

## 14. Data analysis and interpretation

*Concepts and techniques for managing, editing, analyzing and interpreting data from epidemiologic studies.*

### Key concepts/expectations

This chapter contains a great deal of material and goes beyond what you are expected to learn for this course (i.e., for examination questions). However, statistical issues pervade epidemiologic studies, and you may find some of the material that follows of use as you read the literature. So if you find that you are getting lost and begin to wonder what points you *are* expected to learn, please refer to the following list of concepts we expect you to know:

- Need to edit data before serious analysis and to catch errors as soon as possible.
- Options for data cleaning – range checks, consistency checks – and what these can (and can not) accomplish.
- What is meant by data coding and why is it carried out.
- Basic meaning of various terms used to characterize the mathematical attributes of different kinds of variables, i.e., nominal, dichotomous, categorical, ordinal, measurement, count, discrete, interval, ratio, continuous. Be able to recognize examples of different kinds of variables and advantages/disadvantages of treating them in different ways.
- What is meant by a “derived” variable and different types of derived variables.
- Objectives of statistical hypothesis tests (“significance” tests), the meaning of the outcomes from such tests, and how to interpret a p-value.
- What is a confidence interval and how it can be interpreted.
- Concepts of Type I error, Type II error, significance level, confidence level, statistical “power”, statistical precision, and the relationship among these concepts and sample size.

Computation of p-values, confidence intervals, power, or sample size will not be asked for on exams. Fisher’s exact test, asymptotic tests, z-tables, 1-sided vs. 2-sided tests, intracluster correlation, Bayesian versus frequentist approaches, meta-analysis, and interpretation of multiple significance tests are all purely for your edification and enjoyment, as far as EPID 168 is concerned, not for examinations. In general, I encourage a nondogmatic approach to statistics (*caveat*: I am not a “licensed” statistician!).

## Data analysis and interpretation

Epidemiologists often find data analysis the most enjoyable part of carrying out an epidemiologic study, since after all of the hard work and waiting they get the chance to find out the answers. If the data do not provide answers, that presents yet another opportunity for creativity! So analyzing the data and interpreting the results are the “reward” for the work of collecting the data.

Data do not, however, “speak for themselves”. They reveal what the analyst can detect. So when the new investigator, attempting to collect this reward, finds him/herself alone with the dataset and no idea how to proceed, the feeling may be one more of anxiety than of eager anticipation. As with most other aspects of a study, analysis and interpretation of the study should relate to the study objectives and research questions. One often-helpful strategy is to begin by imagining or even outlining the manuscript(s) to be written from the data.

The usual analysis approach is to begin with descriptive analyses, to explore and gain a “feel” for the data. The analyst then turns to address specific questions from the study aims or hypotheses, from findings and questions from studies reported in the literature, and from patterns suggested by the descriptive analyses. Before analysis begins in earnest, though, a considerable amount of preparatory work must usually be carried out.

### ***Analysis - major objectives***

1. Evaluate and enhance data quality
2. Describe the study population and its relationship to some presumed source (account for all in-scope potential subjects; compare the available study population with the target population)
3. Assess potential for bias (e.g., nonresponse, refusal, and attrition, comparison groups)
4. Estimate measures of frequency and extent (prevalence, incidence, means, medians)
5. Estimate measures of strength of association or effect
6. Assess the degree of uncertainty from random noise (“chance”)
7. Control and examine effects of other relevant factors
8. Seek further insight into the relationships observed or not observed
9. Evaluate impact or importance

### ***Preparatory work – Data editing***

In a well-executed study, the data collection plan, including procedures, instruments, and forms, is designed and pretested to maximize accuracy. All data collection activities are monitored to ensure adherence to the data collection protocol and to prompt actions to minimize and resolve missing

and questionable data. Monitoring procedures are instituted at the outset and maintained throughout the study, since the faster irregularities can be detected, the greater the likelihood that they can be resolved in a satisfactory manner and the sooner preventive measures can be instituted.

Nevertheless, there is often the need to “edit” data, both before and after they are computerized. The first step is “manual” or “visual editing”. Before forms are keyed (unless the data are entered into the computer at the time of collection, e.g., through CATI – computer-assisted telephone interviewing) the forms are reviewed to spot irregularities and problems that escaped notice or correction during monitoring.

Open-ended questions, if there are any, usually need to be coded. Codes for keying may also be needed for closed-end questions unless the response choices are “precoded” (i.e., have numbers or letters corresponding to each response choice). Even forms with only closed-end questions having precoded responses choices may require coding for such situations as unclear or ambiguous responses, multiple responses to a single item, written comments from the participant or data collector, and other situations that arise. (Coding will be discussed in greater detail below.) It is possible to detect data problems (e.g., inconsistent or out of range responses) at this stage, but these are often more systematically handled at or following the time of computerization. Visual editing also provides the opportunity to get a sense for how well the forms were filled out and how often certain types of problems have arisen.

Data forms will usually then be keyed, typically into a personal computer or computer terminal for which a programmer has designed data entry screens that match the layout of the questionnaire. For small questionnaires and data forms, however, data can be keyed directly into a spreadsheet or even a plain text file. A customized data entry program often checks each value as it is entered, in order to prevent illegal values from entering the dataset. This facility serves to reduce keying errors, but will also detect illegal responses on the form that slipped through the visual edits. Of course, there must be some procedure to handle these situations.

Since most epidemiologic studies collect large amounts of data, monitoring, visual editing, data entry, and subsequent data checks are typically carried out by multiple people, often with different levels of skill, experience, and authority, over an extended period and in multiple locations. The data processing procedures need to take these differences into account, so that when problems are detected or questions arise an efficient routing is available for their resolution and that analysis staff and/or investigators have ways of learning the information that is gained through the various steps of the editing process. Techniques such as “batching”, where forms and other materials are divided into sets (e.g., 50 forms), counted, possibly summed over one or two numeric fields, and tracked as a group, may be helpful to avoid loss of data forms. Quality control and security are always critical issues. Their achievement becomes increasingly complex as staff size and diversity of experience increase.

### ***Preparatory work – Data cleaning***

Once the data are computerized and verified (key-verified by double-keying or sight-verified) they are subjected to a series of computer checks to “clean” them.

### *Range checks*

Range checks compare each data item to the set of usual and permissible values for that variable. Range checks are used to:

1. Detect and correct invalid values
2. Note and investigate unusual values
3. Note outliers (even if correct their presence may have a bearing on which statistical methods to use)
4. Check reasonableness of distributions and also note their form, since that will also affect choice of statistical procedures

### *Consistency checks*

Consistency checks examine each pair (occasionally more) of related data items in relation to the set of usual and permissible values for the variables as a pair. For example, males should not have had a hysterectomy. College students are generally at least 18 years of age (though exceptions can occur, so this consistency check is “soft”, not “hard”). Consistency checks are used to:

1. Detect and correct impermissible combinations
2. Note and investigate unusual combinations
3. Check consistency of denominators and “missing” and “not applicable” values (i.e., verify that skip patterns have been followed)
4. Check reasonableness of joint distributions (e.g., in scatterplots)

In situations where there are a lot of inconsistent responses, the approach used to handle inconsistency can have a noticeable impact on estimates and can alter comparisons across groups. Authors should describe the decision rules used to deal with inconsistency and how the procedures affect the results (Bauer and Johnson, 2000).

## **Preparatory work – Data coding**

Data coding means translating information into values suitable for computer entry and statistical analysis. All types of data (e.g., medical records, questionnaires, laboratory tests) must be coded, though in some cases the coding has been worked out in advance. The objective is to create variables from information, with an eye towards their analysis. The following questions underlie coding decisions:

1. What information exists?
2. What information is relevant?
3. How is it likely to be analyzed?

## Examples of coding and editing decisions

- A typical criterion for HIV seropositivity is a repeatedly-positive ELISA (enzyme linked immunosorbent assay) for HIV antibody confirmed with a Western blot to identify the presence of particular proteins (e.g., p24, gp41, gp120/160). Thus, the data from the laboratory may include all of the following:
  - a. An overall assessment of HIV status (positive/negative/indeterminant)
  - b. Pairs of ELISA results expressed as:
    - i. ++ / +- / -- / indeterminate
    - ii. optical densities
  - c. Western Blot results (for persons with positive ELISA results) expressed as:
    - i. (+ / - / indeterminate)
    - ii. specific protein bands detected, e.g., p24, gp41, gp120/160

How much of this information should be coded and keyed?

- How to code open-ended questionnaire items (e.g., “In what ways have you changed your smoking behavior?”, “What are your reasons for quitting smoking?”, “What barriers to changing do you anticipate?”, “What did you do in your job?”)
- Closed-end questions may be “self-coding” (i.e., the code to be keyed is listed next to each response choice), but there can also be:
  - a. Multiple responses where only a single response is wanted – may be
    1. Inconsistent responses (e.g., “Never” and “2 times or more”)
    2. Adjacent responses indicating a range (e.g., “two or three times” and “four or five times”, by a respondent who could not choose among 2-5 times).
  - b. Skipped responses – should differentiate among
    1. Question was not applicable for this respondent (e.g., age at menarche for male respondents)
    2. Respondent declined to answer (which respondents sometimes may indicate as “N/A”!)
    3. Respondent did not know or could not remember
    4. Respondent skipped without apparent reason

It is necessary to achieve a balance between coding the minimum and coding “everything”.

- Coding is much easier when done all at once.
- One can always subsequently ignore coded distinctions not judged as meaningful.

- Information not coded will be unavailable for analysis (e.g., date questionnaire received, which questionnaires were randomly selected for 10% verification survey).
- More detail means more recodes for analysis means more programming means more opportunities for error.
- Decisions deferred have to be made sometime, so why not decide up front (e.g., When a respondent circles adjacent response choices, such as “3. Once or twice” and “4. Two to five times”, what should be coded – 3?, 4?, 3.5? a missing value code? a code to be replaced at a later date when a decision is made?)

It is important to document how coding was done and how issues were resolved, so that consistency can be achieved and the inevitable questions (“How did we deal with that situation?”) answered.

### **Types of variables - levels or scales of measurement**

Constructs or factors being studied are represented by “variables”. Variables (also sometimes called “factors”) have “values” or “levels”. Variables summarize and reduce data, attempting to represent the “essential” information.

#### **Analytic techniques depend upon variable types**

Variables can be classified in various ways. A **continuous variable** takes on all values within its permissible range, so that for any two allowable values there are other allowable values in between. A continuous variable (sometimes called a “measurement variable”) can be used in answer to the question “how much”. Measurements such as weight, height, and blood pressure can, in principle, be represented by continuous variables and are frequently treated as such in statistical analysis. In practice, of course, the instruments used to measure these and other phenomena and the precision with which values are recorded allow only a finite number of values, but these can be regarded as points on a continuum. Mathematically, a **discrete variable** can take only certain values between its maximum and minimum values, even if there is no limit to the number of such values (e.g., the set of all rational numbers is countable though unlimited in number). Discrete variables that can take any of a large number of values are often treated as if they were continuous. If the values of a variable can be placed in order, then whether the analyst elects to treat it as discrete and/or continuous depends on the variable’s distribution, the requirements of available analytic procedures, and the analyst’s judgment about interpretability.

#### **Types of discrete variables**

1. **Identification** – a variable that simply names each observation (e.g., a study identifying number) and which is not used in statistical analysis;
2. **Nominal** – a categorization or classification, with no inherent ordering; the values or the variable are completely arbitrary and could be replaced by any others without affecting the results (e.g., ABO blood group, clinic number, ethnicity). Nominal variables can be **dichotomous** (two categories, e.g., gender) or **polytomous** (more than two categories).

3. **Ordinal** – a classification in which values can be ordered or ranked; since the coded values need only reflect the ranking they can be replaced by any others with the same relative ranking (e.g., 1,2,5; 6,22,69; 3.5,4.2, 6.9 could all be used in place of 1,2,3). Examples are injury severity and socioeconomic status.
4. **Count** – the number of entities, events, or some other countable phenomenon, for which the question “how many” is relevant (e.g., parity, number of siblings); to substitute other numbers for the variable’s value would change its meaning. In epidemiologic data analysis, count variables are often treated as continuous, especially if the range is large.

### Types of continuous variables

1. **Interval** – differences (intervals) between values are meaningful, but ratios of values are not. That is, if the variable takes on the values 11-88, with a mean of 40, it is meaningful to state that subject A’s score of 60 is “twice as far from the mean” as subject B’s score of 50. But it is not meaningful to say that subject A’s score is “1.5 times the mean”. The reason is that the zero point for the scale is arbitrary, so values of the scores have meaning only in relation to each other. Without loss of information, the scale can be shifted: 11-88 could be translated into 0-77 by subtracting 11. Scale scores can also be multiplied by a constant. After either transformation, subject A’s score is still twice as far from the mean as is subject B’s, but subject A’s score is no longer 1.5 times the mean score. Psychological scales (e.g., anxiety, depression) often have this level of measurement. An example from physics is temperature measured on the Fahrenheit or Celsius scale.
2. **Ratio** – both differences and ratios are meaningful. There is a non-arbitrary zero point, so it is meaningful to characterize a value as “x times the mean value. Any transformation other than multiplying by a constant (e.g., a change of units) will distort the relationships of the values of a variable measured on the ratio scale. Physiological parameters such as blood pressure or cholesterol are ratio measures. Kelvin or absolute temperature is a ratio scale measure.

Many variables of importance in epidemiology are dichotomous (i.e., nominal with two levels) – case vs. noncase, exposed vs. unexposed. For an apparently ordinal or continuous variable, the phenomenon itself may not warrant treatment as such. It is necessary to ask such question as: “Is ‘more’ really more?” and “Are thresholds or discontinuities involved?” Again, the underlying reality (or, rather, our conceptual model of it) determines the approach to quantification. Variable values are often collapsed into a small number of categories for some analyses and used in their original form for others.

### Preparatory work – Data reduction

Data reduction seeks to reduce the number of variables for analysis by combining single variables into compound variables that better quantify the construct. Variables created during coding attempt to faithfully reflect the original data (e.g., height, weight). Often these variables can be used directly for analysis, but it is also often necessary to create additional variables to represent constructs of interest. For example, the construct overweight is often represented by a variable derived from the values for height and weight. Data reduction includes simplifying individual variables (e.g., collapsing

six possible values to a smaller number) and deriving compound variables (e.g. “socioeconomic status” derived from education and occupation).

In general:

- Simpler is better
- Avoid extraneous detail
- Create additional variables, rather than destroy the original ones (never overwrite the raw data!).
- Inspect detail before relying on summaries
- Verify accuracy of derived variables and recodes by examining crosstabulations between the original and derived variables.
- Take into account threshold effects, saturation phenomena, and other nonlinearities
- Categorize based on the nature of the phenomenon (e.g., a study of Down’s syndrome can collapse all age categories below 30 years; a study of pregnancy rates will require a finer breakdown below 30 years and even below 20 years).

### **Types of derived variables**

**Scales** - In a pure scale (e.g., depression, self-esteem) all of the items are intended as individual measures of the same construct. The scale score is usually the sum of the response values for the items, though items with negative valence (e.g., “I feel happy” in a depression scale) must be inverted. The purpose of deriving a scale score by having multiple items is to obtain a more reliable measure of the construct than is possible from a single item. Scale reliability (internal consistency) is typically assessed by using Cronbach’s *coefficient alpha*, which can be thought of as the average of all of the inter-item correlations. If the items did indeed measure the same construct in the same way and were indeed answered in an identical manner, then the only differences in their values should be due to random errors of measurement. Cronbach’s alpha gives the proportion of the total variation of the scale scores that is not attributable to random error. Values of 0.80 or greater are considered adequate for a scale that will be used to analyze associations (if the scale is used as a clinical instrument for individual patients, its alpha should be at least 0.90 – see Nunally’s textbook, *Psychometrics*). When the scale consists of separate subscales, internal consistency may be more relevant for the individual subscales than for the scale as a whole. Analyses of relationships between individual items (inter-item correlation or agreement), between each item and the remaining items (item-remainder correlation), between each item and the total scale (item-scale correlation), and among groups of items (factor analysis) are standard methods of analyzing item performance.

**Indexes** - An index consists of a group of items that are combined (usually summed) to give a measure of a multidimensional construct. Here, each of the items measures a different aspect or dimension, so that internal consistency measures like Cronbach’s alpha are either not relevant or require a different interpretation. Examples of indexes derived from several variables include

socioeconomic status (e.g., occupation, income, education, neighborhood), social support (e.g., marital status, number of close family members, number of close friends), sexual risk behavior (number of partners, types of partners, use of condoms, anal intercourse). Items may have different weights, depending upon their relative importance and the scale on which they were measured.

**Algorithms** - A procedure that uses a set of criteria according to specific rules or considerations, e.g., major depressive disorder, “effective” contraception (I have not seen this term used to designate a type of variable before, but I am not aware of any other term for this concept).

### **Preparatory work – Exploring the data**

Try to get a “feel” for the data – inspect the distribution of each variable. Examine bivariate scatterplots and cross classifications. Do the patterns make sense? Are they believable?

- Observe shape – symmetry vs. skewness, discontinuities
- Select summary statistics appropriate to the distribution and variable type (nominal, ordinal, measurement)

*Location* - mean, median, percentage above a cut-point

*Dispersion* - standard deviation, quantiles

- Look for relationships in data
- Look within important subgroups
- Note proportion of missing values

### **Preparatory work – Missing values**

Missing data are a nuisance and can be a problem. For one, missing responses mean that the denominators for many analyses differ, which can be confusing and tiresome to explain. Also, analyses that involve multiple variables (e.g., coefficient alpha, crosstabulations, regression models) generally exclude an entire observation if it is missing a value for any variable in the analysis (this method is called *listwise deletion*). Thus, an analysis involving 10 variables, even if each has only 5% missing values, could result in excluding as much as 50% of the dataset (if there is no overlap among the missing responses)! Moreover, unless data are *missing completely at random (MCAR)* – equivalent to a pattern of missing data that would result from deleting data values throughout the dataset without any pattern or predilection whatever), then an analysis that makes no adjustment for the missing data will be biased, because certain subgroups will be underrepresented in the available data (a form of selection bias).

#### ***Imputation for missing values - optional topic***

As theory, methods, and computing power have developed over the years, analytic methods for handling missing data to minimize their detrimental effects have improved. These

methods seek to *impute* values for the missing item responses in ways that attempt to increase statistical efficiency (by avoiding the loss of observations which have one or a few missing values) and reduce bias. Earlier methods of imputation, now out of favor, include replacing each missing value by the mean or median for that variable. Even though such practices enable all observations to be used in regression analyses, these methods do not reduce bias and tend to introduce additional distortion. More sophisticated methods reduce bias from missing data while minimizing distortion from imputation. These methods derive imputations that make use of the values of variables for which data are present and which are related to the variable being imputed.

Typically, *complete data cases* (observations that have no missing values for the variables of interest) serve as the raw material for the imputations. Factors that are theoretically related to the variables to be imputed and with which they are associated in the complete data cases are used to develop “predictive” models for the imputed variables. These models are then applied to the remaining observations, providing predicted (“imputed”) values for their missing responses. The resulting imputations are said to be *conditioned on* the variables in the model.

For example, suppose the available data show a positive correlation between blood pressure and age. By conditioning imputations on age, we impute (on average) higher blood pressures to older subjects with missing blood pressure data and lower blood pressures to younger subjects missing blood pressure data. This technique preserves the relationship between age and blood pressure that exists in the complete data cases. Moreover, if older subjects are more likely to be missing a blood pressure reading, then the conditioning reduces the bias from analyzing only the complete data cases.

If the process that led to the missing data is uniformly random except for being positively related to identifiable factors (e.g., subject’s age), then the missing data process is called *missing at random (MAR)*, rather than MCAR. In such a situation, the overall mean blood pressure for the complete data cases is biased downwards (due to underrepresentation of older subjects), but the overall mean based on imputations conditioned on age is not.

If predicted values are simply substituted for missing values, however, then although bias will be reduced so will standard errors. The reason is that the imputation models were created based on (imperfect) associations between the conditioning variables and the variables being imputed. In contrast, the predicted values are directly computed from the model as if, in our example, blood pressure were completely determined by age. In effect, the model functions as a “self-fulfilling prophecy”. To avoid this problem a source of random variability is introduced into the imputation process. For example, rather than substituting the predicted values themselves for the missing data, the imputed values may be sampled from distributions whose means are the predicted values (e.g., if the estimated mean for a yes-no response were 0.30 [where 1=“yes” and 0=“no”], then the imputed value would be generated randomly from a binomial distribution with a proportion of “successes” of 0.30).

In addition, by using multiple imputations (typically five), the analyst can adjust the standard errors to reflect the uncertainty introduced by the imputation process. Carrying out multiple imputations means repeating the imputation process to create multiple versions of the dataset (one for each imputation), analyzing each dataset separately, and combining the results according to certain procedures.

Imputation causes the least distortion when the proportion of missing data is small, and data are available for variables that are strongly associated with the variable being imputed. Perversely, however, imputation is most needed when the proportion of missing data is large. Also, unfortunately, the available data may provide little guidance about whether the missing process is MCAR, MAR, or “nonignorable”. Attention to causes of missing responses during data collection can be helpful (Heitjan, 1997).

[I would like to thank Michael Berbaum and Ralph Folsom for their patient explanations of imputation and for reading over earlier versions of this section.]

## ***Descriptive analyses***

Exploration of the data at some point becomes descriptive analysis, to examine and then to report measures of frequency (incidence, prevalence) and extent (means, survival time), association (differences and ratios), and impact (attributable fraction, preventive fraction). These measures will be computed for important subgroups and probably for the entire study population. Standardization or other adjustment procedures may be needed to take account of differences in age and other risk factor distributions, follow-up time, etc.

## ***Evaluation of hypotheses***

After the descriptive analyses comes evaluation of the study hypotheses, if the study has identified any. Here there will be a more formal evaluation of potential confounding, other forms of bias, potential alternative explanations for what has been observed. One aspect of both descriptive analysis and hypothesis testing, especially of the latter, is the assessment of the likely influence of random variability (“chance”) on the data. Much of the field of statistics has grown up to deal with this aspect, to which we will now turn.

## ***Evaluating the role of chance - inference***

Whether or not we believe, in Albert Einstein’s words, that “the Lord God doesn’t play dice with the universe”, there are many events in the world that we ascribe to “chance”. When we roll a die, the resulting number is generally unpredictable and does not (or at least, should not) follow any evident pattern. Similarly, when we draw five cards from a freshly-shuffled, unmarked deck, we know that some outcomes are more or less likely than others (e.g., a pair is more likely than three of a kind), but we cannot predict what cards we will draw. The theories of probability and statistics were born in the gaming parlors of Monte Carlo and came of age in the fields of the British countryside. The computer revolution put their power, for good or for whatever, into the hands of any of us who can click a mouse.

The basis for the incorporation of the fruits of the theory of probability and statistics into medical and epidemiologic research has been recounted by Austin Bradford Hill as follows:

“Between the two world wars there was a strong case for emphasizing to the clinician and other research workers the importance of not overlooking the effects of the play of chance upon their data. Perhaps too often generalities were based upon two men and a laboratory dog while the treatment of choice was deduced from a difference between two bedfuls of patients and might easily have no true meaning. It was therefore a useful corrective for statisticians to stress, and to teach the need for, tests of significance merely to serve as guides to caution before drawing a conclusion, before inflating the particular to the general.” (pg 299 in *The environment and disease: association or causation. Proceedings of the Royal Society of Medicine*, 1965: 295-300)

From this innocent and commonsensical beginning, statistical procedures have (like kudzu?? – just kidding!) virtually engulfed the thinking of researchers in many fields. Hill continues:

“I wonder whether the pendulum has not swung too far – not only with the attentive pupils but even with the statisticians themselves. To decline to draw conclusions without standard errors can surely be just as silly? Fortunately I believe we have not yet gone so far as our friends in the USA where, I am told, some editors of journals will return an article because tests of significance have not been applied. Yet there are innumerable situations in which they are totally unnecessary - because the difference is grotesquely obvious, because it is negligible, or because, whether it be formally significant or not, it is too small to be of any practical importance. What is worse the glitter of the t table diverts attention from the inadequacies of the fare. . . .”

He admits that he exaggerates, but he suspects that the over-reliance on statistical tests weakens “our capacity to interpret data and to take reasonable decisions whatever the value of P.” Hill is referring to tests of significance, which are probably the most common procedures for assessing the role of chance, or perhaps more precisely, the amount of numerical evidence that observed differences would not readily arise by chance alone.

### ***Illustration of a statistical test***

Consider the following data, from the first study to report an association between adenocarcinoma of the vagina and maternal use of diethylstilbestrol (DES). During the 1960’s, a handful of cases of adenocarcinoma of the vagina were observed in young women, a highly unusual occurrence. Investigation into the histories of the affected women revealed that in most cases the girl’s mother had taken diethylstilbestrol (DES) while she was carrying the girl in her uterus. At that time DES had been prescribed in the belief that it might prevent premature delivery in women who had lost pregnancies. In how many patients would this history have to emerge for it before the investigators could be confident that it was not a chance observation? This question is usually answered by means of a statistical test.

**Prenatal exposure to diethylstilbestrol (DES)  
among young women with adenocarcinoma of the vagina**

	Exposed to diethylstilbestrol?		
	Yes	No	Total
Cases	7	1	8
Controls	0	32	32
Total	8	33	40

Source: Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *New Engl J Med* 1971; 284:878-881. [From Schlesselman JJ. *Case-Control Studies*. New York, Oxford, 1982: 54]

All but one of the cases had a positive history for intrauterine exposure to DES. In contrast, not one of 32 controls did. The relative risk from this table cannot be calculated directly, because of the zero cell, but adding 0.5 to all four cells yields a relative risk (OR) of 325, a stronger association than most of us will ever encounter in our data. However, this study has only eight cases. Could these results be due to chance?

A statistical test of significance is a device for evaluating the amount of numerical data on which an observed pattern is based, to answer a question like, “How often could such a strong association arise completely by chance in an infinite number of analogous experiments with the same number of subjects and the same proportion of cases (or of exposed)?” This question is not quite the same as “How likely is it that chance produced the association in the table?” nor as “How much of the association is due to chance?”. But if such a strong association would arise only very infrequently by chance alone, then it is reasonable to suppose that at least some potentially identifiable factor has contributed to the observed association. That factor could be bias, of course, rather than the exposure, but at least it would be something other than chance. Conversely, it is also possible that much stronger associations could readily arise by chance and yet the one we observed might reflect a causal process. The significance test simply evaluates the strength of numerical evidence for discounting chance as a likely sufficient explanation.

In order to conduct a test of significance, we need to operationalize the concept of “analogous experiment”. There’s the rub. What kind of experiment is analogous to an epidemiologic study, all the more so an observational one? For the above table, the significance test that would be used is Fisher’s Exact Test. The analogous experiment (*probability model*) here is equivalent to the following:

Suppose that you have 40 pairs of socks – 7 pairs of red socks, and 33 pairs of blue socks. You want to pack 8 pairs of socks in your travel bag, so without looking you take 8 pairs at random and put them in your bag. How many red pairs have you packed for your trip?

When this “analogous experiment” is repeated a sufficient number of times, the proportion of trials in which the bag has 7 red pairs will provide the probability that chance alone would produce a situation in which you had packed 7 pairs of red socks. This probability is the “p-value” for the significance test of the relationship between adenocarcinoma of the vagina and maternal DES in the above table.

Fortunately, the distribution of the number of red pairs in the bag has already been worked out theoretically, so that the exact probability can be computed without having to carry out what in this case would be a VERY large number of trials. The formula for the (hypergeometric) distribution is:

$$\Pr(A=j) = \frac{\binom{n_1}{j} \binom{n_0}{(m_1-j)}}{\binom{n}{m_1}} = \frac{n_1!n_0!m_1!m_0!}{n!j!(n_1-j)!(m_1-j)!(n_0-m_1-j)!}$$

where  $\Pr(A=j)$  is the probability of obtaining  $j$  red pairs of socks in the travel bag and  $m_0, m_1, n_0, n_1,$  and  $n$  are the row and column totals in the table:

	Color		
	Red	Blue	Total
Travel bag	$j$	$m_1 - j$	$m_1$
In drawer	$n_1 - j$	$n_0 - m_1 - j$	$m_0$
Total	$n_1$	$n_0$	$n$

Here is how the formula is applied:

	Red	Blue	Total
	(DES)		
Packed (cases)	7	1	8
In drawer (controls)	0	32	32
Total	7	33	40

Possible outcomes (Colors of pairs of socks in travel case)		Probability of each outcome
Red	Blue	
0	8	.181
1	7	.389
2	6	.302
3	5	.108
4	4	.019
5	3	.0015
6	2	.00005
7	1	$4.3 \times 10^{-7}$
8	0	0
		1.0000

}

7! 33! 8! 32!

40! 5! 2! 3! 30!

}

p-value

### Comments on the “red socks” model:

1. A model is a system or structure intended to represent the essential features of the structure or system that is the object of study. The above model is a very simplified representation!
2. The model is derived on the basis of certain constraints or assumptions (e.g., in this case, 8 cases, 7 DES-exposed mothers, and 40 subjects in all – “fixed marginals” – as well as “all permutations are equally likely”).
3. The model underlying hypothesis testing assumes a repeatable experiment and an *a priori* specification of the “hypothesis” being tested – a “null” hypothesis [this is embodied in the model of “equally likely” permutations] and an “alternative hypothesis” [this deals with what results we would regard as inconsistent with the null hypothesis].
4. The above model is tedious to compute for large tables, though computers have solved that problem.

### Concept of hypothesis testing (tests of significance)

What we really want to know is: “Is the observed association due to chance?”, or “How likely is it that the observed association is due to chance?”. This probability is sometimes referred to as the “posterior [*a posteriori*] probability”, the probability that the hypothesis is true given the observed results. (The “prior [*a priori*] probability” that the hypothesis is true is our belief in the hypothesis before we have the results in question). The frequentist school of statistics, from which significance testing derives, cannot answer this question directly. Instead, significance tests and p-values attempt to provide an indirect answer, by reformulating the question as: “How often would an association as strong as that observed occur by chance alone?”. The role of chance is played by a suitable probability model, chosen to represent the probability structure of the data and the study design. But most epidemiologic studies deviate rather markedly from the probability models on which statistical

tests are based (e.g., see Sander Greenland, Randomization, statistics, and causal inference), so although statistical theory is extremely precise, it must be thoughtfully applied and thoughtfully interpreted.

A compromise version of the question that underlies a significance test is “How consistent are the numerical data with what would be expected ‘by chance’ – as played by a suitable probability model”. The probability model is most often one that assumes no systematic difference between groups, partly because such models are easier to derive and also because it is often convenient for the hypothesis-testing framework. The result of the significance test is a probability (the *p-value*), which provides a quantitative answer to this compromise question. (Note: The statistical “null hypothesis” is rarely of interest from a substantive perspective. A study hypothesis should be stated in terms of no association only if that is indeed what the investigator hopes to demonstrate. In fact, it is quite difficult to demonstrate the absence of association, since the evidence for no association is related to the Type II error probability (1 – statistical power) for the study, which is generally considerably greater than the significance level – see below).

The p-value itself can be regarded as a descriptive statistic, a piece of evidence that bears on the amount of numerical evidence for the association under study. When a decision is called for, though, then some method of assigning an action to the result of the significance test is needed. Decision-making entails the risk of making errors. Ideally the loss function (the costs of errors of various types) is known explicitly. Under broadly applicable assumptions, though, the theory of decision-making provides a technique for decision-making based on the results of a statistical test. That technique is statistical hypothesis testing.

As noted, the hypothesis to be tested is generally a “null hypothesis” (usually designated  $H_0$ ).  $H_0$  is the probability model that will play the role of chance (for example, the red socks model). In the present context, that model will be based on the premise that there is no association. If there is sufficient numerical evidence to lead us to reject  $H_0$ , then we will decide that the converse is true, that there is an association. The converse is designated as the “alternate hypothesis” ( $H_A$ ). The decision-making rule is to reject  $H_0$ , in favor of  $H_A$ , if the p-value is sufficiently small, and to otherwise accept  $H_0$ .

Since we have a decision between two alternatives ( $H_0$  and  $H_A$ ) we can make two kinds of errors:

**Type I error.** Erroneously reject  $H_0$  (i.e., conclude, incorrectly, that data are not consistent with the model)

**Type II error.** Erroneously fail to reject  $H_0$  (i.e., conclude, incorrectly, that data are consistent with the model)

(The originator of these terms must have been more prosaic than the originators of the terms “significance”, “power”, “precision”, and “efficiency”) Traditionally, the Type I error probability has received more attention and is referred to as the “*significance level*” of the test.

In a strict decision-making mode, the result of the significance test is “Reject null hypothesis” or “Fail to reject null hypothesis”. (Note that “fail to reject the null hypothesis” is not equivalent to declaring that the null hypothesis is true.) However, rarely must a decision be made based on a single study, so it is preferable to report the calculated p-value (probability that the assumed probability model would produce data as extreme or more so). The p-value gives more information than the statement “results were significant at the 5% level”, since it quantifies the degree to which the data are incompatible with “chance” (as played by the probability model), allowing the reader to apply his/her tolerance for a Type I error. Note that the p-value is not a direct index of the strength of an association in an epidemiologic sense nor of its biologic, clinical, or epidemiologic “significance”. The p-value simply assesses the compatibility of the observed data with the assumed probability model that serves to represent  $H_0$ .

There are many methods for obtaining a p-value or conducting a test of statistical significance. The choice depends upon the level of measurement of the variables (dichotomous, nominal polytomous, ordinal, continuous), the sampling design from which the data came, and other factors. The statistical test illustrated above is an “exact” test (Fisher’s exact test), since it is based on a model that considers all possible outcomes and in how many ways each can occur. In an exact test, the probability model is readily apparent.

### **Illustration of an asymptotic test**

More commonly-used, because they are much simpler to compute, are *asymptotic tests* (e.g., chi-square, t-test). Asymptotic tests are approximations whose accuracy improves as the sample size increases, and the underlying probability model on which they are based tends to be more abstract. Typically, asymptotic tests are based on the “normal” (Gaussian) distribution. Why the Gaussian distribution? Because it offers a number of analytic advantages and, most especially, because of the Central Limit Theorem (“one of the most remarkable theorems in the whole of mathematics”, Mood and Graybill, 1963:149). The Central Limit Theorem holds that if we take large enough random samples from any distribution with a finite variance, the means of those samples will have an approximately Gaussian distribution.

The general form for such a test is (see Rothman, *Modern epidemiology*, p. 139 or Kleinbaum, Kupper, and Morgenstern, *Epidemiologic research*):

$$Z = \frac{a - E(a)}{\sqrt{\text{var}(a)}}$$

where “a” is the observed value (e.g., the number of exposed cases),  $E(a)$  is the expected value for “a” under the null hypothesis (a.k.a. analogous experiment), and  $\text{var}(a)$  is the variance of “a” under the null hypothesis. Thus, Z is the number of standard deviations by which “a” differs from what would be expected if there were no association and has an approximate unit normal distribution. (Z is occasionally written as  $\chi$  (called “chi” [pronounced “KAI”], a unit normal distribution is the same as the square root of a one-degree-of-freedom chi-square distribution).

The probability associated with being “Z” standard deviations away from the mean of a normal distribution can be computed and is readily available in statistical tables (see table excerpt below). The value of a normally-distributed random variable is usually (i.e., probability 95%) less than two standard deviations from its mean, so if Z exceeds 1.96 we say “ $p < .05$ ”, or more precisely, we take the value we have calculated for Z, look it up in a table of the normal distribution and read off the corresponding p-value.

The table excerpt below shows various probabilities derived from the unit normal distribution. For example, the probability associated with a distance of 1.645 standard deviations above the mean is shown in column B (0.05) and is identical to the probability associated with a distance of 1.645 standard deviations below the mean (since the normal distribution is symmetrical). The probability associated with obtaining a value of z that is either above or below a distance of 1.645 standard deviations from the mean is shown in column D (0.10). So if using the formula above (or one of those below) we obtain a value of Z equal to 1.645, then the p-value is either 0.05 or 0.10, depending upon the alternative hypothesis.

**Excerpt from a table of the Normal Distribution**

z	H	A	B	C	D	E
0.00	0.3989	0.0000	0.5000	0.0000	1.0000	0.5000
0.01	0.3989	0.0040	0.4960	0.0080	0.9920	0.5040
0.02	0.3989	0.0080	0.4920	0.0160	0.9840	0.5080
...	...	...	...	...	...	...
0.8416	0.2800	0.30	0.20	0.60	0.40	0.80
...	...	...	...	...	...	...
1.282	0.1755	0.40	0.10	0.80	0.20	0.90
...	...	...	...	...	...	...
1.645	0.1031	0.45	0.05	0.90	0.10	0.95
...	...	...	...	...	...	...
1.960	0.0585	0.475	0.025	0.95	0.05	0.975
...	...	...	...	...	...	...
2.576	0.0145	0.495	0.005	0.99	0.01	0.995
...	...	...	...	...	...	...
3.090	0.0034	0.499	0.001	0.998	0.002	0.999
...	...	...	...	...	...	...

Legend:

z = number of standard deviations to the right of the mean

h = height of the normal curve for that number of standard deviations from the mean

A = area between the mean and z

B = area to the right of z (or to the left of -z)

C = area between  $-z$  and  $+z$

D = area beyond  $|z|$  (i.e., to the left of  $-z$  and the right of  $+z$ )

E = area to the left of  $z$

(Source: National Bureau of Standards – Applied Mathematics Series–23, U.S. Government Printing Office, Washington, D.C., 1953, as abstracted in Table A-4 in Richard D. Remington and M. Anthony Schork, *Statistics with applications to the biological and health sciences*. Englewood Cliffs, NY, 1970.)

### **One-sided versus two-sided p-values**

Recall that the p-value is the probability of obtaining an association as strong as (or stronger than) the association that was observed. It turns out, though, that the phrase “as strong as (or stronger than)” is ambiguous, because it does not specify whether or not it is intended to include inverse associations, i.e., associations in the opposite direction from the putative association that motivated the study. For example, if we observe a 2.5 relative risk, does “as strong” mean only relative risks of 2.5 or larger, or does it also mean relative risks of 0.4 or smaller? If the former (only 2.5 and larger), then the corresponding p-value is “one-sided” (or “one-tailed”). In contrast, if  $H_A$  is “either greater than or equal to 2.5 or [inclusive] less than or equal to 0.4”, then a two-sided p-value is indicated. [Only one-sided p-values can be interpreted as the “probability of observing an association as strong or stronger under the chance model” (Rothman and Greenland,185).]

The issue of one-sided versus two-sided p-values can arouse strong emotions. For a given calculated value of  $Z$ , the one-sided p-value is exactly half of the two-sided p-value. Proponents of two-sided p-values argue that a one-sided p-value provides an inflated measure of the statistical significance (low probability of obtaining results by chance) of an association. Appropriate situations for using one-sided p-values are sometimes characterized as ones where the investigator has no interest in finding an association in the opposite direction and would ignore it even it occurred. However, a posting on the EPIDEMIOLOG-L listserv asking for situations of this sort produced very few persuasive examples.

Here is a dramatical presentation of some of the issues in choosing between 1-sided and 2-sided p-values:

The wife of a good friend of yours has tragically died from lung cancer. Although she was a life-long nonsmoker, your friend used to smoke quite heavily. Before her death she had become an anti-smoking activist, and her last wishes were that your friend bring suit against R.J. Morris, Inc., the manufacturer of the cigarette brand your friend used to smoke. Knowing that he cannot afford expert consultation, your friend turns to you and prevails upon you to assist him with the lawsuit.

In preparation for the trial, the judge reviews with both sides the standard of evidence for this civil proceeding. She explains that for the court to find for the plaintiff (your side) it must

conclude that the association is supported by “a preponderance of the evidence”, which she characterizes as “equivalent to a 90% probability that R.J. Morris’ cigarettes caused the disease”. The R.J. Morris attorney objects, declaring that, first of all, only the probability that cigarettes can cause disease can be estimated, not the probability that cigarettes did cause the disease. As the judge is about to say that the judicial interpretation of probability permits such a conclusion, the R.J. Morris attorney raises her second objection: since the plaintiff is basing his case on scientific evidence, the plaintiff’s case should be held to the conventional standard of evidence in science, which requires a significance level of 5%. [Recall that the significance level is the probability of a Type I error, which in this case would mean finding the company responsible even though your friend’s lung cancer was really due to chance. If the court were to fail to find the tobacco company responsible, even though the company’s cigarettes did cause the cancer, that would be or a Type II error.]

Seeing an opportunity, you pass a note to your friend, who passes it on to his attorney. Upon reading it, his attorney says to the judge “Your Honor, my client is prepared to accept the R.J. Morris’ insistence on a 5% significance level, provided that it is based on a one-sided alternative hypothesis.” Beginning to regret that she introduced the probability metaphor, the judge turns to the R.J. Morris attorney, who is now hastily conferring with her biostatistician. After a quick consultation, the R.J. Morris attorney charges indignantly that plaintiff’s attorney is trying, through deception, to obtain a lower standard of evidence. A 5% one-tailed significance level, she charges, is actually a 10% significance level, since everyone knows that two-tailed tests are more appropriate. Your friend’s attorney senses that this charge will be a telling point with the judge and anxiously looks back to you for advice on how to respond.

With your coaching, your friend’s attorney replies that a two-tailed test is warranted only when the appropriate alternative hypothesis ( $H_A$ ) is two-sided. The question in this case is whether R.J. Morris is or is not liable, i.e., whether their cigarettes did or did not cause the cancer. This question corresponds to a one-sided  $H_A$ , i.e., the court can (1) reject  $H_0$  (no causation) in favor of the alternative that R.J. Morris is liable or (2) fail to reject  $H_0$ , if the court finds the evidence insufficient. “May it please the court,” she continues, “there is no issue here that the cigarette smoke could have acted to prevent the cancer from occurring, so requiring a two-tailed alternative hypothesis is tantamount to imposing a significance level of 2.5%, which is closer to the standard for a criminal, rather than a civil, trial”.

With the benefit of further consultation, the R.J. Morris attorney “strenuously objects”. “Plaintiff may see this case as involving a one-sided  $H_A$ , but notwithstanding the tobacco settlement, as far as the R.J. Morris Company is concerned the relationship between smoking and cancer has not been proved. Therefore a finding that cigarette smoking can in fact prevent cancer is just as relevant as plaintiff’s contention that the cigarettes were responsible.”

You are naturally outraged by the R.J. Morris lawyer’s assertion that the relationship between smoking and cancer is not proved, but you have to put that aside as your friend’s lawyer asks you is it not correct that the significance level is simply a mechanism for deciding how many standard deviations away from the mean are required to exclude chance as an explanation. Usually, people exclude chance when the statistical test comes out two standard deviations from the center of a normal distribution (actually, 1.96 standard deviations, which corresponds to a

two-tailed 5% significance level). If the judge does accept the one-tailed 5% significance level, even with a good argument that because the appropriate  $H_A$  is one-sided so that the Type I error probability is really only 5%, a decision that meets the test of being only 1.65 standard deviations from the mean (corresponding to a one-tailed 5% significance level) may be vulnerable on appeal. Since the scientific evidence is firm, would it be better to agree to the two-tailed test?

The judge looks at her watch, and you see beads of perspiration breaking out on your friend's attorney's forehead. Meanwhile you're trying to sort through the issues. You've only just received your epidemiology degree, and you aren't yet sure that it works. It's true that an appeals court might reject the idea of a one-tailed test, since appellate judges tend to be conservative, and R.J. Morris will certainly appeal an adverse judgment. But then a dark thought jumps into your mind. What if R.J. Morris has concocted evidence that will somehow make it appear that your friend is responsible for his wife's death from lung cancer? You know that that is outlandish, but what if they could? With a two-sided  $H_A$ , the court could reject  $H_0$  and find your friend responsible, thereby destroying him financially and emotionally. "One-sided!", you cry out ... and then suddenly you wake with a start. The professor and your fellow students are looking at you with puzzlement, wondering what question you thought that you were responding to. As you emerge from your daydream you hope that you have not slept through too much of the lesson and vow to go to bed earlier in the future.

### **Significance testing in a two-by-two table**

For a two-by-two table, the formula can be more easily expressed for computational purposes by defining "a" as the contents of a single cell in the table, conventionally the "a" (upper left corner) cell, so that  $E(a)$  is the value expected for "a" under the null hypothesis ( $n_1 m_1 / n$ ), and  $\text{Var}(a)$  is the variance of "a" under the null hypothesis  $\{(n_1 n_0 m_1 m_0) / [n^2 (n-1)]\}$ , based on the hypergeometric distribution. The test statistic is then simply:

$$Z = \frac{a - n_1 m_1 / n}{\sqrt{\{(n_1 n_0 m_1 m_0) / [n^2 (n-1)]\}}}$$

An equivalent, but more easily remembered computation formula, is:

$$Z = \sqrt{X^2} = \sqrt{\frac{(ad - bc)^2 (n-1)}{n_1 n_0 m_1 m_0}}$$

[Note: you may also see the above formula with  $n$ , instead of  $(n-1)$  [e.g., Hennekens and Buring, p. 251 uses  $T$  instead of  $(n-1)$ ]. The reason is that the above formula gives a Mantel-Haenszel chi-square statistic (based on the hypergeometric distribution) instead of the Pearson chi-square statistic (based on the normal distribution). For large samples the two are essentially equivalent. There are parallel formulas for person-time data.]

	Exposed to diethylstilbesterol?		
	Yes	No	Total
Cases	a	b	m <sub>1</sub>
Controls	c	d	m <sub>0</sub>
Total	n <sub>1</sub>	n <sub>0</sub>	n

Whatever misgivings we may have about the statistical model and its application, results with as small a p-value as that obtained in this study will be very satisfying to practically any investigator who obtains them. But to appreciate the dynamics of this procedure, and the problems of interpretation that arise in more equivocal circumstances, let us analyze what underlies a small p-value.

A small p-value (i.e., low probability that results similar to those observed would be produced by “chance” [as played by a given statistical model]) reflects:

- a strong observed association (or a large observed difference)

or

- a large sample size (roughly speaking).

Therefore, when the p-value is not small, there are two possibilities (ignoring the possibilities of systematic error, inappropriate statistical model, etc.):

1. the observed association or difference is not strong.
2. the observed association is of a respectable size but the study size was too small to judge it “significant.”

How we interpret a failure to obtain a low p-value depends upon our judgment of the magnitude of the observed association and of the statistical power of the study to detect an important real difference.

If the p-value is small (e.g., less than five percent [typical], ten percent [less common], or one percent [for the demanding or rich in data]), the observed results are somewhat inconsistent with an explanation based on chance alone, so we are inclined to view them as having some origin worth inquiring about (e.g., systematic influences from the way the study was designed or conducted, biological or psychosocial processes related to the factors under study, etc.). If the observed difference or association is too small to be scientifically or clinically significant (as opposed to statistically significant), we will not care to pursue the matter regardless of the p-value.

If the p-value is not small (i.e., the results are “not significant”), was an association observed? If no association was observed, then the appropriate characterization of the finding is “no association was observed” (but see below). If an association was observed, then we can say “an association was observed but the data were insufficient to discount chance as an explanation” [not “there was no association”!].

If no association was observed, then we also need to ask how precise is our estimate of the association. (Note: the earlier version of this chapter suggested that computing the statistical power from the study results – “*post-hoc* power” could provide guidance about how much evidence the study might have provided against the existence of a strong association. However, Sander Greenland (2012) explains that nonsignificance plus high statistical power does not imply support for the null hypothesis over the alternative. Instead, the confidence interval (see below) provides the best estimate of the association and how precisely this study has been able to estimate it.

If the observed association is strong enough to be important if it is not due to chance, then the only conclusion we can draw is that the data do not provide sufficient evidence to discount an explanation of chance alone – this is not equivalent to a conclusion that “an association was not observed” [since one was] or that “the observed association is due to chance” [which no one knows]. Other characterizations often stated are also unfortunate:

“The observed association is not significant” [which tends to impugn it]

“The association did not reach statistical significance” [which implies that the association should have been stronger – it may be as strong as it should be but be based on too few subjects.]

Better to say “an association of \_\_\_\_ was observed, but the data were too few to discount an explanation based on chance” or some similar expression. [Note: Any result can become “nonsignificant” if we stratify enough.]

An alternate possibility is that the observed association was too weak to be meaningful even if it had been associated with a small p-value. If the confidence interval is narrow, then our estimate suggests that if an association does exist, it is not strong. If the confidence interval is wide, however, then there really is not much we can say or conclude, except that our failure to find an association could well be due to chance (i.e., we may well have made a “Type II error”). This inability is one of the reasons for discouraging researchers from conducting small studies except as a pilot study to develop procedures and instruments. A more precise estimate puts us in a better position to interpret our results as evidence against the existence of a real association. But increasingly a single study becomes just a data point for a meta-analysis that combines the estimates across multiple studies to obtain the best overall estimate from existing studies.

## ***Statistical power and sample size***

Statistical power refers to the ability to detect an association of interest in the face of sampling error. Suppose that there is a true association of a certain type and degree, but that through the workings of chance our studies will observe the association to be weaker or stronger. In order to be

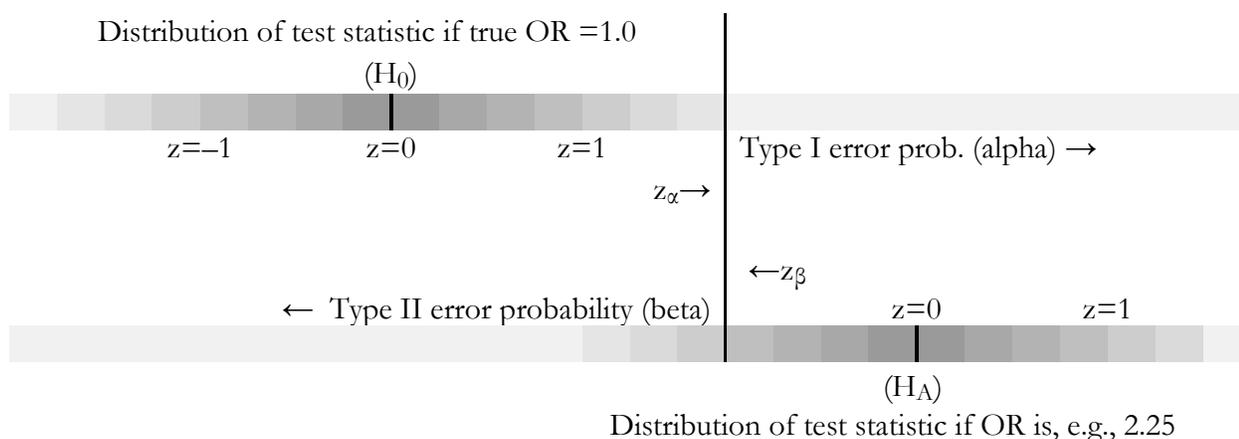
reasonably certain that our study will detect the association, the study has to be large enough so that sampling error can be contained.

For example, suppose that we are comparing a group of Alzheimer’s disease cases to a group of controls to see whether the cases are different in respect to presence of a specific gene. Suppose that this gene is actually present in 20% of cases and in 10% of the population from which the cases arose (i.e., the OR in a large, unbiased case-control study would be 2.25). If we study 20 cases and 20 controls, we may well find 4 cases with the gene and 2 controls with the gene, so that we correctly estimate the prevalence of the gene in cases and in the population and the OR.

With such few subjects, however, we could very easily get only 3 cases with the gene and 3 controls with the gene, completely failing to detect the difference in prevalence (OR = 1.0). In fact, we might even get 4 controls with the gene and only 2 cases with the gene, so that the gene appear to be protective (OR = 0.44). Of course, we would not want to react to a difference or an OR that may readily be due to chance, so we will test whatever result we observe to make sure that it is greater than is expected to occur by chance alone (i.e., ”significant”). That means that we will discount any association that we regard as within chance expectation. (Or recalling our courtroom fantasy, a “preponderance of the evidence”, not merely suspicion.)

Therefore, in order to detect an association, we must both (1) observe it in our study and (2) decide that chance would not likely have created it. Each of these requirements places a demand on the size of the study. We need at least some minimum number of subjects so that (1) we have a reasonable expectation of observing an association if one exists (i.e., that we will not make a type II error), and (2) we will think it unlikely that chance could produce an association of that size.

### Statistical power to detect an OR ≠ 1.0 with a one-tailed significance test



This diagram illustrates the overlap between the central portions of the distributions of a test statistic (e.g.,  $Z$ ) expected under the null hypothesis (e.g., true OR is 1.0) and alternate hypothesis (e.g., true OR is 2.25). When we obtain the results from the study we will compute the test statistic (e.g.,  $Z$ ) and compare it to its distribution under the  $H_0$  (the upper of the two distributions in the

diagram). If the calculated value of  $Z$  is smaller than  $z_\alpha$ , i.e., it falls to the left of the cutpoint we have set (defined by the Type I error probability, alpha), then we will conclude that the data we observed came from the upper distribution (the one for no association, true OR=1.0). Even if the OR we observed was greater than 1.0 (which implies that  $Z$  was greater than 0), because  $Z$  was not greater than our cutpoint we regard the observed OR as a chance deviation from 1.0. If the unseen truth is that there really is no association, then our conclusion is correct. If instead the true OR is really 2.25, so that the data we observed really came from the lower distribution, then our conclusion represents a Type II error. The area to the left of the cutpoint on the lower distribution represents the probability of making a Type II error, “beta”. Statistical power – the probability of detecting a true difference – is equal to one minus beta (i.e.,  $1 - \text{beta}$ ).

Conversely, if we observe a value of  $Z$  to the right of the cutpoint, we will conclude that the data we observed did not come from the upper distribution and that therefore the true OR is greater than 1.0. If we are incorrect – if the association we observed was in fact simply a chance finding – then our conclusion represents a Type I error. The area to the right of the cutpoint on the upper distribution represents the probability of making a Type I error, “alpha”.

If we abhor making a Type I error, we can move the cutpoint to the right, which reduces alpha – but increases beta. If we prefer to reduce beta, we can move the cutpoint to the left – but that increases alpha. What we would really like to do is to reduce both alpha and beta, by making the distributions narrower (so that more of the shading is located at the center of the each distribution, symbolizing greater precision of estimation). The width of the distribution is controlled by the sample size. With a powerful light we can easily distinguish between, for example, a snake and a stick. But with a weak light, we cannot be certain what we are seeing. We can elect to err on one side or the other, but the only way to reduce our chance of error is to get a more powerful light.

Commonly used values for alpha and beta are, respectively, 0.05 and 0.20 (power=0.80), for a total probability of error of 0.25. If the study size is limited due to the low incidence of the disease, the low prevalence of the exposure, or the low amount of the budget, then our study estimates will be imprecise – the distributions in the above diagram will be wide. The total error probability will be below 0.25 only when the lower distribution is farther to the right, i.e., corresponds to a stronger association.

In essence, intolerance for error (i.e., small alpha and beta) and desire to detect weak associations must be paid for with sample size. In our courtroom daydream, the better the chance we want of winning the case against R.J. Morris (our power) and/or the more R.J. Morris can persuade the judge to raise the standard of evidence (the significance level), the higher the price we will have to pay for our legal representation (more study subjects.) The Appendix contains a section that translates these concepts into estimated sample sizes.

## Small studies bias

In crude terms, big studies are powerful; small studies are weak. The concept of “small studies bias” illustrates the importance of having an understanding of statistical power when interpreting epidemiologic studies.

The idea behind small studies bias (Richard Peto, Malcolm Pike, et al., *Br J Cancer* 34:585-612, 1976) is that since small studies are easier to carry out than large studies, many more are carried out. Small studies that do not find a “significant” result are often not published. The journals tend not to be interested, since as explained above, there is not much information in a negative study that had low power. In fact, the investigators may not even write up the results – why not just conduct another study. In contrast, large studies are expensive and involve many investigators. Whatever the results from a large study, there is more interest on everyone’s part to publish it.

To the extent that this scenario describes reality, the body of published studies contains primarily small studies with “significant” results and large studies with “significant” and “nonsignificant” results. However, if there are many small (i.e., easy, inexpensive) studies going on, then a 5% probability of making a Type I error translates into a large number of positive findings and resultant publications. Thus, many of the small studies that appear in the literature are reporting Type I errors rather than real associations.

The following example, based on randomized trials of new treatments, comes from the article by Peto, Pike, et al. Assume that there are 100 large and 1,000 small trials of treatments that are not really different, and 20 large and 200 small trials of treatments which are really different. The large trials have statistical power of 95%; the small trials have statistical power of 25%. The significance level is 5%, and only trials reporting significant results are published. These somewhat pessimistic, but perhaps very realistic, assumptions lead to the following hypothetical scenario for the number of treatment trials in progress that will be “statistically significant” ( $p < 0.05$ ):

Planned trial size	True death rate in		Postulated # of trials	Expected number to find	
	Control	Treatment		$p > 0.05$	$p < 0.05$
250	50%	50%	100	95 (TN)*	5 (FP)*
250	50%	33%	20	1 (FN)	19 (TP)
25	50%	50%	1,000	950 (TN)	50 (FP)
25	50%	33%	1,000	150 (FN)	50 (TP)

\* TN, FP, FN, TP are for analogy with sensitivity and specificity (see below).

In this scenario, 100 small trials with “significant” results will be published, but only half of them will reflect a real difference between treatments. Peto, Pike *et al.*'s conclusion is to pay attention only to large trials, particularly ones that are large enough to be published even if they do not find a significant difference in treatments.

These results can be thought of in terms of the concepts of sensitivity, specificity, and predictive value. In this conceptualization, sensitivity corresponds to the statistical power to detect a true difference (95% for large trials, 25% for small trials), specificity corresponds to one minus the significance level – the probability of correctly identifying a chance result (95% specificity for a 5% significance level), and positive predictive value is the probability that a “significant” result in fact reflects a true difference in treatment effectiveness.

### Large trials (e.g., 250 deaths)

True death rate in treatment group (assuming 50% death rate in control group)			
P < 0.05	33%	50%	Total
Yes	19	5	24
No	1	95	96
Total	20	100	120

Thus, the predictive value of a  $p < 0.05 = 19/24 = 79\%$

### Small trials (e.g., 25 deaths)

True death rate in treatment group (assuming 50% death rate in control group)			
P < 0.05	33%	50%	Total
Yes	50	50	100
No	150	950	1,100
Total	200	1,000	1,200

Predictive value of a  $P < .05 = 50/100 = 50\%$

## ***Evaluating the role of chance - interval estimation***

[EPID 168 students are responsible for these concepts, but not for the computations]

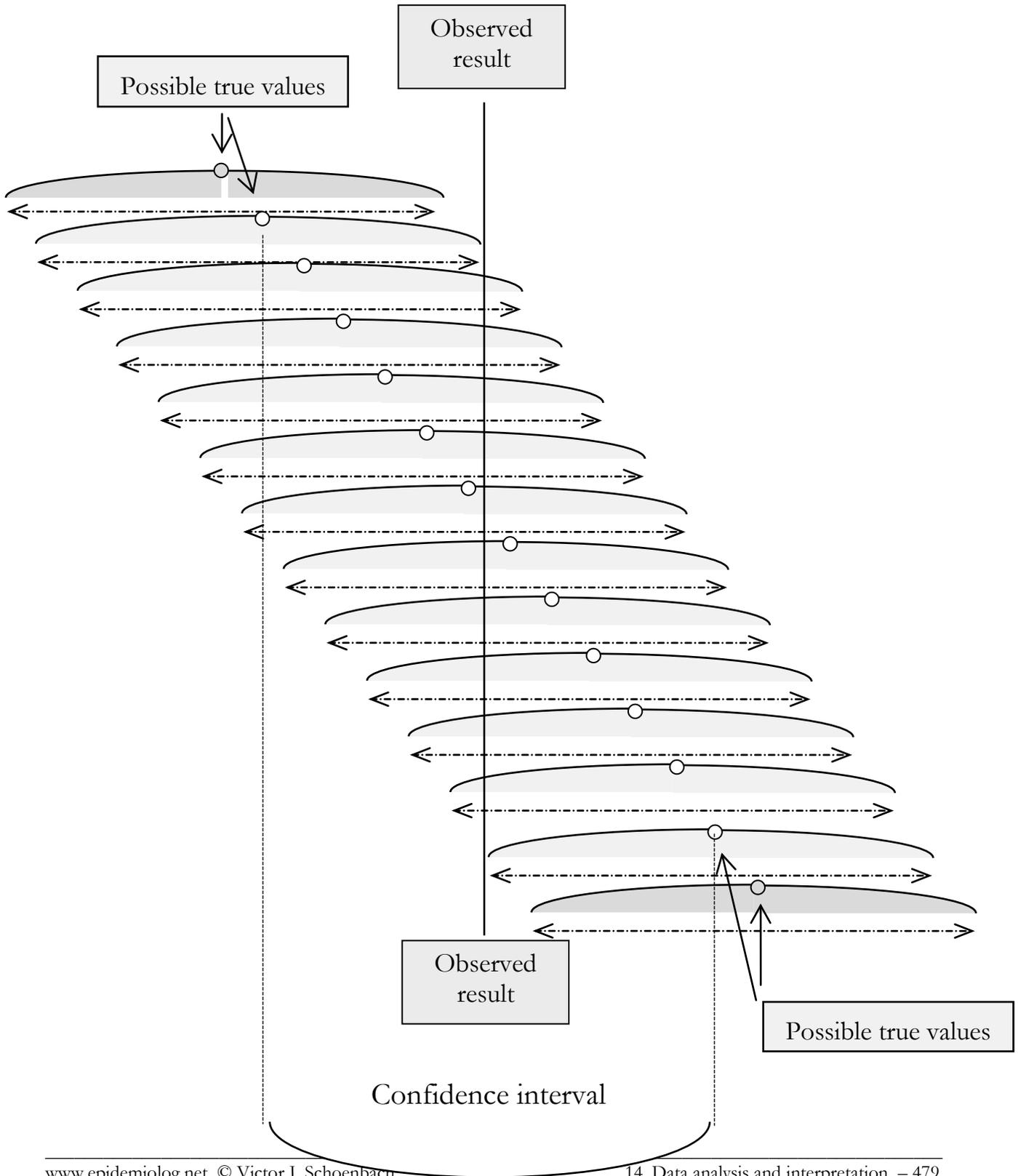
Statistical significance testing, with its decision-making orientation, has fallen somewhat out of favor for reporting data from epidemiologic investigations. On the premise that an epidemiologic study is essentially a measurement process (see Rothman), it is argued that the more appropriate statistical approach is one of estimation (e.g., of a measure of effect) rather than significance testing. Of course, there is still a need to quantify the role of chance, but in an estimation framework chance is

quantified by a confidence interval or confidence limits about the point estimate. Confidence limits quantify the amount of uncertainty in an estimate by defining an interval which should cover the population parameter being estimated (e.g., measure of effect) a known percentage of the time. Various authors have argued that confidence intervals are superior to p-values as a means of quantifying the degree of random error underlying an association.

Confidence intervals address the question, “what possible values for a population parameter (e.g., incidence density ratio) are consistent with the observed results?” Stated another way, “what is the range of true values which, when distorted by haphazard influences, could well have produced the observed results?” Confidence intervals provide a statement about the precision of an estimate or estimates based on the amount of data available for the estimate. If a “significant” association was not observed, then the confidence interval can give some idea of how strong an association might nevertheless exist but, due to the luck of the draw, not be observed.

The nature of a confidence interval and what it does and does not provide, however, is a little tricky (judging from a discussion of confidence intervals on the STAT-L internet listserv that continued for weeks and drew a host of responses and counter-responses). The frequentist view is that a “95% confidence interval” is an interval obtained from a procedure that 95% of the time yields an interval containing the true parameter. Ideally, a 95% confidence interval would be one that “contains the parameter with 95% probability”. But frequentists argue that the interval is set by the data, and the population parameter already exists in nature. The parameter is either in the interval or it is not. There is no probability about it. All that can be said is that 95% of the time the procedure will yield an interval that embraces the value of the parameter (and therefore 5% of the time the procedure will yield an interval that does not). In this view, a 95% confidence interval is like a student who typically scores 95% – the probability that he/she will give the correct answer to a question is 95%, but the answer he/she gave to any particular question was either correct or incorrect.

# The concept behind the confidence interval



## Computing a confidence interval for a ratio measure of effect

Introductory biostatistics courses cover the method for obtaining a 95% confidence interval for the estimate of a population proportion  $p$ . If the sample is large enough so that  $np > 5$  and  $n(1-p) > 5$ , then the confidence limits are:

$$p \pm 1.96 \sqrt{\text{var}(p)}$$

$$p \pm 1.96 \sqrt{p(1-p)/n}$$

where  $p$  is the observed proportion,  $\text{var}(p)$  is the variance of the estimate of  $p$  (so  $\sqrt{\text{var}(p)}$  is the standard error), and  $n$  is the number of observations. For a proportion,  $\text{var}(p)$  equals  $p(1-p)/n$ .

This method can be used to estimate confidence intervals for prevalence, cumulative incidence, and other simple proportions. Many epidemiologic measures, however, are ratios (e.g., CIR, IDR, and OR). Since ratio measures of effect have a highly skewed distribution (most of the possible values lie to the right of the null value of 1.0), the usual approach is to first estimate the confidence interval for the natural logarithm [ $\ln(\text{CIR})$ ,  $\ln(\text{IDR})$ , or  $\ln(\text{OR})$ ] and then take the anti-log (exponent) of the confidence limits:

$$95\% \text{ CI for } \ln(\text{OR}) = \ln(\text{OR}) \pm 1.96 \sqrt{\text{var}[\ln(\text{OR})]}$$

$$\begin{aligned} 95\% \text{ CI for OR} &= \exp\{\ln(\text{OR}) \pm 1.96 \sqrt{\text{var}[\ln(\text{OR})]}\} \\ &= \text{OR} \exp\{\pm 1.96 \sqrt{\text{var}[\ln(\text{OR})]}\} \end{aligned}$$

To obtain the variance of the  $\ln(\text{OR})$ , we use a simple formula (that has been derived by means of a Taylor series approximation to the  $\ln[\text{OR}]$ ):

$$\text{var}[\ln(\text{OR})] = 1/a + 1/b + 1/c + 1/d$$

which works well if  $a$ ,  $b$ ,  $c$ ,  $d$  are all at least 5.

The 95% confidence interval for the  $\ln(\text{OR})$  is therefore:

$$\ln(\text{OR}) \pm 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}$$

and the 95% confidence interval for the OR is:

$$\text{OR} \exp\{\pm 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}\}$$

or

$$\text{OR} e^{\pm 1.96 \sqrt{(1/a + 1/b + 1/c + 1/d)}}$$

Formulas for confidence intervals for the CIR and IDR can be found in Kleinbaum, Kupper and Morgenstern and Rothman and Greenland. Of course, if the study population is highly-selected (i.e., unrepresentative of any other population of interest), how useful is the value of the estimate?

IMPORTANT CAVEAT: Everything in this section, of course, has been based on the assumption of perfect (unbiased, independent) sampling and measurement. Anything other than an unbiased simple random sample and any error in measurement will invalidate the above at least to some extent.

## **Meta-analysis**

Meta-analysis is a quantitative approach to summarizing and synthesizing the findings from different studies of a particular relationship of interest. Meta-analysis proceeds from the recognition that the failure to find “significant results” can be due as much to the limited statistical power of individual studies as to the absence of a relationship. Combining the information from multiple studies can yield a more precise and definitive assessment of the existence and strength of a relationship than is available from any one study or, it is argued, from a nonquantitative distillation of the literature.

There are four steps in carrying out a meta-analysis: 1) formulating the problem, 2) identifying the studies (published and unpublished), 3) coding and evaluating the studies, and 4) statistical analysis. Steps 2) and 3) are critical for the validity of the meta-analysis, since the judgments from the meta-analysis will depend upon the adequacy with which the evidence about the relationship is represented by the studies that are finally analyzed (the possibility of publication bias against “negative” studies implies that some effort should be made to locate unpublished studies). The strategy for statistical analysis can be similar to that for stratified analysis, regarding each study as a separate “stratum”. More refined approaches recognize that the studies themselves can be regarded as a sample from some universe of possible studies, so that the weighting scheme needs to take into account inter-study variability as well as intra-study variability (as in the random-effects model of analysis of variance).

In its pure form, meta-analysis is predicated on the assumption that the collection of studies represents a random sample of equivalently-obtained observations of an association, so that the differences across the studies can be regarded as random (sampling) variability. Hence, a summary constructed by combining the studies gives a more precise estimate of the true association. In actual practice, however, epidemiologic studies are rarely equivalent, since they often differ in the population(s) studied, measurements taken, and analysis approaches. Even studies that appear to be equivalent (e.g., “unmatched population-based case-control studies with a physiologic measure of exposure and control for the same set of potential confounders”) will differ in less obvious ways: the populations likely differ in unknown and unmeasured ways, the disease ascertainment systems may differ across populations, response factors in the selection of controls may differ, collection processes and laboratory analysis of the exposure may differ in subtle ways that can nevertheless affect the results (e.g., see examples involving HIV tests and homocysteine analyses in *J Clin Epidemiol* 2001(5)), and collection of data and analytic handling of potential confounders can differ. An exploration of heterogeneity in the meta-analysis of studies of SIDS and sleeping positions (Dwyer et al, 2001) illustrates some of these issues.

## ***Interpretation of results***

### **Key questions**

1. How good are the data?
2. Could chance or bias explain the results?
3. How do the results compare with those from other studies?
4. What theories or mechanisms might account for findings?
5. What new hypotheses are suggested?
6. What are the next research steps?
7. What are the clinical and policy implications?

## Bibliography

### General

Ahlbom, Anders. *Biostatistics for epidemiologists*. Boca Raton, Florida, Lesis Publishers, 1993, 214 pp., \$45.00 (reviewed in *Am J Epidemiol*, April 15, 1994).

Bailar, John C., III; Thomas A. Louis, Philip W. Lavori, Marcia Polansky. Studies without internal controls. *N Engl J Med* 1984; 311:156-62.

Bauer UE, Johnson TM. Editing data: what difference do consistency checks make. *Am J Epidemiol* 2000;151:921-6.

Bulpitt, C.J. Confidence intervals. *The Lancet* 28 February 1987: 494-497.

Dwyer, Terence; David Couper, Stephen D. Walter. Sources of heterogeneity in the meta-analysis of observational studies: The example of SIDS and sleeping position. *J Chron Dis* 2001;54:440-447.

Feinstein, Alvan R. The fragility of an altered proportion: a simple method for explaining standard errors. *J Chron Dis* 1987; 40:189-192.

Feinstein, Alvan R. X and iprr: An improved summary for scientific communication. *J Chron Dis* 1987; 40:283-288.

Frank, John W. Causation revisited. *J Clin Epidemiol* 1988; 41:425-426.

Gerbarg, Zachary B.; Ralph I. Horwitz. Resolving conflicting clinical trials: guidelines for meta-analysis. *J Clin Epidemiol* 1988; 41:503-509.

Glantz, Stanton A. *Primer of biostatistics*. NY, McGraw-Hill, 1981.

Godfrey, Katherine. Comparing means of several groups. *N Engl J Med* 1985;313:1450-6.

Hertz-Picciotto, Irva. What you should have learned about epidemiologic data analysis. *Epidemiology* 1999;10:778-783.

Northridge, Mary E.; Bruce Levin, Manning Feinleib, Mervyn W. Susser. Statistics in the journal—significance, confidence, and all that. Editorial. *Am J Public Hlth* 1997;87(7):1092-1095.

Powell-Tuck J, MacRae KD, Healy MJR, Lennard-Jones JE, Parkins RA. A defence of the small clinical trial: evaluation of three gastroenterological studies. *Br Med J* 1986; 292:599-602.

Ragland, David R. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology* 1992;3:434-440

Rothman - *Modern Epidemiology*, Chapters 9, 10, 14.

Schlesselman - *Case-control studies*, Chapters 7-8. (Especially the first few pages of each of these chapters).

Woolf SH, Battista RN, Anderson GM, Logan AG, et al. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *J Clin Epidemiol* 1990; 43:891-905.

Zeger, Scott L. Statistical reasoning in epidemiology. *Am J Epidemiol* 1991; 134(10):1062-1066.

### **The role of statistical hypothesis tests, confidence intervals, and other summary measures of statistical significance and precision of estimates**

Allan H. Smith and Michael N. Bates. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology* 1992;3:449-452

Browner, Warren S.; Thomas B. Newman. Are all significant P values created equal? *JAMA* 1987; 257:2459-2463.

Fleiss, Joseph L. Significance tests have a role in epidemiologic research: reactions to A.M. Walker (*Am J Public Health* 1986; 76:559-560). See also correspondence (587-588 and 1033).

George A. Diamond and James S. Forrester. Clinical trials and statistical verdicts: probable grounds for appeal. *Annals of Internal Medicine* 1983; 93:385-394

Greenland, Sander. Randomization, statistics, and causal inference. *Epidemiology* 1990;1:421-429.

Maclure, Malcome; Greenland, Sander. Tests for trend and dose response: misinterpretations and alternatives. *Am J Epidemiol* 1992;135:96-104.

Mood, Alexander M. and Franklin A. Graybill. *Introduction to the theory of statistics*. 2ed. NY, McGraw-Hill, 1963.

Oakes, Michael. *Statistical inference*. Chestnut Hill, Mass., Epidemiology Resources, 1986.

Peace, Karl E. The alternative hypothesis: one-sided or two-sided? *J Clin Epidemiol* 1989; 42(5):473-477.

Poole, Charles. Beyond the confidence interval *Am J Public Health* 1987; 77:195-199.

Poole, C. Confidence intervals exclude nothing *Am J Public Health* 1987; 77:492-493. (Additional correspondence (1987; 77:237)).

Savitz DA, Tolo KA, Poole C. Statistical significance testing in the *American Journal of Epidemiology*, 1970-1990. *Am J Epidemiol* 1994;139:1047-.

Thompson, W. Douglas. Statistical criteria in the interpretation of epidemiologic data *Am J Public Health* 1987; 77:191-194.

Thompson, W.D. On the comparison of effects *Am J Public Health* 1987; 77:491-492.

Walker, Alexander M. Reporting the results of epidemiologic studies *Am J Public Health* 1986; 76:556-558.

Woolson, Robert F., and Joel C. Kleinman. Perspectives on statistical significance testing. *Annual Review of Public Health* 1989(10).

### **Sample size estimation**

Donner A, Birkett N, and Burk C. Randomization by Cluster: sample size requirements and analysis. *Am J Epidemiol* 1981; 114:706

Snedecor GW, Cochran WG. *Statistical Methods*, 1980 (7th ed) see pages 102-105, 129-130 (Table A is from page 104).

### **Imputation**

Heitjan, Daniel F. Annotation: what can be done about missing data? Approaches to imputation. *Am J Public Hlth* 1987;87(4):548-550.

Little RJA, Rubin DB. *Statistical analysis with missing data*. NY, Wiley, 1987.

### **Interpretation of multiple tests of statistical significance**

Bulpitt, Christopher J. Subgroup analysis. *Lancet* 1988 (July 2);31-34.

Cupples, L. Adrienne; Timothy Heeren, Arthur Schatzkin, Theodore Coulton. Multiple testing of hypotheses in comparing two groups. *Annals of Internal Medicine* 1984; 100:122-129.

Holford, Theodore R.; Stephen D. Walter, Charles W. Dunnett. Simultaneous interval estimates of the odds ratio in studies with two or more comparisons. *J Clin Epidemiol* 1989; 42(5):427-434.

Jones, David R. and Lesley Rushton. Simultaneous inference in epidemiological studies. *Int J Epidemiol* 1982;11:276-282.

Lee, Kerry L., Frederick McNeer, Frank Starmer, et al. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 61:508-515, 1980.

Stallones, Reuel A. The use and abuse of subgroup analysis in epidemiological research. *Preventive Medicine* 1987; 16:183-194 (from Workshop on Guidelines to the Epidemiology of Weak Associations)

See also Rothman, *Modern Epidemiology*.

### **Interpretation of “negative” studies**

Freiman, Jennie A., Thomas C. Chalmers, Harry Smith, Jr., and Roy R. Kuebler. The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;299:690-694.

Greenland, Sander. Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol* 2012;22:364-368.

Hulka, Barbara S. When is the evidence for ‘no association’ sufficient? Editorial. *JAMA* 1984; 252:81-82.

### **Meta-analysis**

Light, R.J.; D.B. Pillemer. *Summing up: the science of reviewing research*. Cambridge MA, Harvard University Press, 1984. (very readable)

Longnecker M.P.; J.A. Berlin, M.J. Orza, T.C. Chalmers. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 260(5):652-656. (example)

Wolf, F.M. *Meta-Analysis: quantitative methods for research synthesis*. Beverly Hills, CA, Sage, 1986.

### **Bias**

Greenland, Sander. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; 112:564-569.

Walter, Stephen D. Effects of interaction, confounding and observational error on attributable risk estimation. *Am J Epidemiol* 1983;117:598-604.

## Appendix

### ***Estimating sample size to compare two proportions or means***

(Adapted from a summary provided by of Dana Quade, UNC Department of Biostatistics, June 1984)

Let  $N$  be the number of subjects (observational units) required in **each** of two groups to be compared. Then

$$N = I \times D \times C$$

Where:

$I$  = Intolerance for error, which depends on:

- Alpha = Desired significance level that we want to use for our hypothesis test (e.g., 5%, two-sided)
- Beta = Type II error (e.g., 10 – same as 1 - power)

Formula:  $I = (Z_{\alpha} + Z_{\beta})^2$

$Z_{\alpha}$  and  $Z_{\beta}$  are, respectively, the critical values corresponding to alpha and beta from the normal distribution (see Table A on next page)

$D$  = Difference to detect, which depends on the narrowness of the difference between the true proportions or means, in relation to the standard deviation of that difference.  $D$  can be regarded as the inverse of the “signal-to-noise ratio” – the softer the signal or the louder the noise, the more subjects needed

$$D = \frac{\text{“noise”}}{\text{“signal”}} \quad \text{OR} \quad \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_1 - p_2)^2} \quad \text{OR} \quad \frac{2(\sigma^2)}{(\mu_1 - \mu_2)^2}$$

(for differences in proportions, where  $p_1$  and  $p_2$  are the two proportions – see table on next page)

(for differences in means, where  $\mu_1$  and  $\mu_2$  are the two means, and  $\sigma^2$  is the variance of the difference)

C - Clustered observations, which depends on whether observations are selected independently or in clusters.

- If all observations are sampled independently,  $C = 1$ .
- If observations are sampled in clusters (e.g., by households, schools, worksites, census tracts, etc.), then sample size must be increased to offset the fact that observations within a cluster are more similar to each other than to observations in other clusters. If rho is the intracluster correlation among observations within clusters, then:

$$C = 1 + (m-1) \rho$$

where m is the average cluster size (i.e.,  $n = km$ , where k is the number of clusters). C is often referred to as the *design effect*. If the clusters are large or if the people in them tend to be very similar, then individual subjects contribute little information and you therefore need to study a very large number of them. If you choose “independent thinkers”, you will learn more from each one.

**Table A: Intolerance for error**

Desired power	Two-Tailed Tests			One-Tailed Tests		
	<u>Significance Level</u>			<u>Significance Level</u>		
	0.01	0.05	0.10	0.01	0.05	0.10
0.80	11.7	7.9	6.2	10.0	6.2	4.5
.90	14.9	10.5	8.6	13.0	8.6	6.6
0.95	17.8	13.0	10.8	15.8	10.8	8.6

**Table B: Difference to be detected**

		p2					
		.10	.20	.30	.40	.50	.60
p1	.05	55	9.2	4.1	2.4	1.5	1.0
	.10	–	25	7.5	3.7	2.1	1.3
	.15	87	115	15	5.9	3.1	1.8
	.20	25	–	37	10	4.6	2.5
	.25	12.3	139	159	19	7.0	3.5

**Complications**

*1) Unequal sample sizes*

Let n be the **average** sample size =  $(n_1+n_2)/2$

Let  $\lambda_1 = n_1/2n$ ,  $\lambda_2 = n_2/2n$  ( $\lambda_1 + \lambda_2 = 1$ )

$$D = \frac{\frac{p_1(1-p_1)}{2\lambda_1} + \frac{p_2(1-p_2)}{2\lambda_2}}{(p_1-p_2)^2} \quad \text{OR} \quad \frac{\frac{\sigma^2_1}{2\lambda_1} + \frac{\sigma^2_2}{2\lambda_2}}{(\mu_1-\mu_2)^2}$$

*2) Covariables*

If statistical tests are to be conducted separately within each stratum then n as determined above is required for each stratum.

If results for different strata are to be tested only for an overall average association, it is probably best not to try to allow for them in the sample size formulas explicitly, but make a modest overall increase in n.

Note: more “precise” formulas can be found in the literature, but the parameters needed for factors D and C are never really known.

### **Sample size for interval estimation**

The tolerable width for a confidence interval can be used as the target for estimating the required sample size for a study population. Suppose, for example, that an investigator wishes to estimate the proportion (p) of condom use in a clinic population. If the investigator can obtain a simple random sample of that population, then her estimate of the proportion of condom users would be  $p = u/n$ , where u is the number of users in the sample and n is the size of her sample. As noted above, if  $np > 5$  and  $n(1-p) > 5$ , then a 95% confidence interval for p is :

$$p \pm 1.96 \sqrt{[p(1 - p)/ n]}$$

For example, if p is 0.50, then the confidence interval is:

$$0.5 \pm 1.96 \sqrt{[(0.5)(0.5)/n]} = 0.5 \pm 1.96 \frac{(0.5)}{\sqrt{n}}$$

[The square root of (0.5)(0.5) is, of course, 0.5]

Since  $1.96 \times 0.5$  is approximately 1, for practical purposes this expression is equivalent to:

$$0.5 \pm 1/\sqrt{n}, \text{ so that the confidence limits are } (0.5 - 1/\sqrt{n}, 0.5 + 1/\sqrt{n})$$

For example, suppose that n, the sample size, is 100. Then the 95% confidence interval around the point estimate of 0.5 is:

$$\begin{aligned} & (0.5 - 1/\sqrt{100}, 0.5 + 1/\sqrt{100}) \\ = & (0.5 - 1/10, 0.5 + 1/10) \\ = & (0.5 - 0.1, 0.5 + 0.1) \\ = & (0.4, 0.6) \end{aligned}$$

Imprecision is often quantified in terms of the half-width of the interval, i.e., the distance between the point estimate and the interval’s upper (or lower) limit, which we will refer to here as the “margin of error”. The half-width of the above interval is 0.1 (i.e., the square root of n) in absolute terms or 20% (0.1/0.5) in relative terms. A 0.1 absolute or 20% relative margin of error is adequate for a “ballpark” estimate of a proportion, but not much more.

Since the above expressions involve the square root of the sample size, progressive narrowing of the interval width involves substantially greater increases in sample size. For example, to obtain a 0.05 absolute or 10% relative margin of error, sample size must be quadrupled, to 400:

$$\begin{aligned}
 & (0.5 - 1/\sqrt{400}, 0.5 + 1/\sqrt{400}) \\
 = & (0.5 - 1/20, 0.5 + 1/20) \\
 = & (0.5 - 0.05, 0.5 + 0.05) \\
 = & (0.45, 0.55)
 \end{aligned}$$

Similarly, a sample size of 900 yields confidence limits one-third as wide as from a sample of 100, a sample of 2,500 yields limits one-fourth as wide as for n=100, etc.

These numbers apply to a point estimate of 0.5, which produces the widest error margin in absolute terms. A smaller or greater point estimate will have a narrower (in absolute terms) interval, because the square root of p(1 - p) cannot exceed 0.5 (try it! – or use calculus). The relative margin of error, on the other hand, is inversely related to the size of the point estimate. Examine the following table:

Point estimate	Sample size	Margin of error (rounded)	
		Absolute *	Relative ** (%)
0.1	100	0.06***	60***
0.2	100	0.080	40
0.3	100	0.090	30
0.4	100	0.096	24
0.5	100	0.100	20
0.6	100	0.096	16
0.7	100	0.090	12
0.8	100	0.080	9.8
0.9	100	0.060	6.5
0.1	400	0.030	30
0.2	400	0.040	20
0.3	400	0.045	15
0.4	400	0.048	12
0.5	400	0.050	10
0.6	400	0.048	8.0
0.7	400	0.045	6.4
0.8	400	0.040	4.9
0.9	400	0.030	3.2

\* Approximate half-width of 95% confidence interval in absolute terms

\*\* Approximate half-width of 95% confidence interval in absolute terms, relative to the size of the point estimate

\*\*\* Calculation:  $1.96 \sqrt{[(0.01)(1-0.01) / 100]} = 1.96 (0.03) = 0.0588 \approx 0.06$  absolute error margin

This table illustrates that:

1. quadrupling sample size halves the margin of error.
2. absolute error margin decreases as the point estimate moves away from 0.5
3. relative error margin is inversely – and very strongly – related to the size of the point estimate

For very small point estimates, as illustrated in the following table, very large samples are required to obtain a small relative margin of error. Even a sample size of 2,500 still produces a relative error margin of 17% for a proportion of 0.05.

Point estimate	Sample size	Margin of error (rounded)	
		Absolute *	Relative * (%)
0.5	100	0.10	20
0.5	400	0.05	10
0.5	900	0.033	6.6
0.5	1,600	0.025	5.0
0.5	2,500	0.020	4.0
0.05	100	0.043	85
0.05	400	0.021 **	43 **
0.05	900	0.014	28
0.05	1,600	0.011	21
0.05	2,500	0.009	17

\* See previous table

\*\* Calculation:  $1.96 \times \sqrt{[(0.05)(0.95)/400]} = 1.96 \times 0.0109$

= 0.0214  $\approx$  0.021 absolute error margin

Relative =  $0.0214 / 0.05 = 0.427 = 42.7\%$  (approximately 43%)

Recall that this formula requires that  $nP \geq 5$ , which is just met for  $P=0.05$  and  $n=100$ .

How large a sample is large enough? If the objective is to set an upper or lower bound on a proportion, then a small absolute margin of error may suffice. For example, if one is testing for hepatitis C antibody and wants to be reassured that the seroprevalence is below 5%, then a sample

size of 900 will produce an interval with an absolute error margin no wider than 0.033 (for a point estimate of 0.5 – see above table) and more likely 0.011 (for a point estimate of 0.05) or smaller. Since we expect the seroprevalence to be very small, then the 0.011 is much more relevant than the 0.033. If when we carry out the study we obtain a point estimate of exactly 0.05, then the 95% confidence interval will be (0.039,0.061) which will tell us that the true value is at least not likely to be greater than 6%. If the point estimate is below 0.04, then the upper confidence limit will be below 5% and we are reassured that the seroprevalence is no greater than that value.

Note that the above is all based on the assumption of perfect (unbiased) simple random sampling and measurement. Anything other than an unbiased simple random sample and any error in measurement will invalidate the above at least to some extent.

### ***Meditations on hypothesis testing and statistical significance***

The statistical theory of hypothesis testing and assessment of statistical “significance” proceeds from an analysis of decision-making with respect to two competing hypothesis: a “null” hypothesis and an alternative hypothesis. Two types of errors are possible:

Type I: Erroneously reject the “null hypothesis” ( $H_0$ ), in favor of the alternate hypothesis ( $H_A$ ), i.e., erroneously reject chance as a sufficient explanation for the observed results.

Type II: Erroneously fail to reject  $H_0$ , i.e., erroneously accept chance as an explanation. [A parallel dichotomy will be seen later in the course when we discuss sensitivity and specificity.]

Traditionally, the Type I error probability has received more attention and is referred to as the “significance level” of the test. The Type I error presumably owes its prominence to the scientific community’s desire to avoid false alarms, i.e., to avoid reacting to results that may readily have been chance fluctuations. Also the Type I error probability is easier to estimate, since the Type II error probability depends on stating the size of true difference that one seeks to detect.

During recent decades, the calculation and presentation of “p-values” (which give information about the Type I error probability) have become *de rigueur* in the empirical scientific literature. Indeed, a significant (!) number of people refuse to pay any attention to results that have p-values greater than .05 (5% probability of a Type I error).

Such a stance is a considerable labor-giving device, but is perhaps a bit brutal. After all, a result with a p-value of .10 would result from a chance process in only one in ten trials. Should such a finding be dismissed? Moreover, since the p-value reflects the number of subjects as well as the size of the observed difference, a small study will have very small p-values only for large (and perhaps unrealistic?) observed differences. If the size of the observed difference is unreasonably large, then we may be suspicious of the finding despite a small p-value. If the observed difference is plausible, but because the study is small the p-value is “not significant”, we may nevertheless want to pay some attention.

Another reason for a reflective, rather than a reflexive, approach to p-values (and statistical inference generally) is that the probability estimates themselves are accurate only with respect to the models that underlie them. Not only may the mathematical models not adequately capture the real situation, but the context in which they are used clouds the issue. One critical assumption is that of random sampling or randomization (as in a randomized controlled trial). Although this assumption is the basis for virtually all of the statistical theory of hypothesis testing and confidence intervals, it is rarely met in observational studies and the limitations that it imposes on the interpretation of statistical tests are often underappreciated (Greenland S. Randomization, statistics, and causal inference *Epidemiology* 1990;1:421-249).

Even in randomized trials problems of interpretation exist. For example, the p-value for a single result in a single study may be 5 percent. But that means that 20 independent studies of two identical phenomena would observe, on the average, one difference that was “significant” at the five percent level. A prolific investigator who conducts 200 studies in his/her professional life can expect to have ten that are “significant” by chance alone. Moreover, a study will often examine multiple outcomes, including multiple ways of defining the variables involved.

Such “multiple comparisons” increase the likelihood of chance differences being called “significant”. But the statistical procedures for dealing with this “significance inflation” tend, like measures to suppress price inflation or grade inflation, to produce recession or even depression [of study findings]. Should an investigator be required to take an oath that he/she had (1) fully specified an a priori hypothesis, including the procedures for defining and manipulating all variables, decisions about all relationships to examine, what factors to control, etc; (2) proceeded directly to the pre-specified statistical test without looking at any other data; and (3) will not perform any further statistical tests with the same data? (See *Modern Epidemiology* for more on these points.)

What about so called “fishing expeditions” in which an investigator (or her computer) pores over a dataset to find “significant” relationships. Is such a procedure better characterized as “seek and ye shall find” or as “search and destroy”? Some analysts recommend adjusting the significance level to take account of such “multiple comparisons”, but an energetic investigator can easily perform enough tests so that the adjusted significance level is impossible to satisfy. Other writers (e.g., Rothman, Poole) assert that no adjustment is required – that once the data are in, the number of tests is irrelevant. Others (e.g., Greenland) have proposed more sophisticated approaches to adjustment. Perhaps the best course at this time is twofold:

(1) If you are conducting a study, for example, a randomized trial, in which you have a good chance of satisfying the assumptions for a statistical hypothesis test and are hoping to test a specific hypothesis, especially one that may lead to some decision, then it is probably better to adhere to the Neyman-Pearson hypothesis testing format as much as possible. This approach ensures maximum impact for your results;

(2) If you are conducting an inquiry with few of the above characteristics, or have already completed the a priori hypothesis test, analyze all that you like but be candid in describing how you proceeded. Then readers can interpret as they judge most appropriate.

Apparent (calculated) power is rarely achieved because it often assumes no errors in classification of subjects. A study with advertised power of 90% could well have much less probability of detecting a given true difference because of dilution by information bias. Similarly we can in principle improve the effective power of a study if we can increase the precision with which important variables are measured.

Louis Guttman has written that estimation and approximation, never forgetting replication, may be more fruitful than significance testing in developing science. [Louis Guttman. What is not what in statistics. *The Statistician* 25(2):81-107.]

Independent replication is the cornerstone of scientific knowledge.

### ***Bayesian approach to p-value interpretation***

The application of the concepts of sensitivity, specificity, and predictive value to interpreting statistical hypothesis tests suggests an analogy between statistical tests and diagnostic tests (see Browner and Newman, 1987; Diamond and Forrester, 1983; and Feinstein, *Clinical Biostatistics*). Just as the interpretation of a diagnostic test depends upon disease prevalence (the “*a priori* likelihood that the patient has the disease”), the interpretation of statistical tests can be regarded as dependent upon “truth prevalence”, i.e., on the reasonableness of the hypothesis.

As noted earlier, we would like statistical inference to provide an estimate of the probability that a hypothesis of interest (H) is true given the observed results. The p-value provides instead the probability of observing an extreme result under a null hypothesis (typically the inverse of the hypothesis of interest). The Bayesian approach to interpreting p-values tries to provide an answer that comes closer to the original objective. In the Bayesian approach, we begin with a prior probability for the truth of the hypothesis and then adjust that probability based on the results of a study, to obtain a posterior probability. The effect that the study results have on our assessment of the credibility of the hypothesis depends on our original assessment of its credibility.

Let T mean that a statistical test is “significant”. According to Bayes Theorem, if Pr(H) is the “*a priori*” probability of H, i.e., the probability that H is true based only on previous information, then the *a posteriori* probability of H (the probability that H is true based on previous information and the current test result) is:

$$\Pr(H|T) = \frac{\Pr(H) \Pr(T|H)}{\Pr(H) \Pr(T|H) + \Pr(h) \Pr(T|h)}$$

[where Pr(T|h) means the probability of a positive test given that the hypothesis is not true] which can be written:

$$\Pr(H|T) = \frac{\Pr(H) \Pr(T|H)}{\Pr(H) \Pr(T|H) + [1 - \Pr(H)] \Pr(T|h)}$$

Since  $\Pr(T|H)$  is the statistical power (the probability of a positive test given that the hypothesis is true) and  $\Pr(T|h)$  is the p-value (the probability of a positive test given that the hypothesis is not true), the posterior probability can be written:

$$\Pr(H|T) = \frac{\Pr(H) \text{ (power)}}{\Pr(H) \text{ (power)} + [1 - \Pr(H)] \text{ (p-value)}}$$

$\Pr(H|T)$  is therefore a function of the “*a priori*” probability of the hypothesis, the statistical power, and the p-value. Therefore the p-value has more impact on  $\Pr(H|T)$  when  $\Pr(H)$  is small (i.e., when a hypothesis is not supported by prior research or laboratory data) (see Diamond and Forrester).

To get an idea of how these formulas work with typical values for the various elements, take a look at the following table:

**Evaluation of posterior probability based on  
prior probability, statistical power, and p-value**

	Prior probability (Before the study) Pr(H)	Statistical power of the study Pr(T H)	P-value (Findings of the study) Pr(T h)	Posterior probability (After the study) Pr(H T)	
Credible hypotheses	0.60	0.8	0.100	0.92	High power
	0.60	0.8	0.050	0.96	
	0.60	0.8	0.001	1.00	
	0.60	0.5	0.100	0.88	Low power
	0.60	0.5	0.050	0.94	
	0.60	0.5	0.001	1.00	
Long shot hypotheses	0.05	0.8	0.100	0.30	High power
	0.05	0.8	0.050	0.46	
	0.05	0.8	0.001	0.98	
	0.05	0.5	0.100	0.21	Low power
	0.05	0.5	0.050	0.34	
	0.05	0.5	0.001	0.96	

In this table, for example, a very strong p-value (e.g., 0.001) gives high credibility (posterior probability) even for a long shot hypothesis examined in a study of low statistical power. A p-value that is “just significant”, however, does not make a hypothesis highly credible unless it was judged more likely than not before the study. Even a “nonsignificant” p-value (e.g., 0.10) provides some increase in credibility of the hypothesis, so in the Bayesian framework a p-value of 0.10 would not be regarded as a “negative” result casting doubt on the existence of an association. Meta-analysis, in which results are combined across studies to obtain a quantitative assessment of an association from the full body of evidence, also takes into account evidence for the association from studies that observed an association but had a p-value greater than 0.05. Formal use of Bayesian methods in everyday work, however, is somewhat constrained by the absence of an obvious method for obtaining a prior probability.

We can also use these concepts to show why the frequent temptation to interpret a small p-value as the probability that the result was due to chance is mistaken. When confronted with a p-value of, e.g., 0.05, many statistical novices are tempted to interpret the result as meaning that “the probability that the finding was due to chance is only 5%”, which is equivalent to a posterior probability that the hypothesis is true of  $1 - 0.05 = 0.95 = 95\%$ . From the above table, we see that the only situations in which the posterior probability is close to one minus the p-value occur when the hypothesis is more likely than not to be true (*a priori* probability of 0.60). For long shot hypotheses, a p-value of 0.05 corresponds to a posterior probability of much less than 95%.

## ***More meditations on interpreting statistical significance tests***

Some of the concepts in the interpretation of statistical tests of significance can perhaps be illustrated through an example based on one glorious origin of probability theory – games of chance. Suppose that our friend tells you that he has an intuition about roulette wheels. By watching the operator spin the wheel, your friend can, he claims, predict where the ball will land within a very small margin. If, for simplicity, the wheel has numbers 1-100 on it, your friend says he can predict the numbers where the ball will land. He wants you to put up some money to send him to Monte Carlo to make our fortunes.

Naturally you're excited by the prospect of instant riches but also a bit skeptical. To verify your friend's claim, you undertake a statistical test. You give your friend \$5 to prove his prowess at the local gambling casino, and you wait to see how he does.

The null hypothesis for your statistical test is that your friend has no special ability, so that his chances of predicting the resting place of the ball on any one try are simply 1 out of 100 (.01). The 1-sided alternate hypothesis is that your friend does have this ability and can predict the correct number more often than 1 out of 100 times. [The 2-sided alternate hypothesis is that your friend will predict the resting place of the ball either more than would be expected by chance or less than would be expected.]

Your friend returns with \$400. Knowing that the probability of his being correct on a given try by chance alone was only 1%, you are impressed. His performance was “significant at the .01 level”! Do you underwrite his trip to Monte Carlo? How do you interpret his correct prediction?

Is it correct to say that there is only a 1% chance that his accurate prediction was due to “luck”? Not quite. According to the frequentist interpretation, the prediction was made and the roulette wheel has already been spun. The accuracy was due either to “chance” (“luck”) or your friend's ability, but only one of these was actually responsible that time. So the probability that his correct prediction was due to chance is either zero (i.e., your friend can predict) or one (your friend cannot predict). The only trouble is, you don't know which!

You can say (before the wheel was spun and assuming it was a balanced wheel) that if your friend had no special ability there was only a one percent probability of his making a correct prediction and that therefore his winning is evidence against the null hypothesis (of no ability) and in favor of the alternate hypothesis (ability to predict). If you have to decide that day, you might figure that it would be worth underwriting his trip to Monte Carlo, but you would be aware that his correct prediction could have been due to chance because there was a one percent probability that in the absence of any clairvoyance his prediction would have been correct (not quite the same as a one percent probability that his correct prediction was due to chance). So you give your friend \$2,000. He thanks you profusely, and in parting, tells you that it actually took him 30 tries to make a correct prediction – he borrowed the money for the other 29 tries.

That information gives you pause. Certainly you would not have been so impressed if he had told you he could make a correct prediction in 30 tries. If the probability of a correct prediction (i.e., a correct guess) in the absence of any special ability is 0.01, then the probability of one or more correct guesses in 30 tries is 0.26 (1.0 minus the quantity 0.99 raised to the 30th power). Twenty-six percent is still less than 50 percent, i.e., the probability of winning a coin flip, but not so impressively. The evidence against the null hypothesis is now not nearly so strong. This change in your interpretation illustrates the issue that arises in connection with multiple significance tests and small studies bias.

It is possible, using statistical theory, to adjust significance levels and p-values to take into account the fact that multiple independent significance tests have been done. But there are various practical problems in applying such procedures, one of which is the lack of independence among multiple tests in a particular set of data. For example, if your friend explained that he so rarely makes an incorrect prediction that when he did he became so upset that it took him a whole hour (and 29 more predictions) to regain his predictive ability, then even if you remained skeptical you would be hard-put to calculate an adjusted p-value for your test if you thought he was telling the truth. Similarly, in a given dataset, does the fact that an investigator tested the same difference in various ways (e.g., obesity as indexed by weight/height<sup>2</sup> [Quetelet's index], weight/height<sup>3</sup> [ponderal index], percent above ideal weight, skinfold thickness, and body density) weaken the findings for each test? If she also looked at blood pressure differences, would that weaken the credibility of statistical significance of differences in obesity?

“You pays your money, and you takes your choice.”