This book will give ophthalmologists a simple, clear presentation of epidemiologic and statistical techniques relevant to conducting, interpreting, and assimilating the commonest types of clinical research. Such information has become all the more important as clinical ophthalmic research has shifted from anecdotal reports and uncritical presentations of patient data to carefully controlled and organized studies.

The text is divided into two parts. The first part deals with epidemiologic concepts and their application and with the various types of studies (for example, prospective or longitudinal) and their organization (use of controls, randomness, sources of bias, choice of sample size, standardization and reproducibility). The second half of the text is devoted to the selection and use of the relevant statistical manipulations: determining sample size and testing the statistical significance of results. The text is illustrated with many examples covering topics such as blindness registry data, incidence of visual field loss with different levels of ocular hypertension, sensitivity and specificity of tests for glaucoma, and sampling bias in studies of the safety of intraocular lenses. References and several helpful appendices are included.

The Author

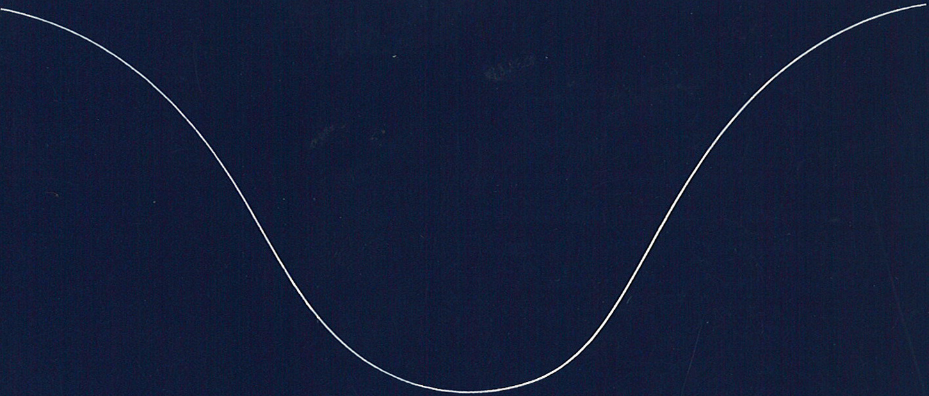
Alfred Sommer received his M.D. in Ophthalmology from Harvard University Medical School, and his M.H. Sc. degree in Epidemiology from Johns Hopkins School of Hygiene and Public Health. At the Wilmer Institute, he is the Director of the International Center for Epidemiologic and Preventive Ophthalmology. He is the author of *Field Guide to the Detection and Control of Xerophthalmia.*

Oxford University Press, New York

Cover design by Joy Taylor

ISBN 0-19-502656-X

Epidemiology *and* Statistics *for the* Ophthalmologist



Alfred Sommer

Epidemiology and Statistics for the Ophthalmologist

ALFRED SOMMER, M.D., M.H.Sc.
The Wilmer Ophthalmological Institute
of Johns Hopkins University,
and Helen Keller International

New York Oxford
OXFORD UNIVERSITY PRESS
1980

LIBRARY OF CONGRESS IN PUBLICATION DATA

Sommer, Alfred, 1942–
Epidemiology and statistics for the
ophthalmologist.

Bibliography: p.
Includes index.
1. Ophthalmology—Statistical methods.
2. Epidemiology—Statistical methods. I. Title.
[DNLM: 1. Epidemiologic methods. 2. Statistics.
WA950 S697e]
RE58.S67 617.7′001′82 79-17792
ISBN 0-19-502656-X

Printed in the United States of America

To Nat and Joe

Preface

There are three kinds of lies: lies, damned lies, and statistics.

BENJAMIN DISRAELI

Until recently, epidemiologic and statistical principles were routinely ignored in ophthalmic research. Fortunately this is no longer the case, since their use results in better designed, more efficient, and meaningful studies—usually with little additional work. It is essential that every clinical investigator be familiar with the concepts involved: where outside epidemiologic and statistical assistance is available, he must be able to recognize what help to seek and questions to ask; and where not available, he must be able to carry out the necessary procedures on his own. It is equally essential for every informed clinician who wishes to be better equipped to evaluate the significance and value of published—and sometimes conflicting—reports.

This is a practical primer for busy clinicians who have neither the time nor inclination to pursue formal courses in epidemiology and statistics. For simplicity and brevity we confine ourselves to principles and techniques required for conducting and interpreting the most common types of clinical studies: descriptive reports seeking new etiologic agents, evaluations of diagnostic and screening procedures, and therapeutic trials.

I can only hope my colleagues will ultimately disagree with Disraeli, and find the subject interesting and above all, useful.

London, England
September 1979

Alfred Sommer

Acknowledgments

This guide is the outgrowth of what now seems an obvious and natural merging of my interests in epidemiology and ophthalmology. Yet when I originally and quite accidentally considered combining these two disciplines I encountered little precedent and even less encouragement. Were it not for Dr. A. Edward Maumenee, then Director of the Wilmer Institute, and Dr. Abraham M. Lilienfeld, then Chairman of the Department of Epidemiology at Hopkins I would likely have ended up pursuing one of these disciplines to the exclusion of the other. For their continuing friendship, advice, and encouragement I am deeply grateful.

I would also like to thank Drs. Frank Newell, Frederick Blodi, and Argie Hillis for reviewing early drafts and making many helpful suggestions, and Mr. David Andrews of the Wilmer Institute for his editorial assistance.

Epidemiology and Statistics for the Ophthalmologist

Contents

Epidemiology

Epidemiology has two overriding characteristics: a preference for rates rather than absolute numbers, and a peculiarly thoughtful approach to studies amounting to applied common sense.

RATES: THEIR MEANING AND USE

Attack rate

Epidemiologists almost always present data in the form of rates: the *proportion* of individuals with a particular disease or characteristic. A common example is the *attack rate*. Christy and Sommer (1) were interested in determining which antibiotic regimen, if any, offered the best protection against postoperative endophthalmitis. They divided their patients into three groups (Table 1). The first group received infrequent preoperative topical antibiotics; the second intraoperative periocular penicillin; and the third intensive preoperative topical chloramphenicol combined with intraoperative periocular penicillin. The absolute number of cases in each group tells us little because the size of the groups varied widely. Absolute numbers only become meaningful after adjusting for the size of their respective group. This adjustment, the common *attack rate*, is calculated as follows:

$$\text{Attack rate} = \frac{\text{Number of individuals who develop the disease}}{\text{Number of individuals at risk of developing the disease}} \times 1000$$

In each group the "individuals who develop the disease" are the cases of endophthalmitis, whereas those "at risk of developing the disease" were all individuals who underwent cataract extraction. This adjustment almost always results in a tiny fraction, the number of cases of disease per person at risk. For convenience this is multiplied by 1000 (or some other appropriate number) and the results expressed as rate of occurrence per 1000 individuals. In this particular example it became apparent that introduction of combined chloramphenicol and penicillin prophylaxis resulted in a marked drop in the infection rate, whereas penicillin alone had little effect.

Another interesting example is provided by glaucomatous blindness registry data (2). The total number of nonwhites and whites registered as blind in the Model Reporting Area were

Table 1
INCIDENCE OF POSTOPERATIVE ENDOPHTHALMITIS

	Prophylactic regimen			*Infections*	
Series	*Penicillin*	*Chloramphenicol-sulphadimidine*	*Operations (number)*	*Number*	*Rate per 1000*
Ia Jan. '63–Dec. '67	–	–	9714	54	5.6
Ib Jan. '68–May '72	–	–	12,340	55	4.5
II May '72–Dec. '72	+	–	2071	9	4.3
III Jan. '73–March '77	+	+	21,829	30	1.4

Modified from Christy and Sommer (1).

Table 2
PERSONS REGISTERED BLIND FROM GLAUCOMA

	Number	*Population*[1]	*Rate per 100,000*[2]
White	2832	32,930,233	8.6
Nonwhite	3227	3,933,333	72.0

[1] Fourteen Model Reporting Area states
[2] Population adjusted to a standard (equivalent) age distribution
Modified from Hiller and Kahn (2).

roughly comparable (Table 2). Adjusting these numbers for the size of their respective populations, however,

$$\begin{array}{l}\text{Rate of registered}\\\text{blindness per}\\\text{100,000 population}\end{array} = \frac{\begin{array}{c}\text{Number of individuals}\\\text{registered as blind}\end{array}}{\begin{array}{c}\text{Number of individuals}\\\text{in the population}\end{array}} \times 100{,}000$$

reveals a strikingly different picture: the blindness rate among nonwhites was over 8 times that among whites.

Relative risk

In the example above it was natural, almost without thinking, to compare the *rate* of registered blindness in the two racial groups and thus recognize that the rate among nonwhites was 8 times that among whites. There is a simple term to express this concept: *relative risk*.

In our previous example, the rate of postoperative endophthalmitis among patients not receiving prophylaxis was 4.9 per 1000, while among those receiving combined prophylaxis it was only 1.4 per 1000. Individuals not covered by combined prophylaxis ran a risk of postoperative endophthalmitis 4.9/1.4 or 3.5 times greater than those receiving combined prophylaxis. Their risk of endophthalmitis, *relative* to those receiving such prophylaxis, was 3.5:1.

$$\text{Relative risk} = \frac{\text{Rate of disease in group 1}}{\text{Rate of disease in group 2}}$$

Similarly, the relative risk of registrable glaucoma blindness among nonwhites was 8.4 times that among whites.

A "relative risk" is the ratio of two rates: the risk of disease in one group to that in another. It is therefore not an absolute figure. The denominator, of course, assumes a value of 1. Hence the relative risk of the group in the denominator is 1 vis-à-vis the rate of disease in the group in the numerator.

Similarly, the relative risk for the numerator group is relative to this particular denominator group. If a different denominator group is used, with a different rate (risk) of disease, then of course the *relative* risk of the group in the numerator will also change, even though its *absolute* rate (risk) of disease remains the same.

Group-specific rates

The rate at which a particular disease occurs within a group is a summary, overall statistic. It does not mean that each and every individual in that group is actually at identical risk of disease. Individuals vary markedly, and some of these differences might be important factors influencing the occurrence of the disease. Our two previous examples make this point nicely. The overall rate of postoperative endophthalmitis among all patients in the series was 3.3 per 1000. This does not mean that all patients ran an identical, 3.3 in 1000 chance of developing endophthalmitis. Slicing up the baloney appropriately, we found that the rate varied from a high of 4.9 per 1000 among those who did not receive prophylactic antibiotics to a low of 1.4 per 1000 for those who received combined prophylaxis. Had we been sufficiently clever in slicing it up further, we might have determined which factors were responsible for endophthalmitis in the first place. As it was, we were not. We classified patients by whether or not they suffered vitreous loss, iris prolapse, extracapsular extraction, etc. and then calculated endophthalmitis attack rates for each group. Unfortunately, the rates were roughly the same, the relative risk for each comparison (extracapsular extraction versus intracapsular, iris prolapse versus no prolapse, etc.) being approximately 1:1. Since the relative risks were all "1", none of these conditions added appreciable risk to the development of endophthalmitis. They were therefore not significant risk factors.

The situation with registered glaucomatous blindness was quite different. The overall rate of glaucomatous blindness was 16.4 per 100,000. As we've already seen, race-specific rates indicated that such blindness was more common among nonwhites than among whites. Race is apparently an important risk factor. This does not necessarily mean that nonwhites suffer more glaucoma or have a genetic predisposition to the disease: the cause may be a lack of health care and delay in diagnosis, greater access to registration, and the like. We will discuss these epidemiologic inferences later.

Race, age, and sex are so often related to disease that their relative risk almost always requires evaluation. Simultaneous race-, age-, and sex-specific glaucomatous blindness registration rates are presented in Table 3. Not surprisingly, blindness rates increase with age. What is astonishing, however, is the extraordinary rate of disease among nonwhites between the ages of 45 and 64. Quick calculation indicates that their risk of glaucomatous blindness registration is 15 times greater than that among whites of similar age. This does not prove that middle-aged nonwhites are actually more prone to glaucoma or glaucomatous blindness, but identifies an exceptional event requiring further investigation. Potential high-risk factors identified in this way are often the earliest clues to the etiology of a condition.

Prevalence and incidence

These two terms are almost always misused.

Prevalence is the rate or frequency with which a disease or trait is found in the group or population under study at a particular point in time.

Table 3
RACE-, AGE-, AND SEX-SPECIFIC GLAUCOMA BLINDNESS REGISTRATIONS

		Rate per 100,000		
Sex	*Age*	*White*	*Nonwhite*	*Ratio nonwhite/white*
Male	20–44	1.2	9.5	7.9
	45–64	9.3	154.9	16.7
	65–74	42.9	430.2	10.0
	75–84	104.5	637.2	6.1
	85+	285.7	707.7	2.5
Female	20–44	0.6	5.6	9.3
	45–64	8.3	111.1	13.4
	65–74	30.9	356.0	11.5
	75–84	92.5	534.4	5.8
	85+	284.6	773.7	3.1

Modified from Hiller and Kahn (2).

$$\text{Prevalence} = \frac{\begin{array}{c}\text{Number of individuals with the trait}\\ \text{at the time of examination}\end{array}}{\text{Number of individuals examined}}$$

In strict epidemiologic parlance we can say that the prevalence of elevated intraocular pressures (21 mm Hg or above) in the general adult population of Ferndale, Wales was 9%, while the prevalence of glaucomatous field loss was only 0.4% (3). In less strict usage, modified for clinical series, we might say the prevalence of severe malnutrition among children admitted to the hospital with active vitamin A deficient corneal disease was 66% (4), and the prevalence of unrecognized intraocular malignant melanomas among eyes enucleated with opaque media was 10% (5). Prevalence concerns a condition already present at the time of examination, regardless of when that condition arose.

Incidence is the frequency with which new cases of a disease or other characteristic arise over a defined period of time:

$$\text{Incidence} = \frac{\text{Number of individuals who developed the condition over a defined period of time}}{\text{Number of individuals initially lacking the condition who were followed for the defined period of time}}$$

After the adult population was examined for glaucomatous visual field loss, ocular hypertensives without such loss were followed for 5 to 7 years, and the rate at which new cases of field loss occurred was calculated (6). On the average, 5 new cases of field loss occurred among every 1000 ocular hypertensives during each year of follow-up (an incidence of 5 per 1000 per year). In another study, with an average follow-up of 43 months, the incidence of glaucomatous field loss clearly increased with increasing level of initial intraocular pressure, confirming the familiar clinical observation that people with higher pressures are at greater risk of developing visual field loss (Table 4) (7).

Prevalence rates and incidence rates are obviously interrelated. If a condition (such as glaucomatous field loss) is permanent, and if people with it have the same mortality rate as the rest of the population, the prevalence of the condition in the

Table 4
INCIDENCE OF GLAUCOMATOUS VISUAL FIELD LOSS IN HYPERTENSIVE EYES

Initial IOP (mm Hg)	*Total number of eyes*	*Developed visual field defect*		
		number	*percent*	*incidence*[1]
21-25	75	2	3	8
26-30	25	3	12	34
>30	17	7	41	114

[1] Per 1000 per year grossly approximated by applying average follow-up of 43 months to all groups. More accurate analysis would have employed an IOP-specific life-table analysis.

Modified from David et al. (7).

population will be the sum total of its past incidence. Thus, if the overall incidence of glaucomatous field loss among ocular hypertensives is 5 per 1000 per year, 50 new cases will occur among 1000 ocular hypertensives followed for 10 years. At the end of that period, the prevalence of visual field loss among the original population of ocular hypertensives will be 50 per 1000. As usual, life is not always that simple. Individuals with the condition do not always accumulate in the population: the mortality rate among those with the condition might be greater than among those without, or the condition itself might be reversible. The overall incidence of active, irreversible, xerophthalmic corneal destruction among preschool Indonesian children is 4 per 1000 per year (8). We would therefore expect the prevalence of corneal scars among five year olds to be about 20 per 1000. Instead it is half that, indicating that the mortality rate among affected children must have been twice that of the others. Similarly, we would expect a large proportion of patients with intraocular melanoma to succumb to their disease, and the prevalence of intraocular melanomas in the community to tell us little about the true incidence of this malignancy.

The prevalence of night blindness and Bitot's spots (potentially reversible manifestations of vitamin A deficiency) in Indonesian children is 7 per 1000. This represents the net cumulative effect of an annual incidence of 10–14 per 1000 and spontaneous cure rate of 30–70% (4, 8).

Familiar examples of incidence include the rate of secondary hemorrhages among patients undergoing different treatments for traumatic hyphema (although rarely indicated as such, these are usually new events over a 2–4-week observational period); and the rate of cystoid macular edema (during the first postoperative week, month, etc.) following cataract extraction.

These are the only *rates* required to describe most clinical observations. Does blunt trauma lead to chronic glaucoma? Simply compare the *incidence* of new cases of glaucoma among patients who experienced blunt trauma in the past with the incidence among appropriately matched controls. Alternatively, compare the *prevalence* of glaucoma in the two groups at a particular point in time. Does intraocular lens insertion increase the risk of cystoid macular edema? Compare the incidence of CME in matched groups of patients who either did or did not receive an IOL at the time of cataract surgery over a given postoperative interval, or the prevalence of CME in the two groups at one or more postoperative points in time (one month, one year, etc). Does miotic therapy reduce the risk of visual field loss in ocular hypertensives? Compare the incidence of field loss in two well-matched groups, one treated, one not. As we shall see, the two groups can be composed of separate individuals, or, preferably, opposite eyes of the same individuals.

In every instance we simply compare the rate at which the disease or characteristic is found (prevalence) or occurs over time (incidence) in one population to the rate in another. If the incidence of glaucomatous field loss turns out to be lower in the miotic-treated group, this group was at lower *risk* of disease. How much lower? The incidence in the miotic-treated group, divided by that in the control group, provides the answer: *the relative risk* of field loss in the treated versus the control group.

Sensitivity and specificity

Two additional rates, fundamental to evaluating diagnostic procedures and criteria, deserve mention: sensitivity and specificity. As with the rates already discussed (attack rate, prevalence, incidence, and relative risk), we are already familiar with the underlying concepts. When we subject patients with elevated intraocular pressure to perimetry, tonography, and even long-term miotic therapy, we do so on the assumption that patients with elevated intraocular pressure are likely to have glaucoma, while those with normal pressure are not. Similarly, a vertically oval cup is said to "suggest" true glaucoma (9). The question, however, is not whether vertical ovality "suggests" true glaucoma, but whether it is sufficiently more common among glaucomatous patients than among normals to serve as a distinguishing characteristic. What we really need to know is the regularity with which glaucomatous patients have vertically oval cups, and nonglaucomatous patients lack such cups (10). The former is the *sensitivity* of the criterion or test, the latter the *specificity*.

The *sensitivity* is therefore the proportion of abnormal individuals detected as being abnormals by the parameter in question (screen positive):

$$\text{Sensitivity} = \frac{\text{Number of abnormals who screen positive}}{\text{Total number of abnormals}}$$

The *specificity* is the proportion of normals detected as being normal (screen negative):

$$\text{Specificity} = \frac{\text{Number of normals who screen negative}}{\text{Total number of normals}}$$

Common, shorthand notations for this analysis are given in Table 5. The sensitivity and specificity of an ideal screening test would be 100%, a level rarely achieved.

Tonometry is the commonest form of glaucoma screening, because it is quick and simple to perform, and elevated pressure is presumed to cause the optic atrophy characteristic of the disease. Many individuals with elevated IOP, however, never develop such atrophy. It has therefore become popular to define glaucoma by the presence of classical visual field loss. With this

Table 5
ANALYSIS OF SCREENING PARAMETERS

Result of screening	*Presence of disease*		*Total*
	Yes	*No*	
Positive	a	b	$a + b$
Negative	c	d	$c + d$
Total	$a + c$	$b + d$	$a + c + b + d$

Sensitivity = $a/a + c$
Specificity = $d/b + d$
False positive rate = $b/a + b + c + d$
False negative rate = $c/a + b + c + d$

as our definitive criterion, we can evaluate the accuracy of tonometric screening. Hollows and Graham found that 9% of individuals over the age of forty had an intraocular pressure of 21 mm Hg or higher on a single reading (3). Only 0.4% of this same population had glaucomatous field loss, and only 13 of the 20 persons with field loss also had an elevated pressure (Fig. 1).

SCREENING EFFICIENCY OF TONOMETRY

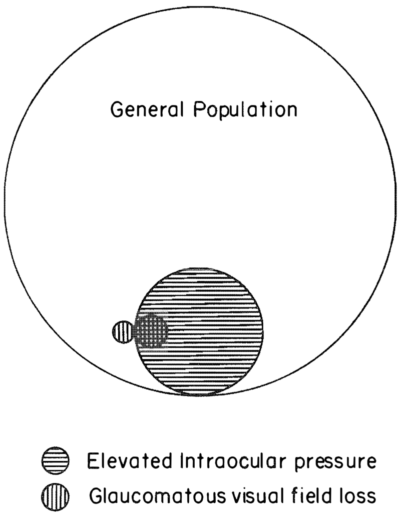

General Population
▤ Elevated Intraocular pressure
▥ Glaucomatous visual field loss

Figure 1. Out of every 1000 individuals examined in a general population, approximately 90 will have an intraocular pressure above 21 mm Hg. Only 3 of these 90, however, will have glaucomatous visual field loss. Almost an equal number of ocular normotensive individuals will also have glaucomatous field loss.

The sensitivity of tonometric screening was therefore 13/20, or 65%, and the specificity, 91.7%. One-third of individuals with established field loss screened negative and would have been denied urgently needed therapy, while over 8% of all normals screened positive, and would have been referred for expensive, possibly anxiety-provoking examinations. Of course a percentage of these "ocular hypertensives" will eventually develop true glaucoma, but the incidence, 1–5 per 1000 per year, is so small, and drop-out rates during follow-up so high, that keeping track of such patients may not be worth the effort.

One can frequently improve the sensitivity of a test by lowering the screening criterion, say from 21 mm Hg to 15. Unfortunately, this almost always results in an even more drastic decline in specificity, making the test even less efficient, and often totally unmanageable.

A valid clinical sign need not always be a useful screening parameter. By narrowing the criterion, one may raise the specificity dramatically, to a point where any patient fulfilling the criterion has a high likelihood of having the disease (10). For example, most patients with an IOP over 40 mm Hg are likely to have, or soon develop, glaucomatous visual field loss (11, 12). This is a useful clinical sign of established or impending glaucomatous field loss, but this heightened specificity is accompanied by a marked loss in sensitivity: many patients with the disease would not satisfy this criterion and therefore would be missed (13).

One of the most important characteristics of the sensitivity/specificity analysis is that the results are entirely independent of the actual proportion of normals and abnormals in the study population. As can be seen in Table 5, each analysis is column specific: sensitivity only involves abnormals, specificity only normals. Such analyses are therefore extremely versatile, and the results in one population are easily compared with those of another, even where the proportions of abnormals in the populations differ. Such is not the case in more traditional false positive/false negative analyses.

False positives and false negatives

The number (or rate) of false positives and negatives depends not only on the sensitivity and specificity of the test, but also on the proportion of abnormals to normals in the study population. As can be seen in Table 5, the analysis is no longer column specific.

In Hollows and Graham's study there were almost 9 false positives and 0.2 false negatives for every 100 individuals screened. This information is not particularly meaningful. Firstly, there is no way of knowing that fully one-third of all the abnormals had been missed. Secondly, these false positive and false negative rates apply only to the distribution of abnormals in this particular population. If the proportion of abnormals in the population had been only half what it was, the false negative rate would have been 0.1%, whereas drawing the study population from a consultant's practice (a common event), where a third or more of the patients may have established field loss, would yield a false negative rate of 11%, even though the sensitivity of the test remained unchanged.

False positive/false negative analyses can be used in two other ways, one useful, the other not. We shall begin with the latter, the common claim that a diagnosis "was correct 91% of the time." This is a summary statistic of little value that conveys even less information than the actual rate of false positives and negatives. In analyzing the value of radioactive phosphorus testing for malignant melanoma, we are interested in learning how many normal eyes screened positive, and were therefore at risk of inappropriate enucleation, and how many abnormals eyes screened negative, and were therefore at risk of going untreated, and the patient perhaps dying. We don't really care that the diagnosis was correct 999 out of 1000 times, since that might simply mean 999 eyes known not to have melanomas (controls) all screened appropriately negative, while the single case with a melanoma inappropriately screened negative as well. The result, 99.9% correct diagnosis, sounds extremely accurate, even though the test was worthless: it was only correct in the huge proportion of patients never suspected of having a melanoma in the first place, and missed the single case of melanoma in the series (Fig. 2).

False positive/false negative rates can be useful, however, in determining the potential efficiency of a screening test. The specificity of tonometric screening, 91%, seems quite high. It would be if it were not for the rarity of true glaucoma in the

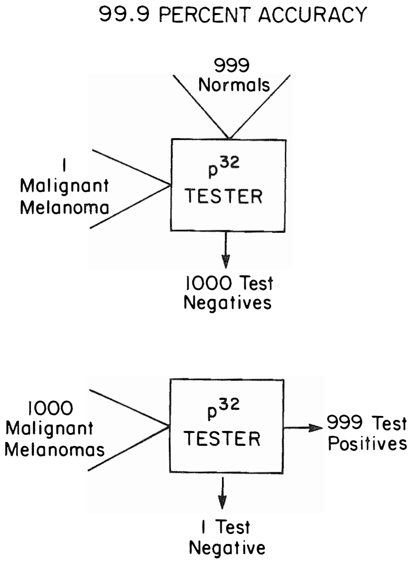


Figure 2. "99.9% accuracy" is a useless statistic which can mean anything from having missed the only abnormal tested to correctly identifying 999 out of every 1000 abnormals examined.

population. Since only 3% of Hollows and Graham's ocular hypertensives had established field loss, 32 "normals" screened positive, and would be referred for expensive, potentially anxiety-provoking evaluation for every 1 abnormal detected, a wholly unsatisfactory ratio. Results based on our consultant's practice would appear more efficient because of the higher proportion of abnormals, hence the higher ratio of true positives to false positives than would be found in the general population.

CLINICAL STUDIES: TECHNIQUES, COMMON SENSE, AND A FEW MORE DEFINITIONS

Most clinical studies fall into one of two categories: prospective or retrospective. Despite a common misconception the dif-

Table 6
COMPARISON OF PROSPECTIVE AND RETROSPECTIVE STUDIES

Prospective study	*Retrospective study*
Initial disease-free group (A) is followed over time, and the number to develop disease (D) and remain free of disease (C) determined:	Initial group with the disease (D) and controls (C) are examined, and the number *without* (D_0; C_0) and with (D_t; C_t) the trait in question determined:
$\underset{\text{Initial disease-free group}}{A} = \underset{\text{Developed disease}}{D} + \underset{\text{Remained free of disease}}{C}$	$\underset{\text{All individuals with the disease}}{D} = \underset{\text{Disease, without trait}}{D_0} + \underset{\text{Disease, with trait}}{D_t}$ $\underset{\text{Controls (no disease)}}{C} = \underset{\text{Controls, without trait}}{C_0} + \underset{\text{Controls, with trait}}{C_t}$
Attack rate (A.R.) $= D/A$	Proportion with trait $= C_t/C;\ D_t/D$
Relative risk of developing disease between two groups, A_t and A_0 (treated and not treated, respectively) is the ratio of their incidence or attack rates:	Relative risk of developing disease between two groups, those *with* and *without* the trait:
Relative Risk $A_t\!:\!A_0 \left\{ \begin{array}{l} = \dfrac{\text{A.R.}_t}{\text{A.R.}_0} \\ = \dfrac{D_t/A_t}{D_0/A_0} \end{array} \right.$	Relative Risk $t\!:\!0 = \dfrac{(D_t)(C_0)}{(D_0)(C_t)}$

ference between them has nothing to do with when the data are collected or analyzed. It is far more fundamental (Table 6).

Prospective studies

A *prospective* or *longitudinal* study begins with a group of individuals free of the disease or trait in question and determines

Table 7
RATE OF SECONDARY HEMORRHAGE IN TRAUMATIC HYPHEMA

Regimen	*Total (number)*	*Secondary hemorrhage*	
		number	*percent*
A	66	12	18
B	71	18	25
Total	137	30	22

Modified from Read and Goldberg (14).

the rate at which it occurs over time (Fig. 3). In other words, prospective studies determine *incidence;* whenever incidence rates are generated, one is dealing with a prospective study. Read and Goldberg (14) found that the rate (incidence) of secondary hemorrhage in traumatic hyphema was not influenced by the treatment regimen (Table 7); Peyman et al. (15) found that the rate (incidence) of postoperative endophthalmitis was significantly reduced by the use of intraocular gentamycin (Table 8). The collaborative Diabetic Retinopathy Study (DRS) (16) showed that photocoagulation retarded the progression of retinopathy and eventual loss of vision (Table 9).

In all instances the authors followed two (or more) groups of

Table 8
INTRAOCULAR GENTAMYCIN AND DEVELOPMENT OF POSTOPERATIVE ENDOPHTHALMITIS

Regimen	*Total eyes (number)*	*Eyes with endophthalmitis*	
		number	*rate/1000*
Gentamycin	1626	6	3.7
No gentamycin	400	11	27.5

Modified from Peyman et al. (15).

Table 9
CUMULATIVE RATES OF FALL IN VISION TO LESS THAN 5/200 AMONG PATIENTS WITH DIABETIC RETINOPATHY

	Therapeutic regimen			
	Photocoagulation		*Control*	
Duration of follow-up (months)	*No. of eyes followed*	*Rate of visual loss (per 100)*	*No. of eyes followed*	*Rate of visual loss (per 100)*
12	1588	2.5	1582	3.8
20	1200	5.3	1166	11.4
28	707	7.4	651	19.6
36	232	10.5	204	26.5

Modified from The Diabetic Retinopathy Study Research Group (16).

individuals and compared the rates at which an unwanted event occurred. The greater the difference between the rates, the more meaningful (clinically significant) it is. The traditional method of expressing this difference is the *relative risk*, which in prospective studies is simply the ratio of the incidence rates.

$$\text{Relative risk} = \frac{\text{Incidence of disease in group 1}}{\text{Incidence of disease in group 2}}$$

The rate of postoperative endophthalmitis among patients denied gentamycin was 3%, versus only 0.4% among those who received it. Those denied gentamycin had 8 times as much risk of endophthalmitis as those who received it (3/0.4). Conversely, gentamycin reduced the risk of endophthalmitis by 87% (100 − (0.4/3 × 100)).

When prospective data are analyzed during the course of a study, it is a *concurrent* prospective study. All the studies above were of this type. When, instead, the data are analyzed considerably later, often in a manner for which they were never intended, it is a *nonconcurrent* prospective study. Our earlier example of endophthalmitis rates among 46,000 postoperative

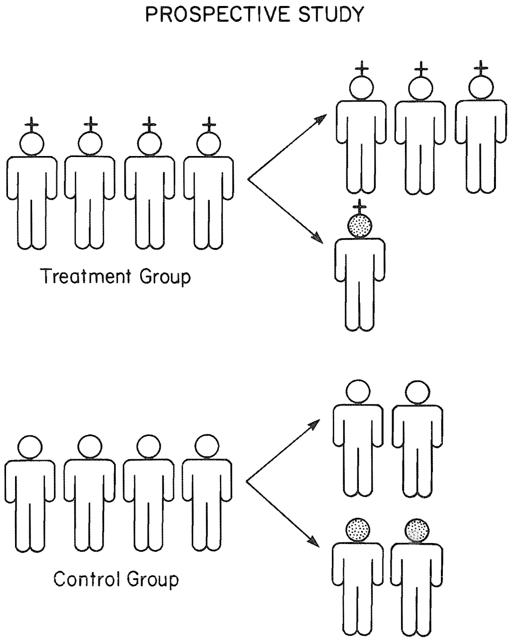


Figure 3. A prospective or longitudinal study begins with a group of individuals free of the disease or trait in question, who are then followed, over time, for its appearance. In this particular example 25% of the treatment group and 50% of the controls developed the disease (spotted faces) during the follow-up period.

cataract patients was a nonconcurrent prospective analysis of data accumulated over a 15-year period.

A classic, nontherapeutic prospective study was the Framingham Heart Study. Smoking habits, serum lipid levels, and a variety of other data were collected on adults being followed for the development of cardiovascular disease. Many years later, another group of investigators took advantage of this accumulated data by searching for etiologic (risk) factors in the development of ocular disease (17). Both the heart and eye studies were prospective, but the heart study was concurrent because the analysis proceeded along with the accumulation of data; the eye study was nonconcurrent, since the analysis took place years later and the data used were not originally collected for that purpose.

Retrospective studies

A *retrospective* or *case control* study always contains at least two groups of individuals: one in which all the individuals already have the disease, and a control group in which they do not. Instead of being followed over a period of time, they are often examined only once, and the frequency with which different factors or characteristics occur in the two groups compared (Fig. 4). If a factor occurs more frequently among abnormals than controls, it is said to be *associated* with the disease and may or may not be of etiologic significance. A good example is the classic study of histoplasmin skin sensitivity among patients with what we now call presumed ocular histoplasmosis syndrome (18) (Table 10). The proportion of patients with classical choroidal lesions who had positive skin tests was significantly higher than among patients with other forms of retinal and uveal disease, and the possible role of histoplasmosis in the etiology of the condition strengthened.

A more recent example is the comparison of diabetes rates among patients hospitalized for cataract extraction versus those hospitalized for other reasons (controls) (19). Diabetes was more common among patients undergoing cataract extraction (Table

Table 10
HISTOPLASMIN SKIN SENSITIVITY AMONG PATIENTS WITH AND WITHOUT OCULAR "HISTOPLASMOSIS"

	Tested (number)	*Positive (number)*	*Positive (percent)*
Classical lesion present	61	57	93
Other forms of retinal-uveal disease	190	48	25

Modified from Van Metre and Maumenee (18).

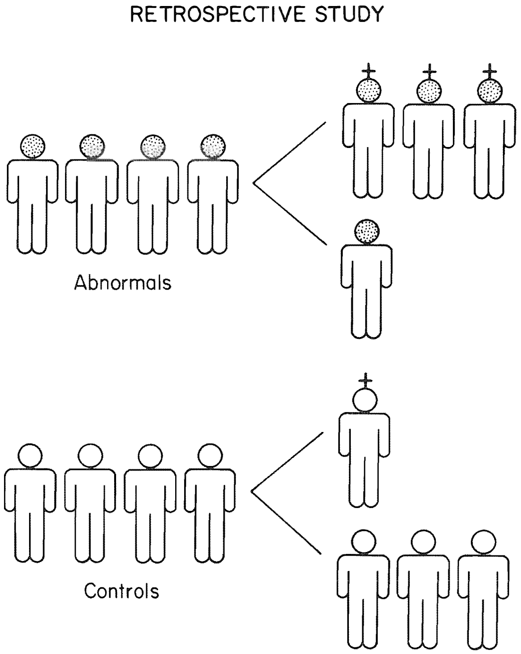


Figure 4. A retrospective or case-control study always begins with individuals who already have the disease, and a closely matched group who do not. Both groups are examined for the presence of one or more characteristics thought to be associated with (perhaps the cause of) the disease. In this particular example, 75% of abnormals but only 25% of controls have the trait in question.

Table 11
PREVALENCE OF DIABETES AMONG HOSPITALIZED PATIENTS 40-49 YEARS OF AGE

Reason hospitalized	*Total patients (number)*	*Number with diabetes*	*Prevalence of diabetes (per 100)*
Senile cataract extraction	60	7	12
Fractures, etc.	1098	30	3

Modified from Hiller and Kahn (19).

11), suggesting that diabetics are more likely to undergo cataract extraction than are nondiabetics. How much more? Unlike the prospective study, the retrospective study does not produce incidence or attack rates: the groups are enrolled on the basis of whether or not they already have the disease. Absence of incidence data forces us to resort to a more complex, less intuitive calculation of relative risk than that used so far (Table 12).

Retrospective studies are usually less expensive and time consuming than prospective ones, but they are also less powerful: it is more difficult to choose appropriate controls; there is increased risk of hidden bias; and they do not produce incidence rates. While they can be a useful means of choosing between several avenues of investigation, design, and analysis are best left to experienced epidemiologists.

Perhaps the single most important topic in this manual, and

Table 12
RELATIVE RISK OF SENILE CATARACT EXTRACTION IN DIABETICS CALCULATED FROM RETROSPECTIVE (CASE-CONTROL) STUDY[1]

	Patient classification	
Diabetes present	*Cataract extraction ("disease")*	*Fracture, etc. ("control")*
yes	D_t (7)	C_t (30)
no	D_0 (53)	C_0 (1068)
Relative risk (diabetics: controls)	$= \dfrac{(D_t)(C_0)}{(D_0)(C_t)}$	
	$= \dfrac{(7)(1068)}{(53)(30)}$	
	= 4.7:1	

[1] Raw data shown in Table 11.

Modified from Hiller and Kahn (19).

the most intuitive, is the epidemiologist's common sense approach to study design. Simply stated, it means never losing sight of the many extraneous factors that can affect a study's outcome. Foremost among them is bias.

Bias and its control

Results are said to be biased when they reflect extraneous, often unrecognized, influences instead of the factors under investigation. Potential sources of bias include the choice and allocation of subjects, their perceptions, and the investigator's expectations.

Table 13 lists each of these sources of bias and methods for their control.

SAMPLING BIAS

Samples chosen for comparison should be as alike as possible except for the factors under investigation. In a prospective comparison of two topical antihypertensive agents for example, the two groups of ocular hypertensives should differ only in the

Table 13
BIAS AND ITS CONTROL

Source	*Methods of control*
Selection bias	Eligibility criteria rigidly fixed and followed
	Subjects allocated only after enrollment
	Randomization
	Matching and stratification
	Tracing those lost to follow-up
Patient bias	Placebos/cross-over study
	Masking
	Firm, objective endpoints
Observer bias	Controls
	Masking
	Firm, objective, well-defined endpoints
	Standardization
	Measurement of reproducibility

medication they use. When this is not the case, and the samples differ in some meaningful, systematic way that influences the results, sampling bias is present.

Retrospective studies are particularly prone to sampling bias. In comparing the prevalence of diabetes among cataract patients and controls the investigators assumed that the two groups were comparable except for the cataract operation and whatever etiologic or risk factors were related to it (diabetes). If they had been less careful, the controls (fracture patients) might have been younger than the abnormals (cataract patients). Since the prevalence of diabetes increases with age, the prevalence of diabetes among the older, cataract group would have been higher than among controls regardless of whether or not diabetics are more likely to require cataract extraction.

Other investigators attempted to show, through an unfortunate confusion of retrospective and prospective techniques, that keratoconus was more likely to follow the use of hard contact lenses than soft contact lenses (20). They concluded that the use of hard contact lenses increased the risk of developing keratoconus. In the absence of careful matching, this was a hazardous comparison. Many keratoconus patients undoubtedly become symptomatic from irregular or high degrees of astigmatism before the true nature of their disease is recognized. Since irregular or high degrees of astigmatism require use of hard instead of soft contact lenses, any general population of hard contact-lens wearers will automatically contain a higher proportion of future keratoconus patients than a population of patients using soft contact lenses. In other words the deck was loaded: use of hard contact lenses was bound to be associated with keratoconus, whether or not it contributed to the development of that disease.

In another analysis, the rate of metastatic deaths among patients who underwent enucleation for choroidal melanomas (of all sizes and degrees of invasiveness) was unfavorably compared with that among patients without enucleation who were followed (21). The authors concluded that enucleation was responsible for the increased mortality in the first group, despite the fact that those without enucleation almost invariably harbored small tumors of questionable malignancy.

Sampling bias can also occur in prospective studies, through "bad luck" or the investigator's subconscious bias in recruitment and allocation of subjects. For example, an investigator already convinced of the danger of intraocular lenses in selected conditions might unknowingly but consistently assign eyes with significant pathology to the nonimplantation group. Regardless of the true complication rate in the two procedures, the non-IOL group would be already weighted with less favorable results.

The same conditions occur when a patient determines his own therapy, or when two independent series are compared. In a particularly careful, but nonconcurrent, prospective study of the safety of intraocular lenses, patients who had received implants at the time of cataract extraction were matched with those who had not, since selection criteria for the implant group had been stricter (22). Even so, the authors acknowledged at least one potential source of bias: patients with extensive corneal guttata were more likely to undergo routine cataract extraction than intraocular lens insertion. In other words, the two groups had originally differed by more than just the choice of operation. Although intraocular lenses were still associated with a higher incidence of postoperative corneal edema, this difference would probably have been even greater had the two series been more closely matched.

RANDOMIZATION

Perhaps the greatest virtue of the concurrent prospective study is the availability of a powerful technique for minimizing selection bias: randomization. Randomization ensures that every subject has exactly the same chance of being assigned to each of the study groups. This takes the decision out of the hands of the clinician; it is the only sampling scheme amenable to routine statistical manipulation; and it distributes patients (on the average) equally between the groups irrespective of personal attributes—both obvious (like age and sex) and unrecognized.

It is always a good idea to test the success of randomization, since bad luck or hidden bias could have been present. While it is impossible to keep track, or even be aware, of every potentially important parameter by which the two groups can differ, the distribution of readily recognized factors should be compared. If the randomization was successful, these nonmatched attributes should be evenly distributed.

Despite careful matching in Jaffe's series (22), the prevalence of senile macular choroidal degeneration (SMCD) among those undergoing routine cataract extraction was twice that among IOL recipients, even though SMCD is a common indication for receiving an IOL. Since this unmatched variable, SMCD, differed appreciably in the two groups, one wonders what other factors, unrecognized but pertinent to the study's outcome, might also have varied, influencing the results.

"Random" is not synonymous with "haphazard." Even when haphazard assignment does not appear at first glance to have any consistent pattern, it almost invariably does. Some biases are obvious: a series of patients receiving a radical new procedure compared with a group who refused it. Some more subtle: in one sampling scheme patients admitted on Mondays, Wednesdays, and Fridays received one form of therapy; those admitted on Tuesdays, Thursdays, and Saturdays another. Referring physicians quickly learned which therapy was given on which days, and arranged for their patients to arrive on a day when they would receive the therapy the physician preferred.

This form of bias is best controlled by having collaborating clinicians agree in advance on uniform criteria for all therapeutic regimens. Once these are met, and the patient agrees to enter the study regardless of the treatment assigned, he or she is randomly allocated to one of the groups.

The selection process need not be left entirely to chance (random assignment). Although the composition of the different study groups is likely to be equivalent when large numbers of patients are involved, chance variation (a euphemism for bad luck) may lead to less successful allocations with smaller samples. If a particular attribute is likely to have an important influence on the outcome of the trial, the two groups can be deliberately matched for that attribute. For example, we would expect patient age to influence any comparison of extracapsular and intracapsular cataract extraction. Rather than risk the chance that most of the younger patients will end up undergoing one procedure and most older patients the other, we can begin the allocation by first separating all patients into two strata: those under and those over thirty years of age. The first patient registered in each stratum is randomly assigned to one of the two operative techniques, the next person in the same age stratum automatically receiving the other technique. The process is then repeated: the third patient is randomly assigned to one technique, the fourth automatically going to the other, etc. Study groups can be matched on age, sex, race, all three, or any combination of factors that seems important, although in practice the small number of subjects involved usually limits the number of variables on which they can be matched.

Various methods are available for randomizing a series. Perhaps the simplest is to assign each patient a number from a serial list of random numbers (Appendix 1). Since every digit has an equal chance of appearing at every position on the list, there is no pattern or bias in their arrangement. If the patients are being divided into two groups, those whose random numbers end in an odd digit go to one group, those with an even digit to the other. If three groups are involved, patients whose numbers end in 1, 2, or 3 go to the first group, 4, 5, or 6 to the second, and 7, 8, or 9 to the third. Random numbers ending in zero are skipped and not assigned to any study subjects.

STANDARDIZATION AND ADJUSTMENT

Despite attempts at randomization, or more commonly where it was not or could not be used, the different groups (abnormals and controls, treated and untreated, etc.) might well vary in ways that could influence the outcome. We've already discussed examples in which a difference in age distribution, or in the proportion of patients with irregular astigmatism, might have existed and led to biased, potentially erroneous results. Proper comparisons require age-specific or astigmatism-specific analyses, where comparisons are made between subgroups of abnormals and controls of the same age, refractive error, and the like. Alternatively, when these individual subgroups contain too few cases for comparison, the entire group of abnormals and controls can be adjusted to the same "standard" distribution. Both age-specific and age-adjusted rates were used in the analysis of glaucomatous blindness registry data (see Tables 2 and 3).

LOSS TO FOLLOW-UP

One of the most common sources of sampling bias, unrelated to the selection process itself, is loss to follow-up. Except for small, captive populations, all prospective studies will lose patients over time. At best, this only reduces the number of subjects left to work with—the reason we usually recruit more than the number required for purely statistical purposes. At worst, however, loss to follow-up can result in a horribly distorted sample and biased conclusions. In an obvious, if exaggerated, example, any significant loss to follow-up would seriously compromise a comparison of photocoagulation versus enucleation in the management of choroidal melanomas. Missing patients merely recorded as "lost to follow-up" may well have died from metastatic disease. If all such metastatic deaths occurred in the nonsurgical group, and the fact that they died went unrecognized, photocoagulation would appear safer than it really is (perhaps even safer than enucleation, when the facts might be quite the opposite).

Good doctor-patient rapport with frequent emphasis on the necessity of remaining in contact with the project tends to minimize loss to follow-up. Arrangements can be made for those who have moved outside the study area to be followed by a local physician. Even so, some participants will inevitably disappear. It is then important to determine whether the rate at which this occurs, and the characteristics of those who have disappeared, varies from one study group to another, and, if so, if such variation is likely to affect the study's outcome. Those who remain in the study should also be compared with those who disappear, and intensive efforts should be made to trace at least a random sample of the missing subjects (by visiting their last known address, questioning neighbors as to their whereabouts and health status, searching death reports, etc.).

CONTROLS

Concerns about randomization, selection bias, and the like all presuppose the use of controls. Except for unusual case reports, any study of substance should be expected to contain suitable controls. The history of our profession is replete with uncontrolled studies reporting significant therapeutic advances, many of which turned out, on controlled examination, to be no better than placebos. The frequently repeated argument that concurrent controls were unnecessary to prove penicillin effective in pneumococcal pneumonia is spurious. Few ophthalmic diseases yield so dramatically to a single intervention. More commonly we deal with conditions having highly variable outcomes and treatments providing only a modicum of benefit. To date, there has not been a single well-controlled study demonstrating that photocoagulation benefits patients having presumed ocular histoplasmosis (POH) or that Daraprim is effective in treating human toxoplasmic chorioretinitis. To the contrary, despite numerous testimonials to the benefit of photocoagulation in POH, recent nonconcurrent comparisons suggest that little is gained from this mode of therapy (23).

As already noted, controls ideally should match the treatment group in all respects except for the therapeutic regimen. The closer the match, the more meaningful the results. In most instances we construct matching groups of individuals, as alike as possible in age, sex, race, and whatever characteristics are peculiar to the disease in question, such as level of intraocular pressure and degree of field loss; size and location of the subretinal neovascular net; degree, form, and location of proliferative retinopathy; and size and characteristics of the choroidal melanoma.

No matter how good the match, it is never perfect. Individuals always differ. Depending on the size of these differences, this innate variation, or "background noise," can mask small but definite therapeutic advantages. In some instances we can eliminate individual variation by matching the two eyes of a single subject, using a hard contact lens in one, a soft lens in the other; one form of antihypertensive medication in one, an alternative form in the other; retinoic acid in one, placebo in the other (24), etc. But the two eyes of the same individual can differ. Even this difference can be eliminated by applying first one agent and then the other to the same eye, a so-called *crossover* study. A single eye then becomes its own matched control. Examples include testing various topical antihypertensive agents in the same eye (25); various carbonic anhydrase inhibitors in the same person (26); etc. To ensure that the effect of the first drug doesn't influence the second, different eyes or individuals should initiate the series with different drugs. By minimizing extraneous variables (background noise), these *paired* comparisons are usually stronger than group (independent) comparisons, and may demonstrate statistical and clinical significance which might otherwise be missed.

Nonconcurrent, historical controls are obviously much weaker, since many other factors (patient selection, operative techniques, medications, length of hospitalization, etc.) may change simultaneously with the factors under investigation. For example, traumatic hyphemas treated with a new agent were compared with those treated many years before (27). The investigator acknowledged that many other variables changed in the interim, but chose to ignore them, attributing all the benefit to the new agent. The reader might seriously question his conclusions.

Factors that change in the interim need not be obvious. Initial search of Christy's cataract series identified only a single change in technique: introduction of periocular penicillin prophylaxis in May 1972 (ref. 1 and N. Christy, A. Sommer, unpublished data). The series was therefore divided into two periods, pre- and postpenicillin. For interest, each subseries was divided further, at its midpoint. Analysis indicated that there was a drop in infection rates coinciding with the introduction of penicillin prophylaxis (Table 14). Surprisingly, however, the rates continued to decline during the ensuing period. Further investigation revealed an additional, previously overlooked change: intensive preoperative topical application of chloramphenicol-sulfadimidine was instituted in January 1973. Dividing the postpenicillin subseries at this point revealed that penicillin alone had not influenced endophthalmitis rates in the least. All the improvement followed the addition of topical chloramphenicol to the prophylactic regimen (see Table 1). A prospective, concurrently controlled

Table 14
POSTOPERATIVE ENDOPHTHALMITIS AND INTRODUCTION OF PENICILLIN PROPHYLAXIS

			Infections	
Series	*Prophylactic penicillin*	*Operations (number)*	*number*	*rate per 1000*
Ia Jan. '63–Dec. '67	–	9714	54	5.6
Ib Jan. '68–May '72	–	12,340	55	4.5
II May '72–Dec. '74	+	12,957	30	2.3[1]
III Jan. '75–Dec. '76	+	10,481	9	0.8[1]

[1] Compare with rates shown in Table 1, where series II and III are divided at a different point in time.

Modified from Christy and Sommer (1) and N. Christy, A. Sommer (unpublished data).

Table 15
INCIDENCE OF POSTOPERATIVE ENDOPHTHALMITIS IN MASKED RANDOMIZED PROSPECTIVE TRIAL

	Prophylactic regimen			*Infections*	
Series	*Penicillin*	*Chloramphenicol-sulfadimidine*	*Operations (number)*	*number*	*rate per 1000*
IV	–	+	3309	15	4.5
V	+	+	3309	5	1.5

Compare with rates for differing regimens in nonconcurrent series, Table 1.
Modified from Christy and Sommer (1).

trial then demonstrated that chloramphenicol alone was equally ineffective, all the benefit arising from combined prophylaxis (Table 15) (1).

PATIENT BIAS

A patient's expectations can seriously influence the results of therapy. If he believes that he is receiving the latest, most technologically advanced treatment (flashing lights, darkened rooms, and nervous doctors all contribute to his perceptions) he is likely to experience the greatest subjective improvement. At least two techniques are useful in controlling this form of bias. The first is *masking.* The patient is not told which treatment (if any) he is receiving. The use of a *placebo* (or alternative medication) helps to keep the patient masked. A *crossover* study, in which the patient first uses one and then another drug (in masked fashion) achieves the same effect with the additional advantage of testing two (or more) agents instead of one, with a minimum of background noise. Of course it is sometimes difficult for placebo or sham regimens to duplicate the conditions of actual therapy (e.g., laser burns). The second is reliance on "hard," objective criteria, e.g., actual visual acuity or millimeters of proptosis rather than how the patient "feels."

OBSERVER BIAS

An investigator's conscious or subconscious (unconscious?) expectations can greatly affect a study. We've already discussed how he can influence the selection and allocation of subjects. He can also influence the patient's (and his own) perceptions of the benefits of treatment. If already convinced of the value of a particular mode of therapy, the investigator can always coax another line from the Snellen chart. Whenever possible, the person compiling the clinical observations should be masked: not know whether he is examining an abnormal or control, a treated or untreated patient. It is obviously impossible to keep a clinician from knowing that an intraocular lens was inserted or panperipheral ablation performed. But it is possible to keep this information from the technician recording best-corrected acuity or intraocular pressure.

When treatment assignments are masked from both patient and observer, we have the classic *double blind* (preferably *double masked*) study. Closely related to observer bias is observer variation.

Observer variation and reproducibility

Just as two clinicians examining the same patient may arrive at different diagnoses, two observers will not necessarily record the same findings; nor, necessarily, will the same observer examining the patient a second time. The former is known as *interobserver*, the latter *intraobserver* variation. Forty stereofundus photographs read twice in masked fashion by the same individual demonstrated the (intraobserver) variation shown in Tables 16 and 17 (28). Relatively little can be done to decrease innate variability of this sort short of switching to other, more objective criteria.

Interobserver variation is usually greater; the problem is compounded by biases peculiar to each observer. One observer may be more apt to diagnose cataracts, macular pigment disturbance, myopic cups, and the like than another, or consistently estimate

Table 16
VARIABILITY OF REPEAT ESTIMATIONS OF CUP RADII BY A SINGLE OBSERVER

	Differences (in tenths of a disc diameter)[1]	
	Mean	*Standard deviation*
Temporal	0.31	0.40
Nasal	0.27	0.48
Superior	0.19	0.22
Inferior	0.27	0.34

[1] Differences between two readings for each of 40 eyes.

Modified from Sommer et al. (28).

cup/disc ratios as larger (Table 18) (29). Clear, detailed criteria (including reference photos where appropriate) and frequent standardization will reduce these biases and minimize interobserver variation.

Reproducibility should be repeatedly tested, in masked fashion, to ensure maintenance of standardization, and to quantify

Table 17
VARIABILITY OF REPEAT ESTIMATIONS OF CUP RADII BY A SINGLE OBSERVER

	Frequency distribution of difference (in tenths of a disc diameter)		
	≤ 1.0	*≤ 2.0*	*≤ 3.0*
Temporal	38	2	0
Nasal	40	0	0
Inferior	39	1	0
Superior	40	0	0

Modified from Sommer et al. (28).

Table 18
INTEROBSERVER VARIABILITY FOR DIFFERENT CRITERIA

	Percent of patients diagnosed positive by observer				
	1	*2*	*3*	*4*	*5*
Horizontal C/D < 0.3	78	62	64	43	69
Macular pigment disturbance	41	5	19	27	19
Myopic cup	53	2	7	3	0

Modified from Kahn et al. (29).

the magnitude of variation attributable solely to the lack of perfect reproducibility. Obviously, criteria with firm, quantifiable end points (intraocular pressure) will be far more reproducible than those more subjective and difficult to define (e.g., presence of a cataract).

With all their variability, the observers listed in Table 18 worked from a common manual in which all diagnostic criteria were precisely defined. If this is not the case, variability is likely to be much larger. One man's cup (or cataract) is not always another's, a major problem in comparing the results of independent studies (or investigators).

THE STATISTICAL INTERFACE

To complete our discussion of epidemiologic principles and study design, we must introduce two major uses of statistics: choice of sample size and tests of "significance." Both will be discussed in greater detail later. They are the easiest, quickest, most straightforward part of any study.

Sample size

One of the earliest, most important determinations an investigator can make is the size of the sample required to test his hypothesis. Long before the grant application is complete, a single form designed, a research assistant hired, or patient examined, he will learn whether the study can be smaller than originally anticipated, or (more commonly) must be far larger. In fact, the sample size required may prove so large as to be impractical. Better to discover this at an early stage than many frustrating years later.

Sample size is just as important to the critical reader. Numerous investigators have reported the absence of a statistically significant difference between treatment regimens, implying that they were of equal efficacy, when the samples were insufficient to demonstrate all but the most spectacular of results. In a study of unilateral versus bilateral ocular patching for traumatic hyphema (30), the sample size was so small that bilateral patching would have had to reduce the incidence of secondary hemorrhages by over 80% to have had a reasonable expectation of being proven significant. Since most therapeutic benefits are considerably smaller, it is hardly surprising that the authors found no statistical difference.

Test of significance

For the present, it is sufficient to recognize that we commonly employ statistical tests for a single purpose: to determine the likelihood or probability that the difference observed between two (or more) groups (e.g., treated versus placebo) might have arisen purely by chance. The famous notation "$p < .05$" is simply shorthand for "the likelihood that we would have observed this large a difference between the two groups, when in fact there was no real difference between them, is less than 5 in 100." When $p < .01$, the likelihood is even smaller, less than 1 in 100.

Purely by convention (nothing sacred or magical about it), we begin to consider that some factors other than chance may have been responsible for the difference when the likelihood of its being due to chance alone is less than 5 in 100. Our confidence in this "other factor," i.e., its *statistical significance*, increases as the likelihood of the difference being due to chance recedes ($p < .05, < .01, < .001$, etc.).

An important distinction, often overlooked, is that there is absolutely no way of *proving* that a new treatment is beneficial, only that the observed difference is unlikely to have arisen by chance. Conversely, with small samples even large differences can occur purely by chance (e.g., $p < .50$). This does not mean that the treatment is not beneficial; only that the possibility of chance producing a difference of this size is so large that it is impossible to demonstrate the "significance" of the treatment effect.

Clinical versus statistical significance

In a rush to conform with new scientific standards, many articles conclude that "the differences are highly significant." In regard to what? In most instances the author´means they are *statistically* significant. Rarely indicated is whether the differences, real or not, are large enough to make any practical difference. An operation that has a success rate of 85% may be *statistically* significantly better than one with a success rate of 84.6% (that is, the 0.4% improvement is likely to be real), but the *clinical* (or practical) significance is nil (especially since the "superior" procedure may have other mitigating characteristics in cost, complexity, speed, and the like). Similarly, one drug might heal herpes simplex ulcers in 6.7 days, while a placebo takes 7.0 days. The improvement is real, but is it *clinically* significant? Fuller Albright was probably expressing this principle when he said of statistical methods: ". . . if you have to use them, I don't believe it" (31). In general, if there is a meaningful difference it should already be obvious. A statistical test of significance merely establishes the risk entailed in assuming that the difference was not due to chance. One should examine the level of benefit in light of competing aspects (cost, side effect, etc.), before accepting any new treatment as a meaningful therapeutic advance.

Statistical associations and epidemiologic inferences

The goal of most studies is to determine whether two (or more) parameters are associated with one another. We have seen that combined prophylaxis was associated with a lower rate of postoperative endophthalmitis; diabetes with a higher rate of cataract extraction; and presumed ocular histoplasmosis syndrome with a higher prevalence of histoplasmin skin-test sensitivity. In each instance the association was statistically significant: the probability was less than 5 in 100 that it could have arisen by chance alone. We must now consider why two parameters might appear to be associated, and what inferences and conclusions can be drawn from that fact.

ASSOCIATION

Every association has several possible explanations (Table 19). Firstly, the association might not be real: with typical "bad luck" we may be dealing with one of those 5 in 100, or even 1 in 1000 instances in which the association is due entirely to chance. A statistically significant event at the .05 level will occur by chance alone once in every 20 observations. This may have been one of them. Unfortunately, no method is really adequate for dealing with this problem. When Weber et al. (32) made 460 comparisons (at the $p < .05$ level) between viral antibody titers and various forms of uveitis, they were at risk of detecting 23 (460/20) mean-

Table 19
POSSIBLE EXPLANATIONS FOR A STATISTICAL ASSOCIATION

A. Spurious
   1. Chance event
   2. Biased study
B. Real
   1. Indirect (linked through common third factor)
   2. Direct (possibly etiologic)

ingless associations. They handled the problem by "deleting many statistically significant correlations that did not seem sensible." Investigators studying diabetic retinopathy (16) handled the problem somewhat differently. Rather than assign probabilities to the apparent associations, they simply reported the values of the statistical calculations (e.g., chi-square). The reader was provided the opportunity of considering the alternatives for himself.

Alternatively, it may have been bias rather than chance that caused the spurious association. As already discussed, retrospective studies have far greater potential for spurious associations than prospective studies, since randomization, which leads to a more uniform distribution of unrecognized but potentially important factors, cannot be employed.

Assuming that the association is real, the question still remains as to whether it is direct, with possible etiologic significance, or indirect, with the two characteristics being linked through their common association with some third factor. Early in this century a xerophthalmia epidemic in Denmark was traced to increased margarine consumption (33). The association was real but not really causal. Poorer segments of society had substituted margarine, devoid of vitamin A, for more expensive dairy products. It was this lack of vitamin A, rather than any toxic substance in the margarine, that caused the epidemic.

Similarly, trachoma is associated with hot, dry climates. While the association is real, it is not direct. Rather it is a somewhat complex interaction between the dry climate, lack of water, and inadequate personal hygiene, and their effect on transmission of the trachoma agent and contributory conjunctivitis.

To borrow a classic example from outside the realm of ophthalmology, early epidemiologic studies demonstrated that bronchogenic carcinoma of the lung was confined almost exclusively to men. The association between a person's sex and risk of cancer however was not direct, and the malignancy not related to male genes or hormones. Instead, early in the century cigarette smoking was a male habit and females rarely indulged. Maleness was associated with smoking, and of course smoking was associated with bronchogenic carcinoma. Although the association between maleness and bronchogenic carcinoma was real, it was indirect, through their mutual association with smoking. Just as smoking habits among women have since approached those of men, so has their risk of lung cancer.

In summary, an association may be spurious or real; if real, it may be indirect or direct. Numerous methods exist for evaluating the association further: replicating the experiment using different populations; employing different, randomized controls; approaching the relationship from a different perspective, using multiple graduated comparisons and correlated laboratory experiments, etc. For the present, it is sufficient merely to keep in mind that a statistically significant association can represent many different things.

INFERENCE

An association must be precisely described and no more must be claimed than was actually demonstrated. A striking error in this regard is the oft-repeated assertion that senile cataracts are more common among diabetics than nondiabetics, implying that impaired glucose metabolism is important in the etiology of senile cataracts. In fact, there is little epidemiologic evidence to support this contention. What most studies have shown is that diabetics are more likely to undergo cataract extraction than nondiabetics. While the distinction may seem subtle, it has the greatest implications (34). If, in fact, senile cataracts were more common among diabetics, a massive effort should be launched to identify a useful prophylactic agent. Since all we really know is that diabetics have a higher rate of cataract extraction, we must first determine whether they are at greater risk of developing a cataract, or only of having it removed. It is reasonable to suspect that diabetics are referred to and examined by ophthalmologists more frequently than are nondiabetics, and hence more likely to have their cataracts identified and removed. Had investigators kept careful sight of what they had actually demonstrated (diabetics are more likely to undergo cataract extraction), and not confused it with what they inferred (diabetics are at increased risk of developing cataracts), the latter, merely an inference, would not now be established "fact."

We did not fall into this trap with an observation already discussed: the inordinate risk of registerable glaucoma blindness among blacks. No one claims that blacks are at higher risk of glaucoma. They might be, but they might also have a more severe form of glaucoma, be less responsive to medications, make less use of health services, etc., or be more assiduous in registering their blindness than are other segments of society.

Having demonstrated and precisely defined an association that appears to be real and direct (e.g., a new therapeutic agent), how widely applicable are the results? No one goes to the trouble of doing a study just to prove something about those who participated in it. The object is to extrapolate the findings to other peoples and places. When Hiller and Kahn (19) demonstrated that patients with diabetes were at increased risk of cataract extraction, they were not just interested in those few patients studied, but in inferring a fundamental principle about diabetics in general.

Ideally, a random sample of the entire relevant population should be studied. In practice, this is rarely feasible. Instead, we assume that similar persons will respond in similar ways. One must take every precaution, however, to ensure that they are indeed similar. For example, a history of night blindness proved to be an effective tool for xerophthalmia screening in Java (35). But recognizing and expressing an accurate history of night blindness is probably culturally dependent. There's no guarantee that it will work equally well in India. The demonstration that intraocular gentamycin could reduce the incidence of postoperative endophthalmitis in Indian cataract camps from 3% to 0.4% does not mean it will cause any reduction in postoperative endophthalmitis in modern, well-equipped hospitals where the rate is already 0.1% and the source and character of causative agents are probably different. The first report from the DRS indicated that photocoagulation was effective in retarding neovascular proliferation and visual loss (36)—not in all diabetics, nor in all diabetics with neovascularization, but only in a highly selected subgroup. Had the patients not been rigidly stratified, the beneficial affects of photocoagulation might have been missed entirely, and even if not, the results may have been inappropriately applied.

PUTTING IT ALL TOGETHER

Having completed our discussion of epidemiologic principles and techniques, we should pause for a brief review. As stated at the outset, epidemiology is characterized by a preference for rates (summarized in Table 20), rather than absolute numbers,

Table 20
BASIC RATES

1. Attack rate	$\frac{\text{number who develop attribute}}{\text{number followed}}$
2. Incidence	$\frac{\text{number who develop attribute over specified period of time}}{\text{number followed for specified period of time}}$
3. Prevalence	$\frac{\text{number with attribute}}{\text{number examined}}$
4. Relative risk	$\frac{\text{incidence in group 1}}{\text{incidence in group 2}}$
(retrospective study)[1]	$\frac{\text{prevalence in group 1}}{\text{prevalence in group 2}}$
5. Sensitivity	$\frac{\text{number of abnormals screening abnormal}}{\text{total number of abnormals}}$
6. Specificity	$\frac{\text{number of normals screening normal}}{\text{total number of normals}}$

[1] Only in special instances. True relative risk requires the more complex analysis shown in Table 6.

and a particularly thoughtful approach to the conduct and interpretation of studies. More detailed discussions of epidemiologic principles and techniques can be found in references E-1 through E-5 of Appendix 5.

Organization of clinical studies—getting started and staying on track

All clinical investigations adhere to Murphy's law: "Everything that can go wrong, will." It operates at every step and phase of a study, the potential for problems (especially in the critical areas of selection, patient and observer bias) increasing geometrically with the number of investigators, and exponentially with the number of centers involved.

Although Murphy's law can never be fully circumvented, careful attention to detail will minimize its impact. Some of the more critical steps in planning and executing clinical studies are summarized in Table 21.

Is it really worth quoting?

Whereas the investigator's job is to reduce the impact of Murphy's law the reader's job is to identify every instance in which he failed. Before all else, the reader must satisfy himself that a study was conducted and interpreted properly and that its conclusions, regardless of how dramatic and potentially important they seem, are likely to be valid. This is only possible where study methods are described in sufficient detail for critical evaluation. Where they are not, the reader must be wary of accepting the results.

Table 21 highlights areas where most mistakes occur and to which the reader should pay the greatest attention. Was the sample size really adequate to disprove the value of the drug or procedure, and at what level of potential benefit (10% or 80%)? Were the observers adequately masked and standardized, and how large was the interobserver variation? Were techniques for allocating patients and investigating those lost to follow-up suf-

Table 21
CRITICAL STEPS IN A CLINICAL STUDY

Step	*Comment*
1. Define specific goal(s)	Diffuse fishing expeditions often get nowhere
2. Review literature thoroughly	May discover question already answered, better ways to design study, other areas worth considering, and background data required to determine sample size.
3. Select sample size	May need more or fewer patients than originally anticipated. Required sample size may prove so large that study is impractical
4. Establish, standardize, and quantify reproducibility of forms, procedures, and personnel before starting study (pilot trial) and at frequent intervals thereafter	Poor, unreproducible questions and procedures can be modified or replaced without loss of study data; magnitude of intra- and interobserver variation must be known for analysis of results.
5. Prepare *detailed* protocol	Ready reference of all procedures; and basis of "introduction," "methods," and "discussion" sections in final report.
6. If therapeutic trial, every patient meeting criteria is offered participation. Only after they accept are they randomized.	Randomization before enrollment introduces potential bias.
7. If case-control study, examine matching carefully to rule out inappropriate or biased controls	Selection of controls most critical part of study; can easily result in biased sample and results.
8. Repeatedly check monitorable data (age, sex, race, etc.) between groups to ensure randomization (if applied) is functioning properly.	A breakdown of randomization can occur at any time.
9. Determine whether masking remains effective.	If code inadvertently broken, observations may be biased.
10. Conduct repeated, specific searches for bias.	If bias is discovered only after study completed, study may not be salvageable.

Table 21 (cont.)

Step	*Comment*
11. Review cases lost to follow-up for consistent pattern that might explain results. Where possible, trace a random subsample to establish definite outcome.	Patients lost to follow-up represent a potentially important source of selection bias.
12. Where indicated, subject all results to rigorous statistical tests.	Required to prove point and necessary for publication. But do not disregard obvious or potential differences just because they are not statistically significant. Sample size may simply be too small because of error in original assumptions.
13. Estimate *clinical*, as well as statistical, significance of results	The two are not synonymous and should not be confused.
14. Consider alternative explanations for any apparent associations.	Not all associations are real, meaningful, or of direct, etiologic or therapeutic significapce.
15. Inferences should be strictly grounded in actual observations.	The greater the distance between inference and actual observation, the more hypothetical and less meaningful the inference.

ficient to rule out sizeable selection bias? Could the degree of patient, observer, or selection bias account for the results? Were all reasonable explanations for apparent associations explored or only those the author originally hypothesized? Were the author's conclusions warranted by his data, or did his inferences wander far afield? Passing muster with a journal's referees is no guarantee that an article proves what it claims (37-39).

Statistics

SIMPLE CONCEPTS AND COOKBOOK FORMULAS

Statistics are a rather simple matter, at least as regards the vast majority of clinical studies; they are absolutely critical to any well-designed study with the least of pretensions; and they have become a rather practical necessity as the better journals begin to impose a semblance of science on our communications.

What's it about and why do I need it?

Statistics are simply a means of expressing probabilities. Why bother? Because we never deal in absolute truths (a probability of 1.0), or at least we need some way of recognizing how close we come to them. In the commonest example, we wish to know whether a difference observed between two groups, say the rate of postoperative endophthalmitis or histoplasmin skin test sensitivity, is real and likely to recur in repeated experiments, or simply the result of chance variation—likely to vary from experiment to experiment and disappear entirely when averaged over repeated investigations. If a hundred people flip the proverbial coin 10 times, they will, on the average, end up with 5 heads and 5 tails. While true of the average, this is not true for every individual person. We'd be very surprised indeed if some people didn't get 6, 7, 8, or even occasionally 10 heads on the basis of chance variation, although intuitively we recognize that the greater the deviation from the average, the less frequently it should occur. One person in one hundred tossing 8 heads and 2 tails is not inconsistent with an expected average 5:5 split, although ten out of the same hundred people would suggest a crooked coin (or tossers). Before throwing the coin's owner in jail, we might reasonably ask a statistician how often this many large deviations from the expected 5:5 split would occur on the basis of chance alone. The statistician might say that this is an uncommon event, which should occur less than 5% of the time ($p < .05$). That is, if we repeated the experiment 100 times, each time having the same one hundred people toss the coin 10 times, we might expect ten honest people to toss 8 or more heads in less than 5 of those 100 experiments. A law-and-order judge might consider that occurrence so unlikely as to prove the fellow's guilt. But an appeals court might overturn the verdict, arguing that 5 in 100 is too frequent to dismiss the possibility that it was due to chance alone, and only if its likelihood of occurring purely by chance were less than 1 in 100 ($p < .01$), or 1 in 1000 ($p < .001$) should the coin's owner be convicted of fraud.

Our usual, run-of-the mill statistical tests are therefore simply a means of determining the likelihood that our observations *could* have occurred by chance alone. Conversely, when the likelihood of their being a chance event is small, we consider the results (e.g., a difference between two groups) to be real, i.e., significant. For no very good reason (except historical accident), an event with less than a 5% probability of occurring by chance alone ($p < .05$) is routinely accepted as statistically significant. We feel even more confident, however, if the likelihood of a chance occurrence is even lower (e.g., $p < .01$).

Alpha error

Whenever we "accept" an observation as a real occurrence, we do so at risk of being wrong: regardless of how unlikely it is to have arisen from chance alone, it always *could* have. If p is less than .05, the risk is 1 in 20 that an observation which is not real (e.g., the difference between two treatment effects) will prove statistically significant. This risk of being wrong and accepting an observation as true when in fact it was merely due to chance, is known as the alpha error. Almost all statistical tests present their "level of significance" in terms of the alpha error.

TESTS OF STATISTICAL SIGNIFICANCE

In usual clinical practice we wish to compare observations in one group of individuals with those in another, and determine whether any apparent differences are likely to be real (i.e., were they unlikely to be due to chance). Choice of an appropriate statistical test depends, for the most part, on whether we are comparing attribute-type data (the proportion of individuals who went blind, developed nerve fiber bundle defects, had relapses, etc.) or measurement-type data (average number of days to corneal healing, mean intraocular pressure, etc.).

Attribute data

Here we are dealing with the proportion of individuals with a particular *attribute* (the rate at which a characteristic occurs in two groups of individuals).

THE NORMAL DEVIATE (z)

p_1 is the proportion of individuals in the first group who have that attribute, p_2 the proportion in the second.

$$p = \frac{\text{number of individuals with the attribute}}{\text{number of individuals examined in that group}}$$

n = the number of people in that group

$q = 1 - p$ = the proportion of individuals in the group without the attribute

$\bar{p}$ = average proportion positive in the 2 groups

$$= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$\bar{q} = 1 - \bar{p}$ = the average proportion negative in the 2 groups

The general formula is:

$$z = \frac{p_1 - p_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

The resultant number z is known as the normal deviate. The probability that the difference between the groups ($p_1 - p_2$) could have arisen by chance is shown in tables of normal deviates opposite the appropriate z value. It is always safest to use a "two-tailed test," since p_1 could be either larger or smaller than p_2. A two-tailed table of z is provided in Appendix 2.

A quick glance at the formula reveals two obvious facts:

1. The larger the difference between p_1 and p_2, the larger the z value
2. Similarly, the larger the size of the two groups (n_1 and n_2) the larger z

The larger z is, the less likely is the possibility that the observed difference is due to chance alone (i.e., the more statistically significant it is).

In practical terms, this means that large differences require fewer patients to prove statistical significance, whereas smaller differences require larger samples. Once again, a sufficiently large population can establish statistical significance for a very small difference, even one that is not the least bit "significant" from a clinical standpoint. Conversely, too small a sample can mask a moderate though clinically meaningful difference.

Illustrative example: The prevalence of superficial punctate keratopathy (SPK) among patients with conjunctival xerosis was

75% and among controls 7% (40). Is the difference statistically significant?

	Group 1 *Conjunctival xerosis*	*Group 2* *Controls*
n	63	58
number with SPK	47	4
p	$47/63 = .75$	$4/58 = .07$
q	$1.00 - .75 = 0.25$	$1.00 - .07 = 0.93$
$\overline{p}$	$\frac{47 + 4}{121} = 0.42$	
$\overline{q}$	$1.00 - 0.42 = 0.58$	

$$\text{Test: } z = \frac{p_1 - p_2}{\sqrt{\overline{p}\overline{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.75 - .07}{\sqrt{(0.42)(0.58)\left(\frac{1}{63} + \frac{1}{58}\right)}}$$

$$= 7.6$$

Looking up the z value of 7.6 in Appendix 2, we find it completely off the table, indicating $p < .001$. Had the z value been only 1.97, p would have been $< .05$. If z had been 2.8, p would have been $< .01$.

A small bit of complexity might be added at this point. This particular test of significance requires a "correction for continuity," especially if $n < 50$. What that means is unimportant. What you must do is narrow the difference observed between the two groups slightly, by subtracting 0.5 from the numerator used in arriving at p of the larger proportion, and adding 0.5 to the numerator used in arriving at p of the smaller proportion. Assuming $p_1 > p_2$, the corrected formula is:

$$z_c = \frac{\left(p_1 - \frac{0.5}{n_1}\right) - \left(p_2 + \frac{0.5}{n_2}\right)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

In general I find the normal deviate to be a more useful test than chi-square, because it deals directly with the rates or proportions in which we are interested. The chi-square, on the other hand, deals with absolute numbers, which have little inherent meaning. The statistical basis of the two tests, however, is identical.

The normal deviate (z) test is not appropriate under the following circumstances:

1. Where the size of the sample is small ($n < 20$ or larger—between 20 and 40—but the smallest number with the attribute is less than 5) employ *Fisher's exact test* (S-1 and S-2, Appendix 5).
2. To determine the likelihood that the proportion of individuals with the attribute in three or more groups was due to chance, rather than comparing one group with another, we use the chi-square test.

CHI SQUARE (χ^2)

As already noted, χ^2 is statistically equivalent to the z test. It also has the same "small sample" limitations. For our purposes its sole advantage is the ability to compare attribute data in three or more groups.

We begin by calculating the overall distribution of individuals with and without the attribute in the study as a whole, and then the degree to which the distribution within each group differs from the overall distribution. The chi-square test determines the probability that this (cumulative) amount of variation could have arisen by pure chance. If the probability is low ($p < .05$, $p < .01$, etc.), we conclude that one or more of the groups differs from the others in some meaningful way. To determine which groups (and how many) differ, we can resort again to simpler two-group (z) comparisons.

The general formula for chi-square is as follows:

$$\chi^2 = \frac{\sum(f - F)^2}{F}$$

where

f = the number of individuals *observed* to have (or not have) the attribute in that group

F = the number of individuals *expected* to have (or not have) the attribute in that group (if the overall proportion of people with and without the attribute in the study as a whole applied)

$\sum$ = sum of the calculations for all of the groups

In longhand, we say χ^2 is equal to the sum of the square of the difference between the observed and expected frequency divided by the expected frequency in every group.

Our first task is to construct a *contingency table*, simply a convenient means of displaying the data. As an example we will use additional data from the study on prevalence of punctate keratopathy among children with clinical xerophthalmia. Alternatively, one could be dealing with incidence of visual loss among patients on various antihypertensive medications, with various degrees of diabetic retinopathy, etc.

Four groups of children were studied: normals, and children with night blindness, conjunctival xerosis, or corneal xerosis. The observed number (f) of children with and without punctate keratopathy in each group is shown in Table 22. This is known as a 2 × 4 contingency table, since it distributes the entire study population among two classes (rows) and four categories (columns), resulting in eight "cells."

Next, we determine the expected number (F) of individuals in each cell. The expected number is derived by applying the overall distribution of positives and negatives to the total number

Table 22
CONTINGENCY TABLE ANALYSIS

I. Number of Eyes Observed (f)

SPK	*Corneal xerosis*	*Conjunctival xerosis*	*Night blindness*	*Normal controls*	*Total*
Present	47	47	10	4	108
Absent	0	16	18	54	88
Total	47	63	28	58	196

of children in each of the four groups. A simplified formula for determining F for any cell is:

$$F = \frac{\text{(row total)(column total)}}{\text{grand total}}$$

Note that if one carries out this calculation for a single cell in each of three different columns, the five remaining cells can be computed by simple subtraction. Results are shown in Table 23.

The difference between the observed and expected numbers $(f - F)$ in each cell appears in Table 24. To calculate χ^2 we simply plug in the appropriate numbers:

$$\chi^2 = \frac{(f - F)^2}{F} = \frac{(21.10)^2}{25.90} + \frac{(-21.10)^2}{21.10}$$

$$+ \frac{(12.29)^2}{34.71} + \frac{(-12.29)^2}{28.29}$$

$$+ \frac{(-5.43)^2}{15.43} + \frac{(5.43)^2}{12.57}$$

$$+ \frac{(-27.96)^2}{31.96} + \frac{(27.96)^2}{26.04}$$

$$= 106.26$$

We now have the value for χ^2. What in the world do we do

Table 23
CONTINGENCY TABLE ANALYSIS

II. Number of Eyes Expected (F)

SPK	*Corneal xerosis*	*Conjunctival xerosis*	*Night blindness*	*Normal controls*	*Total*
Present	25.90	34.71	15.43	31.96	108.00
Absent	21.10	28.29	12.57	26.04	88.00
Total	47.00	63.00	28.00	58.00	196.00

Table 24
CONTINGENCY TABLE ANALYSIS

III. Number of Eyes Observed Minus Number Expected ($f - F$)

SPK	*Corneal xerosis*	*Conjunctival xerosis*	*Night blindness*	*Normal controls*
Present	21.10	12.29	−5.43	−27.96
Absent	−21.10	−12.29	5.43	27.96

with it? Simply turn to Appendix 3, a table of probability values for χ^2. This is used in much the same way as the two-tailed table of z except for one added annoyance: you must choose from a column labeled "degrees of freedom" (d.f.). Whereas the concept involved is not germane to our discussion, choosing the correct degrees of freedom is. This is simply calculated from our contingency table:

$$\text{d.f.} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

In this particular example, we were dealing with a 2 × 4 contingency table, hence

$$\text{d.f.} = (2 - 1)(4 - 1) = 3$$

and the p value for our study, with χ^2 of 106.62, and 3 d.f. is <

Table 25
CHI-SQUARE FOR 2 × 2 TABLE

Simplified Version Already Employing Yates Correction

I. Format of table

	Column I	*Column II*	*Total*
Row I	a	b	r_1
Row II	c	d	r_2
Total	k_1	k_2	N

II. Simplified formula for χ_c^2

$$\chi_c^2 = \frac{N(|ad - bc| - ½N)^2}{(k_1)(k_2)(r_1)(r_2)}$$

where $(|ad - bc| - ½N)$ indicates reducing the absolute value of $ad - bc$ by one-half the sum of the observations (N).

.001. We really do not have to prove this when the differences are as large and obvious as in this case.

The good news is that the χ^2 test can be used even if some values of F (expected number) are as low as 1, as long as most are *substantially* larger (at least 5); and that when χ^2 has more than 1 d.f., the "correction for continuity" (necessary in the z, or its equivalent, the 2 × 2 χ^2 test) is unnecessary.

If you insist on using the χ^2 test instead of the normal deviate (z), it is always safest to employ the Yates correction factor. This is already incorporated in the simplified determination of χ^2 applicable to 2 × 2 contingency tables shown in Table 25.

Measurement data

Statistical tests of the difference between measurement data (dealing with means instead of proportions) in two groups are a bit trickier; they require familiarity with two additional concepts: *standard deviation* (of the observations) and *standard error* (of the mean). Neither is particularly complex.

STANDARD DEVIATION

If we were to measure the intraocular pressure of 100 eyes randomly selected from the general population, we would find that they varied considerably. The lowest measurement might be 8, the highest 32, and the remainder somewhere in between. The *mean,* calculated by adding all the measurements together and dividing by the total number of eyes examined (100 in this instance) would be about 16. The *standard deviation* (S.D.) is simply a technique for indicating the degree of variation of the individual measurements about this mean. The mean ±1 S.D. would include two-thirds of all the individual measurements. A common standard deviation of IOP among normal eyes is 2.4. Therefore, 66% of the population measured would have an IOP within the mean ±1 S.D. (i.e., 16 ± 2.4; 13.5–18.4 mm Hg). The mean ±2 S.D. encompasses 95% of all measurements. So 95% of the individuals in the group have an IOP between 16 ± 2 S.D. (16 ± 4.8; 11.2–20.8 mm Hg). In fact, this is exactly the way 21 was chosen as the standard upper limit of "normal" IOP. Not because there is anything magic about 21, but because over 95% of the general population has a pressure equal to or below this amount.

STANDARD ERROR

The *standard error of the mean* (S.E.) relates not to the variation of individual measurements about the mean for that particular study or group, but the variation of *mean* measurements, each from a different study, about the mean for all the studies. For example, the mean IOP for the first examination of 100 eyes might be 16.5. The second series of measurements on the same eyes might have a mean of 17.4, and the third 15.9. The standard error of the mean (S.E.) would describe the variation of these individual study means (16.5, 17.4, and 15.9) about the mean for all three studies together (16.6). It should be intuitively obvious that the variation of sample or study means about the "true" mean (i.e., the mean of all the samples or studies together) will be smaller than the variation of individual measurements about a single study's mean. Since each study's mean has already averaged out the extreme highs and lows of the individual measurements within that study, study means are less likely to vary by a large amount from one another, or the overall mean. As with the S.D., the overall sample mean ±1 S.E. will include two-thirds of all individual sample means, and the overall sample mean ±2 S.E. will include 95%.

CALCULATING THE STANDARD DEVIATION

For years, students of elementary statistics have memorized the formula for the standard deviation. This is no longer necessary, and it will not be given here. Almost any inexpensive electronic calculator can do it automatically, and with greater precision and ease than was previously possible.

CALCULATING THE STANDARD ERROR

Unlike the standard deviation, which is calculated directly from the actual observations of which it is composed, the standard error is rarely determined from multiple replications of the same experiment. Since the S.E. is always smaller than the S.D., and to a reasonably predictable degree, it is calculated directly from the S.D. as follows:

$$\text{S.E.} = \frac{\text{S.D.}}{\sqrt{n}}$$

where n = the total number of observations from which the S.D. was calculated.

COMPARISON OF TWO INDEPENDENT SAMPLES

With a simple calculator in hand, we can now determine whether the difference in measurement data (e.g., IOP) between two samples is statistically significant.

1. For each of the two samples, calculate its mean ($\bar{x}$), standard

deviation (S_x), and number of participants (n). $\bar{x}_1$ is the mean for sample 1, S_1 the standard deviation for sample 1, etc.

2. For each of the two samples, convert the standard deviation (S_x) to $\sum x^2$ (the "sum of the squares of the differences of each individual measurement from the sample mean") as follows:

$$\sum x^2 = (S_x)^2(n - 1)$$

That is, simply square the standard deviation and multiply it by 1 less than the number of observations in that sample.

3. Determine s^2 pooled (the "pooled variance for the comparison"). Forget what it means, just do the following:

$$s^2 \text{ pooled} = \frac{\sum x_1^2 + \sum x_2^2}{(n_1 - 1) + (n_2 - 1)}$$
$$= \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}$$

4. Determine $S_{1\text{-}2}$ (the standard error of the difference of the means of the two samples) as follows:

$$S_{1\text{-}2} = \sqrt{s^2 \text{ pooled} \left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$
$$= \sqrt{\frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} \left(\frac{n_1 + n_2}{n_1 n_2}\right)}$$

5. Compute the t statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{1\text{-}2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} \left(\frac{n_1 + n_2}{n_1 n_2}\right)}}$$

That's all there is to it, except for one little hitch: what in the world is t? In fact, t is no more than our old friend the normal deviate (z) in a new guise. Why use t? Because by consulting Student's t table (Appendix 4) the answer is already corrected for small samples (especially important where $n < 30$). As with χ^2, you must choose the correct degrees of freedom. In this instance,

$$\text{d.f.} = (n_1 - 1) + (n_2 - 1)$$

The calculations above are for the general situation, where n_1 need not equal n_2, and where the samples are *unpaired*, that is, independent. Patients randomly allocated to receive or not receive a new oral agent for reducing intraocular pressure form two independent samples, and the test above is the one that is appropriate. Similarly, a comparison of serum vitamin A levels between children with and without xerophthalmia might be handled in the same way.

Illustrative example: The mean serum vitamin A level among children with night blindness and their matched controls was 13.9 and 17.6 μg/100 ml, respectively (35). Is the difference in vitamin A levels statistically significant? Using my pocket calculator, I determined n, $\overline{x}$, and S_x for each of the two groups:

	Group 1 *Night blindness*	*Group 2* *Controls*
n	174	161
$\overline{x}$	13.9	17.6
S_x	5.9	7.8

Now for a little algebra (using my calculator of course):

$\sum x^2$	$(5.9)^2(173)$ $= 6022$ (for group 1)	$(7.8)^2(160)$ $= 9734$ (for group 2)

s^2 pooled $$\frac{6022 + 9734}{173 + 160} = 47.3$$

$S_{1\text{-}2}$ $$\sqrt{47.3\left(\frac{174 + 161}{(174)(161)}\right)} = 0.75$$

t $$\frac{17.6 - 13.9}{0.75} = 4.93$$

d.f. $$(174 - 1) + (161 - 1) = 333$$

Consulting Student's t table under 333 degrees of freedom under the infinity classification at the bottom of the page, we find that a t value of 4.93 is larger than any value listed, and $p < .001$. If, for example, there had been 25 degrees of freedom, and t equaled 2.18, then p would have been $<.05$.

COMPARISON OF TWO PAIRED SAMPLES

Special circumstances exist, especially in ophthalmology, where "pairing" between samples is possible: e.g., treating one eye of an individual with a new topical antihypertensive agent, while the other eye receives an old (or no) medication. Since the two eyes of the same individual are more likely to be alike than two eyes from unrelated individuals (the situation in nonpaired, independent samples), natural variations in biological responsiveness are also likely to be less, and the standard deviations smaller. With chance variation (S.D.) reduced, it is easier to detect differences due to the drug itself. The statistical significance of paired data is also easier to calculate:

1. Find the difference within each pair (D) with the sign intact (i.e., always subtract Eye 1 from Eye 2).
2. Punch these differences into your handy calculator and read off $\overline{\mathrm{D}}$ (the mean difference) and $\mathrm{S_D}$ (the standard deviation of the individual differences).
3. Compute $\mathrm{S.E._D}$ (the standard error of the mean difference) as follows:

$$\mathrm{S.E.}_{\overline{\mathrm{D}}} = \frac{\mathrm{S_D}}{\sqrt{n}} \quad \text{(where } n = \text{number of } \textit{pairs}\text{)}$$

4. Calculate t

$$t = \frac{\overline{\mathrm{D}}}{\mathrm{S.E.}_{\overline{\mathrm{D}}}}$$

i.e.,
$$t = \frac{\text{Mean difference}}{\text{S.E. of mean differences}}$$

$$\text{d.f.} = n - 1 \quad \text{(where } n = \text{number of } \textit{paired} \text{ observations)}$$

Illustrative example: In our previous example we compared mean serum vitamin A levels among children who were night blind and controls as if they were independent samples. In fact, the abnormals were actually matched to an age/sex/neighborhood-specific control. This matching permits a paired analysis of the data.

Again, n (the number of pairs), $\bar{D}$ (the mean difference), and $S_{\bar{D}}$ (the standard deviation of the individual differences) are read directly from the calculator. In this case n is smaller than in the grouped comparisons because blood samples were not always available for both members of each pair.

n	$\bar{D}$	$S_{\bar{D}}$
(number of pairs)	(mean difference between members of each pair)	(S.D. of the difference between pair members)
160	3.77	9.0

$$\text{S.E.}_{\text{D}} = \frac{9.0}{\sqrt{160}} = 0.71$$

$$t = \frac{3.77}{0.71} = 5.3$$

$$\text{d.f.} = 159$$

Consulting the t table for $t = 5.3$ and d.f. = 159, we find $p < .001$.

If you think a paired analysis is in order, do it. Nothing is lost if you are wrong; results will be equivalent to the standard grouped (independent sample) analysis. If you are right, however, the standard deviation will be smaller, and you might detect a statistically significant difference missed by the grouped comparison.

COMPARISON OF THREE OR MORE SAMPLES

Student's t test is a simple, convenient method for comparing the means of two different samples. It is equally useful and applicable when one is dealing with more than two samples: e.g., in searching for statistically significant differences in the reduction of mean IOP with different carbonic anhydrase inhibitors (26). One simply repeats the *t* test for every combination of agents (methazolamide versus placebo, acetazolamide versus placebo, methazolamide versus acetazolamide, etc.) for which there appears, on casual examination, to be a potentially significant difference.

Occasionally, however, we wish to compare three or more sample means simultaneously (analogous to the use of the chi-square test for attribute data, with contingency tables larger than 2 × 2). A powerful method is *analysis of variance* (ANOVA). Because it is rarely required, and the computations and principles involved somewhat more complex than what we've dealt with to date, it will not be presented here. It is discussed in detail in references S-1, S-4, and S-5 of Appendix 5.

Choosing a sample size

From a certain viewpoint, we should have begun the statistical section with this discussion, since it is the first statistical tool the investigator employs and the reader reviews. But the other sections will already have introduced the concepts involved, and we can now proceed more rapidly.

As we've already seen, the likelihood of detecting a statistically significant difference between two groups depends upon the size of the groups and of the difference. As an obvious example, we can't possibly expect to demonstrate a reduction in postoperative endophthalmitis from 4.5 per 1000 to 1.5 per 1000 (a difference of only 3 cases out of every 1000 operations) by treating 100 people with a new antibiotic and another 100 with a placebo. On the average, one could not expect even a single case of endophthalmitis in the control group, let alone in the treatment group. Unfortunately, numerous studies have been undertaken in just this manner: the more perceptive investigators discover that they have wasted a lot of time and effort; less perceptive investigators publishing the results anyway, claiming (quite right-fully) that "there was no statistically significant difference" between the groups. Unfortunately, they wrongly interpret this to mean no 'significant clinical difference," which is something else entirely and would require many thousands (rather than hundreds) of operations to prove.

If, on the other hand, one were interested in the proportion of ocular hypertensives who respond to pilocarpine, far fewer patients would be needed. In short, there is a direct relationship between the level of difference one wishes to establish and the number of subjects required in each of the groups (one receiving the drug, the other placebo, etc.). The smaller the expected difference, the larger the number of subjects required.

The first, and most important, step is determining the smallest possible difference you would like to be able to demonstrate. From a practical standpoint, begin with the lowest level deemed clinically significant. This level will vary with the disease. As indicated earlier, a drug that speeds healing of herpetic ulcers from 7 to 6 days is probably no great advance; but one that heals the ulcer in 2 or 3 days might well be. Simply plug this difference into the appropriate formula below and read off the number of patients required. One is often astonished at the size of the required sample. If it is beyond practical means, one can raise the level of difference, recognizing that smaller differences are likely to be missed. For example, instead of hoping to show a fall in the incidence of postoperative endophthalmitis of as little as 10%, from 4.5 per 1000 to 4 per 1000, one might have to settle for a much larger "minimal" drop, say of 65% (from 4.5 to 1.5 per 1000). You may discover, to your dismay, that even this sample size is impractical: after all, how many cataract surgeons can accumulate 10–20,000 operations in a reasonable period of time. It's far better to learn that the project is futile *before* beginning it—rather than after years of fruitless labor.

Before using the formulas, there are two additional (and less flexible) considerations that require discussion: the levels of alpha and beta error. As already mentioned, the *alpha error* is the likelihood of accepting an apparent difference between two treatments as real, when in fact the regimens are equally effective. Our familiar, generally accepted level of $p < .05$ (two-tailed) is usually used. The *beta error* is the flip side of the coin: the risk we run of being wrong in concluding that there was no real difference. Just as chance alone *could* produce a "statistically significant" difference not really due to the treatment's effect (alpha error), chance alone could mask a *real* difference of the magnitude that we are seeking. We could erroneously conclude that no significant difference exists in the two therapeutic regimens, when in fact a sizeable difference does exist. For a variety of reasons—most notably a belief it is somehow less harmful to fail to prove that a new regimen is better, when in fact it is, than to claim that a new regimen is superior, when in fact it isn't—we usually use a less stringent standard for our beta than for our alpha error. It is common practice to accept a one-tailed beta error of 0.2 (i.e., there is a 20% chance we will miss a real difference of the size we are after), though in some circumstances a risk of this size might be unacceptable. The smaller the alpha or beta error chosen, the larger n will become.

That's all there is to it. Determine the difference you wish to detect, assume the alpha and beta errors you can live with (usually a two-tailed alpha error of .05 and a one-tailed beta error of 0.2), and calculate n. If n is too large, you can raise the level of difference, or increase the size of the alpha and beta error. Of course, this means that you are more likely to fail to detect a smaller difference, call a difference real when it is not, or the converse, reject a difference as not significant when in fact it is.

Attribute data

$$n = \frac{(z_\alpha + z_\beta)^2(p_1q_1 + p_2q_2)}{(p_2 - p_1)^2}$$

n = number of subjects needed in *each* of the two groups
p_1 = estimated proportion with the attribute (e.g., will experience a fall in intraocular pressure, develop postoperative endophthalmitis, etc.) in group 1
p_2 = same calculation for group 2

$q_1 = 1 - p_1$

$q_2 = 1 - p_2$

$p_2 - p_1$ = *minimal* level of difference you wish to detect between the two groups, treatment and control, for example (you will automatically detect any differences that are larger).

z_α = normal deviate of your alpha error (for .05, two-tailed $z_\alpha = 1.96$).

z_β = normal deviate of your beta error (for 0.2, one-tailed, $z_\beta = 0.84$).

As n becomes smaller, the formula above becomes less accurate. Under most circumstances, however, it provides a reasonable ballpark figure within which to work.

Another convenient method for calculating attribute n makes use of relative risks and the Poisson distribution (41).

Illustrative example: Let us return to our earlier example, the prevalence of superficial punctate keratopathy (SPK) among children with conjunctival xerosis, and see what sample size would have been required for the degree of difference actually found.

p_1 = estimated proportion of normal (control) children with SPK = 0.07.

$q_1 = 1.00 - 0.07 = 0.93$

p_2 = estimated proportion of abnormal children with SPK = 0.75

$q_2 = 1.00 - 0.75 = 0.25$

z_α = 1.96 (.05, two-tailed)

z_β = 0.84 (.20, one-tailed)

$$n = \frac{(1.96 + 0.84)^2[(.07)(.93) + (.75)(.25)]}{(.75 - .07)^2}$$

$n = 4.3$

Even with a few extras for safety, 10 children in each group would have been adequate to demonstrate that punctate keratopathy was statistically significantly more common ($p < .05$) among children with conjunctival xerosis than among controls.

All we accomplished by examining many more children was to demonstrate it at the .000 . . . 1 level and, of course, waste a lot of time. In all fairness, we had no idea what the difference in prevalence would be. In fact, the observation had never been made before.

Now for a more common, hypothetical example. Let us assume that 10% of sighted diabetics with neovascularization ordinarily go blind within a year of examination. We are interested in learning whether panretinal photocoagulation can reduce the rate to at *least* half, or 5%.

$$p_1 = 0.10 \qquad p_2 = 0.05$$

$$q_1 = 0.90 \qquad q_2 = 0.95$$

$$z_\alpha = 1.96 \qquad z_\beta = 0.84$$

$$n = \frac{(1.96 + 0.84)^2[(.10)(.90) + (.05)(.95)]}{(.10 - .05)^2}$$

$$n = 431$$

Given the usual loss to follow-up, etc., we will need 500 to 550 sighted patients (or eyes) with diabetic neovascularization in each of the two groups (treatment and control).

Measurement data

As usual, computations of measurement-type data are more complex.

PAIRED SAMPLES

The basic formula for paired samples is as follows:

1.

$$n = \frac{(z_\alpha + z_\beta)^2(2\sigma)}{D^2}$$

where

n = size of the sample required in *each* group

z_α = normal deviate of your alpha error (for .05, two-tailed, $z_\alpha = 1.96$)

z_β = normal deviate of your beta error (for 0.2, one-tailed, $z_\beta = 0.84$)

D^2 = square of the expected difference between the two groups (i.e., if you expect a drug to lower mean IOP from 25 to 20, $D^2 = (25 - 20)^2 = 25$)

σ = the standard deviation actually observed in a prior experiment (for example the S.D. of IOP in a group similar to that to be studied). The term 2σ is actually an estimate of what we expect the (squared) standard error of the difference between our study groups to be.

2. Solve main formula for n.
3. We are not yet finished. For a variety of reasons not germane to our discussion, n must now be adjusted as follows:
   a. Find n_1 ($n_1 = n$ of sample 1) (i.e., step 2 above)
   b. Calculate the total degrees of freedom (d.f.)

   $$\text{d.f.} = (n_1 - 1) + (n_2 - 1) = 2(n - 1)$$

   c. Adjust n as follows:

   $$n_c = n_1 \frac{\text{d.f.} + 3}{\text{d.f.} + 1}$$

   (For those who are interested, this adjustment makes use of the t distribution: as n increases, the adjustment has less effect.)

INDEPENDENT SAMPLES

Do exactly the same thing for independent samples, but substitute the expression $2\sigma^2$ for 2σ in expression 1:

$$n = \frac{(z_\alpha + z_\beta)^2(2\sigma^2)}{D^2}$$

Illustrative example: This time, let us return to our comparison of serum vitamin A levels among children with night blindness and matched controls. For convenience, we will use the actual standard deviation (σ) observed in that study, although many other published examples could be used.

D = the minimum difference in mean vitamin A levels we wish to detect = 3.7 μg/dl

σ = standard deviation observed in either of those populations = 7.8

$z_\alpha = 1.96 \qquad z_\beta = 0.84$

$$n = \frac{(1.96 + 0.84)^2(2)(7.8)^2}{(3.7)^2} = 70$$

$$\text{d.f.} = 2(70 - 1) = 138$$

$$n_c = 70\left(\frac{138 + 3}{138 + 1}\right) = 71$$

It is obvious that as n increases, the correction n_1 (d.f. + 3/ d.f. + 1) assumes less importance.

CONFIDENCE LIMITS

Till now, we've concerned ourselves with the probability that an *observed* difference between two groups is real. We've said nothing about how large the *true* difference is likely to be. Dahlen et al. (42) found that on the average the intraocular pressure among patients treated with acetazolamide was 7 mm lower than among controls, and that this difference was statistically significant ($p < .05$). Statistically, it is unlikely that this large a difference could have resulted from chance alone. We therefore conclude that acetazolamide is an effective agent for reducing intraocular pressure. We have little information, however, about the magnitude of its effect. In this particular study, the mean reduction was 7 mm Hg. Will the average of all subsequent studies also be 7 mm? That would be unlikely. You will recall that the standard error of the mean describes the variability of the mean values of repeated studies around the average mean of all the studies, and that this average mean ±(2 times) its standard error includes 95% of all individual study means. None of which gets us very far, since we rarely repeat the same experiment many times. To be useful, this concept must be turned on its head: we calculate the standard error from the standard deviation observed in this single experiment. The observed mean ±2 S.E. then has a 95% chance of including within it the "true" mean for this sample. This is the 95% confidence level of our mean. Our sample mean ±3 S.E. has a 99% chance of encompassing the "true" mean, whatever that elusive figure is, and is therefore the 99% confidence interval for the mean. As we widen the interval, we decrease the likelihood of missing the "true" mean, but of course we also reduce the precision of our estimate.

Confidence intervals also have another use. If our sample was representative of all similar individuals in the population at large (e.g., we had selected our ocular hypertensives randomly from among all ocular hypertensives), the confidence interval would include the "true" mean for all ocular hypertensives in the population.

Illustrative example: As an example, let us turn again to our study of serum vitamin A levels. The mean serum vitamin A level of randomly sampled children was 20 μg/100 ml. The S.E. of this mean was:

$$\text{S.E.}_{\cdot x} = \frac{\text{S.D.}}{\sqrt{n}} = \frac{7.86}{\sqrt{268}} = 0.48$$

The 95% confidence limits are 20.0 ± 2(.48), or 19.0 and 21.0. The interval between 19.0 and 21.0 μg/100 ml had a 95% chance of including the true mean serum vitamin A level of 6000 children: the population from which the 268 who contributed blood samples had been randomly selected. The 99% confidence interval would be 18.6 to 21.4 μg/100 ml, which increases our assurance but lowers our precision.

If, instead of the mean level, we are interested in the mean *difference* between two groups, we simply substitute the standard error of the mean difference for the standard error of the mean level. In our paired comparison between night-blind and normal children, the mean difference was 3.77 μg/100 ml and the S.E. of this difference (S.D./$\sqrt{n}$ = 9.0/$\sqrt{160}$) was 0.71. The 95% confidence limits are therefore 2.35 and 5.19. Similarly, in our independent (unpaired) comparison of roughly these same children, the difference between the means ($\overline{x_1} - \overline{x_2}$) was 3.7, and the standard error of the difference of the means of the two samples ($S_{1\text{-}2}$) was 0.75. The 95% confidence interval for the difference between the means of the two samples is therefore 2.2 to 5.2. Had these been random samples of all night-blind children (and their matched controls) in Indonesia, then the "true" difference between all 1 million night-blind children and matched controls in Indonesia would probably (95% chance) lie between these limits.

To make use of confidence limits, we need only calculate the standard error.

For *measurement data*, as in the above examples:

$$\text{Standard error} = \frac{\text{standard deviation}}{\sqrt{n}}$$

the standard deviation being read off a handy pocket calculator. Calculation of the S.E. of a *difference* between two independent samples is slightly more complex, and was already covered under tests of significance (p. 59).

For *attribute data* (rates and proportions), the standard deviation and standard error are identical and calculated as follows:

$$\text{S.E. (or S.D.)} = \sqrt{pq/n}$$

where

p = proportion with the attribute
q = 1 − p (the proportion without the attribute)
n = size of the sample

Illustrative example: To turn again to the prevalence of punctate keratopathy among children with Bitot's spots, the observed rate was 0.75. Had this been a representative sample of all children in Indonesia with Bitot's spots (it was not), the true, overall population prevalence of SPK among them would have a 95% chance of lying within 0.75 ± 2 S.E.

$$\text{S.E.} = \sqrt{\frac{(.75)(.25)}{63}} = .055$$

$$\text{Confidence limits} = 0.75 - 2(.055) = 0.64$$

$$0.75 + 2(.055) = 0.86$$

SOME PARTING ADVICE

Congratulations! Having completed (and presumably understood) this section, you now have at your disposal all the statistical tools required to carry out and evaluate the most common types of clinical studies. You can estimate sample size, determine whether the difference between two groups (of diseases, treatments, etc.) is statistically significant, and estimate how large that difference really is. These procedures are outlined in Table 26. Nonetheless, it is a good idea to consult with

Table 26
APPLICATION OF BASIC STATISTICAL PROCEDURES

A. *Tests of statistical significance*	
1. Rates and proportions (attribute data)	
a. Two samples	Normal deviate (z) Fisher's exact test (for small samples) Chi-square (χ^2)
b. Three or more samples	Chi-square (χ^2)
2. Means and averages (measurement data)	
a. Paired samples	Student's t (paired analysis: $\overline{D}$)
b. Two independent samples	Student's t (unpaired analysis: $\bar{x}_1 - \bar{x}_2$)
c. Three or more samples	Student's t (nonsimultaneous comparisons) Analysis of variance (ANOVA) (simultaneous comparisons)

Table 26 (cont.)

B. *Estimation of true magnitude*	Confidence limits
1. Rates and proportions (attribute data)	$p \pm (a)$ S.E. S.E. $= \sqrt{pq/n}$
2. Means or differences (measurement data)	$\bar{x} \pm (a)$ S.E. $\bar{D} \pm (a)$ S.E. S.E. $=$ S.D.$/\sqrt{n}$ $a = 2$, C.L. $= 95\%$ $a = 3$, C.L. $= 99\%$
C. *Determination of sample size*	See text
1. Rates and proportions (attribute data)	
2. Means and averages (measurement data)	
a. Paired samples	
b. Independent samples	

an epidemiologist/statistician whenever possible, especially when undertaking large studies utilizing many subgroups, risk factors, end points, or sequential analyses.

Individuals involved in numerous studies will want to use more sophisticated calculators, capable of carrying-out many of these statistical manipulations automatically. A word of caution! The formulations that they employ are not always adequate. For example, the normal deviate (z) is rarely corrected for continuity, and chi-square (for 2 × 2 tables) rarely employs the Yates correction. Before relying on any prepackaged program, be sure that its formulation is adequate for your purposes. If not, you can always construct your own program for use in programmable calculators.

The statistical theory behind these various procedures, their mathematical derivations, and additional, more complex manipulations are described in detail in references S-1 through S-6, Appendix 5.

References

1. Christy, N. E., Sommer, A. Antibiotic prophylaxis of postoperative endophthalmitis. Ann. Ophthalmol. 11: 1261–1265, 1979.
2. Hiller, R., Kahn, A. H. Blindness from glaucoma. Am. J. Ophthalmol. 80:62–69, 1975.
3. Hollows, F. C., Graham, P. A. Intraocular pressure, glaucoma, and glaucoma suspects in a defined population. Brit. J. Ophthalmol. 50:570–586, 1966.
4. Sommer, A., Sugana, T., Hussaini, G., Emran, N., Tarwotjo, I. Xerophthalmia-determinants and control. Proceed. XXIII. Int. Cong. Ophthalmol., Shimizu, K., Oostehuis, J. A., eds. Amsterdam, Excerpta Medica, 1979, pp. 1615–1618.
5. Kirk, H. O., Petty, R. W. Malignant melanoma of the choroid: a correlation of clinical and histological findings. Arch. Ophthalmol. 56:843–860, 1956.
6. Perkins, E. S. The Bedford glaucoma survey. I. Long-term follow-up of borderline cases. Brit. J. Ophthalmol. 57:179–185, 1973.
7. David, R., Livingston, D. G., Luntz, M. H. Ocular hypertension—a long-term follow-up of treated and untreated patients. Brit. J. Ophthalmol. 61: 668–674, 1977.

8. Indonesian Nutritional Blindness Prevention Project. Annual Report. Bandung and New York: Helen Keller International, 1978.
9. Kirsch, R. E., Anderson, D. R. Clinical recognition of glaucomatous cupping. Am. J. Ophthalmol. 75:442–454, 1973.
10. Sommer, A., Pollack, I., Maumenee, A. E. Optic disc parameters and onset of glaucomatous field loss. 2. Static screening criteria. Arch. Ophthalmol. 97: 1449–1454, 1979.
11. Stromberg, U. Ocular hypertension. Acta. Ophthalmol. Suppl. 69, p. 62. Copenhagen, Munksgaard, 1962.
12. Pohjanpelto, P. E. J., Palva, J. Ocular hypertension and glaucomatous optic nerve damage. Acta. Ophthalmol. 52:194–200, 1974.
13. Bankes, J. L. K., Perkins, E. S., Tsolakis, S., Wright, J. E. Bedford glaucoma survey. Brit. Med. J. 1:791–796, 1968.
14. Read, J., Goldberg, M. F. Comparison of medical treatment for traumatic hyphema. Tr. Am. Acad. Ophthal. Otol. 78: 799–815, 1974.
15. Peyman, G. A., Sathar, M. L., May, D. R. Intraocular gentamycin as intraoperative prophylaxis in South India eye camps. Brit. J. Ophthalmol. 61: 260–262, 1977.
16. The Diabetic Retinopathy Study Research Group. Photocoagulation treatment of proliferative diabetic retinopathy: the second report of diabetic retinopathy study findings. Ophthalmol. 85:82–106, 1978.
17. Kahn, H. A., Liebowitz, H. M., Ganley, J. P., Kini, M. M., Colton, T., Nicherson, R. S., Dawber, T. R. The Framingham eye study. II. Association of ophthalmic pathology with single variables previously measured in the Framingham heart study. Am. J. Epidem. 106:33–41, 1977.
18. Van Metre, T. E., Maumenee, A. E. Specific ocular lesions in patients with evidence of histoplasmosis. Arch. Ophthalmol. 71:314–324, 1965.
19. Hiller, R., Kahn, H. A. Senile cataract extraction and diabetes. Brit. J. Ophthalmol. 60:283–286, 1976.

20. Gasset, A. R., Houde, W. L., Garcia-Bengochea, M. Hard contact lens wear as an environmental risk in keratoconus. Am. J. Ophthalmol. 85:339-341, 1978.

21. Zimmerman, L. E., McLean, I. W., Foster, W. D. Does enucleation of the eye containing a malignant melanoma prevent or accelerate the dissemination of tumor cells? Brit. J. Ophthalmol. 62: 420-425, 1978.

22. Jaffe, N. S., Eichenbaum, D. M., Clayman, H. M., Light, D. S. A comparison of 500 Binkhorst implants with 500 routine intracapsular cataract extractions. Am. J. Ophthalmol. 85: 24-27, 1978.

23. Klein, M. L., Fine, S. L., Patz, A. Results of argon laser photocoagulation in presumed ocular histoplasmosis. Am. J. Ophthalmol. 86:211-217, 1978.

24. Sommer, A., Emran, N. Topical retinoic acid in the treatment of corneal xerophthalmia. Am. J. Ophthalmol. 86:615-617, 1978.

25. Moss, A. P., Ritch, R., Hargett, N. A., Kohn, A. N., Smith, H., Podos, S. M. A comparison of the effects of timolol and epinephrine on intraocular pressure. Am. J. Ophthalmol. 86:489-495, 1978.

26. Lichter, P. R., Newman, L. P., Wheeler, N. C., Beall, O. V. Patient tolerance to carbonic anhydrase inhibitors. Am. J. Ophthalmol. 85:495-502, 1978.

27. Yasuna, E. Management of traumatic hyphema. Arch. Ophthalmol. 91:190-191, 1974.

28. Sommer, A., Pollack, I., Maumenee, A. E. Optic disc parameters and onset of glaucomatous field loss. 1. Methods and progressive changes in disc morphology. Arch. Ophthalmol. 97: 1444-1448, 1979.

29. Kahn, H. A., Leibowitz, H., Ganley, J. P., Kini, M., Colton, T., Nickerson, R., Dawber, T. R. Standardizing diagnostic procedures. Am. J. Ophthalmol. 79:768-775, 1975.

30. Edwards, W. C., Layden, W. E. Monocular versus binocular patching in traumatic hyphema. Am. J. Ophthalmol. 76: 359-362, 1973.

31. Casscells, W., Taylor, S. Damned Lies. Lancet ii:687–688, 1978.
32. Weber, J. C., Schlagel, T. F., Golden, B. Statistical correlation of uveitis syndromes with virus titers. Am. J. Ophthalmol. 78:948–951, 1974.
33. Blegvad, O. Xerophthalmia, keratomalacia and xerosis conjunctivae. Am. J. Ophthalmol. 7:89–117, 1924.
34. Sommer, A. Cataracts as an epidemiologic problem. Am. J. Ophthalmol. 83:334–339, 1977.
35. Sommer, A., Hussaini, G., Muhilal, Tarwotjo, I., Susanto, D., Sarosa, S. History of nightblindness: a simple tool for xerophthalmia screening. Am. J. Clin. Nutr. (in press).
36. The Diabetic Retinopathy Study Research Group. Preliminary report on effects of photocoagulation therapy. Am. J. Ophthalmol. 81:383–396, 1976.
37. Sommer, A. Toward a better understanding of medical reports. Am. J. Ophthalmol. 79:1053–1056, 1975.
38. Sommer, A. Keratoconus in contact lens wear. Am. J. Ophthalmol. 86:442–443, 1978.
39. Ederer, F., Ferris, F. Studying the role of an environmental factor in disease etiology. Am. J. Ophthalmol. 87:434–435, 1979.
40. Sommer, A., Emran, N., Tamba, T. Vitamin A responsive punctate keratopathy in xerophthalmia. Am. J. Ophthalmol. 87:330–333, 1979.
41. Diamond, E. L., Temple, B. Tables of the power of the conditional test for equality of two Poisson parameters. Dept. of Epidem., Johns Hopkins University School of Public Health.
42. Dahlen, L., Epstein, D. L., Grant, W. M., Hutchinson, B. T., Prien, E. L., Krall, J. M. A repeated dose-response study of methazolamide in glaucoma. Arch. Ophthalmol. 96:2214–2218, 1978.

APPENDIX 1

TABLE OF RANDOM NUMBERS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	6	2	8	3	5	7	6	4	9	0	7	6	6	8	0
2	3	4	2	5	2	0	3	0	5	1	5	1	3	5	7	1
3	3	4	7	4	1	5	8	8	9	9	4	0	3	4	3	6
4	4	7	5	0	4	8	3	3	0	5	7	4	8	4	5	9
5	9	3	5	6	8	1	1	7	2	0	7	8	3	5	8	6
6	8	6	1	5	7	5	3	7	6	6	4	9	5	0	7	1
7	2	2	2	3	2	7	1	2	4	4	3	6	2	6	5	0
8	2	3	3	4	7	5	8	2	0	2	8	7	4	4	1	8
9	2	0	4	2	6	0	5	7	9	4	8	5	4	6	0	3
10	6	5	3	3	1	1	0	3	6	9	0	2	7	3	1	7
11	3	9	2	9	8	9	5	4	4	6	4	6	8	6	3	3
12	7	2	2	1	8	4	5	9	5	6	5	9	2	5	3	2
13	7	4	0	7	3	7	4	2	6	8	6	5	3	1	8	9
14	9	7	2	2	8	0	3	9	9	8	1	5	7	4	7	9
15	1	9	9	8	9	3	9	4	4	2	2	1	4	6	5	7
16	7	2	9	4	6	1	6	7	9	8	7	5	3	7	4	6
17	9	1	5	2	3	0	2	6	5	8	1	2	2	3	7	9
18	6	9	3	4	5	2	8	0	6	2	4	7	9	2	9	6
19	6	2	1	6	5	6	2	9	5	3	2	7	4	1	0	8
20	0	7	4	1	1	6	0	6	2	1	8	2	7	8	3	7
21	3	6	7	6	7	2	6	0	2	7	7	2	5	6	8	3
22	4	9	3	0	8	5	6	9	5	9	4	9	7	5	4	3
23	1	4	1	2	0	3	3	6	7	0	1	4	4	1	5	1
24	7	5	5	6	9	4	1	6	0	8	9	2	6	0	7	0
25	7	4	0	6	5	5	8	4	6	7	3	6	5	2	6	5
26	2	1	4	1	0	4	6	1	2	0	8	5	2	2	7	1
27	7	0	0	2	6	9	1	0	3	7	4	5	9	5	9	4
28	4	6	4	7	1	2	4	6	9	6	9	1	1	1	7	9
29	6	1	2	9	8	0	3	9	5	0	7	4	8	6	2	3
30	2	9	1	0	8	6	7	4	5	2	9	5	6	2	1	5
31	3	7	9	8	0	9	7	1	9	1	3	8	7	7	3	8
32	9	6	5	0	5	1	0	6	9	7	1	5	4	7	5	9
33	2	2	9	3	1	1	0	5	1	5	8	4	4	9	7	6
34	5	8	9	9	9	7	1	0	7	9	6	9	4	3	4	6
35	1	9	8	0	6	6	5	2	4	1	0	7	1	0	1	6
36	6	6	9	3	9	0	9	3	3	5	6	6	9	0	3	0
37	3	1	7	4	7	0	0	5	9	6	9	4	5	3	0	2
38	9	8	0	3	4	9	1	2	4	0	7	7	6	9	6	1
39	9	9	7	3	1	0	3	3	8	8	2	2	4	3	4	6
40	6	3	8	2	0	7	2	6	1	6	4	3	1	1	1	8

Reproduction of Table I of A. Bradford Hill: *Principles of Medical Statistics*, London, Lancet, by permission of the author and publishers.

TABLE OF RANDOM NUMBERS

	1	2	3'	4	5	6	7	8	9	10	11	12	13	14	15	16
1	2	8	6	9	3	0	9	6	6	3	9	2	9	6	6	5
2	4	0	4	0	5	8	7	3	9	4	3	7	7	6	6	4
3	1	6	0	2	7	7	3	1	0	4	9	9	4	2	7	9
4	8	4	1	3	1	8	5	0	5	6	3	7	4	7	2	9
5	8	5	7	5	3	7	7	0	3	2	4	9	4	0	1	5
6	4	4	3	4	8	5	0	2	6	6	2	5	8	6	8	0
7	3	7	2	3	0	4	6	0	3	0	7	3	4	0	1	8
8	7	0	0	9	8	0	7	4	9	2	6	6	6	9	1	9
9	9	7	6	5	6	0	9	7	4	4	7	0	8	0	5	8
10	3	2	5	9	9	3	9	7	8	3	6	1	8	1	0	4
11	9	8	3	6	0	3	8	9	7	4	5	0	4	9	4	2
12	1	8	2	9	0	1	3	2	1	4	6	8	2	6	9	8
13	2	1	2	6	4	9	8	3	0	4	6	1	9	8	0	6
14	9	5	1	4	7	5	6	4	1	4	0	3	2	7	4	3
15	0	5	1	0	5	5	2	9	4	8	8	7	7	8	2	1
16	2	8	8	4	5	9	7	8	7	4	2	3	3	7	4	9
17	6	5	6	3	2	6	0	5	0	0	4	9	6	6	7	0
18	8	8	8	0	1	6	9	6	1	8	6	8	6	3	3	3
19	3	1	9	3	5	3	3	6	5	0	9	6	5	0	1	8
20	4	1	4	6	6	7	1	1	4	4	5	1	0	0	5	9
21	4	7	4	0	7	5	0	6	8	5	6	6	4	4	4	2
22	1	8	7	5	4	8	2	6	7	1	3	[illegible]	6	2	3	7
23	8	0	3	6	6	5	2	5	9	9	3	9	0	8	8	9
24	4	4	7	0	2	1	8	1	9	7	8	5	7	5	3	5
25	2	9	1	9	8	6	2	0	4	5	0	3	5	4	4	1
26	0	3	4	2	5	9	4	8	6	2	1	5	7	2	7	2
27	9	1	5	9	4	6	8	6	4	5	2	0	4	8	7	6
28	0	1	9	6	8	5	3	7	3	1	5	9	4	7	0	8
29	6	1	6	2	0	1	3	6	9	6	6	0	1	1	8	7
30	5	9	3	6	0	5	4	9	4	8	9	2	9	1	8	5
31	0	9	0	2	7	8	9	9	0	4	6	7	1	2	0	7
32	7	5	0	3	5	8	7	2	7	6	8	3	8	7	4	5
33	6	4	0	4	7	3	6	1	3	7	2	7	1	2	7	4
34	3	8	5	1	4	5	2	4	5	0	8	2	2	9	1	5
35	4	4	8	1	9	7	6	9	4	0	5	7	4	6	2	9
36	3	1	2	3	9	6	2	2	1	4	6	8	8	5	1	2
37	1	4	1	9	4	7	1	8	6	4	7	3	1	3	2	6
38	3	3	3	9	5	5	6	0	5	3	2	0	6	7	6	3
39	1	8	4	8	5	6	3	8	4	3	7	8	2	2	7	7
40	8	8	3	5	8	6	3	9	0	6	0	3	4	7	4	5

Reproduction of Table II of A. Bradford Hill: *Principles of Medical Statistics*, London, Lancet, by permission of the author and publishers.

APPENDIX 2

TWO-TAILED PROBABILITY VALUES FOR THE NORMAL DEVIATE (z)

z	P
1.645	0.1
1.960	0.05
2.326	0.02
2.576	0.01
3.291	0.001

Condensed from Table A-1 of P. Armitage: *Statistical Methods in Medical Research,* Oxford, Blackwell Scientific, by permission of the author and publishers.

APPENDIX 3

PROBABILITY VALUES FOR CHI-SQUARE (χ^2)

Degrees of freedom	*Probability*			
	0.10	0.05	0.01	0.001
1	2.706	3.841	6.635	10.827
2	4.605	5.991	9.210	13.815
3	6.251	7.815	11.345	16.268
4	7.779	9.488	13.277	18.465
5	9.236	11.070	15.086	20.517
6	10.645	12.592	16.812	22.457
7	12.017	14.067	18.475	24.322
8	13.362	15.507	20.090	26.125
9	14.684	16.919	21.666	27.877
10	15.987	18.307	23.209	29.588
11	17.275	19.675	24.725	31.264
12	18.549	21.026	26.217	32.909
13	19.812	22.362	27.688	34.528
14	21.064	23.685	29.141	36.123
15	22.307	24.996	30.578	37.697
16	23.542	26.296	32.000	39.252
17	24.769	27.587	33.409	40.790
18	25.989	28.869	34.805	42.312
19	27.204	30.144	36.191	43.820
20	28.412	31.410	37.566	45.315

Adapted from Table IV of R. A. Fisher and F. Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, London, Longmans Group, by permission of the authors and publisher.

APPENDIX 4

TWO-TAILED PROBABILITY VALUES FOR STUDENT'S *t* TEST

Degrees of freedom	*Probability* 0.10	0.05	0.01	0.001
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.941
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.859
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.405
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.767
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
27	1.703	2.052	2.771	3.690
28	1.701	2.048	2.763	3.674
29	1.699	2.045	2.756	3.659
30	1.697	2.042	2.750	3.646
∞	1.645	1.960	2.576	3.291

Adapted from Table III of R. A. Fisher and F. Yates: *Statistical Tables for Biological, Agricultural and Medical Research*, London, Longmans Group, by permission of the authors and publisher.

APPENDIX 5

BACKGROUND AND SOURCE MATERIAL

Epidemiology

E-1. Lilienfeld, A. M. *Foundations of Epidemiology.* New York: Oxford University Press, 1976.

E-2. Lilienfeld, A. M., Pederson, E., Dowd, J. E. *Cancer Epidemiology: Methods of Study.* Baltimore: Johns Hopkins University Press, 1967.

E-3. McMahon, B., Pugh, T. F. *Epidemiology, Principles and Methods.* Boston: Little, Brown, 1970.

E-4. Fox, J. P., Hall, C. E., Elverback, L. R. *Epidemiology: Man and Disease.* New York: Macmillan, 1970.

E-5. Slonim, M. J. *Sampling.* New York: Simon & Schuster, 1960.

Statistics

S-1. Armitage, P. *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications, 1971.

S-2. Swinscow, T. D. V. *Statistics at Square One.* London: British Medical Association, 1976.

S-3. Fliess, J. L. *Statistical Methods for Rates and Proportions.* New York: John Wiley & Sons, 1973.

S-4. Snedecor, G. W., Cochran, W. G. *Statistical Methods.* Ames: Iowa State Press, 1967.

S-5. Neter, J., Wasserman, W. *Applied Linear Statistical Models.* Homewood, Ill.: Richard D. Irwin, 1974.

S-6. Hill, A. B. *A Short Textbook of Medical Statistics.* Philadelphia: Lippincott, 1977.

Index