McGill University

Department of Epidemiology
and Biostatistics

EPIB – 607

Principles of Inferential Statistics in Medicine

Lawrence Joseph

Please Note

The following pages contain a copy of the overhead slides used in the course "Principles of Inferential Statistics in Medicine", in the Department of Epidemiology and Biostatistics at McGill University.

Accordingly, this booklet is not a textbook in biostatistics, and in fact much of the material will become clear only during the lectures.

Table of Contents

Introduction 1

Data Summaries and Descriptive Statistics 19

Probability and Probability Distributions 37

Inferences Concerning Means 77

Inferences Concerning Proportions 133

Nonparametric Statistics 157

Regression and Correlation 170

Assignments 219

Principles of Inferential Statistics in Medicine – EPIB-607 – 4 credits

Instructor: Lawrence Joseph
Email address: Lawrence.Joseph@mcgill.ca (best way to reach me)
Home page: http://www.epi.mcgill.ca/Joseph/
Telephone: 934-1934 X 44713
Address: Division of Clinical Epidemiology
Montreal General Hospital
1650 Cedar Avenue
Room L10 509
H3G 1A4

Course Objectives: To provide consumers/producers of biomedical research with basic principles of statistical inference applicable to clinical and epidemiologic research so that they can: (i) understand how statistical methods are used by others, (ii) apply them in their own research (iii) use them as a base for more advanced biostatistics courses.

Content: See detailed two page outline, pages 5 and 6.

Place and Time: September 4 to December 9, 2003. Tuesdays 9:00–11:00 AM (Room 1/12 Anatomy and Dentistry Building) and Thursdays 9:00–11:00 AM (Room Room 1/12 Anatomy and Dentistry Building).

Assessment: Assignments 5 × 2% each = 10%, Midterm Examination = 30%, Project = 15%, Final Examination = 45%.

Please note that both exams are open book.

Textbooks:

- Moore D and McCabe G. Introduction to the Practice of Statistics, Fourth Edition. 2002. Freeman and Company.
- Armitage P, Berry G and Matthews J. Statistical Methods in Medical Research, Fourth Edition. 2001. Blackwell Scientific Publications.
- Colton T. Statistics in Medicine. 1974. Little Brown.

Equipment: Scientific hand calculator (with square root, log and exponential functions).

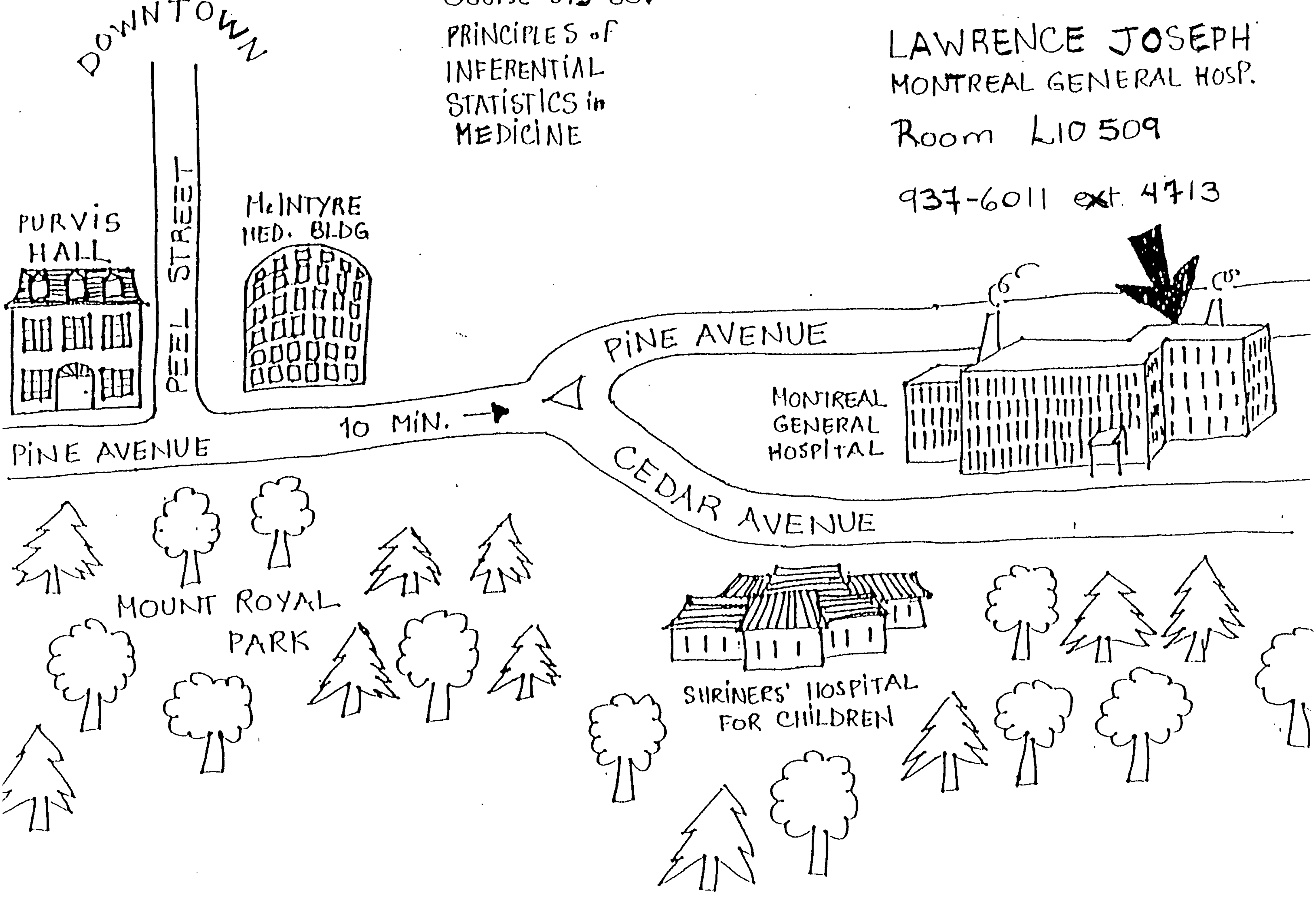
Prerequisites: Differential and integral calculus.

Midterm Exam: Thursday October 23, 9:00 AM to 11:00 AM, Room 1/12.
Final Exam: Tuesday December 9, 9:00 AM to 12:00 PM, Room TBA.

Course 513-607

PRINCIPLES of INFERENTIAL STATISTICS in MEDICINE

LAWRENCE JOSEPH
MONTREAL GENERAL HOSP.
Room L10 509
937-6011 ext. 4713



Principles of Inferential Statistics in Medicine

Other textbooks of interest

- B. Rosner. Fundamentals of Biostatistics. Duxbury 1994. *Another basic book on biostatistics, with lots of examples.*
- Michael Oaks, Statistical Inference. Epidemiology Resources, 1990. *Excellent overview of the meaning of statistical tests, and different schools of statistical inference.*
- G. Friedman, A Primer of Epidemiology. McGraw-Hill, 1974. *A quick introduction to epidemiology.*
- Bailar, J.C. and Mosteller F. (Eds.) Medical uses of Statistics. *From the NEJM series on statistics.*
- Moore D. Statistics: Concepts and Controversies. Freeman, 1985. *Presents ideas of statistics, rather than techniques. Nontechnical introduction to statistics.*
- Ingelfinger, J. and Mosteller, et al. Biostatistics in Clinical Medicine. Macmillan 1983. *Statistics are explained through single patient management.*
- E. Lehmann and H. D'Abrera . Nonparametrics: Statistics based on ranks. Prentice Hall, 1998. *Revised Edition of the classic text on nonparametric statistics.*
- Sprent P. Applied nonparametric statistical methods. Chapman Hall, 1989. *A another more applied nonparametrics book.*
- Gelman, A. et al. Bayesian Data Analysis. Chapman and Hall, 1995. *An introductory book on Bayesian analysis.*
- Rosenberg L, Joseph L, Barkun A. Surgical Arithmetic: Epidemiological, Statistical and Outcomes-Based Approach to Surgical Practice. Landes Biosciences, 2000. *Chapter 2 of this book is an attempt to put (almost) the entire 607 course into a single 50 page book chapter, with separate chapters on diagnostic tests and regression. Introductory book on many issues of interest to epidemiology students, including statistics, basic epidemiology, decision analysis, clinical trials, survival analysis, meta-analysis, and technology assessment.*

Statistical Project

You are asked to find three instances of the use of statistical methods. One of the sources should be a journal article, one should be from a newspaper or magazine, and the last could be from any source (can again use a journal, newspaper or magazine article, but can also be from advertising, etc.). For each, provide a concise (maximum three DOUBLE SPACED pages each, but can be less) commentary on each. At least two of the three should relate to the use of statistics in medicine, but application to another area is allowed for one of the articles.

Your comments can include an explanation of the methods and calculations, the assumptions required by the methods, and, *most importantly*, comments on whether the major conclusions follow from the data and methods presented. You should take into consideration the source, for example, one cannot expect a brief newspaper article to have a complete description of the methods when reporting about a medical finding.

This project counts for 15% of the final grade for the course, so that each article is worth about 5%. In assessing the quality of the exercise, I will consider the extent to which you demonstrate understanding of important statistical concepts via the examples, and your judgement in evaluating the conclusions. Credit will also be given for ingenuity in the use of the available information. For example, if an article provides only a p-value, you may be able to derive an approximate confidence interval, which would usually be more informative (for reasons we will see during the course).

Deadline: The exercise is due on Thursday, November 27, 2003, although I would strongly encourage you to collect items of interest throughout the term to avoid the end-of-term rush.

Please hand in complete copies of all articles on which the exercises are based.

Principles of Inferential Statistics in Medicine

Course Outline – EPIB–607, September – December 2003

	General Area	Specific Topics	Dates	Colton	Moore and McCabe	Armitage and Berry
1	Introduction	– course description and evaluation – introduction to statistical analysis in medicine – math background	Sept 4	Chapter 1 pp 1–7	Not covered.	Chapter 1 pp 1–7
2	Data Summaries and Descriptive Statistics	– types of data – histograms – stemplots – boxplots – means – medians – variance – relocating/rescaling	Sept 9 – Sept 11	Chapter 2 pp 11–44 Boxplots and stemplots not covered.	Chapter 1 pp 1–55 & Ch2. 106–112	Chapter 1 pp 8–46
3	Probability and Probability Distributions	– laws of probability – discrete and continuous random variables – expectation and variance of r.v.'s – diagnostic tests and conditional probabilities – Bayes Theorem – Normal distribution – area under Normal curve – binomial distribution – Normal approximation to the binomial – Poisson distribution	Sept 16 – Sept 23	Chapter 3 pp 63–92	Chapter 1 pp 63–78 Chapter 3 pp 260–269 Chapter 4 pp 279–352 Chapter 5 pp 365–383 Diagnostic tests and Poisson not covered.	Chapter 2 pp 47–82 Chapter 19 pp 692–698
4	Inference Concerning Means	– random sampling – hypothesis testing for means – type I and type II errors – p-values – confidence intervals for means – t distribution – paired and unpaired samples – Bayesian inference – sample size calculations and power	Sept 25 – Oct 16	Chapter 4 pp 99–146 Bayes not covered	Chapter 5 pp391–400 Chapter 6 pp 415–479 Chapter 7 pp 491–543 Bayes not covered	Chapter 4 pp 83–112 pp 137–141 Chapter 6 pp 165–174 Chapter 16 pp 528–538

Midterm Exam: Thursday October 23, 2003, 9:00 AM – 11:00 AM, Room 1/12, Dentistry and Anatomy Building. Tuesday October 21, 2003 will be used as a review day, to go over old exams and answer questions.

Principles of Inferential Statistics in Medicine

Course Outline – EPIB–607, September – December 2003

	General Area	Specific Topics	Dates	Colton	Moore and McCabe	Armitage and Berry
5	Inference concerning proportions and counts	– hypothesis testing for proportions – sample size calculations and power – paired and unpaired samples – χ^2-test to compare 2 or more proportions – Fishers exact test – Bayesian inference – Bayesian inference – Mantel-Haenzel to combine 2 × 2 tables – relative risk and odds ratios – inference for count data	Oct 28 – Nov 11	Chapter 5 pp 151–183 Bayes, Mantel-Haenzel, counts, relative risk and odds ratio not covered.	Chapter 8 pp 571–595 Fishers exact test, Bayes, Mantel-Haenzel, relative risk, odds ratios and counts not covered.	Chapter 4 pp 112–137 Chapter 6 pp 175–179 Chapter 19 pp 667–676
6	Nonparametric Statistics	– sign test – Rank sum test – Wilcoxon signed rank test – CI for median	Nov 13 – Nov 18	Chapter 7 pp 219–226 Sign test and CI not covered.	Chapter 14 (On CDROM) pp 1–26 Chap 7 553–559	Chapter 10 pp 272–285
7	Regression and Correlation	– difference between regression and correlation – scatter plots – linear regression – least squares method – estimation of parameters in regression – Bayesian inference in regression –basic design in regression – other types of regression – Pearson's correlation – Spearman's correlation	Nov 20 – Nov 27	Chapter 6 pp 189–214 Bayes not covered	Chapter 2 pp 126–168 Chapter 10 pp 657–691 Bayes not covered	Chapter 7 pp 187–207 Chapter 16 pp 538–543

Final Exam: Tuesday December 9, 2003, 9:00 AM – 12:00 PM, Room TBA. Tuesday December 2, 2003 will be used as a review day, to go over old exams and answer questions.

WHY STUDY STATISTICS IN MEDICINE?

- Medicine and Epidemiology are becoming increasingly quantitative.
- Knowledge of statistics is required to design experiments that will satisfactorily answer medical questions.
- To understand medical literature.

CANADA POST SHOWS IMPROVEMENT

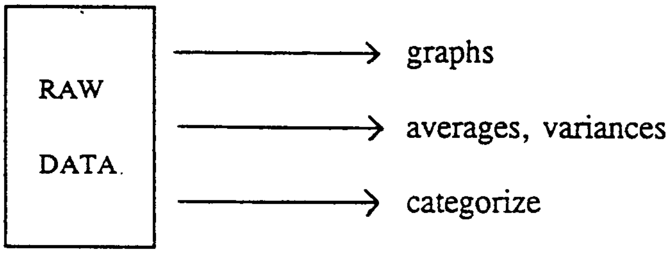
Percentage of the first class mail delivered on time:

December 1989	June 1990
85%	95%
$N_1 = 1000$	$N_2 = 2000$

Overall for the six months: 91.6%

Aspirin		Tylenol	
Cured	Not Cured	Cured	Not Cured
5	5	5	5
6	4	5	5
6	4	4	6
7	3	4	6
8	2	4	6
8	2	3	7
9	1	3	7
⋮	⋮	⋮	⋮
10	0	0	10

DESCRIPTIVE STATISTICS

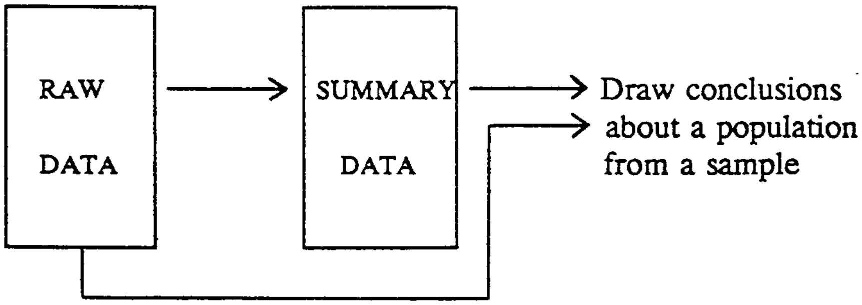


RAW DATA → graphs

→ averages, variances

→ categorize

INFERENTIAL STATISTICS



RAW DATA → SUMMARY DATA → Draw conclusions about a population from a sample

NEJM, 1983, Volume 309, pp. 709-713.

10

Table 2. Statistical Content and Accessibility of *Journal* Articles.

Procedure	Articles Containing Methods	Accumulated by Article	Accessibility by Article-Method
	no. (%)	*no.* (%)	(%)
✓ No statistical methods or descriptive statistics only	443 (58)	443 (58)	(40)
✓ t-test	179 (24)	509 (67)	(56)
✓ Contingency tables	112 (15)	551 (73)	(66)
✓ Non-parametric tests	45 (6)	571 (75)	(70)
✓ Epidemiologic statistics	39 (5)	585 (77)	(73)
✓ Pearson correlation	55 (7)	598 (79)	(78)
✓ Simple linear regression	37 (5)	621 (82)	(81)
Analysis of variance	33 (4)	636 (84)	(84)
✓ Transformation	26 (3)	650 (86)	(87)
✓ Non-parametric correlation	15 (2)	662 (87)	(88)
Life table	24 (3)	674 (89)	(90)
Multiple regression	19 (3)	686 (90)	(92)
✓ Multiple comparisons	13 (2)	698 (92)	(93)
Other methods	17 (2)	708 (93)	(94)
Adjustment and standardization	13 (2)	718 (95)	(96)
Multiway tables	12 (2)	728 (96)	(97)
✓ Power	13 (2)	737 (97)	(98)
Other survival analysis	11 (1)	747 (98)	(99)
Regression for survival	6 (1)	753 (99)	(99)
Cost–benefit analysis	6 (1)	758 (100)	(100)
Sensitivity analysis	2 (0)	760 (100)	(100)
Totals:			
Article-methods used	1120		
Articles		760	

11

MYOCARDIAL ISCHEMIA CAUSED BY DISTAL CORONARY-ARTERY CONSTRICTION IN STABLE ANGINA PECTORIS

Giuseppe Pupita, M.D., Attilio Maseri, M.D., F.R.C.P., Juan Carlos Kaski, M.D.,
Alfredo R. Galassi, M.D., Stavros Gavrielides, M.D., Graham Davies, M.D., F.R.C.P.,
and Filippo Crea, M.D.

Abstract *Background.* In patients with stable coronary artery disease, the ischemic threshold for the production of effort-related angina is often quite variable. Although this feature is commonly attributed to changes in the caliber of coronary arteries at the site of stenosis, it could also be caused by the constriction of distal vessels, collateral vessels, or both.

Methods. In order to test this hypothesis, we studied 11 patients with stable angina, total occlusion of a single coronary artery that was supplied by collateral vessels, normal ventricular function, no evidence of coronary-artery spasm, and no other coronary stenoses. These conditions precluded the modulation of coronary flow by vasomotion at the site of the coronary stenosis.

Results. The ischemic threshold — assessed by multiplying the heart rate by the systolic blood pressure at a 1-mm depression of the ST segment during exercise testing — increased by 19 percent after the administration of nitroglycerin (P<0.05) and decreased by 18 percent after the administration of ergonovine (P<0.01). Ambulatory electrocardiographic monitoring of the patients when not receiving treatment detected 73 ischemic episodes that, in keeping with the history, showed variations of 25 to 52 beats per minute in the heart rate at a 1-mm depression of the ST segment; 12 episodes of sinus tachycardia exceeded the lowest ischemic heart rate by a mean (±SD) of 22±13 beats per minute without ST-segment depression. Furthermore, 21 ischemic episodes occurred at a heart rate more than 25 beats per minute below that at a 1-mm depression of the ST segment during exercise testing. Delayed and reduced filling of collateral and collateralized vessels associated with depression of the ST segment similar to that observed during ambulatory monitoring was detected on angiographic evaluation after the intracoronary administration of ergonovine in three patients.

Conclusions. We propose that the constriction of distal coronary arteries, collateral vessels, or both may cause myocardial ischemia in patients with chronic stable angina. (N Engl J Med 1990; 323:514-20.)

Statistical Analysis

Continuous data are presented as means ±SD. Statistical analysis was performed with Student's t-tests for paired and unpaired data, as appropriate. A value of P<0.05 was considered to indicate statistical significance.

EXPOSURE TO HOUSE-DUST MITE ALLERGEN (*Der p* I) AND THE DEVELOPMENT OF ASTHMA IN CHILDHOOD

A Prospective Study

Richard Sporik, M.R.C.P., Stephen T. Holgate, M.D., F.R.C.P.,
Thomas A.E. Platts-Mills, M.D., Ph.D., and Jeremy J. Cogswell, M.D., F.R.C.P.

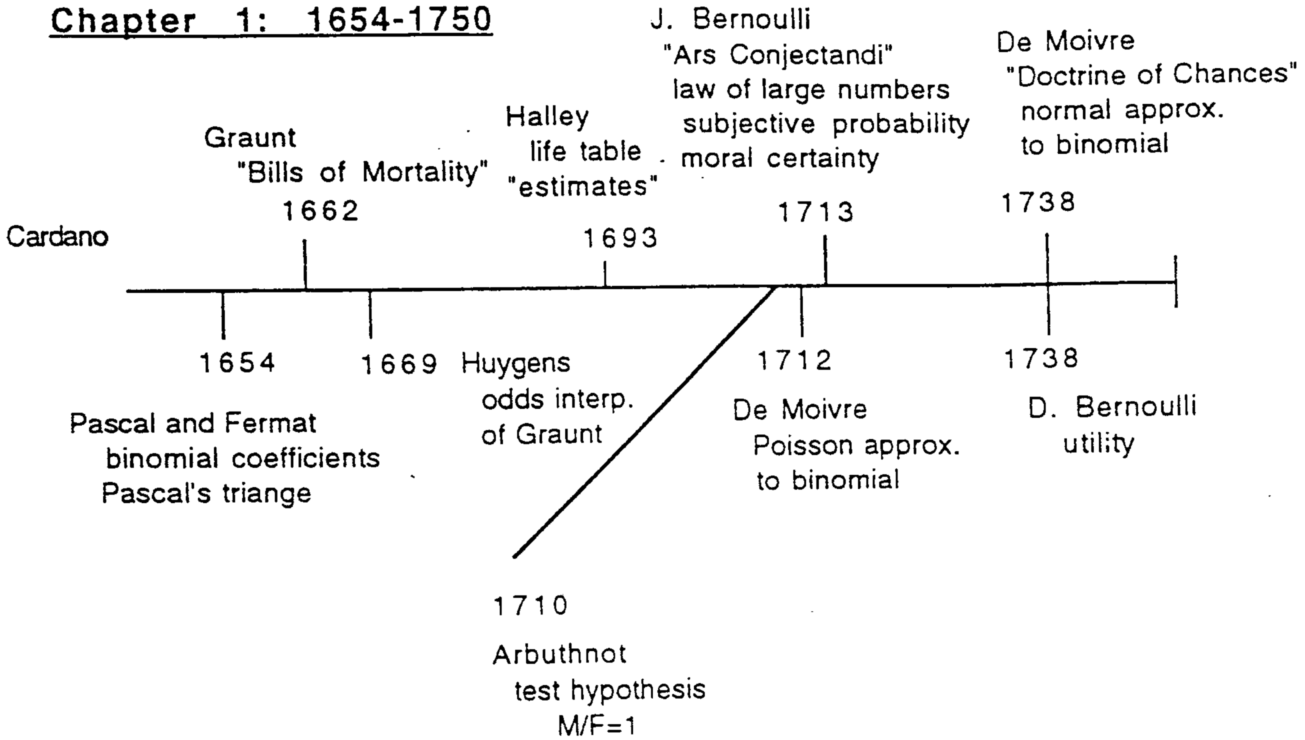
Abstract *Background and Methods.* Children with asthma commonly have positive skin tests for inhaled allergens, and in the United Kingdom the majority of older children with asthma are sensitized to the house-dust mite. In a cohort of British children at risk for allergic disease because of family history, we investigated prospectively from 1978 to 1989 the relation between exposure to the house-dust mite allergen (*Der p* I) and the development of sensitization and asthma.

Results. Of the 67 children studied in 1989, 35 were atopic (positive skin tests), and 32 were nonatopic. Of the 17 with active asthma, 16 were atopic ($P<0.005$), all of whom were sensitized to the house-dust mite, as judged by positive skin tests and levels of specific IgE antibodies ($P<0.001$). For house-dust samples collected from the homes of 59 of the children in 1979 and from 65 homes in 1989, the geometric means for the highest *Der p* I exposure were, respectively, 16.1 and 16.8 μg per gram of sieved dust. There was a trend toward an increasing degree of sensitization at the age of 11 with greater exposure at the age of 1 ($P = 0.062$). All but one of the children with asthma at the age of 11 had been exposed at 1 year of age to more than 10 μg of *Der p* I per gram of dust; for this exposure, the relative risk of asthma was 4.8 ($P = 0.05$). The age at which the first episode of wheezing occurred was inversely related to the level of exposure at the age of 1 for all children ($P = 0.015$), but especially for the atopic children ($r = -0.66$, $P = 0.001$).

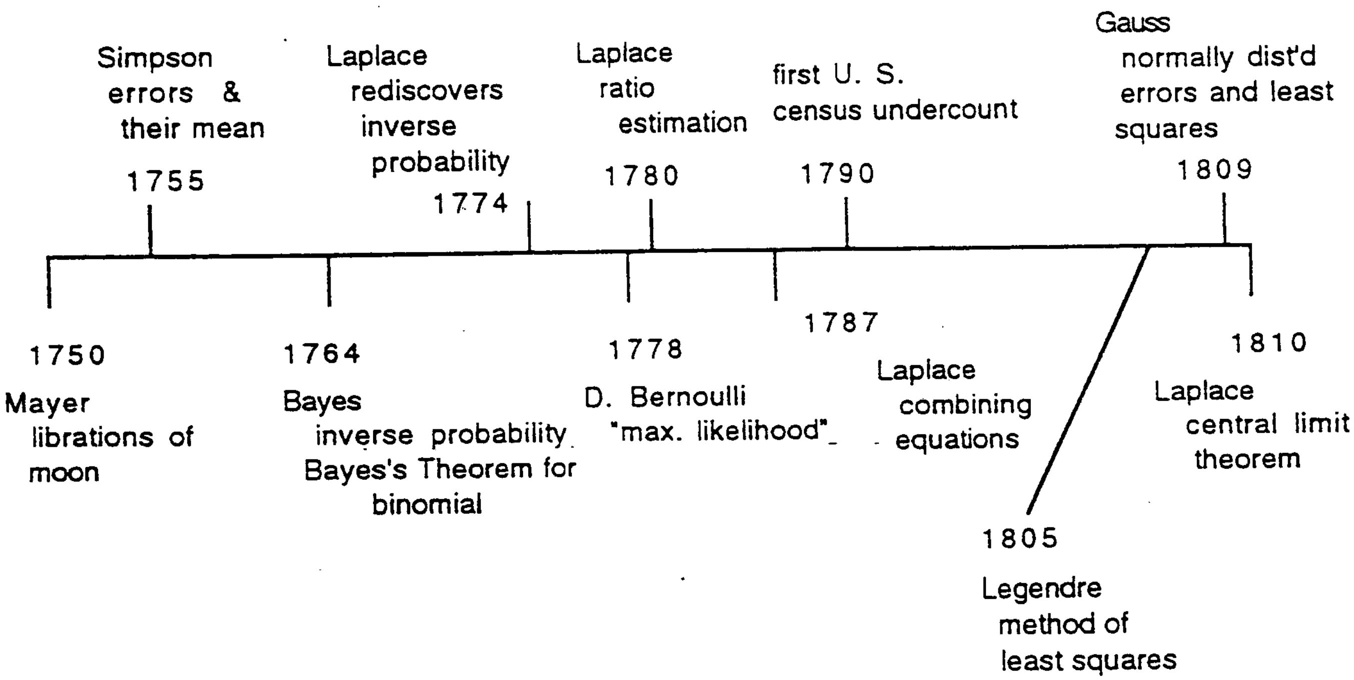
Conclusions. In addition to genetic factors, exposure in early childhood to house-dust mite allergens is an important determinant of the subsequent development of asthma. (N Engl J Med 1990; 323:502-7.)

Comparisons between the clinical groups were made by nonparametric methods. Contingency tables were analyzed by the chi-square method and a two-tailed Fisher's exact test. Data sets were analyzed by Spearman's rank-correlation test. Relative risk was calculated as the ratio of prevalence of disease among exposed children to the prevalence of disease among unexposed children, as described by Schlesselman.[28] The significance of the relative risk was calculated by the chi-square test. The relation between exposure and sensitization to mites was analyzed for linear trend in the proportion of sensitive children.[29]

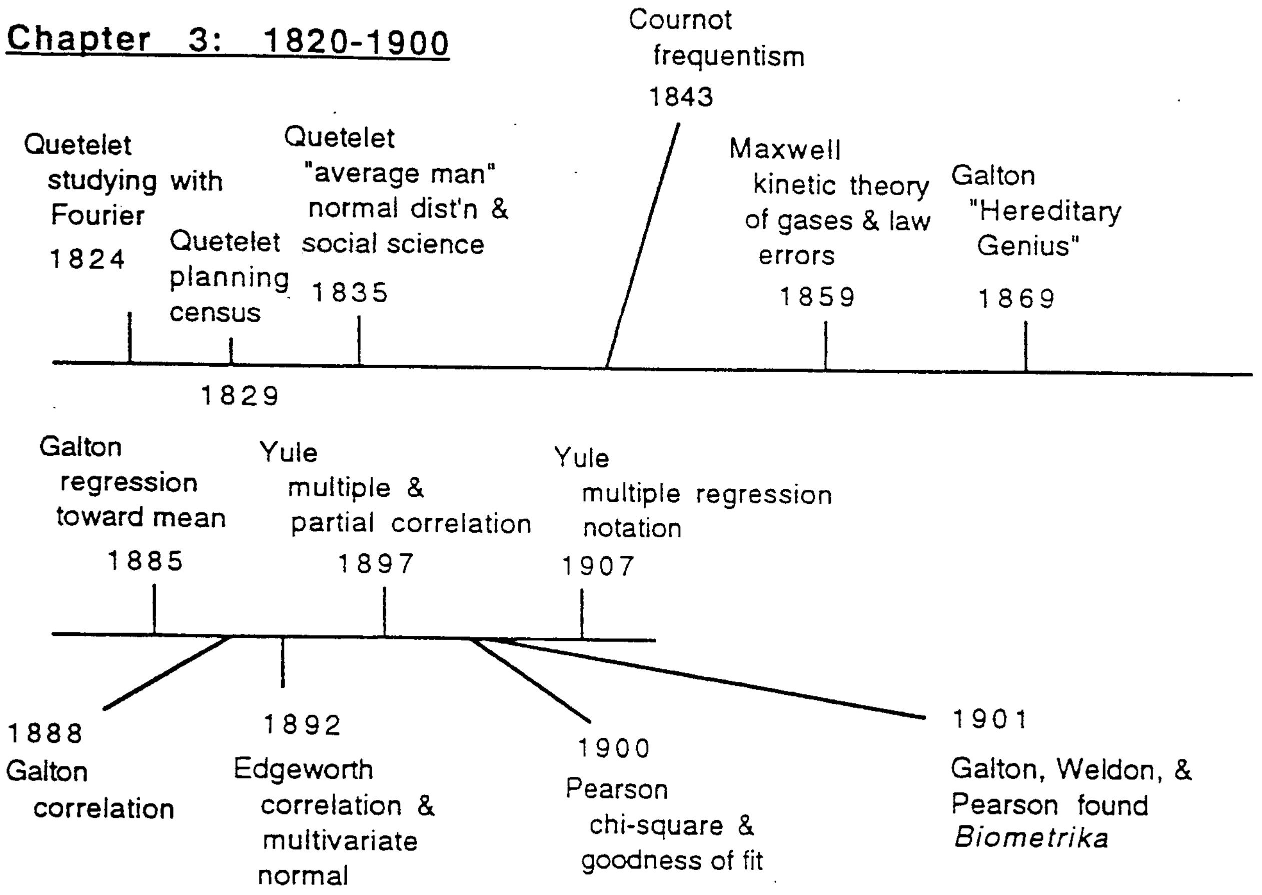
Chapter 1: 1654-1750



Chapter 2: 1750-1820



Chapter 3: 1820-1900



1824 Quetelet studying with Fourier

1829 Quetelet planning census

1835 Quetelet "average man" normal dist'n & social science

1843 Cournot frequentism

1859 Maxwell kinetic theory of gases & law errors

1869 Galton "Hereditary Genius"

1885 Galton regression toward mean

1888 Galton correlation

1892 Edgeworth correlation & multivariate normal

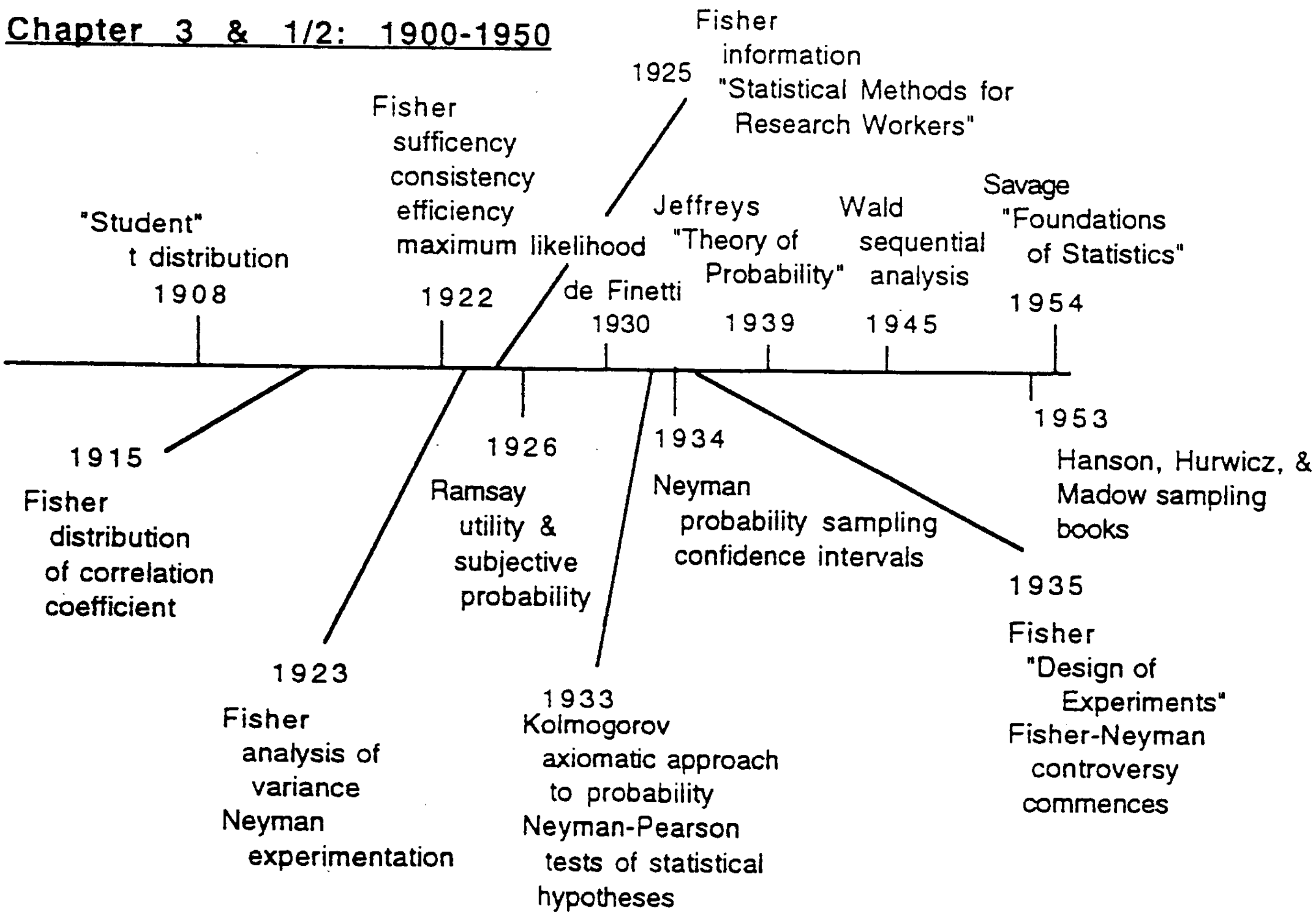
1897 Yule multiple & partial correlation

1900 Pearson chi-square & goodness of fit

1901 Galton, Weldon, & Pearson found *Biometrika*

1907 Yule multiple regression notation

Chapter 3 & 1/2: 1900-1950



1908 "Student" t distribution

1915 Fisher distribution of correlation coefficient

1922 Fisher sufficency consistency efficiency maximum likelihood

1923 Fisher analysis of variance Neyman experimentation

1925 Fisher information "Statistical Methods for Research Workers"

1926 Ramsay utility & subjective probability

1930 de Finetti

1933 Kolmogorov axiomatic approach to probability Neyman-Pearson tests of statistical hypotheses

1934 Neyman probability sampling confidence intervals

1935 Fisher "Design of Experiments" Fisher-Neyman controversy commences

1939 Jeffreys "Theory of Probability"

1945 Wald sequential analysis

1953 Hanson, Hurwicz, & Madow sampling books

1954 Savage "Foundations of Statistics"

Principles of Inferential Statistics in Medicine

Mathematical Background

QUESTIONS:

1. (a) What is a function?
   (b) Why do we need functions in statistics?
2. (a) What is the derivative of a function?
   (b) Why do we need derivatives in statistics?
3. (a) What is the indefinite integral of a function?
   (b) Why do we need indefinite integrals in statistics?
4. (a) What is the definite integral of a function?
   (b) Why do we need definite integrals in statistics?

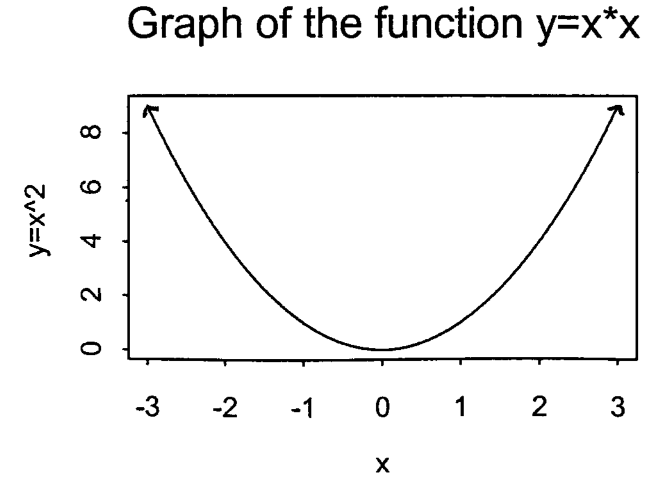
> Note: The following are *very* non-rigorous definitions designed to suit the purpose of our course. Refer to any calculus textbook for the exact definitions and/or more information.

1. (a) For our purposes, a *function* assigns a unique numerical value to each number in a specified set. For example, the function

$$f(x) = x^2, \; -\infty < x < +\infty$$

assigns the value x^2 to each x, $-\infty < x < +\infty$. Thus $x = 1$ is assigned the value 1, $x = 2$ is assigned the value 4, and $x = -2.1$ is assigned the value $+4.41$, etc. A function is defined over a set of values, which here is the set of all real numbers.

Functions are often easily understood by looking at the *graph* of the function.



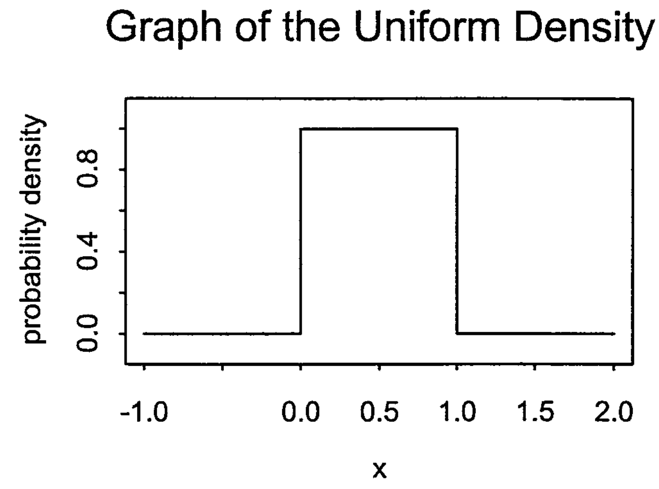
Graph of the function y=x*x

(b) Functions are used in statistics to describe probability (density) functions (among many other things). We will discuss probability functions starting in Section 3 of the course, but we can look ahead now to some examples:

(i) The Uniform probability (density) function describes the experiment of choosing a random number between 0 and 1. The function is

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise}, \end{cases}$$

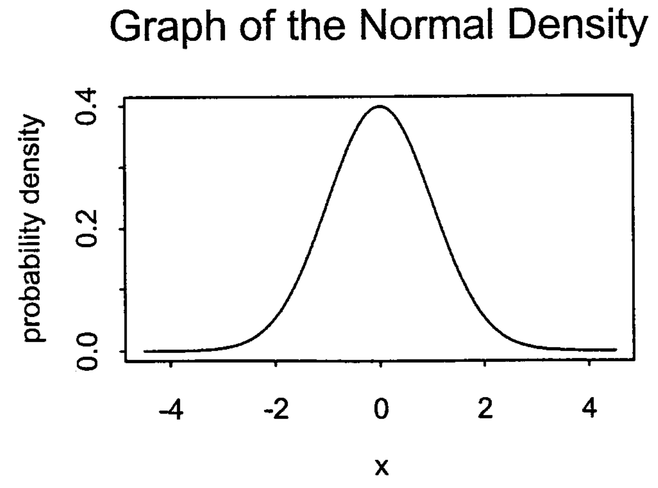
and the graph is shown below:



(ii) The standard Normal probability (density) function is used extensively in virtually every discipline where statistics are used, including medicine. The function is

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\left\{ -\frac{x^2}{2} \right\}, \; -\infty < x < +\infty$$

and the graph is shown below:



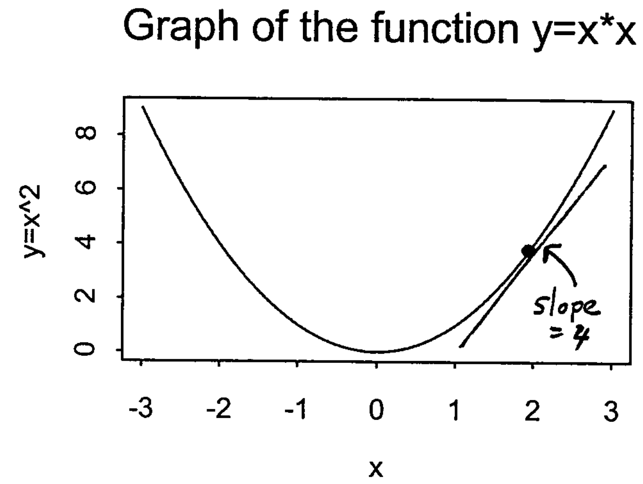
2. (a) The *derivative* of a function measures the slope of the tangent line to the graph of the function at a given point. For example, if

$$f(x) = x^2,$$

then the derivative is given by

$$f'(x) = 2 \times x.$$

For example, this means that the slope of the tangent line at the point $x = 2$ (with $f(x) = y = 4$) is $2 \times 2 = 4$.



You may recall the following useful facts relating to derivatives:

1. The slope of a line is a measure of how quickly the function is rising or falling as x increases in value.
2. If a function has a maximum or minimum value, the the derivative is usually equal to 0 at that point. In the above, the function has a minimum at $x = 0$, where the value of the derivative is zero.

(b) Derivatives will be used when maximum likelihood estimators are discussed, in Section 4 of the course.

3. (a) The *indefinite integral* is a synonym for "anti-differentiation". In other words, when we calculate the indefinite integral of a function, we look for a function that when differentiated, returns the function under the integral sign. For example, the indefinite integral of the function $f(x) = x^2$ is given by the

$$\int x^2 \, dx = \frac{1}{3} \times x^3$$

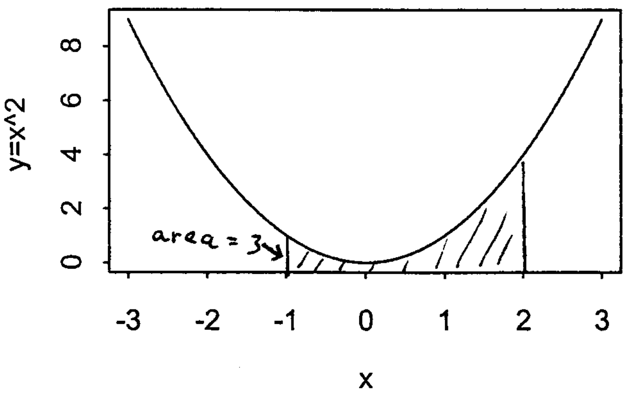
because the derivative of $\frac{1}{3} \times x^3$ is x^2.

(b) Indefinite integrals are used in many places in statistics, but we will see them only in the context of regression. When we want to look at the probability density of a regression coefficient, for example, we use a definite integral to go from a *joint density* (many variables at once) to a *marginal density* (of a single variable).

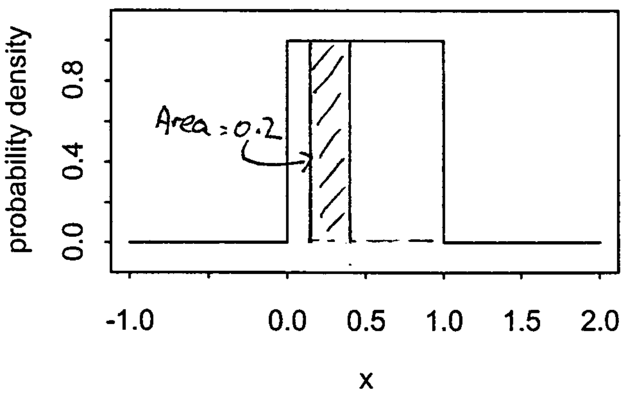
4. (a) The *definite integral* of a function is the area under the graph of that function. This area can be approximated directly from the graph, but exact mathematical formulae are also available from calculus. For example, the area under the the curve ranging from -1 to +2 of the function $f(x) = x^2$ is given by the following definite integral formula:

$$\int_{-1}^{+2} x^2 \, dx = \frac{1}{3} \times x^3 \Big|_{-1}^{+2} = \frac{2^3}{3} - \frac{(-1)^3}{3} = \frac{8}{3} + \frac{1}{3} = 3.$$

Graph of the function y=x*x



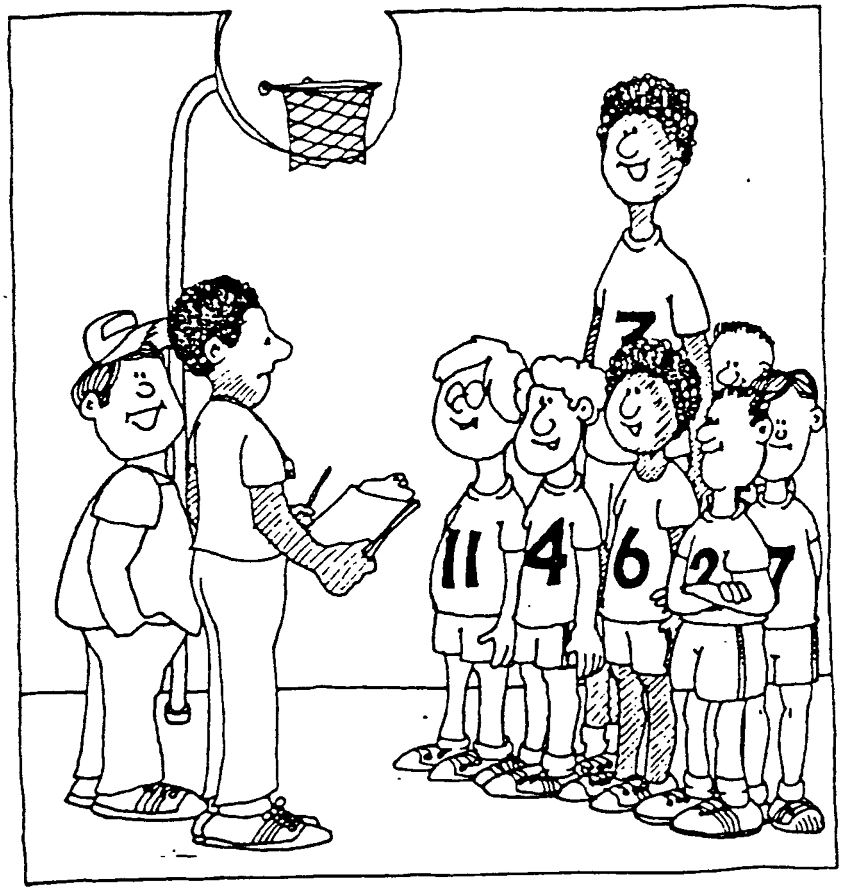
Graph of the Uniform Density



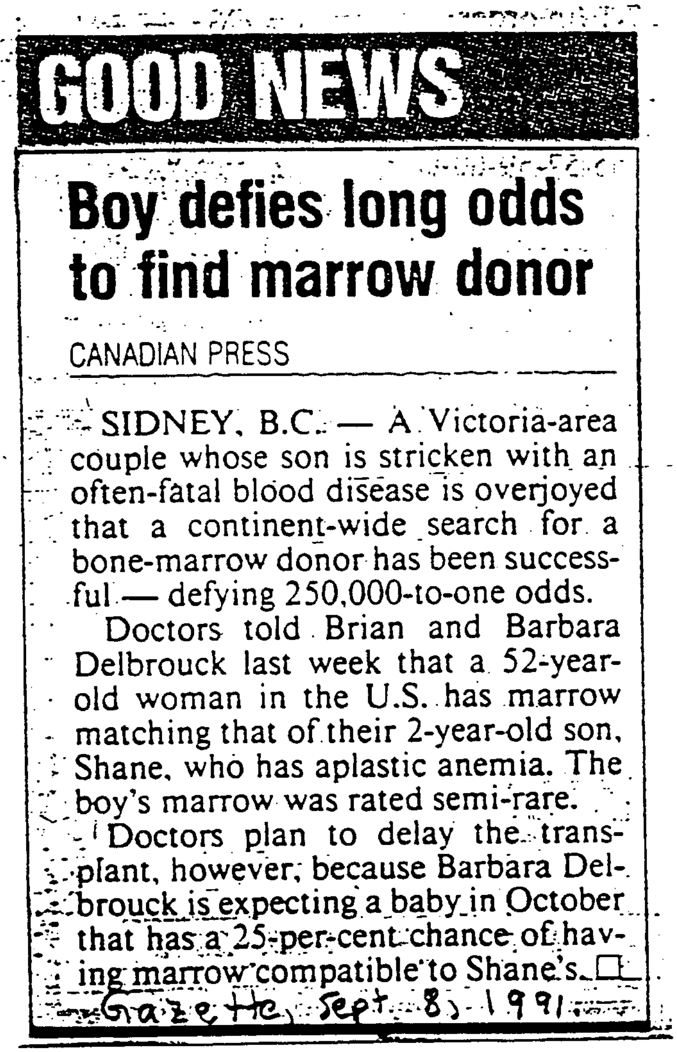
(b) The area under a curve of a probability density function gives the probability of getting values in the region of the definite integral. For example, supposed we wished to calculate the probability that in choosing a random number between 0 and 1 (Uniform density function) the particular number we choose falls between 0.2 and 0.4. This is calculated by the definite integral

$$\int_{0.2}^{0.4} 1 \, dx = x \Big|_{0.2}^{0.4} = 0.4 - 0.2 = 0.2.$$

We will also see definite integrals in the context of calculating means and variances of random variables.



" SHOULD WE SCARE THE OPPOSITION BY ANNOUNCING OUR MEAN HEIGHT OR LULL THEM BY ANNOUNCING OUR MEDIAN HEIGHT ? "



GOOD NEWS

Boy defies long odds to find marrow donor

CANADIAN PRESS

SIDNEY, B.C. — A Victoria-area couple whose son is stricken with an often-fatal blood disease is overjoyed that a continent-wide search for a bone-marrow donor has been successful — defying 250,000-to-one odds.

Doctors told Brian and Barbara Delbrouck last week that a 52-year-old woman in the U.S. has marrow matching that of their 2-year-old son, Shane, who has aplastic anemia. The boy's marrow was rated semi-rare.

Doctors plan to delay the transplant, however, because Barbara Delbrouck is expecting a baby in October that has a 25-per-cent chance of having marrow compatible to Shane's. □

Gazette, Sept. 8, 1991

What are the real odds?

If 1 person was recruited: 1/250,000=0.000004

If 2 persons were recruited: 0.000008

If 10 persons were recruited: .0000399992

If 100 persons were recruited: .0003999228

If 1000 persons were recruited: .0039920351

If 10,000 persons were recruited: .0392106073

If 100,000 persons were recruited: .3296819238

If 250,000 persons were recruited: .6321229463

If 1,000,000 persons were recruited: .9816848366

Descriptive Statistics

Blood Pressure Data

Patient #	Diastolic Pressure	Age	Age Category	Sex 0=M, 1=F
1	100	58	6	1
2	75	38	4	0
3	102	59	6	0
4	87	51	5	1
5	72	45	5	1
6	74	45	5	0
7	61	35	4	0
8	95	52	5	0
9	55	36	4	0
10	101	51	5	1
11	66	42	4	1
12	69	41	4	1
13	74	42	4	0
14	73	46	5	1
15	62	41	4	0
16	78	49	5	1
17	59	38	4	0
18	108	61	6	1
19	74	42	4	0
20	62	37	4	0
21	71	38	4	1
22	51	34	3	0
23	115	59	6	1
24	92	51	5	0
25	70	39	4	1
26	47	32	3	0
27	105	58	6	1
28	103	57	6	0
29	43	25	3	1
30	77	43	4	0
31	87	44	4	1
32	47	27	3	1
33	99	55	6	1
34	89	48	5	1
35	84	50	5	1
36	87	44	4	1
37	69	40	4	1
38	83	44	4	0
39	96	56	6	0
40	104	52	5	1
41	70	42	4	1
42	91	55	6	1
43	76	44	4	0
44	66	37	4	0
45	74	40	4	1
46	68	43	4	0
47	96	52	5	0
48	108	61	6	1
49	90	53	5	0
50	98	56	6	1
Mean	80.06	45.76	4.62	0.54
Variance	17.91	8.81	0.92	0.5
Min	43	25	3	0
Max	115	61	6	1

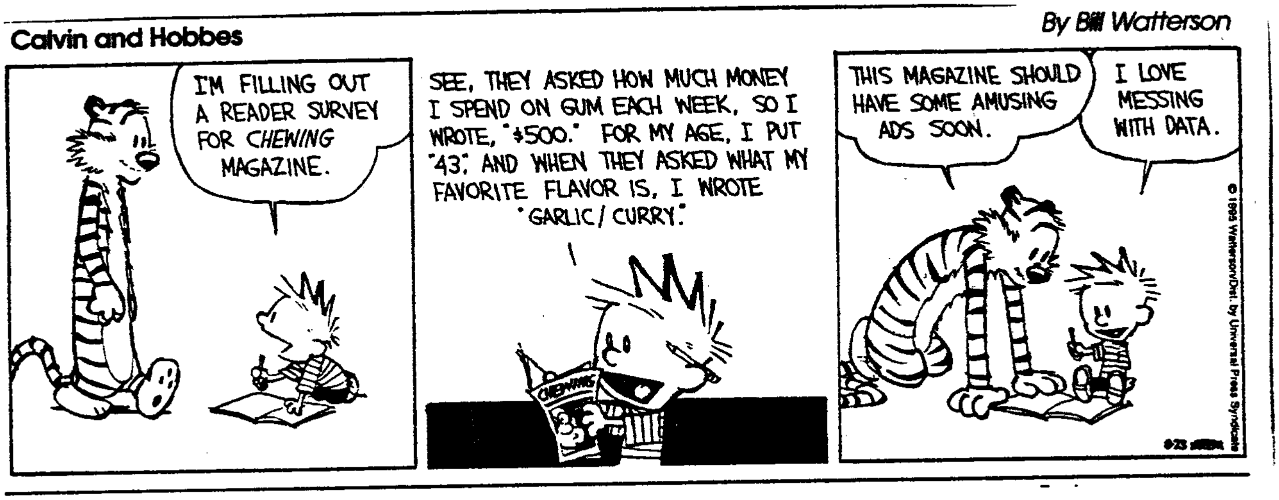
Blood Pressure Data (Sorted)

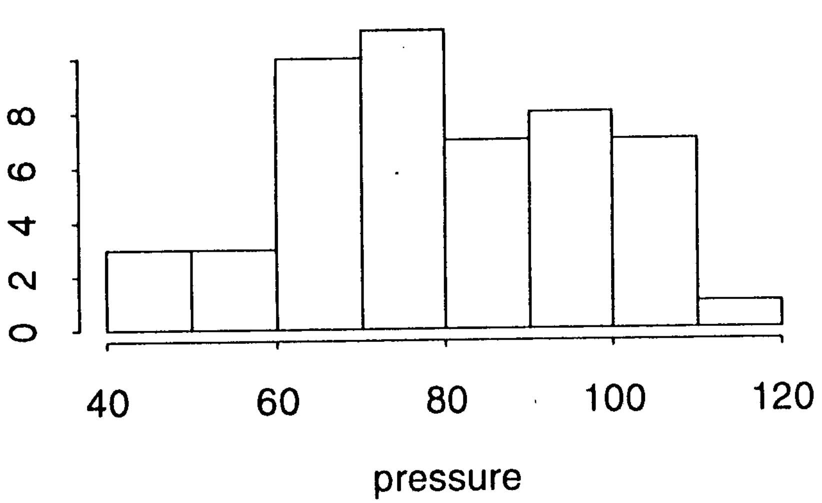
Rank	Patient #	Diastolic Pressure	Age	Age Category	Sex 0=M, 1=F
1	29	43	25	3	1
2	32	47	27	3	1
3	26	47	32	3	0
4	22	51	34	3	0
5	9	55	36	4	0
6	17	59	38	4	0
7	7	61	35	4	0
8	20	62	37	4	0
9	15	62	41	4	0
10	44	66	37	4	0
11	11	66	42	4	1
12	46	68	43	4	0
13	37	69	40	4	1
14	12	69	41	4	1
15	25	70	39	4	1
16	41	70	42	4	1
17	21	71	38	4	1
18	5	72	45	5	1
19	14	73	46	5	1
20	45	74	40	4	1
21	19	74	42	4	0
22	13	74	42	4	0
23	6	74	45	5	0
24	2	75	38	4	0
25	43	76	44	4	0
26	30	77	43	4	0
27	16	78	49	5	1
28	38	83	44	4	0
29	35	84	50	5	1
30	31	87	44	4	1
31	36	87	44	4	1
32	4	87	51	5	1
33	34	89	48	5	1
34	49	90	53	5	0
35	42	91	55	6	1
36	24	92	51	5	0
37	8	95	52	5	0
38	47	96	52	5	0
39	39	96	56	6	0
40	50	98	56	6	1
41	33	99	55	6	1
42	1	100	58	6	1
43	10	101	51	5	1
44	3	102	59	6	0
45	28	103	57	6	0
46	40	104	52	5	1
47	27	105	58	6	1
48	18	108	61	6	1
49	48	108	61	6	1
50	23	115	59	6	1

DATA CLEANING

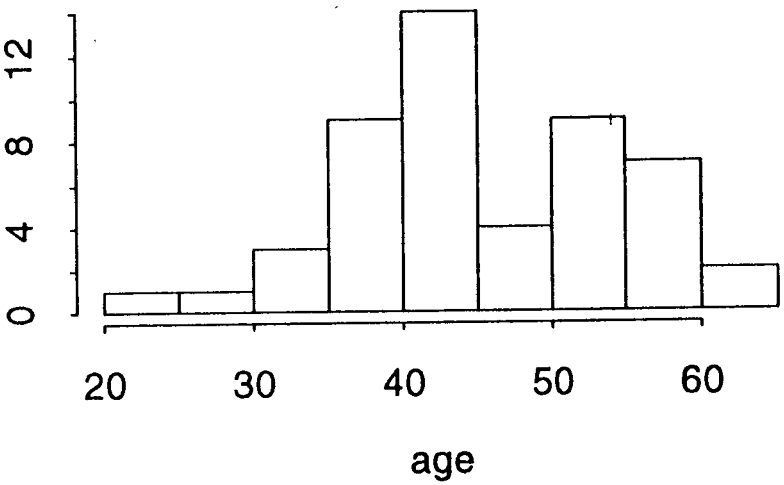
In 1985 British scientists reported a hole in the ozone layer of the earth's atmosphere over the South Pole. This is disturbing, since ozone protects us from cancer-causing ultraviolet radiation. The British report was at first disregarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software used to analyze the data had automatically suppressed these values as erroneous outliers! Readings dating back to 1979 were reanalyzed and showed a large and growing hole in the ozone layer that is unexplained and possibly dangerous.[5] Computers analyzing large volumes of data are often programmed to suppress outliers as protection against errors in the data. As the example of the hole in the ozone layer illustrates, suppressing an outlier without investigating it can keep valuable information out of the sight. ■

- M & M page 16.

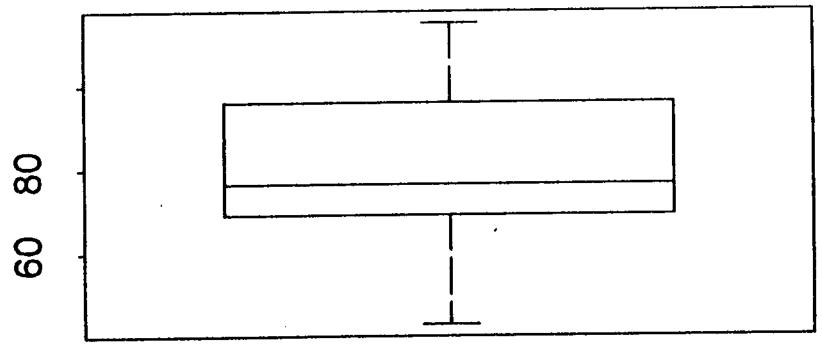

Calvin and Hobbes
By Bill Watterson
I'M FILLING OUT A READER SURVEY FOR *CHEWING* MAGAZINE.
SEE, THEY ASKED HOW MUCH MONEY I SPEND ON GUM EACH WEEK, SO I WROTE, "$500." FOR MY AGE, I PUT "43," AND WHEN THEY ASKED WHAT MY FAVORITE FLAVOR IS, I WROTE "GARLIC / CURRY."
CHEWING
THIS MAGAZINE SHOULD HAVE SOME AMUSING ADS SOON.
I LOVE MESSING WITH DATA.
© 1995 Watterson/Dist. by Universal Press Syndicate



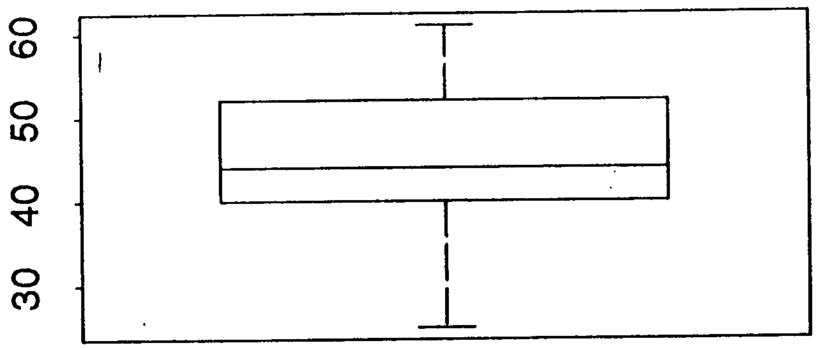
Histogram of Blood Pressure



Histogram of Age

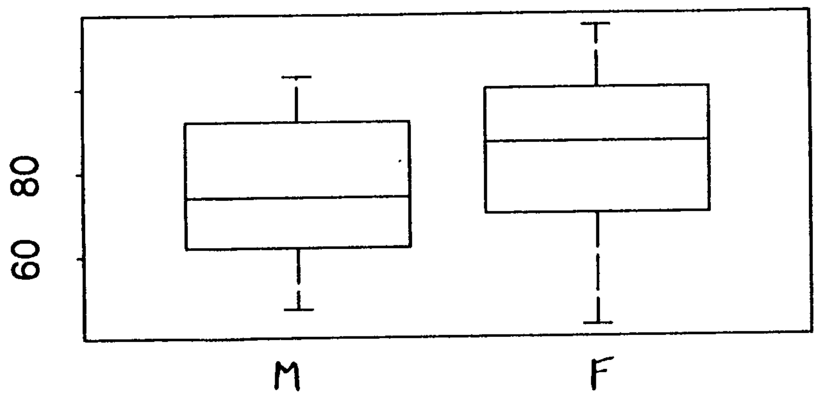


Boxplot of Pressure

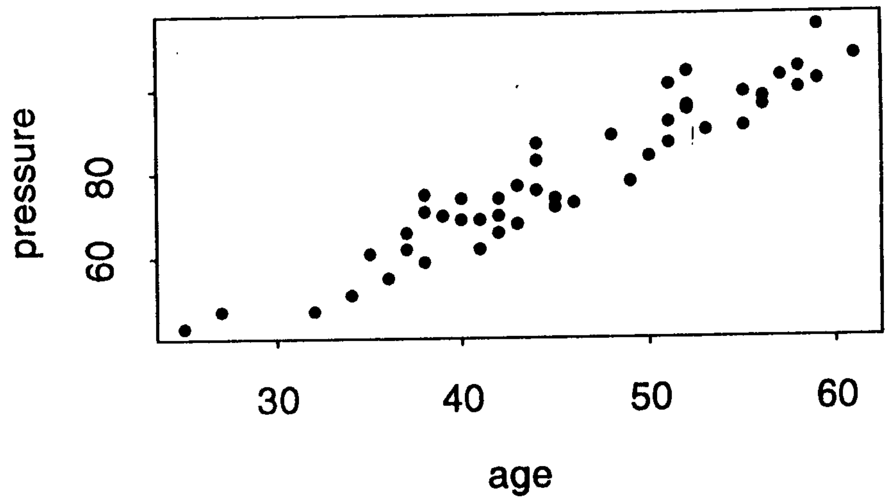


Boxplot of Age

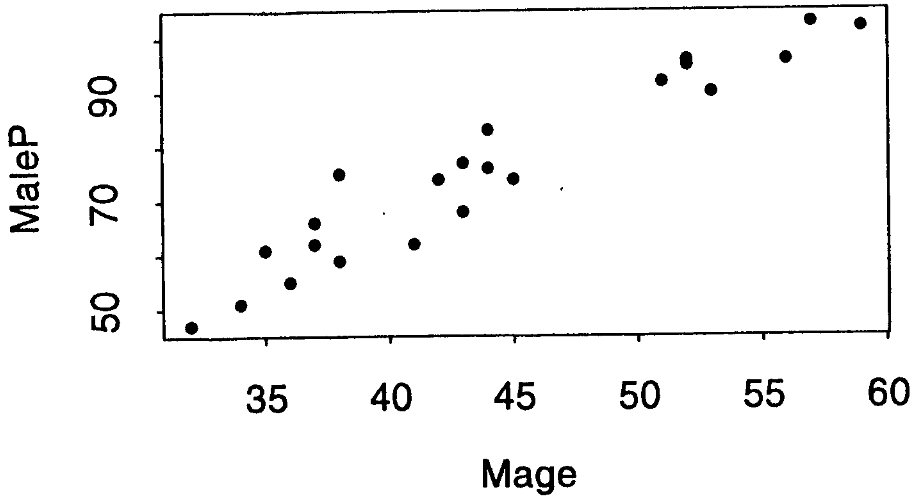
Boxplots of Male vs Female Blood Pressure



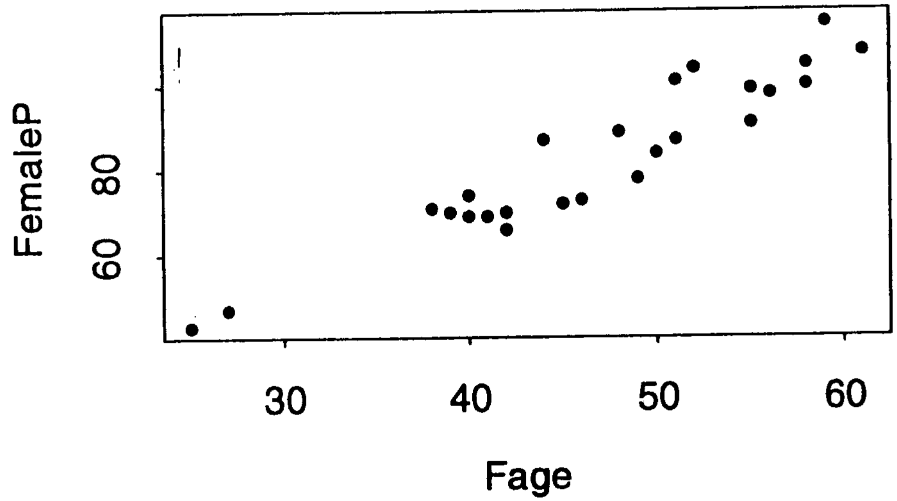
Scatterplot of age vs Pressure



Plot of Age vs Pressure (Males)



Plot of Age vs Pressure (Females)



Stemplots of Age and Blood Pressure

```
> stem(pressure)

N = 50   Median = 76.5
Quartiles = 69, 96

    4 : 377
    5 : 159
    6 : 12266899
    7 : 0012344445678
    8 : 347779
    9 : 01256689
   10 : 01234588
   11 : 5
```

```
> stem(MaleP)

N = 23   Median = 74
Quartiles = 62, 92

    4 : 7
    5 : 159
    6 : 12268
    7 : 444567
    8 : 3
    9 : 02566
   10 : 23
```

```
> stem(FemaleP)

N = 27   Median = 87
Quartiles = 70, 100

    4 : 37
    5 :
    6 : 699
    7 : 0012348
    8 : 47779
    9 : 189
   10 : 014588
   11 : 5
```

```
> stem(age)

N = 50   Median = 44
Quartiles = 40, 52

   2 : 57
   3 : 24
   3 : 56778889
   4 : 00112222334444
   4 : 55689
   5 : 01112223
   5 : 556678899
   6 : 11
```

```
> stem(Mage)

N = 23   Median = 43
Quartiles = 37, 52

   3 : 24
   3 : 567788
   4 : 1223344
   4 : 5
   5 : 1223
   5 : 679
```

```
> stem(Fage)

N = 27   Median = 48
Quartiles = 41, 55

   2 : 57
   3 :
   3 : 89
   4 : 0012244
   4 : 5689
   5 : 0112
   5 : 556889
   6 : 11
```

Early detection and treatment of hyperlipidemia: physician practices in Canada

CAN MED ASSOC J 1990; 143 (9)

Terry N. Tannenbaum, MD, MPH; John S. Sampalis, PhD; Renaldo N. Battista, MD, ScD; Ellen R. Rosenberg, MD; Lawrence Joseph, PhD

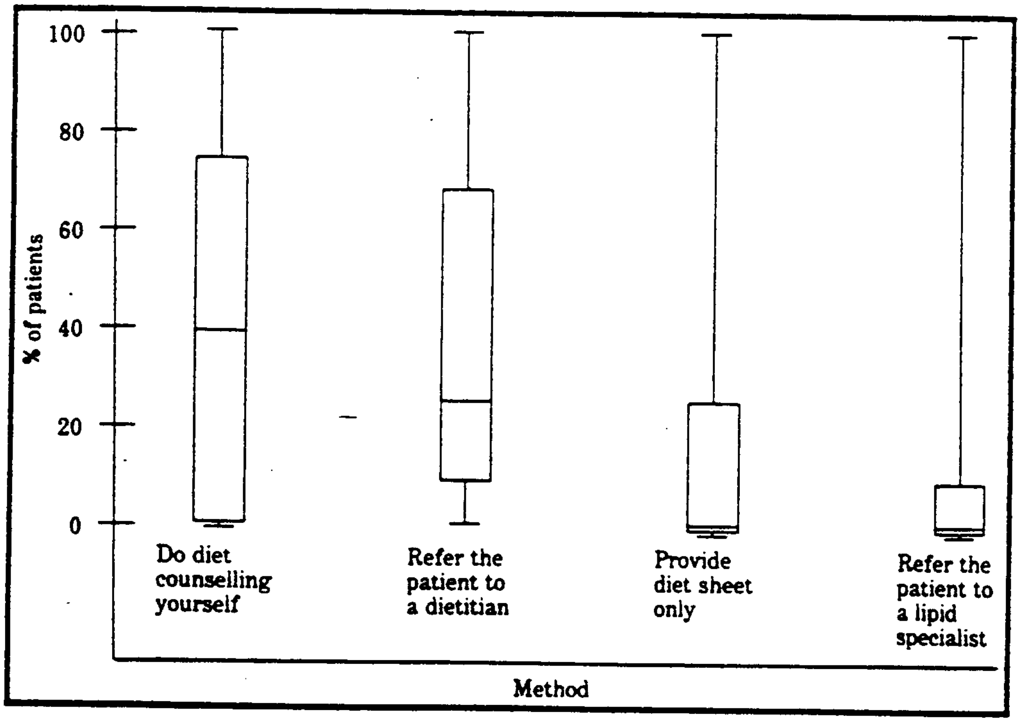


Fig. 1: Frequency of use by primary care physicians of methods for providing dietary therapy for hyperlipidemia. Horizontal lines from top to bottom represent 100th, 75th, 50th, 25th and 0 percentiles of responses; 25th and 75th percentiles form top and bottom frame of box, 50th percentile being within box.

Six Types of Data

1. zero–one, dichotomous, attribute data
2. unordered data
3. ordered classification data
4. ranked data
5. numerical discrete data
6. numerical continuous data

Principles of Inferential Statistics in Medicine

Some common and useful mathematical notation

Referring to data:

Particular values of the variable X in a given data set are usually referred to using lower case letters with subscripts. For example, if the data set $X = \{34, 43, 56, 52, 38\}$ is given, where X may refer to the age of a group of five patients, then $x_3 = 56$ refers to the age of the third data point in the data set. In general, the i^{th} data point is referred to by the notation x_i. The number of subjects in a data set, or the sample size, is often notated by n. In the above, $n = 5$. Thus we can say that we have observed the data $x_i,\ i = 1, 2, \ldots, 5$.

Sums and Products:

The notation $\sum$ refers to the sum of a group of numbers. Thus we can write

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_{n-1} + x_n.$$

In the above example,

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 34 + 43 + 56 + 52 + 38 = 223.$$

Similarly, the notation $\prod$ refers to the product of a group of numbers. Thus we can write

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_{n-1} \times x_n.$$

In the above example,

$$\prod_{i=1}^{5} x_i = x_1 \times x_2 \times x_3 \times x_4 \times x_5 = 34 \times 43 \times 56 \times 52 \times 38 = 161,779,072.$$

Median

To compute the median of a distribution:

1 Arrange all observations in order of size, from smallest to largest.

2 If the number n of observations is odd, the median M is the center observation in the ordered list. The location of the median is found by counting $(n + 1)/2$ observations up from the the bottom of the list.

3 If the number n of observations is even, the median M is the average of the two center observations in the ordered list. The location of the median is again found by counting $(n + 1)/2$ observations up from the bottom of the list.

Mean

If n observations are denoted by $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

or in more compact notation

$$\bar{x} = \frac{1}{n}\sum x_i \tag{1.1}$$

Variance and standard deviation

The variance of n observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

or more compactly

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 \tag{1.2}$$

The standard deviation s is the square root of the variance s^2.

Computing formula for the variance* If you do statistical calculations with a basic calculator, you will need to know alternative formulas that are designed for easier calculation of such quantities as s^2. Equation 1.2 is the *defining formula* for the variance. That is, this equation shows how s^2 measures spread about the mean. But Equation 1.2 is awkward to use because you must first subtract the mean $\bar{x}$ from each individual observation. A bit of algebra shows that an equivalent formula is

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right] \tag{1.3}$$

This is a *computing formula* for the variance; it obscures the meaning of s^2 but leads to much shorter calculations. Equation 1.3 uses the basic quantities $\sum x_i$ and $\sum x_i^2$, which can be obtained quickly on a calculator with a memory and a square button without the need to write down intermediate results.

Numerical Descriptive Statistics of Blood Pressure Data

	All Cases			
	Pressure	Age	Sex	Agecat
Number of Cases	50	50	50	50
Minimum	43	25	0	3
Maximum	115	61	1	6
Average	80	45.8	0.54	4.6
Standard Dev	17.9	8.8	0.5	0.92
Variance	320.8	77.7	0.25	0.85

	Males		
	Pressure	Age	Sex
Number of Cases	23	23	23
Minimum	47	32	0
Maximum	103	59	0
Average	75.6	44	0
Standard Dev	16.6	7.9	0
Variance	275.2	62.5	0

	Females		
	Pressure	Age	Sex
Number of Cases	27	27	27
Minimum	43	25	1
Maximum	115	61	1
Average	83.9	47.3	1
Standard Dev	18.4	9.4	0
Variance	338.7	88.1	0

Table 1: Baseline Patient Characteristics.

haracteristic	Treatment Group Enoxaparin (n = 66)	 Placebo (n = 65)
ge - yr (mean ± SD)	67.48 ± 8.49	68.78 ± 8.88
ex		
Male	23	29
Female	43	36
Type of surgery		
Tibial osteotomy	15	10
Unicompartmental arthroplasty	4	3
Bicompartmental arthroplasty	47	52
Type of anesthesia		
General	62	54
Epidural	4	11
Reason for knee surgery		
Osteoarthritis	59	57
Rheumatoid arthritis	5	7
Complication of prosthesis	0	1
Knee pain	1	0
Avascular necrosis of bone	1	0
Cemented prosthesis	45	44
Operation time - min (mean ± SD)	139.20 ± 50.45	150.31 ± 51.05
Tourniquet time during operation - min (mean ± SD)	79.92 ± 39.29	83.98 ± 31.74
Post-operative continuous passive motion - days (mean ± SD)	6.14 ± 3.61	6.00 ± 3.14
Days to completion of study (mean ± SD)	8.30 ± 3.09	7.83 ± 2.98

Why divide by $n - 1$ instead of n?

Suppose that the population consists of only $n = 4$ members, with values {1,2,3,4}. Then the true mean of the population is

$$\mu = \frac{1+2+3+4}{4} = 2.5,$$

and the true population variance is

$$\sigma^2 = \frac{(1-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2}{4} = 1.25,$$

so that the true standard deviation is

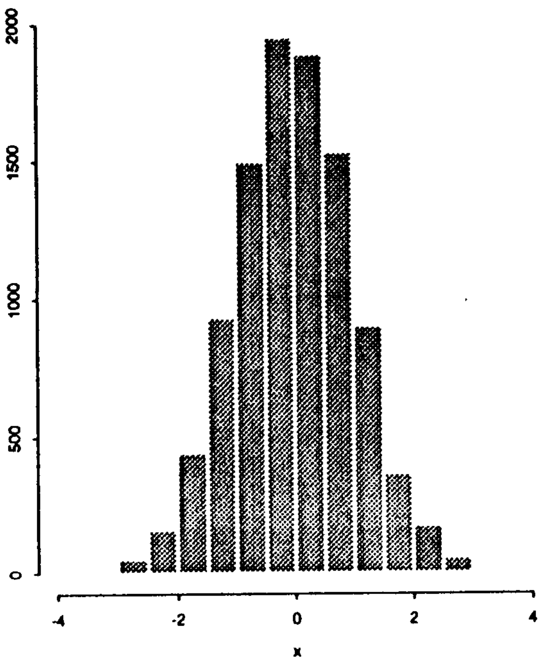
$$\sigma = \sqrt{\sigma^2} = \sqrt{1.25} \approx 1.118.$$

Now suppose that we cannot view the whole population, but instead take a sample of size two. Below we list all of the possible samples that we could take from this population, together with the calculations for mean and variance, where we calculate the variance and standard deviations both with divisors $n - 1$ and n:

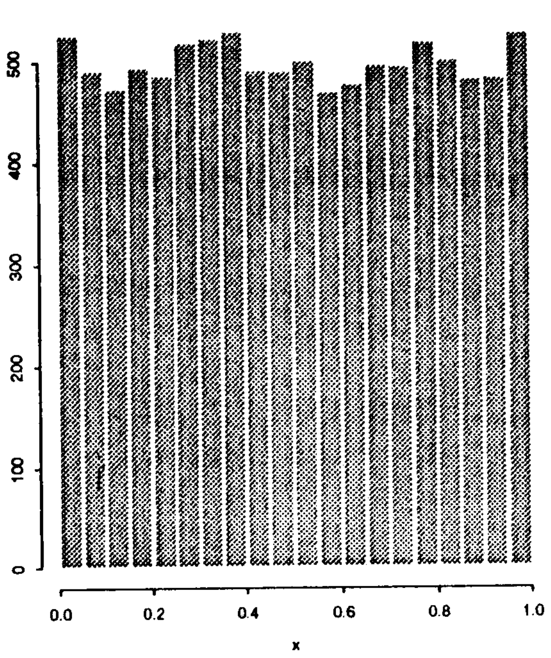
sample (x_1, x_2)	sample mean $\overline{x} = \frac{(x_1+x_2)}{2}$	sample var (n-1) $s^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2}{2-1}$	sample sd (n-1) $s = \sqrt{s^2}$	var (n) $s^2_{pop} = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2}{2}$	sd (n) $s_{pop} = \sqrt{s^2_{pop}}$
(1,2)	1.5	0.5	0.707	0.25	0.5
(1,3)	2.0	2.0	1.414	1.0	1.0
(1,4)	2.5	4.5	2.212	2.25	1.5
(2,3)	2.5	0.5	0.707	0.25	0.5
(2,4)	3.0	2.0	1.414	1.0	1.0
(3,4)	3.5	0.5	0.707	0.25	0.5
(2,1)	1.5	0.5	0.707	0.25	0.5
(3,1)	2.0	2.0	1.414	1.0	1.0
(4,1)	2.5	4.5	2.212	2.25	1.5
(3,2)	2.5	0.5	0.707	0.25	0.5
(4,2)	3.0	2.0	1.414	1.0	1.0
(4,3)	3.5	0.5	0.707	0.25	0.5
(1,1)	1	0	0	0	0
(2,2)	2	0	0	0	0
(3,3)	3	0	0	0	0
(4,4)	4	0	0	0	0
avg	2.5	1.25	1.118	.625	.791

Hence, on average, the divisor $n - 1$ gives the correct estimate, while the divisor n underestimates the value.

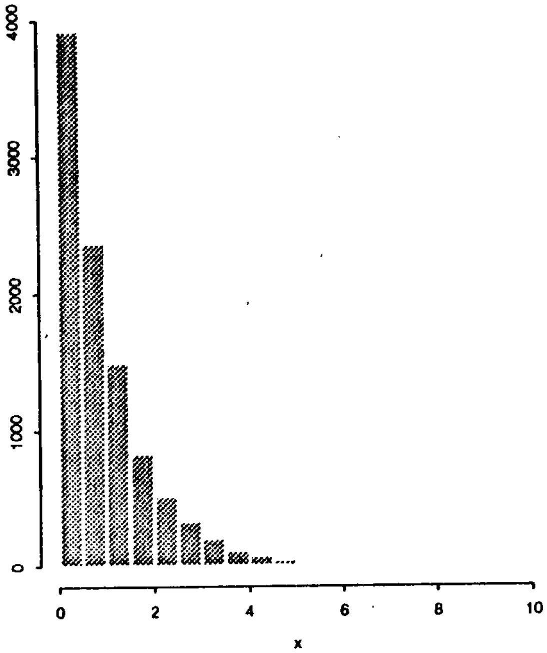
Normal Shape



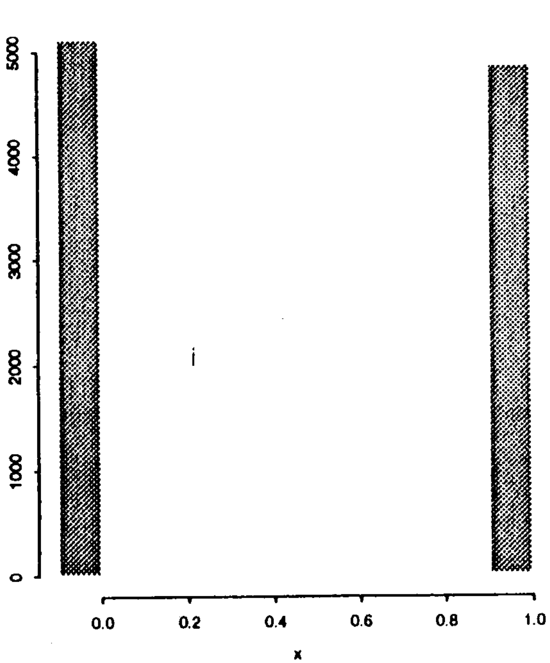
Uniform Shape



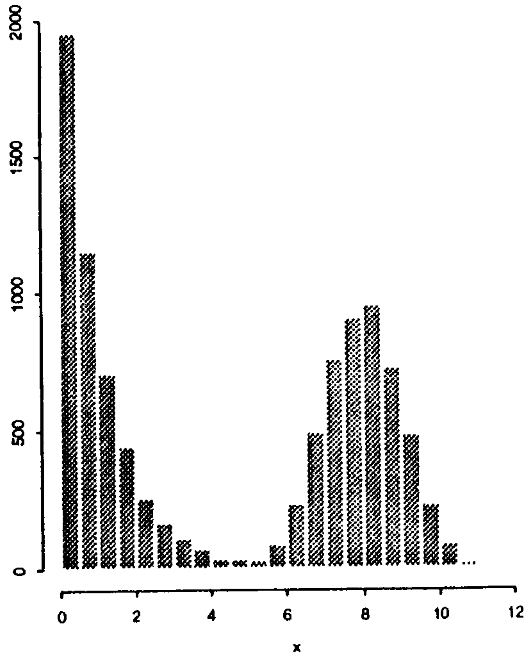
Skewed (Exponential) Shape



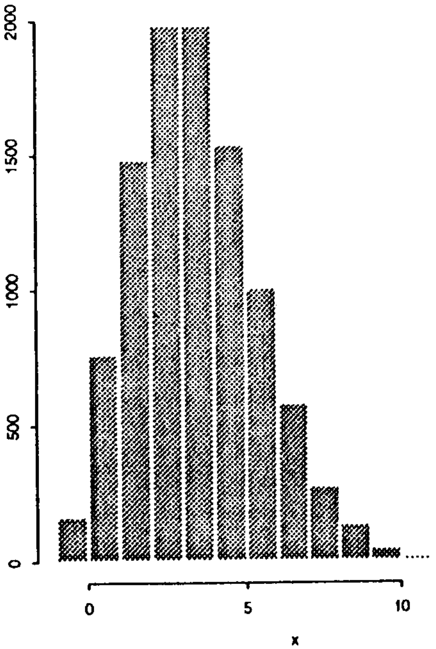
Two Outcome Shape



Bi-Modal Shape



Mildly Skewed Shape



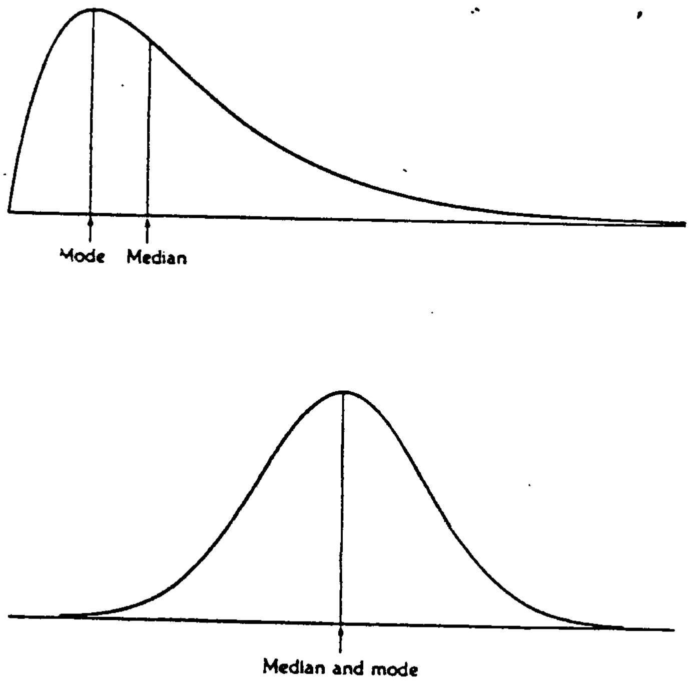


Figure 5-1 The mode and median of a frequency distribution. The mode is the point at which the frequency curve attains its highest value. The median is the point that divides the area under the curve into two equal parts to the left and to the right of it. The median is the center point of any symmetric frequency curve. This normal curve is highest at the center, so the center point is also the mode.

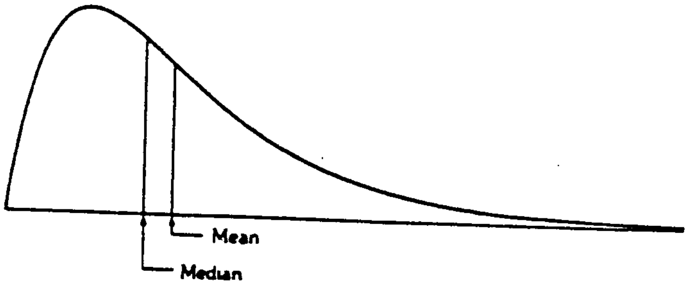


Figure 5-3 The mean and median of a skewed distribution. The mean is located farther toward the long tail of a skewed frequency curve than is the median.

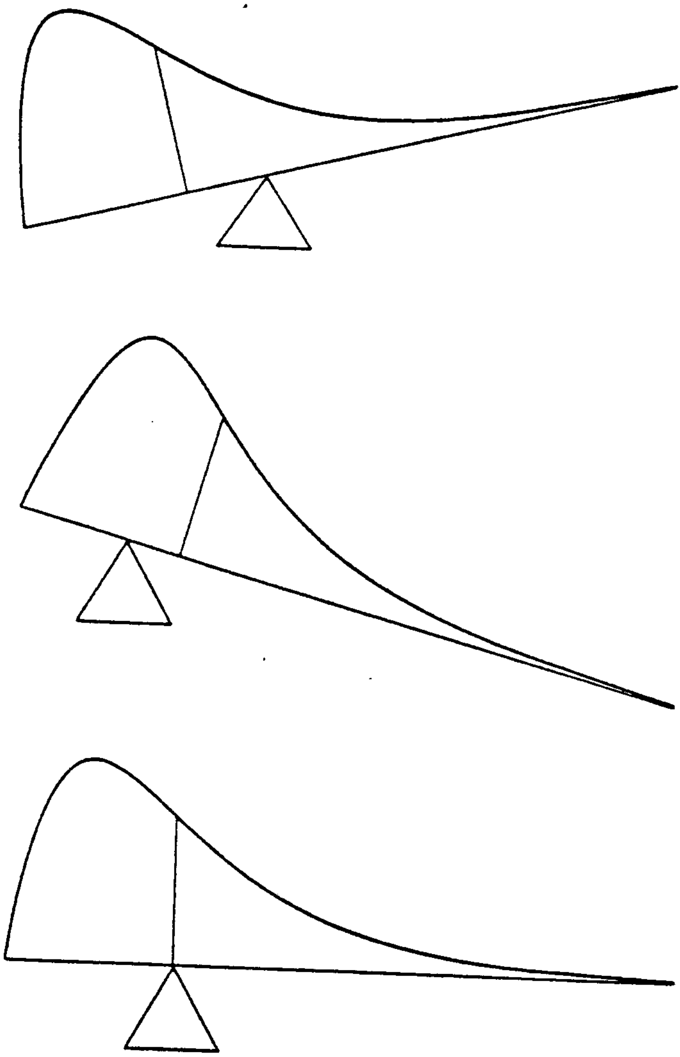


Figure 5-2 The mean of a frequency distribution. The mean is the cen of gravity of the frequency curve, the point about which the curve wou balance on a pivot placed beneath it.

Principles of Inferential Statistics in Medicine

Relocating and rescaling numbers

Suppose that we have a set of data, $X = \{x_1, x_2, \ldots, x_n\}$, which has mean $\overline{x}$ and variance s^2, or standard deviation s. If we transform each value in X by adding or multiplying by a constant, what happens to the mean and standard deviation of the "new" transformed sample? Call the new transformed variable Y=f(X), and refer to the following table:

Type of transformation	New mean	New standard deviation
Original (no transformation) $Y = X$	$\overline{x}$	s
Add the constant a $Y = X + a$	$\overline{x} + a$	s
Multiply by the constant b $Y = b \times X$	$b \times \overline{x}$	$b \times s$
Add then Multiply $Y = b \times (X + a)$	$b \times (\overline{x} + a)$	$b \times s$
Multiply then Add $Y = (b \times X) + a$	$(b \times \overline{x}) + a$	$b \times s$

These rules can all be "proved" using basic algebraic results.

Note: All of the above also applies to a population from a distribution, where X is a random sample from the population or density, and μ and σ^2, the population mean and variance replaces their data equivalents, $\overline{x}$ and s^2.

The above formulae are used at various places in biostatistics. One of the most common uses is in "standardizing" a variable, which means transforming it to have a mean of 0 and a standard deviation (and variance) of 1. Using the third rule above, if X has a mean μ_X and a standard deviation of σ_X, then letting $a = -\mu_X$ and $b = \frac{1}{\sigma_X}$, we will have

$$Y = \frac{X - \mu_X}{\sigma_X}$$

which has $\mu_Y = 0$ and $\sigma_Y = 1$.

Montreal Gazette, August 2000

Site probes Canada's brain drain

ANGELA PACIENZA
Canadian Press

TORONTO – While the debate over whether Canada is experiencing a true brain drain to the U.S. has gone on for years, the creator of a Web site thinks she has an easy way for people to discuss what's alluring and frightening about moving south.

With real-life stories on the brain drain, Martha Fusca said she believes www.canadasbraindrain.ca can determine what it takes to entice a Canadian to move to the U.S.

"We want to find out what's so intriguing about the U.S.," said Fusca, president of Stornoway Communications. "People should be talking and thinking about the issue in an interactive setting."

Visitors are asked to fill out a questionnaire about likes and dislikes of working in Canada or the U.S. For instance, would a $10,000 pay hike persuade you to move south? How about $20,000?

The results, says Fusca, will provide a more accurate picture of how the average Canadian sees the issue and what lures Canadians south.

Preliminary results, tabulated in real time by a built-in calculator, show that three of every 10 respondents would consider moving if offered $20,000 more than they make now. Nine out of every 10 would move if given $900,000 more.

Fourteen per cent of respondents who have thought about moving stay because they like Canada's "conservative" attitude, whereas 27 per cent said they like the more "risk-taking" attitude of Americans. Another 30 per cent say they prefer Canada's ethnic mix.

Fusca said the survey will paint a clearer picture of how Canadians feel about brain drain than previous studies because of the large number of Internet users. She expects the sample size will be large enough to surpass earlier studies within a couple of weeks.

"We hope the survey will provide us with the wherewithal of what kind of policy government should put in place," said Fusca. "We have talented people and we want to keep them here."

The site, born out of a documentary produced by Fusca in 1999, offers data on subjects such as green cards, immigration lawyers and costs of living, health care and taxation in the two countries. Visitors can also have their say on the site's discussion forum.

But while Fusca wants to learn more about the supposed brain drain, she says she's not oblivious to statistics that show Canada is gaining more brains than it's losing.

Statistics Canada reported in May that Canada receives far more highly educated immigrants from abroad every year than it loses. For every university grad who moved to the U.S., four came to Canada, the report said.

"We're not saying Canadians are leaving in droves, but there are a significant number going," she said.

Between 22,000 and 35,000 Canadians moved annually to the United States in the '90s, including about 10,000 university graduates, Statistics Canada said.

Fusca has toyed with the idea of moving south but her attachment to Canada, a country she came to from Italy as a child, is too strong.

"I hope the site encourages people to stay in Canada," she said. "I love this country so much."

Montreal Gazette,
Spring 2000

BACK T

Make visors mandatory

Bryan Berard's National Hockey League career came to a sudden and tragic end Saturday night and I can only hope that his misfortune will lead the NHL to adopt mandatory facial protection for every player.

The Toronto defenceman lost the sight in his right eye when he was struck by Ottawa's Marian Hossa's stick. Hossa received a double -minor penalty for high-sticking on the play but, by all accounts, this was an accident. Hossa went to take a shot, missed the puck and hit Berard in the face. He was rushed to hospital, but emergency surgery performed overnight failed to save his sight.

While doctors said there was a slim chance Berard would regain his sight, a promising career is over less than a week after Berard's 23rd birthday.

The hope now is that this incident will move the NHL and the NHL Players' Association to take the necessary steps to require that all players wear proper facial protection. Their neglect to do so in the past borders on the criminal.

If you think that's an overstatement, consider the view of Emile J. Therien, the president of the Canada Safety Council. In a telephone conversation from Ottawa yesterday, Therien said this was an issue of safety in the workplace. In any other industry, workers are required to wear protective equipment and the employer is liable if he doesn't enforce the use of that equipment.

PAT HICKEY

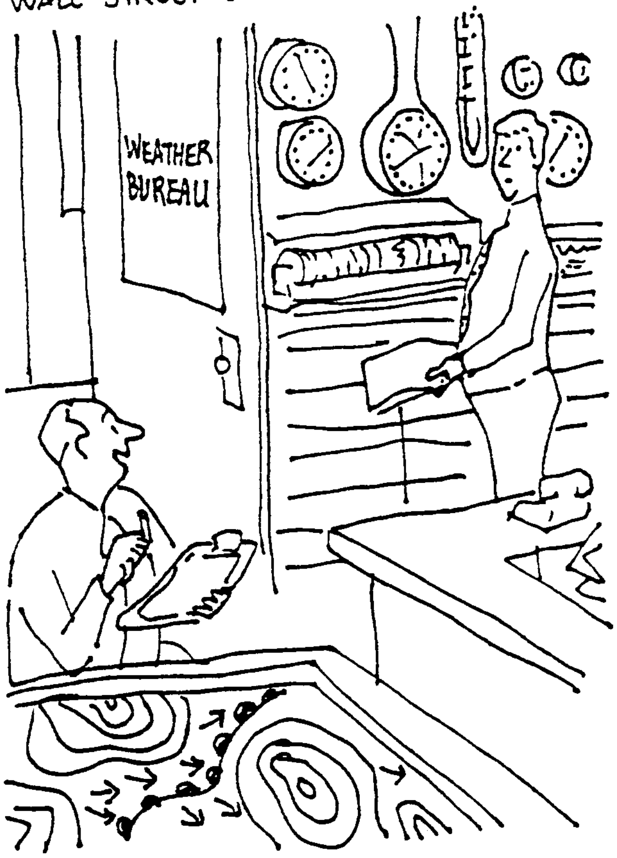
This is not a knee-jerk reaction by Therien. He has expressed this view in the past and he's eminently qualified to speak on the subject. His son, Chris, is a defenceman with the Philadelphia Flyers. Emile and his wife, Pat, have long been concerned because, as he so eloquently put it yesterday: "My dummy doesn't wear a visor."

Chris Therien is among the majority of the players in the NHL who don't wear visors despite the fact that there are dozens of facial and eye injuries each season. Statistics compiled by the team physicians in the NHL show that 95 per cent of those injuries are suffered by players who don't wear visors, despite the overwhelming statistical evidence that this equipment prevents injuries.

Please see **HICKEY**, *Page C2*

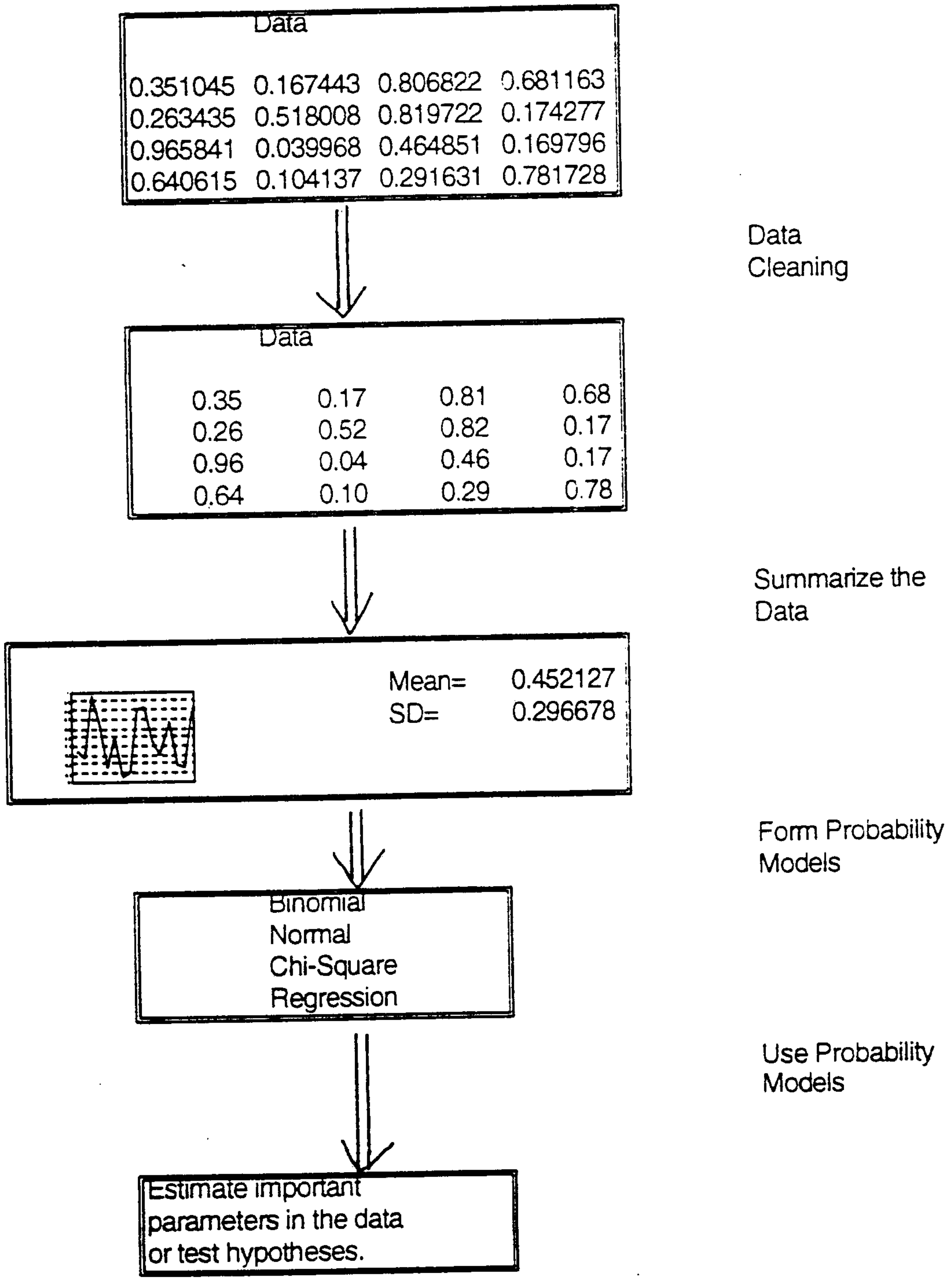
Vision in Bedard's right eye in doubt. *Page C2*

© WALL STREET JOURNAL



« I FIGURE THERE'S A 40%
CHANCE OF SHOWERS, AND
A 10% CHANCE WE KNOW
WHAT WE'RE TALKING ABOUT »

Overview



Data

0.351045	0.167443	0.806822	0.681163
0.263435	0.518008	0.819722	0.174277
0.965841	0.039968	0.464851	0.169796
0.640615	0.104137	0.291631	0.781728

Data Cleaning

Data

0.35	0.17	0.81	0.68
0.26	0.52	0.82	0.17
0.96	0.04	0.46	0.17
0.64	0.10	0.29	0.78

Summarize the Data

Mean= 0.452127
SD= 0.296678

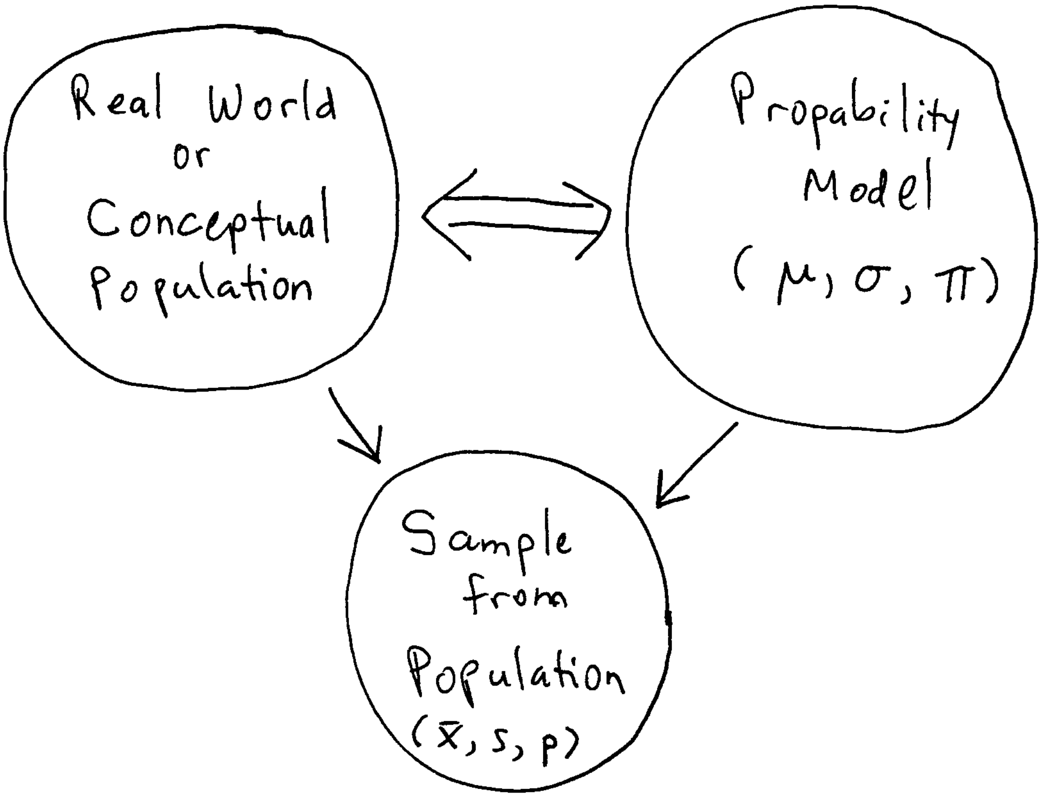
Form Probability Models

Binomial
Normal
Chi-Square
Regression

Use Probability Models

Estimate important parameters in the data or test hypotheses.

Populations and Samples

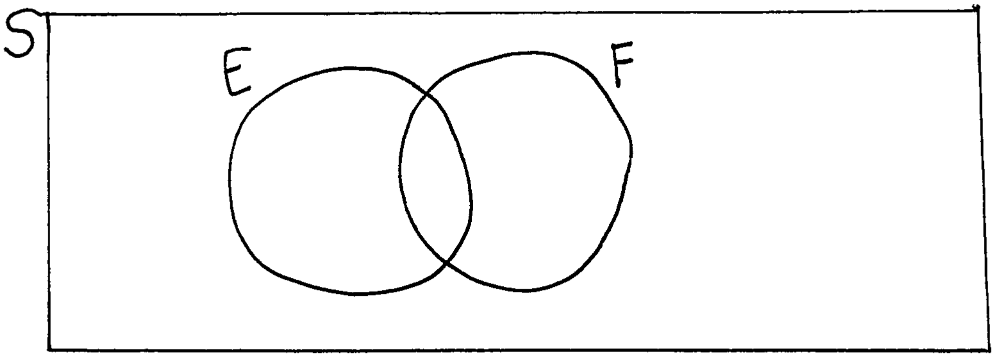


Notation: Letters from the Greek alphabet (μ, σ, π, etc.) will be used to denote the true population values. These values are typically unknown, and are the parameters of the probabilistic model. "Ordinary" letters ($\overline{x}$, s, p) will indicate the corresponding sample quantities.

Rules of Probability

The *sample space, S,* is the collection of all possible outcomes of an experiment. An event, E, is any subset of the sample space. If we let E be any event, then we have the following rules of probabilities:

1. $0 \leq P(E) \leq 1$, for any event E.
2. $P(S) = 1$
3. (Addition rule) If E and F are *disjoint* events, then $P(E \text{ or } F) = P(E) + P(F)$.
4. (Complement Rule) For any event E, $P(E^c) = 1 - P(E)$
5. (Multiplication Rule) If events E and F are *independent*, then $P(E \text{ and } F) = P(E) \times P(F)$.



Conditional Probability

The conditional probability of event E given that event F has happened, is defined to be

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)}.$$

This is interpreted as "Given that F has occurred, calculate the probability E will also occur." Note that we can also write

$$P(E \text{ and } F) = P(F) \times P(E|F),$$

even if E and F are not independent.

Some Examples

(1) The probability of surviving a certain transplant operation is 0.5. If a patient survives the operation, the probability that his or her body will reject the transplant within a month is 0.2. What is the probability of surviving both of these critical stages?

(2) If 40 percent of the mice used in an experiment will become very aggressive within one minute after having been administered an experimental drug, find the probability that exactly six of fifteen mice which have been administered the drug will become very aggressive within one minute, using

(a) the formula for the binomial distribution;

(b) Table E.

(c) What is the expected number of mice that will become agressive? Variance?

(3) In a certain community, 8 percent of all adults over 50 have diabetes. If a health service in this community correctly diagnoses 95 percent of all persons with diabetes as having the disease and incorrectly diagnoses 2 percent of all persons without diabetes as having the disease, find the probabilities that

(a) the community health service will diagnose an adult over 50 as having diabetes;

(b) a person over 50 diagnosed by the health service as having diabetes actually has the disease.

(4) Associated with three diseases A, B, and C, there are the respective chances .20, .30, and .50 of being hospitalized. If an individual has all three diseases (A, B, and C) and if these exert their influences independently, what is the chance of his being hospitalized?

(5) An investigator develops a screening test for cancer. He uses this screening test on known cancer patients and known noncancer patients, and he finds the test has a 5 percent false positive rate (i.e., positive test results for noncancer patients) and a 20 percent false negative rate (i.e., negative test results for cancer patients). He is now going to apply this test to a population in which he knows 2 percent have undetected cancer. Using Bayes' theorem, find the chance that someone with a positive test actually has cancer; also, find the chance that someone with a negative test actually has cancer. In considering the use of this test, what other (nonstatistical) issues are relevant?

(6) Test any drug on six patients. If none of the patients shows remission, reject the drug. Under this program, a drug which produces remissions in 20 percent of a large population of patients has about 3 chances in 4 of passing the screen, while one with a 30 percent remission rate has nearly nine chances in ten of passing the screen. Prove the above two statements.

Binomial Distribution

$$\underset{1/2}{\mathrm{H}} \qquad\qquad \underset{1/2}{\mathrm{T}}$$

$$\underset{1/4}{\mathrm{HH}} \qquad \underbrace{\mathrm{TH} \qquad \mathrm{HT}}_{1/2} \qquad \underset{1/4}{\mathrm{TT}}$$

$$\underset{1/8}{\mathrm{HHH}} \quad \underbrace{\mathrm{HHT} \quad \mathrm{HTH} \quad \mathrm{THH}}_{3/8} \quad \underbrace{\mathrm{HTT} \quad \mathrm{THT} \quad \mathrm{TTH}}_{3/8} \quad \underset{1/8}{\mathrm{TTT}}$$

$$Pr\{X \text{ heads in } N \text{ flips}\} = \frac{N!}{(N-X)!\,X!}(\frac{1}{2})^X(1-\frac{1}{2})^{(N-X)}$$

More generally,

$$Pr\{X \text{ successes in } N \text{ trials}\} = \frac{N!}{(N-X)!\,X!}(\pi)^X(1-\pi)^{(N-X)}$$

43

Entry is $P(X = k) = \binom{n}{k} p^k(1 - p)^{n-k}$

		p								
n	*k*	.01	.02	.03	.04	.05	.06	.07	.08	.09
2	0	.9801	.9604	.9409	.9216	.9025	.8836	.8649	.8464	.8281
	1	.0198	.0392	.0582	.0768	.0950	.1128	.1302	.1472	.1638
	2	.0001	.0004	.0009	.0016	.0025	.0036	.0049	.0064	.0081
3	0	.9703	.9412	.9127	.8847	.8574	.8306	.8044	.7787	.7536
	1	.0294	.0576	.0847	.1106	.1354	.1590	.1816	.2031	.2236
	2	.0003	.0012	.0026	.0046	.0071	.0102	.0137	.0177	.0221
	3				.0001	.0001	.0002	.0003	.0005	.0007
4	0	.9606	.9224	.8853	.8493	.8145	.7807	.7481	.7164	.6857
	1	.0388	.0753	.1095	.1416	.1715	.1993	.2252	.2492	.2713
	2	.0006	.0023	.0051	.0088	.0135	.0191	.0254	.0325	.0402
	3			.0001	.0002	.0005	.0008	.0013	.0019	.0027
	4									.0001
5	0	.9510	.9039	.8587	.8154	.7738	.7339	.6957	.6591	.6240
	1	.0480	.0922	.1328	.1699	.2036	.2342	.2618	.2866	.3086
	2	.0010	.0038	.0082	.0142	.0214	.0299	.0394	.0498	.0610
	3		.0001	.0003	.0006	.0011	.0019	.0030	.0043	.0060
	4						.0001	.0001	.0002	.0003
	5									
6	0	.9415	.8858	.8330	.7828	.7351	.6899	.6470	.6064	.5679
	1	.0571	.1085	.1546	.1957	.2321	.2642	.2922	.3164	.3370
	2	.0014	.0055	.0120	.0204	.0305	.0422	.0550	.0688	.0833
	3		.0002	.0005	.0011	.0021	.0036	.0055	.0080	.0110
	4					.0001	.0002	.0003	.0005	.0008
	5									
	6									
7	0	.9321	.8681	.8080	.7514	.6983	.6485	.6017	.5578	.5168
	1	.0659	.1240	.1749	.2192	.2573	.2897	.3170	.3396	.3578
	2	.0020	.0076	.0162	.0274	.0406	.0555	.0716	.0886	.1061
	3		.0003	.0008	.0019	.0036	.0059	.0090	.0128	.0175
	4				.0001	.0002	.0004	.0007	.0011	.0017
	5								.0001	.0001
	6									
	7									
8	0	.9227	.8508	.7837	.7214	.6634	.6096	.5596	.5132	.4703
	1	.0746	.1389	.1939	.2405	.2793	.3113	.3370	.3570	.3721
	2	.0026	.0099	.0210	.0351	.0515	.0695	.0888	.1087	.1288
	3	.0001	.0004	.0013	.0029	.0054	.0089	.0134	.0189	.0255
	4			.0001	.0002	.0004	.0007	.0013	.0021	.0031
	5							.0001	.0001	.0002
	6									
	7									
	8									

Entry is $P(X = k) = \binom{n}{k}p^k(1 - p)^{n-k}$

		p								
n	*k*	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5		.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937
	6			.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6		.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7				.0001	.0002	.0006	.0016	.0037	.0078
8	0	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313
	2	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6		.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7			.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8					.0001	.0002	.0007	.0017	.0039

Binomial Distribution in Practice

The assumptions behind the use of the binomial distribution may not always be perfectly satisfied in practice. For example:

Let X represent the number of females in four children, among all couples in Canada with exactly four children.

The "Real World" data and the data predicted by a binomial distribution model with $\pi = 4$ are:

X	Predicted	Observed
	Proportion	Proportion
0	0.0625	0.08
1	0.25	0.26
2	0.375	0.31
3	0.25	0.27
4	0.0625	0.08

For the predicted scores, we have used:

$$Pr\{X = k\} = \frac{4!}{k!(4-k)!}\pi^k(1-\pi)^{4-k} \ ,$$

where $\pi = 0.5$ and $k = 0, 1, 2, 3, 4$.

Why do you think that the observed values differ (slightly) from those predicted by a binomial model? Which assumption of the binomial model may be violated here?

Don McGillivray

It's easy to lie with statistics

OTTAWA — In the early 1950s, Darrell Huff wrote a book called *How to Lie With Statistics.*

It's still a classic exposure of tricks with numbers.

Huff unmasked the errors lurking in averages. A man with his head in a hot oven and his feet in a freezer may be suffering the tortures of the damned. But on average he's quite comfortable.

Huff pointed out the wrong conclusions produced by biased samples. *Literary Digest,* for example, forecast the election of Landon, the Republican presidential candidate over Roosevelt, the Democrat, in 1936 by a poll that involved phoning subscribers.

The folks who could still afford telephones and magazines in that Depression year tended to be Republicans. But there were more Democratic voters.

Numbers, Huff showed, can be made to dance to anyone's tune by the cunning manipulation of the base year for comparisons.

Let's take a modern example. Last year's federal budget deficit was $30.5 billion. You want that to look bad? Compare it with the $11.5-billion deficit 10 years earlier. The deficit has jumped by 165 per cent in a decade.

Hard to tell the truth

Do you want to make the $30.5 billion look good? Compare it with the $38.5 billion in 1984-85, the Mulroney government's first year in power. Suddenly the deficit is 20 per cent down, not 165 per cent up.

Lying with statistics is ridiculously easy. The hard thing is telling the truth with statistics.

The world is a complex and puzzling place. It's hard to tell cause from effect. It's hard to tell when something is just a coincidence and has nothing to do with something else.

Take, for instance, the statistics much bandied about these days about the number of days kids in various countries go to school in a year. Canadian schools are open about 185 days compared with 180 in the United States, 226 to 240 in Germany and 243 in Japan.

This is supposed to explain why schooling isn't what it used to be.

Now I don't know that it isn't. Older folks have been saying that kind of thing for as long as I can remember. My own contacts suggest today's kids are as able to cope with reading, writing and numbers as I was at their various ages.

But suppose there is a problem. The cause is probably a lot more complicated than the number of days in the school year.

The difficulty of telling the truth with statistics is also shown in attempts to debunk the myth that the 1950s and 1960s were the "good old days" compared with today's economic conditions. Of course, things were far from perfect and it's easy to show that per-capita income or wealth was lower then, even after allowing for inflation.

Follow a political line

But it's hard to capture in statistics the feeling of hope and confidence of the generation that came out of the Depression of the 1930s and saw evil put to flight in the 1940s.

It wasn't just a matter of what you had or how much you made. It was the feeling of an unlimited future. Economic mismanagement and rising taxes have cost us that future, a loss that cannot be measured in dollars.

Part of the damage is a loss of confidence in the people who should be trying to tell us the truth with statistics.

Statistics Canada still deserves credit for attempting truth-telling within its reduced resources. But other agencies of the federal government, and especially the Finance Department, seem to follow a political line.

Numbers put out by government agencies are not wrong. But a correct number can still be selected and manipulated to give a misleading impression.

The Goods and Services Tax Consumer Information Office put out a report the other day saying that, in the first three months of 1991, "savings from the elimination of the federal sales tax are being passed on to consumers."

Nowhere in the document was there a straight-out admission that the numbers were from the depth of the second-worst recession since the 1930s, a time of downward pressure on prices. This may not be lying. But is it telling the whole truth? □

Means and Variances for Probability Distributions
Discrete Case

Suppose that X is a discrete valued random variable taking on values in the set $\{a_1, a_2, \ldots, a_n\}$, with corresponding probabilities $\{p_1, p_2, \ldots, p_n\}$. The *expected value* or mean for a discrete variable is defined to be

$$E(X) = \sum_{i=1}^{n} a_i \times p_i \tag{1}$$

The variance of X is defined as

$$Var(X) = \sum_{i=1}^{n} (a_i - E(X))^2 \times p_i \tag{2}$$

Example 1: Suppose that entering a certain lottery costs \$1. The top prize is \$10,000,000, which has a $\frac{1}{13{,}983{,}816}$ chance of being won, while second prize is \$100,000, with probability of $\frac{1}{1{,}906{,}884}$, and third prize is \$100, which is won with a probability of $\frac{1}{18{,}424}$. What is the expected gain of someone who buys a ticket?

Note that there is a

$$1 - \frac{1}{13,983,816} - \frac{1}{1,906,884} - \frac{1}{18,424} = \frac{20974573}{20975724} \approx .9999451270$$

chance of winning nothing. Using the above formula, we calculate

$$\begin{aligned}
\text{Expected Winnings} &= (10,000,000 - 1) \times \frac{1}{13,983,816} + (100,000 - 1) \times \frac{1}{1,906,884} \\
&\quad +(100 - 1) \times \frac{1}{18,424} + (0 - 1) \times \frac{20974573}{20975724} \\
&= -\frac{103519}{455994} \approx -.2270183380.
\end{aligned}$$

Example 2: Find the expected value and variance for a binomial random variable, where there are 10 trials and the probability of sucess on each trial is $\pi = 0.6$.

The set of possible outcomes is

$$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

The probabilities for these outcomes, obtained either directly from binomial tables, or from the formula for binomial probabilities, are

$$\begin{aligned} \{ \; & 0.0001048576 & , 0.0015728640, 0.0106168320, 0.0424673280, 0.1114767360, \\ & 0.2006581248 & , 0.2508226560, 0.2149908480, 0.1209323520, 0.0403107840, \\ & 0.0060466176 & \}. \end{aligned}$$

Multiplying these two sets of numbers as in (1) produces $E(X) = 6$, and using (2), we find that $Var(X) = 2.4$.

Note: There are shortcut formulæ for binomial (and many other) random variables. Let X be a binomial random variable with n trials and probability of success π. Then:

1. $E(X) = n \times \pi$, and
2. $Var(X) = n \times \pi \times (1 - \pi)$

Bayes Theorem (Discrete Case)

Suppose we are considering a test for cancer (see Question 5 on a previous sheet).

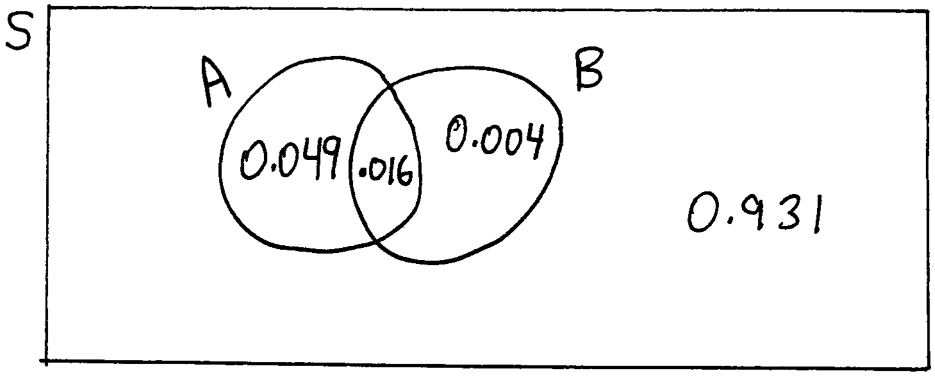
Let A = the event that a test is positive.
Let B = the event of actually having cancer.

Suppose we know that:

- $P(A|B^c) = 0.05$, and so $P(A^c|B^c) = 1 - 0.05 = 0.95$
- $P(A^c|B) = 0.20$, and so $P(A|B) = 1 - 0.20 = 0.80$
- $P(B) = 0.02$, and so $P(B^c) = 0.98$

(a) What is the probability of cancer given that the test is positive?
(b) What is the probability of cancer given that the test is negative?

We can draw a diagram as below:



From the diagram, we see that

$$P(B|A) = \frac{0.016}{0.016 + 0.049} = .2462$$

and

$$P(B|A^c) = \frac{0.004}{0.004 + 0.931} = .0043$$

Alternatively, we can use Bayes Theorem, which states:

$$P(B|A) = \frac{P(B) \times P(A|B)}{P(B) \times P(A|B) + P(B^c) \times P(A|B^c)}$$

Plugging in the numbers, we can check that the solutions are the same. For example,

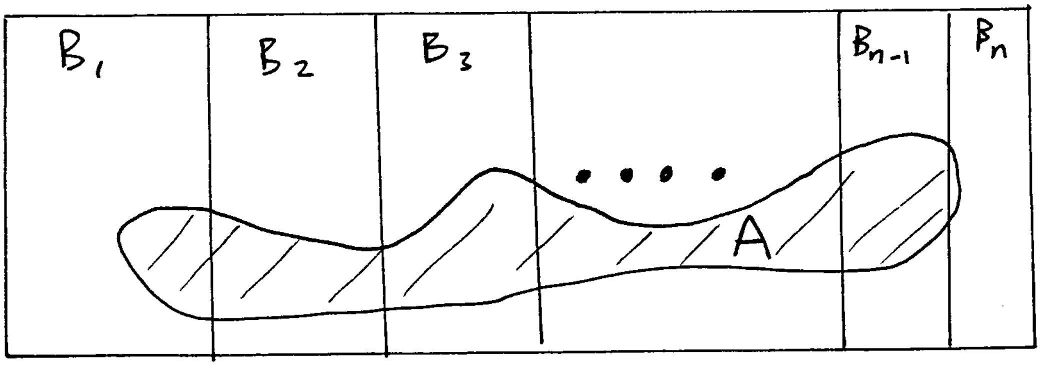
$$P(B|A) = \frac{P(B) \times P(A|B)}{P(B) \times P(A|B) + P(B^c) \times P(A|B^c)} = \frac{0.02 \times 0.80}{0.02 \times 0.80 + 0.98 \times 0.05} = .2462.$$

Switching the roles of A and A^c in the above formula yields

$$P(B|A^c) = \frac{P(B) \times P(A^c|B)}{P(B) \times P(A^c|B) + P(B^c) \times P(A^c|B^c)} = 0.0043$$

Note that before the test is performed, the probability that a person has cancer is 0.02, but that these probabilities are "updated" in a natural way, once the test results become available.

Bayes Theorem may be generalized to the case where the event B has more than two possible outcomes, say $B_1,\ B_2, \ldots, B_n$.



In this case, the Bayes Theorem is

$$P(B_k|A) = \frac{P(B_k) \times P(A|B_i)}{\sum_{i=1}^{n} P(B_i) \times P(A|B_i)},\ k = 1, 2, \ldots, n.$$

We will cover another extension of Bayes Theorem, to the case where B is a continuous outcome, later in the course.

A DICTIONARY OF
EPIDEMIOLOGY

SECOND EDITION

Edited for the
International Epidemiological Association
by

John M. Last

BINARY VARIABLE A variable having only two possible values, e.g. on or off, 0 or 1. See also BIT.

BINOMIAL DISTRIBUTION A probability distribution associated with two mutually exclusive outcomes, e.g., presence or absence of a clinical or laboratory sign, death, or survival. The probability distribution of the number of occurrences of a binary event in a sample of n independent observations. The binomial distribution is used to model CUMULATIVE INCIDENCE RATES and PREVALENCE RATES. The BERNOULLI DISTRIBUTION is a special case of the binomial distribution with $n = 1$.

PROBABILITY

1. The limit of the relative frequency of an event in a sequence of N random trials as N approaches infinity, i.e., the limit of

$$\frac{\text{Number of occurrence of the event}}{N}$$

2. A measure, ranging from zero to 1, of the degree of belief in a hypothesis or statement.

PRIOR PROBABILITY Probability calculated or estimated from theory or belief, before a study is done. See BAYES' THEOREM.

PROBABILITY DENSITY The frequency distribution of a continuous random variable.

PROBABILITY DISTRIBUTION For a discrete random variable, the function that gives the probabilities that the variable equals each of a sequence of possible values. Examples include the binomial and Poisson distributions. For a continuous random variable, often used synonymously with the probability density function.

PROBABILITY SAMPLE (Syn: random sample) See SAMPLE.

PROBABILITY THEORY The branch of mathematics dealing with the purely logical properties of probability. Its theorems underly most statistical methods.

Diagnostic Tests

SENSITIVITY AND SPECIFICITY (of a screening test) *Sensitivity* is the proportion of truly diseased persons in the screened population who are identified as diseased by the screening test. Sensitivity is a measure of the probability of correctly diagnosing a case, or the probability that any given case will be identified by the test (Syn: true positive rate).

Specificity is the proportion of truly nondiseased persons who are so identified by the screening test. It is a measure of the probability of correctly identifying a nondiseased person with a screening test (Syn: true negative rate). The relationships are shown in the following fourfold table, in which the letters *a, b, c,* and *d* represent the quantities specified below the table.

Screening test results	True status		TOTAL
	Diseased	Not diseased	
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

a. Diseased individuals detected by the test (true positives)
b. Nondiseased individuals positive by the test (false positives)
c. Diseased individuals not detectable by the test (false negatives)
d. Nondiseased individuals negative by the test (true negatives)

$$\text{Sensitivity} = \frac{a}{a+c} \qquad \text{Specificity} = \frac{d}{b+d}$$

$$\text{Predictive value (positive test result)} = \frac{a}{a+b}$$

$$\text{Predictive value (negative test result)} = \frac{d}{c+d}$$

FALSE NEGATIVE Negative test result in a subject who possesses the attribute for which the test is conducted. The labeling of a diseased person as healthy when screening in the detection of disease. See also SCREENING; SENSITIVITY AND SPECIFICITY.

FALSE POSITIVE Positive test result in a subject who does not possess the attribute for which the test is conducted. The labeling of a healthy person as diseased when screening in the detection of disease. See also SCREENING; SENSITIVITY AND SPECIFICITY.

PREDICTIVE VALUE In screening and diagnostic tests, the probability that a person with a positive test is a true positive (i.e., does have the disease) is referred to as the "predictive value of a positive test." The predictive value of a negative test is the probability that a person with a negative test does not have the disease. The predictive value of a screening test is determined by the sensitivity and specificity of the test, and by the prevalence of the condition for which the test is used. See also SCREENING; SENSITIVITY AND SPECIFICITY.

Diagnostic Tests

Prevalence Rate=10%

Truth

		Diseased	Non-Diseased	
Test	+	80	90	170
	−	20	810	830
		100	900	1000

Prevalence Rate=2%

Truth

		Diseased	Non-Diseased	
Test	+	16	98	114
	−	4	882	886
		20	980	1000

Prevalence Rate=50%

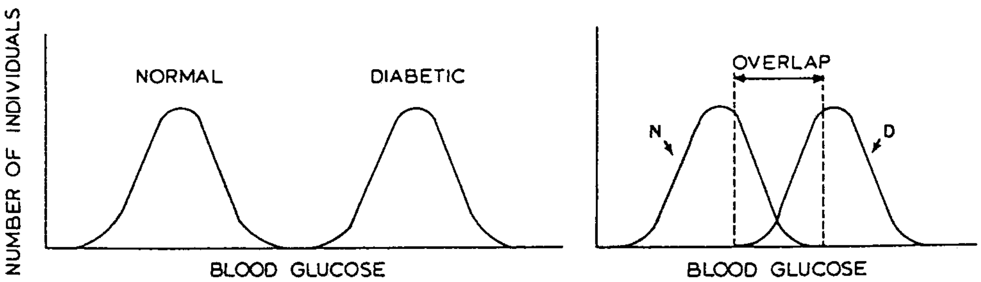
Truth

		Diseased	Non-Diseased	
Test	+	400	50	450
	−	100	450	550
		500	500	1000

Continuous Diagnostic Tests

Many tests record outcomes on a continuous scale, rather than just providing "positive" and "negative" test results. Examples include blood pressure, cholesterol levels, bone mineral density, and blood glucose levels. In this case, a cut-off limit must be chosen in order to classify individuals into "positive" or "negative" categories.

NORMAL LIMITS The limits of the "normal" range of a test or measurement, in the sense of being indicative of or conducive to good health. One way to determine normal limits is to compare the values obtained when the measurements are made in two groups, one that is healthy and has been found to remain healthy, the other ill, or subsequently found to become ill. The result may be two overlapping distributions, as illustrated. Outside the area where the distributions overlap, a given value clearly identifies the presence or absence of disease or some other manifestation of poor health. If a value falls into the area of overlap, the individual may belong to either the normal or the abnormal group. The choice of the normal limits depends upon the relative importance attached to the identification of individuals as healthy or unhealthy. See also FALSE NEGATIVE; FALSE POSITIVE; SENSITIVITY AND SPECIFICITY.

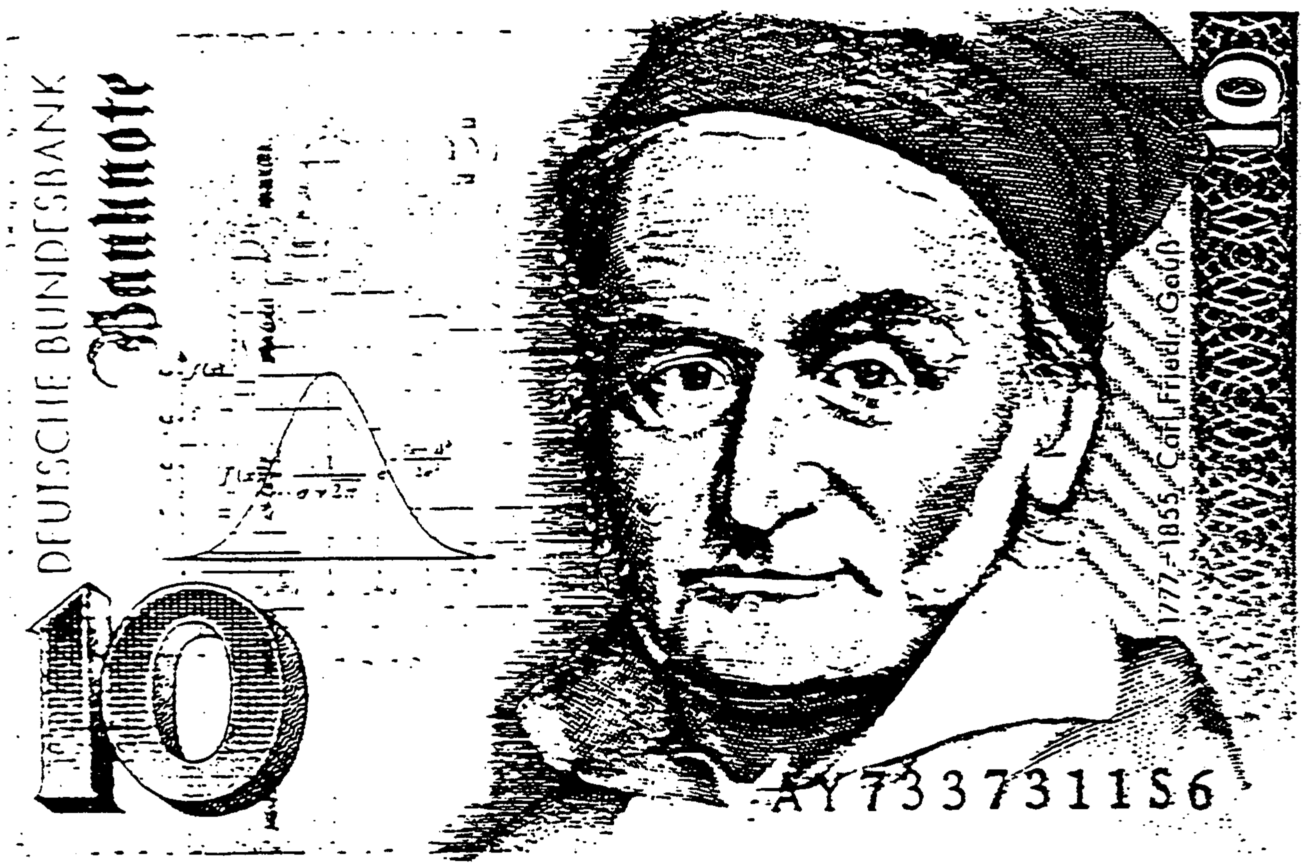


Hypothetical distribution of normal and diabetic glucose levels.
From Lilienfeld and Lilienfeld, 1979.

Selecting different cut-off values provides different sensitivity and specificity levels. Note that in this setup, one cannot increase sensitivity without simultaneously decreasing specificity, and vice versa. We will see an analogous situation arising in classical hypothesis testing.

The Normal Distribution

Germany honors Carl F. Gauss on a 10 Deutsche Mark bill. The Normal Distribution is often called the Gaussian Distribution. Gauss also may have been the first to use least squares regression to fit a model to data, among many other accomplishments.



NORMAL DISTRIBUTION (Syn: Gaussian distribution) The continuous frequency distribution of infinite range represented by the equation

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/2\sigma^2}$$

where x is the abscissa, $f(x)$ is the ordinate, μ is the mean, l is the natural logarithm, 2.718 and σ the standard deviation.

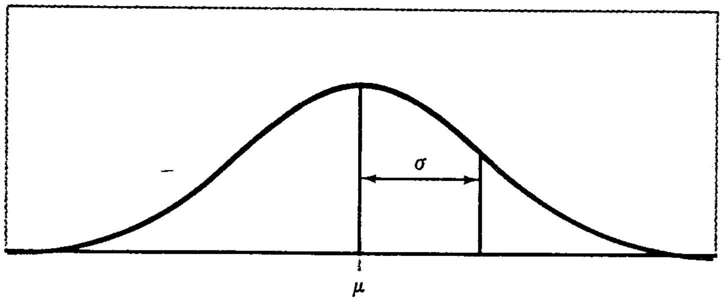
The Normal Distribution

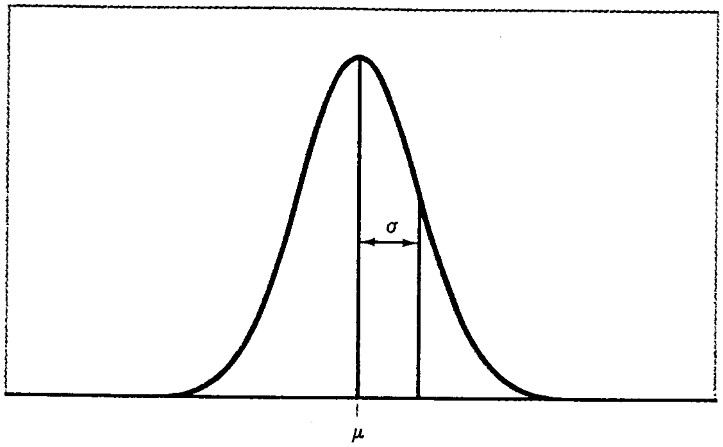
Density:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

Area under the Normal curve:

$$Pr\{a \le x \le b\} = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2} dx$$





The Normal Distribution

Integration of the Normal density to find area under the curve is difficult (you may recall from calculus that integration of exp $(-x^2)$ is difficult), so specially constructed tables are usually used. Increasingly, computer programs are used rather than tables, which provide the areas under the curve using numerical algorithms devised for this purpose. The next four pages provide examples of such tables.

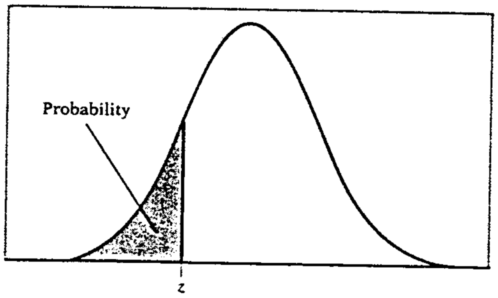


Table entry is probability at or below *z*.

Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

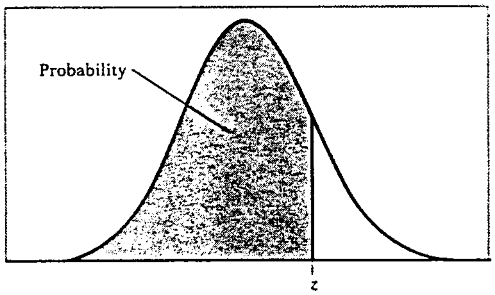
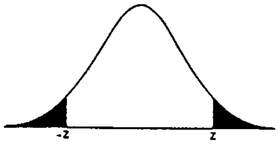


Table entry is probability at or below *z*.

Standard normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

This table shows
the shaded areas

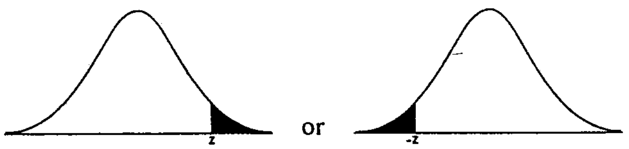


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.000	.992	.984	.976	.968	.960	.952	.944	.936	.928
0.1	.920	.912	.904	.897	.889	.881	.873	.865	.857	.849
0.2	.841	.834	.826	.818	.810	.803	.795	.787	.779	.772
0.3	.764	.757	.749	.741	.734	.726	.719	.711	.704	.697
0.4	.689	.682	.674	.667	.660	.653	.646	.638	.631	.624
0.5	.617	.610	.603	.596	.589	.582	.575	.569	.562	.555
0.6	.549	.542	.535	.529	.522	.516	.509	.503	.497	.490
0.7	.484	.478	.472	.465	.459	.453	.447	.441	.435	.430
0.8	.424	.418	.412	.407	.401	.395	.390	.384	.379	.373
0.9	.368	.363	.358	.352	.347	.342	.337	.332	.327	.322
1.0	.317	.312	.308	.303	.298	.294	.289	.285	.280	.276
1.1	.271	.267	.263	.258	.254	.250	.246	.242	.238	.234
1.2	.230	.226	.222	.219	.215	.211	.208	.204	.201	.197
1.3	.194	.190	.187	.184	.180	.177	.174	.171	.168	.165
1.4	.162	.159	.156	.153	.150	.147	.144	.142	.139	.136
1.5	.134	.131	.129	.126	.124	.121	.119	.116	.114	.112
1.6	.110	.107	.105	.103	.101	.099	.097	.095	.093	.091
1.7	.089	.087	.085	.084	.082	.080	.078	.077	.075	.073
1.8	.072	.070	.069	.067	.066	.064	.063	.061	.060	.059
1.9	.057	.056	.055	.054	.052	.051	.050	.049	.048	.047
2.0	.046	.044	.043	.042	.041	.040	.039	.038	.038	.037
2.1	.036	.035	.034	.033	.032	.032	.031	.030	.029	.029
2.2	.028	.027	.026	.026	.025	.024	.024	.023	.023	.022
2.3	.021	.021	.020	.020	.019	.019	.018	.018	.017	.017
2.4	.016	.016	.016	.015	.015	.014	.014	.014	.013	.013
2.5	.012	.012	.012	.011	.011	.011	.010	.010	.010	.010
2.6	.009	.009	.009	.009	.008	.008	.008	.008	.007	.007
2.7	.007	.007	.007	.006	.006	.006	.006	.006	.005	.005
2.8	.005	.005	.005	.005	.005	.004	.004	.004	.004	.004
2.9	.004	.004	.004	.003	.003	.003	.003	.003	.003	.003
3.0	.003									

60

Table A1. Areas in one tail of the standard normal curve

This table shows
the shaded area



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.500	.496	.492	.488	.484	.480	.476	.472	.468	.464
0.1	.460	.456	.452	.448	.444	.440	.436	.433	.429	.425
0.2	.421	.417	.413	.409	.405	.401	.397	.394	.390	.386
0.3	.382	.378	.374	.371	.367	.363	.359	.356	.352	.348
0.4	.345	.341	.337	.334	.330	.326	.323	.319	.316	.312
0.5	.309	.305	.302	.298	.295	.291	.288	.284	.281	.278
0.6	.274	.271	.268	.264	.261	.258	.255	.251	.248	.245
0.7	.242	.239	.236	.233	.230	.227	.224	.221	.218	.215
0.8	.212	.209	.206	.203	.200	.198	.195	.192	.189	.187
0.9	.184	.181	.179	.176	.174	.171	.169	.166	.164	.161
1.0	.159	.156	.154	.152	.149	.147	.145	.142	.140	.138
1.1	.136	.133	.131	.129	.127	.125	.123	.121	.119	.117
1.2	.115	.113	.111	.109	.107	.106	.104	.102	.100	.099
1.3	.097	.095	.093	.092	.090	.089	.087	.085	.084	.082
1.4	.081	.079	.078	.076	.075	.074	.072	.071	.069	.068
1.5	.067	.066	.064	.063	.062	.061	.059	.058	.057	.056
1.6	.055	.054	.053	.052	.051	.049	.048	.048	.046	.046
1.7	.045	.044	.043	.042	.041	.040	.039	.038	.038	.037
1.8	.036	.035	.034	.034	.033	.032	.031	.031	.030	.029
1.9	.029	.028	.027	.027	.026	.026	.025	.024	.024	.023
2.0	.023	.022	.022	.021	.021	.020	.020	.019	.019	.018
2.1	.018	.017	.017	.017	.016	.016	.015	.015	.015	.014
2.2	.014	.014	.013	.013	.013	.012	.012	.012	.011	.011
2.3	.011	.010	.010	.010	.010	.009	.009	.009	.009	.008
2.4	.008	.008	.008	.008	.007	.007	.007	.007	.007	.006
2.5	.006	.006	.006	.006	.006	.005	.005	.005	.005	.005
2.6	.005	.005	.004	.004	.004	.004	.004	.004	.004	.004
2.7	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003
2.8	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002
2.9	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001
3.0	.001									

THE 68–95–99.7 RULE

In any normal distribution:

- 68% of the observations fall within σ of the mean μ.
- 95% of the observations fall within 2σ of μ.
- 99.7% of the observations fall within 3σ of μ.

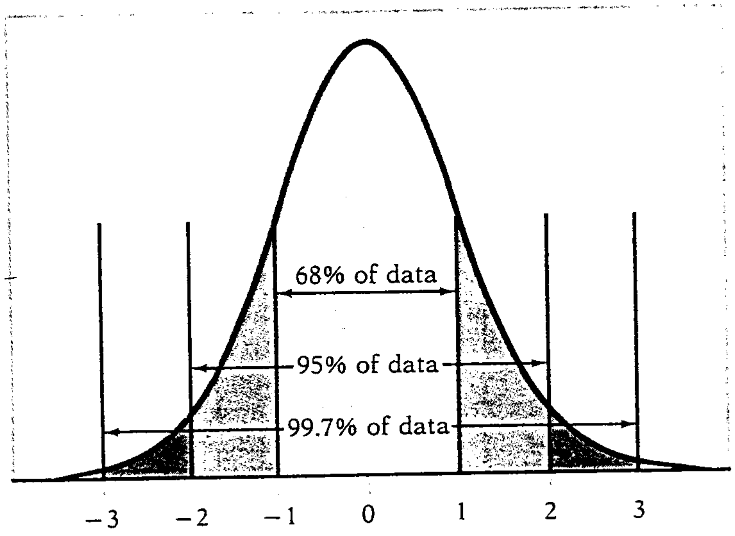


FIGURE 1.16 The 68–95–99.7 rule for normal distributions.

STANDARD NORMAL DISTRIBUTION

If a variable X has a normal distribution with mean μ and standard deviation σ, then the standardized variable

$$Z = \frac{X - \mu}{\sigma} \qquad (1.5)$$

has the normal distribution $N(0, 1)$ with mean 0 and standard deviation 1. This is called the standard normal distribution.

62 Examples on Normal Inferences (from M&M)

5.23 The law requires coal mine operators to test the amount of dust in the atmosphere of the mine. A laboratory carries out the test by weighing filters that have been exposed to the air in the mine. The test has a standard deviation of $\sigma = 0.08$ milligram in repeated weighings of the same filter. The laboratory weighs each filter 3 times and reports the mean result. What is the standard deviation of the reported result?

5.29 Judy's doctor is concerned that she may suffer from hypokalemia (low potassium in the blood). There is variation both in the actual potassium level and in the blood test that measures the level. Judy's measured potassium level varies according to the normal distribution with $\mu = 3.8$ and $\sigma = 0.2$. A patient is classified as hypokalemic if the potassium level is below 3.5.

(a) If a single potassium measurement is made, what is the probability that Judy is diagnosed as hypokalemic?

(b) If measurements are made instead on 4 separate days and the mean result is compared with the criterion 3.5, what is the probability that Judy is diagnosed as hypokalemic?

5.41 A study of working couples measures the income X of the husband and the income Y of the wife in a large number of couples in which both partners are employed. Suppose that you knew the means μ_X and μ_Y and the variances σ_X^2 and σ_Y^2 of both variables in the population.

(a) Is it reasonable to take the mean of the total income $X + Y$ to be $\mu_X + \mu_Y$? Explain your answer.

(b) Is it reasonable to take the variance of the total income to be $\sigma_X^2 + \sigma_Y^2$? Explain your answer.

5.43 The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This distribution is discrete and so is certainly not normal.

(a) Let $\bar{x}$ be the mean number of accidents per week at the intersection during a year (52 weeks). What is the approximate distribution of $\bar{x}$ according to the central limit theorem?

(b) What is the approximate probability that $\bar{x}$ is less than 2?

(c) What is the approximate probability that there are fewer than 100 accidents at the intersection in a year? (Hint: Restate this event in terms of $\bar{x}$.)

5.45 The level of nitrogen oxide (NOX) in the exhaust of a particular car model varies with mean 1.4 g/mile and standard deviation 0.3 g/mile. A company has 125 cars of this model in its fleet. If $\bar{x}$ is the mean NOX emission level for these cars, what is the level L such that the probability that $\bar{x}$ is greater than L is only 0.01?

Means and Variances for Probability Distributions Continuous Case

Suppose that x is a continuous valued random variable taking on values in the range $(-\infty, +\infty)$, with probability density function $f(x)$. The *expected value* or mean for a continuous variable is defined to be

$$E(X) = \mu = \int_{-\infty}^{+\infty} x \times f(x)\; dx \tag{1}$$

The variance of x is defined as

$$Var(x) = \sigma^2 = \int_{-\infty}^{+\infty} (x - E(X))^2 \times f(x)\; dx \tag{2}$$

Example 1: The Uniform Distribution. Suppose that

$$f(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

Then using (1),

$$\begin{aligned} E(x) &= \int_{-\infty}^{+\infty} x \times f(x)\; dx \\ &= \int_{0}^{1} x \times 1\; dx \\ &= \frac{1}{2} \times x^2 \Big|_{0}^{+1} \\ &= \frac{1}{2} \times (1-0) = \frac{1}{2} \end{aligned}$$

Similarly, using (2), the variance is

$$\begin{aligned} Var(x) &= \int_{-\infty}^{+\infty} (x - E(X))^2 \times f(x)\ dx \\ &= \int_{0}^{+1} (x - \frac{1}{2})^2 \times 1\ dx \\ &= \int_{0}^{1} (x^2 - x + \frac{1}{4})\ dx \\ &= (\frac{1}{3}x^3 - \frac{1}{2}x^2 + \frac{x}{4})\Big|_{0}^{+1} \\ &= \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12} = 0.08333333. \end{aligned}$$

Thus the standard deviation of a Uniform random variable is $\sqrt{\frac{1}{12}} = 0.288675$.

Example 2: Find the expected value and variance for a Normal random variable, with parameters μ and σ^2.

By definition, the density is

$$\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad -\infty < x < +\infty.$$

Thus, using (1), we have

$$\begin{aligned} E(x) &= \int_{-\infty}^{+\infty} x \times f(x)\ dx \\ &= \int_{-\infty}^{+\infty} x \times \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)\ dx \\ &= \mu \quad \text{(after much algebraic manipulations)} \end{aligned}$$

Similarly, using (2)

$$
\begin{aligned}
Var(x) &= \int_{-\infty}^{+\infty} (x - E(X))^2 \times f(x)\; dx \\
&= \int_{-\infty}^{+\infty} (x - \mu)^2 \times \frac{1}{\sqrt{2\pi}\sigma} \exp\,(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})\; dx \\
&= \sigma^2 \quad \text{(after much algebraic manipulations)}
\end{aligned}
$$

Note: The two parameters for a Normal distribution, μ and σ^2, are in fact the true mean and variance for that distribution. This is not always the case. For example, the Uniform has two parameters, a and b, marking the ends of the interval (so that $a = 0$ and $b = 1$ in the above), but these parameters are not directly the mean and variance, although they are related. For another example, a Gamma distribution, another very common distribution in medicine, has two parameters, α and β, but the mean is $\alpha \times \beta$, and the variance is $\alpha \times \beta^2$.

Sums of Random Variables

Let X and Y be two arbitrary random variables. Then:

1. $E(X+Y) = E(X) + E(Y)$.
2. $Var(X+Y) = Var(X) + Var(Y)$, if X and Y are independent.
3. $E(aX+bY) = aE(X) + bE(Y)$.
4. $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y)$, if X and Y are independent.
5. If $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$, and X and Y are independent, then $(X+Y) \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

Some examples:

1. If $X \sim N(0,1)$, $Y \sim N(3,4)$, and X and Y are independent, then
$$X + Y \sim N(3,5).$$
2. If $X_1, X_2, \ldots, X_n \sim N(0,1)$, independent, then
$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n \sim N(0,n).$$
3. . . . and then
$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N(0, \frac{1}{n}).$$
4. If $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, independent, then
$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2).$$
5. . . . and then
$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim N(\mu, \frac{\sigma^2}{n}).$$

The central limit theorem

The sampling distribution of $\bar{x}$ is normal if the underlying population itself has a normal distribution. What happens when the population distribution is not normal? It turns out that *as the sample size increases, the distribution of $\bar{x}$ becomes closer to a normal distribution*. This is true no matter what the population distribution may be, as long as the population has a finite standard deviation σ. This famous fact of probability theory is called the *central limit theorem*.* For large sample size n, we can regard $\bar{x}$ as having the $N(\mu, \sigma/\sqrt{n})$ distribution.

Moore & McCabe

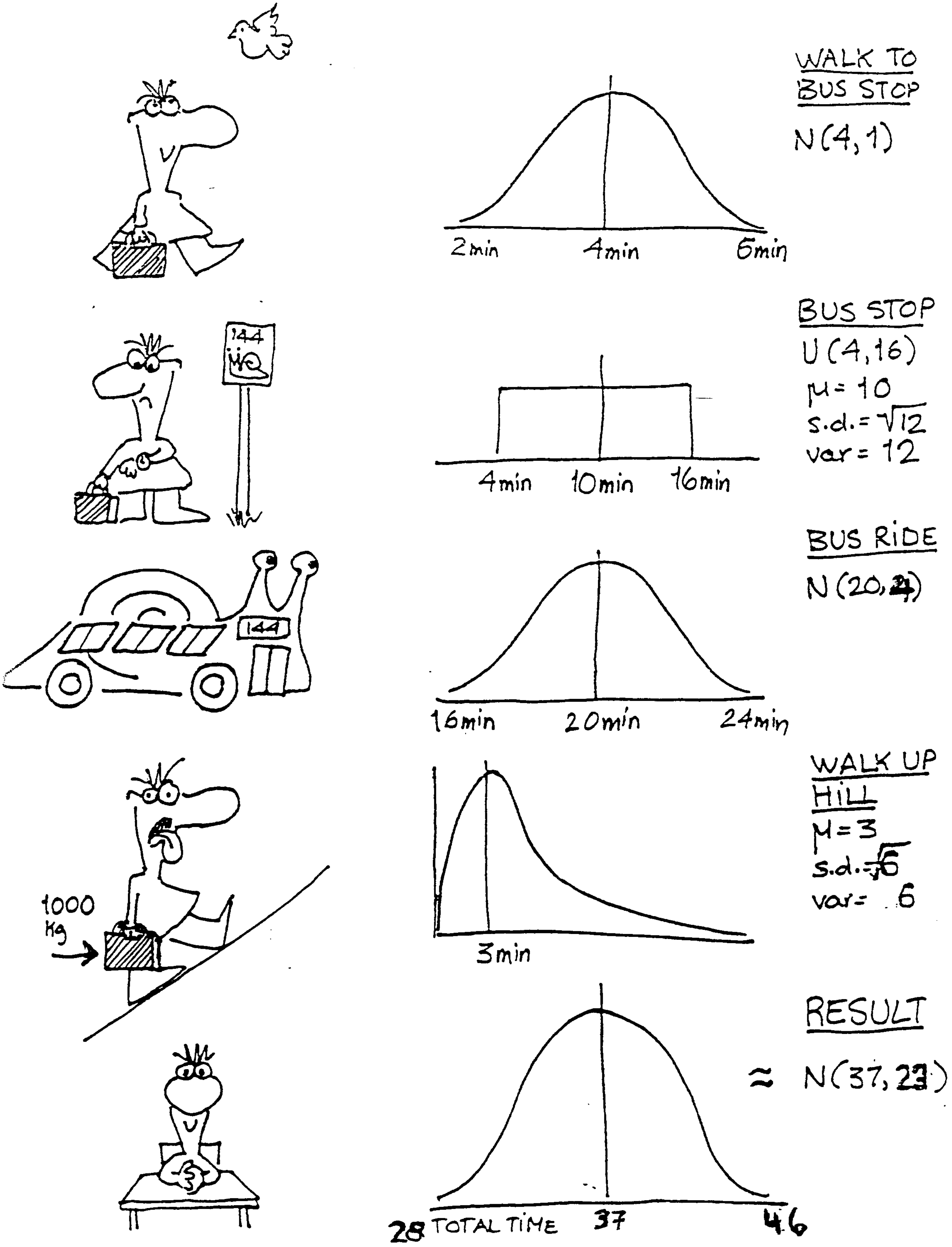
1. The mean of the sampling distribution of means is the same as the population mean, μ.
2. The SD of the sampling distribution of means is $\sigma/\sqrt{n}$.
3. The shape of the sampling distribution of means is approximately a normal curve, regardless of the shape of the population distribution and provided n is large enough.

Colton

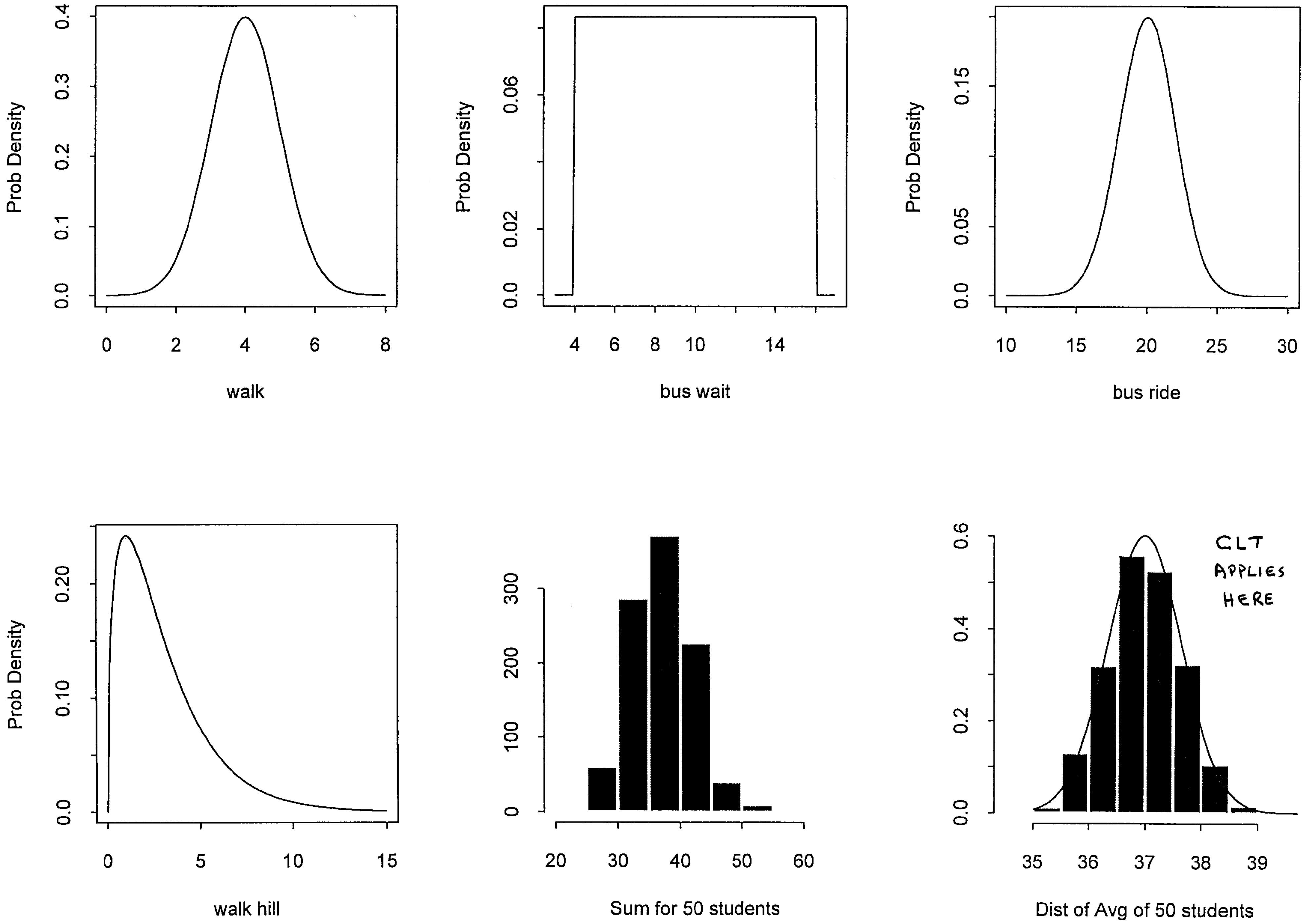
3 If the distribution of x is normal, so will be the distribution of $\bar{x}$. Much more importantly, even if the distribution of x is not normal, that of $\bar{x}$ will become closer and closer to the normal distribution with mean μ and variance σ^2/n as n gets larger. This is a consequence of a mathematical result known as the *central limit theorem*, and it accounts for the central importance of the normal distribution in statistics.

Armitage & Berry

CENTRAL LIMIT THEOREM *IN ACTION*

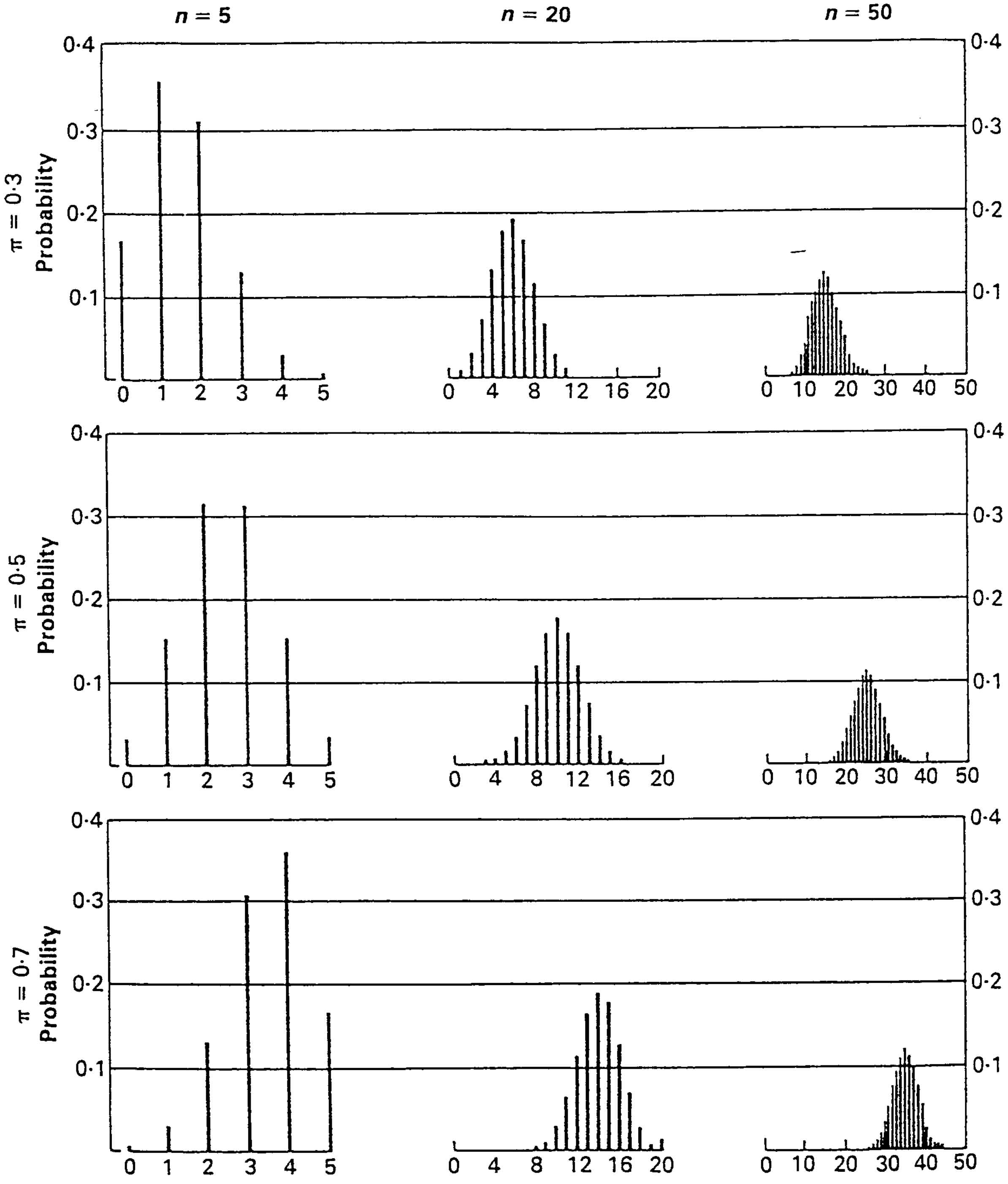


Central Limit Theorem in Action



CLT APPLIES HERE

70



n = 5
n = 20
n = 50
π = 0·3
π = 0·5
π = 0·7
Probability
0·1
0·2
0·3
0·4
0
1
2
3
4
5
0
4
8
12
16
20
0
10
20
30
40
50

Normal Approximation to the Binomial Distribution

Recall that the formula for the binomial distribution is given by

$$Pr\{X \text{ successes in } N \text{ trials}\} = \frac{N!}{(N-X)!\,X!}(\pi)^X(1-\pi)^{(N-X)}.$$

We can either calculate this directly, by plugging numbers in the above formula, or look up the result in tables of the binomial distribution. What happens, however, if we wish to know the probability of getting $X = 80$ or more successes in $N = 150$ trials, with $\pi = 0.6$? Binomial tables do not generally go that high, and calculations seem infeasible, as, for example, 150! is a 263 digit number, and 0.6^{80} is a very small number, and most calculators/computer programs do not handle these numbers very well. In addition, one would have to sum 70 of these numbers to get the final answer.

As the graphs on the previous page seems to indicate, we can approximate the binomial probabilities by a Normal distribution, and then look up the probabilities using tables of Normal probabilities. We proceed as follows:

1. Find the mean and variance of the binomial distribution of interest. In the above example,

   $$\mu = N \times \pi = 150 \times 0.6 = 90$$

   and

   $$\sigma^2 = N \times \pi \times (1-\pi) = 150 \times 0.6 \times 0.4 = 36.$$

2. Then perform the following calculation:

$$\begin{aligned}
Pr\{x \geq 80|binomial(150, 0.6)\} &= Pr\{x \geq 79.5|binomial(150, 0.6)\} \\
&\approx Pr\{x \geq 79.5|N(90, 36)\} \\
&= Pr\{\frac{x-90}{6} \geq \frac{79.5-90}{6}|N(90, 36)\} \\
&= Pr\{Z \geq -1.75|Z \sim N(0, 1)\} \\
&= 0.9599
\end{aligned}$$

The change from 80 to 79.5 is called the *continuity correction*, and it is used to make the approximation slightly more accurate. Without the continuity correction, in this example, the probability would have been 0.9525.

The final answer is usually quite similar with or without the continuity correction, but there is one place where it is absolutely crucial, as the example below illustrates.

Find the probability of exactly $X = 80$ successes out of $N = 150$ trials, where $\pi = 0.6$.

Using the same logic as above, without the continuity correction, we have

$$\begin{aligned} Pr\{x = 80|binomial(150, 0.6)\} &\approx Pr\{x = 80|N(90, 36)\} \\ &= Pr\{\frac{x-90}{6} = \frac{80-90}{6}|N(90, 36)\} \\ &= Pr\{Z = -1.6667|Z \sim N(0, 1)\} \\ &= 0, \end{aligned}$$

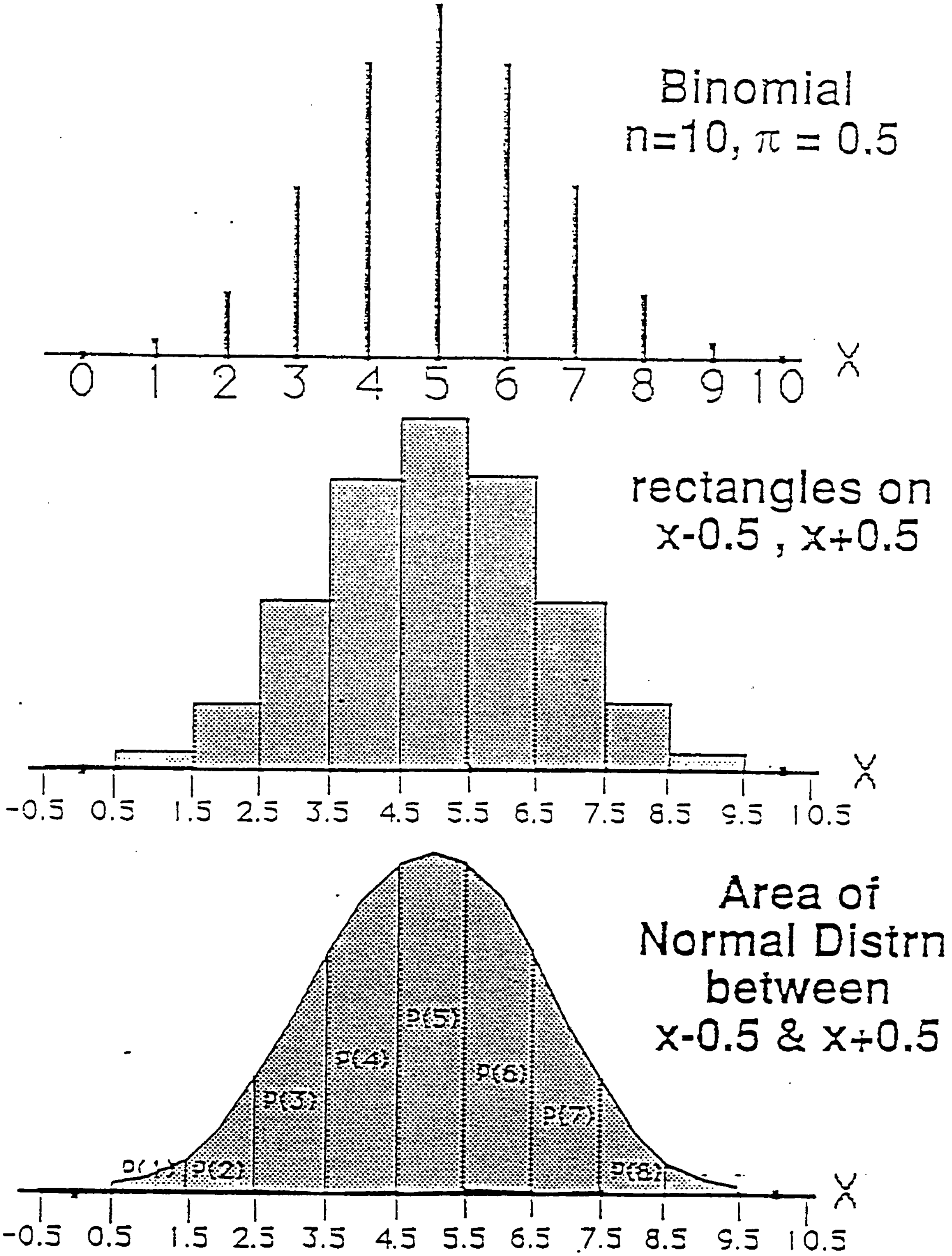
since the probabilty that a Normal variable equals any number exactly is zero.

However, with the continuity correction, we have

$$\begin{aligned} Pr\{x = 80|binomial(150, 0.6)\} &= Pr\{79.5 \le x \le 80.5|binomial(150, 0.6)\} \\ &\approx Pr\{79.5 \le x \le 80.5|N(90, 36)\} \\ &= Pr\{\frac{79.5-90}{6} \le \frac{x-90}{6} \le \frac{80.5-90}{6}|N(90, 36)\} \\ &= Pr\{-1.75 \le Z \le -1.5833|Z \sim N(0, 1)\} \\ &= 0.9599 - .9430 \\ &= 0.0169. \end{aligned}$$

The exact answer, using the binomial formula, is 0.01659816, so that the approximation is correct to 3 decimal places.

NORMAL (GAUSSIAN) APPROXIMATION TO BINOMIAL



Poisson Distribution

Suppose that we would like to calculate probabilities relating to numbers of cancers over a given period of time in a given population. In principle, we can consider using a binomial distribution, since we are talking about numbers of events in a given number of trials. However, the numbers of events may be enormous (number of persons in the population times the number of time periods). Furthermore, we may not even be certain of the denominator, but may have some idea of the rate (per year, say) of cancers in this population from previous data. In such cases, where we have COUNTS of events through time rather than counts of "successes" in a given number of trials, we can consider using the POISSON distribution. More precisely, we make the following assumptions:

1. The probability of an event (say, a cancer) is proportional to the time of observation. We can notate this as $Pr\{$ cancer occurs in time $t\} = \lambda \times t$, where λ is the rate parameter, indicating the event rate in units of events per time.
2. Two events cannot occur simultaneously.
3. The event rate λ is constant through time (homogeneous Poisson process).
4. Events (cancers) occur independently.

If all of these assumptions are true, the we can derive the distribution of the number of counts in any given period of time. Let $\mu = \lambda \times t$ be the rate times time, which is the Poisson mean number of events in time t. Then the Poisson distribution is given by:

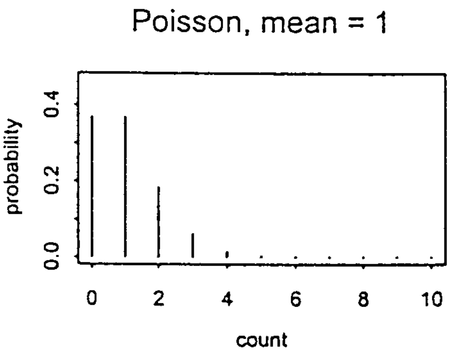
$$Pr\{\ x \text{ events occur in time } t\} = \frac{e^{-\mu}\mu^x}{x!}$$

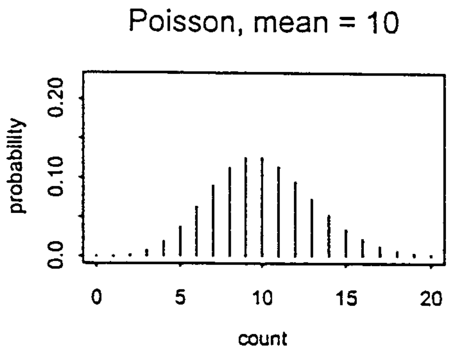
where e = 2.71828..., and $x!$ denotes factorial of x (same as in the binomial distribution).

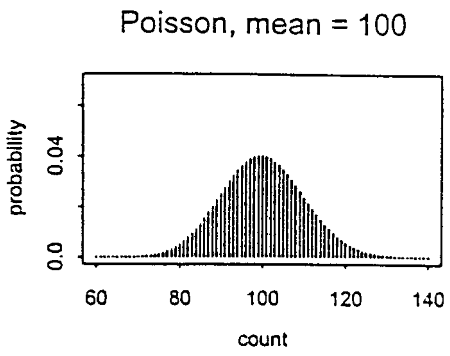
It is quite easy to prove that both the mean and the variance of the Poisson distribution are equal to μ. **Optional exercise: Try to prove this using the formulae of page 47.**

Poisson Examples

1. Let's look at the graphs of Poisson distribution with different values for μ (note that the Poisson is a DISCRETE distribution, which takes on positive integer values only, 0,1,2,3,...):







Note as with many distributions, larger values of μ mean the shape goes towards a normal distribution.

2. How are the values in the above graphs calculated?

$$\begin{aligned}
Pr\{0 \text{ events occur in time t with } \mu = 1\} &= \frac{e^{-1}1^0}{0!} = 1/e = 0.3679 \\
Pr\{1 \text{ event occurs in time t with } \mu = 1\} &= \frac{e^{-1}1^1}{0!} = 1/e = 0.3679 \\
Pr\{2 \text{ events occur in time t with } \mu = 1\} &= \frac{e^{-1}1^2}{2!} = 0.1839 \\
Pr\{99 \text{ events occur in time t with } \mu = 100\} &= \frac{e^{-100}100^{99}}{99!} = 0.0398
\end{aligned}$$

There is a normal approximation to the Poisson, using a continuity correction as in the binomial:

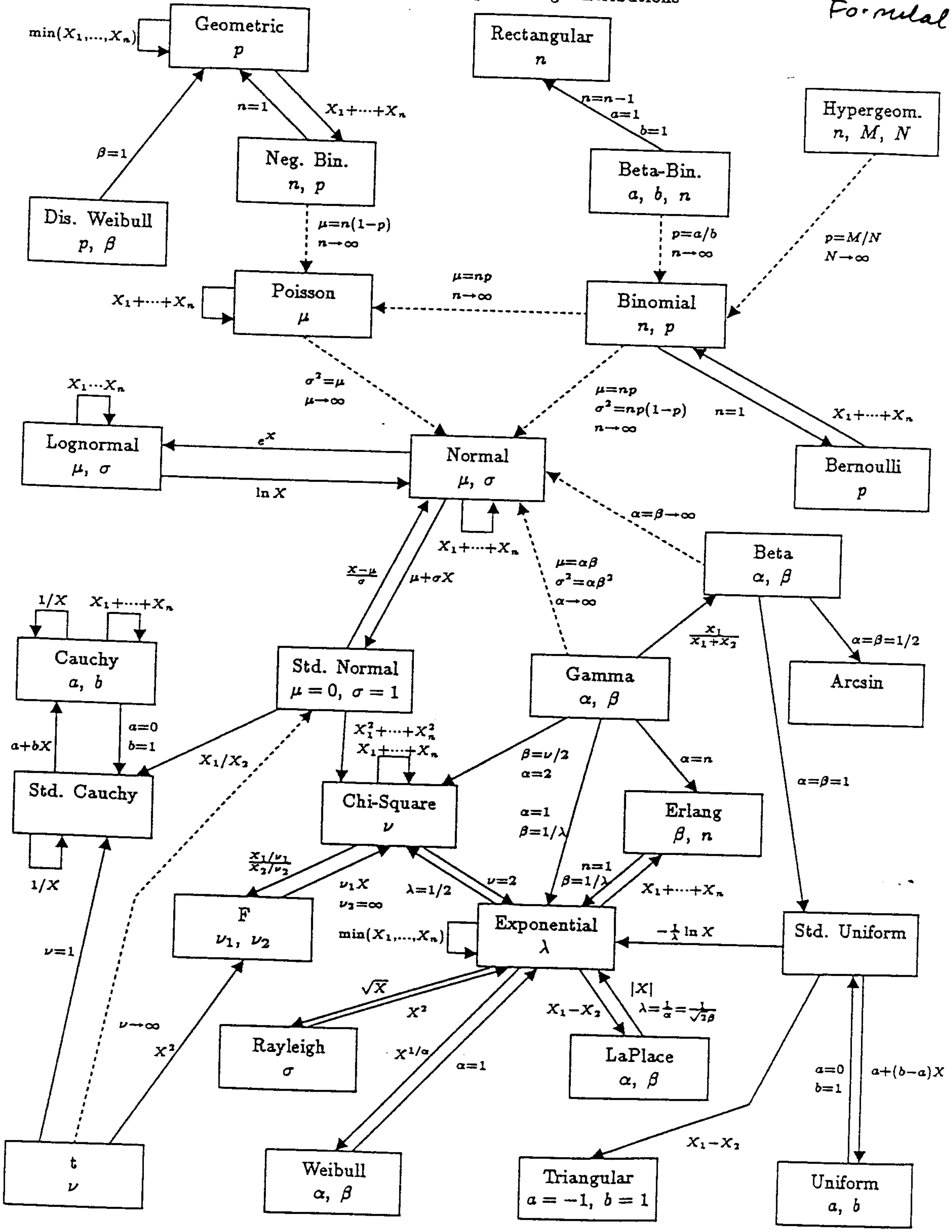
$$\begin{aligned}
& Pr\{99 \text{ events occur in time t with } \mu = 100\} \\
= \; & Pr\{98.5 < x < 99.5 | x \sim N(100, 100)\} \\
= \; & Pr\{\frac{98.5 - 100}{10} < z < \frac{99.5 - 100}{10} \\
= \; & Pr\{-0.15 < z < -0.05 | z \sim N(0, 1)\} = 0.0397 \approx 0.0398
\end{aligned}$$

3. Suppose the number of sudden deaths due to myocardial infarction in Quebec is 250 per year. What is the probability that there will be exactly 135 deaths in the next six months?

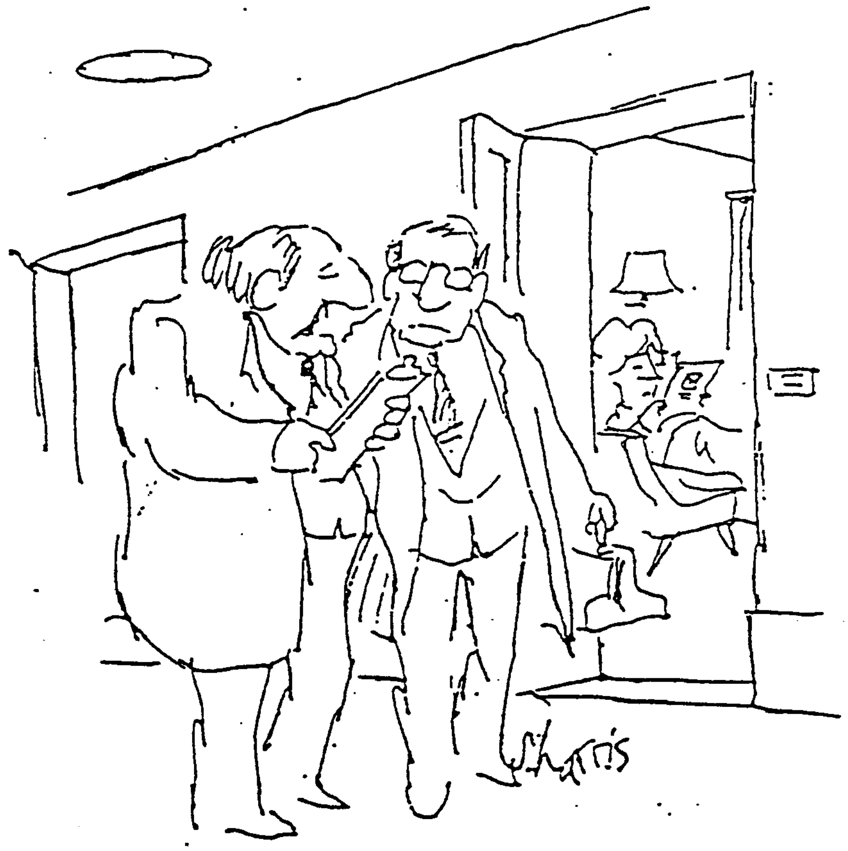
Solution: Let $t = 1$, then $\mu = 250$. If $t = 0.5$, then $\mu = 125$. Using the Poisson distribution, $\frac{e^{-125}125^{135}}{135!} = 0.0232$. **Exercise: check this using the normal approximation (should get 0.0239).**

From: Kokoska and Nevison, Statistical Tables and Formulae.

Table 3. Relationships Among Distributions



THE NULL HYPOTHESIS

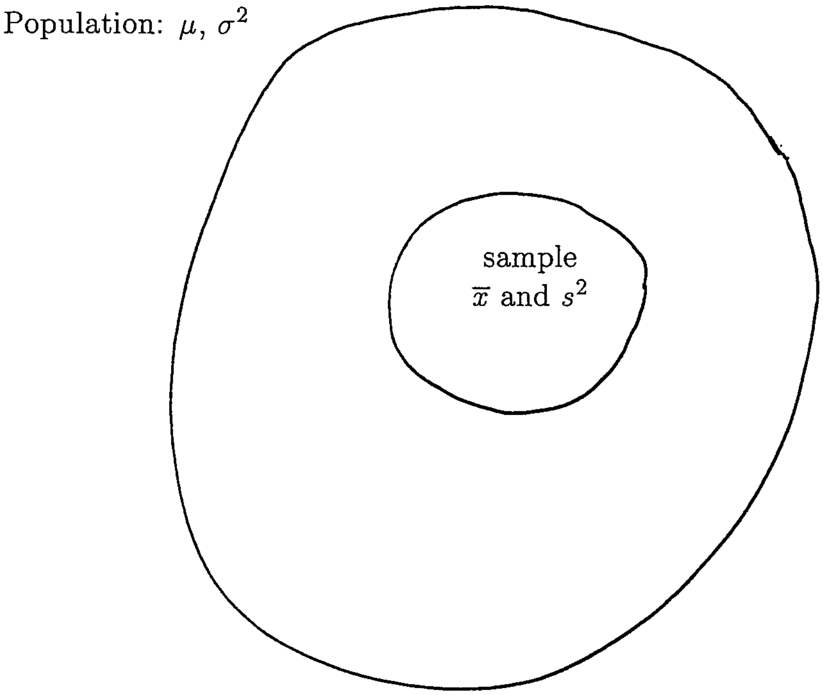


"FIND OUT WHO SET UP THIS EXPERIMENT. IT SEEMS THAT HALF OF THE PATIENTS WERE GIVEN A PLACEBO, AND THE OTHER HALF WERE GIVEN A DIFFERENT PLACEBO."

American Scientist 1982;70:25.

Inference for a single population mean

Suppose that a single population has true (population) parameters μ and σ^2, representing, respectively, the true mean value and variance in that population for a quantity of interest. Given a random sample from that population of values, we do not directly observe μ and σ^2, but rather their sample equivalents, $\overline{x}$ and s^2. Schmatically, we have:

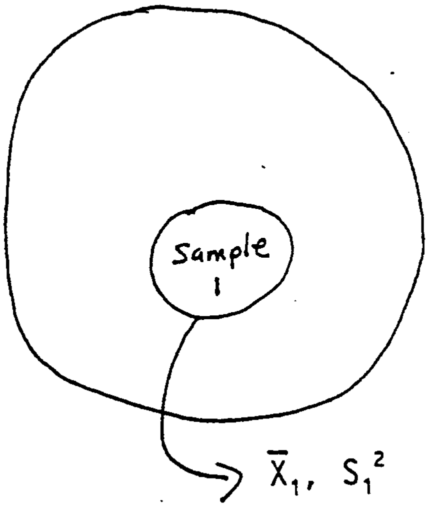


We would like to test whether this value is equal to some prespecified constant. That is, we would like to test:

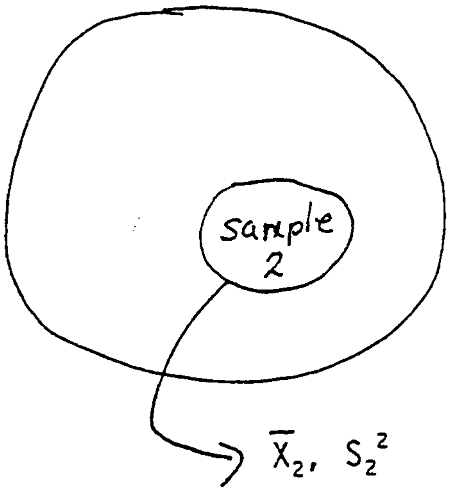
$$\begin{aligned} H_0 : \quad \mu &= \mu_0 \text{ versus} \\ H_A : \quad \mu &\neq \mu_0 \left(\mu > \mu_0 \text{ or } \mu < \mu_0\right) \end{aligned}$$

COMPARING TWO POPULATION MEANS

Population 1: μ_1, σ_1^2



Population 2: μ_2, σ_2^2



H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$
(two-sided)

(or $\mu_1 > \mu_2$, or $\mu_1 < \mu_2$)
(one-sided)

Cases

- Paired or Unpaired?
- Assume $\sigma_1^2 = \sigma_2^2$ or not?

HYPOTHESIS TESTING FOR MEANS

		BEFORE	AFTER	DIFFERENCE
CASE	1	211.000	198.000	-13.000
CASE	2	180.000	173.000	-7.000
CASE	3	171.000	172.000	1.000
CASE	4	214.000	209.000	-5.000
CASE	5	182.000	179.000	-3.000
CASE	6	194.000	192.000	-2.000
CASE	7	160.000	161.000	1.000
CASE	8	182.000	182.000	0.000
CASE	9	172.000	166.000	-6.000
CASE	10	155.000	154.000	-1.000
CASE	11	185.000	181.000	-4.000
CASE	12	167.000	164.000	-3.000
CASE	13	203.000	201.000	-2.000
CASE	14	181.000	175.000	-6.000
CASE	15	245.000	233.000	-12.000
CASE	16	146.000	142.000	-4.000

TOTAL OBSERVATIONS: 16

	BEFORE	AFTER	DIFFERENCE
N OF CASES	16	16	16
MINIMUM	146.000	142.000	-13.000
MAXIMUM	245.000	233.000	1.000
MEAN	184.250	180.125	-4.125
STANDARD DEV	24.909	22.562	4.064

WE WISH TO TEST THE HYPOTHESIS:

DOES THE DIET LEAD TO LOWER WEIGHTS?

DIAGNOSTIC TEST

True State of Disease

		+	-		
Test	+	a ✓	b false pos.	a + b	sens = a / (a+c)
Result	-	c false neg.	d ✓	c + d	spec = d / (b+d)
		a + c	b + d		+ 've = a / (a+b)
					- 've = d / (c+d)

a,b,c,d represent individual persons undergoing diagnostic test

HYPOTHESIS TEST

True State of "Nature"

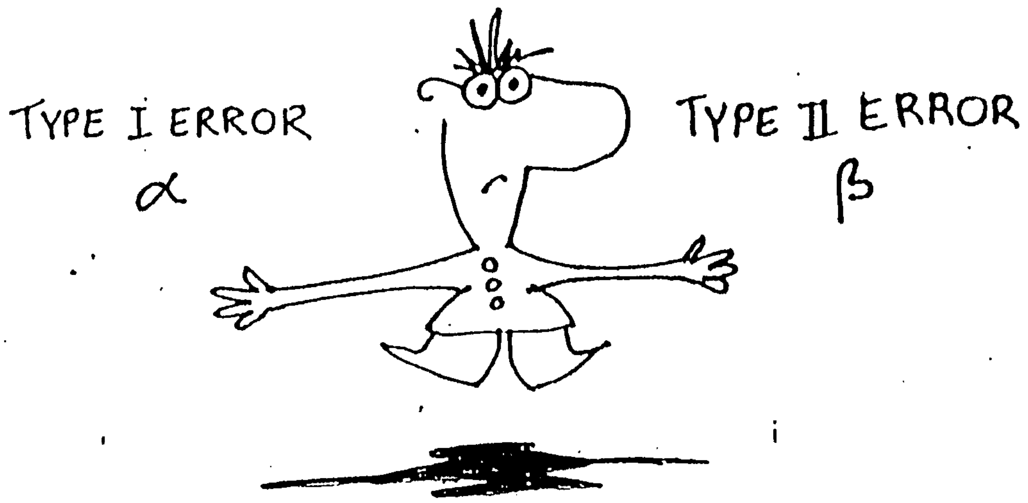
		H_A + Diet effective	H_0 - Diet ineffective	
Statistical Test	+	a ✓	b Type I Error (α)	a + b
Result	-	c Type II Error (β)	d ✓	c + d
		a + c	b + d	

a,b,c,d represent not individuals but counts of tests

P (correct decision | + [nature]) = a / (a+c)
P (correct decision | - [nature]) = d / (b+d)

P (correct decision | + [stat]) = a / (a+b)
P (correct decision | - [stat]) = d / (d+c)

By convention, type I error (α) is simply set to 0.05 (most often), and type II error (ß) is then fixed "automatically" for a given sample size.



The hypothesis on diet may be formally stated as:

H_0 (null hypothesis): The diet does not lead to decreased weight.

H_A (alternative hypothesis): The diet does lead to decreased weight

If we assume that the weight differences (AFTER-BEFORE=DIFF) are normally distributed, then we can say DIFF ~ $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.

Then: H_0: $\mu = 0$
H_A: $\mu < 0$ (one-sided)
[H_A: $\mu \neq 0$ (two-sided)]

Since μ, and σ^2 are unknown, we can estimate them using: (n=16)

$$\text{estimate} \longleftarrow \hat{\mu}_{DIFF} = \bar{X}_{DIFF} = \frac{\sum_{i=1}^{n} X_i}{n} = -4.125$$

$$\hat{\sigma}^2_{DIFF} = S^2 = \frac{\sum (X_i - \overset{\rightarrow 4.125}{\bar{X}})^2}{n-1} = (4.064)^2$$

Now: If DIFF ~ $N(\mu, \sigma^2)$, and If H_0 is true, then DIFF ~ $N(0, \sigma^2)$ and $\bar{X}_{DIFF}$ ~ $N(0, \frac{\sigma^2}{n})$, n=16.

If we wish to be wrong only $\alpha = 5\%$ of the time when H_0 is true, we should reject H_0 as reasonable if $\overline{X}_{DIFF} < -1.67$ (one-sided) or $| \overline{X}_{DIFF} | > 2$ (two-sided). These are our REJECTION REGIONS.

Since we observed $\overline{X}_{DIFF} = -4.125$ which is < -1.67, we fall into our rejection region, and hence the conclusion is to REJECT H_0 at the $\alpha = 5\%$ level.

What is the probability that we are wrong?

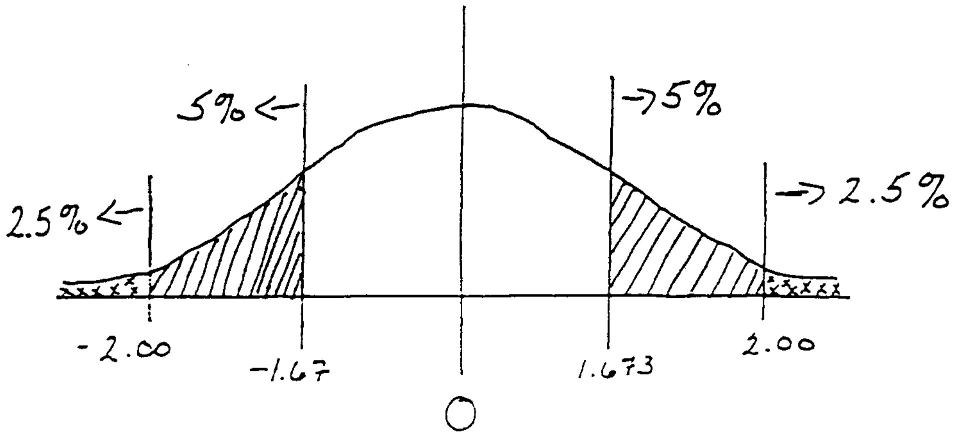
We do not know σ^2, but we have estimated it to be $(4.064)^2$, so that (approximately)

$$\text{DIFF} \sim N(0,\ (4.064)^2)$$

$$\overline{X}_{DIFF} \sim N(0,\ (4.064)^2/16) \quad (\sigma_{\overline{X}_{DIFF}} = 1.02)$$

<u>IF</u> H_0: $\mu = 0$ is true

HENCE the picture is:

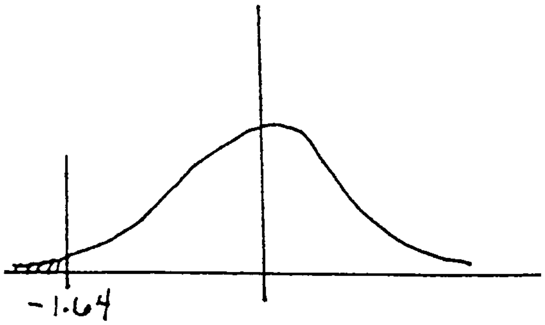


SUMMARY OF EXAMPLE

1. H_0: The diet does not lead to increased weight. ($\mu \geq 0$)

   H_A: The diet does lead to increased weight ($\mu < 0$)

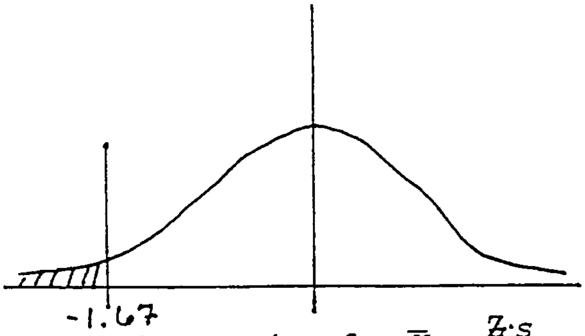
2. Rejection Region: (Find 5% point under H_0)



Rejection region for $Z = \dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}\ (or\ \ s/\sqrt{n})}$

N (0,1)

$\boxed{\text{OR}}$



Rejection region for $\overline{X} = \dfrac{Z\cdot s}{\sqrt{n}}+\mu_0$

N(0,1.02)

3. *Calculate* $\overline{X}$, s^2, s^2/n, $s/\sqrt{n}$ *as required.*

   $\overline{X} = -4.125,\ \dfrac{s^2}{\sqrt{n}} = 1.02$

   Since $\dfrac{\overline{X}-\mu}{s/\sqrt{n}} \leq -1.64$, ($\boxed{or}$ $\overline{X} < -1.67$),

4. We can reject H_0, or find p-value:

$$P = P\left\{\overline{X}<-4.125\right\} = P\left\{\frac{\overline{X}-\mu_0}{s/\sqrt{n}}<\frac{-4.125-0}{1.02}\right\} = P\left\{Z<4.04\right\} = 0.0000267.$$

STATISTICAL SIGNIFICANCE

If the P-value is as small or smaller than α, we say that the data are *statistically significant at level α.*

FIXED SIGNIFICANCE LEVEL z TESTS FOR A POPULATION MEAN

To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the z test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Reject H_0 at significance level α against a one-sided alternative

$$H_a\colon \mu > \mu_0 \quad \text{if} \quad z \geq z^*$$
$$H_a\colon \mu < \mu_0 \quad \text{if} \quad z \leq -z^*$$

where z^* is the upper α critical value from Table D. Reject H_0 at significance level α against a two-sided alternative

$$H_a\colon \mu \neq \mu_0 \quad \text{if} \quad |z| \geq z^*$$

where z^* is the upper $\alpha/2$ critical value from Table D.

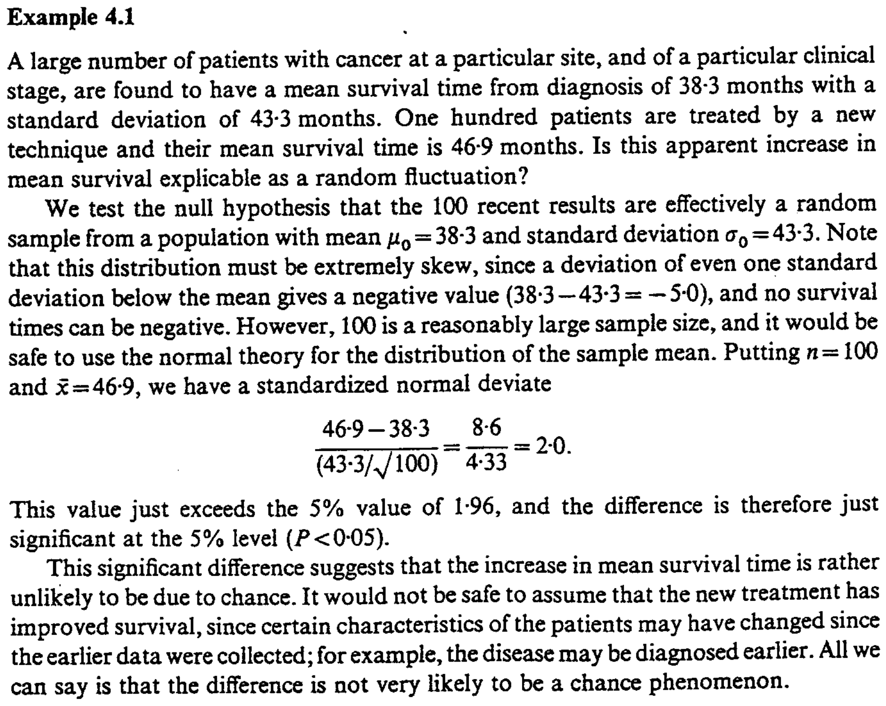
Summary of Testing Procedure

1. State H_0 and H_A. State the α error you think is appropriate for the problem.

2. Find the rejection region.

3. From the data, check whether the observed data fall into the rejection region or not.

4. If the data fall into the rejection region, conclusion is that there is enough evidence to reject the null hypothesis H_0 in favour of the alternative H_A. If the data do not fall into the rejection region, can only say that there is no evidence to reject the null hyupothesis.



Example 4.1

A large number of patients with cancer at a particular site, and of a particular clinical stage, are found to have a mean survival time from diagnosis of 38·3 months with a standard deviation of 43·3 months. One hundred patients are treated by a new technique and their mean survival time is 46·9 months. Is this apparent increase in mean survival explicable as a random fluctuation?

We test the null hypothesis that the 100 recent results are effectively a random sample from a population with mean $\mu_0 = 38{\cdot}3$ and standard deviation $\sigma_0 = 43{\cdot}3$. Note that this distribution must be extremely skew, since a deviation of even one standard deviation below the mean gives a negative value ($38{\cdot}3 - 43{\cdot}3 = -5{\cdot}0$), and no survival times can be negative. However, 100 is a reasonably large sample size, and it would be safe to use the normal theory for the distribution of the sample mean. Putting $n = 100$ and $\bar{x} = 46{\cdot}9$, we have a standardized normal deviate

$$\frac{46{\cdot}9 - 38{\cdot}3}{(43{\cdot}3/\sqrt{100})} = \frac{8{\cdot}6}{4{\cdot}33} = 2{\cdot}0.$$

This value just exceeds the 5% value of 1·96, and the difference is therefore just significant at the 5% level ($P < 0{\cdot}05$).

This significant difference suggests that the increase in mean survival time is rather unlikely to be due to chance. It would not be safe to assume that the new treatment has improved survival, since certain characteristics of the patients may have changed since the earlier data were collected; for example, the disease may be diagnosed earlier. All we can say is that the difference is not very likely to be a chance phenomenon.

6.61 You are designing a computerized medical diagnostic program. The program will scan the results of routine medical tests (pulse rate, blood pressure, urinalysis, etc.) and either clear the patient or refer the case to a doctor. The program will be used as part of a preventive medicine system to screen many thousands of persons who do not have specific medical complaints. The program makes a decision about each patient.

(a) What are the two hypotheses and the two types of error that the program can make? Describe the two types of error in terms of "false positive" and "false negative" test results.

(b) The program can be adjusted to decrease one error probability, at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why? (This is a matter of judgment. There is no single correct answer.)

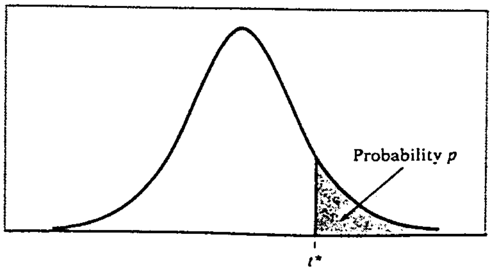


Table entry is the point t^* with given probability p lying above it.

t distribution critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Confidence level C

Example 4.3

In a small clinical trial to assess the value of a new tranquillizer on psychoneurotic patients, each patient was given a week's treatment with the drug and a week's treatment with a placebo, the order in which the two sets of treatments were given being determined at random. At the end of each week the patient had to complete a questionnaire, on the basis of which he was given an 'anxiety score' (with possible values from 0 to 30), high scores corresponding to states of anxiety. The results are shown in Table 4.1.

Table 4.1. Anxiety scores recorded for ten patients receiving a new drug and a placebo in random order

Patient	Anxiety score		Difference d_i
	Drug	Placebo	(drug–placebo)
1	19	22	−3
2	11	18	−7
3	14	17	−3
4	17	19	−2
5	23	22	1
6	11	12	−1
7	15	14	1
8	19	11	8
9	11	19	−8
10	8	7	1
			−13

$H_0: \mu_{\text{Placebo}} = \mu_{\text{DRUG}}$ (DIFF = 0)

$H_A: \mu_{\text{PLACEBO}} \neq \mu_{\text{DRUG}}$ (DIFF ≠ 0)

Use a paired t-test:

$$\bar{X}_{\text{DIFF}} = -1.3$$

$$s^2 = \frac{\sum_{i=1}^{10} (\text{DIFF}_i - \bar{X}_{\text{DIFF}})^2}{n-1} = 20.68$$

$$t_{\text{DF}} = \frac{\bar{x} - 0}{s/\sqrt{n}} \Rightarrow t_{df=9} = \frac{-1.3 - 0}{\sqrt{20.68}/\sqrt{10}} = -0.904$$

From t tables, $p > 0.10$ (Exact value = $\frac{p}{2} = 0.195$)

$p = 0.389$

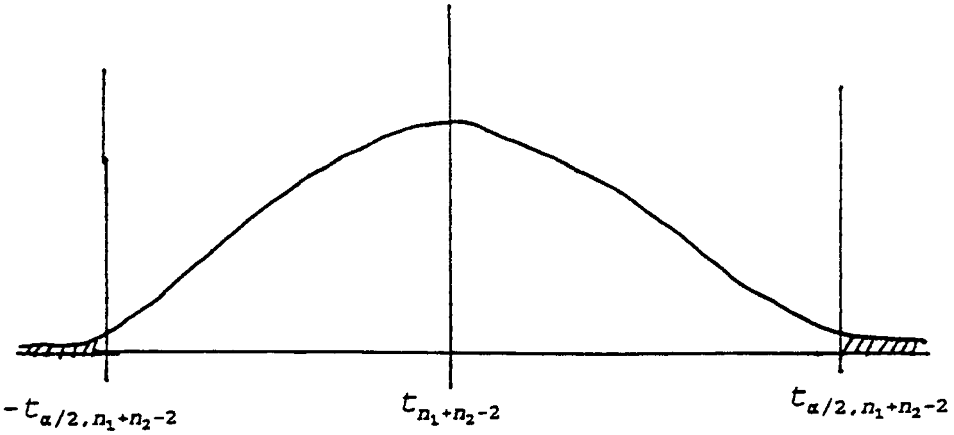
Unpaired, $\sigma_1^2 = \sigma_2^2$ assumption.

1. H_0: $\mu_1 = \mu_2$
   H_A: $\mu_1 \neq \mu_2$

2. $\overline{x}_1 - \overline{x}_2 \sim t_{n_1+n_2-2}$ if $H_o: \ \mu_1 = \mu_2$ is true , $s^2 = s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

where $s_p^2 = \dfrac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}$

s_p^2 = "pooled variance".



Reject if $\left|\dfrac{\overline{X_1}-\overline{X_2}}{s}\right| > t_{\alpha/2,n_1+n_2-2}$

3. Check if value of $\dfrac{\overline{X_1}-\overline{X_2}}{s}$ falls in rejection region.

4. Draw a conclusion or give a p-value,

$$p = P\left\{\frac{\overline{X_1}-\overline{X_2}}{s} > \textit{observed value} \mid H_0 \textit{ is true}\right\}.$$

Summary of Testing for Means

One Population

H_0	Normality?	Variance?	Variance Estimate	Test Statistic	DF
$\mu = \mu_0$	Yes	known	—	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	—
$\mu = \mu_0$	Yes	unknown	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	n-1
$\mu = \mu_0$ large n	No	known	—	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	—
$\mu = \mu_0$ large n	No	unknown	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	n-1

Two Populations

The two population *paired* case can be reduced to the single population case by treating the paired differences as a single population. In this case, it then does not matter if the variances of the original populations are assumed equal or different.

Unpaired Case

H_0	Normality?	Variances?	Variance Estimate	Test Statistic	DF
$\mu_1 - \mu_0 = \Delta$	Yes	known & equal	—	$z = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{\sigma^2/n_1+\sigma^2/n_2}}$	—
$\mu_1 - \mu_0 = \Delta$	Yes	known & different	—	$z = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	—
$\mu_1 - \mu_0 = \Delta$	Yes	unknown & equal	$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$t = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{s_p^2/n_1+s_p^2/n_2}}$	$n_1 + n_2 - 2$
$\mu_1 - \mu_0 = \Delta$	Yes	unknown & different	usual s_1^2 and s_2^2	$t = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$	min(n_1-1,n_2-1) (conservative)
$\mu_1 - \mu_0 = \Delta$ large n	No	known & equal	—	$z = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{\sigma^2/n_1+\sigma^2/n_2}}$	—
$\mu_1 - \mu_0 = \Delta$ large n	No	known & different	—	$z = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	—
$\mu_1 - \mu_0 = \Delta$ large n	No	unknown & equal	$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$t = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{s_p^2/n_1+s_p^2/n_2}}$	$n_1 + n_2 - 2$
$\mu_1 - \mu_0 = \Delta$ large n	No	unknown & different	usual s_1^2 and s_2^2	$t = \frac{\bar{x}_1-\bar{x}_2-\Delta}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$	min(n_1-1,n_2-1) (conservative)

Notes:

1. In most cases, Δ is taken to be 0.
2. For large n, the t distribution is approximated by the Normal.

Example 4.5

A suspension of virus particles is prepared at two dilutions. If the experimental techniques are perfect, Preparation B should have 10 times as high a concentration of virus particles as Preparation A. Equal volumes from each suspension are inoculated onto the chorioallantoic membrane of chick embryos. After an appropriate incubation period the membranes are removed and the number of pocks on each membrane is counted. The numbers are as follows:

Preparation	A	B/10
Counts	x_1	x_2
	0	1·0
	0	1·3
	1	1·3
	1	1·4
	1	1·9
	1	2·0
	2	2·1
	2	2·6
	3	2·9

$n_1=9$ $\qquad$ $n_2=9$

$\bar{x}_1=1{\cdot}2222$ $\qquad$ $\bar{x}_2=1{\cdot}8333$

$s_1^2=0{\cdot}9444$ $\qquad$ $s_2^2=0{\cdot}4100.$

$H_0: \mu_A = \mu_B$

$H_A: \mu_A \neq \mu_B$ (two sided)

Variances seem different, so:

$$t_8 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1.22 - 1.83}{\sqrt{\frac{.944}{9} + \frac{.410}{9}}} = -1.57$$

From t-tables, $0.05 < p < 0.10$ (one sided), so we conclude we cannot reject H_0 $(0.10 < p < .20)$

(Exact $p = 0.155$).

P-VALUE

The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the *P*-value of the test. The smaller the *P*-value, the stronger the evidence against H_0 provided by the data.

SIGNIFICANCE AND TYPE I ERROR

The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true.

CONFIDENCE INTERVALS AND TWO-SIDED TESTS

A level α two-sided significance test rejects a hypothesis H_0: $\mu = \mu_0$ exactly when the value μ_0 falls outside a level $1-\alpha$ confidence interval for μ.

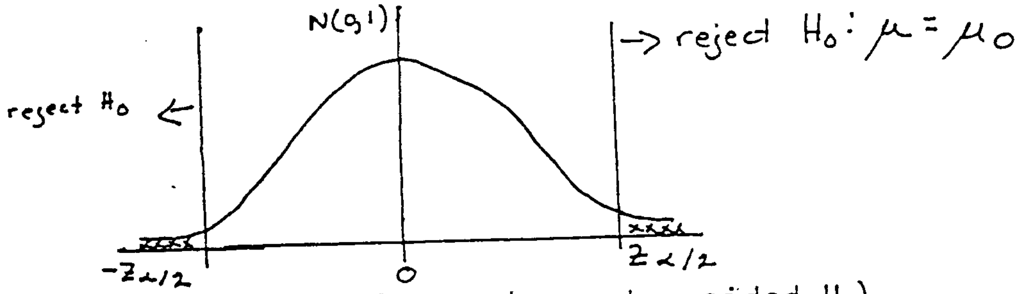
POWER AND TYPE II ERROR

The power of a fixed level test against a particular alternative is 1 minus the probability of a Type II error for that alternative.

NOTES ON SIGNIFICANT TESTS

1. We have seen two different "philosophies" of testing:

a) Decide α in advance, check to see if calculations from data fall into critical region:


N(0,1)
$\rightarrow$ reject $H_0: \mu = \mu_0$
reject H_0 $\leftarrow$
$-z_{\alpha/2}$
0
$z_{\alpha/2}$

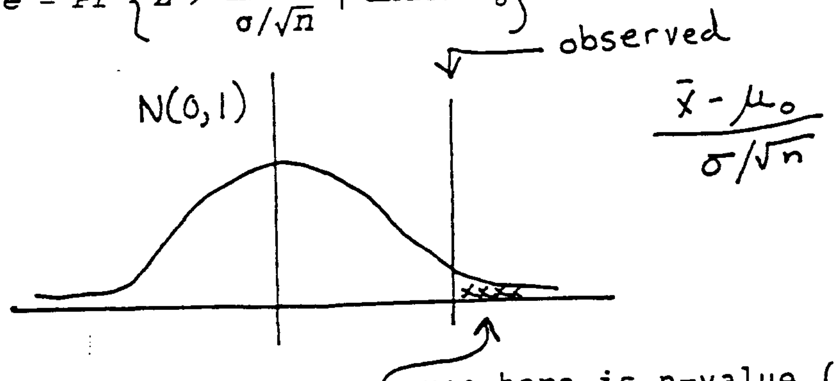
(Illustration is for σ known two-sided H_A).

Reject if $\left| \frac{\bar{X}-\mu_o}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$

- α is often 0.05, but should be chosen according to the problem.

b) Calculate $\frac{\bar{X}-\mu_o}{\sigma/\sqrt{n}}$, and report the

$$p\text{-value} = Pr\left\{ z > \frac{\bar{X}-\mu_o}{\sigma/\sqrt{n}} \mid under\ H_o \right\}$$



area here is p-value (multiply by 2 if H_A is two-sided).

a) implies a decision must be made, from this data alone. If not, b) is more informative.

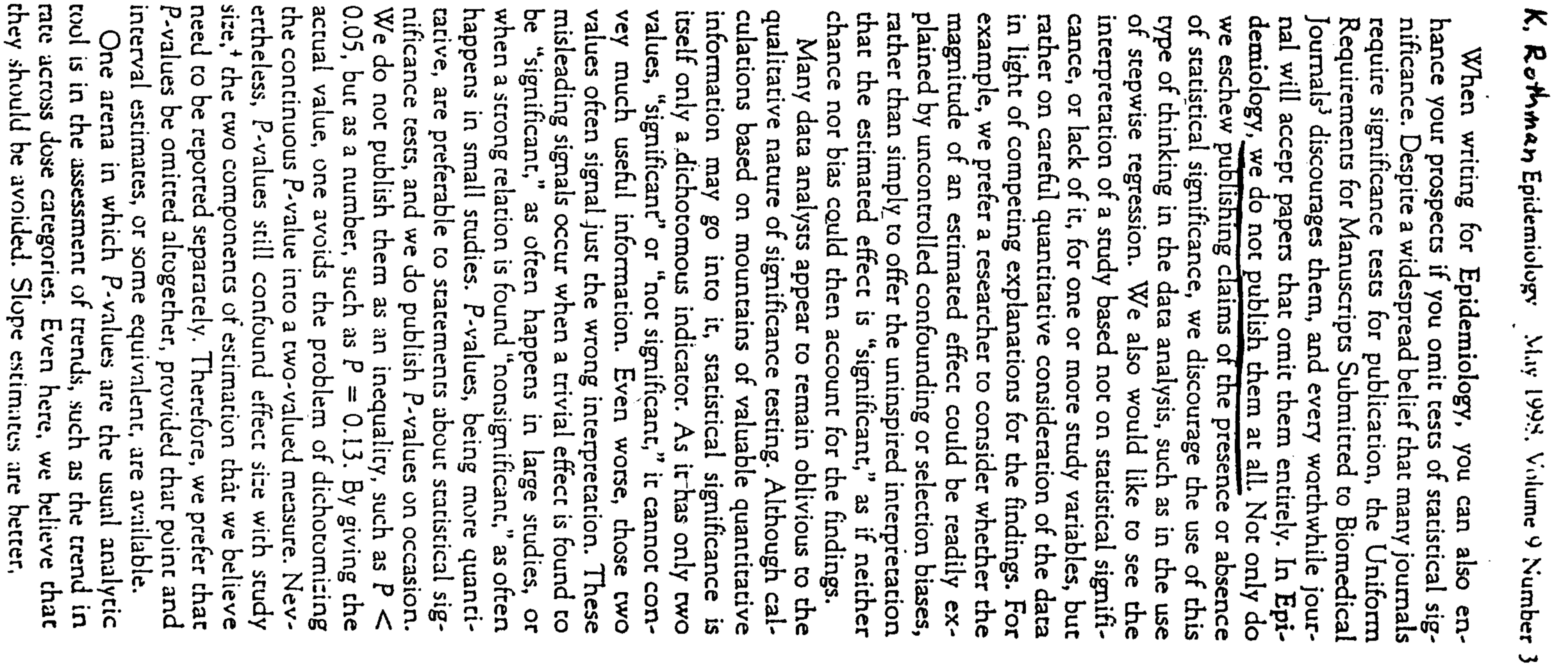
2. There is nothing magic about $\alpha = 0.05$. There is no practical difference if $p = 0.049$ or $p = 0.051$.

3. Even a very small p-value does not guarantee H_0 is false. Repeating the study is usually necessary for further proof, or to vary the conditions or population.

4. Statistical significance (small p-value) is not the same as practical significance.

5. The p-value is not everything. Must also examine your data carefully, data cleaning for outliers, etc. Remember - all tests carry assumptions (Normal distribution, simple random sample from a population, equal variances) which can be thrown off by outliers.

6. Reporting a confidence interval for an effect is more informative than reporting a p-value.

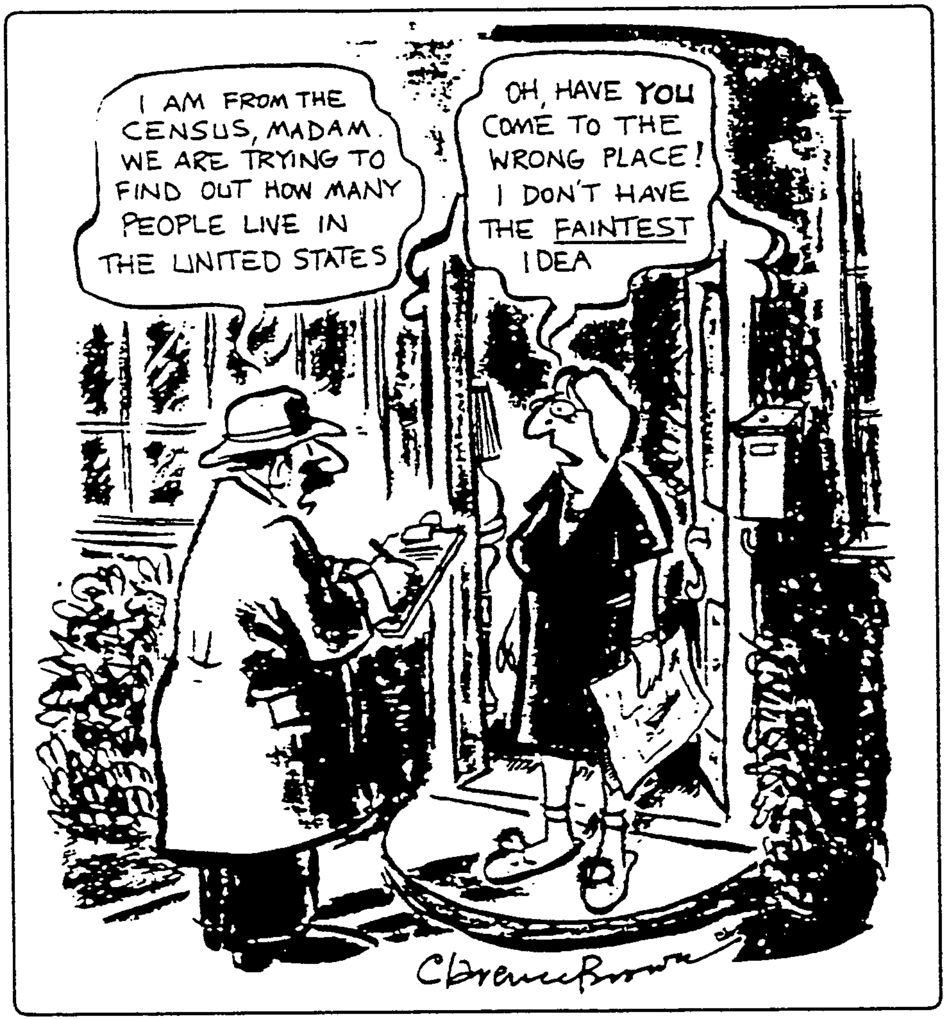


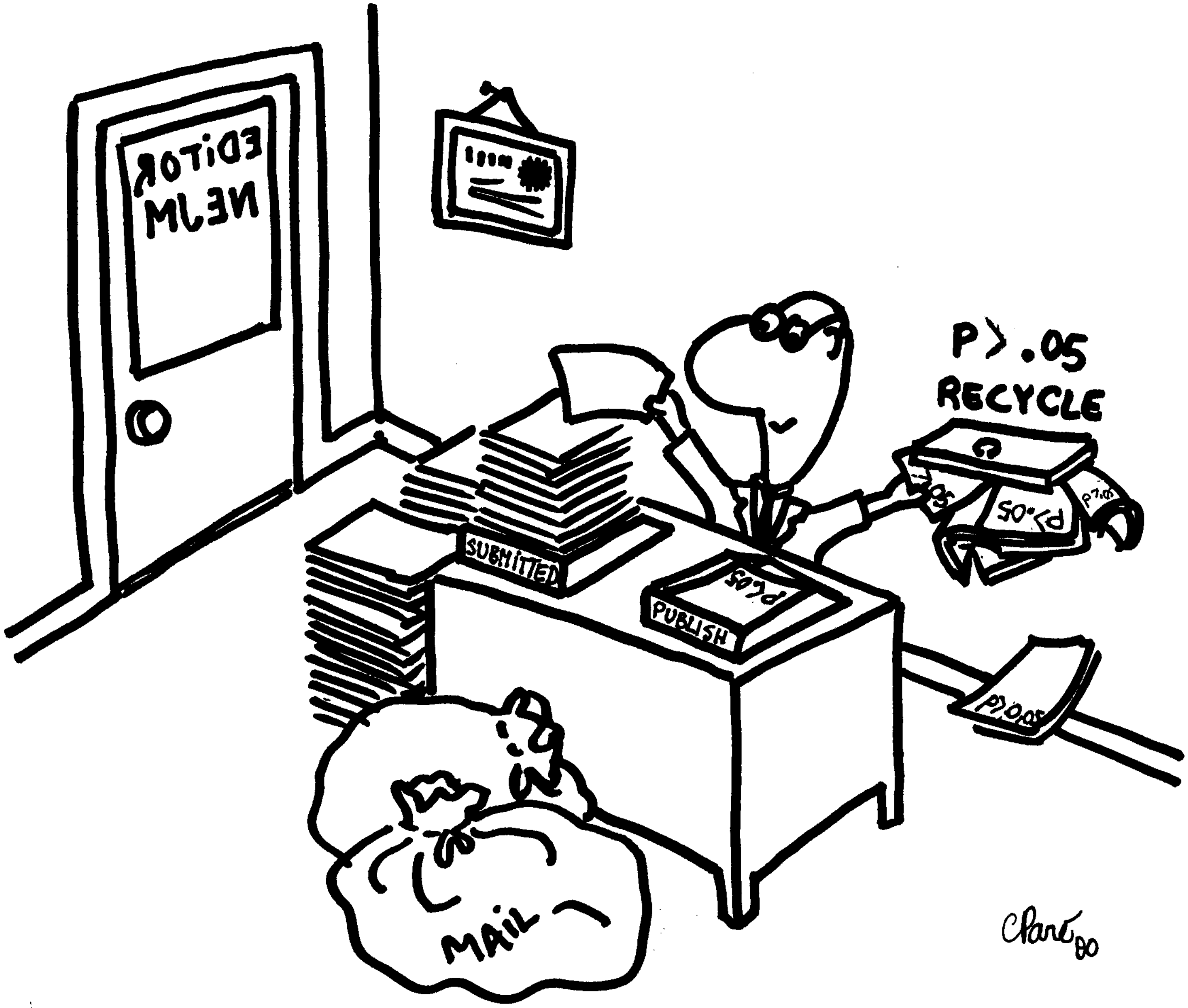
K. Rothman Epidemiology May 1998, Volume 9 Number 3

When writing for **Epidemiology**, you can also enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals[3] discourages them, and every worthwhile journal will accept papers that omit them entirely. In **Epidemiology**, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression. We also would like to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, we prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is "significant," as if neither chance nor bias could then account for the findings.

Many data analysts appear to remain oblivious to the qualitative nature of significance testing. Although calculations based on mountains of valuable quantitative information may go into it, statistical significance is itself only a dichotomous indicator. As it has only two values, "significant" or "not significant," it cannot convey much useful information. Even worse, those two values often signal just the wrong interpretation. These misleading signals occur when a trivial effect is found to be "significant," as often happens in large studies, or when a strong relation is found "nonsignificant," as often happens in small studies. *P*-values, being more quantitative, are preferable to statements about statistical significance tests, and we do publish *P*-values on occasion. We do not publish them as an inequality, such as $P < 0.05$, but as a number, such as $P = 0.13$. By giving the actual value, one avoids the problem of dichotomizing the continuous *P*-value into a two-valued measure. Nevertheless, *P*-values still confound effect size with study size,[4] the two components of estimation that we believe need to be reported separately. Therefore, we prefer that *P*-values be omitted altogether, provided that point and interval estimates, or some equivalent, are available.

One arena in which *P*-values are the usual analytic tool is in the assessment of trends, such as the trend in rate across dose categories. Even here, we believe that they should be avoided. Slope estimates are better,


I AM FROM THE CENSUS, MADAM. WE ARE TRYING TO FIND OUT HOW MANY PEOPLE LIVE IN THE UNITED STATES
OH, HAVE **YOU** COME TO THE WRONG PLACE! I DON'T HAVE THE FAINTEST IDEA
ClarenceBrown


EDITOR
NEW
P>.05
RECYCLE
SUBMITTED
PUBLISH
P<.05
P>.05
P>.05
P>0.05
MAIL

PUBLICATION BIAS

...a survey of four journals of the American Psychological Association showed that of 294 articles using statistical tests, only eight (8) did not attain the 5% significance level.

THIS IS DANGEROUS!

- "Non-significant" results may be "highly significant" if a common belief is not verified - this <u>should</u> be published.

- If 20 studies are done to test similar hypotheses, by chance alone, expect one to be significant. This should (for obvious reasons) not be the only study of the 20 to be published.

- Interesting and potentially fruitful ideas may be lost if the study is small and had low power.

- Publishing all studies helps to carry out meta-analyses which attempts to summarize data from many experiments.

SEARCHING FOR SIGNIFICANCE/MULTIPLE TESTING

Twenty variables are collected as possible risk factors for heart disease. Suppose that none are true explanatory variables, but that each is tested separately by and $\alpha = 0.05$ level test.

Then:

P {none are significant} = $(0.95)^{20}$ = 36%

Pr {exactly one is significant} = $20!/(1!19!)\ (.95)^{19}\ (.05)^{1}$ = 38%

Pr (exactly 2 are significant} = $20!/(2!18!)\ (.95)^{18}\ (.05)^{2}$ = 19%

Pr (exactly 3 are significant} = $20!/(3!17!)\ (.95)^{17}\ (.05)^{3}$ = 6%

Pr {at least one is significant} = 1-36% = 64%

Number of Tests (n)	Probability of at least one type 1 (α) error $1 - (0.95)^n$
1	5%
2	10%
5	23%
10	40%
15	54%
20	64%
50	92%
100	99%

Bonferoni: adjustment: Let $\alpha = \alpha/n$, n = # tests.

Hypothesis Generating
vs.
Hypothesis Verifying

Verifying: Know hypothesis ahead of time, look at data for verification.

Generating: Let data suggest hypothesis. Cannot test a hypothesis on data that first suggest hypothesis.

Power

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the power of the test against that alternative.

7.51 Example 7.11 gives a test of a hypothesis about the SAT scores of California high school students based on an SRS of 500 students. The hypotheses are:

$$H_0\colon \mu = 450$$
$$H_{\alpha}\colon \mu > 450$$

Assume that the population standard deviation is σ = 100. The test rejects H_0 at the 1% level of significance when z , where

$$z = \frac{\bar{x} - 450}{100/\sqrt{500}}$$

Is this test sufficiently sensitive to usually detect an increase of 10 points in the population mean SAT score? Answer this question by calculating the power of the test against the alternative μ = 460.

$$\text{POWER} = Pr\,\{\textit{rejecting } H_0 \,|\, \mu = 460\}$$

$$= Pr\left\{\frac{\bar{X} - 450}{100/\sqrt{500}} \geq 2.326 \;|\; \mu = 460\right\}$$

$$= Pr\left\{\frac{\bar{X} - 450}{100/\sqrt{500}} - \frac{10}{100/\sqrt{500}} \geq 2.326 - \frac{10}{100/\sqrt{500}} \;|\; \mu = 460\right\}$$

$$= Pr\left\{Z = \frac{\bar{X} - 460}{100/\sqrt{500}} \geq 0.090 \;|\; \mu = 460\right\}$$

$$= Pr\,\{\, Z \geq 0.090 \,\} = 0.464 = 46.4\%$$

Power and Sample Size

True state of Nature

		H_A, +	H_0, −
Test	+ (Reject H_0)	$1-\beta$	α
	– (Do not Reject)	β	$1-\alpha$

$$\begin{aligned}
\alpha &= Pr\{\text{ rejecting } H_0 | H_0 \text{ is true}\} = \text{type I error} \\
1-\alpha &= Pr\{\text{not rejecting } H_0 | H_0 \text{ is true}\} \\
\beta &= Pr\{\text{not rejecting } H_0 | H_A \text{ is true}\} = \text{type II error} \\
1-\beta &= Pr\{\text{ rejecting } H_0 | H_A \text{ is true}\} = \text{Power}
\end{aligned}$$

While one can always work out the power by following the definition (as on previous page), "plug-in" formulae have been worked out for various special cases. Let N be the sample size of the experiment (N = total sample size if there are two groups), let the power be given by β, and let the type I error be given by α. Then we can derive the power for the following situations:

1. To test $H_0 : \mu = \mu_0$ versus a two-sided alternative,

$$N = \frac{\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_0 - \mu_A)^2} \; ,$$

   where μ_A is the particular alternative under discussion, σ^2 is the assumed known variance in the population, and the $z_{1-\alpha/2}$ and $z_{1-\beta}$ are normal quantiles corresponding to the type I error rate and power, respectively. From this, one can solve for $z_{1-\beta}$, which gives the power. In particular,

$$z_{1-\beta} = \frac{\sqrt{N}|\mu_0 - \mu_A| - \sigma \times z_{1-\alpha/2}}{\sigma}$$

For example, plugging in $N = 500$, $\mu_0 = 450$, $\mu_A = 460$, $\sigma = 100$, and $z_{1-\alpha} = 2.326$ (one sided test, in this case), gives

$$z_{1-\beta} = \frac{\sqrt{500}|10| - 100 \times 2.326}{100} = -0.090,$$

exactly as on the previous page (except for the negative sign, since now have < rather than >).

2. For a two sample test, $H_0 : \mu_1 - \mu_2 = 0$ versus a two-sided alternative, $H_0 : \mu_1 - \mu_2 \neq 0$, where $\sigma_1^2 = \sigma_2^2$,

$$N = \frac{4 \times \sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_1 - \mu_2)^2} \ ,$$

where $\mu_1 - \mu_2$ is the particular alternative under discussion, and σ^2 is the assumed equal variance in the populations.

From this, one can solve for $z_{1-\beta}$, which gives the power. In particular,

$$z_{1-\beta} = \frac{\sqrt{N}|\mu_1 - \mu_2| - 2 \times \sigma \times z_{1-\alpha/2}}{2 \times \sigma}$$

Example

What size is needed for 80% and 90% power, in the case of equal variances ($\sigma_0^2 = \sigma_1^2 = 1$, say) for selected values of $\Delta = \mu_1 - \mu_2$? Assume a two-sided test at the $\alpha = 0.05$ level.

Δ	N for 80% power	N for 90% power
$\mu_1 - \mu_2$	$Z_\beta = 0.84$	$Z_\beta = 1.28$
.1	3136	4199
.2	784	1050
.3	348	467
.4	196	262
.5	125	167
.6	87	116
.7	64	86
.8	49	66
.9	39	52
1.0	31	42

Reference: JM Lachin. Introduction to sample size determination and power analysis for clinical trials. (1981) Controlled Clinical Trials, 2, 93–113.

106

Standard Deviation, Standard Error

Which 'Standard' Should We Use?

George W. Brown, MD

• **Standard deviation (SD) and standard error (SE) are quietly but extensively used in biomedical publications. These terms and notations are used as *descriptive statistics* (summarizing numerical data), and they are used as *inferential statistics* (estimating population parameters from samples). I review the use and misuse of SD and SE in several authoritative medical journals and make suggestions to help clarify the usage and meaning of SD and SE in biomedical reports.**

(*Am J Dis Child* 1982;136:937-941)

Standard deviation (SD) and standard error (SE) have surface similarities; yet, they are conceptually so different that we must wonder why they are used almost interchangeably in the medical literature. Both are usually preceded by a plus-minus symbol (±), suggesting that they define a symmetric interval or range of some sort. They both appear almost always with a mean (average) of a set of measurements or counts of something. The medical literature is replete with statements like, "The serum cholesterol measurements were distributed with a mean of 180±30 mg/dL (SD)."

In the same journal, perhaps in the same article, a different statement may appear: "The weight gains of the subjects averaged 720 (mean)±32 g/mo (SE)." Sometimes, as discussed further, the summary data are presented as the "mean of 120 mg/dL±12" without the "12" being defined as SD or SE, or as some other index of dispersion. Eisenhart[1] warned against this "peril of shorthand expression" in 1968; Feinstein[2] later again warned about the fatuity and confusion contained in any $a \pm b$ statements where b is not defined. Warnings notwithstanding, a glance through almost any medical journal will show examples of this usage.

Medical journals seldom state why SD or SE is selected to summarize data in a given report. A search of the three major pediatric journals for 1981 (*American Journal of Diseases of Children*, *Journal of Pediatrics*, and *Pediatrics*) failed to turn up a single article in which the selection of SD or SE was explained. There seems to be no uniformity in the use of SD or SE in these journals or in *The Journal of the American Medical Association* (*JAMA*), the *New England Journal of Medicine*, or *Science*. The use of SD and SE in the journals will be discussed further.

If these respected, well-edited journals do not demand consistent use of either SD or SE, are there really any important differences between them? Yes, they are remarkably different, despite their superficial similarities. They are so different in fact that some authorities have recommended that SE should rarely or never be used to summarize medical research data. Feinstein[2] noted the following:

> A standard error has nothing to do with standards, with errors, or with the communication of scientific data. The concept is an abstract idea, spawned by the imaginary world of statistical inference and pertinent only when certain operations of that imaginary world are met in scientific reality.[2(p336)]

Glantz[3] also has made the following recommendation:

> Most medical investigators summarize their data with the standard error because it is always smaller than the standard deviation. It makes their data look better . . . data should never be summarized with the standard error of the mean.[3(pp25-26)]

A closer look at the source and meaning of SD and SE may clarify why medical investigators, journal reviewers, and editors should scrutinize their usage with considerable care.

DISPERSION

An essential function of "descriptive statistics" is the presentation of condensed, shorthand symbols that epitomize the important features of a collection of data. The idea of a *central value* is intuitively satisfactory to anyone who needs to summarize a group of measurements or counts. The traditional indicators of a central tendency are the *mode* (the most frequent value), the *median* (the value midway between the lowest and the highest value), and the *mean* (the average). Each has its special uses, but the mean has great convenience and flexibility for many purposes.

The dispersion of a collection of values can be shown in several ways; some are simple and concise, and others are complex and esoteric. The *range* is a simple, direct way to indicate the spread of a collection of values, but it does not tell how the values are distributed. Knowledge of the mean adds considerably to the information carried by the range.

Another index of dispersion is provided by the differences (deviations) of each value from the mean of the values. The trouble with this approach is that some deviations will be positive, and some will be negative, and their sum will be zero. We could ignore the sign of each deviation, ie, use the "absolute mean deviation," but mathematicians tell us that working with absolute numbers is extremely difficult and fraught with technical disadvantages.

A neglected method for summarizing the dispersion of data is the calculation of percentiles (or deciles, or quartiles). Percentiles are used more frequently in pediatrics than in other branches of medicine, usually in growth charts or in other data arrays that are clearly not symmetric or bell shaped. In the general medical literature, percentiles are sparsely used, apparently because of a common, but erroneous, assumption that the mean±SD or SE is satisfactory for summarizing central tendency and dispersion of all sorts of data.

From the Los Lunas Hospital and Training School, New Mexico, and the Department of Pediatrics, University of New Mexico School of Medicine, Albuquerque.

Reprint requests to Los Lunas Hospital and Training School, Box 1269, Los Lunas, NM 87031 (Dr Brown).

STANDARD DEVIATION

The generally accepted answer to the need for a concise expression for the dispersion of data is to square the difference of each value from the group mean, giving all positive values. When these squared deviations are added up and then divided by the number of values in the group, the result is the *variance*.

The variance is always a positive number, but it is in different units than the mean. The way around this inconvenience is to use the square root of the variance, which is the population standard deviation (σ), which for convenience will be called SD. Thus, the SD is the square root of the averaged squared deviations from the mean. The SD is sometimes called by the shorthand term, "root-mean-square."

The SD, calculated in this way, is in the same units as the original values and the mean. The SD has additional properties that make it attractive for summarizing dispersion, especially if the data are distributed symmetrically in the revered bell-shaped, gaussian curve. Although there are an infinite number of gaussian curves, the one for the data at hand is described completely by the mean and SD. For example, the mean+1.96 SD will enclose 95% of the values; the mean±2.58 SD will enclose 99% of the values. It is this symmetry and elegance that contribute to our admiration of the gaussian curve.

The bad news, especially for biologic data, is that many collections of measurements or counts are not symmetric or bell shaped. Biologic data tend to be skewed or double humped, J shaped, U shaped, or flat on top. Regardless of the shape of the distribution, it is still possible by rote arithmetic to calculate an SD although it may be inappropriate and misleading.

For example, one can imagine throwing a six-sided die several hundred times and recording the score at each throw. This would generate a flattopped, ie, rectangular, distribution, with about the same number of counts for each score, 1 through 6. The mean of the scores would be 3.5 and the SD would be about 1.7. The trouble is that the collection of scores is not bell shaped, so the SD is not a good summary statement of the true form of the data. (It is mildly upsetting to some that no matter how many times the die is thrown, it will never show its average score of 3.5.)

The SD wears two hats. So far, we have looked at its role as a *descriptive statistic* for measurements or counts that are representative only of themselves, ie, the data being summarized are not a sample representing a larger (and itself unmeasurable) universe or population.

The second hat involves the use of SD from a random sample as an *estimate* of the population standard deviation (σ). The formal statistical language says that the sample *statistic*, SD, is an unbiased estimate of a population *parameter*, the population standard deviation, σ.

This "estimator SD" is calculated differently than the SD used to describe data that represent only themselves. When a sample is used to make estimates about the population standard deviation, the calculations require two changes, one in concept and the other in arithmetic. First, the mean used to determine the deviations is conceptualized as an estimate of the mean, x̄, rather than as a true and exact population mean (μ). Both means are calculated in the same way, but a population mean, μ, stands for itself and is a parameter; a sample mean, x̄, is an estimate of the mean of a larger population and is a statistic.

The second change in calculation is in the arithmetic: the sum of the squared deviations from the (estimated) mean is divided by n−1, rather than by N. (This makes sense intuitively when we recall that a sample would not show as great a spread of values as the source population. Reducing the denominator [by one] produces an estimate slightly larger than the sample SD. This "correction" has more impact when the sample is small than when n is large.)

Formulas for the two versions of SD are shown in Fig 1. The formulas follow the customary use of Greek letters for population parameters and English letters for sample statistics. The number in a sample is indicated by the lowercase

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

SD of Population

μ = Mean of Population

N = Number in Population

$$SD = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

Estimate of Population SD From Sample

X̄ = Mean of Sample

n = Number in Sample

Fig 1.—Standard deviation (SD) of population is shown at left. Estimate of population SD derived from sample is shown at right.

$$SD_{X} = \frac{SD}{\sqrt{n}} = SEM \qquad SE_{p} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$$

SEM

SD = Estimate of Population SD

n = Sample Size

SE of Proportion

p = Proportion Estimated From Sample

q = (1 − p)

n = Sample Size

Fig 2.—Standard error of mean (SEM) is shown at left. Note that SD is estimate of population SD (not σ, actual SD of population). Sample size used to calculate SEM is n. Standard error of proportion is shown at right.

"n," and the number in a population is indicated by the capital "N."

The two-faced nature of the SD has caused tension between medical investigators on the one hand and statisticians on the other. The investigator may believe that the subjects or measurements he is summarizing are self-contained and unique and cannot be thought of as a random sample. Therefore, he may decide to use the SD as a descriptive statement about dispersion of his data. On the other hand, the biostatistician has a tendency, because of his training and widespread statistical practice, to conceive of the SD as an estimator of a parameter of a population. The statistician may hold the view that any small collection of data is a stepping-stone to higher things.

The pervasive influence of statisticians is demonstrated in the program for calculating the SD that is put into many handheld calculators; they usually calculate the estimator SD rather than the "descriptor SD."

In essence, the investigator and his statistical advisor, the journal reviewers, and the editors all confront a critical decision whenever they face the term "standard deviation." Is it a descriptive statistic about a collection of (preferably gaussian) data that stand free and independent of sampling constraints, ie, is it a straightforward indication of dispersion? Or, is the SD being used as an estimate of a population parameter? Although the SD is commonly used to summarize medical information, it is rare that the reports indicate which version of the SD is being used.

STANDARD ERROR

In some ways, standard error is simpler than the SD, but in other ways, it is much more complex. First, the simplicities will be discussed. The SE is always smaller than the SD. This may account for its frequent use in medical publications; it makes the data look "tighter" than does the SD. In the previously cited quotation by Glantz,[3] the implication is that the SE might be used in a conscious attempt at distortion or indirection. A more charitable view is that many researchers and clinicians simply are not aware of the important differences between SD and SE. At first glance, the SE looks like a measure of dispersion, just as the SD does. The trouble is that the dispersion implied by the SE is different in nature than that implied by the SD.

The SE is always an estimator of a population characteristic; it is not a descriptive statistic—it is an inferential statistic. The SE is an estimate of the interval into which a population parameter will probably fall. The SE also enables the investigator to choose the probability that the parameter will fall within the estimated interval, usually called the "confidence interval."

Here is a statement containing the SE: The mean of the sample was 73 mg/dL, with an SE of the mean of 3 mg/dL. This implies that the mean of the population from which the sample was randomly taken will fall, with 95% probability, in the interval of $73 \pm (1.96 \times 3)$, which is from 67.12 to 78.88. Technically the statement should be: 95 out of 100 confidence intervals calculated in this manner will include the population parameter. If 99% probability is desired, the confidence interval is $73 \pm (2.58 \times 3)$, which is from 65.26 to 80.74.

As Feinstein[2] notes, the SE has nothing to do with standards or with errors; it has to do with predicting confidence intervals from samples. Up to this point, I have used SE as though it meant only the SE of the mean (SEM). The SE should not be used without indicating what parameter interval is being estimated. (I broke that rule for the sake of clarity in the introduction of the contrast between SD and SE.)

Every sample statistic can be used to estimate an SE; there is an SE for the mean, for the difference between the means of two samples, for the slope of a regression line, and for a correlation coefficient. Whenever the SE is used, it should be accompanied by a symbol that indicates which of the several SEs it represents, eg, SEM for SE of the mean.

Figure 2 shows the formula for calculating the SEM from the sample; the formula requires the estimator SD, ie, the SD calculated using $n-1$, not N. It is apparent from the formula for the SEM that the larger the sample size, the smaller the SEM and, therefore, the narrower the confidence interval. Stated differently, if the estimate of a population mean is from a large sample, the interval that probably brackets the population mean is narrower for the same level of confidence (probability). To reduce the confidence interval by half, it is necessary to increase the sample size by a multiple of four. For readers who know that the SD is preferred over the SEM as an index for describing dispersion of gaussian data, the formula for the SEM can be used (in reverse, so to speak) to calculate the SD, if sample size is known.

The theoretical meaning of the SEM is quite engaging, as an example will show. One can imagine a population that is too large for every element to be measured. A sample is selected randomly, and its mean is calculated, then the elements are replaced. The selection and measuring are repeated several times, each time with replacement. The collection of means of the samples will have a distribution, with a mean and an SD. The mean of the sample means will be a good estimate of the population mean, and the SD of the means will be the SEM. Figure 2 uses the symbol $SD_{\bar{x}}$ to show that a collection of sample means ($\bar{x}$) has a SD, and it is the SEM. The interpretation is that the true population mean (μ) will fall, with 95% probability, within ±1.96 SEM of the mean of the means.

Here, we see the charm and attractiveness of the SEM. It enables the investigator to estimate from a sample, at whatever level of confidence (probability) desired, the interval within which the population mean will fall. If the user wishes to be very confident in his interval, he can set the brackets at ±3.5 SEM, which would "capture" the mean with 99.96% probability.

Standard errors in general have other seductive properties. Even when the sample comes from a population that is skewed, U shaped, or flat on top, most SEs are estimators of nearly gaussian distributions for the statistic of interest. For example, for samples of size 30 or larger, the SEM and the sample mean, $\bar{x}$, define a nearly gaussian distribution (of sample means), regardless of the shape of the population distribution.

These elegant features of the SEM are embodied in a statistical principle called the Central Limit Theorem, which says, among other things:

> The mean of the collection of many sample means is a good estimate of the mean of the population, and the distribution of the sample means (if n = 30 or larger) will be nearly gaussian regardless of the distribution of the population from which the samples are taken.

The theorem also says that the collection of sample means from large samples will be better in estimating the population mean than means from small samples.

Given the symmetry and usefulness of SEs in inferential statistics, it is no wonder that some form of the SE, especially the SEM, is used so frequently in technical publications. A flaw occurs, however, when a confidence interval based on the SEM is used to replace the SD as a descriptive statistic; if a description of data spread is needed, the SD should be used. As Feinstein[2] has observed, the reader of a research report may be interested in the span or range of the data, but the author of the report instead displays an estimated zone of the mean (SEM).

An absolute prohibition against the use of the SEM in medical reports is not desirable. There are situations in which the investigator is using a truly random sample for estimation purposes. Random samples of children have been used, for example, to estimate population parameters of growth. The essential element is that the investigator (and editor) recognize when descriptive statistics should be used, and when inferential (estimation) statistics are required.

SE OF PROPORTION

As mentioned previously, every sample statistic has its SE. With every statistic, there is a confidence interval that can be estimated. Despite the widespread use of SE (unspecified) and of SEM in medical journals and books, there is a noticeable neglect of one important SE, the SE of the proportion.

The discussion so far has dealt with measurement data or counts of elements. Equally important are data reported in proportions or percentages, such as, "Six of the ten patients with zymurgy syndrome had so-and-so." From this, it is an easy step to say, "Sixty percent of our patients with zymurgy syndrome had so-and-so." The implication of such a statement may be that the author wishes to alert other clinicians, who may encounter samples from the universe of patients with zymurgy syndrome that they may see so-and-so in about 60% of them.

The proportion—six of ten—has an SE of the proportion. As shown in Fig 2, the SE_p in this situation is the square root of (0.6×0.4) divided by ten, which equals 0.155. The true proportion of so-and-so in the universe of patients with zymurgy syndrome is in the confidence interval that falls symmetrically on both sides of six of ten. To estimate the interval, we start with 0.6 or 60% as the midpoint of the interval. At the 95% level of confidence, the interval is $0.6 \pm 1.96\ SE_p$, which is $0.6 \pm (1.96 \times 0.155)$, or from 0.3 to 0.9.

If the sample shows six of ten, the 95% confidence interval is between 30% (three of ten) and 90% (nine of ten). This is not a very narrow interval. The expanse of the interval may explain the almost total absence of the SE_p in medical reports, even in journals where the SEM and SD are used abundantly. Investigators may be dismayed by the dimensions of the confidence interval when the SE_p is calculated from the small samples available in clinical situations.

Of course, as in the measurement of self-contained data, the investigator may not think of his clinical material as a sample from a larger universe. But often, it is clear that the purpose of publication is to suggest to other investigators or clinicians that, when they see patients of a certain type, they might expect to encounter certain characteristics in some estimated proportion of such patients.

JOURNAL USE OF SD AND SE

To get empiric information about pediatric journal standards on descriptive statistics, especially the use of SD and SE, I examined every issue of the three major pediatric journals published in 1981: *American Journal of Diseases of Children*, *Journal of Pediatrics*, and *Pediatrics*. In a less systematic way, I perused several issues of *JAMA*, the *New England Journal of Medicine*, and *Science*.

Every issue of the three pediatric journals had articles, reports, or letters in which SD was mentioned, without specification of whether it was the descriptive SD or the estimate SD. Every issue of the *Journal of Pediatrics* contained articles using SE (unspecified) and articles using SEM. *Pediatrics* used SEM in every issue and the SE in every issue except one. Eight of the 12 issues of the *American Journal of Diseases of Children* used SE or SEM or both. All the journals used SE as if SE and SEM were synonymous.

Every issue of the three journals contained articles that stated the mean and range, without other indication of dispersion. Every journal contained reports with a number ± (another number), with no explanation of what the number after the plus-minus symbol represented.

Every issue of the pediatric journals presented proportions of what might be thought of as samples without indicating that the SE_p (standard error of the proportion) might be informative.

In several reports, SE or SEM is used in one place, but SD is used in another place in the same article, sometimes in the same paragraph, with no explanation of the reason for each use. The use of percentiles to describe nongaussian distributions was infrequent. Similar examples of stylistic inconsistency were seen in the haphazard survey of *JAMA*, the *New England Journal of Medicine*, and *Science*.

A peculiar graphic device (seen in several journals) is the use, in illustrations that summarize data, of a point and vertical bars, with no indication of what the length of the bars signifies.

A prevalent and unsettling practice is the use of the mean ± SD for data that are clearly not gaussian or not symmetric. Whenever data are reported with the SD as large or larger than the mean, the inference must be that several values are zero or negative. The mean ± 2 SDs should embrace about 95% of the values in a gaussian distribution. If the SD is as large as the mean, then the lower tail of the bell-shaped curve will go below zero. For many

biol
valu
zym
exis
A
enti
Med
tion
in re
cert
of 4.
ship
nece
state
cles
from
way,
zero
In
as a
rang
woul
De
ment
mear
show
data.
was 1
to 14
so 99
±2.5
and
belie

Th
est a
impr
cal li
point
taint
P val
Berw
tionn
ness
many
demic
ever,
atten
confu
In
urge t
trials
signif
ble to
look
sampl
thorit
to det

110

biologic data, there can be no negative values; blood chemicals, serum enzymes, and cellular elements cannot exist in negative amounts.

An article by Fletcher and Fletcher[4] entitled "Clinical Research in General Medical Journals" in a leading publication demonstrates the problem of ±SD in real life. The article states that in 1976 certain medical articles had an average of 4.9 authors ± 7.3 (SD)! If the authorship distribution is gaussian, which is necessary for ±SD to make sense, this statement means that 95% of the articles had 4.9±(1.96×7.3) authors, or from −9.4 to +19.2. Or stated another way, more than 25% of the articles had zero or fewer authors.

In such a situation, the SD is not good as a descriptive statistic. A mean and range would be better; percentiles would be logical and meaningful.

Deinard et al[5] summarized some mental measurement scores using the mean ± SD and the range. They vividly showed two dispersions for the same data. For example, one set of values was 120.8 ± 15.2 (SD); the range was 63 to 140. The SD implies gaussian data, so 99% of the values should be within ±2.58 SDs of the mean or between 81.6 and 160. Which dispersion should we believe, 63 to 140 or 81.6 to 160?

ADVICE OF AUTHORITIES

There may be a ground swell of interest among research authorities to help improve statistical use in the medical literature. Friedman and Phillips[6] pointed out the embarrassing uncertainty that pediatric residents have with *P* values and correlation coefficients. Berwick and colleagues,[7] using a questionnaire, reported considerable vagueness about statistical concepts among many physicians in training, in academic medicine, and in practice. However, in neither of these reports is attention given to the interesting but confusing properties of SD and SE.

In several reports,[8-10] the authors urge that we be wary when comparative trials are reported as not statistically significant. Comparisons are vulnerable to the error of rejecting results that look negative, especially with small samples, but may not be. These authorities remind us of the error of failing to detect a real difference, eg, between controls and treated subjects, when such a difference exists. This failure is called the "error of the second kind," the Type II error, or the beta error. In laboratory language, this error is called the false-negative result, in which the test result says "normal" but nature reveals "abnormal" or "disease present." (The Type I error, the alpha error, is a more familiar one; it is the error of saying that two groups differ in some important way when they do not. The Type I error is like a false-positive laboratory test in that the test suggests that the subject is abnormal, when in truth he is normal.)

In comparative trials, calculation of the Type II error requires knowledge of the SEs, whether the comparisons are of group means (requiring SEM) or comparisons of group proportions (requiring SE_p).

At the outset, I mentioned that we are advised[2,3] to describe clinical data using means and the SD (for bell-shaped distributions) and to eschew use of the SE. On the other hand, we are urged to examine clinical data for interesting confidence intervals,[11,12] searching for latent scientific value and avoiding a too hasty pronouncement of not significant. To avoid this hasty fall into the Type II error (the false-negative decision), we must increase sample sizes; in this way, a worthwhile treatment or intervention may be sustained rather than wrongly discarded.

It may be puzzling that some authorities seem to be urging that the SE should rarely be used, but others are urging that more attention be paid to confidence intervals, which depend on the SE. This polarity is more apparent than real. If the investigator's aim is description of data, he should avoid the use of the SE; if his aim is to estimate population parameters or to test hypotheses, ie, inferential statistics, then some version of the SE is required.

WHO IS RESPONSIBLE?

It is not clear who should be held responsible for data displays and summary methods in medical reports. Does the responsibility lie at the door of the investigator-author and his statistical advisors, with the journal referees and reviewers, or with the editors? When I ask authors about their statistical style, the reply often is, "The editors made me do it."

An articulate defender of good statistical practice and usage is Feinstein,[2] who has regularly and effectively urged the appropriate application of biostatistics, including SD and SE. In his book, *Clinical Biostatistics*, he devotes an entire chapter (chap 23, pp 335-352) to "problems in the summary and display of statistical data." He offers some advice to readers who wish to improve the statistics seen in medical publications: "And the best person to help re-orient the editors is you, dear reader, you. Make yourself a one-person vigilante committee."[2(p349)]

Either the vigilantes are busy in other enterprises or the editors are not listening, because we continue to see the kind of inconsistent and confusing statistical practices that Eisenhart[1] and Feinstein[2] have been warning about for many years. I can only echo what others have said: When one sees medical publications with inappropriate, confusing, or wrong statistical presentation, one should write to the editors. Editors are, after all, the assigned defenders of the elegance and accuracy of our medical archives.

References

1. Eisenhart C: Expression of the uncertainties of final results. *Science* 1968;160:1201-1204.
2. Feinstein AR: *Clinical Biostatistics*. St Louis, CV Mosby Co, 1977.
3. Glantz SA: *Primer of Biostatistics*. New York, McGraw-Hill Book Co, 1981.
4. Fletcher R, Fletcher S: Clinical research in general medical journals: A 30-year perspective. *N Engl J Med* 1979;301:180-183.
5. Deinard A, Gilbert A, Dodd M, et al: Iron deficiency and behavioral deficits. *Pediatrics* 1981;68:828-833.
6. Friedman SB, Phillips S: What's the difference?: Pediatric residents and their inaccurate concepts regarding statistics. *Pediatrics* 1981;68:644-646.
7. Berwick DM, Fineberg HV, Weinstein MC: When doctors meet numbers. *Am J Med* 1981;71:991-998.
8. Freiman JA, Chalmers TC, Smith H Jr, et al: The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 'negative' trials. *N Engl J Med* 1978;299:690-694.
9. Berwick DM: Experimental power: The other side of the coin. *Pediatrics* 1980;65:1043-1045.
10. Pascoe JM: Was it a Type II error? *Pediatrics* 1981;68:149-150.
11. Rothman KJ: A show of confidence. *N Engl J Med* 1978;299:1362-1363.
12. Guess H: Lack of predictive indices in kernicterus—or lack of power? *Pediatrics* 1982;69:383.

Br Heart J 1988;60:177–80

Editorial

The end of the p value?

STEPHEN J W EVANS,* PETER MILLS, JANE DAWSON

From the Departments of Clinical Epidemiology and Cardiology, The London Hospital; and the British Heart Journal

The application of statistical methods to medical data has been undergoing a sea-change. This is of particular importance in cardiology because the current methods that statisticians recommend express the results of studies in terms that are directly relevant to the clinical use to which they may be put. In March 1986 the *British Medical Journal* nailed its colours firmly to the mast, telling readers that "authors . . . will be expected to calculate confidence intervals whenever the data warrant this approach"[1,2] and the *Lancet*,[3,4] *Annals of Internal Medicine*, and *American Journal of Public Health* are among other journals that have endorsed the new orthodoxy. We expect that studies reported in the *British Heart Journal* will increasingly reflect this approach. The nuts and bolts of calculating the confidence intervals of various types of data are described in a series of articles in the *British Medical Journal*,[5-9] and below we review some aspects of the approach that are particularly relevant to papers published in the *British Heart Journal*.

Towards estimation and away from hypothesis testing

The null hypothesis generally states that there is no relation between the variables under study. For example, when the change in cardiac output before and after intervention is analysed the null hypothesis proposes that the average change is zero. It follows that calculation of the p value, which is based on the null hypothesis, is frequently an inappropriate statistical method for summarising the analysis of cardiological data. Many published studies do not seriously consider the possibility that an intervention has no effect. When a test intervention has been used the question usually being asked is "how great is its effect?" rather than "does it have an effect?"

This point may be illustrated by comparing cardiac output before and after administration of an inotropic drug. A paired *t* test with a p value starts with the hypothesis that the inotrope has no effect. It is unlikely that the drug would be under investigation if no effect on cardiac output were really expected. The questions for the clinician are "on average, how great is the change produced by the intervention" and "with what precision has the average change been estimated?" These questions are answered by the calculation of confidence intervals, whereas hypothesis testing can give only the answer "yes" or "no" to the question "Is there a change?"

Figure 1a is an example of data that arise in such a study. The paired *t* test gives a value of $t = 3\cdot3$ ($p = 0\cdot01$). It indicates that the rise is statistically significant but does not indicate the size of the rise. The appropriate 95% confidence interval which is shown in fig 1a is based on the mean change and two standard errors on either side of the mean and suggests the likely interval within which the true mean lies. Thus the confidence interval centred on the mean change of 0·6 l/min extends from a mean change of +0·2 l/min to one of +1·0 l/min. This implies that the true mean value could lie anywhere between 0·2 and 1·0 and that the data are unlikely to be consistent with a mean change of zero. A confidence interval that does not include zero is equivalent to a test with a statistically significant p value. When the confidence interval includes zero, as for example when the interval is from −0·2 l/min to +1·4 l/min then although the mean change remains the same, at +0·6 l/min, the possibility must be considered that the intervention causes a fall rather than a rise or that it causes no change at all. If the data arise from a smaller sample (see fig 1b in which n = 7 instead of n = 9) or if their standard deviation is larger (see fig 1c in which the standard deviation has increased by 25%) the confidence interval will be wider.

When the confidence interval includes zero the

*Statistical adviser to the *British Heart Journal*.

Requests for reprints to Jane Dawson, *British Heart Journal*, BMA House, Tavistock Square, London WC1H 9JR.

112

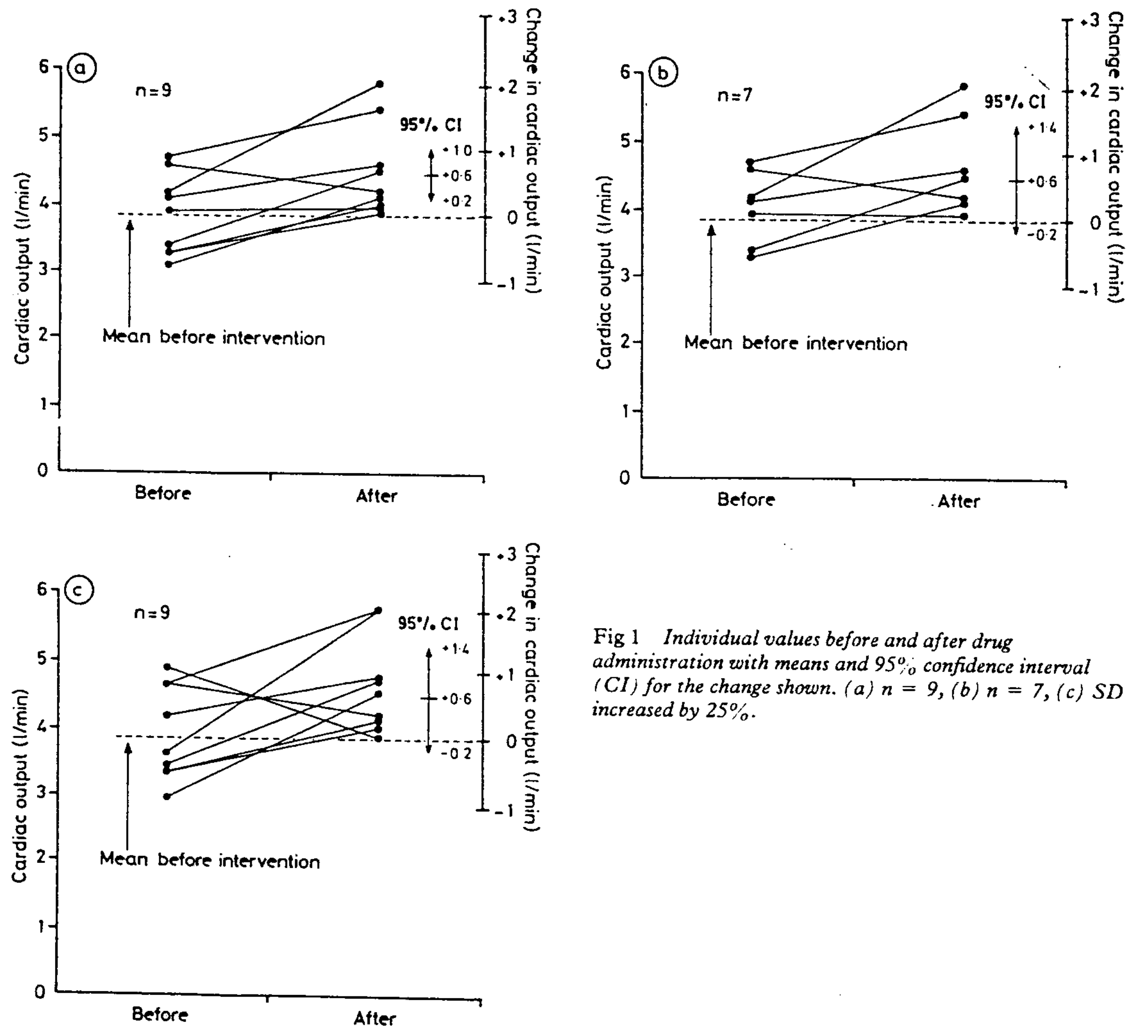


Fig 1 *Individual values before and after drug administration with means and 95% confidence interval (CI) for the change shown. (a) n = 9, (b) n = 7, (c) SD increased by 25%.*

result is equivalent to a significance test that gives a non-significant result. Relying on this feature alone is no better than the use of p values, but the upper limit of the confidence interval (+1·4) draws attention to the possibility that the average increase *might* be clinically useful.

The obvious advantage of a confidence interval is that it expresses results in the units in which the measurements were made, and so allows the reader to consider critically the clinical relevance of the results. If the sample size is small the confidence interval will be wide. The clinician must then examine the extremes of the interval. Do these extremes indicate that the clinical relevance of the results is consistent with the conclusions drawn from the analysis? If the conclusion is drawn that the drug has "no effect" because p is not statistically significant but the 95% confidence interval reaches 1·4 l/min (fig 1b) then it is clear that the drug may well have a positive effect that has not been demonstrated by this study. The confidence interval (of say +0·05 to +0·1 l/min) can also make it clear that a difference which is statistically significant (based on p values) is of no clinical relevance because the statistical significance of the result has been produced spuriously by a very large sample of say about 2000. In such a study even the upper value of the confidence interval suggests that such a change is too small to be of clinical benefit despite its statistical significance.

When the effects of two different drugs on cardiac output are being compared the appropriate test is an independent samples *t* test and the equivalent 95% confidence interval may also be calculated. In fig 2, drug A gives the same results as shown in fig 1a, while

The end o

the effec
statistical
test. Th
whether
the chan

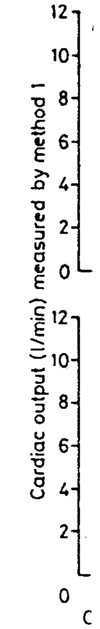
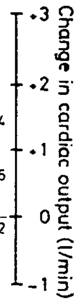


Fig 3 *(a*
two metho
values "in



rval
(c) SD

hen it is
fect that
y. The
nin) can
s statis-
clinical
e of the
ry large
ven the
ests that
benefit

cardiac
est is an
nt 95%
In fig 2,
a, while

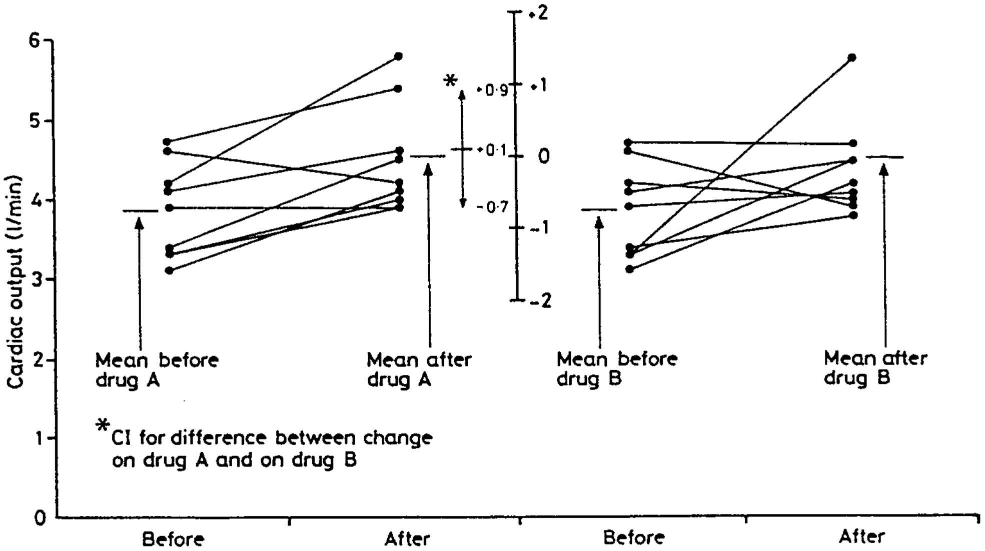


Fig 2 *Individual and mean values before and after administration of two drugs. The 95% confidence interval for difference between the changes with drug A and the changes with drug B is −0·7 to +0·9 l/min.*

the effect of drug B on cardiac output is "not statistically significant" when assessed by a paired *t* test. The relevant question is no longer simply whether each drug alters cardiac output but whether the change with drug A (δCO_A) is importantly different from that with drug B (δCO_B). The 95% confidence interval for *the difference between the changes with A and with B* (δCO_{AB}) is −0·7 to +0·9 l/min. This shows that although the change with A is statistically significant ($p < 0{\cdot}05$) and that with B is not statistically significant, there is insufficient evidence that the change with A is different from the change with B (because the 95% confidence interval for the difference in changes between A and B includes zero). At the same time confidence intervals show that potentially clinically important differences (for example of 0·9 l/min) between the drugs may not have been detected because the sample sizes were too small to produce significant p values.

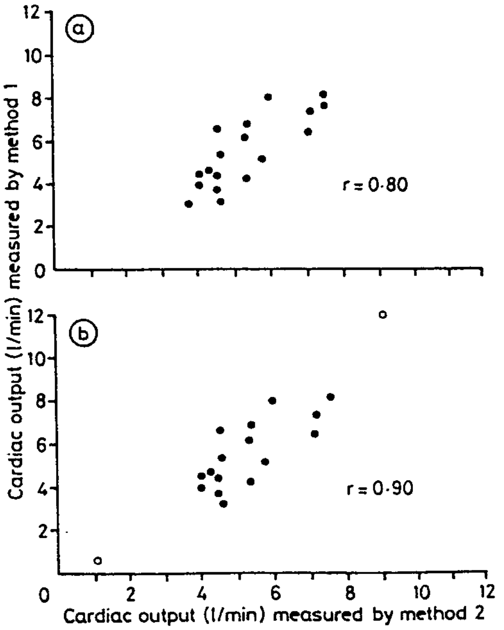


Fig 3 *(a) Traditional scatter diagram for comparison of two methods of measuring cardiac output. (b) Two outlier values "improve" the correlation.*

Thus confidence intervals provide all the information that significance tests give us and also indicate the clinical relevance of the information. Confidence intervals require little more calculation than the appropriate significance test.

Method comparison and the null hypothesis

Many investigations in cardiology compare two methods of measuring the same variable—for example cardiac output determined by Doppler echocardiography and by the Fick principle at cardiac catheterisation. In the past, such data have been summarised by correlation or regression coefficients and calculation of a p value (fig 3a). In both these methods the null hypothesis is tested. But the null hypothesis, which states there is no association between two variables, is not relevant to the

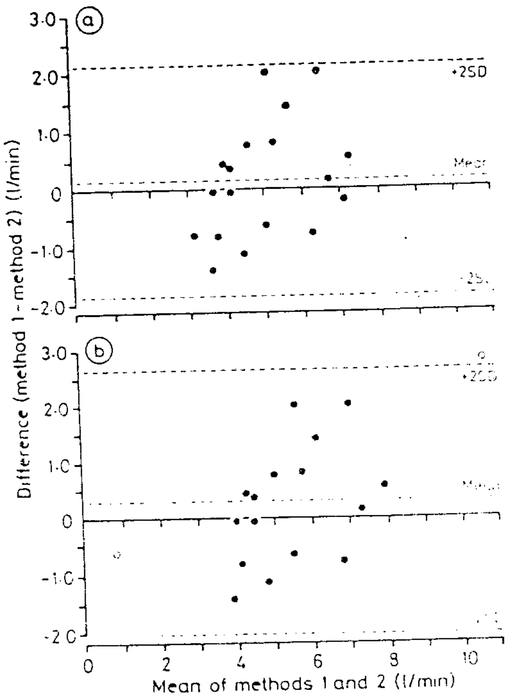


Fig 4 (a) *Diagram showing differences versus mean for data in fig 3a. (b) The two outliers (circles) in fact produce a worse agreement between the methods.*

measurement of the same variable by two different methods.[10] The magnitude of the correlation coefficient is strongly influenced by the range of values under study. In addition, its "significance" is increased simply by increasing the number of subjects studied. The apparently stronger correlation between the variables for the data shown in fig 3b (r = 0·90 *v* r = 0·80) is purely the result of the inclusion of two outliers. The correlation coefficient gives neither the magnitude of any possible discrepancy between the two methods nor whether such discrepancy is consistent over the range of values.

Like confidence intervals the method of analysis advocated by Bland and Altman[10] emphasises clinical relevance, which is determined by understanding the extent to which the two methods give different results—not by confirming that they show a little better than chance agreement when used to measure cardiac output. So fig 4a shows that method 1 gives slightly higher values than method 2 (the mean of the difference is higher than zero) and fig 4b shows that inclusion of the outlying values reduces the agreement rather than improves it (the standard deviation has increased from 1·00 to 1·17).

The key questions are what is the variability of a single observation (including its measurement error) and how much disagreement is there between the two methods of measurement? It is also important to be aware of systematic variation in the answers over the range of interest; for example when the two sets of measurements are examined does one method yield high values at the upper end of the range and low values at the lower end of the range? If it does, are these discrepant values genuine or are they spurious? Investigation of the methods together with some understanding of the possible clinical applications will be necessary to decide which is the better method of measurement. Lastly, while the two methods may agree over a wide range of values including those of normal individuals, does this degree of agreement between the techniques extend into the range of values commonly encountered in disease?

For some time now the *British Heart Journal* has been informally persuading authors who inappropriately use correlation and regression coefficients to use the method of Bland and Altman to examine agreement between methods. The fact that, over many years, correlation and regression have been misused is no reason to perpetuate a bad practice. Comparison of the r values obtained in different studies is meaningless. In addition, the *British Heart Journal* also recommends that confidence intervals should be given where relevant for studies that assess the effects of interventions, compare the effects of different drugs, or evaluate non-invasive techniques.

References

1 Langman MJS. Towards estimation and confidence intervals. *Br Med J* 1986;**292**:716.

2 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;**292**:746–50.

3 Anonymous. Report with confidence. *Lancet* 1987; **i**:488.

4 Bulpitt CJ. Confidence intervals. *Lancet* 1987;**i**:494–7.

5 Gardner MJ, Altman DG. Statistics in medicine: confidence intervals: estimating with confidence. *Br Med J* 1988;**296**:1210–1.

6 Altman DG, Gardner MJ. Statistics in medicine: calculating confidence intervals for regression and correlation. *Br Med J* 1988;**296**:1238–42.

7 Morris JA, Gardner MJ. Statistics in medicine: calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *Br Med J* 1988;**296**:1313–6.

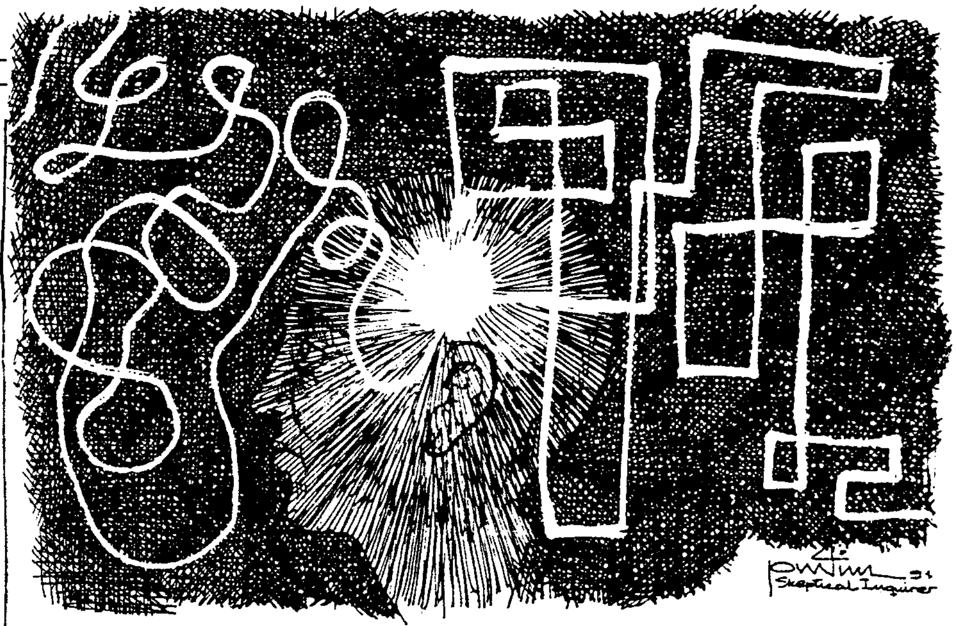
8 Machine D, Gardner MJ. Statistics in medicine: calculating confidence intervals for survival time analyses. *Br Med J* 1988;**296**:1369–71.

9 Campbell MJ, Gardner MJ. Statistics in medicine: calculating confidence intervals for some nonparametric analyses. *Br Med J* 1988;**296**:1454–6.

10 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.

Coincidences

JOHN ALLEN PAULOS



Coincidences fascinate us. They seem to compel a search for their significance. More often than some people realize, however, they are to be expected and require no special explanation. Surely no cosmic conclusions may be drawn from the fact that I recently and quite by accident met someone in Seattle whose father had played on the same Chicago high-school baseball team as my father had and whose daughter is the same age and has the same name as my daughter. As improbable as this *particular* event was, that *some* event of this vaguely characterized sort should occasionally occur is very likely.

More precisely it can be shown, for example, that if two Americans sit next to each other on an airplane, more than 99 times out of 100 they will be linked in some way by two or fewer intermediates. (The linkage with my father's classmate was more striking. It was via only one intermediate, my father, and contained other elements.) Maybe, for example, the cousin of one of the passengers will know the other's dentist. Most of the time people won't discover these links, since in casual conversation they can't run through all their 1,500 or so acquaintances as well as all their acquaintances' acquaintances. (I suppose with laptop computers becoming more popular they could compare their own personal databases and even those of people they know. Perhaps exchanging databases might soon be as common as leaving a business card. Electronic networking. Hellacious.)

There is a tendency, however, to home in on likely co-acquaintances. Such connections are thus discovered frequently enough that the squeals of amazement that commonly accompany their discovery are unwarranted. Similarly unimpressive is the "prophetic" dream, which traditionally comes to light after some natural disaster has occurred. Given the half-billion hours of dreaming each night in this country—two hours a night for 250 million people—we should expect as much.

Or consider the famous birthday problem in probability theory. One must gather together 367 people (one more than the number of days in a leap year) in order to ensure that 2 of them share a birthday. But if one is willing to settle for a 50-50 chance of this happening, only 23 people need be gathered. Rephrasing, I note that if we imagine a school with thousands of classrooms each of which contains 23 students, then approximately half of these classrooms will contain 2 students who share a birthday. No time should be wasted trying to explain the meaning of these or other coincidences of similar type. They just happen.

One somewhat different example concerns the publisher of a stock newsletter who sends out 64,000 letters extolling his state-of-the-art database, his inside contacts, and his sophisticated econometric models. In 32,000 of these letters he predicts a rise in some stock index for the following week, say, and in 32,000 of them he predicts a decline in that same index. Whatever happens he sends a follow-up letter, but only to those 32,000 to whom he's made the correct "prediction." To 16,000 of them he predicts a rise in that index for the next week, and to 16,000 a decline. Again, whatever happens he will have sent two consecutive correct predictions to 16,000 people. Iterating this procedure of focusing exclusively on the winnowed list of people who have received only correct predictions, he can create the illusion in them that he knows what he's talking about. After all, the 1,000 or so remaining people who have received six straight correct predictions (by coincidence) have a good reason to cough up the $1,000 the newsletter publisher requests: they want to continue to receive these "oracular" pronouncements.

I repeat that in discussing these and other coincidences it is useful to distinguish between generic sorts of events and particular events. Many situations are such that the particular event that occurs is guaranteed to be rare—a certain individual winning the lottery or a specific bridge hand being dealt—while the generic outcome—someone's winning the lottery or some bridge hand being dealt—is unremarkable. Consider the birthday problem again. If all that we require is that two people have some birthday in common rather than any particular birthday, then 23 people suffice to make this happen with probability 1/2. By contrast, 253 people are needed in order for the probability to be 1/2 that one of them has a specific birthdate, say July 4. Particular events specified beforehand are, of course, quite difficult to forecast, so it's not surprising that predictions by televangelists, quack doctors, and others are usually vague and amorphous (that is, until the events in question have occurred, at which time the prognosticators like to assert that these precise outcomes were indeed foreseen).

This brings me to the so-called Jeane Dixon effect, whereby the few correct predictions (by psychics, disreputable stock newsletters, or whomever) are widely heralded, and the 9,839 or so false predictions made annually are conveniently ignored. The phenomenon is quite widespread and contributes to the tendency we all have to read more significance into coincidences than is usually justified. We forget all the premonitions of disaster we've had that *didn't* predict the future and remember vividly those couple that seemed to do so. Instances of seemingly telepathic thought are reported to everyone we know; the incomparably vaster number of times this doesn't occur is too banal to mention.

Even our biology conspires to make coincidences appear more meaningful than they usually are. Since the natural world of rocks, plants, and rivers doesn't seem to offer much evidence for superfluous coincidences, primitive man had to be very sensitive to every conceivable anomaly and improbability as he slowly developed science and its progenitor "common sense." Coincidences, after all, *are* sometimes quite significant. In our complicated and largely man-made modern world, however, the plethora of connections among us appears to overstimulate many people's inborn tendency to note coincidence and improbability and lead them to postulate causes and forces where there are none. People know more names (not only family members', but also those of colleagues and a myriad of public figures), dates (from news stories to personal appointments and schedules), addresses (whether actual physical ones or telephone numbers, office numbers, and so on), and organizations and acronyms (from the FBI to the IMF, from AIDS to ASEAN) than ever before. Thus, although it is a very difficult quantity to measure, the rate at which coincidences occur has probably risen over the past century or two. Still, for most of them it generally makes little sense to demand an explanation.

In reality, the most astonishingly incredible coincidence imaginable would be the complete absence of all coincidences.

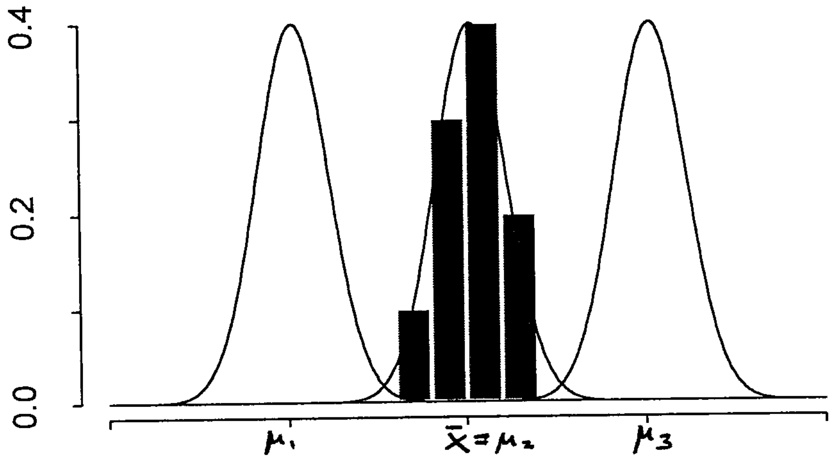
Note

Brief derivations of birthday statements: (1) The probability of 2 people having different birthdays is 364/365; of 3 people having different birthdays, (364/365 × 363/365); of 4 people, (364/365 × 363/365 × 362/365); of 23, (364/365 × 363/365 × 362/365 × . . . × 342/365 × 343/365), which product turns out to equal 1/2. Thus the complementary probability that at least 2 people share a birthday is also 1/2 (one minus the above product). (2) The probability someone does not have a July 4 birthday is 364/365; the probability neither of 2 people has a July 4 birthday is 364/365; the probability neither of 2 people has a July 4 birthday is $(364/365)^2$; none of 3, $(364/365)^3$, none of 253, $(364/365)^{253}$, which turns out to equal 1/2. Thus the complementary probability that at least one of the 253 people has a July 4 birthday is also 1/2, $1 - (364/365)^{253}$.

John Allen Paulos is professor of mathematics at Temple University and author of the best-selling Innumeracy *and of the just-published* Beyond Numeracy: The Ruminations of a Numbers Man *(Knopf, April 1991), from which this is published by permission.*

Maximum Likelihood Estimation

Suppose that we have a sample $x_1, x_2, \ldots, x_n$ from a Normally distributed population, with unknown mean μ, and known variance, $\sigma^2 = 1$. We would like to estimate μ from the data. An obvious answer is to use the sample mean, $\bar{x}$, but why is this a good choice? The answer is that out of all possible estimates of μ that we could choose, $\bar{x}$ maximizes the probability of the observed data, given μ. Thus it is a *maximum likelihood estimate.*



Of all possible choices for μ,
which is most likely, given the observed data?

For each data point, the probability density function is given by

$$f(x_i|\mu) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i - \mu)^2}{2}).$$

Since the data are independent, from the product rule,

$$f(x_1, x_2, \ldots, x_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i - \mu)^2}{2}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}).$$

This function is called the *likelihood function*, as it gives the likelihod or probability of the data for each value of μ.

We wish to maximize this probability function, i.e., pick the value of μ that gives the highest probability to the observed data. From calculus, we know that to maximize a function, we take its derivative, and set it equal to zero. Thus we will solve

$$\frac{df(x_1, x_2, \ldots, x_n|\mu)}{d\mu} = 0,$$

which, using the power rule, rule for exponentials, and the chain rule, becomes

$$\frac{df(x_1, x_2, \ldots, x_n|\mu)}{d\mu} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2}) \times \left[-\frac{1}{2}\sum_{i=1}^n 2(x_i - \mu)(-1)\right] = 0.$$

Since a constant is never zero, and an exponential is never zero, if the derivative is to equal zero, it must be that the term

$$\left[-\frac{1}{2}\sum_{i=1}^n 2(x_i - \mu)(-1)\right] = 0.$$

But this means that

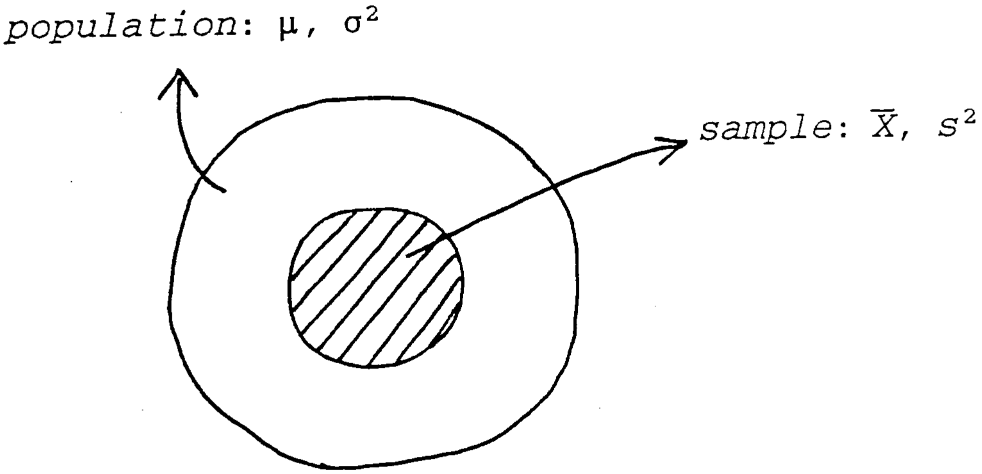
$$\sum_{i=1}^n (x_i - \mu) = \sum_{i=1}^n x_i - \sum_{i=1}^n \mu = 0,$$

so that

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}.$$

Thus the MLE is $\hat{\mu} = \overline{x}$.

PROPERTIES OF ESTIMATORS



1. Unbiased (Assume Random Sample)

$$E(\overline{X}) = \mu, \ E(s^2) = \sigma^2, \text{ unbiased}$$

$$E\left(\frac{\sum(X_i - \overline{X})^2}{n}\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2, \text{ biased}$$

2. Consistent

As sample size gets larger,

$$\overline{X} \to \mu \qquad \text{gets closer and closer}$$

$$s^2 \to \sigma^2$$

3. Maximum Likelihood

What is "most likely" to have been the mean, given the observed data? $\overline{X}$

$\overline{X}$ is maximum likelihood estimate, but s^2 is not.

SURVIVAL DATA FROM RADIATION AND BRAIN CANCER EXPERIMENT

		Group 1	Group 2
		New Schedule	Standard Schedule
		SURV 1	SURV 2
CASE	1	10.	17.
CASE	2	9.	15.
CASE	3	21.	16.
CASE	4	7.	5.
CASE	5	28.	24.
CASE	6	13.	10
CASE	7	1.	9.
CASE	8	11.	9.
CASE	9	20.	19.
CASE	10	16.	3.
CASE	11	1.	12.
CASE	12	2.	18.
CASE	13	1.	27.
CASE	14	5.	25.
CASE	15	10.	8.
CASE	16	3.	19.
CASE	17	8.	17.
CASE	18	1.	9.
CASE	19	19.	7.
CASE	20	40.	16.
CASE	21	10.	4.
CASE	22	6.	15.
CASE	23	1.	18.
CASE	24	11.	9.
CASE	25	7.	16.
CASE	26	11.	20.
CASE	27	22.	18.
CASE	28	13.	12.
CASE	29	11.	19.
CASE	30	7.	11.

TOTAL OBSERVATIONS: 30

SURVIVAL DATA FROM RADIATION AND BRAIN CANCER EXPERIMENT
(continued)

	SURV 1	SURV 2
N OF CASES	30	30
MINIMUM	1.	3.
MAXIMUM	40.	27.
MEAN	10.80	14.34
STANDARD DEV	8.86	6.18

CONFIDENCE INTERVAL FOR μ

$\overline{X}=10.8$ $\qquad$ $\overline{X_2}=14.3$

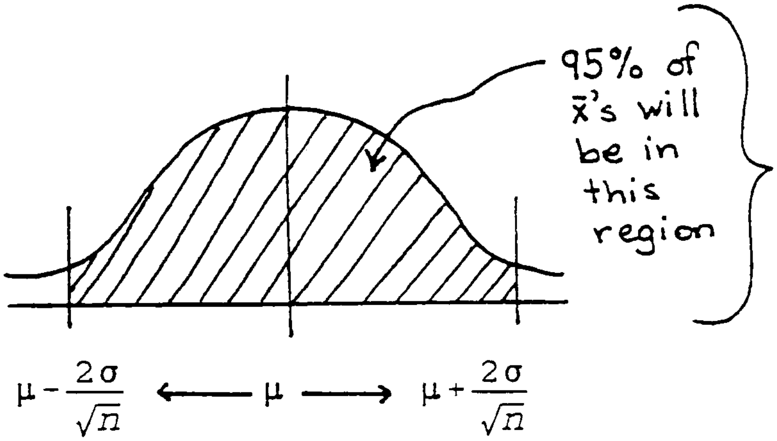
$s_1^2=(8.9)^2$ $\qquad$ $s_2^2=(6.2)^2$

$s.d.(\overline{X_1})=\dfrac{8.9}{\sqrt{30}}=1.6$ $\qquad$ $s.d.(\overline{X_2})=\dfrac{6.2}{\sqrt{30}}=1.13$

What can be said about μ_1, μ_2?

$$\overline{X}\sim N(\mu,\frac{\sigma}{\sqrt{30}})$$

$(2 \approx 1.96)$


95% of
$\bar{x}$'s will
be in
this
region
$\mu-\frac{2\sigma}{\sqrt{n}}$
μ
$\mu+\frac{2\sigma}{\sqrt{n}}$

(unknown)

⇒ so 95% of all intervals $(\overline{X}-\frac{2\sigma}{\sqrt{n}},\ \overline{X}+\frac{2\sigma}{\sqrt{n}})$ should contain μ.

If $s=\sigma$ exactly, then:

Hence: $10.8 \pm \dfrac{2\ (8.9)}{\sqrt{30}}$ $\qquad$ $14.3 \pm \dfrac{2\ (6.2)}{\sqrt{30}}$

$\downarrow$ $\qquad$ $\downarrow$

(7.55, 14.04) $\qquad$ (12.03, 16.56)
95% CI for μ_1 $\qquad$ 95% CI for μ_2

If not, use t table values
e.g., $t_{30,\ 0.025}=2.042$

Interpretation:

"If we used such a procedure repeatedly to form confidence intervals for μ, then 95% of such intervals would contain the true value of μ".

In practice, we use the procedure once, so that either:

(a) The interval does contain μ

(b) The interval does <u>not</u> contain μ, i.e., we had one of the 5% unlucky samples

We never know for any particular calculation of a CI if we are in situation (a) or (b). However, <u>in the long run</u>, we will be in (a) 95% of the time.

95% CONFIDENCE INTERVAL FOR $\mu_2 - \mu_1$

Unpooled: $\overline{X}_2 - \overline{X}_1 \pm t^* \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$,

$$t^* = t_{29, 0.025} = 2.045$$

$$14.3 - 10.8 \pm 2.045 \sqrt{\frac{(8.9)^2}{30} + \frac{(6.2)^2}{30}}$$

$$\Downarrow$$

$$(-0.55 , 7.55)$$

Pooled: $s_p^2 = \dfrac{(29)(8.9)^2 + (29)(6.2)^2}{30+30-2} = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

$$= 58.83 \Rightarrow s_p = \sqrt{58.83} = 7.66$$

$$\overline{X}_2 - \overline{X}_1 \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

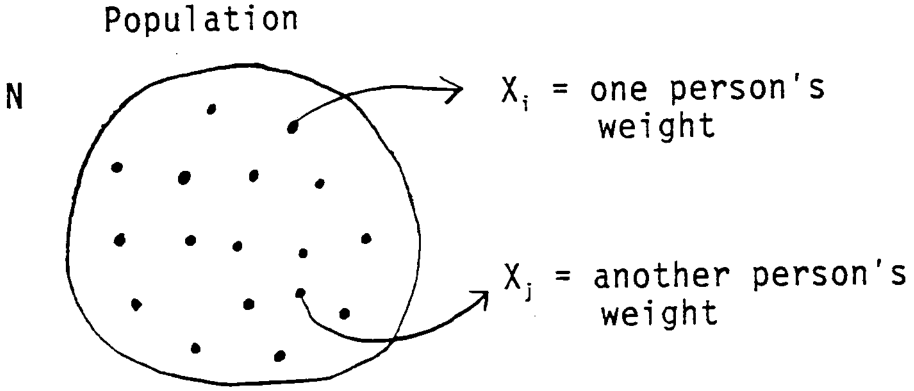
$$14.3 - 10.8 \pm (2.045)(7.66) \sqrt{\frac{1}{30} + \frac{1}{30}}$$

$$\Downarrow$$

$$(-0.55 , 7.55)$$

OVERVIEW: C.I.'s FOR MEANS

X = some measurement
(e.g.) = weight in kg of a person from
a particular population

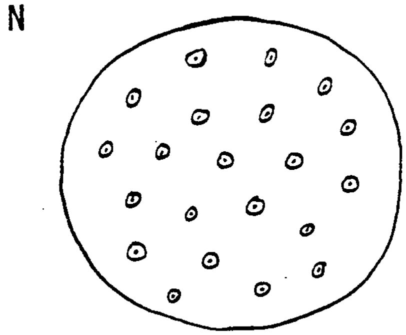


weight of Canadian
population, aged 25+ yrs.

Basic Question: μ = mean = ?

What is average weight?

<u>In theory</u>: "sample" everybody (census) to get exact values

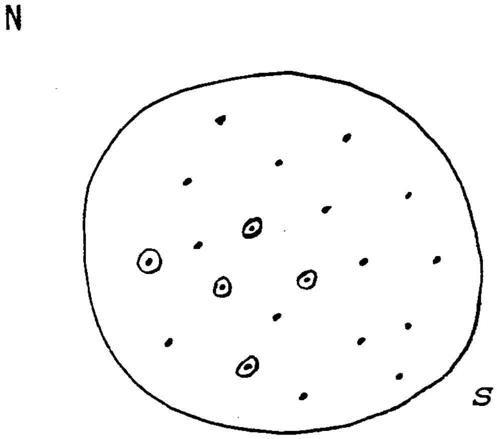


$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \textit{exact mean weight}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N} = \textit{var}$$

TECHNICAL POINT:
"N" is often = infinity

<u>In Practice</u>: Sample only a few from the population.



$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \textit{sample mean}$$

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} = \textit{sample variance}$$

$n = \textit{sample size}$

$X_1, X_2, \ldots, X_n$

	Pop X	Sample $\bar{X}$	Point Estimation
mean	μ	μ	$\bar{X}$ used to estimate μ
var	σ^2	σ^2/n	s^2 " " " σ^2 s^2/n " " " σ^2/n
s.d.	σ	$\sigma/\sqrt{n}$	s " " " σ $s/\sqrt{n}$ " " " $\sigma/\sqrt{n}$ $s/\sqrt{n}$ called "standard error"
increase n = sample size	All ↑ stay the same	σ^2/n $\sigma/\sqrt{n}$ decrease	$\bar{X}$, s^2, s, s^2/n, $s/\sqrt{n}$ are become "more accurate" points estimates.

$\bar{X}$ estimates μ, but rarely will $\bar{X} = \mu$ exactly.

Question: How far off can we be?

μ in interval $(\bar{X}-?, \bar{X}+?)$

CONFIDENCE INTERVALS

Sampling theory says if we take

$$(\overline{X} - t_{\alpha/2,n-1} s/\sqrt{n},\ \overline{X} + t_{\alpha/2,n-1} s/\sqrt{n})$$

then on average, on repeated use of this procedure, μ will fall in this interval 95% of the time ($\alpha = 0.05$).

TECHNICAL POINT:

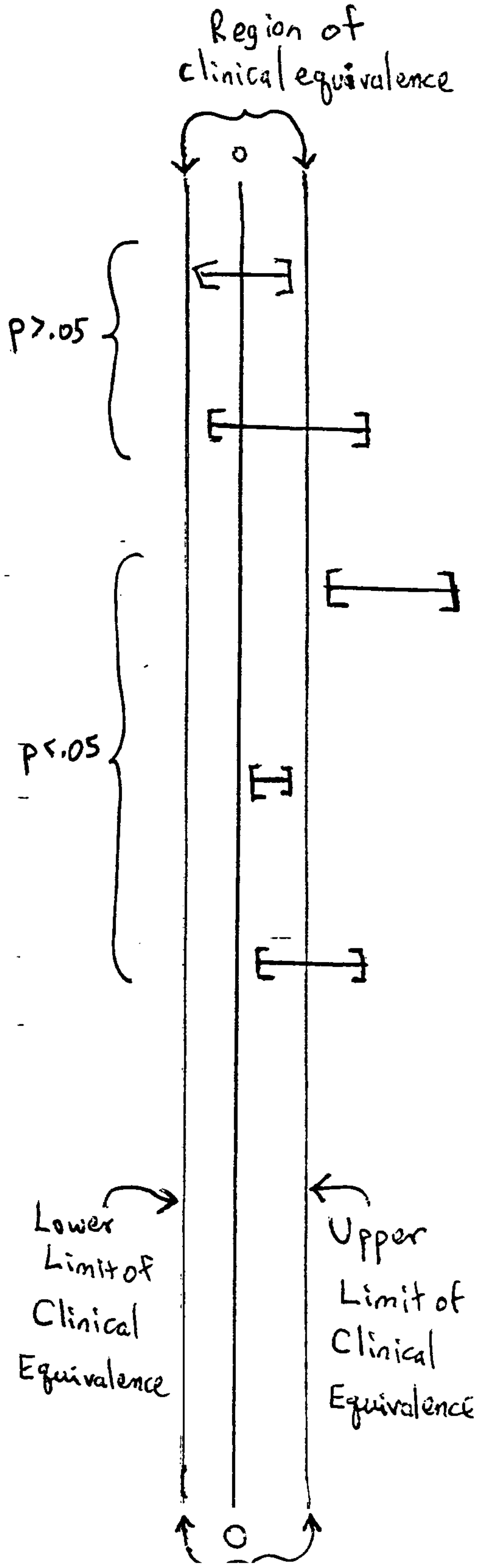
C.I.'s for other types of means may be formed by replacing $\overline{X}$ by $\overline{X}_1 - \overline{X}_2$, or $\overline{(X_1 - X_2)}$, and appropriate "s". If σ is assumed known, replace s by σ and $t_{\alpha/z,\, n-1}$ by $Z_{\alpha/z}$

Interpreting Confidence Intervals in Practice

Suppose that you have just calculated a confidence interval for a certain parameter. There are five possible conclusions that can be drawn, depending on where the upper and lower confidence interval limits fall in relation to the upper and lower limits of the region of clinical equivalence. The region of clinical equivalence, sometimes called the region of indifference, is the region inside of which both treatments would be considered to be the same for all practical purposes.

1. The CI includes zero, and both upper and lower CI limits, if they were the true values, would not be interesting to me clinically. Therefore, this variable has been shown to have no effect.
2. The CI includes zero, but one or both of the upper or lower CI limits, if they were the true values, would be interesting to me clinically. Therefore, the results of this variable in this study is inconclusive, and further evidence needs to be collected.
3. The CI does not include zero, and all values inside the upper and lower CI limits, if they were the true values, would be interesting to me clinically. Therefore, this study shows this variable to be important.
4. The CI does not include zero, but all values inside the upper and lower CI limits, if they were the true values, would not be interesting to me clinically. Therefore, this study shows this variable, while having some small effect, is not clinically important.
5. The CI does not include zero, but only some of the values inside the upper and lower CI limits, if they were the true values, would be interesting to me clinically. Therefore, this study shows this variable has at least a small effect, and may be clinically important. Further study is required in order to better estimate the magnitude of this effect.

Of course, the same thinking applies to credible intervals from Bayesian analyses. If there is little or no prior information, Bayesian credible intervals and frequentist confidence intervals usually are very similar. Futhermore, prior distributions are not very influential for large sample sizes, where again Bayesian credible sets would likely be very similar to CI's. However, Bayesian thinking may be helpful in the same kinds of problems with smaller sample sizes, where prior information may influence decision making.



Sample Sizes via Confidence Intervals

As previously discussed, there has been a strong trend away from hypothesis testing and p-values towards the use of confidence intervals in the reporting of results from biomedical research. Since the design phase of a study should be in sync with the analysis that will be eventually performed, sample size calculations should be carried out on the basis of ensuring adequate numbers for accurate estimation of important quantities that will be estimated in our study, rather than by power calculations.

The question of how accurate is "accurate enough" can be addressed by carefully considering the results you would expect to get (a bit of a "Catch 22" situation, since if you knew the results you will get, there would be no need to carry out the experiment!), and making sure your interval will be small enough to land in intervals numbered 1, 3, or 4 of the previous page. This is a non-trivial exercise, not to be taken lightly.

As with power calculaitons for means, there are two different formulae, depending if you are in a single of two-sample situation. These are derived by solving for the sample size n in the formulae for the confidence intervals.

Single Sample: Let μ be the mean that is to be estimated, and assume that we wish to estimate μ to an accuracy of a total CI width of w (so that the CI will be $\overline{x} \pm d$, where $2 \times d = w$). Let σ be the standard deviation in the population.

Then

$$n = \frac{z_{1-\alpha/2}^2 \sigma^2}{d^2} = \frac{4 \times z_{1-\alpha/2}^2 \sigma^2}{w^2}$$

Two Sample: Let μ_1 and μ_2 be the means of two populations, and that we would like an accurate estimate of $\mu_1 - \mu_2$. Again assume a total CI width of w (so that again $2 \times d = w$). Let σ_1 and σ_2 be the standard deviations in each population, respectively.

Then

$$n = \frac{z_{1-\alpha/2}^2 (\sigma_1^2 + \sigma_2^2)}{d^2} = \frac{4 \times z_{1-\alpha/2}^2 (\sigma_1^2 + \sigma_2^2)}{w^2}$$

where now n represents the required sample size **for each group**. As usual, $z_{1-\alpha/2}$ is 1.96 for a 95% confidence interval, etc.

(See sample size calculator on home page!)

Frequentist versus Bayesian Inference

EXAMPLE: Consider the situation where a group of children are given an intelligence test. Suppose that the data are:

Child #	Score on IQ test
1	105
2	97
3	95
4	100
5	104
6	90
7	116
8	113
9	101
10	109

Then $\overline{x} = 103.0$, and $s^2 = 65.78$.

Let us assume the following:

1. The standard deviation is known *a priori* to be 8 units.
2. The observations come from a Normal distribution, i.e.,

$$x_i \sim N(\mu, \sigma^2 = 8^2), \quad \text{for } i = 1, 2, \ldots, 10.$$

[Note: Neither of these two assumptions are necessary, except that they simplify the problem to make the comparison of the two paradigms easier. We can remove the first assumption by using a t-test instead of a test based on the Normal distribution, and the second can be removed by using the Central Limit Theorem for large sample sizes, or using a non-parametric or other test.]

Suppose that we wish to test the hypotheses:

$$\begin{array}{rl} H_0: & \mu \leq 100 \\ & \text{vs} \\ H_A: & \mu > 100 \end{array}$$

Since

$$x_i \sim N(\mu, \sigma^2 = 8^2), \quad \text{for } i = 1, 2, \ldots, 10,$$

we know that

$$\overline{x} \sim N(\mu, \frac{8^2}{10}), \quad \text{since } N = 10.$$

Thus, if H_0 is correct,

$$\frac{\overline{x} - 100}{8/\sqrt{10}} = 1.185$$

Looking up 1.185 on Normal tables (one-sided test) gives a p-value of 0.118, which is usually classified as evidence not to reject H_0.

Comments

1. μ is regarded as a *fixed* but unknown parameter about which we want to make inference. Since it is fixed, one cannot directly make probability statements about μ.
2. No use of prior information about the children was used, or even discussed. If there was any prior information available, it would usually only be used informally, after looking at the results of the test. For example, if one knew that the children were from a school for "gifted" children, one might reassess the p-value of 0.118 as "close to significance", and "in the expected direction", but this would remain unquantifiable.
3. The p-value (0.118) says nothing about the probability that the null or alternative hypotheses are correct. For this, we must refer to the positive or negative "predictive values" of the test, about which the p-value says nothing.

On the other hand:

1. Bayesian analysis regards μ as a *random* parameter. Since we do not know the value of μ, we can represent our uncertainty in a probability distribution that summarizes what we do know. If we did know the value of μ exactly, then our distribution reduces to a single point with probability one. This will rarely be the case, so in general we consider a range of values, and attach a probability with each subset within that range. Our goal to deduce that distribution (or its parameters, if the form is known, e.g., here it will be Normal). All inference is then based on that distribution.
2. Prior information (when available) is formally incorporated into the model.
3. We can directly calculate the probabilities of the null and alternative hypotheses. Note, however, that this comes at the price of having to specify a prior distribution, which may not always be trivial (or even possible) to do in practice.

Bayesian Approach

Suppose, as before, that the data follow a normal distribution,

$$x_i \sim N(\mu, \sigma^2) = N(\mu, 8^2), \quad \text{for } i = 1, 2, \ldots, 10.$$

Suppose further that we have *a priori* information that the random parameter μ is likely to be in the interval (60,140) according to

$$\mu \sim N(\theta, \tau^2) = N(100, 400). \tag{0.1}$$

Thus we have a "two-stage" procedure: First, a μ is selected from $N(100, 400)$. We do not directly observe this μ. Then, we observe the $x_i \sim N(\mu, 8^2)$.

This choice for a prior is based on any information that may be available at the time of the experiment. In this case, the prior distribution was chosen to have a very large standard deviation ($\tau = 20$) to reflect that we have very little prior information. The prior is centered around $\mu = 100$, so that the prior probabilities of the null and alternative hypotheses are both equal to one half, i.e.,

$$Pr\{\mu \leq 100\} = Pr\{\mu > 100\} = 0.5.$$

We now look at the data, summarized by $\overline{x}$=103. This data, together with the prior distribution, are then combined into a posterior distribution. The combination is carried out by a version of Bayes Theorem, based on the same principle as what we have seen before, but modified by the fact that the distribution is continuous.

The idea is to combine the prior information about μ together with the information provided by the data, represented by the likelihood function, into a final *posterior distribution.* Thus we have

$$\text{posterior distribution} = \frac{\text{prior distribution} \times \text{likelihood of the data}}{\text{a normalizing constant}}$$

The precise formula is

$$f(\mu | x_1, \ldots, x_n) = \frac{f(\mu) \times f(x_1, \ldots, x_n|\mu)}{\int_{-\infty}^{+\infty} f(\mu) \times f(x_1, \ldots, x_n|\mu)\, d\mu} \tag{0.2}$$

In our case, the prior is given by (0.1), and the likelihood function for the data is based on the Normal distribution, i.e.,

$$f(x_1, x_2, \ldots, x_n|\mu) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i - \mu)^2}{2\sigma^2}) \;=\; \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}).$$

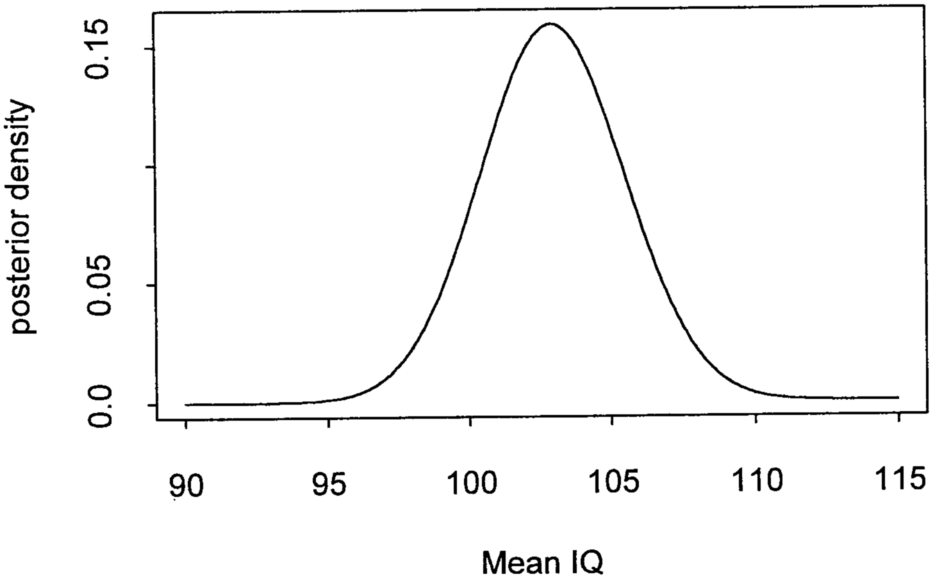
Using (2), the posterior distribution is given by:

$$\text{Posterior of } \mu \sim N(A \times \theta + B \times \overline{x}, \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2})$$

where

$A = \frac{\sigma^2/n}{\tau^2+\sigma^2/n} = 0.0157$
$B = \frac{\tau^2}{\tau^2+\sigma^2/n}$=.9843
$n = 10$
$\sigma = 8$
$\tau = \sqrt{400} = 20$
$\theta = 100$, and
$\overline{x} = 103$

Hence $\mu \sim N(102.95, 6.30)$, so the posterior distribution is



The mean value depends on both the prior mean, θ, and the observed mean, $\overline{x}$.

Note that this is interpreted as the actual probability density of μ, so that we can calculate the probabilities of the null and alternative hypotheses.

$$\begin{aligned} Pr\{H_0 \text{ is true } \mid \text{the data}\} &= Pr\{\mu \leq 100 \mid \mu \sim N(102.95, 6.30)\} \\ &= Pr\{\frac{\mu - 102.95}{\sqrt{6.30}} \leq \frac{100 - 102.95}{\sqrt{6.30}}\} \\ &= Pr\{Z \leq -1.175\} \end{aligned}$$

$$= \quad 0.12$$

Similarly,

$$\begin{aligned} Pr\{H_A \text{ is true } \mid \text{the data}\} &= Pr\{\mu > 100 \mid \mu \sim N(102.95, 6.30)\} \\ &= Pr\{\frac{\mu - 102.95}{\sqrt{6.30}} > \frac{100 - 102.95}{\sqrt{6.30}}\} \\ &= Pr\{Z > -1.175\} \\ &= 0.88 \end{aligned}$$

One can just as easily calculate *credible intervals*, which are the Bayesian analogues to frequentist confidence intervals. Using the fact that $\mu \sim N(102.95, 6.30)$, a 95% posterior credible interval for μ is (98.0, 107.9).

132B

Placing Trials in Context Using Bayesian Analysis

GUSTO Revisited by Reverend Bayes

James M. Brophy, MD, Lawrence Joseph, PhD

Standard statistical analyses of randomized clinical trials fail to provide a direct assessment of which treatment is superior or the probability of a clinically meaningful difference. A Bayesian analysis permits the calculation of the probability that a treatment is superior based on the observed data and prior beliefs. The subjectivity of prior beliefs in the Bayesian approach is not a liability, but rather explicitly allows different opinions to be formally expressed and evaluated. The usefulness of this approach is demonstrated using the results of the recent GUSTO study of various thrombolytic strategies in acute myocardial infarction. This analysis suggests that the clinical superiority of tissue-type plasminogen activator over streptokinase remains uncertain.

(*JAMA*. 1995;273:871-875)

BEFORE any clinical trial results are available, different clinicians will have different opinions regarding the relative benefits of the therapies under study. These opinions will usually range from skepticism to enthusiasm for a new therapy compared with a standard therapy. Regardless of how well it is conducted, no single clinical trial can provide absolutely definitive conclusions. Thus, even after trial results are reported, it is reasonable to expect that a diversity of opinions will persist, although perhaps with some convergence toward the observed trial results. The degree of convergence will depend on the strength of the trial in terms of sample size and scientific rigor in its execution. Therefore, in any medical experiment, clinical researchers must give careful consideration to issues of both design and analysis. Randomized clinical trials are almost universally accepted as the gold standard design for comparative clinical research, since bias and confounding are minimized. Much attention has been directed to the scientific reasoning behind statistical analysis in the medical and statistical literature.[1-3] However, while most clinicians are aware of the importance of good experimental designs, few are aware of the full array of statistical methods available. Some of these methods allow for the reporting of a range of conclusions corresponding to the diversity of prior opinions. They can also answer directly questions of interest to clinicians.

Classical (frequentist) analysis is the most prevalent statistical method used, leading to the ubiquitous *P* values and confidence intervals. *P* values from research trials may be viewed as analogs of false-positive (1−specificity) diagnostic tests. If neither the disease nor the treatment is malignant, we may well accept test specificity of 95% (*P*=.05). However, before accepting a limb amputation for osteosarcoma, we would rightly demand a false-positive value much less than .05. Generally, we are more interested in knowing what is the probability of disease given the test result (analogous to predictive value), and this cannot be supplied from classical statistical considerations alone. Clinicians routinely interpret diagnostic test results in the "clinical context," that is, by considering the background rate of the disease in a given population. In a similar manner, the interpretation of clinical trials should be considered in the light of preexisting knowledge.[1] (The analogy between hypothesis testing and diagnostic testing is completed by noting that statistical power corresponds to the sensitivity of a diagnostic test.)

In the classical approach, model parameters such as population means are fixed (nonrandom) quantities and probability distributions are considered only for test statistics (such as the *t* statistic in a *t* test). The randomness of test statistics arises because frequentists must consider not only the observed data in a given experiment, but also other data that might have occurred had the experiment been repeated. Each of these hypothetical repetitions leads to a different value of the test statistic, and the collection of these form a distribution. It is this distribution that is used to calculate *P* values and confidence intervals.

Rather than directly addressing desired clinical questions, such as "Which treatment is superior?" or "What is the probability of a clinically meaningful treatment difference?," classical analysis usually examines the null hypothesis of no difference between the competing strategies. *P* values denote the probability that a statistic as extreme as or more extreme than the observed test statistic would occur on hypothetical re-

From the Department of Medicine, Centre Hospitalier de Verdun (Quebec) (Dr Brophy); Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec (Drs Brophy and Joseph); and Division of Clinical Epidemiology and Centre for the Analysis of Cost Effective Care, Department of Medicine, Montreal (Quebec) General Hospital (Dr Joseph).

Reprint requests to Department of Medicine, Centre Hospitalier de Verdun, 4000, Boul Lasalle, Verdun, Quebec, Canada H4G 2A3 (Dr Brophy).

Table 1.—Data From GUSTO, GISSI-2, and ISIS-3*

Trial	Agent	No. of Patients	No. (%) of Deaths	No. (%) of Nonfatal Strokes	Combined Deaths or Strokes
GUSTO†	SK	20 173	1473 (7.3)	101 (0.5)	1574 (7.8)
	t-PA	10 343	652 (6.3)	62 (0.6)	714 (6.9)
GISSI-2	SK	10 396	929 (8.9)	56 (0.5)	985 (9.5)
	t-PA	10 372	993 (9.6)	74 (0.7)	1067 (10.3)
ISIS-3	SK	13 780	1455 (10.6)	75 (0.5)	1596 (11.6)
	t-PA	13 746	1418 (10.3)	95 (0.7)	1513 (11.0)

*SK indicates streptokinase; and t-PA, tissue-type plasminogen activator.
†The 10 374 patients who received both SK and t-PA are not included here.

peated trials if the null hypothesis is exactly true. This raises two problems. First, it seems counterintuitive to base statistical inferences on events more extreme than those observed, since these events did not actually occur.[3] Second, one almost never believes that the null hypothesis of exact equivalence is true, and it is consequently usually more relevant to test for a range of equivalence. Such a test is very rarely carried out in practice. *P* values do not measure the true quantity of interest, namely, the probability that the null or alternative hypothesis is true. This contributes to the confusion between the information *P* values provide and the information that is more naturally desired. Therefore, it is not surprising that *P* values are often misinterpreted as the probability that the null hypothesis is true or that $1-P$ represents the probability that the alternative hypothesis is true. Classical statistical analysis does not directly or indirectly provide these probabilities.

Another inherent limitation of *P* values derives from their dependence on sample size. Basically, any difference, no matter how small, can reach statistical significance if the sample size is large enough. For example, an observed difference of only one tenth of a standard deviation will become statistically significant at the .05 level if each group in the trial includes at least 768 subjects and will be nonsignificant otherwise. On the other hand, it is well known that the low power accompanying small trials may lead to *P* values greater than .05 even when clinically meaningful effects are observed in the trial.[4]

All of these limitations of *P* values have prompted an increased use of confidence intervals. Many clinicians do not appreciate that a 95% confidence interval only means that with unlimited repeated experiments, 95% of all the confidence interval limits derived using similar procedures in different studies would contain the true parameter. While this may provide some comfort in the long run, little can be said about the likelihood that, for example, a given treatment is superior or that the true value of the parameter under current study lies in any particular interval.

The shortcomings of classical statistics may obscure the interpretation of even a well-designed and well-executed trial. For example, the recent GUSTO trial (Global Utilization of Streptokinase and Tissue Plasminogen Activator in Occluded Arteries) was a multicenter, randomized study comparing different thrombolytic regimens for the treatment of acute myocardial infarction.[5] This trial is of particular interest since there continues to be controversy over the clinical importance of any treatment differences. In addition, there have been other randomized trials involving large numbers of patients that examine the same question, namely, is tissue-type plasminogen activator (t-PA) superior to streptokinase (SK) in the treatment of acute myocardial infarction.[6,7] The question of therapeutic superiority is of considerable public health importance, since myocardial infarction is a frequent occurrence and t-PA is approximately 10 times more expensive than SK. While many critiques of the GUSTO trial have been published,[8-11] these have mostly centered on design issues and the interpretation of the clinical relevance of the observed mortality differences. This article raises further questions while highlighting some advantages of an alternative (Bayesian) statistical approach. Bayesian analysis has often been dismissed due to its "subjectivity" and because of computational difficulties. While Bayesian analysis can be computationally complex, computer algorithms now exist that make this hurdle more historical than contemporary. As will be seen, Bayesian subjectivity is an asset that can provide an ideal forum for debate, since prior beliefs, including clinical experience, must be formally specified, and one can directly observe how the beliefs are updated in the light of new data. This procedure permits the appreciation of the logic for various a posteriori opinions, which should tend to converge as data accumulate. This process is different from classical meta-analysis, which suffers from all the problems associated with *P* values and confidence intervals mentioned above and furthermore does not permit the incorporation of prior beliefs.[12]

METHODS

Model parameters such as the success rate of a given medical treatment are generally unknown, and therefore experiments are designed to provide information about their values. In virtually any well-designed experiment, more is known about these values after the experiment than before, although at least some information usually exists preexperimentally. A Bayesian statistical analysis is designed to represent this learning process.

The first step in any Bayesian analysis is to obtain a prior distribution over all model parameters. The prior distribution summarizes the preexperimental beliefs about the parameter values. This can be accomplished by using past data, if available, by drawing on expert knowledge, or by a combination of both. This step is nontrivial and can take considerable time and effort. Furthermore, many prior distributions are not unique; clinicians are free to summarize their beliefs into their own prior distribution. Because Bayesian methods can incorporate clinical opinion, they are often labeled "subjective." The experimental data are then used to update the prior distribution to a posterior distribution using Bayes' theorem. This is done through the likelihood function, which provides the probability of obtaining the observed data as a function of the unknown model parameter. This is analogous to using a likelihood ratio (sensitivity/[1−specificity]) to update background probabilities after observing results from a diagnostic test. The posterior distribution represents the postexperimental beliefs about the parameter values, given the new data and the previously stated prior distribution. The two main quantities of interest, namely, the probability that a given treatment is superior and the probability of a clinically meaningful effect, are both directly available from the posterior distribution. Unlike the standard approach, no references to data sets other than those observed are required, since all of the information contained in the data is summarized by the likelihood function.

No one prior distribution is likely to be sufficient to represent the diversity of clinical opinions that exists before a trial is carried out. Indeed, this diversity is usually a prerequisite for ethical randomization. Therefore, trial results should usually be reported starting from a range of prior distributions.[13] The corresponding set of posterior distributions

132 D

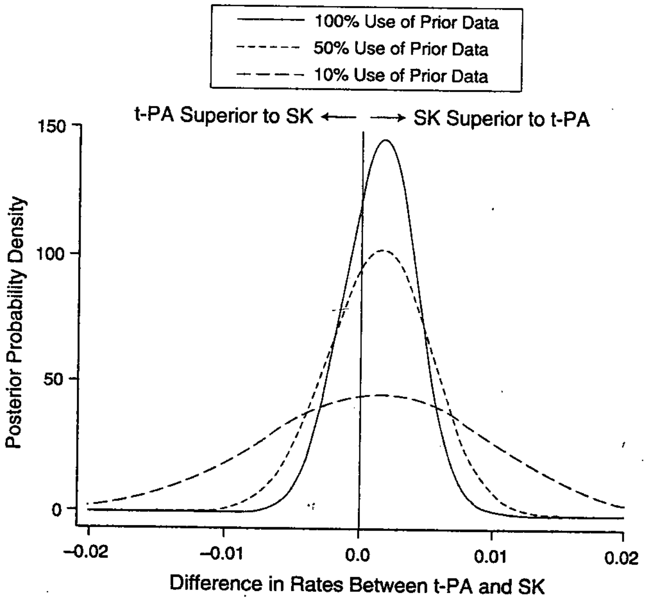


Figure 1.—Plot of the prior distributions for the difference in mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK) using weights of 100%, 50%, and 10% of the GISSI-2 and ISIS-3 data, representing a range in prior beliefs in the relevance of these trials to the GUSTO trial. The area under the curve between any two points on the x-axis is the posterior probability that the difference in mortality rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

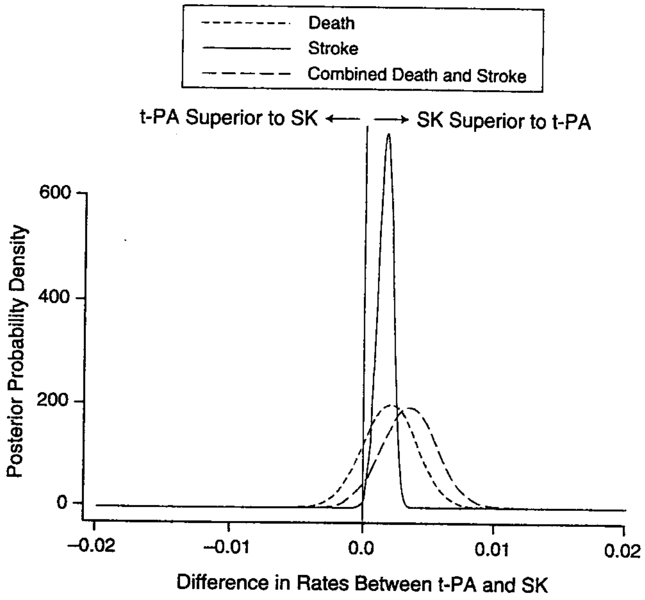


Figure 2.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial, with full prior use of data from the GISSI-2 and ISIS-3 trials. The area under the curve between any two points on the x-axis is the posterior probability that the difference in rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

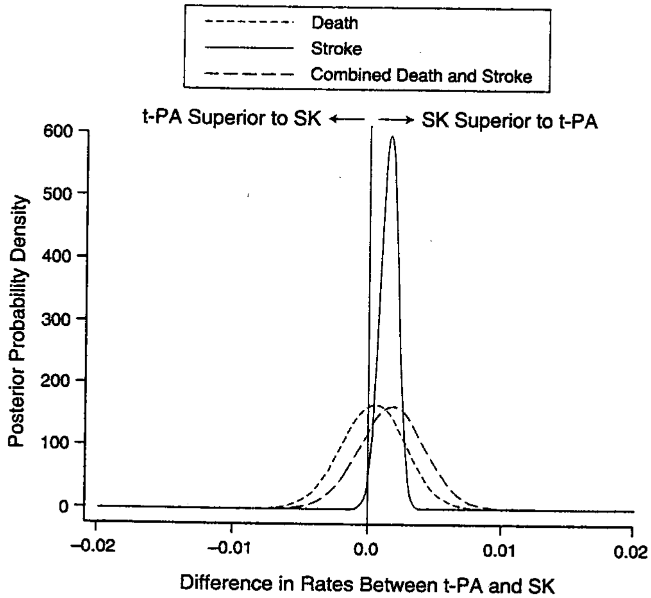


Figure 3.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial, with 50% prior use of data from the GISSI-2 and ISIS-3 trials. The area under the curve between any two points on the x-axis is the posterior probability that the difference in rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

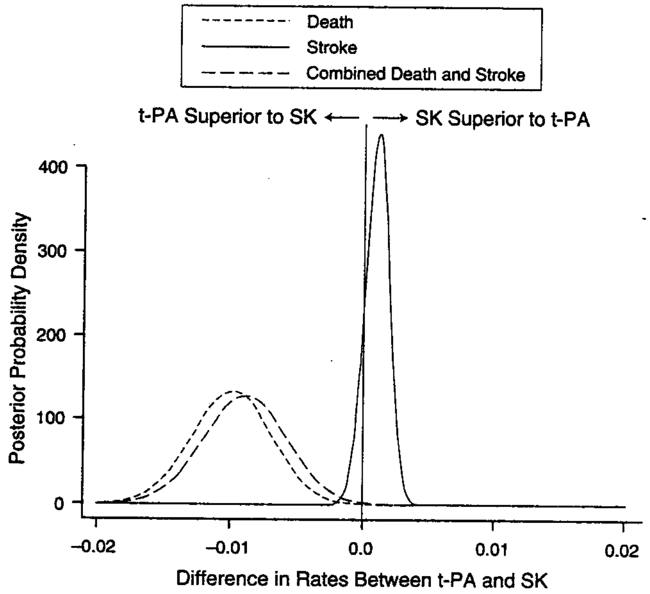


Figure 4.—Plot of the posterior distribution for the difference in mortality, nonfatal stroke, and combined stroke and mortality rates between tissue-type plasminogen activator (t-PA) and streptokinase (SK), using data from the GUSTO trial only. The area under the curve between any two points on the x-axis is the posterior probability that the difference in rates lies between those limits. Numbers to the right of zero indicate the superiority of SK, while those to the left of zero indicate the superiority of t-PA.

then summarizes the range of posttrial beliefs. If this latter set of distributions includes only a sufficiently narrow range of possible effects, conclusions could be drawn with which most clinicians should agree regardless of their initial opinions. Otherwise, the debate continues and further research is indicated.

These methods and their interpretation are illustrated below. Other studies[1,3,13,14] provide fuller descriptions of the use of Bayesian analysis in the con-

Table 2.—Probability of t-PA Superiority as a Function of Prior Belief in GISSI-2 and ISIS-3 Data After Consideration of the GUSTO Data*

Prior Belief in GISSI-2 and ISIS-3, %	Probability of t-PA Mortality Higher Than SK Mortality	Probability of t-PA Net Clinical Benefit Greater Than SK Benefit	Probability of t-PA Net Clinical Benefit Greater Than SK Benefit by at Least 1%
100	.17	.05	<.001
50	.44	.24	<.001
10	.98	.94	.03
0	.999	.998	.36

*See footnote to Table 1 for expansions of abbreviations. Net clinical benefit is the combined death and stroke rate.

text of clinical trials. In this study, posterior distributions for the difference in survival rates between groups of patients receiving two different thrombolytic regimens following acute myocardial infarction are derived and graphically displayed. (Mathematical equations used to derive the Figures are available from the authors on request.)

The GUSTO trial randomized 41 021 patients to four different thrombolytic strategies involving SK, t-PA, or a combination of the two for the treatment of acute myocardial infarction. Compared with SK, the strategy of "front-loaded" or "accelerated" t-PA showed a statistically significant lowered mortality (6.3% vs 7.3%, respectively; P=.001) and combined end point of 30-day mortality or disabling stroke (6.9% vs 7.8%, respectively; P<.006) (Table 1). The interpretation of a P value of .001 is that if the two agents had exactly equivalent mortality rates, then data as extreme as or more extreme than the observed mortality rates would occur once in every 1000 hypothetical repeated trials.

This well-executed clinical trial possesses many of the desirable attributes of a well-done study. The sample size was very large and was designed to have at least 80% power to detect a 15% reduction in mortality or an absolute decrease of 1% between experimental groups. This value has been (somewhat arbitrarily) defined by the GUSTO investigators as the clinically important difference between the two agents. Economic analyses that incorporate patient utilities and health care expenditures may be required to further investigate what difference is clinically meaningful. In this article, we will accept a 1% decrease as the clinically meaningful difference. Potential confounding and bias were minimized by the randomization process. Most clinicians would accept the frequentist analysis of this study as being conclusive (or almost conclusive) proof of the superiority of t-PA, that is, the mortality rate for t-PA was less than that for SK. But is this an adequate summary of the available evidence?

Two previous randomized clinical trials have directly compared SK with t-PA in 48 000 patients. The GISSI-2[6] trial (Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico) compared t-PA (alteplase) and SK both with and without subcutaneous heparin beginning 12 hours after the start of therapy. The 35-day total mortality and nonfatal stroke data are summarized in Table 1. The ISIS-3[7] trial (Third International Study of Infarct Survival) compared t-PA (duteplase) and SK both with and without subcutaneous heparin in a similar factorial design but began heparin 4 hours after the start of therapy. The 35-day mortality and morbidity data are also shown in Table 1.

Although all the trials were randomized with uniform entry criteria and drug dosages, reservations have been expressed about the relevance of any comparisons between these studies. The major sources of controversy are as follows:

- The t-PA used in ISIS-3 was of a slightly different form (although the clinical difference is not believed to be significant).
- Adjunctive therapy accompanying t-PA in GUSTO included more aggressive use of intravenous heparin.
- In GUSTO t-PA was administered in an accelerated fashion.

While there is an abundance of prior information comparing these two agents, there is little consensus as to which agent is superior. Clinicians may vary in their weighting of the importance of the similarities and differences between the trials. This only enhances the utility of a Bayesian analysis, because their uncertainty can be explicitly considered by employing a range of prior beliefs.[13,14]

Figure 1 shows the probability density for the difference in mortality between t-PA and SK as determined from the data of GISSI-2 and ISIS-3. (The area under the probability density curve between two given points on the x-axis represents the probability that a value will fall between the two points.) The difference in mortality rates between t-PA and SK appears along the x-axis (0.01=1% and so forth), and the height of the probability density for this difference is given by the y-axis. The mean of these curves is close to zero (0.0013), suggesting no difference between the two agents. Fully accepting the results of these two trials would suggest almost no possibility of t-PA's being clinically superior to SK (a decrease in the mortality rate with t-PA ≥1% is represented by the area to the left of −0.01, and this area is essentially zero in the case using 100% of the prior data). This leads to a very skeptical prior distribution as to the superiority of t-PA. On the other hand, a clinician who believes that the difference in trial protocols cannot be ignored might elect to only partially consider the earlier results. For example, one could arbitrarily treat the value of each observation in the previous trials as worth only 50% or even 10% of each observation in the GUSTO data. Prior distributions based on these weights also appear in Figure 1. A more extreme position would be that the trials are too dissimilar to be combined and that consequently all previous research should be ignored, thereby assuming that nothing is known about the potential difference in mortality between the two agents (in statistical parlance, this implies a noninformative or uniform prior distribution). Other prior distributions are also possible and are not necessarily derived by a weighting of previous data. Most of these would fall in between the above-mentioned extremes. As the belief in the utility of the prior studies decreases, so increases the possibility that t-PA is a clinically superior agent (widening of the curves and increasing area to the left of −0.01).

RESULTS

The data from Table 1 may be used to derive posterior distributions for stroke, death, and net clinical benefit (death and nonfatal stroke) using Bayes' theorem (the solved equation is available from the authors on request). Figure 2 considers the skeptical prior belief that assigns equal weight to each observation from GISSI-2, ISIS-3, and GUSTO and shows that the mean difference in mortality between t-PA and SK is 0.20% (0.002 in favor of SK), and the final (posterior) probability of t-PA's being superior to SK is only about 17% (area under the curve to the left of 0). Figure 2 also demonstrates that there are 0.15% more nonfatal strokes with t-PA and that the probability that the rate of nonfatal stroke is greater with t-PA exceeds 99.5% (the area to the left of the curve <.005). A similar interpretation of the combined curve suggests that the probability that t-PA is superior to SK is 5.1% with an almost zero probability of exceeding the clinically signficant difference of 1% (area to the left, on the

132 F

combined curve of 0 and −0.01, respectively).

Figure 3, which considers observations from the previous randomized clinical trials to have 50% the value of each observation in GUSTO (a more intermediate prior belief), shows that the probability that t-PA is superior to SK for mortality alone is about 44% (again refer to the area to the left of 0 for the appropriate curve). Further, accepting that a difference of 1% mortality is the minimum clinically significant value, the probability that t-PA is clinically superior remains negligible. The probability of increased stroke with t-PA remains high at almost 98%.

Finally, Figure 4 shows the scenario where all prior data from GISSI-2 and ISIS-3 are considered irrelevant and are ignored. In this case, t-PA is virtually certain to have a lower death rate than SK (99.95%), but the probability that t-PA exceeds the defined clinical superiority is only 48%. The probability of a net clinical benefit exceeding 1% is only 36%, and the probability of increased stroke with t-PA is 86%. The salient elements of Figures 2 through 4 are displayed in Table 2.

COMMENT

The current study demonstrates several advantages of a Bayesian analysis. The most apparent is that the analysis permits the direct answer as to the probability that t-PA is superior to SK. It also permits the calculation of the probability of clinical superiority. The answers, however, can vary since readers must each draw their own conclusions by selecting the posterior distribution that belongs to the prior distribution most closely matching their own initial personal beliefs. The GUSTO investigators suggested a minimum clinical superiority based on economic factors of one life saved per 100 patients treated, but Table 2 could be expanded to include any personalized prior distribution and clinical superiority cut point.

The Bayesian analysis presented herein suggests that restraint in accepting t-PA into routine clinical practice would be appropriate. The same conclusion was reached by Dr Diamond and colleagues,[15] who used a Bayesian point null hypothesis test. When one accepts only partial recognition (50%) of previous randomized clinical trials, the probability that t-PA is superior to SK for mortality or net clinical benefit is only 44% and 24%, respectively. The probability that either mortality or net clinical benefit would exceed clinical importance with the 50% assumption is much less than 1%. Even if one totally ignores all prior studies, the chance that t-PA would exceed the clinical superiority cut point for mortality and net clinical benefit is only 48% and 36%, respectively.

Neither *P* values nor the Bayesian analysis presented herein measures potential bias or confounding. The GUSTO trial was unblinded, which may lead to some degree of confounding. For example, while not reported in the original article, it appears that 9.5% of the t-PA group underwent coronary artery bypass surgery compared with 8.5% in the SK group. This difference may have contributed to the observed mortality differences. A Bayesian approach to adjustments for a wide variety of biases is described by Eddy et al.[12]

In assessing the public health impact of choosing a thrombolytic agent, the following seems clear. *P* values or confidence intervals from conventional statistical analysis are poor tools for formulating public health policy, even when there is a considerable amount of data from the best-designed randomized clinical trials. This is due to the shortcomings of standard significance tests in addressing clinically relevant questions and to the problems in their interpretation, especially across different sample sizes. Furthermore, classical analysis of clinical trials does not easily permit the synthesis of trial results with the range of clinicians' prior beliefs. This makes it difficult to evaluate the coherency of the conclusions and what clinical impact the conclusions should have. Bayesian analyses along the lines presented herein may help to overcome these problems, thereby raising the level of debate following publication of a clinical trial.

References

1. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Intern Med.* 1983;98:385-394.
2. Browner WS, Newman TB. Are all significant *P*-values created equal? the analogy between diagnostic tests and clinical research. *JAMA.* 1987;257:2459-2463.
3. Berger J, Berry D. Statistical analysis and the illusion of objectivity. *Am Scientist.* 1988;76:159-165.
4. Frieman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, type II error and sample size in the randomized control trial: survey of 71 'negative' trials. *N Engl J Med.* 1978;299:690-694.
5. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med.* 1993;329:673-682.
6. The International Study Group. In-hospital mortality and clinical course of 20,891 patients with suspected acute myocardial infarction randomised between alteplase and streptokinase with or without heparin. *Lancet.* 1990;336:71-75.
7. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group. ISIS-3: a randomised comparison of streptokinase vs tissue plasminogen activator vs anistreplase and of aspirin plus heparin vs aspirin alone among 4,299 cases of suspected acute myocardial infarction. *Lancet.* 1993;339:753-770.
8. Rapaport E. GUSTO: assessment of the preliminary results. *J Myocard Ischemia.* 1993;5:15-24.
9. Sleight P. Thrombolysis after GUSTO: a European perspective. *J Myocard Ischemia.* 1993;5:25-30.
10. Ridker PM, O'Donnell C, Marder VJ, Hennekens CH. Large-scale trials of thrombolytic therapy for acute myocardial infarction: GISSI-2, ISIS-3, and GUSTO-1. *Ann Intern Med.* 1993;119:530-532.
11. Ridker PM, O'Donnell C, Marder VJ, Hennekens CH. A response to 'holding GUSTO up to the light.' *Ann Intern Med.* 1994;120:882-884.
12. Eddy DM, Hasselblad V, Shachter R. *Meta-analysis by the Confidence Profile Method.* New York, NY: Academic Press; 1992.
13. Hughes M. Reporting Bayesian analyses of clinical trials. *Stat Med.* 1993;12:1651-1663.
14. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *J R Stat Soc A.* 1994;157:357-416.
15. Diamond GA, Denton TA, Forrester JS, Shah PK. Is tissue plasminogen really superior to streptokinase? *Circulation.* 1993;88:I-452. Abstract.

132 G

For Debate

The statistical basis of public policy: a paradigm shift is overdue

R J Lilford, D Braunholtz

University of Birmingham, Birmingham B16 9PA
R J Lilford, *professor of health services research*

Nuffield Institute for Health, Leeds University, Leeds LS2 9PL
D Braunholtz, *medical statistician*

Correspondence to: Professor Lilford.

BMJ 1996;313:603–7

The recent controversy over the increased risk of venous thrombosis with third generation oral contraceptives illustrates the public policy dilemma that can be created by relying on conventional statistical tests and estimates: case-control studies showed a significant increase in risk and forced a decision either to warn or not to warn. Conventional statistical tests are an improper basis for such decisions because they dichotomise results according to whether they are or are not significant and do not allow decision makers to take explicit account of additional evidence—for example, of biological plausibility or of biases in the studies. A Bayesian approach overcomes both these problems. A Bayesian analysis starts with a "prior" probability distribution for the value of interest (for example, a true relative risk)—based on previous knowledge—and adds the new evidence (via a model) to produce a "posterior" probability distribution. Because different experts will have different prior beliefs sensitivity analyses are important to assess the effects on the posterior distributions of these differences. Sensitivity analyses should also examine the effects of different assumptions about biases and about the model which links the data with the value of interest. One advantage of this method is that it allows such assumptions to be handled openly and explicitly. Data presented as a series of posterior probability distributions would be a much better guide to policy, reflecting the reality that degrees of belief are often continuous, not dichotomous, and often vary from one person to another in the face of inconclusive evidence.

Every five to 10 years a "pill scare" hits the headlines. Imagine that you are the chairperson of the Committee on Safety of Medicines. You have been sent the galley proofs of four case-control studies showing that the leading brands of oral contraceptive, which have been widely used for some five years, are associated with a doubling of the risk of venous thromboembolism. You are surprised; you seem to remember that these new brands contain an "improved" progesterone which has been shown to have no adverse effects on clotting factors—indeed the widespread acceptance of this treatment was predicated on the favourable metabolic effects of the new compound. A literature search and telephone call to local experts confirms your memory. You are aware that case-control studies are often biased. What do you do?

On the one hand you do not wish to over-react. After all, even if the newer brands do carry a higher risk of thrombosis, the risk arising from pregnancy is higher still. Thus widespread alarm may precipitate contraceptive withdrawal in mid-cycle and hence do more harm than good. On the other hand, if you fail to issue a statement advising the profession that a statistically significant doubling of the risk of deep vein thrombosis has been measured then you lay yourself (and others) open to public criticism when, sooner or later, reports of a serious medical mishap are brought to public attention. "Why did you not warn the public so that individuals could make an informed choice? After all, there was a 'statistically significant' doubling of thrombosis rates in the study."

The scenario painted here has an obvious similarity to the recent controversy surrounding oral contraceptives containing new third generation gestagens. Four case-control studies (one nested in a cohort study) have recently been reviewed by McPherson.[1] Taken together they show a statistically significant doubling in the risk of venous thromboembolism. We are not experts in this subject and do not want to add to this particular debate: we want to make a general point about the interpretation of new data in the context of a treatment (or prophylaxis) of which the clinical community has had considerable experience and about which other data exist.

Our thesis is that conventional statistical tests and estimates are an improper basis for public policy for two reasons. Firstly, they dichotomise results according to whether or not they are "significant," thereby tending to produce an off/on response by decision makers. Secondly, they do not take account of additional evidence (generated outside or within the index study) in an explicit way. Such evidence must then be handled implicitly, and this makes it much less useful in defending decisions. The statistically significant result seems "hard" and is explicit, while the notion that our conclusions should be tempered by knowledge of the biochemistry and plausible biases seems "soft" and that knowledge is handled in an implicit manner. Since the statistical analysis does not incorporate these additional factors, they cannot impact explicitly on the conclusions. The chairperson of the Committee on Safety of Medicines is placed on the defensive: she may be seen to be "explaining away" the observed effect if she does not act decisively in the direction predicated by the statistically significant result.

Confronting the difficulty: the Bayesian alternative

But is there another way to proceed: how else can statistics be used to guide policy on an issue of private and public concern? Clearly, if clear cut answers are available then an unambiguous official statement should follow. The effects of the sun's rays on skin cancer and of posture on sudden infant death may be examples where epidemiology has produced sufficiently clear cut answers to provoke specific recommendations. When the situation is less clear cut, however, as in the case of third generation oral contraceptives, conventional

132 H

Bayesian statistics

The key difference between Bayesian and conventional (or frequentist) statistics is the view of what probability is. Frequentists view probability as a relative frequency, or proportion. Thus the probability P of a fair coin landing heads up is 0.5 because in a long series of tosses it lands heads up half the time. Frequentists should not therefore estimate probabilities for one off events—like the probability of President Clinton winning a second term. Strictly, of course, all events are one off, but many events are similar enough to satisfy frequentists' requirements. Bayesians, on the other hand, view probability as a degree of personal belief. Personal belief changes as evidence (data) accrues, but no data at all are necessary. A Bayesian might judge the value of P to be close to 0.5, without the need for any previous experience of coin tossing—on the basis of the physics involved. In fact he or she would want to give a probability distribution for the true value of P. This would be a prior distribution for P, which could then be updated via coin tossing (by means of Bayes's law) to produce a posterior distribution of probabilities.

Bayes's law in itself is uncontentious and is used by frequentists as well as Bayesians, but frequentists use it in much more restricted circumstances. The classic examples are Mendelian genetics and computerised diagnosis, such as that popularised in the UK by the late professor Tim deDombal. Bayes's law as used by Bayesians simply states that the posterior probability distribution is formed by weighting the prior probability distribution by the likelihood.

One practical advantage of the Bayesian approach is that it provides probability distributions for parameters—which is exactly what is needed to inform decisions. As we show in this paper, it also makes the synthesis of new data, and other kinds of evidence, relatively straightforward. Frequentists would argue that the disadvantage is that prior beliefs, being personal, can vary—and conclusions may therefore differ from person to person. Bayesians would respond that that is what real life is like. Also, by carefully doing sensitivity analyses, researchers can assess how robust conclusions are to changes in prior probability distributions, or indeed to changes in the model used to create the likelihood.

Other than in very simple cases (such as that presented here) calculating the posterior probability distribution becomes impossible analytically, and it has to be approximated—for instance, using "Monte-Carlo" methods on computers. This involves generating a large, random, sample from the posterior probability distribution (each number generated may involve substantial computations), and the properties of the posterior probability distribution are "discovered" by analysing this sample.

The advent of fast, cheap computers now makes this feasible for almost anyone, and programs such as BUGS (available from ftp.mrc-bsu.cam.ac.uk) are making it easier to do.

statistics may drive decision makers into a corner (resulting in either false reassurance or excessive caution) and produce sudden, large (and hence potentially harmful) changes in prescribing. The problem does not lie with any of the individual decision makers, but with the very philosophical basis of scientific inference. We propose that conventional statistics should not be used in such cases and that the Bayesian approach is both epistemologically and practically superior.[2]

Here we start with prior belief, which is measured and made explicit. We then incorporate the new data but in so doing we may adjust for the likely extent of bias. We then combine the prior with the adjusted data to obtain a "posterior" probability distribution, using the mathematical theorem associated with the name of the eighteenth century clergyman, Thomas Bayes (see box). Lastly, we carry out a sensitivity analysis, to see what effects different prior beliefs and different assumptions about possible bias might have. Given the data, almost everyone will now have a stronger belief that third generation pills cause clots in the venous system than they had before, but everybody does not have to believe the same thing. Even without considering possible beneficial effects on the risk of heart attack, the health care system can respond incrementally and not precipitate a large scale shift in prescribing practice. There would be little reason for a scare story causing a surge in demand for consultations and in unwanted pregnancies. The principles of Bayesian inference are described in more detail in the box.

Bayesian inference: how it works

We give a worked example, based on McPherson's summary, which shows an odds ratio of 2 for the risk of deep venous thrombosis when the third generation pills were compared with others. Since the risks are small, we can think of the odds ratio as a relative risk. The 95% confidence interval ranges from a relative risk of 1.4 to 2.7. Clearly the 95% confidence interval excludes 1 and the results are therefore significant at the usual $P<0.05$ level. P here is the proportion of times that an effect of this size (or greater) would be measured in an infinite repetition of studies if the true effect was 1—that is, both third generation and older pills were associated with the same risk

However, decision makers want to know the probabilities of thrombosis for the next patient who is eligible for either treatment. A decision maker might ask: "What is the probability that the third generation pills increase the risk when compared to the others; what is the probability that they at least double the risk—as measured in the case-control study; and what is the 'median estimate' (as likely to be too small as too large)?"

The calculations require a prior probability distribution for the true effect. We could obtain this by measuring the collective prior belief of experts. We could contact, say, 25 randomly selected members of the Faculty of Family Planning, probably before they knew about the new data. We would interrogate them to see what their thoughts were on: (*a*) the best estimate of the true relative risk—the effect of the third generation pills on the risk of clotting when compared with the standard pills; (*b*) what values they thought were unlikely for the true relative risk—such that an effect of that size or more extreme would have a chance of being true of less than 0.025. The answers are those that respondents would give if they were forced to set odds and accept any bets while wishing to minimise their losses. For example, they might set odds of 19:1 that the true relative risk would lie within the interval specified at (*b*) above. Imagine that our average respondent thinks that the true relative risk is as likely to be above as below 0.8 (corresponding to a 20 percentage point reduction in risk (relative risk=0.8)) and that a relative risk of 1.6 or greater, or of 0.40 or less, are unlikely to be true. In that case, their prior distribution of probability estimates could be represented on a log relative risk scale as a normal curve— prior distribution 1 in fig 1.

Bayes's theorem allows us to update this prior distribution to take account of McPherson's data, which are converted into a likelihood—likelihood A in fig 2. This updating of the prior distribution by the likelihood would give us the posterior distribution of probabilities referred to as posterior 1A in fig 1. The middle of the posterior distribution corresponds to a relative risk of about 1.69 and the 95% interval (now referred to as a *credible* interval rather than a confidence interval) for the relative risk ranges from 1.3 to 2.3. If asked to state the most likely effect an observer with prior 1 would give a relative risk close to 1 [illegible] of the mathematical point made in the legend to fig 1). For the mathematically minded, likelihood is discussed in more detail in the box below.

Taking into account different beliefs and likely bias: sensitivity analysis

The above figures represent the probabilities for an observer who agrees with the prior distribution of probabilities. We discussed these prior probability distributions with two eminent Leeds gynaecologists with an interest in family planning. Dr Nicholas Johnson agreed with these probability estimates and hence with the posterior probability distributions. Professor James Drife, however, was more sceptical: he was in absolute equipoise[3] before the new data—that is, he thought it equally likely that the third generation or standard oral contraceptives had a higher risk of causing deep vein thrombosis. However, like Johnson, his prior probability distribution was vague, admitting of an equally wide range of plausible values, with a 95% probability that the true relative risk was between 0.5 and 2.0 (curve prior 2 in fig 1). For Drife, the middle relative risk, when both the data and prior belief are taken into account, is 1.76 and the 95% credible interval extends from 1.3 to 2.4—posterior 2A in fig 1. The comparison of Johnson (who was cautiously enthusiastic to start with), Drife (who was sceptical), and yet other experts who may hold more extreme views constitutes a sensitivity analysis.

Sensitivity analysis can be extended to take into account evidence that case-control and other observational studies are often biased and that in this particular case we have reasons to suspect that the measured effect has been overestimated.

Firstly, we could suppose that the particular design and implementation of the studies contributing to McPherson's summary may result in a bias but that this bias is as likely to be positive as negative. We could further suppose that the distribution of this bias was normal on a log relative risk scale, with a standard deviation (SD) of 0.2624 (corresponding to a multiplying, or dividing, factor of 1.3 on the relative risk scale) so that the biased relative risk being estimated from McPherson's summary would be in the range of 60% to 167% of the true relative risk, with probability 0.95. This weakening of the evidence provided by the data results in likelihood B (fig 2) and in a posterior probability distribution closer to the prior distribution, as illustrated in fig 3. Posterior 2A is as in fig 1 (no bias), but posterior 1B and posterior 2B (from Johnson and Drife's prior probability distributions respectively) assume a bias in the included studies distributed as just described.

Secondly, however, it appear that non-randomised studies typically overestimate tr atment effects by about 30%,[4 5] and in this instance we have reason to suspect an overestimate. Firstly, third generation pills may have been given preferentially to higher risk women, and it is never possible to be certain that this has been fully accounted for by statistical adjustment.[6] Secondly, more "modern" general practitioners may both preferentially prescribe newer brands of pill and be especially vigilant in investigating symptoms which could result from venous thromboembolism. Thirdly, women using oral contraceptives which have been in use for a long time are biased with respect to those on newer brands, because many of those with venous thromboembolism (which typically occurs within a few months of starting the pill) will have been screened out—the so called "healthy user effect."[7 8] If we assume a median bias of 30%, given the above, and make no other new assumptions, then the biased relative risk estimated from the summary would be in the range 78% to 217% of the true relative risk, with probability 0.95. The evidence from the data is thus both weakened and shifted—see likelihood C in fig 2. The resulting posterior probability distributions are shown in fig 4, where posteriors 1C and 2C were derived from Johnson's and Drife's prior distributions respectively. The middle of Drife's posterior probability distribution now corresponds to a relative risk of 1.27, while for Johnson a true relative risk of above or below a central value of only 1.16 is equally likely. The probabilities that the relative risks of venous thrombosis are not increased at all with the third generation pills are 15% for Drife and 27% for Johnson. A relative risk of 1.27, calculated on the basis of Drife's original prior probability distribution (which was both equipoised and fairly vague), the data, and (arguably) modest assumptions of bias, translate into 0.4 to 0.8 additional cases of venous thromboembolisms per 10 000 women years (assuming a background risk of between 1.5 and 3 venous thromboembolisms per 10 000 women years on the previous generation of pills).

Manipulation or simply recognising reality?

Some people will feel very uneasy about these and other adjustments in a sensitivity analysis: thejudgmental manipulation of "real" figures may seem wrong. Wrong that is, until we examine the alternative, which is uncritically to accept data which we suspect to be less reliable than, say, the results of a randomised controlled trial. If there is reason to suspect systematic bias then it seems inappropriate not to allow for this in the analysis.[9 10] In this case not only is there empirical

Likelihood

When trying to understand the implications of a dataset researchers usually focus on a few parameters of special interest, which in some way summarise the interesting facets of the data. In this case the parameter of interest is the relative risk. Note that this is not directly observable in the data, but is an intangible idea that we find useful.

Parameters are linked to the data via a model, which describes the sort of data associated with particular values of the parameters. In this case the model we have assumed specifies that the probability distribution for the "observed" log relative risk will be normal with a mean of log (true relative risk) and a known standard deviation. In fact the standard deviation really depends on the sample size and the value of the true relative risk, but in our simple analysis we estimate the standard deviation from the data and then pretend we know it. Of course we have only one dataset, and we do not know the true parameter (relative risk) values. We consider all possible true parameter values, and for each calculate the probability of getting the data actually obtained. These probabilities can be plotted on a graph, and, when thought of as providing information on the likely true value of the parameter given the data, this plot is called the likelihood. Bayesians adhere to the intuitively attractive likelihood principle, which states that information arising from studies or experiments should be based only on the actual data observed. Frequentists often find themselves in conflict with this—for instance, when calculating P values, which take into account the probability of observations more extreme than the actual observations. However, in the case of the normal distribution conventional methods in effect use the likelihood to calculate confidence intervals, and we have used this in converting McPherson's summary into a likelihood: the likelihood is normal on the log (relative risk) scale, centred around log (2.0), and with a standard deviation such that log (2.7) – log (1.4) is 2 × 1.96 SD.

132 J

evidence that observational studies in general may be biased; there are plausible reasons to suspect bias in a particular direction. Thus any bias would be replicated across studies if the confounding factor was typical of the "treatment" in question. An advantage of explicit manipulation of the data, before statistical analysis, is that the process is transparent and hence open to challenge and recalculation on the basis of different assumptions.

Data presented as a series of posterior probability distributions (each based on a respective prior probability distribution and assumption of likely bias) would be a much better guide to policy than results analysed in the conventional way. They would reflect the reality that degrees of belief (*a*) are continuous or incremental, but not dichotomous, and (*b*) vary (quite properly) from one person to another in the face of inconclusive evidence.

On the above scenarios some clinicians might change prescribing habits, while others would be "sensitised" (have a new, more cautious, prior distribution) against the day when yet more data may become available. Women themselves could see that evidence regarding venous thromboembolism was moving against the new pills but would not be alarmed by the notion that harm was proved by "statistics." They would understand the new data (correctly) as merely one more piece of evidence in a complex array. This would encourage women to derive their own estimate of likely risk in consultation with their clinician and make any trade off required by perceptions of countervailing benefit.[11]

In the case of some of the third generation pills there is reason to believe that the risk of heart attack is reduced, in comparison with earlier brands. The newer pills have more favourable effects on blood fats than their second generation cousins. On the basis of this information alone, many rational observers may have formed a prior probability distribution which, while vague, was shifted in the direction of net benefit—that is, many may have had a prior distribution with respect to heart attack similar to that which Johnson had with respect to venous thromboembolism. One of the studies quoted by McPherson does, in fact, give results for heart attacks: the odds ratio is 0.36, suggesting that the risk is indeed lower with third generation pills, but the confidence interval is wide (0.1 to 1.2).[12] Thus, although the latter results are not statistically significant, perhaps because the number of adverse events is still small, they could be used to update a Bayesian prior probability distribution. With any reasonable prior belief and assumptions about bias the posterior probability distribution will be centred on a large reduction in relative risk, but will be widely spread. The uncertainty (corresponding to non-significance in frequentist terms) is, however, no reason to ignore the effect of the newer pills on heart attacks, since that is essentially to assume with complete certainty that there is no effect.

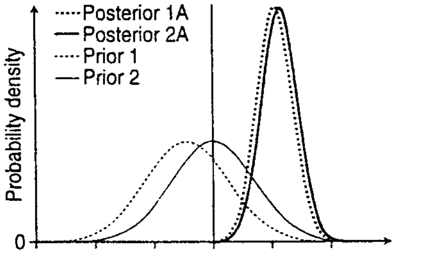


Fig 1—*Probability distributions, on a log (relative risk) scale, of relative risk of venous thromboembolism in third generation contraceptive pills compared with second generation pills. All prior distributions and likelihoods (and hence, owing to the mathematics of Bayes's theorem, posterior distributions) are assumed to be normally distributed on the log scale.*

Priors 1 and 2 are Johnson and Drife's respectively. Both are fairly wide, indicating considerable doubts about the value of the true relative risk. Drife's prior is centred on log(1.0) as (before learning of the new case-control study data) he believed that third generation pills were as likely to be better as to be worse than second generation pills. Johnson was more optimistic that the new pills would have a lower risk of venous thromboembolism, his prior distributions being centred on log(0.80). If McPherson's summary of the various studies is taken at face value (likelihood A in fig 2) and is used to update the experts' prior distributions via Bayes's theorem, posterior distributions 1A and 2A result. These are much narrower than the prior distributions, indicating less doubt about the value of the true relative risk. The data (with an observed odds ratio of 2.0) has influenced the posterior distributions more than the rather vague prior distributions, with the result that they are centred on log(1.69) and log(1.76) respectively, and the probability of the true relative risk being greater than 1 is more than 0.999 in both cases.

Note:The most probable value for the true log (relative risk) is not equal to log (most probable value of relative risk)—that is, the position of the highest points of the probability distributions drawn on log (relative risk) and on relative risk scales do not correspond. For instance the most probable value of log (relative risk) for prior 2 is log(1.0)=0, but the most probable value of relative risk for prior 2 is 0.89, not 1.0. This is because the whole of the negative log (relative risk) axis (and its probability) is "squashed" into the interval (0,1) on the relative risk scale, while the positive log relative risk axis is increasingly stretched out. The centres of distributions are not affected by this problem and have for this reason been used in this paper.

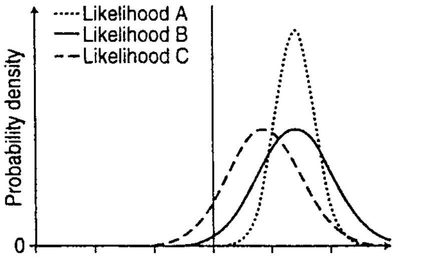


Fig 2—*Posterior distributions are calculated by "weighting" the prior distributions by the data likelihood. The likelihood can be calculated as the probability of the data given varying true values of the parameter (in this case log (relative risk)) but is viewed as the likelihood of the various parameter values given the data. The likelihoods shown here correspond to McPherson's summary of the various studies (relative risk of 2.0, 95% confidence interval 1.4 to 2.7): (A) taken at face value; (B) assuming the summarised data may be biased, with bias drawn at random from a normal distribution on a log (relative risk) scale with mean zero, and SD of log(1.3); (C) as (B) but with mean log bias log(1.3). Clearly an assumption that the data may be biased reduces the information from the data, and if the mean bias is thought to be non-zero, the information is also shifted accordingly.*

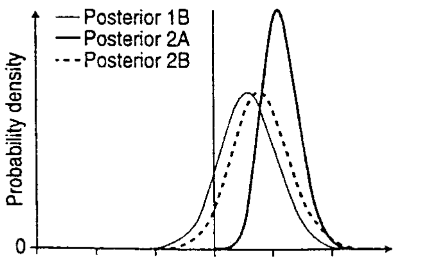


Fig 3—*Posterior 2A is as in fig 1, deriving from Drife's prior distribution and the data taken at face value (likelihood A). Posterior 2B again derives from Drife's prior distribution, this time weighted by likelihood B. The information conveyed by the data is thus much reduced, and the posterior distribution correspondingly wider, and closer to the prior distribution. It is centred on log(1.48), and the probability that the true relative risk is less than 1—that is, that the new pills have reduced risk of venous thromboembolism—is now much increased (though still small) at 0.048. Posterior distribution 1B derives from Johnson's prior distribution and likelihood B. As would be expected, it produces a much higher probability (0.11) that the true relative risk is less than 1.*

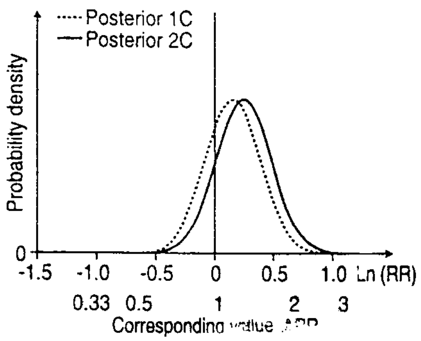


Fig 4—*Both posterior distributions derive from the data summary adjusted by the assumption that the studies which produced the data may have been biased, with the summary bias (on log relative risk scale) sampled from a normal distribution with mean log(1.3) and SD log(1.3)—that is, likelihood C. The information conveyed by the data is thus reduced and shifted. Johnson and Drife's prior distributions, when weighted by likelihood C, result in posterior distributions 1C and 2C respectively centred on log(1.16) and log(1.27). The probabilities that the true relative risk is less than 1—that is, that the new pills actually reduce risk of venous thromboembolism—are further increased, to 0.27 and 0.15, and are now far from negligible in both cases.*

132k

Thomas Bayes

Bayes was a member of the first secure generation of English religious non-conformists. His father, Joshua Bayes FRS, was a respected theologian of dissent; he was also one of the group of six ministers who were the first to be publicly ordained as non-conformists. Privately educated, Bayes became his father's assistant at the presbytery in Holborn, London; his mature life was spent as minister at the chapel in Tunbridge Wells. Despite his provincial circumstances, he was a wealthy bachelor with many friends. The Royal Society of London elected him a fellow in 1742. He wrote little: *Divine Benevolence* (1731) and *Introduction to the Doctrine of Fluxions* (1736) are the only works known to have been published during his lifetime. The latter is a response to Bishop Berkeley's *Analyst*, a stinging attack on the logical foundations of Newton's calculus; Bayes' reply was perhaps the soundest retort to Berkeley then available.

Bayes is remembered for his brief "Essay towards solving a problem in the doctrine of chances" (1763), the first attempt to establish a method to calculate a probability distribution (the probabilities of different events occurring) given a set of data. In so doing he laid the foundations for statistical inference.

Before Bayes there was some understanding of how to reject statistical hypotheses in the light of data, but no one had shown how to measure the probability of statistical hypotheses in the light of data. Bayes began his solution of the problem by noting that sometimes the probability of a statistical hypothesis is given before any particular events are observed; he then showed how to compute the probability of the hypothesis after some observations are made. Bayes was himself too modest to claim that he had solved the basis for the whole of statistical inference, and it was left to Richard Price to submit his work to the Royal Society. However, the great Laplace had no qualms about Bayes's argument; his enormous influence made Bayes's ideas almost unchallengeable until George Boole protested in his *Laws of Thought* (1854). Since then Bayes's technique has been a constant subject of controversy. The controversy relates to deriving the probability of statistical hypotheses (prior probability distributions), especially before any data of the type we want to analyse have been observed.

In *Foundations of Statistics* Leonard J Savage interprets probability in a personal way, as reflecting a person's personal degree of belief; hence, a prior probability distribution is a person's belief before the new observations become available, and a posterior probability distribution is a person's belief after the observations are made available. In the past 10 years or so there has been a sharp revival of interest in Bayes's work, especially its application to medical problems. Researchers in the UK have been in the forefront of this resurgence: they include David Spiegelhalter at the MRC Biostatistics Unit Cambridge, Adrian Smith at Imperial College, London, Deborah Ashby at the University of Liverpool, and, from a philosophical perspective, Peter Urbach at the London School of Economics.

A reasonable approach to answer the relevant question—Are third generation pills preferable to second generation pills?—needs to deal in absolute risks and explicit "costs" to women. The absolute risk of heart attack in users of second generation pills is even lower than that of venous thromboembolism,[12] but a heart attack is typically more serious—so the overall mortality and morbidity due to both may be similar. The combined posterior distribution for the difference between third and second generation pills in total mortality may thus be quite spread out, with a substantial proportion of the area—that is, probability—on both sides of the origin. A summary would conclude that, although it looks fairly probable that venous thromboembolism occurs somewhat more frequently with third generation pills, there is still considerable doubt as to which is safer overall. Such a statement would not have been likely to initiate large scale changes in prescribing, except for women with risk factors for venous thromboembolism. The possibility of collecting more useful data on the safety of third generation pills would not have been all but removed, as McPherson suggests it has been in his editorial.[1] The importance of collecting more data on the safety of third generation pills—to tighten up the posterior distributions—would be emphasised.

Acknowledging imperfections: a better basis for public policy

Bayesian techniques allow all our current knowledge to be explicitly represented and synthesised with new data. If there is little knowledge this is reflected in vague prior probability distributions. If explicit costs and benefits can be assigned to outcomes decision analysis[11] can then be used to trade off the best available estimates of benefit and harm, incorporating preferences for health in the short over the long term. Conventional statistics do not include all the evidence within the calculations. They therefore dichotomise results and tend to result in sensationalism. Faced with data presented in Bayesian and decision analysis terms journalists would have to communicate with the public in a more sophisticated way to show how probabilities vary according to different interpretations of the "starting" information and that the final decision can take account of personal trade offs. Practical actions are based on (often unrecognised) philosophical assumptions. A move from standard to Bayesian statistics would represent a fundamental change in how we think about knowledge and this in turn would affect policy making.

Health issues are now much more complex and the amount of disparate evidence that impacts on belief has increased. Only the Bayesian approach can do justice to all this information and provide the probabilistic basis for action when the results of a particular type of study have not (yet) reached statistical significance or, indeed, for not acting when they have. Sheldon and Smith have advocated this method in the context of environmental effects on health,[13] and a change in approach is overdue in this and other areas of public policy.

We thank Professor Zephne Van Der Spuy and Dr Victoria Lilford, whose dinner party conversation provided the inspiration for this article, and Professor James Drife and Dr Nicholas Johnson for the helpful discussions alluded to in the text.

1 McPherson K. Third generation oral contraception and venous thromboembolism. *BMJ* 1995;312:68-9.

2 Spiegelhalter D, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *Journal of the Royal Statistical Society* 1994;157:357-416.

3 Lilford R J, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;88:552-9.

4 Schulz, K F, Chalmers I, Grimes, D A, Altman, D G. Assessing the quality of randomization from report of controlled trials published in obstetrics and gynaecology journals. *JAMA* 1994;272:125-8.

5 Sacks H, Chalmers T, Smith H. Randomised versus historical controls for clinical trials. *Am J Med* 1982;72:233-40.

6 Buring JE, Glynn RJ, Hennekens CH. Calcium channel blockers and myocardial infarction: a hypothesis formulated but not yet tested. *JAMA* 1995;274:654-5.

7 Reijnen HBM, Atsma WJ. Risk is highest in the first months of use. *BMJ* 1995;311:1637.

8 Arrighi HM, Hertz-Picciotto. The evolving concept of the healthy worker survivor effect. *Epidemiology* 1994;5:189-96.

9 Ashby D, Hutton JL, McGee MA. Simple Bayesian analysis for case-control studies in cancer epidemiology. *The Statistician* 1993;42:385-97.

10 Eddy DM, Hasselbald V, Shachter R. A Bayesian method for synthesising evidence: the confidence profile method. *International Journal of Health Technology Assessment* 1990;6:31-55.

11 Lilford R J, Thornton J. Decision logic in medical practice. *J R Coll Phys* 1992;26:1-20.

12 Lewis MA, Spitzer WO, Heinemann AJ, MacRae KD, Bruppacher r, Thorogood M. Third generation oral contraceptives and risk of myocardial infarction: an international case-control study. *BMJ* 1996;312:88-90

13 Sheldon TA, Smith D. Assessing the health effects of waste disposal sites issues in risk analysis and some Bayesian conclusions. In: Clark M, Smith D, Blowers A. Eds. *Waste location*. London: Routledge, 1990.

(Accepted 17 June 1996)

132 L

5.17 A selective college would like to have an entering class of 1200 students. Because not all students who are offered admission accept, the college admits more than 1200 students. Past experience shows that about 70% of the students admitted will accept. The college decides to admit 1500 students. Assuming that students make their decisions independently, the number who accept has the $B(1500, 0.7)$ distribution. If this number is less than 1200, the college will admit students from its waiting list.

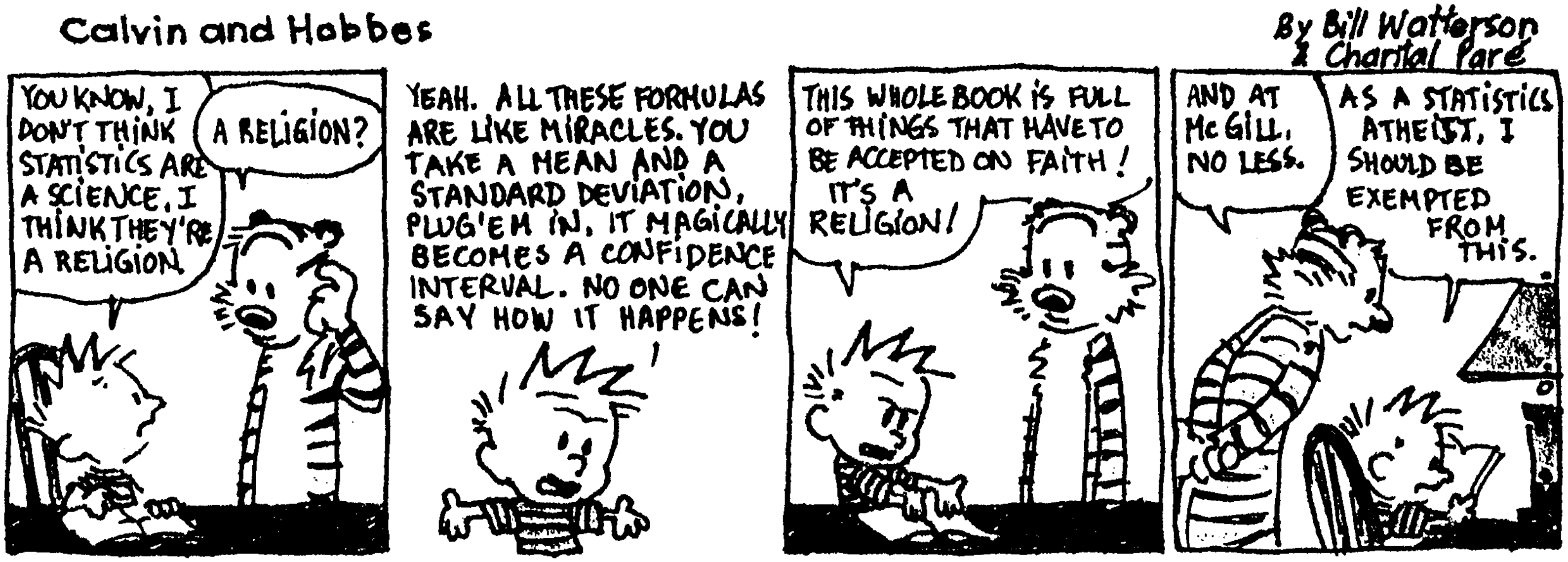
(a) What are the mean and the standard deviation of the number X of students who accept?

(b) Use the normal approximation to find the probability that at least 1000 students accept.

(c) The college does not want more than 1200 students. What is the probability that more than 1200 will accept?

(d) If the college decides to increase the number of admission offers to 1700, what is the probability that more than 1200 will accept?

Inference for Proportions


Calvin and Hobbes
By Bill Watterson & Chantal Paré
YOU KNOW, I DON'T THINK STATISTICS ARE A SCIENCE, I THINK THEY'RE A RELIGION
A RELIGION?
YEAH. ALL THESE FORMULAS ARE LIKE MIRACLES. YOU TAKE A MEAN AND A STANDARD DEVIATION, PLUG'EM IN, IT MAGICALLY BECOMES A CONFIDENCE INTERVAL. NO ONE CAN SAY HOW IT HAPPENS!
THIS WHOLE BOOK IS FULL OF THINGS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND AT McGILL, NO LESS.
AS A STATISTICS ATHEIST, I SHOULD BE EXEMPTED FROM THIS.

Inference for Proportions

There is an analogy between inference for proportions and that already covered for means:

	MEANS	PROPORTIONS
DATA	$\{x_1, x_2, \ldots, x_n\}$	$\{x_1, x_2, \ldots, x_n\} = \{0, 1, \ldots, 1\}$
ESTIMATOR	$\overline{x} = \frac{\sum_{i=1}^n x_i}{n}$	$\hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\text{\# of 1's}}{n}$
SD	$sd = \sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1}}$	$sd = \sqrt{\hat{p}(1-\hat{p})}$
CI	$\overline{x} \pm 1.96 \times \frac{\text{SD}}{\sqrt{n}}$	$\hat{p} \pm 1.96 \times \frac{\text{SD}}{\sqrt{n}}$

Similar analogies hold for other aspects of inferences for proportions, including formulæ for testing, confidence intervals for the difference between two proportions, etc.

The confidence interval for a binomial proportion listed on the previous page is based on the Normal approximation to the binomial distribution. Exact confidence intervals also exist, but are difficult to calculate. Tables and charts have appeared in the literature that list the resulting confidence interval for a given binomial proportion, depending on the sample size. These tools are especially useful for small sample sizes, or for proportions near 0 or 1, where the Normal approximation is less accurate. Examples of such tables and charts appear on the next three pages. We can compare the results found there with those from a Normal approximation to the binomial:

X	N	Exact CI	Normal Aprox. CI
4	10	(0.12, 0.74)	(0.09, 0.71)
8	20	(0.19, 0.64)	(0.18, 0.62)
40	100	(0.31, 0.51)	(0.30, 0.50)
400	1000	(0.37, 0.43)	(0.37, 0.43)

From! Statistical Tables.

136

F.J. Rohlf & R.R. Sokal

(Freeman, 1981)

TABLE 23 Confidence limits for percentages

Y	Confidence coefficients	n = 5	10	15	20	25	30
0	95	0.00 - 45.07	0.00 - 25.89	0.00 - 18.10	0.00 - 13.91	0.00 - 11.29	0.00 - 9.50
		0.00	0.00	0.00	0.00	0.00	0.00
	99	0.00 - 60.19	0.00 - 36.90	0.00 - 26.44	0.00 - 20.57	0.00 - 16.82	0.00 - 14.23
1	95	0.51 - 71.60	0.25 - 44.50	0.17 - 32.00	0.13 - 24.85	0.10 - 20.36	0.08 - 17.23
		20.00	10.00	6.67	5.00	4.00	3.33
	99	0.10 - 81.40	0.05 - 54.4	0.03 - 40.27	0.02 - 31.70	0.02 - 26.24	0.02 - 22.33
2	95	5.28 - 85.34	2.52 - 55.60	1.66 - 40.49	1.24 - 31.70	0.98 - 26.05	0.82 - 22.09
		40.00	20.00	13.33	10.00	8.00	6.67
	99	2.28 - 91.72	1.08 - 64.80	0.71 - 48.71	0.53 - 38.70	0.42 - 32.08	0.35 - 27.35
3	95		6.67 - 65.2	4.33 - 48.07	3.21 - 37.93	2.55 - 31.24	2.11 - 26.53
			30.00	20.00	15.00	12.00	10.00
	99		3.70 - 73.50	2.39 - 56.07	1.77 - 45.05	1.40 - 37.48	1.16 - 32.03
4	95		12.20 - 73.80	7.80 - 55.14	5.75 - 43.65	4.55 - 36.10	3.77 - 30.74
			40.00	26.67	20.00	16.00	13.33
	99		7.68 - 80.91	4.88 - 62.78	3.58 - 50.65	2.83 - 42.41	2.34 - 36.39
5	95		18.70 - 81.30	11.85 - 61.62	8.68 - 49.13	6.84 - 40.72	5.64 - 34.74
			50.00	33.33	25.00	20.00	16.67
	99		12.80 - 87.20	8.03 - 68.89	5.85 - 56.05	4.60 - 47.00	3.79 - 40.44
6	95			16.33 - 67.74	11.90 - 54.30	9.35 - 45.14	7.70 - 38.56
				40.00	30.00	24.00	20.00
	99			11.67 - 74.40	8.45 - 60.95	6.62 - 51.38	5.43 - 44.26
7	95			21.29 - 73.38	15.38 - 59.20	12.06 - 49.38	9.92 - 42.29
				46.67	35.00	28.00	23.33
	99			15.87 - 79.54	11.40 - 65.70	8.90 - 55.56	7.29 - 48.01
8	95				19.10 - 63.95	14.96 - 53.50	12.29 - 45.89
					40.00	32.00	26.67
	99				14.60 - 70.10	11.36 - 59.54	9.30 - 51.58
9	95				23.05 - 68.48	17.97 - 57.48	14.73 - 49.40
					45.00	36.00	30.00
	99				18.08 - 74.30	14.01 - 63.36	11.43 - 55.00
10	95				27.20 - 72.80	21.12 - 61.32	17.29 - 52.80
					50.00	40.00	33.33
	99				21.75 - 78.25	16.80 - 67.04	13.69 - 58.35
11	95					24.41 - 65.06	19.93 - 56.13
						44.00	36.67
	99					19.75 - 70.55	16.06 - 61.57
12	95					27.81 - 68.69	22.66 - 59.39
						48.00	40.00
	99					22.84 - 73.93	18.50 - 64.69
13	95						25.46 - 62.56
							43.33
	99						21.07 - 67.72
14	95						28.35 - 65.66
							46.67
	99						23.73 - 70.66
15	95						31.30 - 68.70
							50.00
	99						26.47 - 73.53

TABLE 23 Confidence limits for percentages

%	Confidence coefficients	n				
		50	100	200	500	1000
0	95	.00- 5.82	.00- 2.95	.00- 1.49	.00- 0.60	.00- 0.30
	99	.00- 8.80	.00- 4.50	.00- 2.28	.00- 0.92	.00- 0.46
1	95	(.02- 8.88)	.02- 5.45	.12- 3.57	.32- 2.32	.48- 1.83
	99	(.00-12.02)	.00- 7.21	.05- 4.55	.22- 2.80	.37- 2.13
2	95	.05-10.66	.24- 7.04	.55- 5.04	1.06- 3.56	1.29- 3.01
	99	.01-13.98	.10- 8.94	.34- 6.17	.87- 4.12	1.13- 3.36
3	95	(.27-12.19)	.62- 8.53	1.11- 6.42	1.79- 4.81	2.11- 4.19
	99	(.16-15.60)	.34-10.57	.78- 7.65	1.52- 5.44	1.88- 4.59
4	95	.49-13.72	1.10- 9.93	1.74- 7.73	2.53- 6.05	2.92- 5.36
	99	.21-17.21	.68-12.08	1.31- 9.05	2.17- 6.75	2.64- 5.82
5	95	(.88-15.14)	1.64-11.29	2.43- 9.00	3.26- 7.29	3.73- 6.54
	99	(.45-18.76)	1.10-13.53	1.89-10.40	2.83- 8.07	3.39- 7.05
6	95	1.26-16.57	2.24-12.60	3.18-10.21	4.11- 8.43	4.63- 7.64
	99	.69-20.32	1.56-14.93	2.57-11.66	3.63- 9.24	4.25- 8.18
7	95	(1.74-17.91)	2.86-13.90	3.88-11.47	4.96- 9.56	5.52- 8.73
	99	(1.04-21.72)	2.08-16.28	3.17-12.99	4.43-10.42	5.12- 9.31
8	95	2.23-19.25	3.51-15.16	4.70-12.61	5.81-10.70	6.42- 9.83
	99	1.38-23.13	2.63-17.61	3.93-14.18	5.23-11.60	5.98-10.43
9	95	(2.78-20.54)	4.20-16.40	5.46-13.82	6.66-11.83	7.32-10.93
	99	(1.80-24.46)	3.21-18.92	4.61-15.44	6.04-12.77	6.84-11.56
10	95	3.32-21.82	4.90-17.62	6.22-15.02	7.51-12.97	8.21-12.03
	99	2.22-25.80	3.82-20.20	5.29-16.70	6.84-13.95	7.70-12.69
11	95	(3.93-23.06)	5.65-18.80	7.05-16.16	8.41-14.06	9.14-13.10
	99	(2.70-27.11)	4.48-21.42	6.06-17.87	7.70-15.07	8.60-13.78
12	95	4.54-24.31	6.40-19.98	7.87-17.30	9.30-15.16	10.06-14.16
	99	3.18-28.42	5.15-22.65	6.83-19.05	8.56-16.19	9.51-14.86
13	95	(5.18-27.03)	7.11-21.20	8.70-18.44	10.20-16.25	10.99-15.23
	99	(3.72-29.67)	5.77-23.92	7.60-20.23	9.42-17.31	10.41-15.95
14	95	5.82-26.75	7.87-22.37	9.53-19.58	11.09-17.34	11.92-16.30
	99	4.25-30.92	6.46-25.13	8.38-21.40	10.28-18.43	11.31-17.04
15	95	(6.50-27.94)	8.64-23.53	10.36-20.72	11.98-18.44	12.84-17.37
	99	(4.82-32.14)	7.15-26.33	9.15-22.58	11.14-19.55	12.21-18.13

TABLE 23 Confidence limits for percentages

%	Confidence coefficients	n				
		50	100	200	500	1000
16	95	7.17-29.12	9.45-24.66	11.22-21.82	12.90-19.50	13.79-18.42
	99	5.40-33.36	7.89-27.49	9.97-23.71	12.03-20.63	13.14-19.19
17	95	(7.88-30.28)	10.25-25.79	12.09-22.92	13.82-20.57	14.73-19.47
	99	(6.00-34.54)	8.63-28.65	10.79-24.84	12.92-21.72	14.07-20.25
18	95	8.58-31.44	11.06-26.92	12.96-24.02	14.74-21.64	15.67-20.52
	99	6.60-35.73	9.37-29.80	11.61-25.96	13.81-22.81	14.99-21.32
19	95	(9.31-32.58)	11.86-28.06	13.82-25.12	15.66-22.71	16.62-21.57
	99	(7.23-36.88)	10.10-30.96	12.43-27.09	14.71-23.90	15.92-22.38
20	95	10.04-33.72	12.66-29.19	14.69-26.22	16.58-23.78	17.56-22.62
	99	7.86-38.04	10.84-32.12	13.26-28.22	15.60-24.99	16.84-23.45
21	95	(10.79-34.84)	13.51-30.28	15.58-27.30	17.52-24.83	18.52-23.65
	99	(8.53-39.18)	11.63-33.24	14.11-29.31	16.51-26.05	17.78-24.50
22	95	11.54-35.95	14.35-31.37	16.48-28.37	18.45-25.88	19.47-24.69
	99	9.20-40.32	12.41-34.35	14.97-30.40	17.43-27.12	18.72-25.55
23	95	(12.30-37.06)	15.19-32.47	17.37-29.45	19.39-26.93	20.43-25.73
	99	(9.88-41.44)	13.60-34.82	15.83-31.50	18.34-28.18	19.67-26.59
24	95	13.07-38.17	16.03-33.56	18.27-30.52	20.33-27.99	21.39-26.77
	99	10.56-42.56	13.98-36.57	16.68-32.59	19.26-29.25	20.61-27.64
25	95	(13.84-39.27)	16.88-34.66	19.16-31.60	21.26-29.04	22.34-27.81
	99	(11.25-43.65)	14.77-37.69	17.54-33.68	20.17-30.31	21.55-28.69
26	95	14.63-40.34	17.75-35.72	20.08-32.65	22.21-30.08	23.31-28.83
	99	11.98-44.73	15.59-38.76	18.43-34.75	21.10-31.36	22.50-29.73
27	95	(15.45-41.40)	18.62-36.79	20.99-33.70	23.16-31.11	24.27-29.86
	99	(12.71-45.79)	16.42-39.84	19.31-35.81	22.04-32.41	23.46-30.76
28	95	16.23-42.48	19.50-37.85	21.91-34.76	24.11-32.15	25.24-30.89
	99	13.42-46.88	17.25-40.91	20.20-36.88	22.97-33.46	24.41-31.80
29	95	(17.06-43.54)	20.37-38.92	22.82-35.81	25.06-33.19	26.21-31.92
	99	(14.18-47.92)	18.07-41.99	21.08-37.94	23.90-34.51	25.37-32.84
30	95	17.87-44.61	21.24-39.98	23.74-36.87	26.01-34.23	27.17-32.95
	99	14.91-48.99	18.90-43.06	21.97-39.01	24.83-35.55	26.32-33.87

TABLE 23 Confidence limits for percentages

%	Confidence coefficients	*n*				
		50	100	200	500	1000
31	95	(18.71-45.65)	22.14-41.02	24.67-37.90	26.97-35.25	28.15-33.97
	99	(15.68-50.02)	19.76-44.11	22.88-40.05	25.78-36.59	27.29-34.90
32	95	19.55-46.68	23.04-42.06	25.61-38.94	27.93-36.28	29.12-34.99
	99	16.46-51.05	20.61-45.15	23.79-41.09	26.73-37.62	28.25-35.92
33	95	(20.38-47.72)	23.93-43.10	26.54-39.97	28.90-37.31	30.09-36.01
	99	(17.23-52.08)	21.47-46.19	24.69-42.13	27.68-38.65	29.22-36.95
34	95	21.22-48.76	24.83-44.15	27.47-41.01	29.86-38.33	31.07-37.03
	99	18.01-53.11	22.33-47.24	25.60-43.18	28.62-39.69	30.18-37.97
35	95	(22.06-49.80)	25.73-45.19	28.41-42.04	30.82-39.36	32.04-38.05
	99	(18.78-54.14)	23.19-48.28	26.51-44.22	29.57-40.72	31.14-39.00
36	95	22.93-50.80	26.65-46.20	29.36-43.06	31.79-40.38	33.02-39.06
	99	19.60-55.13	24.08-49.30	27.44-45.24	30.53-41.74	32.12-40.02
37	95	(23.80-51.81)	27.57-47.22	30.31-44.08	32.76-41.39	34.00-40.07
	99	(20.42-56.12)	24.96-50.31	28.37-46.26	31.49-42.76	33.09-41.03
38	95	24.67-52.81	28.49-48.24	31.25-45.10	33.73-42.41	34.98-41.09
	99	21.23-57.10	25.85-51.32	29.30-47.29	32.45-43.78	34.07-42.05
39	95	(25.54-53.82)	29.41-49.26	32.20-46.12	34.70-43.43	35.97-42.10
	99	(22.05-58.09)	26.74-52.34	30.23-48.31	33.42-44.80	35.04-43.06
40	95	26.41-54.82	30.33-50.28	33.15-47.14	35.68-44.44	36.95-43.11
	99	22.87-59.08	27.63-53.35	31.16-49.33	34.38-45.82	36.02-44.08
41	95	(27.31-55.80)	31.27-51.28	34.12-48.15	36.66-45.45	37.93-44.12
	99	(23.72-60.04)	28.54-54.34	32.11-50.33	35.35-46.83	37.00-45.09
42	95	28.21-56.78	32.21-52.28	35.08-49.16	37.64-46.46	38.92-45.12
	99	24.57-60.99	29.45-55.33	33.06-51.33	36.32-47.83	37.98-46.10
43	95	(29.10-57.76)	33.15-53.27	36.05-50.16	38.62-47.46	39.91-46.13
	99	(25.42-61.95)	30.37-56.32	34.01-52.34	37.29-48.84	38.96-47.10
44	95	30.00-58.74	34.09-54.27	37.01-51.17	39.60-48.47	40.90-47.14
	99	26.27-62.90	31.28-57.31	34.95-53.34	38.27-49.85	39.95-48.11
45	95	(30.90-59.71)	35.03-55.27	37.97-52.17	40.58-49.48	41.89-48.14
	99	(27.12-63.86)	32.19-58.30	35.90-54.34	39.24-50.86	40.93-49.12
46	95	31.83-60.67	35.99-56.25	38.95-53.17	41.57-50.48	42.88-49.14
	99	28.00-64.78	33.13-59.26	36.87-55.33	40.22-51.85	41.92-50.12
47	95	(32.75-61.62)	36.95-57.23	39.93-54.16	42.56-51.48	43.87-50.14
	99	(28.89-65.69)	34.07-60.22	37.84-56.31	41.21-52.85	42.91-51.12
48	95	33.68-62.57	37.91-58.21	40.91-55.15	43.55-52.47	44.87-51.14
	99	29.78-66.61	35.01-61.19	38.80-57.30	42.19-53.85	43.90-52.12
49	95	(34.61-63.52)	38.87-59.19	41.89-56.14	44.54-53.47	45.86-52.14
	99	(30.67-67:53)	35.95-62.15	39.77-58.28	43.18-54.84	44.89-53.12
50	95	35.53-64.47	39.83-60.17	42.86-57.14	45.53-54.47	46.85-53.15
	99	31.55-68.45	36.89-63.11	40.74-59.26	44.16-55.84	45.89-54.11

LARGE-SAMPLE INFERENCE FOR A POPULATION PROPORTION

Draw an SRS of size n from a large population with unknown proportion p of successes. An approximate level C confidence interval for p is

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z^* is the upper $(1-C)/2$ standard normal critical value. To test the hypothesis $H_0\colon p = p_0$, compute the z statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

In terms of a standard normal random variable Z, the approximate P-value for a test of H_0 against

$$\begin{array}{lll} H_a\colon p > p_0 & \text{is} & P(Z \geq z) \\ H_a\colon p < p_0 & \text{is} & P(Z \leq z) \\ H_a\colon p \neq p_0 & \text{is} & 2P(Z \geq |z|) \end{array}$$

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The level C confidence interval for a proportion p will have margin of error approximately equal to a specified value m when the sample size is

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

where p^* is a guessed value for the true proportion.

The margin of error will be less than or equal to m if p^* is chosen to be 0.5. This gives

$$n = \left(\frac{z^*}{2m}\right)^2$$

SIGNIFICANCE TESTS FOR COMPARING TWO PROPORTIONS

To test the hypothesis

$$H_0: p_1 = p_2$$

compute the z statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_p}$$

where

$$s_p = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

In terms of a standard normal random variable Z, the P-value for a test of H_0 against

$$\begin{array}{lll} H_a: p_1 > p_2 & \text{is} & P(Z \geq z) \\ H_a: p_1 < p_2 & \text{is} & P(Z \leq z) \\ H_a: p_1 \neq p_2 & \text{is} & 2P(Z \geq |z|) \end{array}$$

CONFIDENCE INTERVALS FOR COMPARING TWO PROPORTIONS

Draw an SRS of size n_1 from a large population having proportion p_1 of successes and an independent SRS of size n_2 from another population having proportion p_2 of successes. When n_1 and n_2 are large, an approximate level C confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* s_D$$

where

$$s_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and z^* is the upper $(1 - C)/2$ standard normal critical value.

Example

8.31 A clinical trial examined the effectiveness of aspirin in the treatment of cerebral ischemia (stroke). Patients were randomized into treatment and control groups. The study was double-blind in the sense that neither the patients nor the physicians who evaluated the patients knew which patients received aspirin and which the placebo tablet. After 6 months of treatment, the attending physicians evaluated each patient's progress as either favorable or unfavorable. Of the 78 patients in the aspirin group, 63 had favorable outcomes; 43 of the 77 control patients had favorable outcomes. (From William S. Fields et al., "Controlled trial of aspirin in cerebral ischemia," *Stroke*, 8 (1977), pp. 301–315.)

(a) Compute the sample proportions of patients having favorable outcomes in the two groups.

(b) Give a 95% confidence interval for the difference between the favorable proportions in the treatment and control groups.

(c) The physicians conducting the study had concluded from previous research that aspirin was likely to increase the chance of a favorable outcome. Carry out a significance test to confirm this conclusion. State hypotheses, find the *P*-value, and write a summary of your results.

(a) $\frac{63}{78} = 0.808, \quad \frac{43}{77} = 0.588$

(b) $$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= (.808 - .558) \pm 1.96\sqrt{\frac{(.808)(.192)}{78} + \frac{(.558)(.442)}{77}}$$

$$\Rightarrow (0.109, 0.391)$$

(c) $H_0: \pi_1 = \pi_2, \quad H_A: \pi_1 > \pi_2$ (one sided)

$$\boxed{\hat{p} = \frac{63+43}{78+77} = 0.684}$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 3.35, \quad p < 0.0004$$

Conclusion: There is sufficient evidence to reject H_0.

142

DESIGN SENSITIVITY

Statistical Power for Experimental Research

Mark W. Lipsey

SAGE PUBLICATIONS

Chi-Square and the Difference between Proportions

When dependent variables in treatment effectiveness research are categorical rather than continuous, the results are usually presented as a contingency table and tested using Chi-square or some test of the difference between proportions. A typical case, and the only one considered here, is the 2×2 contingency table in which the degree of association between a dichotomous group variable (e.g., treatment vs. control) and a dichotomous dependent variable (e.g., success vs. failure) is tested. Each cell of that table contains a frequency value, that is, the number of subjects in the indicated group with the indicated outcome.

For example, a researcher with 100 subjects evenly divided between treatment and control group and measured on a dependent variable with a "success" baserate of 50% would expect the following results under the null hypothesis:

	Success	Failure
Treatment	25	25
Control	25	25

A treatment effect that altered the success rate to 70% would produce the following table:

	Success	Failure
Treatment	35	15
Control	25	25

To do statistical power analysis for such a situation using the charts in this chapter, the data from the contingency table should first be converted to proportions within each experimental group. That is, the treatment group data should be represented as the proportion in each of the two outcome categories and the control data should be represented likewise. Thus the data for the above example would appear as:

	Success	Failure
Treatment	.70	.30
Control	.50	.50

Power relationships for the situation above can be determined to a close approximation using an effect size index based on the difference between the "success" proportions of the treatment versus control group (or whatever other category of interest is analogous to the success category of the example here). To compute the appropriate effect size, the relevant proportions must be transformed. Cohen (1977) uses the arcsine transformation as follows:

> Let p_t be the success proportion for the treatment group;
> let p_c be the analogous proportion for the control group;
> Let ϕ_t be the arcsine transformation $2\arcsin(\sqrt{p_t})$ and correspondingly, $\phi_c = 2\arcsin(\sqrt{p_c})$.

The effect size index for the difference between p_t and p_c can then be expressed as follows:

$$ES_p = \phi_t - \phi_c \qquad [1]$$

Where ES_p is the effect size formulation for the difference between proportions, and ϕ_t and ϕ_c are the arcsine transformations of the success proportions for the treatment and control populations respectively.

Following convention, we assign ES_p the absolute value of the difference for purposes of determining power, then give it a plus sign if the treatment group results are superior to the control group results, a minus sign if the control group results are superior. Table 4.1 provides the arcsine transformations for proportions from .01 to .99 in increments of .01.

Example. Suppose a medical researcher is considering a study of a new cancer therapy in which the control group survival rate after two years is

143

TABLE 4.1 Arcsine Transformations (ϕ) for Proportions (*p*)

p	ϕ	*p*	ϕ	*p*	ϕ	*p*	ϕ
.01	.200	.26	1.070	.51	1.591	.76	2.118
.02	.284	.27	1.093	.52	1.611	.77	2.141
.03	.348	.28	1.115	.53	1.631	.78	2.165
.04	.403	.29	1.137	.54	1.651	.79	2.190
.05	.451	.30	1.159	.55	1.671	.80	2.214
.06	.495	.31	1.181	.56	1.691	.81	2.240
.07	.536	.32	1.203	.57	1.711	.82	2.265
.08	.574	.33	1.224	.58	1.731	.83	2.292
.09	.609	.34	1.245	.59	1.752	.84	2.319
.10	.644	.35	1.266	.60	1.772	.85	2.346
.11	.676	.36	1.287	.61	1.793	.86	2.375
.12	.707	.37	1.308	.62	1.813	.87	2.404
.13	.738	.38	1.328	.63	1.834	.88	2.434
.14	.767	.39	1.349	.64	1.855	.89	2.465
.15	.795	.40	1.369	.65	1.875	.90	2.498
.16	.823	.41	1.390	.66	1.897	.91	2.532
.17	.850	.42	1.410	.67	1.918	.92	2.568
.18	.876	.43	1.430	.68	1.939	.93	2.606
.19	.902	.44	1.451	.69	1.961	.94	2.647
.20	.927	.45	1.471	.70	1.982	.95	2.691
.21	.952	.46	1.491	.71	2.004	.96	2.739
.22	.976	.47	1.511	.72	2.026	.97	2.793
.23	1.000	.48	1.531	.73	2.049	.98	2.858
.24	1.024	.49	1.551	.74	2.071	.99	2.941
.25	1.047	.50	1.571	.75	2.094		

SOURCE: Computer generated using Microsoft Basic functions

anticipated to be 25% and treatment is expected to improve that to 40%. From Table 4.1, the arcsine transform of .40 is 1.369 and that of .25 is 1.047. The simple difference of these values as per equation [I] above (i.e., 1.369 − 1.047) gives the effect size, ES_p = .32 (rounded). If the researcher has available 100 subjects for each experimental group, Figure 4.5 shows that at α = .10 the power for detecting *ES* = .32 will be about .72.

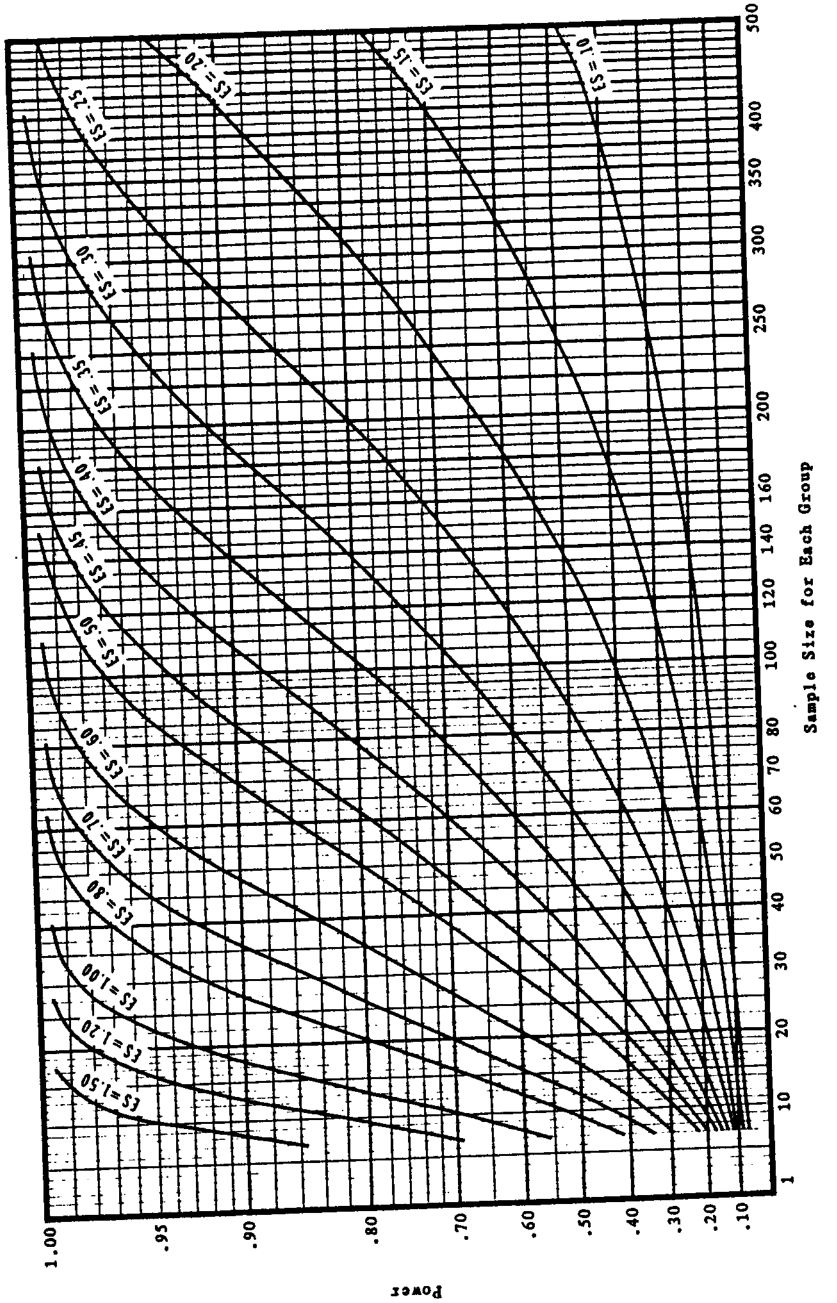


Figure 4.5: Power Chart for α = .10, Two-Tailed or α = .05, One-Tailed

Comparing Two or More Proportions

The generic setup is:

	Category 1	Category 2	...	Category c
Population 1	n_{11}	n_{12}	...	n_{1c}
Population 2	n_{21}	n_{22}	...	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Population r	n_{r1}	n_{r2}	...	n_{rc}

Examples:

1. Use of Stroke Unit versus Medical Unit for acute stroke in the elderly (Taken from Garraway et al, British Medical Journal, 1980).

	Patient Independent	Patient Dependent
Stroke Unit	67	34
Medical Unit	46	45

2. Quality of Sleep before elective operation

	Bad	Reasonably Good	Very Good
Triazolam	2	17	12
Placebo	8	15	8

Example 1 can be handled by the methods for two proportions based on the binomial distribution, which we have already seen. However, it is not possible to directly extend these methods to the case when there are three (or more) outcome categories and/or more than two populations. Furthermore, we have been using the Normal distribution approximation to the binomial, which we know is only valid for "large enough" sample sizes. What can we do if we have a table larger than 2 × 2 or if the sample size is "small"?

Methods to Compare Two or More Proportions

Suppose we wish to test the null hypothesis that $\pi_1 = \pi_2 = \ldots = \pi_N$, that is, we have measured the frequency of occurrence of a dichotomous outcome in N populations, and wish to check if the frequencies are all equal. There are several candidate tests:

Normal approximation (Z) Test: We have seen this test when $N = 2$. The test does not apply when $N > 2$. Alternative hypothesis can be one or two-sided. Requires large samples sizes to be accurate. "Large" is often stated as a criterion like

$$\text{sample size } \times \min\{\pi, (1 - \pi)\} \geq 5.$$

This is somewhat arbitrary, but works reasonably well as a rough guide.

Chi-square (χ^2) Test: The χ^2 test does apply when $N > 2$, but the alternative hypothesis is always two-sided. Requires large samples sizes to be accurate. "Large" is often operationalized as "the expected number of subjects in each cell in the $r \times c$ table must be at least 5". We will see soon how to calculate these expected cell sizes.

Fisher's Exact Test: Both the χ^2 and Z tests require "large" sample sizes to be accurate, but the Fisher's Exact is "exact" for any sample size. The Fisher's Exact Test also applies when $N > 2$, but unlike the χ^2 test, the alternative hypothesis can be one or two-sided.

While it is common practice to use a χ^2 test for large sample sizes and Fisher's Exact Test for smaller sample sizes, a natural question is "Why not just use Fisher's Exact Test all the time, since it is always applicable?" There are two possible answers. The first is that, as we will see, it is computational "expensive" to use Fisher's Exact Test, compared to a χ^2 test. Second, there are different assumptions behind each. As will become clear from the examples on the next few pages, in the Fisher's Exact Test, all "margins" are held fixed ("conditioned upon"), while this is not the case for the Z and χ^2 tests. Thus there is a slightly different inferential philosophy behind each.

One sample χ^2 Test

Suppose we observe the following table of data:

	Success	Failure
Population	x	$n-x$

We would like to test the hypothesis $H_0 : \pi = \pi_0$. For example, we might observe patient survival rates one month following a particular sugery, and would lke to test if the survival rate is 80%. We observe the following data:

	Success	Failure
Population	60	40

We would like to test the hypothesis $H_0 : \pi = 0.80$, where π represents the true one month survival rate.

Procedure: Since we hypothesize $\pi = 0.80$, and since we have 100 subjects, we *expect* 80 survivors and 20 deaths. Observed dicrepancies from these expected values are evidence against the null hypothesis. We calculate:

$$\begin{aligned} X^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(60-80)^2}{80} + \frac{(40-20)^2}{20} \\ &= 400/80 + 400/20 = 25 \end{aligned}$$

Comparing the $X^2 = 25$ value on χ^2 tables with 1 degree of freedom (1 df), we find that $p < 0.0005$, so that we have evidence to reject the null hypothesis.

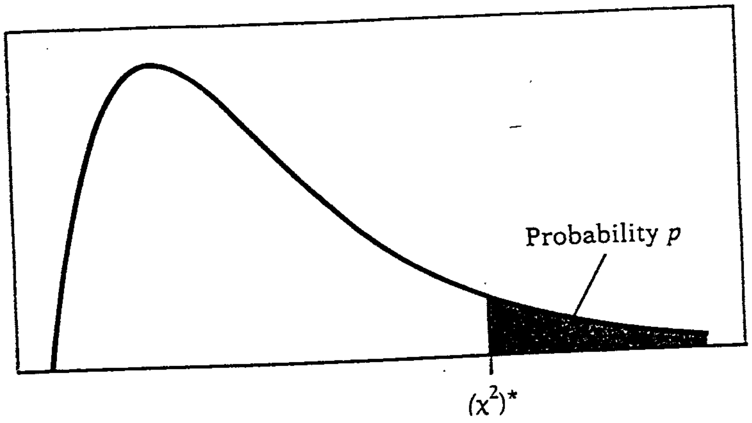


Table entry for p is the point $(X^2)^*$ with probability p lying above it.

Table G χ^2 critical values

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2

Two sample χ^2 Test

Suppose we observe the following table of data, introduced previously:

	Patient Independent	Patient Dependent	Total
Stroke Unit	67	34	101
Medical Unit	46	45	91
Total	113	79	192

We would like to test the hypothesis $H_0 : \pi_1 = \pi_2$; that is, the preportion of independent patients is the same on Medical or Stroke Units.

Procedure: Since we hypothesize $\pi_1 = \pi_2$, we *expect* to observe the following table of data, on average:

	Patient Independent	Patient Dependent	Total
Stroke Unit	59.44	41.56	101
Medical Unit	53.56	37.44	91
Total	113	79	192

Once again, observed dicrepancies from these expected values are evidence against the null hypothesis. We calculate:

$$\begin{aligned} X^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(67-59.44)^2}{59.44} + \frac{(34-41.56)^2}{41.56} + \frac{(46-53.56)^2}{53.56} + \frac{(45-37.44)^2}{37.44} \\ &= 4.9268 \end{aligned}$$

Comparing the $X^2 = 4.9268$ value on χ^2 tables with 1 df, we find that $0.025 < p < 0.05$ (by computer the exact value is 0.0264), so that we have evidence to reject the null hypothesis.

The χ^2 Test for 2 × 3 table

Suppose we observe the following table of data, introduced previously:

	Bad	Reasonably Good	Very Good	Total
Triazolam	2	17	12	31
Placebo	8	15	8	31
Total	10	32	20	62

We would like to test the hypothesis $H_0 : \pi_1 = \pi_2 = \pi_3$; that is, the preportions of patients that experience bad, reasonably good and very good outcomes are the same whether they were given the drug or the placebo.

Procedure: Since we hypothesize $\pi_1 = \pi_2 = \pi_3$, we *expect* to observe the following table of data, on average:

	Bad	Reasonably Good	Very Good	Total
Triazolam	5	16	10	31
Placebo	5	16	10	31
Total	10	32	20	62

As before, observed dicrepancies from these expected values are evidence against the null hypothesis. We calculate:

$$\begin{aligned} X^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(2-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(17-16)^2}{16} + \frac{(15-16)^2}{16}\frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} \\ &= 4.525 \end{aligned}$$

Comparing the $X^2 = 4.525$ value on χ^2 tables with 2 df, we find that $0.10 < p < 0.15$ (by computer the exact value is 0.104), so that we do not have sufficient evidence to reject the null hypothesis at either the 0.05 or 0.10 levels. **Note that in general for an $r \times c$ table, df = (r − 1) × (c − 1).**

Question: We note that the proportion on triazolam increases from 20% to 53% to 60% across outcomes, so it may be a good idea to test for a trend. See Armitage and Berry, page 403.

Fisher's Exact Test

Suppose we observe the following table of data:

	Success	Failure	Total
Group A	4	2	6
Group B	1	6	7
Total	5	8	13

As with the Z and χ^2 tests, we would like to test the null hypothesis $H_0 : \pi_1 = \pi_2$. However, since the sample size is so small, there is doubt about the applicability of these tests to this data set. An "exact" test can be constructed via the following reasoning:

We have observed a total of 5 successes. If groups A and B receive equally effective treatments, then the five successes should be eqaully distributed between the two groups. If the sample sizes were equal, we would expect 2.5 successes in each group, but since the sizes are not equal, we expect the successes to be divided in a 6:7 ratio (almost but not quite half/half). As in the previous tests, discrepancies from this "fair split" indicate departures from the null hypothesis. We calculate:

$$\frac{6}{13} \times 5 = 2.31, \text{ and } \frac{7}{13} \times 5 = 2.69$$

Therefore, approximately 2:3 or 3:2 split is expected, and more extreme splits are evidence against the null hypothesis. How extreme is too extreme to be compatible with the null hypothesis? We will calculate the probability of each possible split:

A	5	1	4	2	3	3	2	4	1	5	0	6
B	0	7	1	6	2	5	3	4	4	3	5	2
	5	8	5	8	5	8	5	8	5	8	5	8
Prob	0.005		0.082		0.326		0.408		0.163		0.016	

The tables with probabilities of $0.005 + 0.082 + 0.016 = 0.103$ have values equal to or more extreme than those observed, so by the definition of the p-value, $p = 0.103$ by the Fisher's Exact Test.

Calculating Probabilities for the Fisher's Exact Test

The probabilities on the previous page were calculated using the **hypergeometric distribution.** In general, if we observe

A	a	b	$a+b$
B	c	d	$c+d$
	$a+c$	$b+d$	N

where $N = a + b + c + d$, then the probability of observing that table is:

$$\text{Prob} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

A less simplified equivalent formulae provides a clue as to how the probability is calculated:

$$\text{Prob} = \frac{\frac{(a+c)!}{a!c!} \times \frac{(b+d)!}{b!c!}}{\frac{N!}{(a+b)!(c+d)!}}$$

Consider the A and B labels as random labels. In how many ways can one choose that all 5 (or 4 or 3 or 2 or 1 or 0) of the A labels happen to end up as "successes"?

Odds Ratios and Relative Risk

Suppose we observe the following 2×2 table of data:

	Disease +	Disease −	Total
Risk Factor +	a	c	$a+c$
Risk Factor −	b	d	$b+d$
Total	$a+b$	$c+d$	N

Then the observed odds ratio is

$$\hat{\psi} = \frac{ad}{bc},$$

and the observed relative risk is

$$\hat{RR} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}.$$

Note that if the risks $\frac{a}{a+c}$ and $\frac{b}{b+d}$ are small, then $\hat{\psi} \approx \hat{RR}$, since $a << c$ and $b << d$.

Confidence interval for Odds Ratios: The distribution of $\hat{\psi}$ is somewhat skew, so that the confidence interval is usually based on a Normal Distribution approximation to $\log \hat{\psi}$. In particular,

$$\mathrm{var}(\log \hat{\psi}) \approx \frac{1}{\mathrm{a}} + \frac{1}{\mathrm{b}} + \frac{1}{\mathrm{c}} + \frac{1}{\mathrm{d}}$$

so that a 95% CI for $\log \psi$ is given by

$$\left(\log \hat{\psi} - 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}, \log \hat{\psi} + 1.96 \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

To convert back to a CI for ψ, one takes the exponent (to the base $e = 2.71828$), to get

$$\left(\exp\left[\log\hat{\psi} - 1.96 \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right], \exp\left[\log\hat{\psi} + 1.96 \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]\right).$$

Similarly,

$$\mathrm{var}(\log \hat{\mathrm{RR}}) \approx \frac{\mathrm{c}}{\mathrm{a(a+c)}} + \frac{\mathrm{d}}{\mathrm{b(b+d)}}$$

so that an approximate 95% CI for a RR is

$$\left(\exp\left[\log\hat{RR} - 1.96 \times \sqrt{\frac{c}{a(a+c)}+\frac{d}{b(b+d)}}\right], \exp\left[\log\hat{RR} + 1.96 \times \sqrt{\frac{c}{a(a+c)}+\frac{d}{b(b+d)}}\right]\right).$$

Example: The following two tables of data are observed from a study on the effects of swimming in polluted water:

	Otitus +	Otitus −	Total
Swimmers	7	72	79
Non-swimmers	2	39	41
Total	9	111	120

	Any symptoms +	Any symptoms −	Total
Swimmers	45	34	79
Non-swimmers	8	33	41
Total	53	67	120

For otitus, $\hat{RR} = \frac{7/79}{2/41} = 1.816$, and $\hat{\psi} = \frac{7/72}{2/39} = 1.896$, so the OR is a good approximation for the RR. A 95% CI for the OR is $(.376, 9.57)$, while a 95% CI for the RR is $(0.395, 8.349)$. For all symptoms, however, $\hat{RR} = \frac{45/79}{8/41} = 2.92$, and $\hat{\psi} = \frac{45/34}{8/33} = 5.46$, so the OR is not an accurate approximation for the RR.

Mantel-Haenzel Estimates

Suppose we observe the following two 2×2 tables of data, the first one representing the relationship between the risk factor and disease in males, the second one for females:

	Disease +	Disease −	Total
Factor +	160	80	240
Factor −	440	320	760
Total	600	400	1000

	Disease +	Disease −	Total
Factor +	240	330	570
Factor −	160	270	430
Total	400	600	1000

Then $\hat{\psi} = 1.45$ in males, and $\hat{\psi} = 1.22$ in females. However, if we construct the combined table,

	Disease +	Disease −	Total
Factor +	400	410	810
Factor −	600	590	1290
Total	1000	1000	2000

we find $\hat{\psi} = 0.95$!! This arises because of confounding, so that the **Mantel-Haenzel** estimator must be used to combine the tables.

Calculating Mantel-Haenzel Estimates

The Mantel-Haenzel estimate is given by:

$$\hat{\psi}_{MH} = \frac{\sum_{(\text{all tables } i)} a_i d_i / n_i}{\sum_{(\text{all tables } i)} b_i c_i / n_i}$$

where each table is represented by

	Disease +	Disease −	Total
Risk Factor +	a_i	c_i	
Risk Factor −	b_i	d_i	
Total			n_i

For the above data,

$$\hat{\psi}_{MH} = \frac{(160 \times 320)/1000 + (240 \times 270)/1000}{(80 \times 440)/1000 + (330 \times 160)/1000} = 1.32$$

Note that unlike the 0.95 value for the "straight" combined estimate, the Mantel Haenzel combined estimate lies in between the separate values for males and females, which is intuitively what would be expected.

See Armitage and Berry for testing and confidence interval formulae.

Inference for Poisson Counts

Let $\mu = \lambda \times t$ be a Poisson parameter. If we observe data x representing a count of events in time t, how can we best estimate μ?

Point estimation: The maximum likelihood estimator of μ is simply x, and the maximum likelihood estimator of λ is simply x/t. These are intuitive estimates, but remember that we can always prove these results by using calculus as on page 117 for normal means.

Confidence intervals: While exact results are available (see, for example, Rosner page 180), we will use a normal approximation. A $(1-\alpha)\%$ confidence for μ is given by

$$\left(x - z_{1-\alpha/2}\sqrt{x}, x + z_{1-\alpha/2}\sqrt{x}.\right)$$

To obtain a confidence interval for λ, simply divide the lower and upper confidence interval limits for μ by t. These results are not terribly accurate, but are fine for our purposes here.

Example: A person-year is defined as a person being followed for one year of time. Suppose that a certain city in the United States has a constant number, 12,000, of children less than 19 years of age. (In other words, we assume that the number of 18 year olds turning 19 in any given year is approximately equal to the number of births, so that the size of the cohort remains constant.) Suppose that 12 cases of leukemia are seen in this city over a 10 year period. What is a 95% confidence interval per 100,000 person-years (a typical way that such data are usually presented)?

Solution: A 95% CI for μ is given by

$$\left(x - z_{1-\alpha/2}\sqrt{x}, x + z_{1-\alpha/2}\sqrt{x}.\right) = \left(12 - z_{1-\alpha/2}\sqrt{12}, 12 + z_{1-\alpha/2}\sqrt{12}\right) = (5.2, 18.8)$$

This would be the CI for the rate per $12,000 \times 10$ person years, so for 100,000 person-years would be

$$\left(5.2 \times \frac{100,000}{120,000}, 18.8 \times \frac{100,000}{120,000}\right) = (4.3, 15.7).$$

The exact interval is (5.2, 17.5), so we are not too far off using the approximation (but the exact method is better if you have tables or a program that implements it).

Note: Bayesian inferences for Poisson parameters typically use the fact that the Gamma distribution is conjugate to Poisson likelihood functions. See Gelman et al. (1995), page 48.

Bayesian Inference for Proportions

Suppose that in a given experiment x successes are observed in N independent Bernoulli trials. Let θ denote the true but unknown probability of success, and suppose that the problem is to find an interval that covers the most likely locations for θ given the data.

The Bayesian solution to this problem follows the usual pattern:

1. Write down the likelihood function for the data.
2. Write down the prior distribution for the data.
3. Use Bayes theorem to derive the posterior distribution.

For the case of a single binomial parameter, these steps are realized by:

1. The likelihood is the usual binomial probability formula, the same one used in the frequentist analysis,

$$L(\theta|x) = Pr\{x \text{ successes in } N \text{ trials}\} = \frac{N!}{(N-x)!\,x!}\theta^x(1-\theta)^{(N-x)}.$$

In fact, all one needs to specify is that

$$L(\theta|x) = Pr\{x \text{ successes in } N \text{ trials}\} \propto \theta^x(1-\theta)^{(N-x)},$$

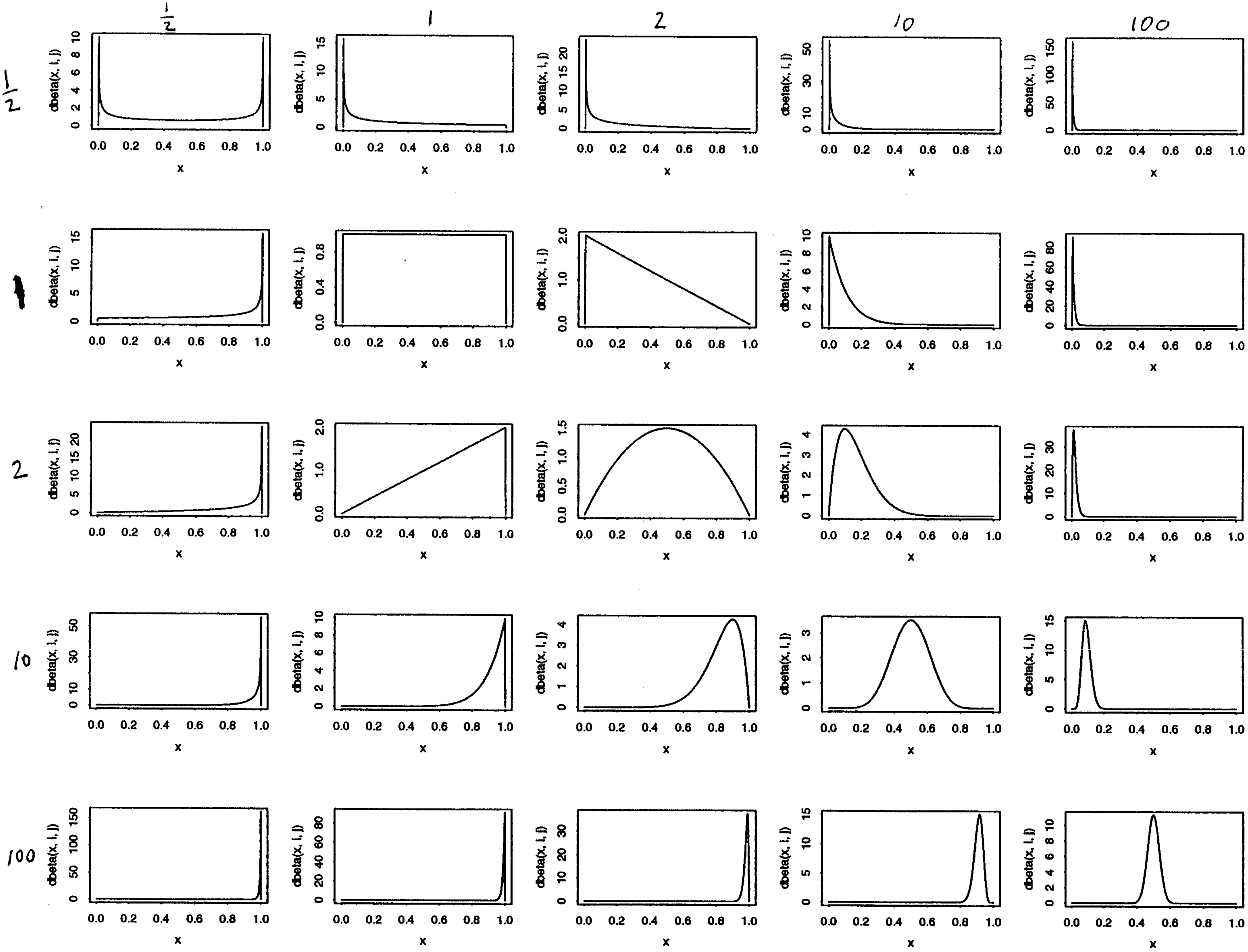
since $\frac{N!}{(N-x)!\,x!}$ is simply a constant that does not involve θ. In other words, inference will be the same whether one uses this constant or ignores it.

2. Although any prior distribution can be used, a convenient prior family is the Beta family, since it is the conjugate prior distribution for a binomial experiment. A random variable, θ, has a distribution that belongs to the Beta family if it has a probability density given by

$$f(\theta) = \begin{cases} \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, & 0 \le \theta \le 1,\ \alpha,\beta > 0, \text{ and} \\ 0, & \text{otherwise}, \end{cases} .$$

[$B(\alpha,\beta)$ represents the Beta function evaluated at (α,β). It is simply the normalizing constant that is necessary to make the density integrate to one, that is, $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$.] The mean of the Beta distribution is given by

$$\mu = \frac{\alpha}{\alpha+\beta},$$



The Beta density can assume a wide variety of shapes.

and the standard deviation is given by

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \quad .$$

Therefore, at this step, one needs only to specify α and β values, which can be done by finding the α and β values that give the correct prior mean and standard deviation values. This involves solving two equations in two unknowns. The solution is

$$\alpha = -\frac{\mu\ (\sigma^2 + \mu^2 - \mu)}{\sigma^2}$$

and

$$\beta = \frac{(\mu - 1)\,(\sigma^2 + \mu^2 - \mu)}{\sigma^2}$$

3. As always, Bayes Theorem says

   posterior distribution $\propto$ prior distribution × likelihood function.

   In this case, it can be shown (by relatively simple algebra which we will omit) that if the prior distribution is $Beta(\alpha, \beta)$, and the data is x successes in N trials, then the posterior distribution is $Beta(\alpha + x, \beta + N - x)$.

Example: Suppose that a new diagnostic test for a certain disease is being investigated. Suppose that 100 persons with confirmed disease are tested, and that 80 of these persons test positively.

(a) What is the posterior distribution of the sensitivity of the test if a Uniform $Beta(\alpha = 1, \beta = 1)$ prior is used? What is the posterior mean and standard deviation of this distribution?

(b) What is the posterior distribution of the sensitivity of the test if a $Beta(\alpha = 27, \beta = 3)$ prior is used? What is the posterior mean and standard deviation of this distribution?

(c) Draw a sketch of the prior and posterior distributions from both (a) and (b).

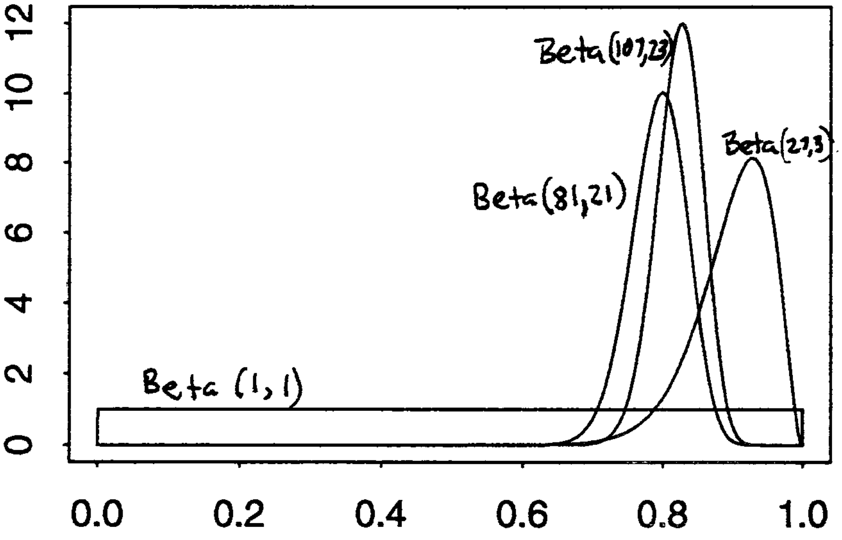
(d) Derive the 95% posterior credible intervals from the two posterior distributions given above, and compare it to the usual frequentist confidence interval for the data. Clearly distinguish the two different interpretations given to confidence intervals and credible intervals.

Solution:

(a) According to the result given above, the posterior distribution is again a Beta, with parameters $\alpha = 1 + 80 = 81$, $\beta = 1 + 20 = 21$. The mean of this distribution is $81/(81 + 21) = 0.794$, and the standard deviation is 0.0398.

(b) Again the posterior distribution is a Beta, with parameters $\alpha = 27 + 80 = 107$, $\beta = 3 + 20 = 23$. The mean of this distribution is $107/(107 + 23) = 0.823$, and the standard deviation is 0.0333.

(c)



(d) From tables of the beta density (contained in many books of statistical tables) or software that includes Bayesian analysis, the 95% credible intervals are (0.71, 0.86) from the Beta(81,21) posterior density, and (0.75, 0.88) from the Beta(107,23) posterior density. The frequentist 95% confidence interval is (0.71, 0.87).

Note that numerically, the frequentist confidence interval is nearly identical to the Bayesian credible interval starting from a Uniform prior. However, their interpretations are very different. Credible intervals are interpreted directly as the posterior probability that θ is in the interval, given the data and the prior distribution. No references to long run frequencies or other experiments are required. On the other hand, confidence intervals have the interpretation that if such procedures are used repeatedly, then $100(1-\alpha)\%$ of all such sets would in the long run contain the true parameter of interest. Notice that there can be nothing said about what happened in this particular case, the only inference is to the long run. To infer anything about the particular case from a frequentist analysis involves a "leap of faith."

Why isn't everybody a Bayesian?

Historically, before about 1925, everybody was a Bayesian. RA Fisher, and Neyman and Pearson in the 1920's and 1930's developed frequentist methods, motivated largely by the desire to get away from subjective prior distributions so as to arrive at "objective" inferences from data. Unfortunately, their methodology also has subjective biases. These are detailed in the article by Berger and Berry, referenced in the JAMA article, and relate to problems with defining the sample space for observations for the "hypothetical repeated trials". It is also unfortunate that frequentist methods are unable to directly address questions of interest to clinicians and other researchers.

So why did frequentist statistics become so popular despite these important deficiencies? There are many reasons, including:

1. Ease of use: To get a p-value, one just needs to plug in data, and software has become so user friendly that one does not even need to know what a p-value is to get them. To get a posterior distribution, however, one needs to carefully assess a prior distribution, an extra step that can be an enormous amount of work if done carefully. Further, one needs to better understand Bayesian methodology to apply it, since the software tends to be more difficult to use, requiring more insight.

2. Misunderstandings about the role of the prior distribution: As explained in the JAMA article, if one understands things more deeply, the prior distribution can be viewed as a great advantage. For example, it allows different researchers to compare their final conclusions based on their initial positions which have been updated by the data. Also, it allows one to assess the importance of the current data set in relation to past data, and come to an overall conclusion based on both (as real scientists must do before stating any conclusions). However, superficially, scientists are wary of the subjectivity implied by having to assess a prior distribution, and worry that others may not accept their analyses as "scientific". (I personally have not found this to be an issue in any of my submissions to medical journals.)

3. Non-uniqueness of prior and therefore of posterior distributions: By simply plugging in the data, everyone can agree (usually, there can be exceptions!)

on what the p-value is. Each scientist, however, will have their own prior distribution, and therefore posterior distributions are not unique. As explained above, this can be a great advantage, but this point is not always appreciated.

4. Education: Most medical researchers are far behind current developments in statistical methodology. Just like it took many years to have journals prefer confidence intervals to p-values, it will take a while for Bayesian methods to become commonplace in medical journals. Many medical researchers are still unaware of Bayesian methodology, but this situation is very rapidly changing.

5. Implementation issues: To a typical medical researcher, Bayesian methods can be more difficult to implement than frequentist methods. Not only is there a prior distribution to elicit, but numerical/computational problems involved with solving the integrals that arise in Bayes Theorem can be difficult. Off the shelf software is rapidly being developed, but is not currently widely available. Programs such as SAS have been very, very slow in adding Bayesian procedures to their programs, perhaps adding to the perception that the methods are not useful, or are not scientifically sound.

6. Is it "too extreme" a position to say that p-values should never be used? For simple models, confidence intervals can usually be calculated just as easily as p-values and provide much more clinically useful information. For more complex models, it is sometimes difficult to calculate confidence intervals (for example, confidence intervals for complex regression equations can be difficult to compute). So one sometimes sees p-values used to assess goodness of fit or a regression parameter in such situations. However, recently developed Bayesian Monte Carlo methods such as the Gibbs sampler make the Bayesian analysis of such complex models easier than a frequentist analysis, and also avoid serious problems with the way nuisance parameters are handled in such models. Hence I would say that p-values should never be used, and that this position is reasonable, and not in any way "extreme".

So, overall, one generally needs to work much harder to get a Bayesian posterior distribution than a p-value or a CI. Is the extra effort worth it? In simple problems, and where there is very little or no prior information, frequentist CI's (but not p-values!) provide very similar inferences to Bayesian credible intervals. In that case, it may not be worth the extra trouble (although in simple problems without prior information, Bayesian methods tend to be easy to use as well). In all other situations, where there may be substantial prior information and/or a complex model (Bayesian methods handle nuisance parameters like normal variances better than other methods), I would say that not only is the extra trouble worthwhile, it is crucial to deriving sound conclusions from data.

Nonparametric Inference

Thus far, statistical inferences on populations have been made by assuming a mathematical model for the population (for example, a Normal distribution), and estimating parameters from that distribution based on a sample. Once the parameters have been estimated (for example, the mean and/or variance for a Normal distribution), the distribution is fully specified. This is known as *parametric inference.*

Sometimes we may be unwilling to specify in advance the general shape of the distribution, and prefer to base the inference only on the data, without a parametric model. In this case, we have *distribution free*, or *nonparametric* methods.

Example: Suppose that a new postsurgical treatment is being compared with a standard treatment by observing the recovery times of n treatment subjects and m controls. Suppose that $m = n = 9$, and that the observed recovery times (in days) are:

Control:	20	21	24	30	32	36	40	48	54
Treatment:	19	22	25	26	28	29	34	37	38

Assume first that these data were matched. A *very* naïve procedure to compare treatment to control is as follows:

C	20	21	24	30	32	36	40	48	54
T	19	22	25	26	28	29	34	37	38
sign	+	−	−	+	+	+	+	+	+

Thus $\frac{7}{9} = 78\%$ were better in the treatment group. Is this likely to be due to chance, or is it "statistically significant"? If the two procedures are truely equivalent (i.e., under the null hypothesis H_0: There is no difference in recovery times between the two treatments), then we would expect roughly equal numbers of +'s and −'s. To test the null hypothesis, we could the calculate

$$p - value \;=\; Pr\{7 \text{ or } 8 \text{ or } 9+\text{'s}|\text{ T and C are equivalent}\}$$

$$= 0.089$$

from binomial tables, with $\pi = 0.5$, and $n = 9$. This is called the *sign test.*

The usual paired t-test gives $p = 0.023$, using the same data. Which procedure should we use?

Distribution Free denotes that we make no assumptions concerning the underlying distribution from which the data arise. However, note that we still used a distribution (here, the binomial) from which we calculated the $p - value$. The main difference between the two procedures is that the t-test requires the assumption that the data arise from a Normal distribution, while the sign test did not. In fact, the sign test made no assumptions about an underlying population, nor the shape of any distribution. In using the sign test, we also did not need to consider degrees of freedom, or whether we had equal variances or not.

Still, the sign test is very wasteful of information, since it assigns each value only a "+" or "–", regardless of the magnitude of the difference. We can take this into account by using a *signed rank test.*

C	20	21	24	30	32	36	40	48	54
T	19	22	25	26	28	29	34	37	38
sign	+	–	–	+	+	+	+	+	+
signed difference	+1	–1	–1	+4	+4	+7	+ 6	+ 11	+16

Ordering them and ranking them in order, we have:

	+1	–1	–1	+4	+4	+6	+ 7	+ 11	+16
ranks	1	2	3	4	5	6	7	8	9
ranks with ties	2	2	2	4.5	4.5	6	7	8	9

Summing up the Positive ranks, we have $T_+ = 41$, and summing the negative ranks, we have $T_- = 4$. If there is no difference in the two groups, then we would expect T_+ and T_- to be approximately equal to each other, so that unequal values indicate departures from the null hypothesis. Significance levels are given in table $A6$. Looking up $T_+ = 41$, we see that $0.02 < p < 0.05$. This nonparametric test is called the *Wilcoxon signed rank test.*

Suppose that the data were not paired, but instead came from two independent samples, as in a clinical trial. We may again order and rank the data:

data	19	20	21	22	24	25	26	28	29	30	32	34	36	37	38	40	48	54
ranks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
group	T	C	C	T	C	T	T	T	T	C	C	T	C	T	T	C	C	C

Now summing the ranks in the control group gives:

$$T_x = 2 + 3 + 5 + 10 + 11 + 13 + 16 + 17 + 18 = 95.$$

On average, *if the null hypothesis is true*, we would expect a sum of

$$E(T_x) = \frac{(m)(m+n+1)}{2} = \frac{(9)(9+9+1)}{2} = 85.5$$

Is the observed number, $T_x = 95$ "significantly" higher than what one would expect if the null hypothesis is true? This depends on the variance,

$$var(T_x) = \frac{mn(m+n+1)}{12} = \frac{(9 \times 9) \times (9+9+1)}{12} = 128.25$$

and thus $sd(T_x) = sqrt(128.25) = 11.32$.

If we assume a large enough sample,

$$z = \frac{T_x - E(T_x)}{sd(T_x)} \sim N(0,1),$$

so that the null hypothesis can be tested using the usual Normal tables.

Here we calculate

$$z = \frac{T_x - E(T_x)}{sd(T_x)} = \frac{95 - 85.5}{11.32} = 0.83,$$

so that $p = 0.41$ from Normal tables. If the sample is too small for the Normal approximation to hold, one can use table $A7$.

This nonparametric test is called the *Wilcoxon rank sum test.* The equivalent unpaired t-test for the same data give a $p - value$ of $p = 0.26$.

Nonparametric Statistics Summary

	paired or matched	unpaired or unmatched
Small Sample	N = number of pairs Use: Wilcoxon signed rank test Statistic: T_+ or T_- Table $A6$	$m + n = N$ Use Wilcoxon rank sum test Statistic: T_x Table $A7$
Large Sample	$E(T_+) = \frac{N(N+1)}{4}$ $var(T_+) = \frac{N(N+1)(2N+1)}{24}$ $z = \frac{T_+ - E(T_+)}{sd(T_+)}$ Use Normal Tables	$E(T_x) = \frac{m(N+1)}{2}$ $var(T_x) = \frac{mn(m+n+1)}{12}$ $z = \frac{T_x - E(T_x)}{sd(T_x)}$ Use Normal Tables

Notes:

- Slightly different formulæ may be used if there are ties in the data.
- The test does not directly test means,
  H_0: There is no treatment effect
  H_A: treatment is more/less effective than control
- Nonparametric confidence intervals are also available, see E. Lehmann (1975): Nonparametrics. Holden Day.

Table A6. Critical values for the signed rank test in the comparison of two groups in paired samples (this table gives the values of the sum of the signed ranks required to achieve statistical significance in a test of the null hypothesis of no difference in the population)

Number of differences	Significance level for two-tail test		
	.05	.02	.01
6	0, 21	—	—
7	2, 26	0, 28	—
8	3, 33	1, 35	0, 36
9	5, 40	3, 42	1, 44
10	8, 47	5, 50	3, 52
11	10, 56	7, 59	5, 61
12	13, 65	9, 69	7, 71
13	17, 74	12, 79	9, 82
14	21, 84	15, 90	12, 93
15	25, 95	19, 101	15, 105
16	29, 107	23, 113	19, 117
17	34, 119	28, 125	23, 130
18	40, 131	32, 139	27, 144
19	46, 144	37, 153	32, 158
20	52, 158	43, 167	37, 173
21	58, 173	49, 182	42, 189
22	66, 187	55, 198	48, 205
23	73, 203	62, 214	54, 222
24	81, 219	69, 231	61, 239
25	89, 236	76, 249	68, 257

Adapted from Tukey [107].

Table A7. Critical values for the rank sum test in the comparison of two groups in independent samples (this table gives, for the sum of the ranks in the smaller of two independent samples, the values required to achieve statistical significance in a test of the null hypothesis of no difference between the populations)

	Significance level, two-tail		
n_1, n_2	.05	.01	.001
2. 8	3. 19		
2. 9	3. 21		
2. 10	3. 23		
2. 11	4. 24		
2. 12	4. 26		
2. 13	4. 28		
2. 14	4. 30		
2. 15	4. 32		
2. 16	4. 34		
2. 17	5. 35		
2. 18	5. 37		
2. 19	5. 39	3. 41	
2. 20	5. 41	3. 43	
2. 21	6. 42	3. 45	
2. 22	6. 44	3. 47	
2. 23	6. 46	3. 49	
2. 24	6. 48	3. 51	
2. 25	6. 50	3. 53	
3. 5	6. 21		
3. 6	7. 23		
3. 7	7. 26		
3. 8	8. 28		
3. 9	8. 31	6. 33	
3. 10	9. 33	6. 36	
3. 11	9. 36	6. 39	
3. 12	10. 38	7. 41	
3. 13	10. 41	7. 44	
3. 14	11. 43	7. 47	
3. 15	11. 46	8. 49	
3. 16	12. 48	8. 52	
3. 17	12. 51	8. 55	
3. 18	13. 53	8. 58	
3. 19	13. 56	9. 60	
3. 20	14. 58	9. 63	
3. 21	14. 61	9. 66	6. 69
3. 22	15. 63	10. 68	6. 72
3. 23	15. 66	10. 71	6. 75
3. 24	16. 68	10. 74	6. 78
3. 25	19. 71	11. 76	6. 81
4. 4	10. 26		
4. 5	11. 29		
4. 6	12. 32	10. 34	
4. 7	13. 35	10. 38	
4. 8	14. 38	11. 41	
4. 9	15. 41	11. 45	
4. 10	15. 45	12. 48	
4. 11	16. 48	12. 52	
4. 12	17. 51	13. 55	
4. 13	18. 54	14. 58	10. 62
4. 14	19. 57	14. 62	10. 66
4. 15	20. 60	15. 65	10. 70
4. 16	21. 63	15. 69	11. 73
4. 17	21. 67	16. 72	11. 77
4. 18	22. 70	16. 76	11. 81
4. 19	23. 73	17. 79	12. 84
4. 20	24. 76	18. 82	12. 88
4. 21	25. 79	18. 86	12. 92
4. 22	26. 82	19. 89	13. 95
4. 23	27. 85	19. 93	13. 99
4. 24	28. 88	20. 96	13. 103
4. 25	28. 92	20. 100	14. 106
5. 5	17. 38	15. 40	
5. 6	18. 42	16. 44	
5. 7	20. 45	17. 48	
5. 8	21. 49	17. 53	
5. 9	22. 53	18. 57	15. 60
5. 10	23. 57	19. 61	15. 65
5. 11	24. 61	20. 65	16. 69
5. 12	26, 64	21. 69	16. 74
5. 13	27. 68	22. 73	17. 78
5. 14	28. 72	22. 78	17. 83
5. 15	29. 76	23. 82	18. 87
5. 16	31. 79	24. 86	18. 92
5. 17	32. 83	25. 90	19. 96
5. 18	33. 87	26. 94	19. 101
5. 19	34. 91	27. 98	20. 105
5. 20	35. 95	28. 102	20. 110
5. 21	37. 98	29. 106	21. 114
5. 22	38. 102	29. 111	21. 119
5. 23	39. 106	30. 115	22. 123
5. 24	40. 110	31. 119	23. 127
5. 25	42. 113	32. 123	23. 132
6. 6	26. 52	23. 55	
6. 7	27. 57	24. 60	
6. 8	29. 61	25. 65	21. 69
6. 9	31. 65	26. 70	22. 74
6. 10	32. 70	27. 75	23. 79
6. 11	34. 74	28. 80	23. 85

Adapted from White [114].

Example

From a group of nine rats available for a study of the transfer of learning, five were selected at random and were trained to imitate leader rats in a maze. They were then placed together with four untrained control rats in a situation where imitation of the leaders enabled them to avoid receiving an electric shock. The results (the number of trials required to obtain ten correct responses in ten consecutive trials) were as follows:[1]

Trained rats:	78	64	75	45	82
Controls:	110	70	53	51	

Find the significance probability of these results when the Wilcoxon test is used.

[1] From Siegel, *Nonparametric Statistics*, McGraw-Hill Book Company, New York, 1956, p. 119. Original data from Solomon and Coles (1954). "A Case of Failure of Generalization of Imitation across Drives and across Situations." *J. Abnorm. Soc. Psychol.* 49:7–13.

Ranks: $n = 5,\ m = 4$

45, 51, 53, 64, 70, 75, 78, 82, 110

T, C, C, T, C, T, T, T, C

1 2 3 4 5 6 7 8 9

$$\text{Sum } T_x = 19$$

From table A7: $4, 5 \Rightarrow 11, 29$

$11 < T_x < 29 \Rightarrow$ not significant.

(p-value = 0.6349 from tables in Lehman)

(p-value for t-test: 0.881.)

Testing for Differences in Spread

Suppose that we observe two samples:

$$x_1 = \{7.6, 7.2, 9.7, 6.7, 10.0, 8.2, 5.5, 7.9, 9.5, 9.0\}$$

and

$$x_2 = \{11.4, 9.8, 10.5, 5.6, 14.1, 7.4, 7.0, 7.1, 12.4, 5.4\}.$$

Here we have $\overline{x}_1 = 8.1,\ s_1^2 = 2.1,\ s_1 = 1.4$, and $\overline{x}_2 = 9.1,\ s_2^2 = 9.0,\ s_2 = 3.0$

Do x_1 and x_2 differ in variance? To answer this question, we can use the *Siegal-Tukey test*, which is a nonparametric test for equality of spread in a population.

Again, we begin by ranking the observations, but in a different ordering, as we wish to track spread.

data	5.4	5.5	5.6	6.7	7.0	7.1	7.2	7.4	7.6	7.9	8.2	9.0	9.5	9.7	9.8	10.0	10.5	11.4	12.4	14.1
ranks	1	4	5	8	9	12	13	16	17	20	19	18	15	14	11	10	7	6	3	2
group	2	1	2	1	2	2	1	2	1	1	1	1	1	1	2	1	2	2	2	2

$$T_x = 72,\ E(T_x) = \frac{10 \times 21}{2} = 105,\ var(T_x) = \frac{10 \times 10 \times 21}{12} = 175$$

$$z = \frac{72 - 105}{\sqrt{175}} = -2.49,\ p = 0.013.$$

Thus we can conclude that the variances are different.

INFERENCE FOR POPULATION SPREAD* (Parametric, From Normal samples)

The two most basic descriptive features of a distribution are its center and spread. In a normal population, these aspects are measured by the mean and the standard deviation. We have described procedures for inference about population means for normal populations and found that these procedures are often useful for nonnormal populations as well. It is natural to turn next to inference about the standard deviations of normal populations. Our advice here is short and clear: Don't do it without expert advice.

There are indeed inference procedures appropriate for the standard deviations of normal populations. We will describe the most common such procedure, the F test for comparing the spread of two normal populations. Unlike the t procedures for means, the F test and other procedures for standard deviations are extremely sensitive to nonnormal distributions. This lack of robustness does not improve in large samples. It is difficult in practice to tell whether a significant F-value is evidence of unequal population spreads or simply evidence that the populations are not normal.

The deeper difficulty that underlies the very poor robustness of normal population procedures for inference about spread already appeared in our work on describing data. The standard deviation is a natural measure of spread for normal distributions, but not for distributions in general. In fact, because skewed distributions have unequally spread tails, no single numerical measure is adequate to describe the spread of a skewed distribution. Thus, the standard deviation is not always a useful parameter, and even when it is (in the normal case), the results of inference are not trustworthy. Consequently, we do not recommend use of inference about population standard deviations in basic statistical practice.[13]

Sometimes equality of standard deviations is tested as a preliminary to performing the pooled two-sample t test for equality of two population means. It is better practice to check the distributions graphically, with special attention to skewness and outliers. The pooled t test is reasonably robust against unequal population standard deviations, at least when the population distributions are roughly symmetric and the two sample sizes are similar. On the other hand, the test for equal standard deviations is often misleading because of its extreme sensitivity to departures from normality.

THE F STATISTIC AND F DISTRIBUTIONS

When s_1^2 and s_2^2 are sample variances from independent SRSs of sizes n_1 and n_2 drawn from normal populations, the F statistic

$$F = \frac{s_1^2}{s_2^2}$$

has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom when H_0: $\sigma_1 = \sigma_2$ is true.

1. Take the test statistic to be

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

This amounts to naming the populations so that s_1^2 is the larger of the observed sample variances. The resulting F is always 1 or greater.

2. Compare the value of F with critical values from Table F. Then *double* the significance levels from the table to obtain the significance level for the two-sided F test.

Using the data from the previous example,

$$H_0: \quad \sigma_1^2 = \sigma_2^2$$

and

$$H_A: \quad \sigma_1^2 \neq \sigma_2^2.$$

$$F = \frac{larger\ s^2}{smaller\ s^2} = \frac{9.0}{2.1} = 4.3.$$

From F-tables,

$$F_{n_1-1,n_2-1} = F_{9,9} \longrightarrow \; 2 \times 0.01 < p < 2 \times 0.025.$$

Thus H_0 can be rejected. However, we do not know if this rejection is due to the populations being non-normal, or whether the variances are truely different.

Degrees of freedom in the denominator

		Degrees of freedom in the numerator								
DFD	*p*	1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48

Confidence Intervals for Medians

Step 1: Form all possible **pairs** of numbers, so if there are n data points, there will be $\frac{n(n+1)}{2}$ possible pairs.

Step 2: Take the **mean** of each of these pairs of numbers (so you should calculate $\frac{n(n+1)}{2}$ means in this step).

The intuition is that each of these means is an estimate of the median, although, of course, some will fall above, and some will fall below. Ordering all of the means from low to high values, we then need to remove the 2.5% that are too small and the 2.5% that are too large, to result in a 95% confidence interval. How does one know how many to remove from the top and bottom?

Step 3: Look up Table A6 (on page 161). The **first** number listed under the heading 0.05 tell you how many need to be removed from each side. For example, with $n = 15$, one would have 105 pairs of means to calculate, and from Table A6 we would remove the lowest 25 pairs and the highest 25 pairs. The lowest and highest **remaining** means form the 95% CI for the median. A good point estimate of the median is the overall means of **all** of the pairs.

Example: Calculate a 95% CI for the median using the following set of 7 numbers: 3, 9, 14, 10, 5, 7, 15.

Steps 1 and 2 can be represented as follows (28 pairs):

	3	5	7	9	10	14	15
3	3	4	5	6	6.5	8.5	9
5		5	6	7	7.5	9.5	10
7			7	8	8.5	10.5	11
9				9	9.5	11.5	12
10					10	12	12.5
14						14	14.5
15						14	15

From Table A6, with 7 numbers, we need to throw away the two highest and two lowest means in the table for a 95% CI. The two highest means are 15 and 14.5, and the two lowest are 3 and 4. After these are removed, a 95% CI formed from ordering the reming ones and looking at the highest and lowest values is (5, 14). (You can check that the usual 95% for the mean here is (4.9, 13.1), agreeing quite closely without result for the median).

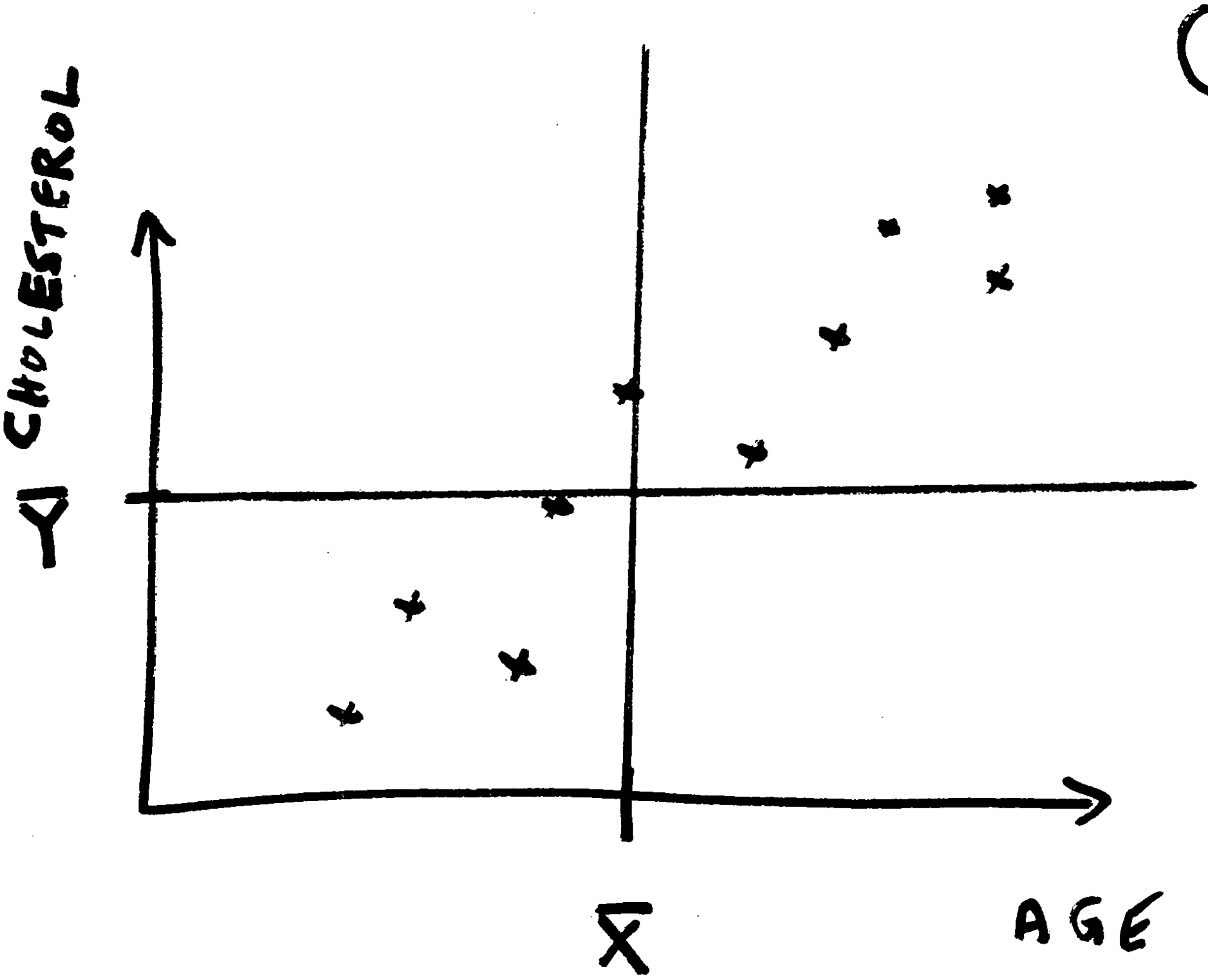
Computers are usually used for these calculations, and there are also normal approximations.

DR. GRANT
NAU MATH U.
GOLFIN'! HAWAII! TOURS
RESULTS
0?-00?
?---0-
0-?
IMPORTANT!
IMP
ORTANT
C. PARE

THIS WILL JUST TAKE A MINUTE!
WHAT TEST DO I USE...

Correlation and Regression

- Are age and cholesterol related to each other?
- How can we measure the strength of such a relationship?
- Given some data, how can we estimate this measure?
- Can we predict the average cholesterol of persons aged 50 years old? How accurately?
- Can we predict an individuals cholesterol, given that his/her age is 50 years old? How accurately?



As X increases, Y increases

Suggested Measure:

$$\text{Covariance} = \sum_{\substack{\text{all} \\ (X_i, Y_i)}} (X_i - \bar{X})(Y_i - \bar{Y}) \Big/ N-1$$

To prevent <u>scale</u> problems, prefer correlation to covariance:

<u>Population:</u>

square root

$$\rho = \frac{\sum_{\substack{\text{all} \\ \text{population}}} (X_i - \mu_x)(Y_i - \mu_Y)}{\left[\sum_{\substack{\text{all} \\ \text{population}}} (X_i - \mu_X)^2 \sum_{\substack{\text{all} \\ \text{population}}} (Y_i - \mu_Y)^2 \right]^{1/2}}$$

<u>Sample:</u>

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2 \right]^{1/2}}$$

$$-1 \leq \rho \leq +1$$

$$-1 \leq r \leq +1$$

4

Example

Exposure (X) Time (min.)	5	10	15	20	25
Number (Y) of Surviving Bacteria (plate counts)	300	210	190	160	140

$\bar{X} = 15.0$ $\bar{Y} = 200$

$sd(x) = S_x = 7.9$ $sd(Y) = S_Y = 62.0$

$$r = \frac{\left[(5-15)(300-200) + (10-15)(210-200) + (15-15)(190-200) + (20-15)(160-200) + (25-15)(140-2\right.}{[7.9 \times 62.0][5-1)}$$

$$= \frac{1850}{(7.9 \times 62.0)4} = -0.94$$

- Prediction for 17 minutes?
- Prediction for 2 minutes?
- Prediction for 35 minutes?

4 Equivalent Correlation Formuli

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\ \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

OR

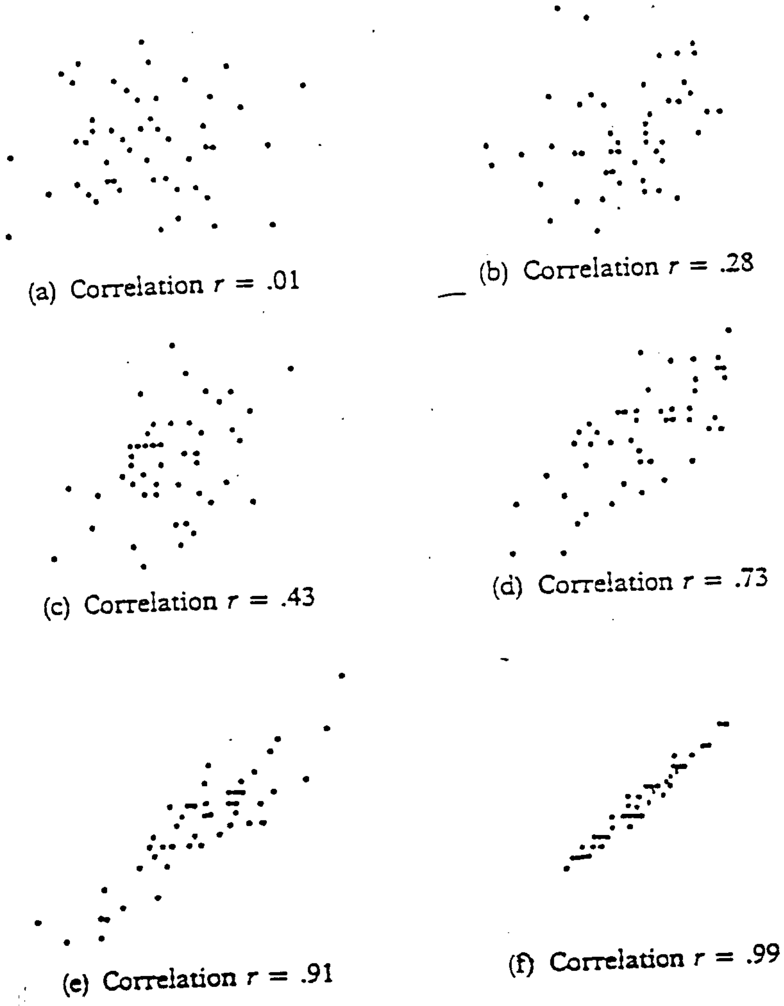
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y \left[n-1\right]}, \text{ where } \begin{cases} S_X = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \\ S_Y = \sqrt{\dfrac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}} \end{cases}$$

OR

$$r = \frac{\sum X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{(n-1) S_X S_Y}$$

Easiest to calculate, as step involving subtracting the mean is not required.

OR $r = b\,\dfrac{S_X}{S_Y}$, where b is the estimated slope of the regression line.



(a) Correlation r = .01

(b) Correlation r = .28

(c) Correlation r = .43

(d) Correlation r = .73

(e) Correlation r = .91

(f) Correlation r = .99

176

The Balloon Rules for a Rough Estimate of the Correlation Coefficient

GUY CHÂTILLON*

By surrounding the sample plot with a kind of birthday balloon, we can guess the value of Pearson's correlation coefficient, r, in a very simple way.

KEY WORDS: Correlation coefficient; Graphical guessing; Sample plot.

1. INTRODUCTION

When discussing the value of Pearson's correlation coefficient, r, it is important to have an idea of the shape of the sample plot. Values of r close to ±1 or 0 are easy to guess. But is it possible, between these extremes, to obtain an intuitive idea of the value of r without calculating it? The aim of this article is to show that such a possibility exists. Since Pearson's r is well known, there is no need to define it here.

2. THE BALLOON IDEA

The basic idea consists of surrounding the sample plot with a kind of "birthday balloon" that is in fact an ellipse. But let us apply this method to an example taken from a well-known volume by Hoel (1971). The sample plot from page 189 of Hoel's book is reproduced in Figure 1.

First, we draw the balloon so as to surround all or most of the points and to fit the plot. Second, we measure the vertical height of the balloon at its center, h, and its vertical height at the extremes, H. Then we compute the formula

$$F = \sqrt{1 - \left(\frac{h}{H}\right)^2}.$$

If the points inside the balloon are "well distributed," then the result of the computation usually gives a fairly good idea of the value of Pearson's correlation coefficient.

For example, with a desk rule, I measured $h \simeq 5.4$ cm and $H \simeq 7.4$ cm in Figure 1. (Perhaps the scale of this figure has been changed in the published version.) So I had

$$F \simeq \sqrt{1 - \left(\frac{5.4}{7.4}\right)^2} \simeq \sqrt{1 - .73^2} \simeq .68.$$

Another way to proceed is by counting the number of little squares. We have

$$h \simeq 16.7 \text{ squares}$$

$$H \simeq 23 \text{ squares}$$

$$\Rightarrow \qquad F \simeq \sqrt{1 - \left(\frac{16.7}{23}\right)^2} \simeq .69.$$

In fact, the exact value of the correlation coefficient is $r \doteq .6317$. So our guess is quite good.

3. THEORETICAL JUSTIFICATION

The formula $F = \sqrt{1 - (h/H)^2}$ may be justified on the basis of two different models.

In the first case, we suppose that the actual sample plot is a particular realization of a pair of random variables (X, Y) whose joint density is the bivariate normal density. In this case the ellipse tends to be close to a level curve of the bivariate normal. We can consider the level curves of the bivariate normal with $\mu_x = \mu_y = 0$ and $\sigma_x = \sigma_y = 1$ since correlations are not affected by changes in scale or by translation. The level curves of this bivariate normal are given by $x^2 - 2\rho xy + y^2 = K(1 - \rho^2)$, where $K > 0$ and ρ is the correlation coefficient of the model. From this, we can obtain $y = \rho x \pm [(K - x^2)(1 - \rho^2)]^{1/2}$. If the balloon is one of these level curves, h is the difference between the two values of y when x is zero, or $h = 2[K(1 - \rho^2)]^{1/2}$. The maximum

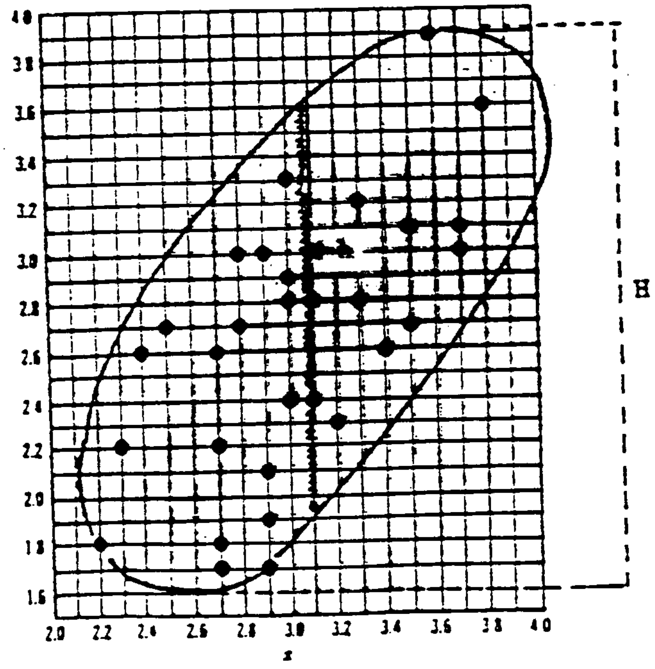
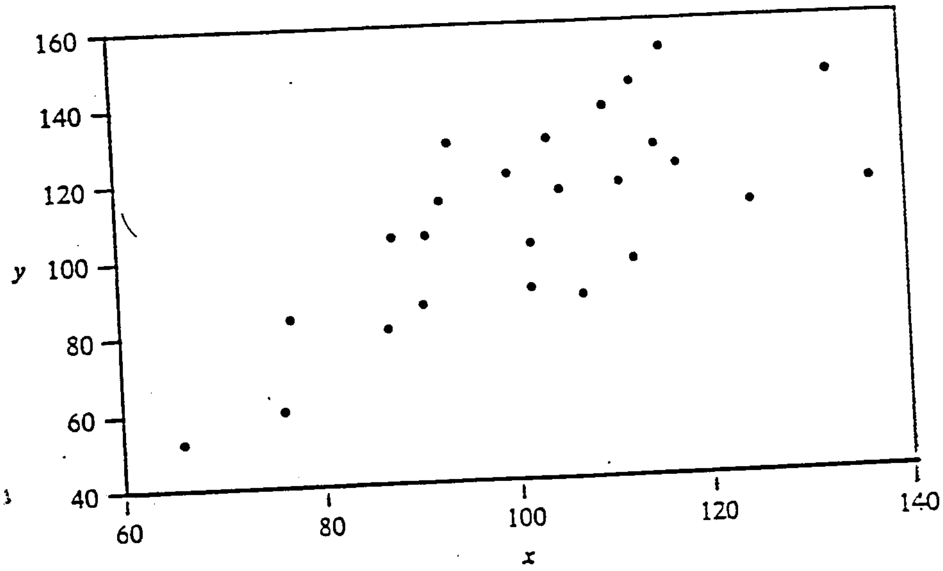
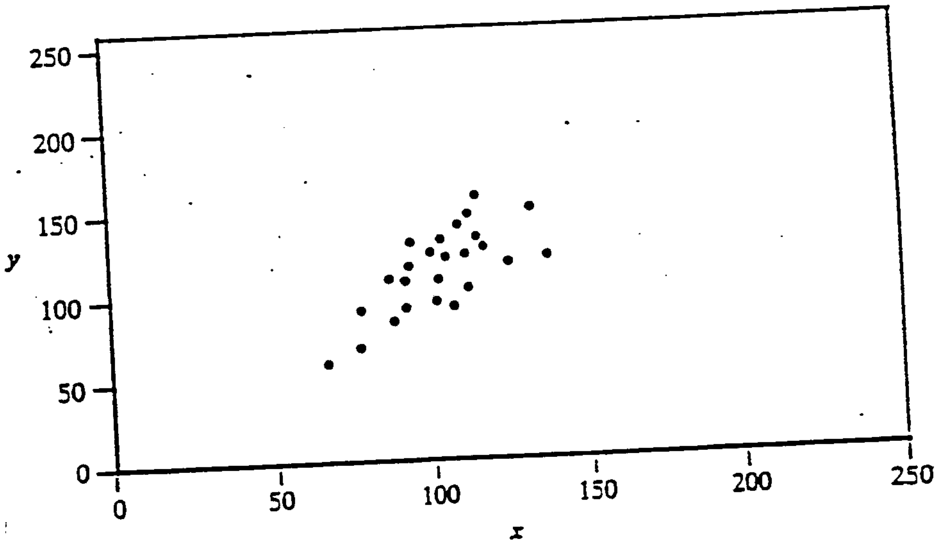
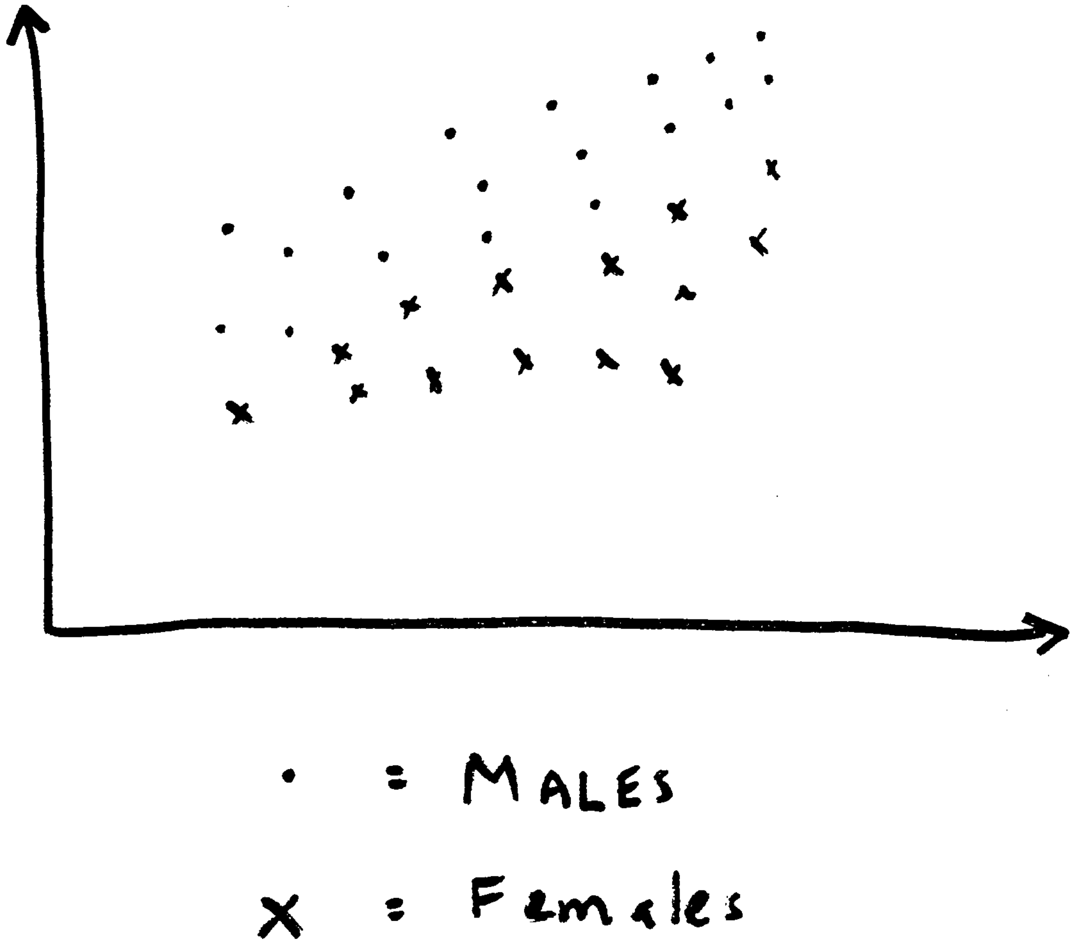


Figure 1. Scatter diagram for grade point averages.

*Guy Châtillon is Professor of Statistics, Department of Mathematics, Université du Québec à Trois-Rivières, Québec, Canada G9A 5H7.


160
140
120
100
80
60
40
y
60
80
100
120
140
x

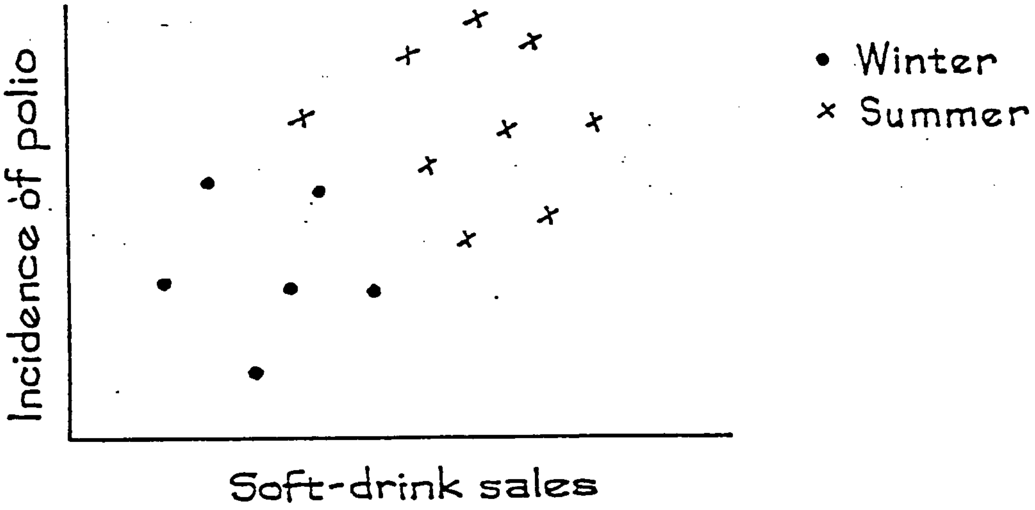

250
200
150
100
50
0
y
0
50
100
150
200
250
x



$r_{\cdot} = ?$

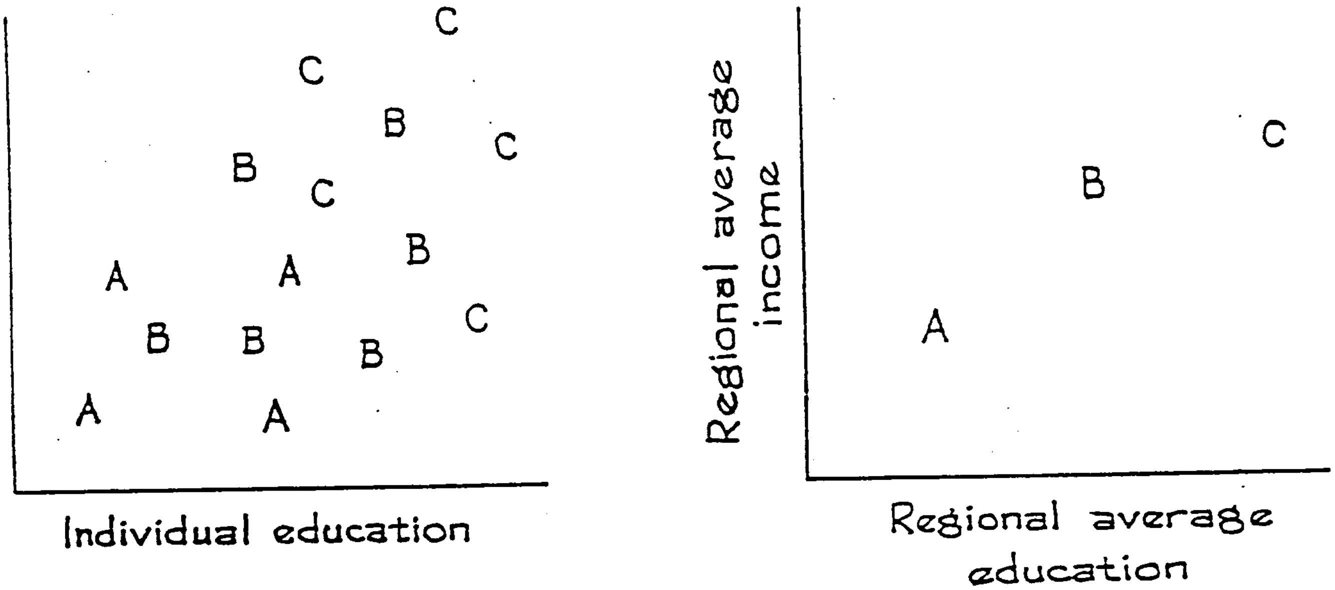
$r_{x} = ?$

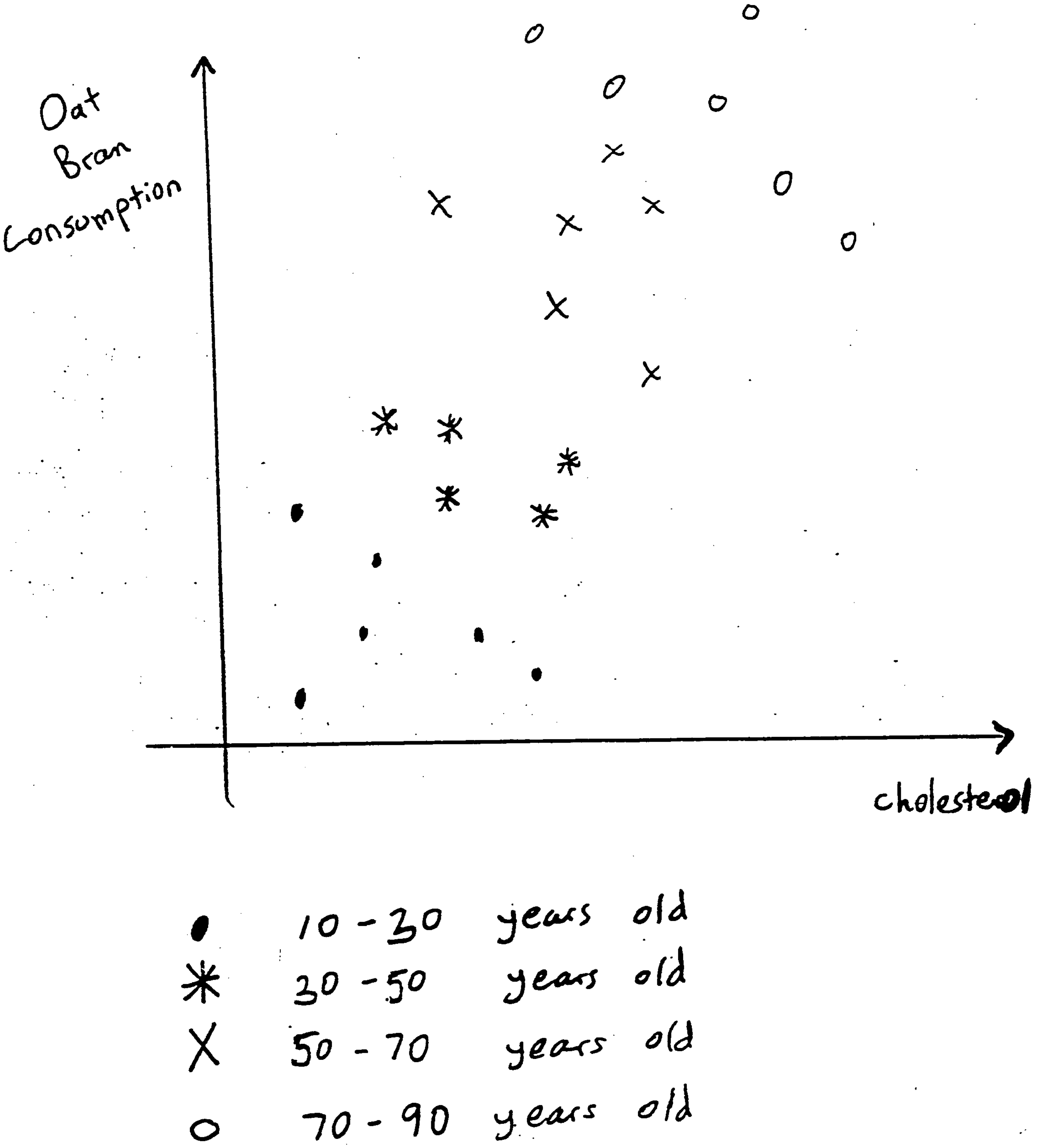
$r_{x \text{ and } \cdot} = ?$



Correlation measures association. But association is not the same as causation.

igure 3. Correlations based on rates or averages are often too big. The anel on the left represents income and education for individuals in three eographic regions, labeled A, B, C. Each individual is marked by the tter showing his region of residence. The correlation is moderate. The anel on the right shows the averages for each region: the correlation etween the averages is almost 1.

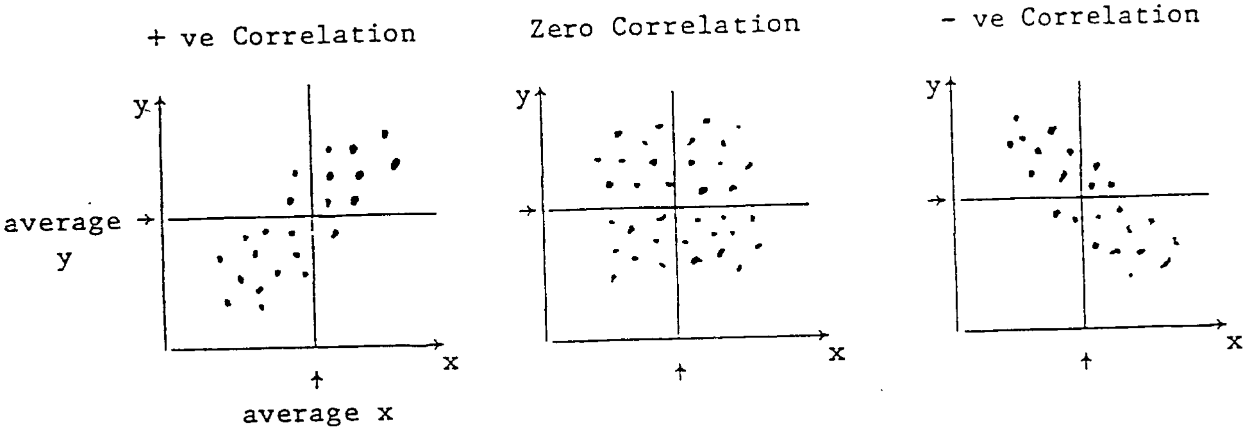



Oat Bran Consumption
cholesterol
● 10 - 30 years old
✳ 30 - 50 years old
X 50 - 70 years old
○ 70 - 90 years old

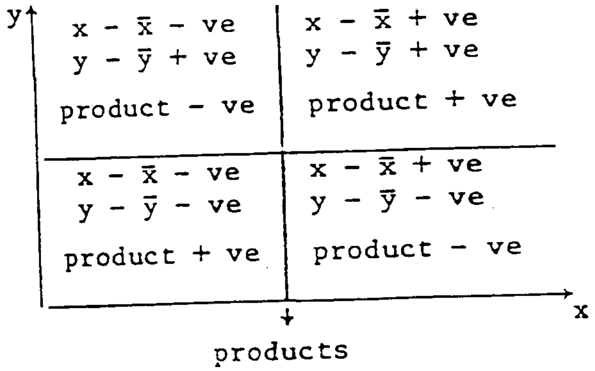
Positive Correlation: Large X's with large Y's and small X's with small Y's

Negative Correlation: Large X's with small Y's and small X's with large Y's

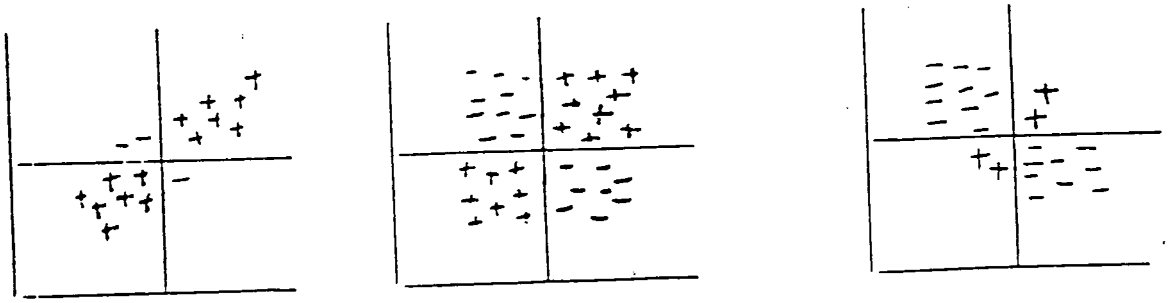
No Correlation: Large X's not more likely to be paired with large Y's than small Y's.



How r ranges from -1 (neg correlation) thru 0 (zero correlation) thru +1 (positive correlation) (r not tied to x or y scale).



$x - \bar{x}$ − ve $y - \bar{y}$ + ve product − ve	$x - \bar{x}$ + ve $y - \bar{y}$ + ve product + ve
$x - \bar{x}$ − ve $y - \bar{y}$ − ve product + ve	$x - \bar{x}$ + ve $y - \bar{y}$ − ve product − ve



Inferences on ρ [using a sample of n (x, y) pairs]

Naturally, the observed r in any particular sample will not exactly match the ρ in the population (i.e. the coefficient one would get if one included everybody in a census). The quantity r would vary from one sample of n to another sample of n. i.e. r is subject to sampling fluctuation about ρ .

1. The most common question asked of one's data is whether there is evidence of a non-zero correlation between 2 variables. To test this, we can set up a null hypothesis H_o that ρ in the population is zero and measure whether our observed r is too discrepant from $\rho=0$ to be just sampling fluctuation.

   This discrepancy of r from zero is measured as

   $$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ and should (if H_o is true) follow

   a student's t distribution with n-2 d. of fr.

   [Colton's table A5 gives the smallest r which would be considered evidence that $\rho \neq 0$. E.g. if n=20, d. of fr. = 18 an observed correlation of 0.444 or higher or between -0.444 and -1 would be considered statistically significant at the p = 0.05 level (2 sided).
   Note: test involves assumption of Bivariate normal distribution.

2. Another common question is: given that r is based only on a sample, what confidence interval should I put around r so that there is a good chance (say 95%) that the interval will include the "true" coefficient ρ .

   A related question, which can use the same technique to answer it is: I observe a certain r_1 ; somebody else observes another value r_2. Are the ρ's in the 2 populations we are studying comparable?

1. The following transformation of r will lead to a statistic which is approximately normal even if ρ is quite a ways from 0:

$$Z = \frac{1}{2}\,\ell n\,\frac{1+r}{1-r} \qquad \left[\begin{array}{l}\ell n = \text{log to base e}\\ \quad = \text{natural log}\end{array}\right]$$

Z is known as Fisher's 'r to Z' transformation;the calculated Z should be compared against a Gaussian distribution with

$$\text{mean} = \frac{1}{2}\ell n\left[\frac{1+\rho}{1-\rho}\right]$$

$$\text{SD} = \sqrt{\frac{1}{n-3}}$$

E.g. $H_o : \rho = \frac{1}{2}$

observe $r = 0.4$ in sample of size $n = 20$

To test if H_o is true

Compute $$\frac{Z - \text{mean}}{\text{SD}} = \frac{\frac{1}{2}\,\ell n\left[\frac{1.4}{0.6}\right] - \frac{1}{2}\,\ell n\left[\frac{1.5}{0.5}\right]}{\sqrt{\frac{1}{17}}}$$

and compare with Gaussian Tables.

Unusually extreme values are evidence against H_o.

2. <u>Confidence Intervals</u> Solve

$$Z - Z_{\alpha/2}\sqrt{\frac{1}{n-3}} \le \frac{1}{2}\,\ell n\,\frac{1+\rho}{1-\rho} \le Z + Z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Example

Suppose that $r = 0.5$ with a sample size of $n = 30$.

a) Derive a 95% confidence interval for ρ, the population correlation coefficient

b) Test the null hypothesis

$$H_0: \rho = 0$$

versus $\quad H_A: \rho \neq 0.$

a) $$Z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right] = \frac{1}{2}\ln\left[\frac{1.5}{0.5}\right] = \frac{1}{2}\ln 3 = 0.549$$

$$\therefore\ .549 - 1.96\sqrt{\frac{1}{27}} \le \frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] \le .549 + 1.96\sqrt{\frac{1}{27}}$$

$$.172 \le \frac{1}{2}\ln\left[\frac{1+\rho}{1-\rho}\right] \le .9265$$

$$.344 \le \ln\left[\frac{1+\rho}{1-\rho}\right] \le 1.85$$

$$1.41 \le \frac{1+\rho}{1-\rho} \le 6.36$$

$$0.170 \le \rho \le .728$$

b)

$$H_0 : \rho = 0$$

$$H_A : \rho \neq 0$$

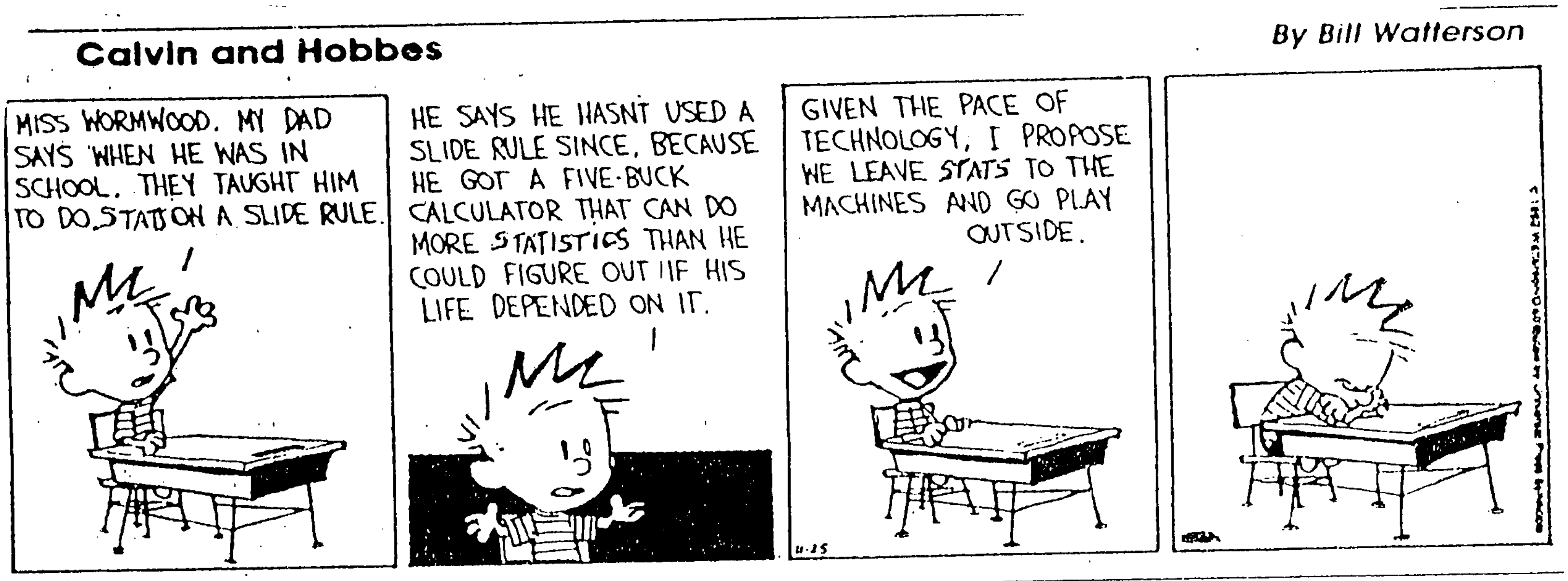
$Z = 0.549$, as before

$$\text{mean} = \frac{1}{2} \ln \left| \frac{1+0}{1-0} \right| = \frac{1}{2} \ln(1) = 0$$

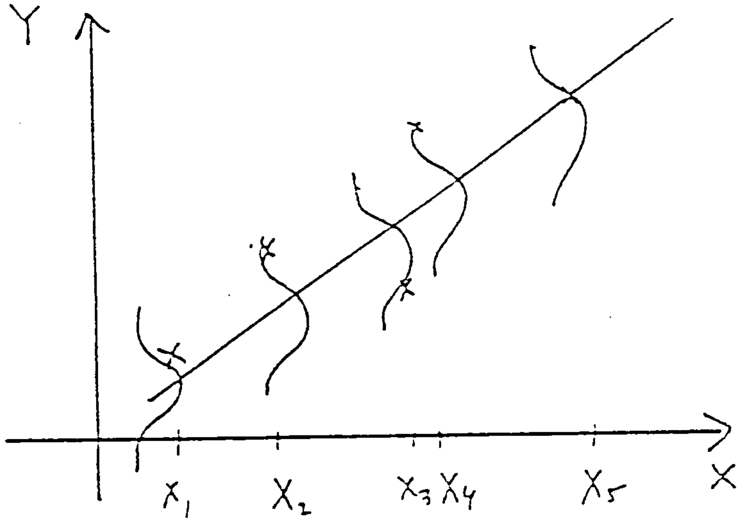
Then: $$Z = \frac{0.549 - 0}{\sqrt{\frac{1}{27}}} = 2.85$$

From Normal Tables, $p = 0.0043$

Regression


Calvin and Hobbes
By Bill Watterson
MISS WORMWOOD. MY DAD SAYS WHEN HE WAS IN SCHOOL, THEY TAUGHT HIM TO DO STATS ON A SLIDE RULE.
HE SAYS HE HASNT USED A SLIDE RULE SINCE, BECAUSE HE GOT A FIVE-BUCK CALCULATOR THAT CAN DO MORE STATISTICS THAN HE COULD FIGURE OUT IF HIS LIFE DEPENDED ON IT.
GIVEN THE PACE OF TECHNOLOGY, I PROPOSE WE LEAVE STATS TO THE MACHINES AND GO PLAY OUTSIDE.
11-25

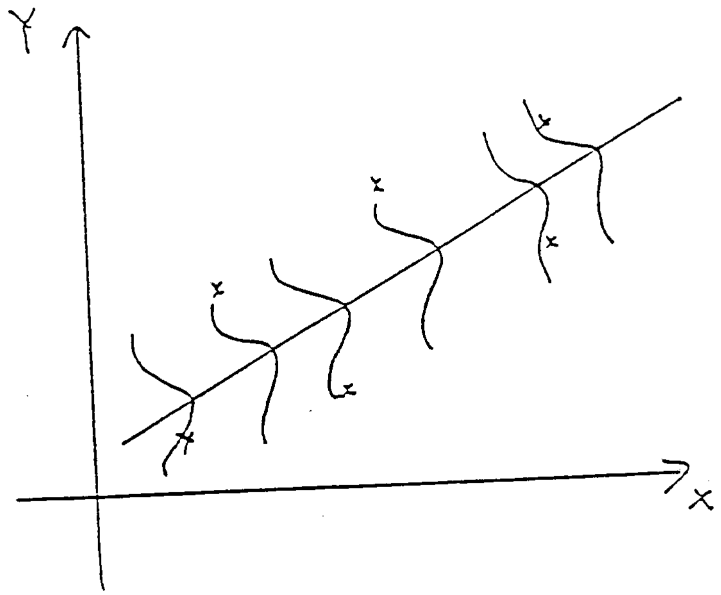
Regression



Model: $Y = \alpha + \beta X + \text{"error"}$

$\boxed{\text{OR}}$ $E(Y) = \alpha + \beta X$

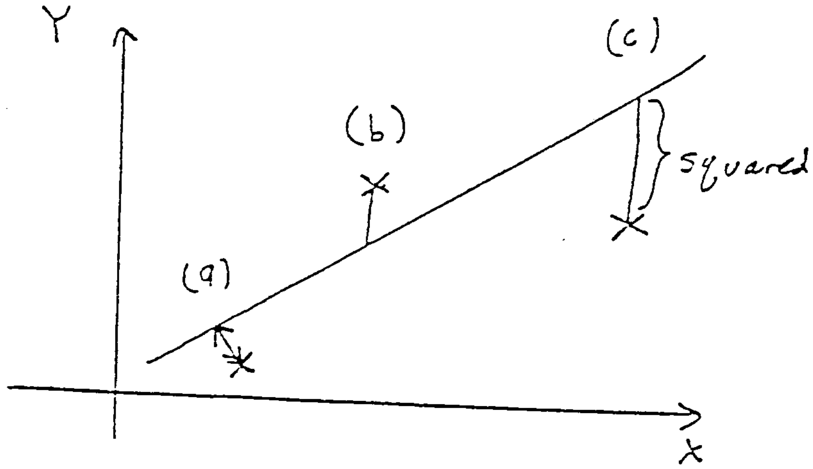
Assumptions:
- "error" is $N(0, \sigma^2)$
- σ^2 is the same throughout the range
- Relationship between X and Y is a straight line.



True values: $Y = \alpha + \beta x +$ "error"

Fitted values: $Y = a + b\underset{\sim}{x} +$ residual.

By what criterion do we judge what is the best estimated line, i.e., best "a" and "b" to choose?



(a) Line with smallest

$\sum_{\text{all pts}}$ (distance to line) ?

(b) Line with smallest

$\sum_{\text{all pts}}$ (vertical distance to line) ?

(c) Line with smallest

$\sum_{\text{all pts}}$ (vertical distance to line)^2 ?

Using calculus + algebra, find:

$$a = \hat{\alpha} = \bar{Y} - b\bar{X}, \text{ where}$$

$$b = \hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

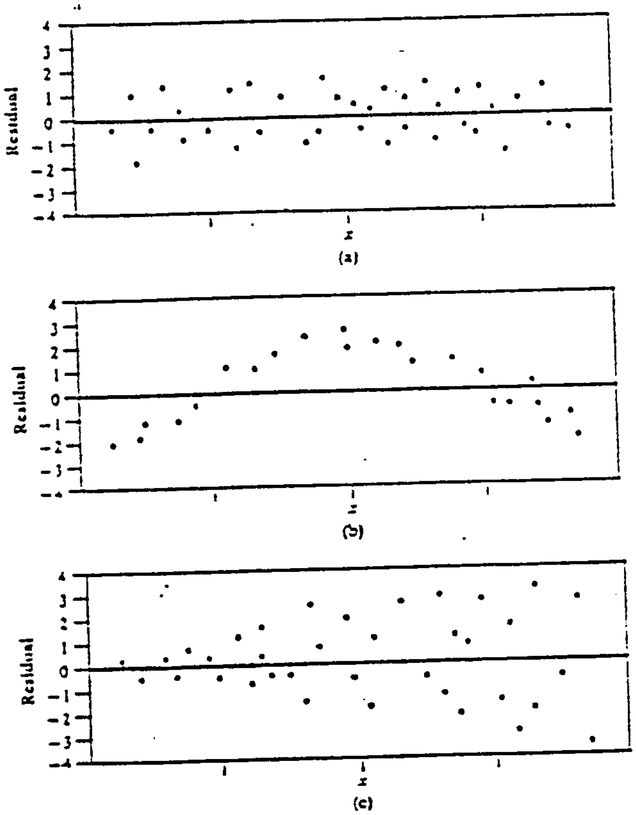
$\boxed{\text{OR}}$
$$\hat{\beta} = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

"Bonus:" Can also estimate σ^2, the assumed "common variance".

$$\hat{\sigma}^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

where $\hat{y}_i$ = "predicted y_i" $= a + b x_i$

Patterns of Residuals



Are assumptions satisfied?

- Normal Residuals
- variance constant ~~throughout~~ range
- straight line fits well.

Inference for Regression Parameters

Confidence intervals are usually of the form

$$\text{estimate } \pm \left\{ \begin{array}{c} z \\ t \end{array} \right\} s.d.(estimate)$$

For example, we have seen

$$\overline{x} \pm \left\{ \begin{array}{c} z \\ t \end{array} \right\} s.d.(\overline{x})$$

or

$$\overline{x} \pm \left\{ \begin{array}{c} z \\ t \end{array} \right\} \frac{s \text{ or } \sigma}{\sqrt{n}}$$

The same basic formulation is followed for inferences for regression parameters, such as α, β, or even when making predictions for future observations,

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \times x_i = a + b \times x_i.$$

Since we already have point estimates for each of these items, all we are missing are the standard deviations, and what values of t or z to use.

Standard Error (Standard Deviation) Formulae

$$SE(\hat{\alpha}) = SE(a) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\Sigma_{i=1}^{n}\,(x_i - \overline{x})^2}}$$

$$SE(\hat{\beta}) = SE(b) = \frac{\sigma}{\Sigma_{i=1}^{n}\,(x_i - \overline{x})^2}$$

$$SE(\text{predicted MEAN at } x) = \sigma\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\Sigma_{i=1}^{n}\,(x_i - \overline{x})^2}}$$

$$SE(\text{predicted INDIVIDUAL at } x) = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\Sigma_{i=1}^{n}\,(x_i - \overline{x})^2}}$$

Problem: We usually do not know σ.

Solution: Estimate σ by

$$\begin{aligned}\hat{\sigma} &= \sqrt{\frac{\text{Residual sum of squares (RSS)}}{n-2}}\\ &= \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - \text{predicted}(y_i))^2}{n-2}}\\ &= \sqrt{\frac{\Sigma_{i=1}^{n}(y_i - [a + b \times x_i])^2}{n-2}}\end{aligned}$$

Confidence Intervals (...and Tests)

Now that we know the standard errors, confidence intervals are easy to compute:

- CI for α: $\hat{\alpha} \pm t_{1-\alpha/2,n-2} \times SE(\hat{\alpha})$
- CI for β: $\hat{\beta} \pm t_{1-\alpha/2,n-2} \times SE(\hat{\beta})$
- CI for predicted mean: $\hat{y}_i \pm t_{1-\alpha/2,n-2} \times SE(\text{predicted MEAN at } x)$
- CI for predicted mean: $\hat{y}_i \pm t_{1-\alpha/2,n-2} \times SE(\text{predicted INDIVIDUAL at } x)$

[Not that we would likely ever want to test after we know the CI, but for completeness, tests of hypotheses about α and β can be similarly constructed:]

To test $H_0: \ \alpha = \alpha_0$, use the fact that

$$\frac{\alpha - \alpha_0}{SE(\hat{\alpha})} \sim t_{n-2},$$

and similarly for β:

To test $H_0: \ \beta = \beta_0$, use the fact that

$$\frac{\beta - \beta_0}{SE(\hat{\beta})} \sim t_{n-2}.$$

Regression Example

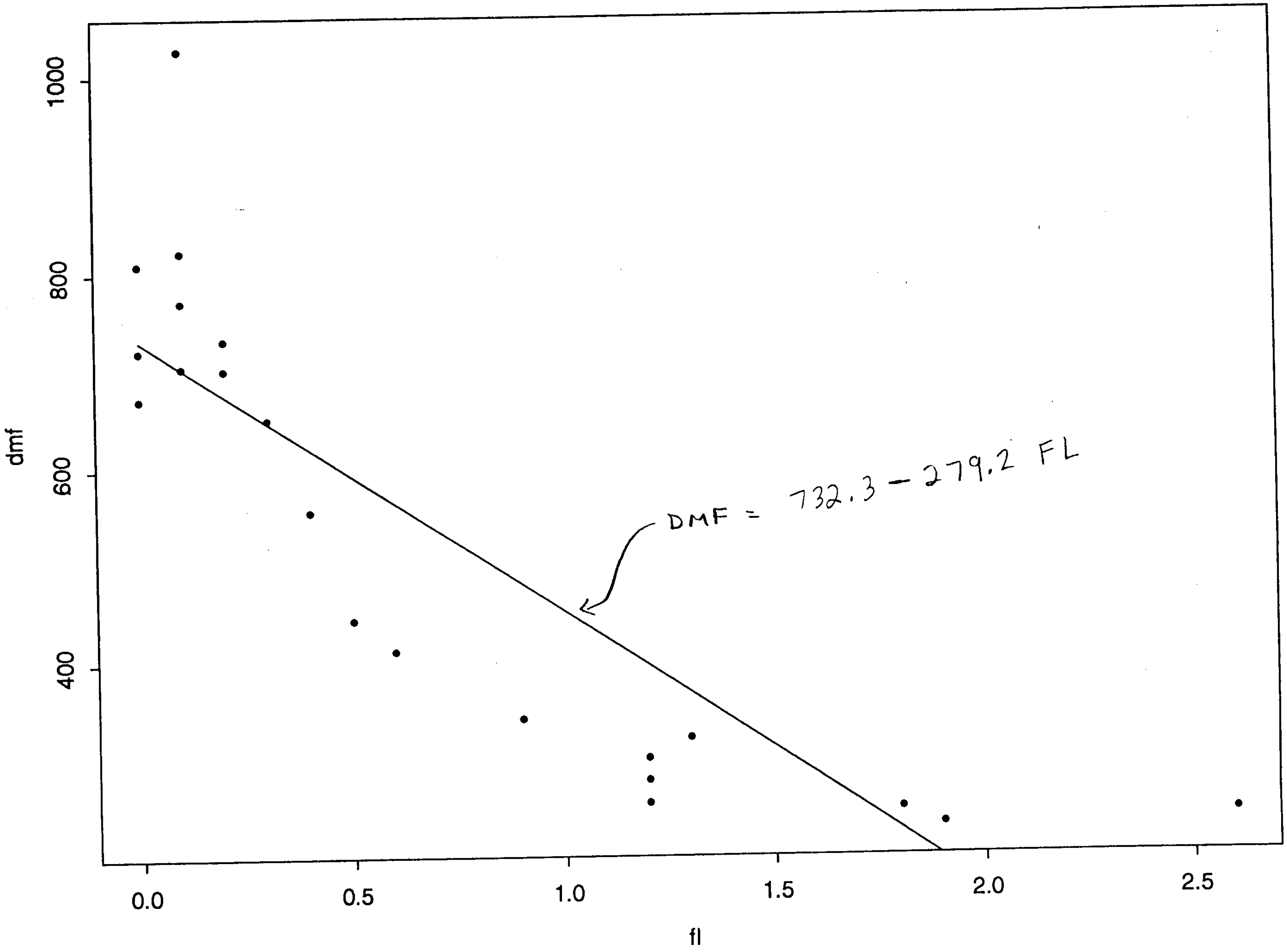
The data on the next page show the caries experience of 7257 children 12–14 years old in 21 communities according to the fluoride concentration of their public water supply. DMF denotes "Decayed, Missing or Filled."

(a) Draw a rough scatter plot to visually examine the association between DMF teeth and fluoride.
(b) Calculate the regression line of DMF teeth on fluoride concentration.
(c) What is the estimate of σ, the residual standard deviation? Calculate a 95% confidence interval for the slope, β of the line calculated in (b).
(d) Give your prediction for the average number of DMF teeth there would be in a community with a fluoride concentration of 1.8 ppm.
(e) What is the 95% confidence interval around your answer in (d)?
(f) Examin the graph of the residuals. Does a linear regression seem appropriate? Even if some other model may be more appropriate, has the linear regression been useful in examining the data?

Community Number	DMF per 100 children	Fluoride Concentration in ppm
1	236	1.9
2	246	2.6
3	252	1.8
4	258	1.2
5	281	1.2
6	303	1.2
7	323	1.3
8	343	0.9
9	412	0.6
10	444	0.5
11	556	0.4
12	652	0.3
13	673	0.0
14	703	0.2
15	706	0.1
16	722	0.0
17	733	0.2
18	772	0.1
19	810	0.0
20	823	0.1
21	1027	0.1

a) Scatter Plot with regression line drawn in.

Fluoride Concentration versus DMF Teeth



(b) $$\hat{\beta} = \frac{\sum_{i=1}^{21} \overset{(FL)}{(X_i - \bar{X})} \overset{(DMF)}{(Y_i - \bar{Y})}}{\sum (x_i - \bar{x})^2}$$

$$= -279.20$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\,\bar{X}$$

$$= 536.9 + 279.2\,(0.7)$$

$$= 732.34$$

Thus the regression line is

$$DMF = 732.34 - 279.20\,(FL)$$

(c) $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{21} (Y_i - \hat{Y}_i)^2}{n-2},$$

Y_i = observed value

$\hat{Y}_i$ = predicted value, from reg. line in (b).

$$= 16210.77,$$

and $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = 127.32$

95% confidence interval for the slope, β, is then

$$\hat{\beta} \pm t_{.025,19} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$= -279.20 \pm (2.09) \frac{127.32}{3.334}$$

$$= -279.20 \pm 38.18\,(2.09)$$

$$\Rightarrow (-359.11, -199.29)$$

(d) $\text{DMF} = 732.34 - 279.20\,(1.8)$

$$= 229.79$$

(e) 95% confidence interval for the predicted value in (d) is

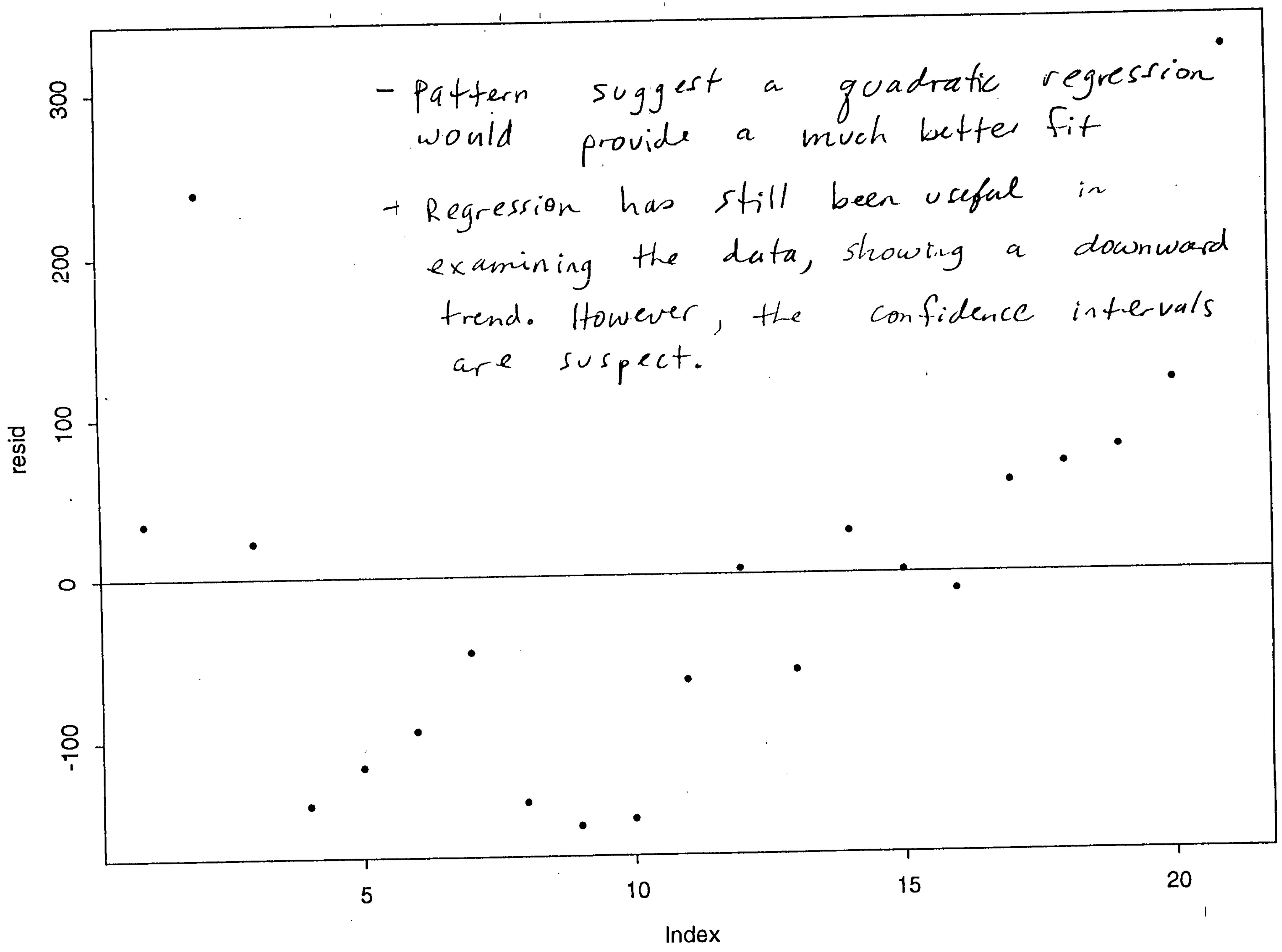
$$229.79 \pm t_{0.025,19}\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(\overset{1.8}{X_i} - \overset{0.7}{\bar{X}})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

$$= 229.79 \pm (50.35)(2.09)$$

$$\Rightarrow (124.39, 335.18)$$

(f)

Residuals from the Fluoride DMF example



- Pattern suggest a quadratic regression would provide a much better fit

+ Regression has still been useful in examining the data, showing a downward trend. However, the confidence intervals are suspect.

```
> dmf<-c(236,246,252,258,281,303,323,343,412,444,556,652,673,703,706,722,733,772,
> fl<-c(1.9,2.6,1.8,1.2,1.2,1.2,1.3,0.9,0.6,0.5,0.4,0.3,0.0,0.2,0.1,0.0,0.2,0.1,0
> length(dmf)
[1] 21
> length(fl)
[1] 21
> mean(dmf)
[1] 536.9048
> mean(fl)
[1] 0.7
> fldiff <- fl - mean(fl)
> dmfdiff <- dmf - mean(dmf)
> betahat <- sum(fldiff*dmfdiff)/(var(fl)*20)
> betahat
[1] -279.1996
> alphahat <- mean(dmf) - betahat * mean(fl)
> alphahat
[1] 732.3445
> regline <- function(fl) {return(732.3445 - 279.1996*fl)}
> sigmahat <- sum((dmf - regline(fl))^2)/19
> sigmahat
[1] 16210.77
> sqrt(sigmahat)
[1] 127.3215
> sebeta<-sqrt(sigmahat)/sqrt(sum(fldiff^2))
> sebeta
[1] 38.18119
> betaupper <- betahat + sebeta*qt(0.975,df=19)
> betalower <- betahat - sebeta*qt(0.975,df=19)
> betalower
[1] -359.1138
> betaupper
[1] -199.2855
> resid <- dmf - regline(fl)
> resid
 [1]   34.13474  239.57446   22.21478 -139.30498 -116.30498  -94.30498
 [7]  -46.38502 -138.06486 -152.82474 -148.74470  -64.66466    3.41538
[13]  -59.34450   26.49542    1.57546  -10.34450   56.49542   67.57546
[19]   77.65550  118.57546  322.57546
> sepredmean <- sqrt(sigmahat)*sqrt(1/21 + ((1.8 - mean(fl))^2/sum(fldiff^2)))
> sepredmean
[1] 50.35756
> 229.7852 + sepredmean * qt(.975,df=19)
[1] 335.1848
> 229.7852 - sepredmean * qt(.975,df=19)
[1] 124.3856
```

Calculations using Splus.

(Could also have just typed data + one line!)

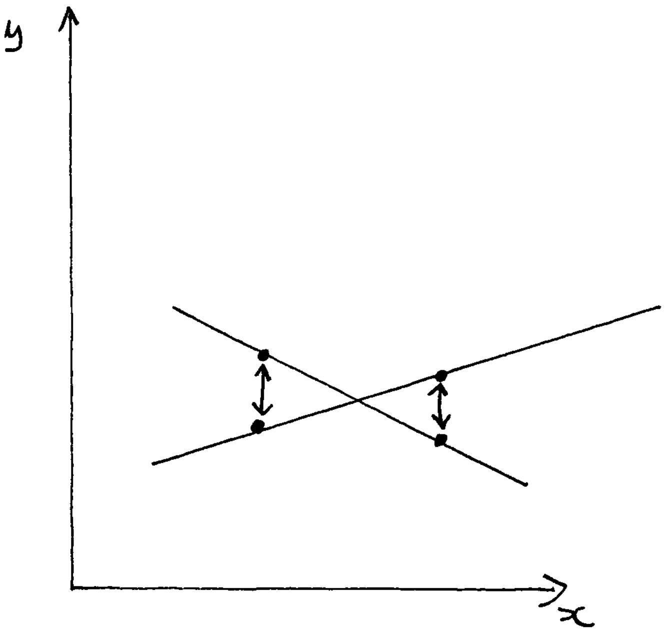
Intuitive Optimal Design for Linear Regression

Suppose we are designing an experiment that will be analysed as a linear regression between two variables, say x and y. We will choose values of x to observe (so we "have control" over values of x), and for each value of x we choose, we will observe the value of y that is associated with it.

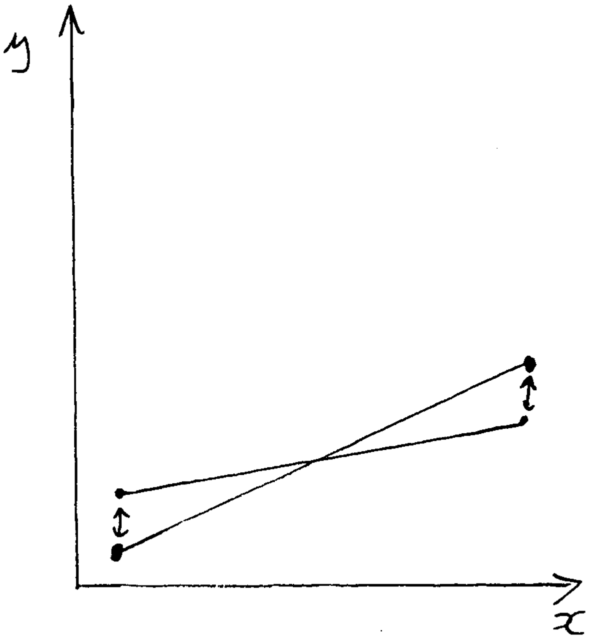
Question: What values of x should we choose in order the "most accurately" estimate the linear regression line?

Points to Consider:

1. Values of x that are further apart from each other tend to produce more stable estimates. Consider the following graphs:



small changes in y values produce large changes in slope.



same changes in y values produce smaller changes in slope

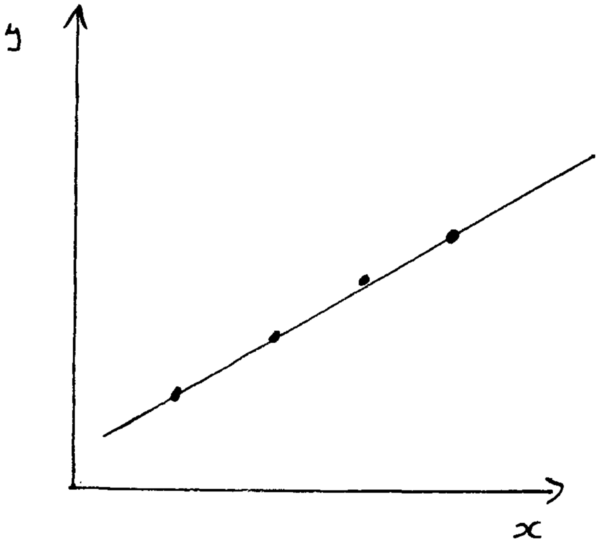
Intuitive Optimal Design for Linear Regression

2. As always, a larger sample size leads to more accurate estimation, since the α and β coefficients will be estimated more accurately.

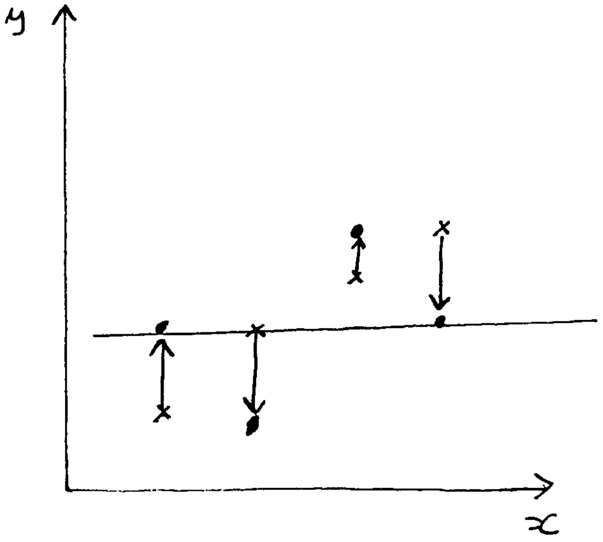
3. If there is a chance that the relationship is not exactly linear (almost always the case), then selecting x values spread out along the feasible range will allow you to explore linearity.

4. Consider taking some repeated measurements at some x values, to allow for investigation of whether σ^2 changes with x.

5. Try to measure x and y as accurately as possible, since **measurement error** can severly compromise your attemps to accurately determine the relationship between x and y. Consider the graphs below:



True relationship seems to be a near perfect straight line



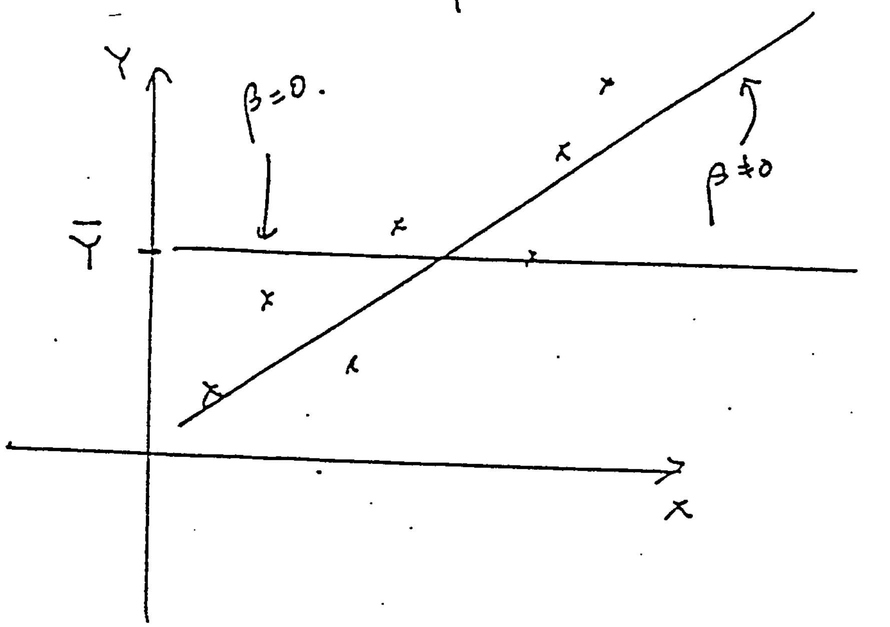
True relationship (x's) becomes obscured by measurement error (•'s).

Comparing Correlation to Linear Regression

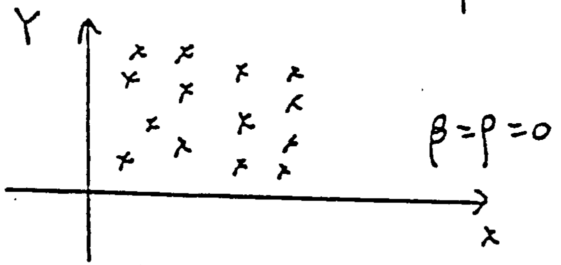
1. Both investigate linear relationships between two variables.
2. A scatter plot is useful in interpreting both corrlation values and regression coefficients.
3. Both can be used descriptively or inferentially.
4. Both try to "explain" the uncertainty of one variable in terms of another.
5. In correlation, there is no distinction between x and y variables, while in regression x and y are used differently (one considers predicting y from x).

One can also define a nonparametric correlation, often called **Spearman's correlation coefficient**. To calculate a Spearman's coefficient from a set of paired data (x_1, y_1), (x_2, y_2), ...,(x_n, y_n), first **rank** the sets $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$, and convert the values of each (x_i, y_i) pair to the values of the ranks of each of x_i and y_i. Then take the usual (Pearson's) correlation coefficient of the resulting ranked version of the data.

Why test $\beta = 0$? (9)



$\beta = 0 \Rightarrow$ X is not useful as a predictor of Y. $(\beta = 0 \Rightarrow \rho = 0.)$

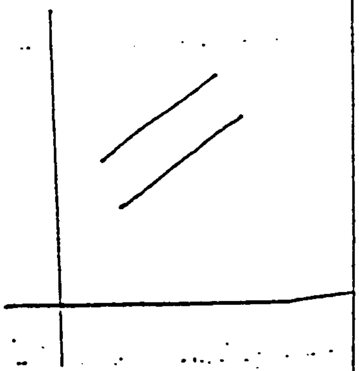


Knowing X does not help predict Y.

16

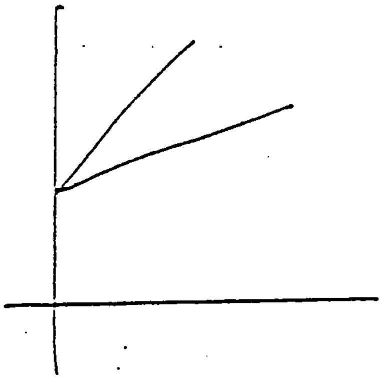
Comparisons of 2 Regression Lines

(a)



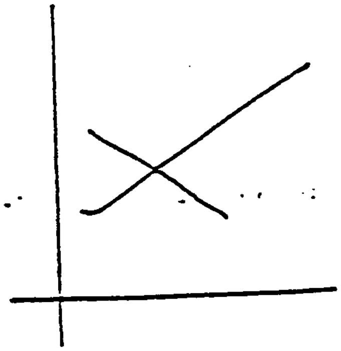
same slope
different intercepts

(b)



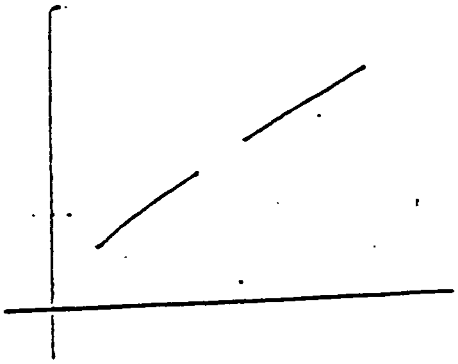
same intercept,
different slope

(c)



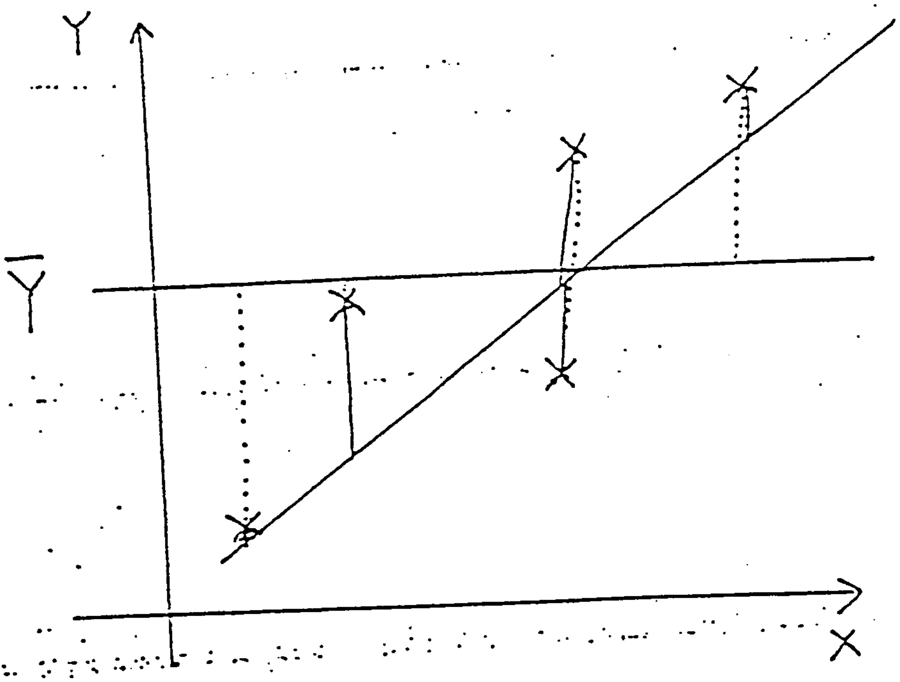
different slope
different intercept

(d)



same slope, same
intercept, different range.

"Goodness of Fit" of a Regression Line



-Total SS = $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$

Residual SS = $\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$

"Good Fit" if Total SS >> Residual

$$R^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}}.$$

$$\boxed{\pm\sqrt{R^2} = r = \text{correlation coefficient}}$$

Bayesian Inference for Regression Parameters

As you might guess by now, Bayesian inference for simple linear regression parameters follows the usual pattern:

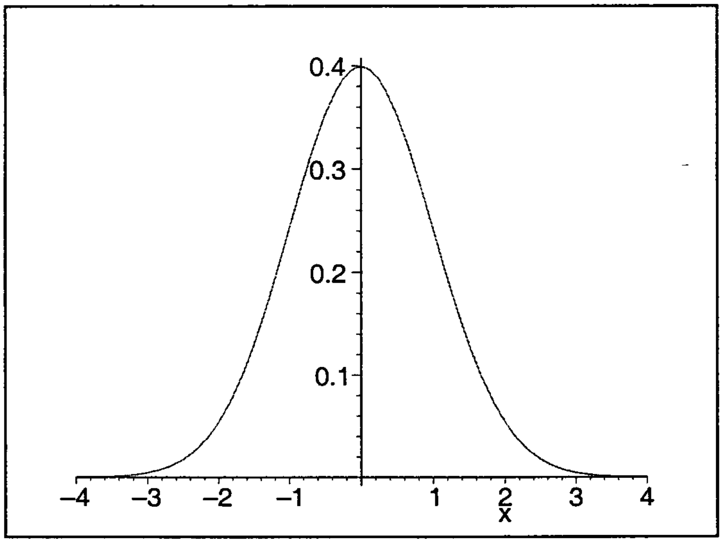
1. Form a prior distribution over all unknown parameters.
2. Write down the likelihood function of the data.
3. Use Bayes theorem to find the posterior distribution of all parameters.

- We have applied this generic formulation so far to problems with binomial distributions, normal means, and to the Poisson parameter. All of these problems involved only one parameter at a time.
- What makes regression different is that we have three unknown parameters, since the intercept and slope of the line, α and β are unknown, and the residual standard deviation, σ is also unknown.
- Hence our Bayesian problem becomes slightly more complicated, since we are in a multi-parameter situation.
- Before detailing the steps involved in Bayes Theorem for regression problems, we need to look at multiparameter problems in general.

2.10

Joint and Marginal Distributions

When we have only one parameter, we speak of its density. For example, if $x \sim N(0,1)$, then the graph of the probability density is:



When we have two or more parameters, we speak of a *joint probability density.* For example, let x and y be *jointly multivariately* normally distributed, which is notated by:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma \right)$$

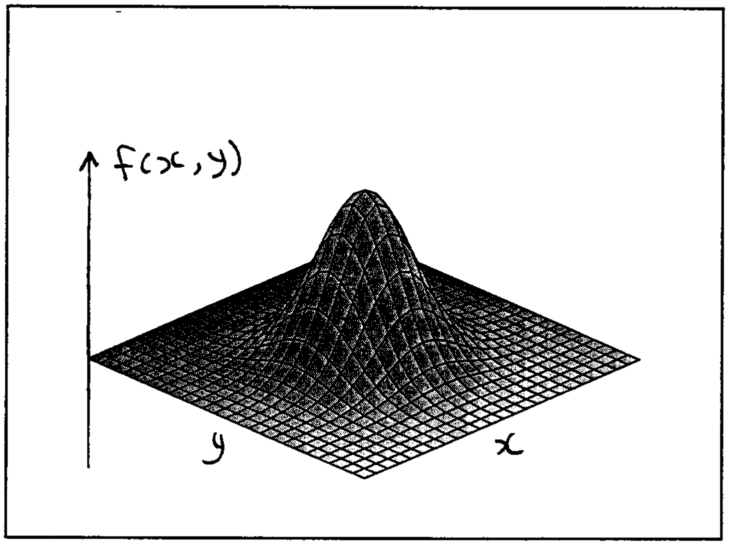
where

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho_{xy} \\ \rho_{xy} & \sigma_y^2 \end{pmatrix}$$

Example: Suppose

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

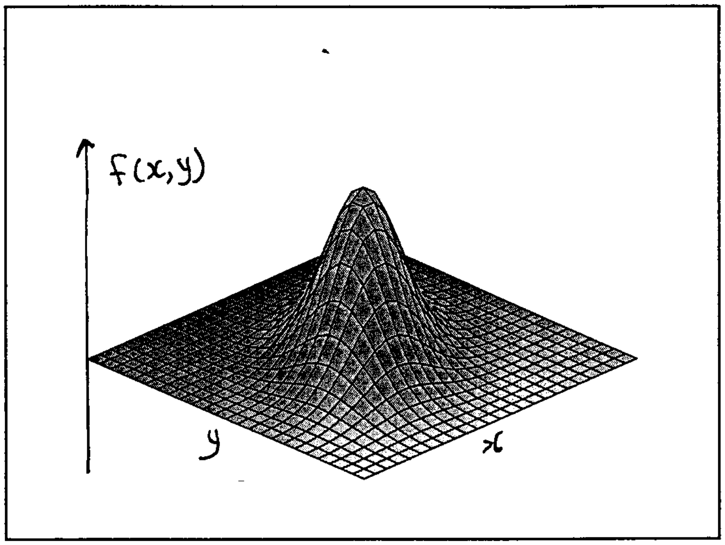
which is equivalent to two independently normally distributed variables, with no correlation between them. Then the picture is:



Note how the "slices" resemble univariate normal densities in all directions. These "slices" are marginal densities, which we will define later. In the presence of correlations, for example a correlation of 0.5, we have

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

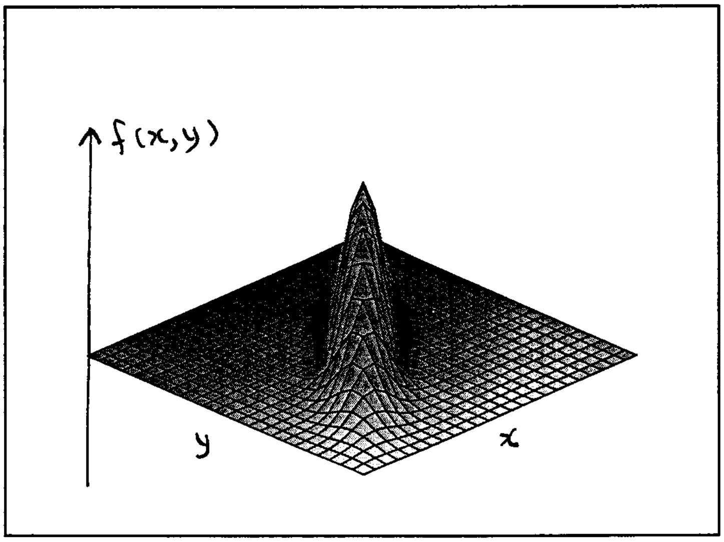
and the picture is:



Similarly, with very high correlation of 0.9, we have

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$$

and the picture is:



The *bivariate* normal density formula is:

$$f(x,y) = \frac{\exp\left\{-\frac{1}{2(1-\rho_{xy}^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho_{xy}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}$$

This is a joint density between two variables, since we look at the distribution of x and y at the same time, i.e., jointly. An example where such a distribution might be useful would be looking at both age and height together. (Another example is looking at the joint posterior distribution of α and β, which is where we are heading with all of this!!)

When one starts with a joint density, it is often of interest to calculate *marginal* densities from the joint densities. Marginal densities look at each variable one at a time, and can be directly calculated from joint densities through integration:

$$f(x) = \int f(x,y)dy, \text{ and}$$

$$f(y) = \int f(x,y)dx.$$

In higher dimensions,

$$f(x) = \int f(x,y,z)dydz,$$

and so on.

Normal marginals are normal

If $f(x, y)$ is a bivariate normal density, for example, it can be proven that both the marginal densities for x and y are also normally distributed. For example, if

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy} \\ \rho_{xy} & \sigma_y^2 \end{pmatrix} \right)$$

then

$$x \sim N(\mu_x, \sigma_x^2)$$

Summary:

- Joint densities describe multi-dimensional probability distributions for two or more variables.
- If one has a joint density, then if it is of interest to look at each variable separately, one can find marginal probability distributions by integrating the joint densities. If one wants the marginal distribution of x, for example, then one would "integrate out" all of the parameters except x, and so on.
- For multivariate normal distributions, all marginal densities are again normal distributions, with the same means and variances as the variables have in the joint density.

Brief Sketch of Bayesian regression

Recall the three steps: prior → likelihood → posterior.

1. We need a joint prior distribution over α, β, and σ. We will specify these as three independent priors [which when multiplied together will produce a joint prior]:

- $\alpha \sim$ uniform$[-\infty, +\infty]$
- $\beta \sim$ uniform$[-\infty, +\infty]$
- $log(\sigma) \sim$ uniform$[-\infty, +\infty]$

Notes:

- The need for the log comes from the fact the the variance must be positive. The prior on σ is equivalent to a density that is proportional to $\frac{1}{\sigma^2}$.
- We specify a non-informative prior distribution of these three parameters. Of course, we can also include prior information when available, but this is beyond the scope of this course.
- Our priors are in fact "improper" because their densities do not integrate to one, since the area under these curves in infinite! In general this is to be avoided since sometimes it can cause problems with posterior distributions. This is *not* one of those problem cases, however, and it is convenient to use a "flat" prior everywhere, so we will use it (it is also the default in First Bayes).

2. Likelihood function in regression:

- As is often the case, the likelihood function used in a Bayesian analysis is the same as the one used for the frequentist analysis.
- Recall that we have normally distributed residuals, $\epsilon \sim N(0, \sigma^2)$
- Recall that the mean of the regression line, given that we know α and β is $y = \alpha + \beta \times x$.
- Putting this together, we have $y \sim N(\alpha+\beta \times x, \sigma^2)$.
- So for a single patient with observed value x_i, we have $y \sim N(\alpha + \beta \times x_i, \sigma^2)$
- So for a single patient, the likelihood function is:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{(y_i - (\alpha + \beta \times x_i))^2}{\sigma^2}\right\}$$

- So for a group of n patients each contributing data (x_i, y_i), the likelihood function is given by

$$\prod_{i=1}^{n} f(y_i) = f(y_1) \times f(y_2) \times f(y_3) \times \ldots \times f(y_n)$$

- So the likelihood function is simply a bunch of normal densities multiplied together...a multivariate normal likelihood of dimension n.

3. Posterior densities in regression

- Bayes theorem now says to multiply the likelihood function (multivariate normal) with the prior $1 \times 1 \times \frac{1}{\sigma^2}$.

- So the posterior distribution simply is:

$$\prod_{i=1}^{n} f(y_i) \times \frac{1}{\sigma^2}.$$

- This is a three dimensional posterior involving α, β, and σ^2.

- By integrating this posterior density, we can obtain the marginal densities for each of α, β, and σ^2.

- After integration (tedious details omitted):
  - $\alpha \sim t_{n-2}$
  - $\beta \sim t_{n-2}$
  - $\sigma^2 \sim$ Inverse Chi-Square (so $1/\sigma^2 \sim$ Chi-Square)

- Note the similar results given by Bayesian and frequentist approaches for α and β.

- Computations usually done by computer programs. You will use First Bayes on assignment #5 to compute Bayesian posterior distributions for a regression problem, and compare the results to frequentist inferences. Of course, as usual, interpretations are different, and one can include prior information in a Bayesian approach.

- Bayes approach also suggests different ways to assess goodness of fit and model selection (beyond scope of course).

Extensions to "Simple Linear Regression"

- σ^2 <u>not</u> constant throughout range

  use <u>Weighted Least Squares</u> to "equalize" the variance

- <u>Polynomial Regression:</u>

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \cdots$$

- <u>Multiple Regression:</u>

$$Y = \alpha + \beta_1 X + \beta_2 Y + \beta_3 Z + \cdots$$

- <u>General Linear Model:</u>

$$g(Y) = \alpha + \beta_1 X + \beta_2 Y + \beta_3 Z + \cdots$$

(e.g., $g(Y) = \log Y \Rightarrow$ loglinear mode

- <u>Multivariable General Linear Model:</u>

$$g\begin{pmatrix} Y_1 \\ \vdots \end{pmatrix} = \alpha + \beta_1 X + \beta_2 Y + \beta_3 Z + \cdots$$

Principles of Inferential Statistics in Medicine

Assignment # 1 EPIB–607, Due: September 25, 2003.

1. A prevalence study indicates that 25% of women over the age of 75 in Canada have low bone mass (osteoporosis). If 10 women over 75 in Canada are randomly chosen and tested, what is the probability that at least one of them will have osteoporosis? (You can assume that the test is perfect, that is, the test is always positive if the patient really has the disease, and the test is always negative if the patient does not have the disease. In general, this assumption is unrealistic, as few if any tests are perfectly accurate.)

2. You have a torn tendon and are facing surgery to repair it. The orthopedic surgeon explains the risks to you: Infection occurs in 6% of such operations, the repair fails in 20%, and both infection and failure occur together in 2%. What percent of these operations succeed and are free from infection?

3. In a large population of mice, 15% of individual mice have a certain genetic anomaly. If a random sample of 10 mice are selected, give the probabilities that:

(a) There are no anomalies at all in the 10 mice.
(b) There is exactly one anomaly.
(c) There are two or less anomalies.
(d) Suppose now that 100 mice are randomly selected from the population. What is the expected number of mice with genetic anomalies in the sample?
(e) What is the (approximate) probability that 20 or more mice have anomalies?

4. Suppose that both parents in a family carry genes for blood types A and B. Then the blood types of their children are independent, and each child has a 1/4 probability of having blood type A. Let X be the number among the 3 children in the family who have blood type A. Compute the distribution of X (that is, the probability of each possible value) and draw a histogram for each possible value.

5. A bone densitometer is a device that measures bone density. A person is considered to have osteoporosis if their bone density is very low (a positive test), and otherwise not (a negative test). It is estimated that 25% of women over the age of 75 are osteoporotic, and that the sensitivity of the test is 80% and the specificity is 50%.

(a) Assuming the above estimates of the prevalence, sensitivity and specificity to be exactly correct, what is the probability that a woman over 75 who tests positive actually has osteoporosis?
(b) Assuming the above estimates of the prevalence, sensitivity and specificity to be exactly correct, what is the probability that a woman over 75 who tests negative actually has osteoporosis?

6. This is a variance contest. You must choose any 4 numbers from the range 0 to 100, with repeats allowed.
(a) Choose the numbers that have the smallest possible variance.

(b) Choose the numbers that have the largest possible variance.
(c) Is more than one choice possible in either (a) or (b)? Explain.

7. Data on the survival times from date of disease onset of 60 persons with lupus are used to create a boxplot describing the distribution of times to death in the sample. The results are:

Minimum value: 30 weeks
First quartile (25% point): 50 weeks
Median: 60 weeks
Third quartile (75% point): 100 weeks
Maximum value: 500 weeks
Mean: 75 weeks.

It is later found that one of the values, 500 weeks, should really have been 50 weeks. Explain how this change would affect

(a) the standard deviation of the sample.
(b) the inter-quartile range of the sample.

8. A person with asthma has a 75% chance of also having allergies. Suppose that the rate of asthma in a certain population is 20%.

(a) Is it possible to determine the probability that a person has asthma but not allergies in that population? If yes, give the probability. If not, state what information is missing.
(b) Is it possible to determine the probability that a person has asthma given that they have allergies in that population? If yes, give the probability. If not, state what information is missing.

9. Calculate the following probabilities, assuming that $X \sim N(\mu = 10, \sigma^2 = 225)$:

(a) $Pr\{X = -13\}$
(b) $Pr\{-8 \leq X \leq +8\}$
(c) $Pr\{X \leq 0\}$

10. The table below contains data on the age, cigarette smoking habits (0=YES, 1=NO), systolic blood pressure and body mass index for 25 patients with heart disease.
(a) Calculate the means and standard deviations for the age, body mass index, and systolic blood pressure variables.
(b) Draw a stemplot for the BMI variable, and comment on the shape. Are there any outlier values?
(c) Draw separate boxplots for systolic blood pressure for smokers and nonsmokers. Using the boxplots, compare the SBP of smokers and non-smokers by commenting on the relative locations of their medians. Does one group seem to have more highly spread values compared to the other?

(Note: If you wish, you may use a computer to help with these exercise, although all can be done by hand.)

Age	CIGS	SBP	BMI
39	1	135	29
57	1	120	25
44	0	150	27
57	1	165	26
41	1	140	24
49	1	150	27
42	0	158	27
36	1	130	29
43	1	230	28
51	0	200	36
37	0	125	25
57	0	235	23
56	1	160	25
57	0	140	30
57	1	115	23
44	0	130	24
58	0	215	28
38	1	158	27
49	0	165	31
56	1	140	27
52	0	130	22
49	1	155	22
55	1	150	26
56	1	148	26
43	1	140	31

222

Principles of Inferential Statistics in Medicine

Assignment # 2 EPIB–607, Due: October 21, 2003.

1. Describe the effects of increasing the sample size (i.e., the number of subjects in the experiment) on each of the following:

(a) The width of a confidence interval for the population mean, from a random sample taken in the population.
(b) The power of a one-sided test for a mean, when H_0 is false and all facts about the population remain unchanged as the sample size increases.

2. A recently developed treatment for gallstones is laparoscopic surgery. One of the factors thought to be predictive of the rate of successful laparoscopic surgery is the age of the patient. Suppose that in a recent trial, the average age of the patients on whom the surgery was successful was 60 years old, while the average among those with unsuccessful surgery was 70. In each group there were 50 patients, and the standard deviation in each group was 10 years.

(a) Test whether the average age differs in the two groups. State the null and alternative hypotheses, the p-value, and your conclusion.
(b) Calculate a 95% confidence interval for the difference in mean age of the two groups. Give the interpretation of this confidence interval.

3. In a two-sided test of a null hypothesis, it is found that the p-value is (p=0.50). State whether each of the following statements are true or false, and explain why.

(a) The null hypothesis has a 50% chance of being true, i.e., the probability that the null hypothesis is true is equal to 0.50.
(b) After carrying out this experiment, there is a one in two chance (0.50) of being wrong if the conclusion is to not reject the null hypothesis.

4. Suppose that the true average value of creatinine clearance in a population of lupus patients is 1.9 ml/sec, with a true standard deviation of 0.025 ml/sec. An investigator takes a random sample of 50 patients from this population, and measures their creatinine clearance.

(a) What is the probability that the sample average of these 50 individuals will be higher than 1.8 ml/sec?
(b) What assumption was necessary in order to answer part (a) of this question?

5. Whether certain mice are black or brown depends on a pair of genes, each of which is denoted as either B or b. A mouse is brown only if it has the pair bb, and otherwise, for the pairs Bb or BB, it is black. The offspring of a pair of mice have two such genes, one from each parent. If a parent has either BB or bb, the offspring receives the same gene as that parent. If the parent has the pair Bb, then the offspring is equally likely to inherit a B or a b from that parent. Suppose that a black mouse results from a mating of a pair of mice each with Bb genes. Suppose further that this mouse is then mated with a brown mouse, and

that all seven offspring turn out to be black. What is the probability that the black parent mouse was BB?

6. Suppose that the weights of a large group of students are normally distributed wth a mean of 50 Kg and a standard deviation of 15 Kg.

(a) Find the proportion of students with weights between 55 and 80 Kg.
(a) Find the proportion of students with weights greater than 50 Kg.

7. When asked to explain the meaning of "statistically significant at the $\alpha = 0.05$ level," a student says, "This means that there is only probability 0.05 that the null hypothesis is true." Is this an essentially correct explanation of statistical significance? Explain your answer carefully.

8. Using First Bayes, calculate the following probabilities:
(a) For a $N(\mu = 0, \sigma^2 = 4)$ distribution, calculate the probability that $-4 \leq X \leq 2$.
(b) For a t distribution with 10 degrees of freedom, calculate the probability that $2 \leq X \leq \infty$. (Hints: Set median = 0, and scale = 1. Also: Since you cannot put ∞ in as a number, substitute a very large number (eg 50 or 100) as a near perfect approximation.)
(c) For a binomial distribution with 20 trials and probability of success equal to 30%, what is the probability of getting exactly 6 (the most likely outcome) successes? What is the probability of getting 6 or more successes?

You need not print them out, but it may be instructive for you to look at the graphs of these distributions as you calculate the probabilities.

9. Consider again the example of children given an intelligence test, first presented on page 129 of the class notes. Using First Bayes, carry out the following analyses:

(a) Confirm (repeat) the analyses given in the notes, that is, find the posterior probability that $\mu \leq 100$ if the prior distribution is such that $\mu \sim N(\theta = 100, \tau^2 = 400)$. (Hints: For this step, you must first enter the data using the file/data menu item, then go to the analysis/normal sample variance known menu item. Enter the prior on the first sheet, and when finished, click on quit to go to the analysis page. Load the data, and do not forget to change the data variance to $8^2 = 64$ in the upper right hand corner.)
(b) Repeat (a), but this time assume that the prior distribution is $\mu \sim N(\theta = 100, \tau^2 = 16)$. (Hint: You do not have to start over, simply click in the edit prior box to go back to the prior page.)
(c) Repeat (a) again, but now assume that the prior distribution is $\mu \sim N(\theta = 110, \tau^2 = 400)$.
(d) Repeat (a) once again, but with prior distribution $\mu \sim N(\theta = 110, \tau^2 = 16)$.

For each of questions (a) through (d), print out the tri-plots (ie, provide a single graph that has on it the prior distribution, the likelihood of the data, and the posterior distribution).

(e) Based on your four answers, give some general observations about the sensitivity of the final conclusions to the choice of prior distribution.

10. The table below gives data on blood pressure before and after treatment for two groups of subjects participating in a clinical trial. One group took a daily calcium supplement, while the other group received a placebo.
(a) Calculate a 95% confidence interval for the difference in blood pressure changes (before minus after) between the two groups. Give the interpretation of this confidence interval.
(b) Carry out a *t*-test of the null hypothesis that there is no difference in blood pressure changes between the two groups. State the null and alternative hypotheses, calculate the test statistic, and state your conclusion.
(c) You now must make a decision regarding whether or not to prescribe calcium supplementation to your patients with mild high blood pressure. In helping you to make this decision, would your answer to part (a) or (b) provide more useful information? Why?

Calcium Group		Placebo Group	
before	after	before	after
107	100	123	124
110	114	109	97
123	105	112	113
129	112	102	105
112	115	98	95
111	116	114	119
107	106	119	114
112	102	112	114
136	125	110	121
102	104	117	118
–	–	130	133

Principles of Inferential Statistics in Medicine

Assignment # 3 EPIB–607, Due November 13, 2003.

1. Suppose that you are planning an experiment to accurately estimate the difference in success rates between a standard and a new treatment. The standard treatment is expected to have a success rate of 20%, while the new treatment is estimated to improve this by at least 10%, that is, a success rate of 30% is expected.

(a) A clinical trial planner thinks that a total CI width of 10% (that is, ±5%) is reasonable to use in a sample size calculation for this trial. Do you agree with this assessment? Why or why not?
(b) Assuming that a total CI width of 10% is reasonable, what sample size in each group would be needed for a 95% confidence interval for the difference in success rates to have this total width?

2. An expert's best guess for the rate of osteoporosis in women over 75 years old is 25%. Suppose that a survey is being conducted to estimate the rate of osteoporosis in women over 75 years old. What sample size would be required to estimate this rate using a 95% confidence interval, such that the interval would have a total length of 4%?

[You may wish to check your answers to numbers 1 and 2 by using the sample size calculator available from my homepage.]

3. The following data are observed in an experiment designed to compare a new treatment to a standard therapy:

	Therapy		
	New	Standard	
Success	6	3	9
Failure	1	4	5
	7	7	14

(a) Test the null hypothesis that there is no difference in success rates between the new and standard therapies. State the null and alternative hypotheses, and calculate a p-value using a χ^2 test. State your conclusion.
(b) Repeat part (a), but use a (two-sided) Fisher's Exact test instead.
(c) How do the p-values calculated by the two different procedures compare?

4. Suppose that there is a 80% chance for a certain operation to be successful each time it is performed. Further suppose that Hospital A performed 1000 such operations last year, and Hospital B performed this operation on 200 patients last year. Without necessarily doing the calculations, state which, if any, of these two hospitals has a better chance of having an observed 70% success rate for the operation last year, and explain why.

5. Consider the following data set from an experiment of heavy versus light coffee drinkers:

	Coffee consumption		
	heavy	light	
stomach cancer	10	60	70
no stomach cancer	110	800	910
	120	860	980

(a) Calculate point estimates of the odds ratio and relative risk.
(b) Calculate an approximate 95% confidence interval for the odds ratio.

6. There exists a test in which pregnant women with a history of polycystic kidney disease can determine if the child they are carrying is also likely to have the disease. The test has a specificity of 80% and a sensitivity of 90%. We can assume that only women who truly have the disease take the test. The genetic pattern of inheritance indicates that the probability of passing the disease on to a child is 50%, ie,

$$Pr\{\text{child will be born with polycystic kidneys} \mid \text{mother has the disease}\} = 0.50$$

Assuming the values given above for sensitivity and specificity to be exactly correct, what is the probability that the child will have the disease given that the mother has the disease, and the test is positive?

7. You have just gone shopping, and received a quarter in change from the cashier.

(a) Assume that your prior probability that the coin will come up heads in any given toss can be expressed by a beta distribution with appropriately chosen α and β parameters. State your prior distribution. (Note: There is no "correct" answer, since each individual will have their own prior distribution. However, you should justify your answer in terms of your prior mean and variance (or standard deviation), that is, check to ensure that the values of α and β give reasonable means and variances. You may wish to imagine a 95% probability interval, and consider that the mean is in the center of that interval, and that four times the standard deviation will equal the length of that interval. See page 156C of the notes.)
(b) Suppose that the coin is now tossed 5 times, and there are no heads. What is your posterior probability for the probability of heads for that coin? What is the 95% highest density interval? (Hint: You will likely wish to use First Bayes for this question. First enter and load a new data set, which simply consists of the number 0 repeated 5 times. Then go to the analyses menu, and choose binomial sample. Enter your prior distribution from part (a) in the first screen, and the "quit" to go to the posterior screen. Load the data set you just entered, and the posterior distribution will then be available.)
(c) If you were using a frequentist approach to analyse the same data (ie, five tails in a row), what would the exact 95% confidence interval be? (Hint: see page 136 of the notes.)
(d) Provide interpretations of the intervals you calculated in parts (c) and (d). Which of the intervals given in (c) or (d) do you prefer? Why?

Principles of Inferential Statistics in Medicine

Assignment # 4 EPIB–607, Due November 20, 2003.

1. The table below gives the birth weights (in pounds) of babies categorized by whether their mother smoked or not.

Smokes		Never Smoked	
4.5	6.9	3.3	6.6
5.4	6.9	6.6	7.4
5.6	7.1	6.6	7.4
5.8	7.0	6.8	7.2
6.1	7.3	6.9	7.4

Use the appropriate nonparametric test to test whether smoking mothers give birth to babies of different weight than non-smoking mothers. State the null and alternative hypotheses, calculate the p-value, and state your conclusion.

2. The following are the weights in kg, before and after of ten persons who stayed on a certain reducing diet for four weeks:

Patient #	BEFORE	AFTER
1	75	70
2	80	78
3	120	116
4	79	70
5	100	101
6	64	67
7	76	70
8	82	75
9	88	80
10	120	110

Use the appropriate nonparametric test to determine whether the weights before are different from the weights after the diet. State the null and alternative hypotheses, calculate the p-value, and state your conclusion. Calculate a 95% CI for the median difference.

3. The table below gives the pressures in mmHg of two different heart pumps:

Pump A		Pump B	
6	16	1	18
7	16	2	20
9	16	3	26
9	17	4	29
10	17	8	30
11	18	11	31
11	19	12	31
12	19	14	32
12	21	15	35
13	23	16	44
15	24	17	45

Use the appropriate nonparametric test to test whether there is a difference in pressure between pumps A and B. State the null and alternative hypotheses, calculate the p-value, and state your conclusion.

4. Using the same data as in the previous question, test whether there is evidence for a difference in spread between the two groups. Use the appropriate nonparameteric test.

Principles of Inferential Statistics in Medicine

Assignment # 5 EPIB–607, Due December 2, 2003.

1. State whether each of the following statements are true or false, and explain why:

(a) A high negative correlation value, for example, $\rho = -0.95$, means that the value of the regression line slope, β must be a large negative number.
(b) If a t-test of the hypotheses

$$H_o \; : \; \rho = 0$$

versus

$$H_a \; : \; \rho \neq 0$$

is rejected at the $\alpha = 0.05$ level, then ρ cannot be a number very near 0.

2. Assume that the mean diastolic blood pressure of Canadian men in their thirties is 84 with a standard deviation of 3, and the mean diastolic blood pressure of women in the same age group is 78, with a standard deviation of 5.

(a) If the correlation coefficient between the diastolic blood pressure of husbands and wives in this age group is 0.6, what is the slope of the regression line of the husbands diastolic blood pressure (y) on the wives diastolic blood pressure (x) for marriages in this age group?
(b) Under the same conditions as (a), can you predict the diastolic blood pressure of a man whose wife has a pressure of 90? If yes, calculate the predicted value.

3. State whether each of the following statements are true or false, and explain why:

(a) If the slope of the regression line between two variables x and y is $\beta = 1$, then the slope between x and $y/2$ (i.e., all y values are divided by two) must be 0.5.
(b) If the intercept of the regression line between two variables x and y is $\alpha = 1$, then the intercept between x and $y/2$ (i.e., all y values are divided by two) must be 0.5.

4. Consider the data in the table below:

Case #	Dosage level	weight gain
1	6	16.2
2	3	11.6
3	5	13.5
4	7	18.6
5	2	7.8
6	8	24.5
7	7	21.0
8	4	13.3
9	4	14.3
10	6	14.0

The data come from an experiment on a drug that is supposed to increase weight. We will analyse the effects of the different dosages on the weights.

(a) Draw a rough scatter plot (by hand) to visually examine the association between the dosage (x-axis) and weight gain (y-axis). Does there appear to be a relationship?
(b) Calculate the regression line for this data, that is, provide the best values for the intercept (α) and slope (β) of the least squares line.
(c) Calculate the estimate of the residual standard deviation, σ.
(d) Calculate 95% confidence intervals for the intercept and slope values you calculated in part (b).
(e) Suppose the next subject that enters the study is given a dosage of 5. What is your prediction for the weight gain for this subject?
(f) Give a 95% confidence interval around your answer in (e).

5. In this question we will reanalyse the data in Question 4 using First Bayes. To do this, we first have to enter the data. Going to the data page, enter two different data sets, one for dosage, and one for weight. Make sure you keep the same order as in the table above. Save each of them, and load both for future use. Going to the analysis menu, choose the "regression data" item.

Unlike the analysis of normal and binomial data, this brings you immediately to the posterior page, rather than a page that first lets you enter prior information. This is because First Bayes always assumes "weak" or non-informative (uniform or close to uniform) prior distributions. [Note: This is not because prior information is not usually available in regression situations (it usually is), but rather because is if difficult to specify prior distributions at an elementary level appropriate to an introductory course. There are ways to get around this limitation in First Bayes, but this is beyond the scope of this course.] Because non-informative prior distributions are used, we would expect to get similar answers to those obtained in Question 1. For example, we expect that the 95% CI's calculated above will be similar (but not identical) to 95% HDI's we will calculate below. Of course, the interpretations of the intervals are different.

(a) On the posterior page for regression, load the dosage data as the x variable, and the weight as the y variable. Clicking on 'scatter" will give you a scatter plot of the data, together with 95% intervals for future predictions. Print out this plot.
(b) You can examin the posterior distributions for the α, β, and σ^2 parameters by clicking on the ˆ icon. The default when you open the screen is β. Calculate the 95% HDI's for α and β. How do they compare to the confidence intervals you calculated above?
(c) Give the mean and median values for the variance. How do they compare to the point estimate of the variance you calculated in Question 4(c)? (Hint: Remember to square the latter value, so you compare a variance to another variance, not a standard deviation). Look at the plot of the posterior distribution for the variance. Is it symmetric or skewed?
(d) Suppose the next subject that enters the study is given a dosage of 5. What is your mean prediction for the weight gain for this subject?
(e) Give a 95% confidence interval around your answer in (d). (Hint: To do the latter two parts, you must click on "Show predictive", which means that the left hand side of the screen now shows information about predictions, rather than about α, β, or σ^2.)
Plot the graph of the residuals. Does a linear regression model seem appropriate for these data?