Analysis and Interpretation of Data

S. James Kilpatrick, PhD and Maurice Wood, MD
Richmond, Virginia

The analysis and interpretation of data is an area of particular difficulty for researchers in primary care. Most primary care physicians have little prior experience in the research field and their tendency is to leave analysis and interpretation of data to the statisticians or to blindly follow statistical advice provided out of context or in a series of corridor consultations with a statistician.

This road leads inevitably to disaster and wasted effort. The statistician should be involved as early as possible in the beginning of and continuously throughout a study.

The presentation format of the results of the study should be determined during the planning stage, and blank frequency tables, histograms, and bar charts created.

As the authors infer later, there is some inappropriate use of statistical tests in family medicine literature. The overuse of statistical tests in research studies should be guarded against as assiduously as the overuse of laboratory tests in patient care.

In this paper, the terms used will be as defined in the Glossary for Primary Care.[1]

Descriptive Studies

Population and Units

Articles in the literature of primary care are frequently ambiguous about the underlying unit being counted or measured. Confusion occurs between the patient and his/her problem (see, for example, Marsh et al[2]) and between the patient and his/her visits (see, for example, a recent editorial in *Patient Care*[3]). Each table in a paper should identify the unit being counted or measured. The focus will change from workload studies[4-7] (in which the physician is the unit of interest), to content of care studies[8,9] (in which the problem is the unit of interest), to epidemiological studies[10-14] (in which the patient is the unit of interest).

A *descriptive study*[5,7-9] is one which, on the basis of records collected, describes some facet of family practice in a particular situation. This description is restricted to a specified population of patients, problems, providers, or practices. Having specified the unit of interest and the population, we can then turn to the *data elements* collected on each unit of interest. From a family practice viewpoint these may be classified as descriptors of patients, problems, providers, or practices; or morbidity, encounter, or service descriptors; or as treatment, intervention, or outcome descriptors. However, it is convenient, for subsequent statistical analysis,[15] to consider data elements on each unit of interest as either qualitative or quantitative. (These terms may be further subdivided into nominal, ordinal, continuous, or discrete counts.)[16]

As will be seen, this classification of the basic data into qualitative or quantitative elements is used to determine both the method by which the data are summarized in a descriptive study and also the test of significance used in an inferential study.

Time Frame

Information on a unit may be collected once only or at different times. Studies may therefore be classified as:

current—information collected at one point in time or over a given (short) period.[4,5]

follow-up—information on the *same* units collected over an extended period.[12,14,17,18]

repeated—information on units (not all the same) collected at different times.[19]

These types of studies, however, merge into one another. Thus a current study[9] over a protracted time period becomes a follow-up study,[18] and with turnover and losses, a follow-up study will become a repeated study. The focus of interest will largely determine whether a study is to be considered a current, follow-up, or repeated study.

Statistical Summaries

Statistical summaries are of various types. Each will be described and illustrative examples given.

One data element from a current study

A *quantitative* variable may be summarized by a one-way frequency table, a histogram, or by a

Dr. Kilpatrick is Professor and Chairman, Department of Biostatistics, Medical College of Virginia/Virginia Commonwealth University, Richmond, Virginia. Dr. Wood is Professor and Director of Research, Department of Family Practice, Medical College of Virginia/Virginia Commonwealth University, Richmond, Virginia.

0094-3509/78/0703-0101$03.00

Table 1. Age Distribution of Women Taking Oral Contraceptives

Age (in years)	Number	Percent
15-19	1,291	5.5
20-24	7,006	29.7
25-29	6,094	25.8
30-34	4,692	19.9
35-39	2,905	12.3
40-44	1,291	5.5
45+	332	1.4
	23,611	100.0

Mean age 28.79 years.

Adapted with permission from the Royal College of General Practioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

Table 2. Status of Women Taking Oral Contraceptives

Status	Number	Percent
'New' takers	4,851	20.5
Previous takers	18,755	79.4
Unknown	5	0.0
	23,611	99.9

Adapted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

summary statistic. Thus, the age distribution of women taking oral contraceptives[14] in the Royal College of General Practitioners' (RCGP) study was shown both as a frequency table and as a mean (Table 1).

A *qualitative* variable may similarily be summarized by a one-way table or bar chart. Thus, the status of the 23,611 women with regard to oral contraceptives at the beginning of the study was as shown in Table 2.[14]

Note that whereas age is a quantitative variable which can be summarized by the mean, the status of women with regard to oral contraceptives at the beginning of the study is a qualitative variable which cannot be summarized by the mean.

Two data elements from a current study

Two *quantitative* variables may be summarized by a two-way frequency table, a scatter plot, or by the correlation coefficient.

Thus, the 23,611 oral contraceptive takers in the RCGP study[14] could have been tabulated by age and the number of cigarettes smoked daily as shown in Table 3.

Table 3 was, in fact, summarized in the RCGP study[14] by stating the mean ($\bar{x}$) in each age group as follows (Table 4).

A scatter plot is inappropriate in this case since it would contain 23,611 plots, each woman contributing a plot of her age against the daily number of cigarettes smoked. Figure 1 shows an example of a scatter plot of two quantitative variables given by Kilpatrick.[20]

Table 3. Age and Smoking History for Women Taking Oral Contraceptives

Age	0	1-4	5-9	10-14	15-19	20+	Total
15-19	670	80	97	175	113	156	1,291
20-24	3,635	436	525	951	651	844	7,006
25-29	3,162	379	457	827	535	734	6,094
30-34	2,435	292	352	637	412	564	4,692
35-39	1,507	181	218	394	255	350	2,905
40-44	670	80	97	175	113	157	1,291
45+	173	21	23	46	28	40	332
Total	12,252	1,469	1,769	3,205	2,071	2,845	23,611

(Note: cell frequencies are hypothetical)
Adapted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

Table 4. Mean Daily Cigarettes for Women Taking Oral Contraceptives

Age	Mean Daily Cigarettes
15-19	6.38
20-24	6.41
25-29	6.34
30-34	6.34
35-39	6.44
40-44	7.75
45+	6.27

Adapted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

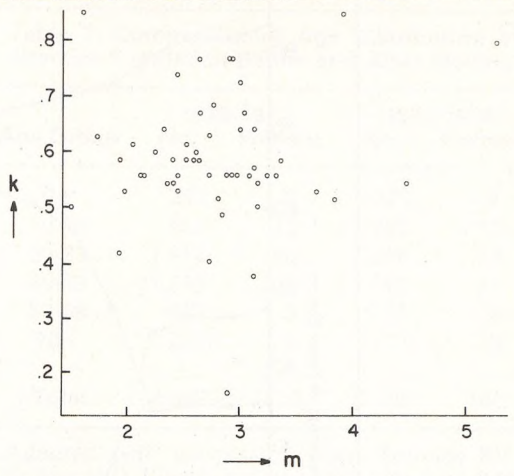

Figure 1. Consultation Frequencies in 53 Practices Summarized by the Parameters *m* and *k* of the Negative Binomial.

Reprinted with permission from *Health Services Research,* Volume 12, Number 3, Fall 1977, p 294. Copyright 1977 by the Hospital Research and Educational Trust, 840 N Lake Shore Drive, Chicago, IL 60611.

This figure shows each practice in the Second National Morbidity Survey represented by a plot of *m* against *k* to summarize the recorded distribution of consultations in that practice. The association (or lack of association) of the two variables plotted may then be summarized by a correlation coefficient which here is 0.11,[20] not significantly different from zero.

Table 5. Analysis of Deaths in the RGCP Oral Contraceptive Study

Cause of Death	Takers	Ex-takers	Controls	Total
Malignancy	6	4	11	21
Violence	13	3	6	22
Vascular	9	3	4	16
Other	2	1	7	10
Total	30	11	28	69

Adapted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

Table 6. Cigarette Smoking by Social Class Among Oral Contraceptive Users

Social Class	Mean Daily Cigarette Consumption
I	4.45
II	5.29
III non-manual	5.58
III manual	7.07
IV	7.26
V	8.22

Adapted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

Two *qualitative* variables may be summarized by a two-way frequency table. (See, for example, Table 5.)

A quantitative and qualitative pair of variables may be summarized by a two-way frequency table or a one-way table showing a summary statistic of the quantitative variable by different levels of the qualitative variable (Table 6).

In Table 6 social class is the qualitative variable and the number of cigarettes smoked daily is the quantitative variable. The original table showing the number of women taking oral contraceptives by social class and the number of cigarettes

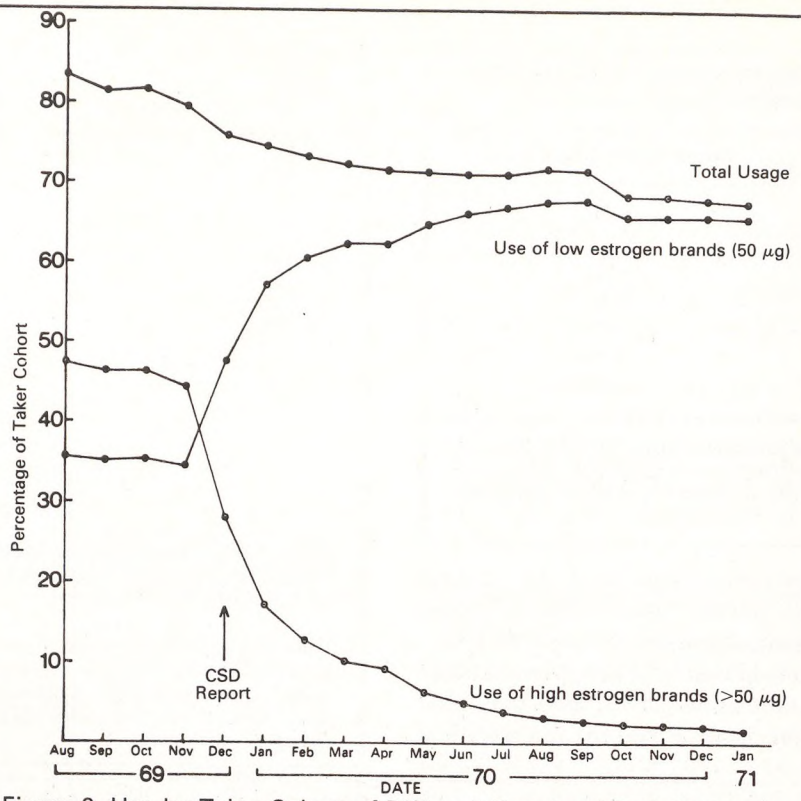

Figure 2. Use by Taker Cohort of Different Brands of Oral Contraceptives (August 1969 to January 1971)

Reprinted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

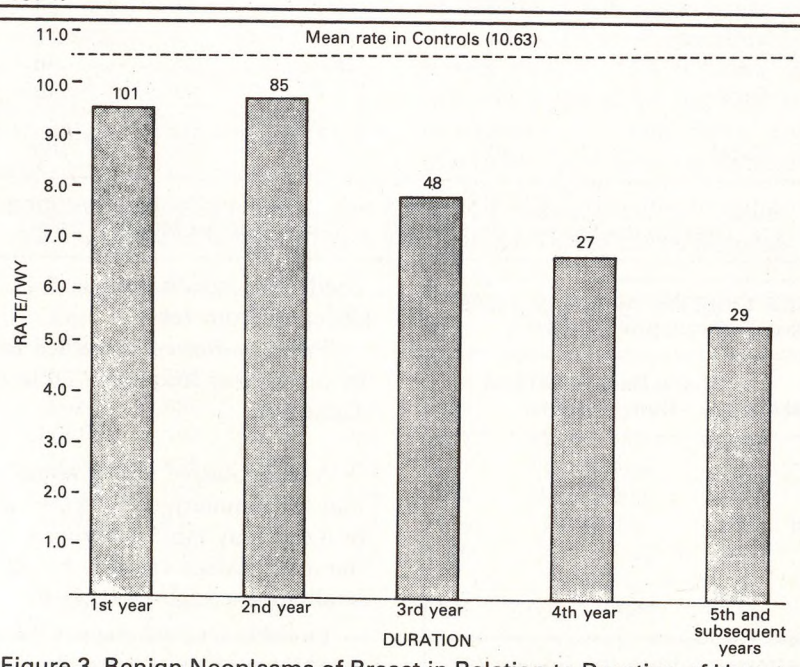

Figure 3. Benign Neoplasms of Breast in Relation to Duration of Usage of Oral Contraceptives (TWY-Thousand Woman Years)

Reprinted with permission from the Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974.

smoked daily has been summarized by tabulating the mean number of cigarettes smoked daily in each social class.

One data element from a follow-up study

A *quantitative* variable followed over time in the same units may be collated by a series of summary statistics for different periods, by an average trend line, or by statistical summaries of the differences between each pair of successive observations. Thus, in a follow-up study of untreated chemical diabetics,[17] the mean fasting blood glucose and its standard deviation are given for each year of follow-up in patients stratified by age at entry to the study (6 under 35 years, 10 aged 35 to 44, and 11 over 44 years).

A trend may be shown graphically, as in the RCGP Oral Contraceptive Study,[14] where the percentage of the cohort using high estrogen oral contraceptives fell off markedly following the publication in December 1969 of a report by the Committee on the Safety of Drugs (CSD report) recommending that only low estrogen brands of contraceptives be taken under normal conditions (Figure 2).

A *qualitative* variable followed over time in the same units may be summarized by a rate for a binomial variate, by a series of one-way frequency tables for a multinomial variate, or by a series of rankings of the units.

Thus the episode rate of benign breast tumors is related to the duration of ingestion of oral contraceptives by plotting rate against duration (Figure 3).

In the same way, treating age as a multiple classification rather than a continuous variate, the change over time of the age distribution in a practice was given as shown in Table 7.

One data element from a repeated study

A *quantitative* variable determined at different times on units which may differ may be collated by a series of summary statistics for different periods. Table 7 applies here especially since it is rare in family practice for the units of interest (patients, practices, or providers) to remain constant from one period to the next.

Likewise, when comparing the most frequent complaints over time, it may be sufficient to compare their ranks, as in Table 8.

Table 7. Comparison of Age Distribution of Practice Population Before and After Moving

Age Group	1970-1971 No.	1970-1971 Percent	1973-1974 No.	1973-1974 Percent
0-9	391	8	229	7
10-19	629	13	490	15
20-39	1,913	40	1,366	43
40-59	1,213	25	667	21
60-69	440	9	274	9
70+	262	5	172	5
Total	4,848	100	3,198	100

Adapted with permission from Rudnick KV, Spitzer WO, Pierce J: Comparison of a private family practice with a university teaching practice. J Med Educ 51: 395, 1976.

Rates: Numerators and Denominators

A rate adjusts a response to a per unit basis. The numerator of a rate should therefore either be the number in the denominator which has some attribute, or the number of times an event occurred to the units in the denominator. The more common rates used in family practice have been previously defined.[1] Depending on the focus of interest, the denominator could be patients, problems, providers, or practices. Patient populations can be the study population or registered population,[1] or can be defined in terms of some patient characteristic or characteristics (active, visiting, attending, temporary).

Rates: Comparison and Adjustment

There is no reason why the same rate from different populations should be the same. Differences among different populations can, however, be examined to study the likely reasons for the observed difference or differences.

Age and sex differences between two populations often explain most of the difference between two rates. Thus, episode rates usually increase with age, and females generally have a higher rate than males.[19] We can attempt to remove the effect of age or sex on an observed difference in rate either by comparing the rate in subgroups which have the same characteristic (all males, all

Table 8. Profiles of the Ten Most Frequent Complaints by Period

Complaint	1970-1971 Rank	1970-1971 % of Visits	1973-1974 Rank	1973-1974 % of Visits
Cold and flu	1	11.1	1	15.9
Sore throat	2	8.5	2	12.2
Extremity pain	3	7.1	4	8.1
Cough	4	6.0	6	5.0
Nausea and vomiting	5	4.8	10	1.1
Rashes and skin lesions	6	4.4	5	7.2
Abdominal pain	7	4.1	8	3.4
Earache	8	3.9	9	3.0
Back pain	9	3.7	7	4.4
Fatigue and depression	10	3.0	3	10.1
All other	11	43.4	11	29.6

Adapted with permission from Rudnick KV, Spitzer WO, Pierce J: Comparison of a private family practice with a university teaching practice. J Med Educ 51: 395, 1976.

females, or a given age group), or by standardizing for age and sex. In direct standardization, specific morbidity rates (for example, age and sex groups) are combined in proportion to the (age and sex) composition of an ideal population.

Standardization or comparison among like subgroups can be used to adjust for differences due to known factors (ie, factors known both to be associated with the level of the morbidity rate and by which each unit in the population may be classified). Unfortunately, we often do not know what caused the observed difference and if we did, we cannot adjust for its influence because that information on the units is missing. Thus, in very few studies can we infer a cause-and-effect relationship.

Standardization, both direct and indirect, is generally explained in textbooks in terms of mortality,[21] but is equally applicable to morbidity. Direct standardization of episode rates in family practice has been used to adjust for the different age and sex composition of patients seen by first, second, and third year residents, and by the faculty in three Virginia practices.[22] Indirect standardization was used in the RCGP Oral Contraceptive Study[14] to adjust for age differences among the three groups compared (takers, ex-takers, and controls). We prefer direct standardization since indirect standardization leads to the Standardized Mortality Ratio, an index which is often used erroneously in multiple comparisons.[23]

Inferential Studies

Sampling

Statistical thinking is usually concerned with how to make inferences from a sample of units. Sampling, ie, the measurement, examination, or testing of a small fraction of the units of a population, is cheap, saves time, and is geared to providing answers which will specifically answer the question with the required accuracy. Sampling is a much neglected tool in family practice research,[24,25] although it is used clinically and diagnostically (blood samples, urine and fecal specimens, skin biopsies).

Studies which use sampling are of two types—those which estimate some characteristic or characteristics of the population sampled and those which are concerned with determining which population the observed sample came from. These latter studies generally end with a test of significance.

Types of Research Studies

Descriptive

If based on the whole population, the descriptive study is called a census. If based on a sample, it is called a sample survey.

Comparative

In a comparative study, two or more groups of

units are compared.[26] Clearly, a comparative study can contrast some characteristic of two or more populations or can compare the characteristic in *samples* taken from the different populations.

No inference is necessary in a comparison of populations. They are either alike or are different with regard to some characteristic. In a comparison based on samples, allowance must be made for sampling variation. Therefore, even if a difference is found in some characteristic between two samples, this does not necessarily mean that the populations are different.

Experimental

Statisticians use this term to mean a study in which either the samples are drawn strictly at random from the populations and/or the treatments are assigned at random. Statistical inference is based on the assumption of randomized allocation and/or sampling. A frequent practice is to draw inferences from a nonrandomized study using conventional statistical tests of significance designed for randomized samples. Such inferences, however, can be misleading and should be made with caution.[27]

Where in this scheme do controlled clinical trials, epidemiological studies, and intervention studies fall? A *controlled clinical trial*[21] is a study in which treatments are allocated strictly at random to study subjects. It is therefore an experiment. An *uncontrolled clinical trial*[28] is one without a control group. An *epidemiological study*[29] does not involve randomization but does involve a comparison of different groups and is therefore a comparative study. An *intervention study* is one in which a baseline is established, a treatment is administered, and a follow-up period ensues; it is a before-and-after study and is, therefore, a comparative study. Just as confidence limits may be calculated on a sample estimate when that sample is not randomly selected, we can also apply statistical tests of significance to comparative studies even though treatments were not allocated at random. This has led to the terms quasi-experiments[30] and pseudo-experiments being used by applied statisticians for comparative studies.

To date, much of the research in family practice has involved descriptive studies (ad hoc descriptions of one practice,[4,12] of many practices,[5,8,9,31-34] or simple comparisons among practices,[2,18] or surveys[19,22]).

Although "the general scientific problem with which we are concerned is that of testing a hypothesis that a certain treatment alters the natural history of a disease for the better,"[35] a review of the family practice literature reveals few experimental studies. Some examples of controlled clinical trials involving randomization in general practice have been published,[36-39] together with the use of control groups,[40] both matched[41] and randomly selected.[42]

Since the family physician provides "continuing comprehensive health maintenance and medical care to the entire family,"[1] it is surprising how few longitudinal or family studies have been published. Examples of long term follow-up studies may be found in the treatment of hypertension,[43] the care of the aged,[44] and in the psychiatric morbidity of John Fry's practice.[45] Likewise, there have been relatively few outcome or intervention studies reported. Kuenssberg and Knox[46] report on following 10,000 pregnancies to term, and some intervention studies[47,48] are concerned with the operation of the practice rather than the effect of a change in treatment.

Excellent reviews and summaries of general practice/family practice are to be found in several texts.[49-51] These same texts, however, omit specific consideration of the need for rigorous research design, as exemplified by Cochrane[35] in controlled clinical trials or by Susser[29] in the inferences which can be made from epidemiological investigations. Most elementary statistical books[16,21,27,30,52] contain a section on the design of experiments and surveys, and any statistician may be consulted to provide specific advice or assistance in creating specific designs to meet the study objectives.

Concepts in Statistical Inference

The statistical concepts described below deal mostly with quantitative variables whereas family practice research mostly uses qualitative variables! This may be one reason for the inappropriate use of tests of significance by some family practice researchers.

Standard Error and Standard Deviation

Both of these summary statistics measure variation, the standard deviation estimating the varia-

tion of the basic quantitative variable, and the standard error the variation from sample to sample of some summary statistic such as the mean. The standard error of a proportion describes the variation from sample to sample of the proportion. The sample proportion is a summary statistic already, combining the yes/no, dead-or-alive observations made for each unit. The standard error and standard deviation are sometimes confused in clinical papers. Often in articles no indication is given of which measure of variation is intended when a figure is given after a ± sign.

In describing a follow-up of a cohort of chemical diabetics (subdivided by duration since the beginning of the study), the mean and the standard deviation were used by Logie et al[17] to describe both the fasting blood glucose (mg/100 ml) and the body weight (kg). The reader can run his eye down these columns of numbers and see that, over the seven years of follow-up, the blood glucose increased in the 11 (n=11) diabetics over 44 years of age, both in terms of the mean (from 75 to 90) and in terms of the standard deviation (from 7.1 to 14.3). Therefore, the overall increase in blood glucose was accompanied by an increase variability in the 11 patients. This could have been the result of some two or three of the patients maintaining a constant blood glucose in contrast to the others whose blood glucose increased. In this same group, body weight varied from 67.6 ± 7.1 at the start to 65.8 ± 8.5 at the end of the study, demonstrating no increase in body weight and little change in the variability of weight over the period.

In another study,[53] on umbilical cord urea in newborns, birth weight (in ounces) and gestation were given as shown in Table 9.

The interest here is in evaluating the differences of the means among the maternal classifications. The standard error is therefore given to indicate how variable these means would be if the study were repeated with the same number of mothers in each group. Note that the largest standard errors occur in the smallest group (Other hypertensive) and the smallest standard errors in the largest group (Normotensive). This is as expected since it is known that the standard error decreases as the sample size increases (in fact is proportional to $1/\sqrt{n}$).

Confidence and Tolerance Limits

Since confidence and tolerance limits are calculated using either the standard error or the standard deviation, these limits can also be confused. Confidence limits attempt to indicate the range in which the estimated parameter (variate of interest) will fall in the population, with a stated probability, and use the standard error. Tolerance limits give limits within which a certain proportion of the population are expected to lie, with a stated probability, and are based on the standard deviation.

Thus, confidence limits for cord urea in children of toxemic and normotensive mothers were calculated and plotted as shown in Figure 4 to demonstrate that the mean cord urea in infants from toxemic mothers is likely to be higher than the mean cord urea in children from normotensive mothers.

If these were tolerance limits, the equivalent statement could be made for a comparison of individual newborns. As it stands, this figure is in terms of confidence limits of means and does *not* rule out that an individual newborn from a toxemic mother may have a low cord urea comparable to that of a child of a normotensive mother.

Sensitivity/Specificity

Specificity means the precision with which the occurrence of one event or value will predict the occurrence of another.[29] Specificity and sensitivity

Table 9. Mean and Standard Errors by Mother's Classification

Maternal Classification	No.	Birth Weight (oz)	Gestation (weeks)
Normotensive	96	94.4 ± 2.4	38.0 ± 0.3
Other hypertensive	15	115.7 ± 6.2	39.9 ± 0.8
Toxemic	44	85.2 ± 3.6	37.2 ± 0.5

Adapted with permission from Kilpatrick SJ, McKay E: Umbilical cord urea in toxaemia of pregnancy. J Obstet Gynaecol Br Commonw 72: 402, 1965.

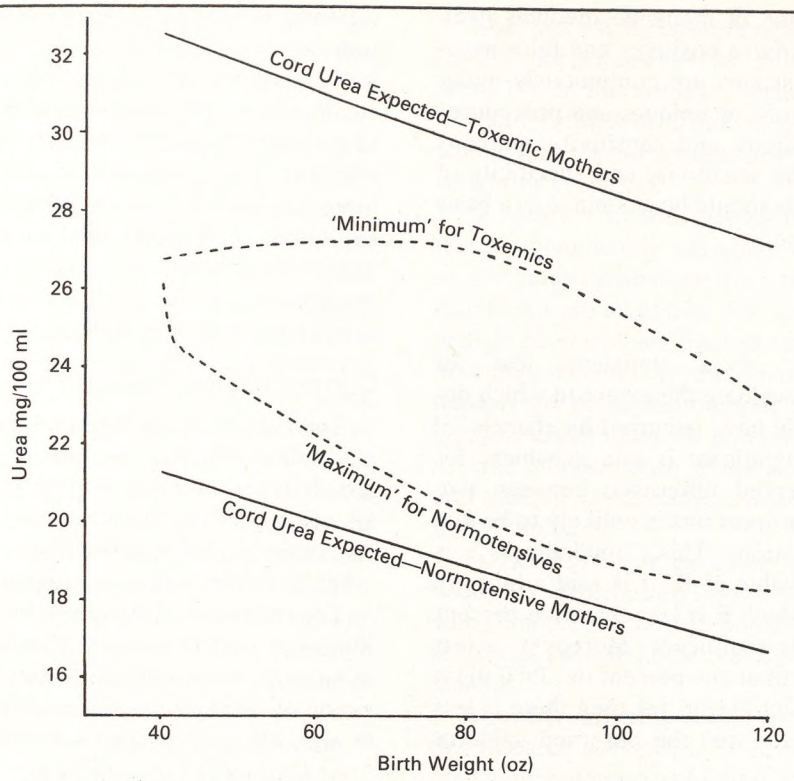

Figure 4. The Relationship Between Cord Urea and Birth Weight for Children of Normotensive and Toxemic Mothers.
(The dotted lines show the lower 95 percent confidence limits [single tail] for the toxemic group and the upper 95 percent confidence limits [single tail] for the normotensive group.)

Reprinted from Kilpatrick SJ, McKay E: Umbilical cord urea in toxaemia of pregnancy. J Obstet Gynaecol Br Commonw 72:402, 1965.

arise together in diagnostic testing in which two types of errors may be made—a false positive or a false negative.

Specificity is, then, the probability that a negative test is, in truth, negative, and sensitivity, the probability that a positive test is, in truth, positive. In other words,[54]

$$\text{Sensitivity (\%)} = \frac{\text{the number of diseased persons with a positive test}}{\text{the total number of diseased persons tested}} \times 100,$$

and

$$\text{Specificity (\%)} = \frac{\text{the number of nondiseased persons with a negative test}}{\text{the total number of all nondiseased persons tested}} \times 100.$$

It follows that the probability of a false positive is one minus the specificity and that the probability of a false negative is one minus the sensitivity.[52] In other words,[54]

$$\text{P (false positive)} = \frac{\text{the number of false positives}}{\text{the total number of all nondiseased persons treated}}$$

$$= 1 - \text{Specificity, and}$$

$$\text{P (false negative)} = \frac{\text{the number of false negatives}}{\text{the total number of all diseased persons tested}}$$

$$= 1 - \text{Sensitivity.}$$

Since the practice of medicine involves decisions, examples of false positives and false negatives abound. Physicians are continuously under pressure to adopt new techniques and procedures of unknown specificity and sensitivity; perhaps before doing so, the sensitivity and specificity of established methods should be examined as a basis for later comparison.

Significance

"The purpose of a statistical test [of significance] is to evaluate the extent to which observed results could have occurred by chance."[14] A test which is significant is one in which, for example, an observed difference between two sample means or proportions is unlikely to be due to sampling variation. This "unlikeliness" is measured by the value of P. It is said arbitrarily that any test for which P is less than five percent (written $P<0.05$) is significant. Moreover, a test for which P is less than one percent (ie, $P<0.01$) is said to be highly significant, for then there is less than 1 chance in 100 that the observed value or difference could have occurred by sampling variation.

Too much emphasis can be placed on a test of significance. "With small numbers it is very easy to give the impression that a treatment is no more effective than a placebo. . . .Alternatively, with large numbers it is often possible to achieve a result that is statistically significant but may be clinically unimportant. . . .Another snag is that [the randomization required for clinical trials may] not always be applicable for ethical reasons."[35]

Common Significance Tests

The nature of the data and the objective of the study determine which statistical test is appropriate. It is impractical here to outline all possible situations, so the authors have listed in Table 10 the more common statistical tests and indicated when and where they are applicable.

For examples of the χ^2 test in the family practice literature see Cooper et al,[45] where the association of sex and age (under and over 40 years) with duration of mental illness is examined, and Rudnick et al,[18] where χ^2 is used to compare standardized

Table 10. Uses of Common Statistical Tests

Test	Types of Data	Some Assumptions or Restrictions
χ^2	A one or two-way table of frequencies or counts	The expected value should not fall below 2
t	Comparison of the means or proportions in one or two samples	Equal variances or common population proportions in two groups; Independence Random selection
Analysis of Variance	Comparison of the means of more than two groups or subgroups	Equal variances Independence Normality Random selection
Significance of (product-moment) Correlation	Quantitative variates	Normal distribution Random selection
Significance of Regression Coefficient(s)	Quantitative response variate(s)	Normally distributed errors Fixed independent variate(s)

rates of hospitalization and the utilization of drugs in the management of indicator diseases. (Review of this latter paper does not make clear how the analysis follows from the basic records.)

The t-test of difference of means is illustrated in Kilpatrick and McKay.[53] The statement, "Underweight infants of normotensive mothers had a significantly higher mean urea than their fellows of normal weight ($t=2.04$, $df=94$, $P<0.05$)," is the authors' summary of a t-test to evaluate the size of the difference between the mean urea (22.6) of a group of 14 underweight babies compared with the mean urea (18.7) of a group of 82 babies of normal weight at birth. The value of t (2.04) is the value of the test statistic. The associated P value is obtained by looking up tables of t, entering the table at a row labeled df=94. (This means degrees of freedom which in this type of test is two less than the number of patients studied [ie, $df=14+82-2=94$].) The value of P associated with a t value of 2.04, and 94 degrees of freedom is less than five percent (written $P<0.05$), and so the authors conclude that the difference is likely to be real and not a chance observation due to sampling variation.

The same study[53] may be used to illustrate testing the significance of correlation coefficients. The correlation coefficient was calculated between urea and weight and gestation in each of three groups, as shown in Table 11.

Note that a significant correlation was found only in the normotensives as indicated by the asterisk (*). In this context "significant" means significantly different from zero, the situation in which there is no association between urea and weight or gestation.

In addition to these common tests, which all test for quantitative variables with the exception of χ^2, many nonparametric, distribution-free, or exact tests are available. The type of data in these tests is obvious from the test statistic (often ranks or comparisons). As their names suggest, they make fewer or no assumptions concerning the underlying distribution. In general, however, they do assume independence of successive observations and random selection from a well-defined population of units.

Russell[24] gives an excellent justification for the need for distribution-free tests of significance when comparing consultation distributions, and he illustrates two methods (the Wilcox and McNemar tests).

A further difficulty is that much of family practice research uses rates. "Because morbidity data are expressed in terms of rates, common statistical tests which demand the use of actual numbers are inapplicable."[14] This difficulty has caused these RCGP investigators to create a new test of significance[14] specifically for their study.

Table 11. Estimated Correlation of Urea with Weight and Gestation by Maternal Classification

Maternal Classification	No.	Correlation of Urea with Weight	Correlation of Urea with Gestation
Normotensive	96	−0.20*	−0.24*
Other hypertensive	15	−0.04	−0.11
Toxemic	44	−0.15	−0.05
Combined	155	−0.18*	−0.15

*Estimates of this absolute size or greater would not occur in more than five percent of random samples from a large population of infants in which the correlation was zero.

Adapted with permission from Kilpatrick SJ, McKay E: Umbilical cord urea in toxaemia of pregnancy. J Obstet Gynaecol Br Commonw 72: 402, 1965.

This further illustrates the absolute necessity for statistical consultation at the very beginning and throughout the period of a research study in family medicine.

References

1. Froom J, Anderson JE, Bass M, et al: A glossary for primary care. J Fam Pract 5:633, 1977
2. Marsh GN, Wallace RB, Whewell J: Anglo-American contrasts in general practice. Br Med J 1:1321, 1976
3. Gordon RD (consultant), Kasanof D (editor): Getting to know your patient family. Patient Care 10(1):159, 1976
4. Collyer JA: A measure of a family doctor's work: Part 1: Setting up a one-practice study. Can Med Assoc J 112:1256, 1975
5. Williams WO: A study of general practitioners' workload in South Wales, 1965-1966. Reports from General Practice of the J R Coll Gen Pract, No. 12, 1970
6. Hodgkin GK: Evaluating the doctor's work. J R Coll Gen Pract 23:759, 1973
7. Labelle C: Alberta workload study. Can Fam Physician 19:100, 1973
8. Last JM, White KL: Content of medical care in primary practice. Med Care 7:41, 1969
9. Marsland DW, Wood M, Mayo F: A data bank for patient care, curriculum, and research in family practice. J Fam Pract 3:37, 1976
10. White KL: Contemporary epidemiology. Int J Epidemiol 3:295, 1974
11. Baikie AG: Epidemiological aspects of the Australian Morbidity Survey. Ann Gen Pract 14:175, 1969
12. Reinhard KR, Greenwalt NI: Epidemiological definition of the cohort of diseases associated with diabetes in southwestern American Indians. Med Care 13:160, 1975
13. Sklaroff SA: Use of NHS general practice records in epidemiological inquiries: Duodenal ulcer—A test case. Br J Prev Soc Med 17:177, 1963
14. Royal College of General Practitioners: Oral Contraceptives and Health. London, Pitman Medical, 1974
15. Bahn AK: Basic Medical Statistics. New York, Grune & Stratton, 1972
16. Colton T: Statistics in Medicine. Boston, Little, Brown, 1974, pp 11-61
17. Logie AW, Stowers JM, Dingwall-Fordyce I: Longitudinal study of untreated chemical diabetes. Br Med J 4:630, 1974
18. Rudnick KV, Spitzer WO, Pierce J: Comparison of a private family practice with a university teaching practice. J Med Educ 51:395, 1976
19. Crombie DL, Pinsent RJFH: Comparison of the first and second national morbidity surveys. J R Coll Gen Pract 25:874, 1975
20. Kilpatrick SJ: Consultation frequencies in general practice. Health Serv Res 12:284, 1977
21. Hill AB: Principles of Medical Statistics, ed 9. London, Lancet, 1971
22. Kilpatrick SJ, Wood M: A comparison of the morbidity recorded in two family practice surveys. J Fam Pract 4:972, 1977
23. Kilpatrick SJ: Mortality comparisons in socioeconomic groups. J R Stat Assoc: Series C: Applied Statistics 12(2):65, 1963
24. Russell I: Matched controls in surveys of the use of general practice. J R Coll Gen Pract 25:801, 1975
25. Bevan JM, Draper GD: Sampling problems in studies of general practice. Med Care 3:168, 1965
26. Hills M: Statistics for Comparative Studies. London, Chapman and Hall, 1974, p 47
27. Mainland D: Elementary Medical Statistics, ed 2. Philadelphia, WB Saunders, 1963
28. Hesbacher P, Rickels K, Zamostien B, et al: A collaborative research model in family practice. J Fam Pract 4:923, 1977
29. Susser M: Causal Thinking in the Health Sciences. New York, Oxford University Press, 1973
30. Campbell DT, Stanley JC: Epidemiological and Quasi-Experimental Designs for Research. Chicago, Rand-McNally College Publishing, 1966
31. Morbidity statistics from general practice: Second National Morbidity Survey 1970-1971. In Office of Population Censuses and Surveys: Studies on Medical and Population Subjects, No. 26. London, Her Majesty's Stationery Office, 1974
32. Logan WPD, Cushion AA: Morbidity statistics from general practice: Volume 1 General: First National Morbidity Survey. In General Register Office: Studies on Medical and Population Subjects, No. 14. London, Her Majesty's Stationery Office, 1958
33. Johnson A: Report on a National Morbidity Survey, February 1962-January 1963. Part 1. Canberra, Australia, Australian National Health and Medical Research Council, 1966
34. Anderson JE, Lees REM: Patient morbidity and some patterns of family practice in southeastern Ontario. Can Med Assoc J 113: 123, 1975
35. Cochrane AL: Effectiveness and Efficiency. London, Nuffield Provincial Hospitals Trust, 1972, pp 20, 23
36. Murphy JE: The treatment of anxiety neurosis in general practice: A double-blind comparison of benzoctamine (Tacitin) with chlordiazepoxide (Librium). Clin Trials J 8(1):18, 1971
37. Bennett AE, Fraser IGP: Impact of a screening program in general practice: A randomized controlled trial. Int J Epidemiol 1:55, 1972
38. Rae RH, Hill GB: Disodium cromoglycate (Intal) in asthma: A controlled general practice trial. Clin Trials J 9(1):41, 1972
39. Howie JGR, Clark GA: Double-blind trial of early demethylchlortetracycline in minor respiratory illness in general practice. Lancet 2:1099, 1970
40. Lloyd G: The value of vaccination against influenza. Practitioner 205:357, 1970
41. Edwards LD: Malnutrition and disease in pre-school aboriginal children in the Walgett area of New South Wales. Med J Aust 1:1007, 1970
42. Watt CM: The estimation of hemoglobin and the diagnosis of anemia in general practice. Practitioner 209:218, 1972
43. Coigley MHF: Essential hypertension and its treatment by the family physician: A study of 61 cases. J R Coll Gen Pract 20:269, 1970
44. Faldauer A: A nine-year study of the aged in a general practice. Huisaits en Wetenschap 16:135, 1973
45. Cooper B, Fry J, Kalton G: A longitudinal study of psychiatric morbidity in a general practice population. Br J Prev Soc Med 23:210, 1969
46. Kuenssberg EV, Knox JDE: Teratogenic effect of anticonvulsants. Lancet 1:198, 1973
47. Morrell DC, Kasap HS: The effect of an appointment system on demand for medical care. Int J Epidemiol 1:143, 1972
48. Eaton M, Scherger J: Health team function: Testing a method for improvement. J Fam Pract 6:101, 1978
49. Eimerl TS, Laidlaw AJ (eds): A Handbook for Research in General Practice, ed 2. Edinburgh, E&S Livingstone, 1969
50. Hicks D: Primary Health Care: A Review. Commissioned by the Department of Health and Social Security. London, Her Majesty's Stationery Office, 1976
51. Hodgkin K: Toward Earlier Diagnosis: A Guide to General Practice, ed 3. Edinburgh, Churchill-Livingstone, 1973
52. Armitage P: Statistical Methods in Medical Research. Oxford, Blackwell, 1971
53. Kilpatrick SJ, McKay E: Umbilical cord urea in toxaemia of pregnancy. J Obstet Gynaecol Br Commonw 72:402, 1965
54. Lusted LE: Introduction to Medical Decision Making. Springfield, Ill, Charles C Thomas, 1968