P A (2009) Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools. 3rd edition, Troubador, Leicester. Available from: http://www.spatialanalysisonline.com/

[FIS1] Fisher W D (1958) On grouping for maximal homogeneity. J. of the American Statistical Association, 53, 789-98

[GLA1] Gladwell M (2008) Outliers - the story of success. Alan Lane/Penguin, London

[JEN1] Jenks G F, Caspall F C (1971) Error on choroplethic maps: Definition, measurement, reduction. Annals of American Geographers, 61, 217-44

[MAR1] Mardia K V, Jupp P E (1999) Directional statistics. 2nd ed., John Wiley, Chichester

[PEA1] Pearson E S (1933) A Survey of the Uses of Statistical Method in the Control and Standardization of the Quality of Manufactured Products. J. Royal Stat. Soc., 96,1, 21-75

[PEA2] Pearson K, Lee A (1903) On the Laws of Inheritance in Man: I. Inheritance of Physical Characters. Biometrika, 2(3), 357-462

[SCO1] Scott D W (1979) On optimal and data-based histograms. Biometrika 66,3, 605–610

[TUK1] Tukey J W (1977) Exploratory data analysis. Addison-Wesley, Reading, MA

The Statistical Method

Many people would regard statistical analysis as a purely technical exercise involving the application of specialized data collection and analysis techniques, but this perception is both incorrect and misleading. Statistical problems should be viewed within the context of a broad methodological framework, and it is the specific nature of this framework that defines "The Statistical Method". Here we are using the terminology and interpretation of MacKay and Oldford (2000, [MAC1]). They carefully examined the nature of statistical analysis by discussing the problem of determining the speed of light, as conducted in the experiments of A A Michelson in 1879. Although they used research that involved a relatively complicated experiment as their example, the conclusions they draw are much more wide-reaching. Essentially they argue that statistical analysis must involve a broad perspective on the task under consideration, from the initial Problem definition stage (P), through Planning and Data collection stages (P,D) through to Analysis (A) and Conclusions (C). This is similar to the "statistical problem solving cycle" as described in the Probability & Statistics leaflet mentioned in our Suggested Reading section and elsewhere, but widens the scope of this methodology.

The elements of this methodological framework are shown in the PPDAC table below - each is discussed in detail in their paper. MacKay and Oldford note that very often the complexity of the analysis phase is greatly reduced if the totality of a problem is addressed in the manner described. As can be seen, the formal analysis stage comes well down the sequence of steps that are involved in producing good quality statistical research. Absolutely crucial to the entire process is the initial problem definition. Only once this is thoroughly understood by all interested parties can a plan for data collection be devised and the data obtained for subsequent analysis.

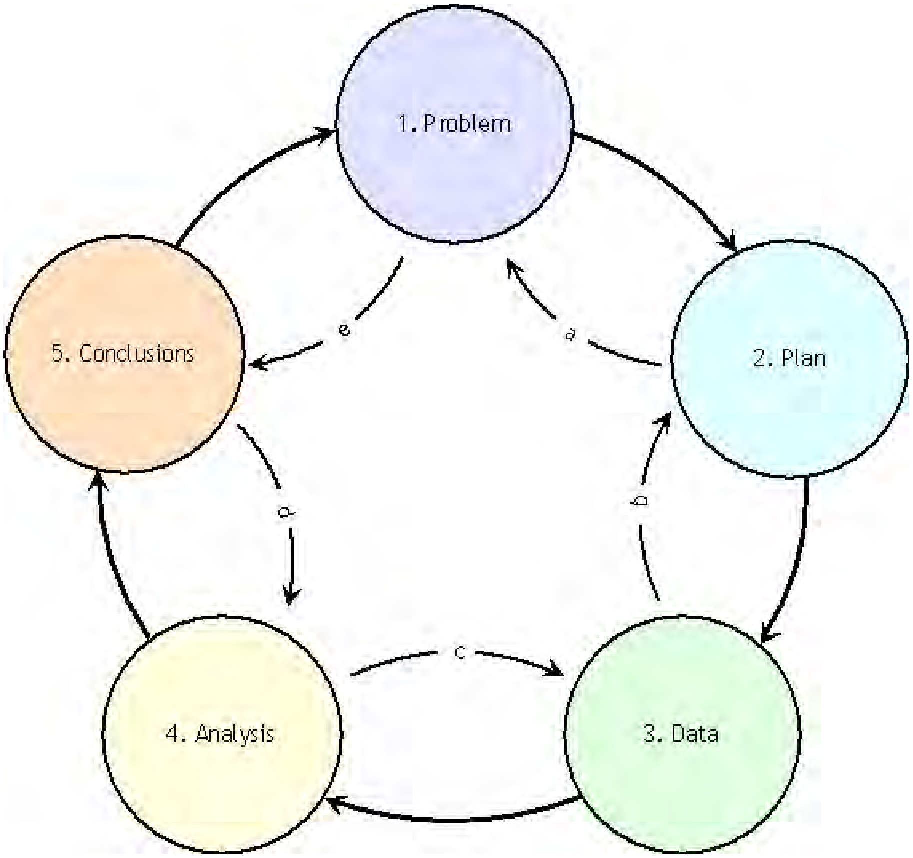
PPDAC: The Statistical Method, after MacKay and Oldford (2000)

	Details	Michelson experiment
Problem	Units & Target Population (Process) Response Variate(s) Explanatory Variates Population Attribute(s) Problem Aspect(s) - causative, descriptive, predictive	Unit: One (measured) light transmission. Population: all such transmissions Response variate: the speed of light in each measured transmission Explanatory variates: a large number of possible factors that might help explain variations in the measured data (e.g. method used, the measurement process) Population attributes: the average speed of light in a vacuum Problem aspect: descriptive (seeking an estimate of a specific value)
Plan	Study Population (Process) (Units, Variates, Attributes) Selecting the response variate(s) Dealing with explanatory variates Sampling Protocol Measuring processes Data Collection Protocol	Study population: The collection of units that could possibly be measured (known as the sampling frame in survey work). Michelson measured the speed of light in air, not in a vacuum - the difference between the study population and the true population is known as the *study error* Response variates: Michelson measured the speed of light indirectly, using distances, rotation speed (of a mirror), timing device (tuning forks) and temperature Explanatory variates: There may be a large number. Where possible Michelson tried to fix those factors he was aware of, and measure or vary others to check if they had an effect on his results Sampling protocol: The detailed procedure followed for sampling the data - in Michelson's case he made sets of measurements one hour after sunrise and one hour before sunset, on a series of days close to mid-summer. He made 1000 measurements, with some made by an independent observer Measuring processes: the equipment, people, and methods used - *measurement error* , which is the difference between the measured value and the true value, is incurred in this step of the procedure Data Collection Protocol: the management and administration (recording etc) of the entire data collection exercise - nowadays this would include data storage and processing considerations
Data	Execute the Plan and record all departures	Execution: Michelson did not record every result, but just the average values for blocks of 10 measurements

	Data Monitoring Data Examination for internal consistency Data storage	Data monitoring: Tracking data as they are obtained helps identify patterns, temporal drift, outliers etc. Michelson did not explicitly do this Data examination: The internal consistency of the data should be checked, for unexpected features (each using EDA techniques), but Michelson did not appear to do this Data storage: simple tabulated results on paper in this instance
Analysis	Data Summary numerical and graphical Model construction build, fit, criticize cycle Formal analysis	In Michelson's case he summarized his data in tables and computed the average of his 100 measured velocities in air, and then corrected for the deflection effect that air would have on his results, making a small adjustment for temperature variations in each case. Formal analysis was limited to analyzing possible source of error and their maximum impact on the results, in order to obtain an estimate of the velocity of light in a vacuum, +/- the estimated errors
Conclusions	Synthesis plain language, effective presentation graphics Limitations of study discussion of potential errors	Michelson presented his central finding and provided a full discussion as to possible sources of error and why many factors could be ignored due to the manner in which the plan was made and executed. Despite this, the true value for the speed of light is actually outside the limits of his estimates at the time, even though his mean result was within 0.05% of the correct figure, hence he slightly underestimated the size of the errors affecting his result

The PPDAC summary table suggests a relatively linear flow from problem definition through to conclusions - this is typically not the case. It is often better to see the process as cyclical, with a series of feedback loops. A summary of a revised PPDAC approach is shown in the diagram below. As can be seen, although the clockwise sequence (1→5) applies as the principal flow, each stage may and often will feed back to the previous stage. In addition, it may well be beneficial to examine the process in the reverse direction, starting with Problem definition and then examining expectations as to the format and structure of the Conclusions (without pre-judging the outcomes!). This procedure then continues, step-by-step, in an anti-clockwise manner (e→a) determining the implications of these expectations for each stage of the process.

PPDAC as an iterative process



We now expand our discussion by examining the components this revised model in a little more detail:

Problem: Understanding and defining the problem to be studied is often a substantial part of the overall analytical process - clarity at the start is obviously a key factor in determining whether a programme of analysis is a success or a failure. Success here is defined in terms of outcomes (or objectives) rather than methods. And outcomes are typically judged and evaluated by third parties - customers, supervisors, employers - so their active involvement in problem specification and sometimes throughout the entire process is essential. Breaking problems down into key components, and simplifying problems to focus on their essential and most important and relevant components, are often very effective first steps. This not only helps identify many of the issues to be addressed, likely data requirements, tools and procedures, but also can be used within the iterative process of clarifying the customer's requirements and expectations. Problems that involve a large number of key components tend to be more complex and take more time than problems which involve a more limited set. This is fairly obvious, but perhaps less obvious is the need to examine the interactions and dependencies between these key components. The greater the number of such interactions and dependencies the more complex the problem will be to address, and as the numbers increase complexity tends to grow exponentially. Analysis of existing information, traditionally described as "desk research", is an essential part of this process and far more straightforward now with the advantage of online/Internet-based resources. Obtaining relevant information from the client/sponsor (if any), interested third parties, information gatekeepers and any regulatory authorities, forms a further and fundamental aspect to problem formulation and specification. Box *et al*. (2005, p13,

[BOX1]) suggest a series of questions that should be asked, particularly in the context of conducting experiments or trials, which we list below with minor alterations from their original. As can be seen, the questions echo many of the issues we raise above:

- what is the objective of this investigation?
- who is responsible?
- I am going to describe your problem - is my description correct?
- do you have any past data? and if so, how were these data collected/in what order/on what days/by whom/how?
- do you have any other data like these?
- how does the equipment work/what does it look like/can I see it?
- are there existing sampling, measurement and adjustment protocols?

Plan: Having agreed on the problem definition the next stage is to formulate an approach that has the best possible chance of addressing the problem and achieving answers (outcomes) that meet expectations. Although the PLAN phase is next in the sequence, the iterative nature of the PPDAC process emphasizes the need to define and then re-visit each component. Thus whilst an outline project plan would be defined at this stage, one would have to consider each of the subsequent stages (DATA, ANALYSIS, CONCLUSIONS) before firming up on the detail of the plan. With projects that are more experimental in nature, drawing up the main elements of the PLAN takes place at this stage. With projects for which pre-existing datasets and analysis tools are expected to be used, the PLAN stage is much more an integrated part of the whole PPDAC exercise. The output of the PLAN stage is often formulated as a detailed project plan, with allocation of tasks, resources, times, analysis of critical path(s) and activities, and estimated costs of data, equipment, software tools, manpower, services etc. Frequently project plans are produced with the aid of formal tools, which may be paper-based or software assisted. In many instances this will involve determining all the major tasks or task blocks that need to be carried out, identifying the interconnections between these building blocks (and their sequencing), and then examining how each task block is broken down into sub-elements. This process then translates into an initial programme of work once estimated timings and resources are included, which can then be modified and fine-tuned as an improved understanding of the project is developed. In some instances this will be part of the Planning process itself, where a formal functional specification and/or pilot project forms part of the overall plan. As with other parts of the PPDAC process, the PLAN stage is not a one-shot static component, but typically includes a process of monitoring and re-evaluation of the plan, such that issues of timeliness, budget, resourcing and quality can be monitored and reported in a well-defined manner. The approach adopted involves consideration of many issues, including:

- the nature of the problem and project — is it purely investigative, or a formal research exercise; is it essentially descriptive, including identification of structures and relationships, or more concerned with processes, in which clearer understanding of causes and effects may be required, especially if predictive models are to be developed and/or prescriptive measures are anticipated as an output?
- does it require commercial costings and/or cost-benefit analysis?
- are particular decision-support tools and procedures needed?
- what level of public involvement and public awareness is involved, if any?

- what particular operational needs and conditions are associated with the exercise?
- what time is available to conduct the research and are there any critical (final or intermediate) deadlines?
- what funds and other resources are available?
- is the project considered technically feasible, what assessable risk is there of failure and how is this affected by problem complexity?
- what are the client (commercial, governmental, academic, personal) expectations?
- are there specifications, standards, quality parameters and/or procedures that must be used (for example to comply with national guidelines)?
- how does the research relate to other studies on the same or similar problems?
- what data components are needed and how will they be obtained (existing sources, collected datasets)?
- are the data to be studied (units) to be selected from the target population, or will the sample be distinct in some way and applied to the population subsequently (in which case, as discussed earlier, one must consider not just *sampling error* but *study error* also)?

When deciding upon the design approach and analytical methods/tools it is often important to identify any relevant available datasets, examine their quality, strengths and weaknesses, and carry out exploratory work on subsets or samples in order to clarify the kind of approach that will be both practical and effective. There will always be unknowns at this stage, but the aim should be to minimize these at the earliest opportunity, if necessary by working through the entire process, up to and including drafting the presentation of results based on sample, hypothetical or simulated data.

Data: In research projects that involve experiments, the data are collected within the context of well-defined and (in general) tightly controlled circumstances, with the response and explanatory variates being clearly included in the design of the experiment. In many other instances data is obtained from direct or indirect observation of variates that do not form part of any controlled experiment. And in survey research, although there will be a carefully constructed sample design, the level of direct control over variates is typically very limited. Key datasets are also often provided by or acquired from third parties rather than being produced as part of the research. Analysis is often of these pre-existing datasets, so understanding their quality and provenance is extremely important. It also means that in many instances this phase of the PPDAC process involves selection of one or more existing datasets from those available. In practice not all such datasets will have the same quality, cost, licensing arrangements, availability, completeness, format, timeliness and detail. Compromises have to be made in most instances, with the over-riding guideline being fitness for purpose. If the datasets available are unsuitable for addressing the problem in a satisfactory manner, even if these are the only data that one has to work with, then the problem should either not be tackled or must be re-specified in such a way as to ensure it is possible to provide an acceptable process of analysis leading to worthwhile outcomes. A major issue related to data sourcing is the question of the compatibility of different data sets: in formats and encoding; in temporal, geographic and thematic coverage; in quality and completeness. In general datasets from different sources and/or times will not match precisely, so resolution of mismatches can become a major task in the data phase of any project. And as part of this process the issue of how and where to store the data arises, which again warrants early consideration, not merely to ensure consistency and retrievability but also for convenient analysis and reporting. Almost by definition no dataset is perfect. All may contain errors, missing values, have a finite resolution, include distortions as a result modeling the real world with discrete mathematical forms, incorporate measurement errors and uncertainties, and may exhibit deliberate or designed adjustment of data (e.g. for privacy reasons, as part of aggregation procedures).

Analysis: The Analysis phase can be seen as a multi-part exercise. It commences with the review of data collected and the manipulation of the many inputs to produce consistent and usable data. Exploratory data analysis (EDA), including the production of simple data summaries, tabulations and graphs is typically the first stage of any such analysis. The impact on research of exceptions - rare events, outliers, extreme values, unusual clusters - is extremely important. Exploratory methods, such as examining individual cases and producing box-plots, help to determine whether these observations are valid and important, or require removal from the study set. This phase then extends into more formal study in order to identify patterns of various kinds that help the researcher to develop new ideas and hypotheses regarding form and process. And this in turn may lead on to the use or development of one or more models within a formal build-fit-criticize cycle. Crawley (2007, p339, [CRA1]) provides the following extremely sound advice regarding model selection (echoing a quote attributed to George Box):

*"It is as well to remember the following truths about models: all models are wrong; some models are better than others [*Box said *more useful]; the correct model can never be known with certainty; and the simpler a model the better it is"!*

Finally the output of the models and analysis is examined, and where necessary the dataset and data gathering plan is re-visited, working back up the PPDAC model chain, prior to moving on to producing the output from the project and delivering this in the Conclusion stage. The application of a single analytical technique or software tool is often to be avoided unless one is extremely confident of the outcome, or it is the analytical technique or approach itself that is the subject of investigation, or this approach or toolset has been specifically approved for use in such cases. If analysis is not limited to single approaches, and a series of outputs, visualizations, techniques and tests all suggest a similar outcome then confidence in the findings tends to be greatly increased. If such techniques suggest different outcomes the analyst is encouraged to explain the differences, by re-examining the design, the data and/or the analytical techniques and tools applied. Ultimately the original problem definition may have to be reviewed.

Conclusions: The last stage of the PPDAC process is that of reaching conclusions based upon the analyses conducted, and communicating these to others. Note that implementation of findings (e.g. actually proceeding with building a bypass, designating an area as unfit for habitation, or implementing a vaccination programme) does not form part of this model process, but lies beyond its confines.

“The purpose of the Conclusion stage is to report the results of the study in the language of the Problem. Concise numerical summaries and presentation graphics [tabulations, visualizations] should be used to clarify the discussion. Statistical jargon should be avoided. As well, the Conclusion provides an opportunity to discuss the strengths and weaknesses of the Plan, Data and Analysis especially in regards to possible errors that may have arisen” Mackay and Oldford (2000)

For many problems this summary is sufficient. For others the conclusions stage will be the start of additional work: re-visiting the problem and iterating the entire process or parts of the process; a new project; implementing proposals; and/or wider consultation. In Michelson's case, he was aware of several imperfections in his research, and in fact spent the rest of his life conducting a series of further experiments in order to progressively improve the accuracy of his estimate of the true speed of light. A full discussion of this revised PPDAC model in the context of spatial analysis is provided in the "Chapter 3: Spatial analysis and the PPDAC model" of de Smith *et al.*, 2009 [DES1] which is available online.

References

[BOX1] Box G E P,Hunter J S, Hunter W G (1978, 2005) Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building. J Wiley & Sons, New York. J Wiley & Sons, New York. The second, extended edition was published in 2005

[CRA1] Crawley M J (2007) The R Book. J Wiley & Son, Chichester, UK

[DES1] de Smith M J, Goodchild M F, Longley P A (2009) Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools. 3rd edition, Troubador, Leicester. Available from: http://www.spatialanalysisonline.com/

[MAC1] MacKay R J, Oldford R W (2000) Scientific Method, Statistical Method and the Speed of Light. Statist. Sci., 15, 3, 254-278. Available from: http://projecteuclid.org/euclid.ss/1009212817

Wikipedia, Speed of Light article: http://en.wikipedia.org/wiki/Speed_of_light

Misuse, Misinterpretation and Bias

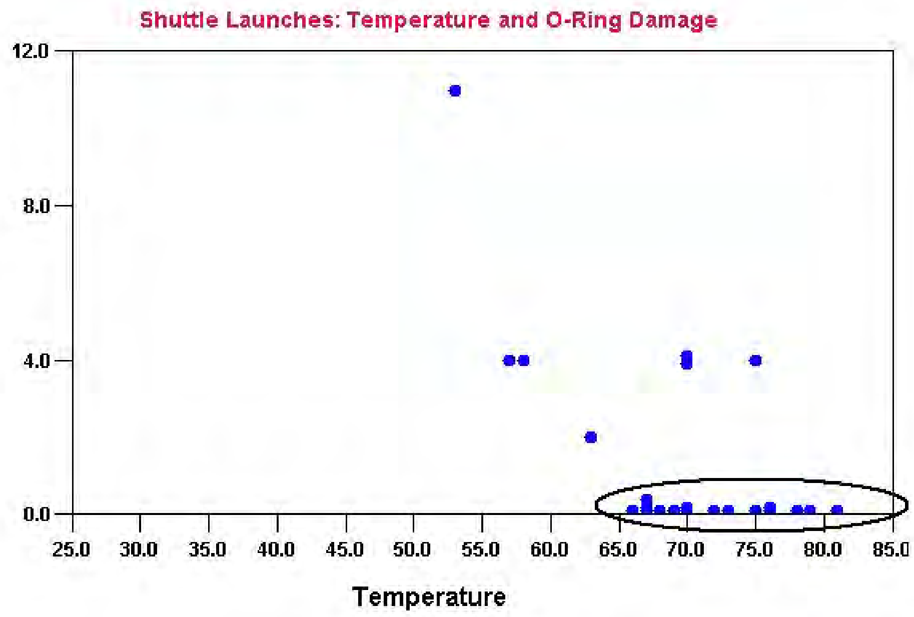
A great deal has been written about the misuse of statistics by pressure groups and politicians, by pollsters and advertising campaigns, by the broadcast media (newspapers, magazines, television, and now the Internet), and even misuse by statisticians and scientists. In some instances the misuse has been simply lack of awareness of the kinds of problems that may be encountered, in others carelessness or lack of caution and review, whilst on occasion this misuse is deliberate. One reason for this has been the growth of so-called *evidence-based* policy making - using research results to guide and justify political, economic and social decision-making. Whilst carefully designed, peer-reviewed and repeatable research does provide a strong foundation for decision-making, weak research or selective presentation of results can have profoundly damaging consequences. In this section we provide guidance on the kinds of problems that may be encountered, and comment on how some of these can be avoided or minimized. The main categories of misuse can be summarized as:

- inadequate or unrepresentative data
- misleading visualization of results
- inadequate reasoning on the basis of results

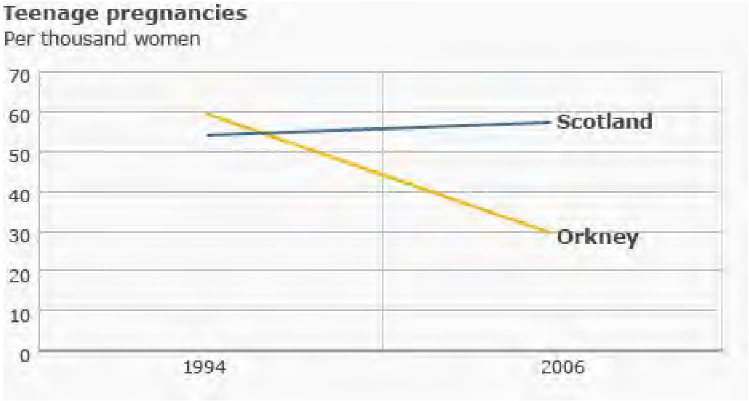
In the subsections of this topic we discuss each of these categories in turn.

Where data is obtained as the result of some form of trial, experiment or survey, careful design can help avoid many (but not all) of the problems identified in the first category (see also Design of Experiments and Bias). This is of particular importance in medical research, and for this reason we have included a separate subsection focusing on this particular application area and the kinds of problems and issues that are encountered.

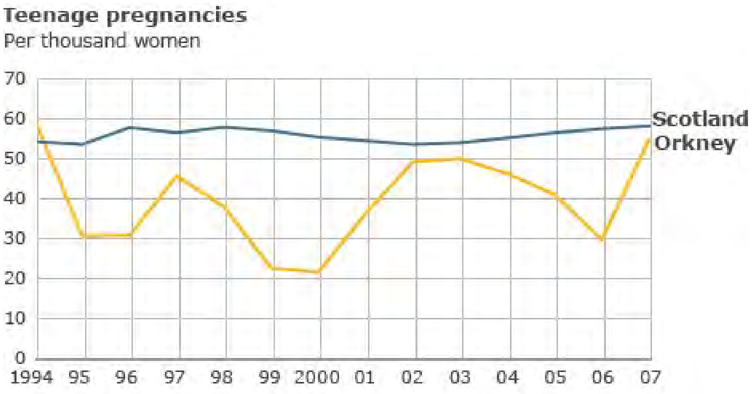
A simple example, which occurs only too frequently, is the presentation and interpretation of data where some data items are omitted. A much reported example of this concerned the analysis of the failure of O-rings on the US space shuttle in 1986. NASA staff and their contractors examined the pattern of failures of O-rings during launches against temperature just prior to the ill-fated shuttle launch on January 28 1986. They concluded that the data showed no apparent relationship between the number of failures and temperature, but as we now know, the low temperature overnight did result in a failure of these components (see graph below) with catastrophic results. What the analysts failed to consider were all those launches that had 0 failures. All the launches with no failures occurred when the ambient temperature at the launch site was much higher, as highlighted in the diagram (see also, the SpaceShuttle dataset and example in the R library, vcd).



In a rather different context highlighted in Jan 2010 by BBC journalist Michael Blastland (see also, our Recommended Reading topic, [BLA1]) reports of declining teenage pregnancy rates, in Orkney off the north coast of Scotland, were shown to be highly misleading. Blastland showed two graphs. The first appears to show a halving of the teenage pregnancy rate between 1994 and 2006, following an intensive programme of education and support:



However, the reports omitted data for the intervening years, and as we know from stock market and many other types of data, rates of change depend very heavily on your start and end date. The data in this case is actually quite cyclical, and choosing 2006 rather than, say 2007, provides a completely misleading picture, as the graph below demonstrates.



In many instances misuse is not deliberate, but leads to biased results and conclusions that cannot be relied upon and the consequences can be very serious. Each of the following subsections examines these issues in more detail.

Our final example concerns the question of independent sampling. On 2nd February 2010 a UK national newspaper, the Daily Mail, reported the story of a woman who had bought a box of 6 eggs and found that every one contained a double-yolk. They argued that because roughly 1 egg in a thousand has a double yolk, the chances of having a box with every one being double-yolks was one in a quintillion (1 in 10^{18}). It was clearly a crazy statement that assumed the occurrence of multiple yolks in a box of eggs represented a set of independent events, and that it was therefore valid to multiply 1:1000 x 1:1000 etc 6 times. In fact the events are in no way independent, for a whole variety of reasons. One respondent to a discussion about this example pointed out that most eggs are boxed in large sorting and packing warehouses, and in some cases eggs are checked against a strong light source to see if they contain a double yolk. If they do, they are put to one side and the staff often take these home for their own use, but if there are too many they are simply boxed up, resulting in boxes of double-yolk eggs.

Inadequate or unrepresentative data

This is probably the most common reason for 'statistics' and statistical analysis falling short of acceptable standards. Problems typically relate to inadequacies in sampling, i.e. in the initial design of the data collection, selection or extraction process. This results in the sample, from which inferences about the population are made, being biased or simply inadequate. The following list includes some of the main situations which lead to such problems:

- **datasets and sample sizes** - there are many situations where the dataset or sample size analyzed is simply too small to address the questions being posed, or is not large enough for use with the proposed statistical technique, or is used in a misleading fashion. Smaller sample sizes are also more prone to bias from missing data and non-responses in surveys and similar research exercises. For example, when examining the incidence of particular diseases recorded in different census districts (or hospital catchment areas etc) we might find that for some diseases recorded cases were quite low in rural districts (<10), but were much higher in urban districts (>100). Does this mean the disease is more likely to occur amongst urban dwellers? Not necessarily, as there are more urban dwellers. To remove the effect of differences in the *population-at-risk* we might decide to compute the incidence (or *rate*) of the disease per 1000 population in each district (perhaps stratified by age and sex). Because of the relatively low population-at-risk in the rural area this might then show the risk appears much higher in the rural areas. Is the risk really higher or is the result a reflection of the relatively small numbers reported? Is reporting of cases for this disease the same in rural and urban areas, or is there some differential in recording perhaps due to differences in the quality of health care available or for social reasons? For a rare disease, a reported 25% increase year-on-year in the incidence of a particular type of cancer in the rural district might simply be the result of an increase of a single new reported case. It is also important to be aware that small samples tend to be much more variable in *relative* terms than large samples. This can result in errors in reasoning, as we discuss later in this section. (see also: Sampling and sample size)

- **clustered sampling** - this issue relates to the collection of data in a manner that is known in advance to be biased, but is not subsequently adjusted for this bias. Examples include the deliberate decision to over-sample minority social groups because of expected lower response rates or due to a need to focus on some characteristic of these groups which is of particular interest - see, for example, the discussion of this issue by Brogan (1998, [BRO1]). A second example applies where the only available data is known to be clustered (in space and/or time) - for example, in order to obtain estimates of the levels of trace elements in groundwater it is often only possible to take samples from existing wells and river courses, which are often spatially clustered. If the samples taken are not subsequently weight-adjusted (or *de-clustered*) results may be biased because some groups or areas are sampled more than others
- **self-selection and pre-screening** - this is a widespread group of problems in sampling and the subsequent reporting of events. Surveys that invite respondents to participate rather than randomly selecting individuals and ensuring that the resulting survey sample is truly representative are especially common. For example, surveys that rely on opting in, such as those placed in magazines, or via the Internet, provide a set of data from those who read the publication or view the Internet site, which is a first category of selection, and from this set the individuals who choose to respond are then self-selecting. This group may represent those with a particular viewpoint, those with strong views (so greater polarization of responses) or simply those who have the time and inclination to respond. Likewise, a survey on lifestyle in the population at large that advertises for participants in a range of lifestyle magazines, or in fitness studios and sports clubs, is likely to result in a significantly biased sample of respondents
- **exclusions** - the process of research design and/or sampling may inadvertently or deliberately exclude certain groups or datasets. An example is the use of telephone interviewing, which effectively pre-selects respondents by telephone ownership. If the proportion of exclusions is very small (e.g. in this example, the current proportion of people with telephones in a given country may be very high) this may not be a significant issue. A different category of exclusion is prevalent where some data is easier to collect than others. For example, suppose one wishes to obtain samples of bacteria in the soil of a study region. Areas which are very inaccessible may be under-sampled or omitted altogether whilst other areas may be over-sampled. In a different context, surveys of individuals may find that obtaining an ethnically representative sample is very difficult, perhaps for social or language reasons, resulting in under-representation or exclusion of certain groups - groups such as the disabled or very young or very old are often inadvertently excluded from samples for this reason. Limitations of time and/or budget are often factors that constrain the extent and quality of data collection and hence relevant and important data may be excluded for reasons of necessity or expediency. Data may also be deliberately or inadvertently excluded as being probably an error or outlier. In May 1985 the existence of the huge 'ozone hole' over the Antarctic (depleted