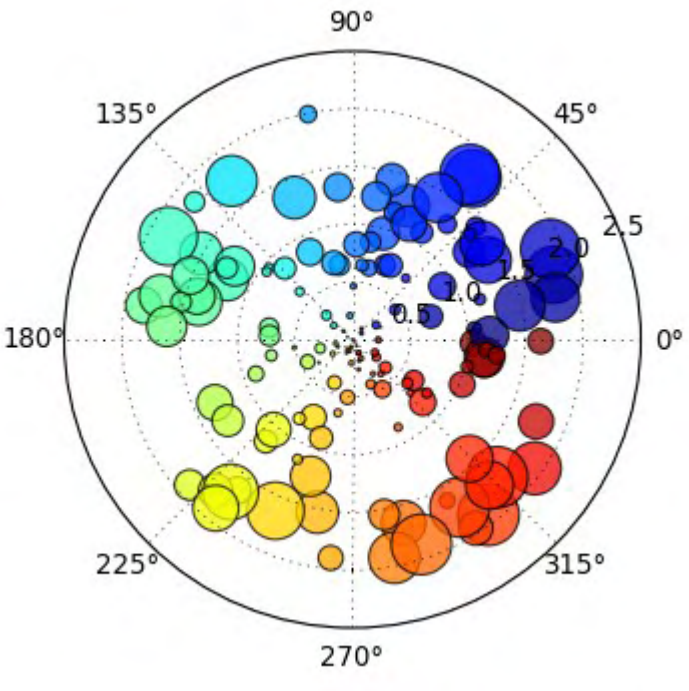
# Statistical Analysis

## A comprehensive handbook of statistical concepts, techniques and software tools

---

*Dr M J de Smith*



# Statistical Analysis

## A comprehensive handbook of statistical concepts, techniques and software tools

---

*by Dr M J de Smith*

# Statistical Analysis

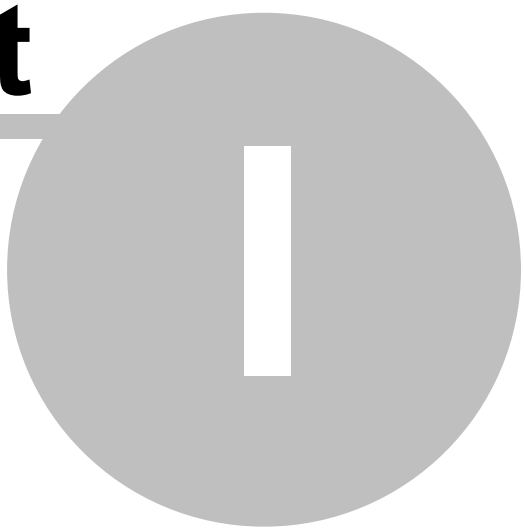


Front cover image: Polar bubble plot (source: MatPlotLib library, Python)

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

# Table of Contents

# Part



# 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 $x_1, x_2, x_3, \ldots x_n$, 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 $X_1, X_2, X_3, \ldots X_n$, 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 $w(x_i)$. 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 $\sum f_i=1$ |
| $\sum$ | 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 |
| $\cup$ | Set union. The notation $P(A \cup B)$ is used to indicate the probability of A *or* B |

| Item | Description |
|---|---|
| $\Delta$ | 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 |
| $\int$ | 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 |
| $\prod$ | 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 |
| $\rightarrow$ | Tends to, typically applied to indicate the limit as a variable tends to 0 or $\infty$ |
| $\bar{}$ | 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 $y \sim N(0,1)$ 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 ${}^{n}C_{r}$, or similar, and shorthand for $n!/[(n-r)!r!]$. |
| $\equiv$ | 'Equivalent to' symbol |
| $\approx$ | 'Approximately equal to' symbol |
| $\propto$ | Proportional to |
| $\in$ | '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 |
| $\leq$ | Less than or equal to, represented in the text where necessary by <= (provided in this form to support display by some web browsers) |
| $\geq$ | Greater than or equal to, represented in the text where necessary by >= (provided in this form to support display by some web browsers) |

| Item | Description |
|---|---|
| $\lfloor x \rfloor$ | 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 |

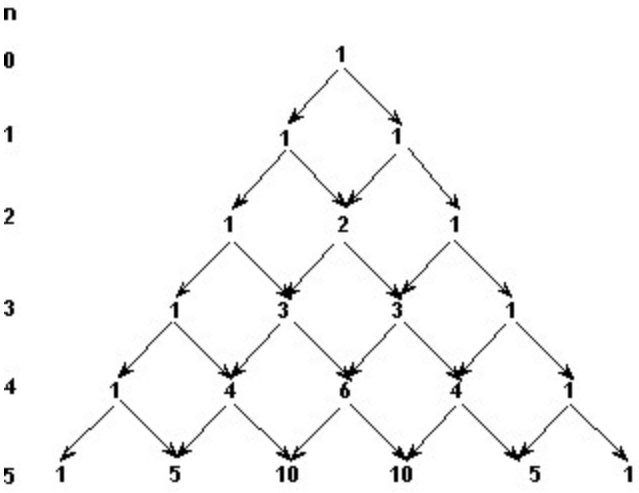
# References

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:

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

This formula is typically expanded as:

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

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 $r \geq m$, 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 $q=1-p=5/6$ ):

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

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 behaviour, whether this be medical research data, social surveys, analysis of purchasing behaviour 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 behaviour 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

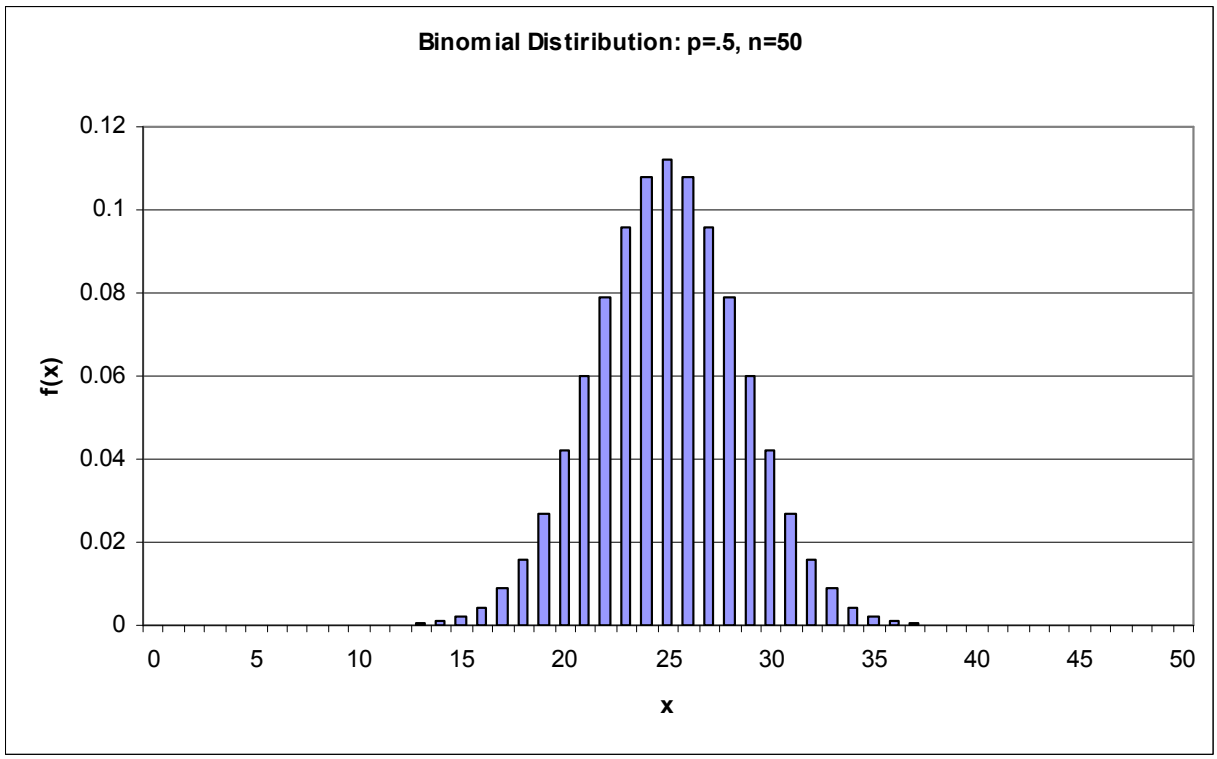
As mentioned in the previous section, the discipline that we now know as Statistics, developed from early work in a number of applied fields. It was, and is, very much an applied science. Gambling was undoubtedly one of the most important early drivers of research into probability and statistical methods 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)

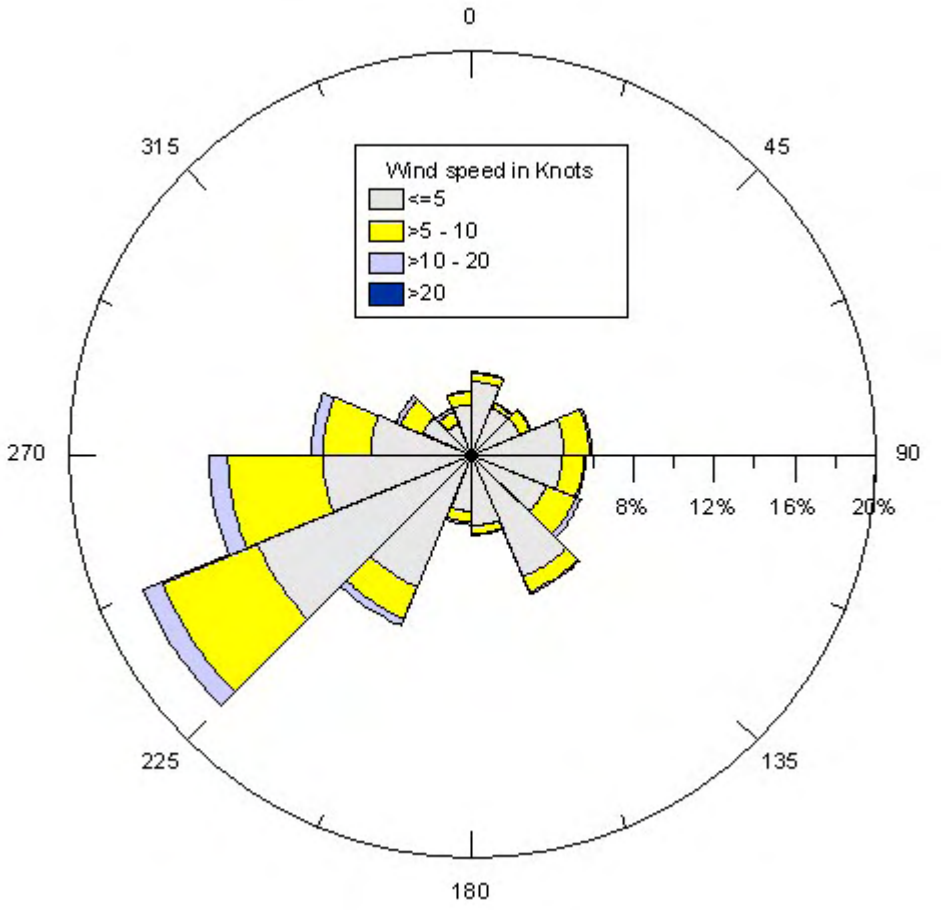
## Measurement scales

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 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. Note that the latter is highly misleading as