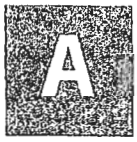
Biostatistics & Epidemiology



Marc B. Schenker, MD, MPH

It is apparent to anyone who reads the medical literature today that some knowledge of biostatistics and epidemiology is a necessity. This is particularly true in occupational and environmental health, in which many of the findings are based on epidemiologic studies of subjects exposed to low levels of an agent. Research has become more rigorous in the area of study design and analysis, and reports of clinical and epidemiologic research contain increasing amounts of statistical methodology. This Appendix provides a brief introduction to some of the basic principles of biostatistics and epidemiology.

I. BIOSTATISTICS

DESCRIPTIVE STATISTICS

Types of Data

Data collected in medical research can be divided into three types: nominal (categorical), ordinal, and continuous.

Nominal (categorical) data are those that can be divided into two or more unordered categories, such as gender, race, or religion. In occupational medicine, for example, many outcome measures, such as cancer rates, are considered separately for different gender and race categories.

Ordinal data are different from nominal data in that there is a predetermined order underlying the categories. Examples of ordinal data include clinical severity, socioeconomic status (SES), or ILO (International Labor Office) profusion category for pneumoconiosis on chest radiographs.

Both nominal and ordinal data are examples of discrete data. They take on only integer values.

Continuous data are data measured on an arithmetic scale. Examples include height, weight, blood lead levels, or forced expiratory volume. The accuracy of the number recorded depends on the measuring instrument, and the variable can take on an infinite number of values within a defined range. For example, a person's height might be recorded as 72 inches or 72.001 inches or 72.00098 inches depending on the accuracy of the measuring instrument.

Summarizing Data

Once research data are collected, the first step is to summarize them. The two most common ways of summarizing data are measures of location, or central tendency, and measures of spread, or variation.

A. Measures of Central Tendency:

1. Mean—The mean ($\bar{x}$) is the average value of a set of interval data observations. It is computed using the following equation:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where n is sample size and x_i is a random variable, such as height, with $i = 1, \ldots, n$.

The mean can be strongly affected by extreme values in the data. If a variable has a fairly symmetric, or bell-shaped, distribution, the mean is used as the appropriate measure of central tendency.

2. Median—The median is the "middle" observation, or 50th percentile; that is, half the observations lie above the median and half below. It can be applied to interval or ordinal data. When there is an odd number of observations, the median is merely the middle observation. For example, for the following series of observations of subjects' weights (in pounds): 124, 138, 139, 152, and 173, the median is 139. When there is an even number of observations, the median is the mean of the two middle numbers. Using a similar example of subject weights, for the following series of weights: 124, 138, 139, 152, 173, and 179, the median is (139 + 152)/2 = 145.5. The median does not have the mathematical niceties of the mean, but it is not as susceptible as the mean to extreme values. If the variable being measured has a distribution that is asymmetric or skewed—that is, if there are a few extreme values at one

407 - 264 - 906162

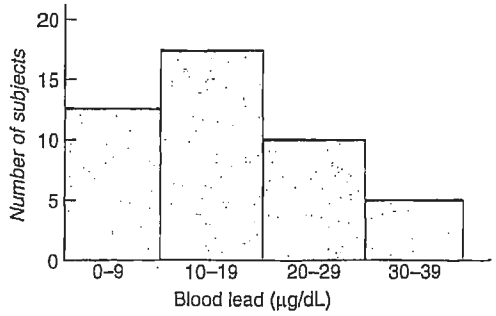


Figure A–1. Frequency distribution of subjects by blood lead category.

end of the distribution—the median is a better descriptor than the mean of the "center" of the distribution.

3. Mode—The mode is the most frequently occurring observation. It is used rarely, except when there are a limited number of possible outcomes.

4. Frequency distribution—In discussing measures of location or spread, we often refer to the frequency distribution of the data. A frequency distribution consists of a series of predetermined intervals (along the horizontal axis) together with the number (or percentage) of observations whose values fall in that interval (along the vertical axis). An example of a frequency distribution is presented in Figure A–1.

B. MEASURES OF VARIATION:

1. Range—The range is the simplest measurement of variation and is defined as the difference between the highest and lowest values. Disadvantages of the range are that it is sensitive to a single extreme value, and it tends to increase in value as the number of observations increases. Furthermore, the range does not provide information about the distribution of values within the set of data. The interquartile range (25–75th percentiles) is sometimes used because it is less influenced by extreme values.

2. Variance—The sample variance (s^2) is a measure of the dispersion about the mean arrived at by calculating the sum of the squared deviations from the mean and dividing by the sample size minus 1. The equation for deriving sample variance is as follows:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance can be thought of as the average of squared deviations from the mean, or more simply, variance tells you how spread out the distribution of the observations is.

3. Standard deviation—The sample standard deviation (s) is equal to the square root of the sample variance. Basically, it tells you how tightly clustered all the observations are around the mean of a set of data.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

See Table A–1 for examples of the calculation of mean, median, mode, variance, and standard deviation.

Variability in data may be a result of the natural distribution of values or of random factors produced by errors in measurement. The variance or standard deviation does not distinguish between different sources of variability.

Table A–1. Calculation of mean, median, mode, variance, and standard deviation ($n = 10$ workers).

x_i = Number of Years of Exposure to Asbestos.

Worker	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1.	$x_1 = 4.0$	−2.2	4.84
2.	$x_2 = 4.5$	−1.7	2.89
3.	$x_3 = 5.0$	−1.2	1.44
4.	$x_4 = 5.0$	−1.2	1.44
5.	$x_5 = 6.0$	−0.2	0.04
6.	$x_6 = 6.5$	+0.3	0.09
7.	$x_7 = 7.0$	+0.8	0.64
8.	$x_8 = 7.5$	+1.3	1.69
9.	$x_9 = 8.0$	+1.8	3.24
10.	$x_{10} = 8.5$	+2.3	5.29
Total:	$\Sigma x_i = 62.0$		$\Sigma(x_i - \bar{x})^2 = 21.6$

Mean: $\bar{x} = \frac{62.0}{10} = 6.2$

Variance $= \Sigma(x_i - \bar{x})^2/(n-1) = 21.6/9 = 2.4$

Standard deviation $= \sqrt{2.4} = 1.55$

Median:

1. Order the observations from lowest to highest.
2. Median $= \frac{1}{2}\left(\left[\frac{n}{2}\right]\text{observation} + \left(\left[\frac{n}{2}\right]+1\right)\right.$ observation $\left.\right) = 1/2$ (5th observation + 6th observation)
3. Therefore, median $= \frac{1}{2}(6.0 + 6.5) = 6.25$

Mode:

Most commonly occurring observation is 5.0, because it occurs twice and all other observations occur once.

Sample versus Population Descriptive Statistics

The descriptive statistics discussed thus far are sample estimates of true population values or parameters. Because we usually do not have the resources to measure the variables of interest on entire populations, we instead select a sample from the population of interest and then estimate the population mean from the sample mean or the population variance from the sample variance. The population mean usually is represented by the Greek letter μ and the population variance by the Greek letter σ^2. One almost never knows the true population values for these parameters and is almost always conducting sample surveys to estimate them.

The Normal Distribution

The most important continuous probability distribution is the normal, or Gaussian, distribution, also known as the *bell-shaped curve.* Many quantitative variables follow a normal distribution, and it plays a central role in statistical tests of hypotheses. Even when one is sampling from a population whose shape departs from the normal distribution, under certain general conditions, it still forms the basis for statistical testing of hypotheses.

We often transform data to make them more normal in distribution. The normal distribution has several nice properties that make it amenable to statistical analysis, and variables that follow a normal distribution are for that reason preferred. For example, in occupational exposure studies, the log dose often is used rather than the actual dose because the log dose more closely approximates a normal distribution. A particular normal distribution is defined by its mean and variance (or standard deviation). Two normal distributions with different means but the same variance will differ in location but not in shape (Figure A–2). Two normal distributions with the same mean but different variances will have the same location but different shapes or "spreads" about the mean value (Figure A–3). Note that the normal distribution is unimodal

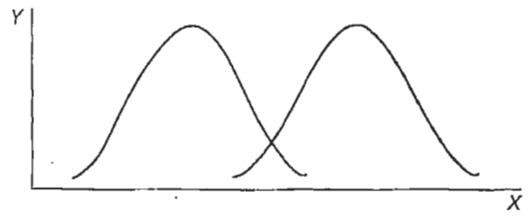


Figure A–2. Two normal distributions with different means but identical standard deviations.

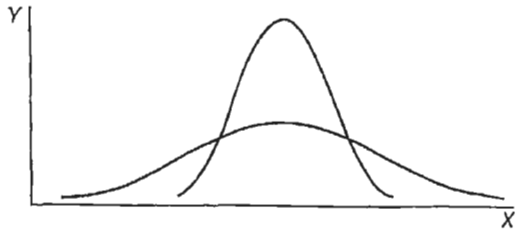


Figure A–3. Two normal distributions with identical means but different standard deviations.

(has one value occurring most frequently), bell-shaped, and symmetric about the mean.

The population encompassed by one standard deviation (σ) on either side of the mean in a normally distributed population will include approximately 67% of the observations in that population (Figure A–4); the population between 2σ on either side of the mean will include approximately 95% of the observations; and that between 3σ on either side of the mean encompasses more than 99% of the observations in the population (see Figure A–4). This property of the normal distribution is particularly useful when a researcher or clinician is trying to identify patients with high or low values in response to a certain test. If one knows the mean for that particular test and has a good estimate of what the standard deviation is, the range within which one would expect (let us say) 95% of patients to fall can be determined, and a patient with values outside this range might need to be examined further.

To use this property of the normal distribution, the sample should be large enough to provide reasonably certain estimates of the mean and standard deviation.

> ***Example I:*** *If the mean hematocrit value in a clinical population is 42% with a standard deviation of 3%—and assuming hematocrit values follow a normal distribution—one would expect 95% of the clinic population to have hematocrit values between 42% ± (2× 3%) or (36, 48)%. A patient falling outside this range could be identified for further testing.*

Another principle relevant to the normal distribution is the central limit theorem, which holds that no

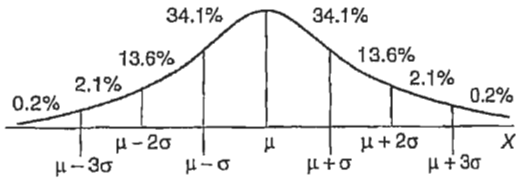


Figure A–4. Standard normal distribution.

matter what the underlying distribution of x, the particular variable of interest, the sample mean ($\bar{x}$) will have a normal distribution if the sample size (n) is large enough. Thus, if $\bar{x}$ itself comes from a population with a mean value μ and population standard deviation s, then $\bar{x}$ (calculated from a sufficiently large sample of size n) will have a normal distribution with the same population mean μ and a smaller population standard deviation equal to $\sigma/\sqrt{n}$. One then can test hypotheses concerning the sample mean $\bar{x}$ because it is known to have a normal distribution, and its mean and standard deviation are also known. The standard deviation of $\bar{x}$ is called the *standard error of the mean* (SEM).

Because one is usually concerned with estimating the true population mean μ from the sample mean $\bar{x}$, it is important to know how good an estimate the sample mean is of the true mean. Every time a sample of size n is selected from the population and $\bar{x}$ is calculated, a different value for $\bar{x}$ will be obtained and thus a different estimate of μ. If this were done over and over again and many $\bar{x}$ values were generated, the $\bar{x}$ values themselves would have a normal distribution centered on with standard deviation equal to $\sigma/\sqrt{n}$. In practice, one does not calculate several $\bar{x}$ values to estimate μ; only one is calculated. The SEM quantifies the certainty with which this one sample mean estimates the population mean. The certainty with which one estimates the population mean increases with sample size, and it can be seen that the standard error decreases as n increases. It also can be seen that the standard error increases as σ increases. This means that the more variability in the underlying population, the more variable will be the estimate of μ. The "true" SEM is $\sigma/\sqrt{n}$, and the sample estimate of the standard error of the mean is $s/\sqrt{n}$, where s is the sample standard deviation. An investigator wanting a more precise estimate of the mean (smaller SEM) could either increase the sample size n or try to decrease σ.

Many investigators summarize the variability in their data with the standard error because it is smaller in value than the standard deviation. However, the standard error does not quantify variability in the population; it quantifies the uncertainty in the estimate of $\bar{x}$, the population mean. An investigator describing the population sampled should use the standard deviation to describe that population. The SEM is used in testing hypotheses about the population mean.

Example II: *Suppose that blood lead is measured in 20 patients. Assume that the sample mean (x) equals 20 μg/dL and that the sample standard deviation (s) equals 5 μg/dL with a sample size (n) of 20. If blood lead has a normal distribution in this sample, one would expect 95% of the population to lie within 2s of the mean. Thus, if the investigator's sample was a representative one, 95% of the population will have blood leads between 20 ± (2 × 5) (i.e., between 10 and 30 μg/dL). These numbers quickly summarize the distribution and give the reader a range against which to compare the reader's own patients. However, investigators often summarize their data with the mean and the standard error of the mean and report, "Blood lead in this sample population was 20 ± [2 × (5/√20)]." This would lead a reader to believe that 95% of blood lead values are expected to fall between 17.8 and 22.2 μg/dL if one did not know the difference between the standard deviation and the standard error of the mean. In reality, 17.8 and 22.2 μg/dL describe a quantity known as the 95% confidence interval for the true mean blood lead ; it does not describe a range of expected values. The reader of the report usually wishes to compare a patient's blood lead with an expected range of values for blood lead, that is, the mean ± 2s.*

INFERENTIAL STATISTICS

In general, there are two steps to be followed in data analysis. The first is to describe the data by using descriptive statistics such as the mean, median, variance, and standard deviation. The second step is to test specific hypotheses that were formulated before conducting the research project. This is done by formulating a null hypothesis and an alternative hypothesis, where the null hypothesis is "no difference exists" and the alternative hypothesis is "difference exists."

An example of a null hypothesis might be, "There is no difference in pulmonary function between groups of underground miners and surface miners." The alternative hypothesis would be, "There is a difference between the two groups."

Once the hypotheses are formulated, the appropriate statistical test can be performed. Some of the most commonly used methods are discussed below.

The Case of Two Groups: The *t*-Test

In many instances an investigator is interested in comparing two groups to determine whether they differ on average for some continuous variable. For example, an investigator might be interested in determining whether exposure to organic solvents has an effect on psychomotor performance such as reaction time. To do this, one would select a sample of a group of industrial painters who are exposed to such solvents and compare their test performances with those of a group of workers not exposed to such solvents. Obviously, even if there are truly no differences between two employee groups in how they perform on such a test, the sample mean test scores probably will be unequal simply because of random fluctuation.

The main question is, "Are the differences larger than one would expect by chance if there truly is no dif-

ference in the reaction times?"—that is, do the samples come from one underlying population, not two? The null hypothesis in this situation is that the true mean reaction time in the painter group equals the true mean reaction time in the nonpainter group.

The alternative hypothesis is that the underlying true means are unequal. This is usually called a *two-sided* alternative hypothesis because we are not specifying the direction of the inequality. In the example, average reaction time in the painter group might be faster or slower than average reaction time in the nonpainter group. Differences in either direction are examined by testing the null hypothesis.

The appropriate statistical test in this situation is the two-sample *t*-test. Two independent samples have been drawn; that is, the individuals in one sample are independent of the individuals in the other. The *t*-test has the following form:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$$

where $\bar{x}_1$ is the sample mean in group 1 and $\bar{x}_2$ is the sample mean in group 2.

Note that the numerator is the difference of sample means, and the denominator is the standard error of this quantity. Dividing by the standard error standardizes the difference in sample means by the variability present in the data. If the difference in the means was very large but the data from which it was calculated were highly variable, the *t*-statistic would reflect this and would be adjusted accordingly.

Use of the *t*-statistic assumes that the two samples have the same underlying population variance s_p^2. Thus a pooled estimate of the variance is calculated and substituted into the *t*-statistic. This pooled estimate s_p^2 has the following form:

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

Therefore, the two-sample *t*-statistic is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_p^2}{n_1}\right) + \left(\frac{s_p^2}{n_2}\right)}}$$

Note that the pooled estimate of the variance is simply a weighted average of the variances from sample 1 and sample 2. Thus, if one sample is much larger than the other, more weight is given to its estimate of σ^2 because it is assumed to be more reliable given that it is based on a larger sample size. Note further that if the two samples are of equal size, the pooled variance is simply the sum of the two sample variances divided by 2. From the format of the *t*-test, one can see that if the two sample means are similar in value, the numerator of *t* will be close to zero—and consequently, the value of *t* would be small—leading to the conclusion that the null hypothesis is true and that there is probably only one underlying distribution from which the two samples come. If one obtains a large value for the *t*-statistic, it is likely that the two samples come from two different underlying distributions, and one therefore would want to reject the null hypothesis.

How large does *t* have to be to reject the null hypothesis? Tables of the t-statistic indicate what value of *t* would cause the null hypothesis to be rejected. Even when the null hypothesis is true and there really is no difference between the groups being compared, there is the possibility that a large value of *t* might occur owing to random chance alone. One would like the probability of this occurrence to be small, that is, less than 5%.

To find the proper cutoff value of *t* (to reject the null hypothesis) for a particular study, it is necessary to know the number of degrees of freedom. The degrees of freedom are equal to $(n_1 + n_2 - 2)$. This may be thought of as the number of observations that are free to vary once the mean is known. Once the degrees of freedom are known, the value of *t* may be obtained from the *t*-table and compared with the *t*-statistic calculated in the study. If the study *t*-statistic is larger than the tabled cutoff value, one can conclude that this is unlikely to have happened under the null hypothesis, which is therefore rejected.

Bear in mind that the alternative hypothesis was the two-sided alternative, meaning that the two group means were simply different but not specifying the direction of the difference. Consequently, in the *t*-table, two cutoff points actually are obtained because both very large negative and very large positive values of *t* are of interest. The *t*-distribution is symmetric, so the two cutoff points are simply $\pm t$. If the study *t*-value is larger than $+t$ or smaller than $-t$, the null hypothesis is rejected.

Example III gives the flavor of the *t*-test and how it is used.

Example III: *Two-sample t-tests. The following tabulation presents the mean change in plasma cholinesterase concentration from baseline levels for 15 pesticide applicators and 14 unexposed controls.*

	N	Mean Decline (%)	Standard Deviation
Applicators	15	25	11
Controls	14	10	8

Do the data present sufficient evidence from which to conclude that the mean decline in cholinesterase is different for the two groups?

The null hypothesis is that there is no difference in cholinesterase change between the two groups. The alternative hypothesis is that there is a difference in cholinesterase change between the two groups.

First calculate s_p^2:

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

$$= \frac{(15-1)11^2+(14-1)8^2}{(15+14-2)}$$

$$= 90.21$$

Substitute into the formula for t:

$$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left(\frac{s_p^2}{n_1}\right)+\left(\frac{s_p^2}{n_2}\right)}}$$

$$= \frac{25-10}{\sqrt{\left(\frac{90.21}{15}\right)+\left(\frac{90.21}{14}\right)}}$$

$$= \frac{15}{\sqrt{12.458}}$$

$$= 4.25$$

Therefore, $t = 4.25$ *and* $df = n_1+n_2-2 = 27$.

The study t-value of 4.25 with 27 degrees of freedom is compared with the tabled t value of ± 2.05, which has a 5% chance of occurring when the null hypothesis is true. Because +4.51 is larger than +2.05, the null hypothesis is rejected; that is, there is a statistically significant difference in the mean change in plasma cholinesterase from baseline between the two study groups. In other words, this difference is unlikely to have occurred by chance.

This result also can be expressed as the confidence interval or maximum range of the true change in cholinesterase. In this case, the 95% confidence interval is 16.5–33.5. Stated another way, the probability is approximately 0.95 that the true mean decline in plasma cholinesterase concentration in the applicators is within the range 16.5–33.5.

Paired *t*-Test

The preceding discussion concerns the two-sample *t*-test and is appropriate for the situation in which two independent groups are being compared. Another common situation occurs when there are paired samples; that is, the two observations are not independent of one another.

For example, suppose that a researcher is measuring change in pulmonary function [e.g., forced expiratory volume in 1 second (FEV_1)] over a work shift and there are 20 subjects in the study (see the example below). The researcher would measure FEV_1 among the subjects before and after the work shift. Clearly, the before and after measurements are not independent, and one would like to take advantage of the fact that all individual (nonexposure) characteristics have been controlled. To do this, the difference in FEV_1 (before – after) is calculated for each subject. Because the difference is the only observation made per subject, the data set now has gone from 40 observations (2 per subject) to 20 observations (1 per subject). If there is no effect of work shift on FEV_1, one would expect the difference in FEV_1 for each subject to be small in value or close to zero. If the null hypothesis is not true and work shift exposure does change FEV_1, the differences will not be close to zero. The *t*-statistic calculated in this situation is known as the *paired t*-statistic and has the following form:

$$t = \frac{\bar{D}}{(s_D/\sqrt{n})}$$

where $\bar{D} = \frac{\Sigma D_i}{n}$ **= average difference and**

s_D **= standard deviation of differences.**

$$= \sqrt{\frac{\sum_{i=1}^{n}(D_i-\bar{D})^2}{n-1}}$$

The appropriate null hypothesis is that the true mean of the differences is zero, and the appropriate alternative hypothesis is that the true mean of the differences is not zero. Again, it is a two-sided alternative, and one is looking for large positive or large negative differences. Small absolute values of the *t*-statistic would indicate that the null hypothesis is probably true, and large absolute values of *t* would lead to rejection of the null hypothesis. One goes to the *t*-table or computer program to determine how large a value of *t* is needed to reject the null hypothesis. To obtain the correct value, one needs to know the appropriate degrees of freedom. In the paired *t* situation, there are $n - 1$ degrees of freedom, or the number of pairs minus one.

Common Errors in Use of the t-Test

EXAMPLE: Paired t-test

A study of painters involved measuring pulmonary function (FEV, liters) at the beginning (A) and end (B) of a work shift. The results were as follows:

Case #	A_1	B_1	$D_1=(A_1-B_1)$	(D_1-D)	$(D_1-D)^2$
1	3.14	3.01	0.13	0.10	0.010
2	2.85	2.80	0.05	0.02	0.000
3	2.50	2.30	0.20	0.17	0.029
4	3.01	3.15	−0.14	−0.17	0.029
5	1.55	1.55	0.00	−0.03	0.001
6	2.21	2.15	0.06	0.03	0.001
7	2.81	2.68	0.13	0.10	0.010
8	3.25	3.34	−0.09	−0.12	0.014
9	2.66	2.56	0.10	−0.07	0.029
10	1.95	1.90	0.05	−0.02	0.000
11	3.50	3.46	0.04	0.01	0.000
12	3.95	4.06	−0.11	−0.14	0.020
13	4.10	3.90	0.20	0.17	0.029
14	3.60	3.56	0.04	0.01	0.000
15	2.80	2.90	−0.10	−0.13	0.017
16	2.50	2.50	0.00	−0.03	0.001
17	2.10	2.16	−0.06	−0.09	0.008
18	3.70	3.61	0.09	0.06	0.004
19	2.92	2.86	0.06	0.03	0.001
20	3.31	3.42	−0.11	−0.14	0.020
			0.54		0.198

$$\bar{D} = \frac{\Sigma D_I}{n} = \frac{0.54}{20} = 0.027$$

$$s_D = \sqrt{\frac{\sum_{I=1}^{n} (D_I - \bar{D})^2}{n-1}}$$

$$= \sqrt{\frac{0.198}{19}} = 0.102$$

$$t = \frac{\bar{D}}{(s_D / \sqrt{n})} = \frac{0.027}{0.102 / \sqrt{20}} = 1.18$$

Compare the calculated t of 1.18 to the tabled t of 2.093. Since the calculated t is less than the t in the table, the null hypothesis (of no change in function over work shift) is not rejected.

A common mistake made with the t-test is known as the *multiple-comparison problem.* The problem arises when an investigator has several groups to compare and proceeds to compare them in groups of two, using the t-test each time. In other words, group 1 is compared against group 2 using the t-test, then group 2 against group 3, then group 1 against group 3, and so on. The problem with proceeding in this fashion is that overall there is *more* than a 5% chance of erroneously rejecting the null hypothesis even though there is only a 5% chance of making this mistake with each individual comparison. This increased probability of making a mistake occurs because multiple tests increase the likelihood that an error will occur. Thus the chance of erroneously rejecting a null hypothesis is greater than the 5% risk of mistakenly rejecting each comparison taken by itself, even if all the hypotheses are true. There are many ways of adjusting for this situation, known as *multiple-comparison procedures.* What is important to remember is that if one does enough of such two-group comparisons, the probability of rejecting the null hypothesis incorrectly at least once increases with the number of such comparisons made and can be quite a bit greater than 5% unless the investigator uses an appropriate adjustment for multiple comparisons.

Analysis of Variance (ANOVA)

When the variables under study are continuous in nature and there are more than two groups being studied, the investigator usually is concerned with whether the means in the groups are different from one another. An appropriate statistical method to answer this question is to use *analysis of variance* (ANOVA).

Suppose that one were studying three groups of workers occupationally exposed to three different gases. One might want to test whether the particular gases affect mean FEV_1 levels differently in the three groups. In this example, individual FEV_1 values would be adjusted for nonexposure determinants (i.e., age, gender, height, or race). The null hypothesis is that the group means for FEV_1 are equal, that is, that a particular exposure has no effect on FEV_1 values. Obviously, there will be differences between the sample means in each group owing to random fluctuations in FEV_1 among individuals.

Are the differences observed in the sample means merely a result of random fluctuations, or are they a consequence of true differences in FEV_1 caused by the gas exposures? To answer this question, one examines whether the data are consistent with the assumption that the gas exposure has no effect and that the three groups are really random samples from the same underlying population. The null hypothesis assumes that any observed differences in the sample means and standard deviations are due simply to random sampling. ANOVA tests this null hypothesis by estimating the true population variance in two different ways and comparing these two estimates of the variance. If the three samples do indeed come from the same underlying population, these two estimates of the variance will be very close in value. If the three samples do not all come from the same underlying population, these two

estimates will be further apart in value, and this variation is what one hopes to detect.

Certain statistical assumptions are made when an ANOVA test is performed on a set of data: (1) It is assumed that groups have been randomly assigned to receive the treatment or exposure and that the groups are independent; (2) the underlying variance (σ^2) in each group is assumed to be identical (even though the true group means may be different and the sample variances may differ slightly); and (3) the random variable under study—for example, FEV_1—has a normal distribution.

Conceptually, the method of ANOVA proceeds as follows: Once the null hypothesis is formulated, the sample variance (s^2) is computed within each exposure group, and each of these s^2 estimates is unaffected by differences among the group means. These s^2 estimates are averaged to obtain one "within group" variance estimate. The values of the individual exposure group means then are used to arrive at a second "between group" variance estimate of σ^2. In this "between group" estimate of σ^2, differences (or variability) among the group means will affect the overall estimate of σ^2. For example, if a particular gas exposure has no effect on FEV_1, both estimates of σ^2 should be similar. To test the null hypothesis, a statistic known as the F statistic is calculated. The value of F is simply the ratio of the "between group" variance estimate to the "within group" variance estimate. Because both numbers estimate the same parameter (σ^2), if the null hypothesis is true, the value of F should be close to 1. If F is significantly larger than 1, you should reject the null hypothesis and conclude that the exposure groups are different with regard to FEV_1.

How does one determine how large F must be in order to reject the null hypothesis? Because of random fluctuations in the data, it is possible that a large F statistic might result even when the null hypothesis is true. However, one would like the chance of this happening to be very small. Tables of the F statistic are available to assist the investigator in selecting a value of F against which the F statistic calculated from the data can be compared. The tabled value of F is one that would occur less than approximately 5% of the time if the null hypothesis were true. If the F statistic calculated from the researcher's data is larger than the one found in the table, the results are less than 5% likely to have occurred by random chance, even if the null hypothesis (no difference in sample groups) is true. Because the observed results therefore are very unlikely to have happened by chance under the null hypothesis, the researcher is justified in rejecting the null hypothesis and saying that there is a difference among the groups. The 5% cutoff point is an arbitrary one, and depending on the individual situation, one could set the cutoff at one or 10%; however, the conventional cutoff point is 5%.

When one is studying more than two groups and the data involved are continuous (e.g., FEV_1 or blood lead concentration) and the question of interest is whether the groups all come from the same underlying population—that is, have the same mean for the variable of interest—ANOVA is the most appropriate method to use for initial testing of the null hypothesis. If one fails to reject the null hypothesis with the F statistic, no further tests of the null hypothesis are necessary. There are no differences among groups. On the other hand, if one performs ANOVA on the data and rejects the null hypothesis, then differences in the outcome (FEV_1 or blood lead level) among the study groups associated with the particular exposure may exist. One then can use multiple-comparison tests to identify exactly which group or groups are significantly different.

This is a simplified discussion of ANOVA meant only to introduce the concept of this important statistical method. We have not provided enough details for the reader to be able to perform this test accurately. The purpose is to identify situations in which ANOVA is appropriate as an initial analytic procedure (see References).

Analyzing Rates & Proportions: The Chi-Square Test

In preceding sections we described methods of analysis for continuous types of data. This section begins a discussion of the analysis of categorical data. The following table of cigarette smoking history and lung cancer cases and controls (persons without cancer) presents an example of categorical data.

	Lung Cancer	Controls
Cigarette smokers	450	225
Nonsmokers	20	225
Total	470	450

It is immediately apparent, without doing any statistical tests, that there is an association of cigarette smoking and lung cancer. The row variable, cigarette smoking, is associated with the column variable, lung cancer. A simple calculation of the proportions of lung cancer cases and control cases who smoked confirms this association. Of the lung cancer cases, 450/470 = 95.7% smoked cigarettes, whereas 225/450 = 50% of the controls smoked cigarettes.

However, suppose that the table was of mesothelioma (a very rare type of lung cancer) and cigarette smoking, and the following results were obtained:

	Mesothelioma	Controls
Cigarette smokers	80	200
Nonsmokers	40	104
Total	120	304

In this example, the ratio of cigarette smokers to nonsmokers among the mesothelioma cases (80/120 = 66.6%) and the controls (200/304 = 65.8%) is nearly the same, with approximately twice as many smokers as nonsmokers for both the case and control groups. In this case, one would say that there is no association between the column variable (mesothelioma) and the row variable (cigarette smoking). The null hypothesis in this example would be that there is no association between mesothelioma and cigarette smoking, and one could not reject the null hypothesis owing to the similarity of the proportions of smokers in the mesothelioma and the control groups.

Most situations with categorical data are not as clear-cut as these two examples. In most cases, one cannot simply "eyeball" the data to determine whether the two variables are independent or not. The statistical test one uses to determine whether or not there is an association in such data is known as the *chi-square test*. Example IV is a situation in which the chi-square test is applied.

Example IV: *Three groups of farm workers are studied for the occurrence of new skin rashes during the growing season. The three groups are involved in growing and harvesting (1) grapes, (2) citrus crops, and (3) tomatoes. The workers are followed for the growing season, and the occurrence of new rashes in the three groups is compared to determine if there is an association between exposure (crop) and outcome (rash).*

Crop 1, N = 100
Crop 2, N = 200
Crop 3, N = 200

	Exposure (Crop)			
Response	1	2	3	Total
Rash	30	40	32	102
No rash	70	160	168	398
Total	100	200	200	500

The null hypothesis in this situation is again the hypothesis of "no difference," only it is phrased as no association between the row variable (rash) and the column variable (crop).

One can quickly compute from the table that the percentage working on crop 1 with a rash is 30/100 = 30%; on crop 2, it is 40/20 = 20%; and on crop 3, it is 32/200 = 16%. By just quickly observing the data, one might think that crop 1 is different from crops 2 and 3. However, the null hypothesis is that there is no association between crop worked and rash development. Thus the question is whether the observed differences in response are simply a result of random variation in the data or are larger than one would expect by chance alone if the null hypothesis were true. To test this, a chi-square statistic is calculated. As with the *t*-test and *F*-test, one determines whether this chi-square value is unlikely to have occurred by chance alone under the null hypothesis. The calculation of the chi-square involves first determining an "expected" value for each cell in the table. The expected value is the value one would "expect" to see in the cell if there were no association between row (rash) and column (crop exposure) variables, that is, that value one would "expect" to see if the null hypothesis were true. The expected value is obtained as follows.

According to the null hypothesis, we would expect the same proportion to develop a rash in each group. If this is true, the best estimate of the expected proportion with rashes in each exposure group comes from the overall information given by the total number of workers with rashes divided by the total number of workers in the study; that would be 102/500 = 0.204. Then, for crop 1, one expects that 0.204 of the 100 people in crop exposure group 1 will develop rashes, that is, 20.4 people; for crop 2, one expects that 0.204 of the 200 people working with crop 2 will develop rashes, that is, 40.8 people; and for crop 3, one expects that 0.204 of the 200 people will develop rashes, that is, 40.8 people. In other words, because under the null hypothesis there is no association between exposure and percentage developing a rash, one expects the same percentage to respond favorably (or unfavorably) in each group. The expected proportion of workers not developing rashes is obtained in the same manner. The best estimate of the proportion not developing a rash in each group is the total number not developing a rash divided by the total number of workers, which equals 398/500 = 0.796. This gives an expected frequency of 100 × 0.796 = 79.6 working with crop 1 not developing rashes, 159.2 working with crop 2 not developing rashes, and 159.2 working with crop 3 not developing rashes. Putting the expected values in parentheses alongside the observed values, the table now looks like this:

	Exposure (Crop)			
Response	1	2	3	Total
Rash	30 (20.4)	40 (40.8)	32 (40.8)	102
No rash	70 (79.6)	160 (159.2)	168 (159.2)	398
Total	100	200	200	500

To test the null hypothesis, one looks at the observed and expected numbers in each cell to see how close together the two values are. If the values are close together, one may decide that the null hypothesis is true. If they are very different, one may decide that the null hypothesis is not true. To decide whether the

observed and expected values are close together, the chi-square statistic is calculated. It has the following form:

$$\chi^2 = \sum_{I=1}^{n}\left[\frac{(O_I - E_I)^2}{E_I}\right]$$

where E_i is the expected value in cell i, O_i is the observed value in cell i, $i = 1, \ldots, n$, and n is the number of cells in the table.

Large chi-square values indicate a lack of agreement between observed and expected values; small chi-square values indicate close agreement.

How does one determine what constitutes a large chi-square value? As in the preceding discussions about *t*- and *F*-tests for continuous data, one consults a table of chi-square values. The table identifies the chi-square value that would occur less than 5% of the time if the null hypothesis (no association) were true, and this is compared with the study chi-square value. If the study-chi square is larger than the table cutoff value, the null hypothesis is rejected because this is known to occur less than 5% of the time when the null hypothesis is true. If the study chi-square value is smaller than the table cut-off value, the null hypothesis is not rejected. Alternatively, one could calculate the exact probability, or *P* value, of the study chi-square statistic. To use the chi-square tables, the degrees of freedom are needed to select the proper value from the table. The degrees of freedom in the chi-square situation are equal to (number of rows − 1) × (number of columns − 1). When there are two rows and three columns in a table, the degree of freedom is (2 − 1) × (3 − 1), which equals 2 degrees of freedom. One thing to remember is that the chi-square statistic works only when the sample is sufficiently large. A rule of thumb is that the chi-square test yields good results when the expected values in each cell are greater than or equal to 5.

Calculating the chi-square statistic for the preceding example, the following results are obtained:

$$\begin{aligned}\chi^2 &= \frac{(70-79.6)^2}{79.6} + \frac{(160-159.2)^2}{159.2} \\ &+ \frac{(168-159.2)^2}{159.2} + \frac{(30-20.4)^2}{20.4} \\ &+ \frac{(40-40.8)^2}{40.8} + \frac{(32-40.8)^2}{40.8} \\ &= 8.08\end{aligned}$$

The tabled value of chi-square to which the calculated value is compared is 5.99. Because 8.08 is larger than 5.99, the null hypothesis is rejected.

Calculating the chi-square statistic is only one method for analyzing categorical data. It is, however, one of the most common statistical tests found in the medical literature.

The *P*-Value & Statistical Significance

An important quantity in all statistical hypothesis tests is the *P*-value. The *P*-value is the probability of observing a particular study result (e.g., *t*-statistic calculated from study data) by chance alone when the null hypothesis is really true. In the examples thus far, the *P*-value of the test statistic actually has been used without calculating its exact value. The procedure has been to calculate, for example, a *t*-statistic from the study data. A computer program then would compare the *t*-statistic observed with the *t*-statistic known to have a *P*-value of 5%.

If the value of the *t*-statistic computed for the sample is smaller than the 5% value, the null hypothesis is not rejected. When the computed sample *t*-statistic has a value larger than the 5% *P*-value, the null hypothesis is rejected. The exact *P*-value of the sample *t*-statistic also can be obtained from tabulated values so that one can report *P*-values less than other cutoff values, for example, 1% ($P < 0.01$). When the *P*-value is less than 5%, the result is commonly referred to as being *statistically significant.* However, statistical significance may not be the same as clinical or public health significance because the former is affected by the size of the study population and may reflect differences that have no biologic importance.

Another way to express the statistical significance of an observed result is the confidence interval (CI). The CI provides a range and the probability that this range includes the true population mean. For example, a 95% CI is calculated as the sample mean plus or minus two standard errors of the mean. The CI is interpreted as giving a 95% probability of including the true population mean. A 99% CI is the sample mean plus or minus three standard errors of the mean.

It should be noted that the width of the CI will decrease as the sample size increases; that is, we are more confident of knowing the true population mean when it is estimated from a larger sample. The degree of certainty is also inversely related to the width of the confidence interval. For example, we can be more precise (narrower CI) in estimating the 95% CI than the 99% CI for the same sample size.

The CI is generally preferred to the *P*-value because it gives the range of values observed with a select level of statistical confidence (e.g., 95%) and not just a single determination of whether the observed result is above or below the *P*-value.

The researcher in a typical study is interested in comparing an exposed group to a control group and using the observed difference in proportions or mean values to estimate the effect of the exposure. For example, let's say one is interested in determining delta (δ),

where δ equals the true mean value of sperm concentration among workers exposed to heavy metals minus the true mean value of sperm concentration in unexposed workers. One then wishes to test whether $\delta = 0$; that is, one may wish to determine whether the (true) proportion with disease from one exposure is equal to the (true) proportion with disease under a second exposure or control. One can then calculate δ as the difference between these two proportions, again testing to see whether $\delta = 0$.

Even if the treatment and control groups in the study are truly being sampled from one underlying population (i.e., if there is no real difference between treatment and control), some differences between the two groups will occur by chance alone. If the observed difference in sample means or proportions has a small probability of occurring by chance alone (assuming no true underlying difference), then the null hypothesis that $\delta = 0$ is rejected. The "rule" for deciding how small that probability has to be before rejecting the null hypothesis is known as the *level of significance* of the statistical test and is designated as alpha (α).

Thus the procedure in a typical study is to formulate a null hypothesis (H_0), and usually,

$$H_0: \mu_1 = \mu_2$$
$$\text{also written as } H_0: \delta = \mu_1 - \mu_2 = 0$$

for example, H_0: mean sperm concentration with exposure 1 (heavy metals) = mean sperm concentration with exposure 2 (no exposure), or

$$H_0: p_1 - p_2 = 0$$
$$\text{also written as } H_0: \delta = p_1 - p_2 = 0$$

in other words H_0: proportion with disease in exposure (P_1) = proportion with disease in exposure (P_2).

The (two-sided) alternative hypothesis is

$$H_A: \mu_1 \neq \mu_2$$
$$\text{also written as } H_A: \delta = \mu_1 - \mu_2 \neq 0$$

that is, H_A: the mean sperm concentrations are not equal under treatments 1 and 2, or

$$H_A: p_1 \neq p_2$$
$$\text{also written as } H_A: \delta = p_1 - p_2 \neq 0$$

that is, H_A: the proportions with disease are not equal under treatments 1 and 2 ($P_1 = P_2$).

After completion of the study, sample estimates of μ (or p) are calculated for the two exposure groups. The probability is calculated that a difference as large as the one observed in the study would occur if the null hypothesis were true. This probability is the *P*-value of the test. If the *P*-value is less than α (the significance level), the null hypothesis is rejected. If the *P*-value is not less than α, the null hypothesis is not rejected. A CI also can be calculated for proportions as it can be for means, and one can determine the probability that the true proportion is within the calculated CI.

THE TYPES OF MISTAKES ONE CAN MAKE IN DOING A RESEARCH STUDY

There are two main categories of errors one can make in deriving inferences from a typical research study. They are known as *type I* and *type II errors.*

Type I Error

A type I error occurs if one decides to reject the null hypothesis and declare the two groups different when in fact they really are from the same underlying population. Type I error is equal to the significance level α, and the significance level must be established before the study is conducted. Thus α equals the probability that one will reject the null hypothesis when the null hypothesis is true, that is, when the investigator decides what chance of making this kind of mistake is acceptable and sets the α level accordingly. For example, an investigator may decide that it is extremely important not to declare that a disease (e.g., cancer) is associated with an exposure unless there is overwhelming evidence of an association from the study. In this case, the α level might be set at 1% instead of 5%, where 5% is the value for α used in most studies.

Type II Error

A type II error occurs if a researcher decides not to reject the null hypothesis when, in fact, there is a difference between the two groups; that is, a true difference between the two groups has been missed. Type II error is usually designated by β.

In a research study, the type II error is not a single value. If the null hypothesis is false, this means that outcomes seen in the exposure group are not equivalent to those seen in the control group; that is, δ is not equal to 0. There are an infinite number of values that this difference could take on. For each value of the difference δ between the exposed and control groups, there is a different value for the type II error. If one is interested in determining the probability that one would miss a true difference between exposure and control groups, the exact value of the difference being examined must be specified. Once this is done, the probability that one would fail to reject the null hypothesis given the true nonzero difference between the two groups can be calculated.

The Power of a Study

One of the most important quantities calculated for a research study is the power of a particular study. The power is the probability that one will correctly reject the null hypothesis when the null hypothesis is truly false. In other words, the power is the probability of correctly recognizing a true difference between the two groups. The power of a study is actually the complement of the type II error β, that is, power = $1 - \beta$. Thus the power of a study is different for every different value of β that occurs. To calculate the power, one must specify a particular alternative. Power is particularly important when one is evaluating a negative study—a study that finds no difference between the groups.

Suppose that the power of a specific study is 40%. This means that the researcher has only a 40% chance of discerning that a true difference exists between the exposure groups. Therefore, if no difference between exposure groups is found and the power of the study is reported as 40%, a reader might wonder whether that particular study had any real chance of finding a difference between exposures even if the exposures were truly associated with the different outcomes. In practice, it is much more common to use 80% or 90% for the power of a study so that you have a reasonably good probability of detecting a difference between exposures if one truly exists.

The power of a statistical test is determined or affected by three quantities: (1) the magnitude of the type I error α, (2) the size of the exposure effect δ the researcher is interested in detecting, and (3) the sample size of the study. Quantities (1) and (2) can be used to estimate the sample size needed in a study for a specified study power.

As the size of the type I error becomes smaller, the power of the study likewise becomes smaller. Remember, the type I error is the probability of incorrectly declaring a difference when none actually exists. As it becomes less likely to make this mistake (i.e., α is smaller), it becomes less likely the null hypothesis will be rejected in general, and power involves correctly rejecting the null hypothesis.

When a study is set up to look for a very large exposure effect δ, it is relatively easy to detect this large effect, and the chances are great that the null hypothesis will be correctly rejected. The opposite occurs when one is looking for a very small δ. Thus power increases as δ increases.

As sample size increases, the variability of the measure of exposure effect decreases. Consequently, the test statistic increases in value, making it easier to exceed the cutoff point for rejecting the null hypothesis. This increases the chances of correctly rejecting the null hypothesis, and so power increases as sample size increases.

A handy table for remembering the quantities discussed in this section is shown below:

	H_0 true (no difference)	H_0 true (difference exists)
H_0 study (declare no difference)	Correct decision	Type II error β
H_0 reject (declare a difference)	Type I error α	Power $1-\beta$

REFERENCES

Centers for Disease Control and Prevention: www.cdc.gov/publications.htm (free software download: epi info, epi map).

Minitab: www.minitab.com (a general statistical program, used for teaching and research; good graphics; PC and Mac).

Stata (Stata Corporation): www.stata.com (general purpose statistical software; PC and Mac).

Statistics.com: www.statistics.com (free software, commercial products, and Web-based resources).

University of Glasgow Department of Statistics: www.stats.gla.ac.uk/cti/links_stats/software.html.

II. EPIDEMIOLOGY

Epidemiology is the study of the distribution and determinants of health- and disease-related conditions in populations. It is concerned with both epidemic (excess of normal expectancy) and endemic (always present) conditions.

The basic premise of epidemiology is that disease is not randomly distributed across populations. Not only is it important to know what sort of disease a particular person has, but it is also necessaty to know what sort of person has a particular disease. While the practice of much of occupational medicine is concerned with the pathogenesis (development) of disease and the treatment of individuals with diseases, the focus of occupational epidemiology is on groups of individuals—with or without diseases—in an attempt to infer the causes that precede specific disease conditions and to determine what occupational or other lifestyle factors can be manipulated to eliminate specific diseases or reduce the prevalence of the disease.

There are three major types of epidemiologic studies: descriptive, analytic, and experimental.

Descriptive epidemiologic studies characterize person, place, and time: (1) Person: What are the characteristics of people who get a particular disease (e.g., age, race, gender, occupation, socioeconomic status, immune status)? (2) Place: Where do they live, work, or travel (e.g., international, national, and local comparisons; urban

Table A–2. Measures of mortality.

$$\text{Crude death rate} = \frac{\text{Number of deaths in year (all causes)}}{\text{Total population}} \times 1000$$

e.g., US 1977 = 8.8 ÷ 1000 population or 878.1 ÷ 100,000 population

$$\text{Cause-specific death rate} = \frac{\text{Number of deaths from specific cause in year}}{\text{Total population}} \times 100{,}000$$

e.g., cancer in US 1977 = 178.7 ÷ 100,000 population

$$\text{Age-specific death rate} = \frac{\text{Number of deaths among persons of specified age group in year}}{\text{Population in specified age group}} \times 100{,}000$$

e.g., cancer in age group 1–14 years = 4.9 ÷ 100,000

$$\text{Infant mortality rate} = \frac{\text{Number of deaths among children younger than 1 year of age in year}}{\text{Number of births in year}} \times 1000$$

e.g., US 1977 = 14.1 ÷ 100,000 live births (12.3 for whites; 21.7 for blacks and others)

versus rural populations; climate; altitude)? (3) Time: When does the illness occur (e.g., temporal variation, seasonal fluctuations)? Descriptive studies are not used to test hypotheses but nevertheless are powerful tools for characterizing disease distributions and associations.

Analytic studies attempt to determine the etiologic factors associated with a disease by calculating estimates of risk: (1) What exposures do people with the disease have in common (e.g., smoking, exogenous hormone use, diet, exposure to radiation or asbestos)? (2) How much is disease risk increased by such exposures (using relative risk as the measure of excess risk? (3) How many cases could be avoided if the exposure were eliminated (using attributable risk as the appropriate measure)? Analytic studies involve testing specific hypotheses.

Experimental studies involve a search for strategies for altering the natural history of disease. Examples of experimental studies are intervention trials to reduce risk factors, screening studies aimed at identifying the early stages of disease, and clinical trials of different treatment modalities to improve prognosis.

MORTALITY & MORBIDITY

The two basic measures of disease in a population are mortality (death) rates and morbidity (disease) rates. Table A–2 provides examples of different types of mortality rates and how each is calculated. Morbidity is measured by calculating either prevalence or incidence rates. *Prevalence* is the number of existing cases of a disease at a given time divided by the population at risk for that disease at that time. This result is commonly multiplied by 100,000 to derive the prevalence rate per 100,000 population.

For purposes of etiology, the *incidence rate* is a more important measure of morbidity and is equal to the number of new cases of a disease occurring over a defined interval divided by the midinterval population at risk for that disease (multiplied by 100,000).

While worldwide mortality data are available—at various degrees of precision depending on the quality of death registration systems—incidence rates can be calculated only for those diseases for which there are population-based registries or for which special studies have been conducted. The National Cancer Institute has a program of cancer registries around the United States that provides information on cancer incidence covering approximately 10% of the U.S. population. Accurate enumeration of the population at risk—available from Census data—is vital for deriving valid estimates of both mortality and morbidity rates. Rates can be specific to any subgroup of interest, defined by age, gender, race, or

other characteristics. For example, the age-adjusted incidence rate for cervical cancer among white women in the United States was 8.7 per 100,000, compared with 11.1 per 100,000 among black women and 15.8 per 100,000 among Hispanic women. One must remember that in calculating a rate, the events in the numerator must be drawn from the population specified in the denominator; that is, those in the denominator must be at risk for the disease. Thus, for cervical cancer, men would not be included in the denominator.

Some problems to keep in mind about current disease data sources include the following:

1. The only complete cause-specific disease registry is for deaths, and the cause-of-death assignment on the death certificate is often inaccurate. In addition, for a disease whose case-fatality ratio is low (i.e., a disease unlikely to result in death when it occurs), the death rate is a gross underestimate of the incidence of the condition in the community. An example of this is nonmelanoma skin cancer, which has a high incidence but low mortality rate.
2. Morbidity reports, even when legally mandated, as is the case for certain infectious diseases (e.g., tuberculosis and sexually transmitted diseases), often are incomplete because of severe underreporting.
3. Complete and accurate population-based morbidity registries are limited in geographic coverage.

ADJUSTMENT OF RATES

In attempting to compare disease rates across population groups or assessing changes in rates over time, the effect of differential age distributions in two populations whose rates are being compared should be taken into account. Disease risk almost always is a function of age; differences in crude rates (i.e., rates not adjusted for age) across populations may reflect age differences rather than differences in occupational or environmental factors of interest.

Age-specific rates are not subject to this drawback, provided the range in each age group, or age stratum, is relatively narrow. It is cumbersome, however, to compare rates among populations across many age strata. Age adjustment or standardization provides a summary measure of disease risk for an entire population that is not influenced by variations in age distribution.

There are two methods for age adjustment: a direct method, which applies observed age-specific rates of death or disease to a standard population, and an indirect method, which applies age-specific rates of death or disease from a standard population to the age distribution of an observed population. In discussing the methods for adjusting rates, cancer will be used as the disease of interest.

The direct method of age adjustment is appropriate when each of the populations being compared is large enough to yield stable age-specific rates. For example, the direct method is used for comparison of cancer rates over time in the United States. Crude mortality rates showing a dramatic increase in cancer over the past few decades would seem to provide strong evidence of a cancer epidemic. It needs to be ascertained, however, to what extent the aging of the country's population has contributed to the apparent epidemic or to what extent other factors, such as an increase in cancer-causing agents in the environment, might be responsible.

The first three columns of Table A–3 show the actual age distributions of the U.S. population in 1940 and 1970, the percentage of the population in each group in the two periods, the corresponding number of actual cancer deaths, and the age-specific death rates. Crude death rates per 100,000 population were 120.2 for 1940 and 163.2 for 1970, an increase of more than 30%. Comparison of the age-specific rates, however, shows only minor increases between the two time periods. It should be noted that the percentage of the population in all age groups over 40 was higher in 1970 than in 1940.

To remove the variable effect of age using the direct method of adjustment, a "standard" population is chosen. The number of people in each age group of the standard population then is multiplied by the appropriate age-specific rate in each of the study populations. This generates the number of deaths or cases of disease one would expect in each age group if the populations had similar age distributions. The expected number of deaths or disease cases then is summed over all age groups, the sum is divided by the total standard population, and the result is multiplied by 100,000. The choice of a standard population is arbitrary; it might be the combined population of the two groups whose rates are being compared, only one of those populations, or any other population.

In our example, the standard was the combined population of the United States in 1940 and 1970, shown in column 5 of Table A–3. The age-specific rates for each period (column 4) were applied for each age group to the standard population, yielding the expected number of deaths shown in column 6. Age-adjusted rates then are calculated by dividing the sum of expected deaths for each period by the total standard population. The resulting adjusted rates are 139.8 per 100,000 for 1940 and 149.9 per 100,000 for 1970. Thus the magnitude of the increase in the crude rates has been reduced from about 30% to 7%. It can be concluded that age is an important factor in the increased cancer rates in the United States, although age alone does not entirely explain changes over time.

Table A–3. Age adjustment by direct method, using cancer mortality data for the United States, 1940 and 1970.

	Actual Population		Number of Cancer Deaths	Age-Specific Death Rates Per 100,000	Standard Population	Expected Number of Cancer Deaths
Age Group	(1)	(2)	(3)	(4)	(5)	(6)
1940						
< 40	87,737,829	66.7	10,283	11.72	217,093,330	25,443
40–49	17,053,068	13.0	18,071	105.97	41,149,961	43,607
50–59	13,100,511	10.0	33,279	254.03	34,177,557	86,821
60–69	8,534,997	6.5	43,686	511.85	24,143,606	123,579
70–79	4,073,514	3.1	38,160	936.78	13,352,179	125,080
80+	1,139,143	0.9	14,721	1,292.29	4,934,355	63,766
Totals	131,639,062	100.0	158,200[2]		334,850,988	468,296[2]
1970						
< 40	129,355,501	63.7	16,096	12.44	217,093,330	27,006
40–49	24,096,893	11.9	26,075	108.21	41,149,961	44,528
50–59	21,077,046	10.4	61,143	290.09	34,177,557	99,146
60–69	15,608,609	7.7	90,099	577.24	24,143,606	139,367
70–79	9,278,665	4.6	88,826	957.31	13,352,179	127,821
80+	3,795,212	1.9	49,333	1,299.87	4,934,355	64,140
Totals	203,211,926	100.0	331,572[1]		334,850,988	502,008[1]

[1]Crude death rate = [sum of column 3 + sum of column 1] × 10^5 = 163.2 per 100,000 population. Age-adjusted death rate = [sum of column 6 + sum of column 5] × 10^5 = 149.9 per 100,000 population.
[2]Crude death rate = [sum of column 3 + sum of column 1] × 10^5 = 120.2 per 100,000 population. Age-adjusted death rate = [sum of column 6 + sum of column 5] × 10^5 = 139.8 per 100,000 population.

When the group of interest is relatively small and thus likely to have unstable age-specific rates, it is more appropriate to use the indirect than the direct method of age adjustment. This is commonly the situation with investigation of cause-specific mortality in an occupational cohort. The indirect method is employed frequently to compare the cancer incidence or follow-up experience of a study group with that expected based on the experience of a larger population or patient series. With the indirect method, the age-specific rates from a standard population are multiplied by the number of person-years at risk in each group in the study series. The number of observed deaths then is compared with the number expected by means of a ratio.

The standardized mortality ratio (SMR) is an example of indirect standardization. In calculating an SMR, the age-specific rates from a standard population (e.g., county, state, or country) are multiplied by the person-years at risk in the study population (e.g., industry employees) to give the expected number of deaths. The observed number of deaths divided by the expected number (times 100) is the SMR (see the example in Table A–4). An SMR also may control for time-specific mortality rates by indirect standardization.

Thus the equation for an SMR is as follows:

$$\text{SMR} = \left[\frac{\Sigma a_i}{\Sigma E(a_i)}\right] \times 100$$

$$= \left[\frac{\textit{Observed}}{\textit{Expected}}\right] \times 100$$

where a_i is the number of people with a specific cause of death in the ith stratum of age, and $E(a_i)$ is the expected number of deaths based on the age-specific rates in the reference population.

The result is multiplied by 100, so when observed deaths equal expected deaths, the SMR is 100, and the differences from 100 represent the percentage difference in mortality in the study population compared with that of the reference population.

Indirect standardization also may be used to adjust incidence rates for age or other factors. Thus incident cases of a disease within a workplace could be expressed as the standardized incidence ratio (SIR), as follows:

$$\text{SIR} = \left[\frac{\textit{Observed number of new cases}}{\textit{Expected number of new cases}}\right] \times 100$$

Table A–4. Age adjustment by indirect method in computation of standardized mortality ratio (SMR).

Age (Years)	Observed Deaths (1)	Person Years (2)	US Population Rates (per 10^5) (3)	Expected Deaths = (2) × (3)
20–29	1	5,000	20.6	0.1
30–39	0	15,000	22.7	0.3
40–49	4	60,000	45.3	2.7
50–59	2	40,000	94.3	3.8
60–69	12	70,000	224.4	15.7
	Σ Obs = 19			Σ Exp = 22.6

SMR = [Σ Obs/Σ Exp] × 100 = [19/22.6]100 = 84

Although it is most common to adjust rates for age and time, the direct and indirect methods of adjustment can be used to adjust for population differences in other factors as well, such as gender, race, socioeconomic status (SES), and stage of disease.

Design Strategies for Analytic & Experimental Studies

Descriptive epidemiology provides disease rates for different groups. It identifies segments of the population—by age, gender, occupation, marital status, geographic area of residence, or other parameters—whose unique experience suggests etiologic hypotheses worthy of pursuit through rigorous analytic studies. Descriptive epidemiology tells who gets the disease where and when and is the basis of analytic epidemiology, which, in turn, focuses on specific questions, such as the following:

- What exposure do people with the disease have in common as compared with people without the disease?
- Why does exposure induce or promote disease?
- How much is disease risk increased by such exposure?
- How many cases might be avoided were the exposure eliminated?

The last question addresses the ultimate objective of epidemiologic research: to identify risk factors so that intervention might either prevent the occurrence of the disease (primary prevention) or lead to early detection (secondary prevention).

The three basic strategies for analytic epidemiology are (1) the cohort study, (2) the case-control study, and (3) the experimental study (clinical trial).

Cohort and case-control studies are observational: The investigator does not control exposure or modify behavior of the study subjects. In the experimental study, the investigator intervenes by introducing treatment or other exposures to study their impact on the disease experience.

TYPES OF EPIDEMIOLOGIC STUDIES

1. The Cohort Study

In the design of a cohort study, a disease-free group of individuals (a cohort) characterized by a common experience or exposure of interest is identified and followed forward over time, or prospectively, to determine whether disease occurs at a rate different from that in a cohort without the exposure. The relative risk (RR) of disease associated with the exposure then can be calculated:

$$RR = \frac{\textit{Incidence rate in the exposed group}}{\textit{Incidence rate in the nonexposed group}}$$

A frequently cited example of the prospective cohort design is the follow-up study of British physicians whose smoking habits were ascertained by means of a mailed questionnaire. The doctors were grouped according to smoking habits, and their deaths were subsequently monitored. Lung cancer rates for those exposed to various levels of smoking then were compared with the rates for nonsmokers by means of the relative risk. Other examples of cohort studies include investigations of long-term cancer incidence among atomic bomb survivors exposed to varying degrees of radiation and deaths among British coal miners.

Theoretically, the prospective cohort study is ideal because the hypothesized cause or exposure precedes the effect or disease. It is also valuable because disease rates and relative risks can be calculated directly, provided that a suitable comparison group is built into the study or is otherwise available for calculation of rates in the nonexposed population. In addition, the exposure of interest can be recorded accurately at the time of exposure; it is not based on recall of past events. This approach has been popular in occupational studies in which the disease experience of workers exposed to putatively hazardous substances has been compared with that of other workers without the exposure or compared with that of the general population.

In practice, however, because of the expense, the time involved, and the number of subjects required, the model prospective cohort study is relatively rare. To avoid some of these constraints, a historical cohort study might be done, whereby a group of persons who in the past experienced an exposure of interest is identified, and their disease record up to the present is investigated. An example is the follow-up of mortality among insulation workers exposed to asbestos. The population of union insulation workers in the 1940s was identified, and their cause-specific mortality rates through the 1970s were determined. Mortality rates for lung cancer and other causes in this

Table A–5. Presentation of data from a cohort study.

		Disease		
		Present	Absent	
Exposure	Yes	*a*	*b*	*a* + *b*
	No	*c*	*d*	*c* + *d*

population were tabulated and compared with those expected on the basis of mortality rates for all U.S. men. Because the historical cohort study is really a retrospective approach, the terms *cohort study* and *prospective study* should not be used synonymously.

Measures of Association in a Cohort Study

Measures of association illustrate the statistical relationship between two or more variables, and three important measures of association will be discussed using the symbols and numbers provided in Tables A–5 and A–6. Let us assume that one is doing a study of smokers and nonsmokers and following them to see who develops lung cancer over a defined period of time.

A. Relative Risk

Relative risk (RR) is the risk of disease among people exposed to a factor relative to the risk among people not exposed and is a measure of the strength of association between an exposure and a disease.

$$RR = \frac{\textbf{Disease rate in the exposed population}}{\textbf{Disease rate in the nonexposed population}}$$

$$= \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{\frac{63}{10^5}}{\frac{7}{10^5}} = 9$$

An RR greater than 1 implies a positive association of the disease with the exposure of interest; an RR less than 1 implies a negative association (or protective effect) between the disease and the exposure.

Table A–6. Example of data collected in a cohort study of lung cancer and smoking.

	Develop Lung Cancer	Do Not Develop Lung Cancer	
Smokers	63	99,937	100,000
Nonsmokers	7	99,993	100,000

The results in the preceding example suggest that the risk of lung cancer among smokers is nine times greater than the risk for nonsmokers. RR is important for testing etiologic hypotheses.

B. Attributable Risk

Attributable risk (AR) is the rate in the exposed population minus the rate in the nonexposed population.

$$AR = \frac{a}{a+b} - \frac{c}{c+d}$$

$$= \frac{63}{10^5} - \frac{7}{10^5} = \frac{56}{10^5}$$

It indicates the rate of occurrence of death or disease that is caused by a specific exposure factor.

Of the 63 lung cancer deaths that occur annually among 100,000 smokers, 56 (89%) are attributable to smoking. Because a disease may have multiple risk factors that interact with each other, the sum of attributable risks may be greater than 100%.

AR can be an important tool for counseling individuals with specific risk factors because it helps give an idea about the amount of disease that could be avoided by reducing risk factors in individuals.

C. Population Attributable Risk Percentage

Population attributable risk (PAR) *percentage* is the proportion of a disease in a population related to (or "attributable to") a given exposure.

$$PAR = \frac{P_e(RR-1)}{P_e(RR-1)+1}$$

where P_e is the proportion of the population exposed to the risk factor, and RR is relative risk.

Assuming that 40% of the population smokes (P_e) and that the relative risk (RR) of lung cancer associated with smoking is 9, then

$$= \frac{0.4(9-1)}{0.4(9-1)+1} = \frac{3.2}{4.2} = 76.2\%$$

That is to say, 76% of cases of lung cancer in the general population are attributable to smoking. PAR is important for public health policy and planning, that is, in estimating what percent of cases in a population could be eliminated by removing an exposure.

2. Case-Control Study

The case-control study is a frequently used design in analytic epidemiology. It determines the risk factors

associated with a particular disease by comparing a group of subjects who have the disease (cases) with one or more groups composed of subjects who do not have the disease (controls). Risk factors studied may be permanent, such as gender or race; they may be current, such as present drug use; or they may be historical, such as previous employment. The difference in the frequency distribution of the risk factors between the case and control groups is examined, and the magnitude of the association of these factors with the disease under study is estimated.

Case-control studies are a commonly used design in occupational epidemiology to evaluate multiple exposures associated with a single outcome. For example, an investigator may be interested in the many occupational and nonoccupational causes of lung cancer. Conversely, a study of many health outcomes associated with a single exposure or workplace would best be investigated using a cohort design.

The case-control study is always retrospective. The investigator starts by identifying diseased and nondiseased individuals (i.e., the effect) and looks backward for the presence or absence of exposures (i.e., the causes) in these individuals.

For example, to study the relationship between asbestos exposure and mesothelioma, a case-control study would compare the history of asbestos exposure in a group of mesothelioma patients with the history of asbestos exposure in a group of subjects who do not have mesothelioma. The cohort study, in contrast, first identifies a group of disease-free individuals classified for absence or presence of the risk factor or exposure of interest and then follows these individuals over time to compare the incidence of disease in the exposed and unexposed groups. A cohort study of the relationship between asbestos exposure and mesothelioma first would classify a group of nondiseased persons according to their asbestos exposure and follow them to determine whether the asbestos-exposed subjects had a higher incidence of mesothelioma over time than the nonexposed subjects.

Case-control studies generally can be done more rapidly and less expensively than cohort studies. The time required to complete the study is the time needed to assemble the necessary data; the investigator does not need to wait for cases of the disease to appear. This usually results in lower costs because fewer study personnel and subjects are necessary to test a hypothesis.

For example, suppose that half the general population is exposed to a risk factor (e.g., cigarette smoking) and half is not. If a disease (e.g., lung cancer) has an annual incidence rate of 100 per 100,000 in the exposed population and 10 per 100,000 in the nonexposed population, a study of 100 cases and 100 controls probably would reveal the increased risk of disease associated with exposure to the factor. Uncovering 100 cases of disease in a cohort study would mean following 10,000 exposed people for 10 years. The more rare the disease, the greater the relative advantage of the case-control study.

Source & Selection of Cases

In defining a case, the diagnostic criteria should be clear and permit selection of a homogeneous group of cases. For example, in cancer studies, microscopic confirmation of the presence of disease and clearly defined criteria for classification by a pathologist of the type of cancer greatly enhance the validity and generalizability of the study findings. The case group usually is composed of (1) all persons with the disease seen at a particular medical facility or group of facilities in a specified period or (2) all persons with the disease found in a community or in the general population in a specified period. Whatever the source of the cases, they should be newly diagnosed (or incident) cases of the disease. Inclusion of prevalent (diagnosed in the past) cases will increase the sample size but can complicate analysis and interpretation of results. Prevalent cases are "survivors" and therefore may not be representative of all people who develop a given disease. Inclusion of prevalent cases inadvertently may identify factors that result from the disease rather than factors that are causally related to its development.

Source & Selection of Controls

The four most common sources of the control group are (1) the general population, (2) hospital patients, (3) relatives of cases, and (4) associates or friends of cases.

The general population control group is appropriate if all or most cases occur in a specific geographic area—for example, a county—because in this situation the controls represent the same target population as the cases. Using general population controls, however, presents certain problems: potentially lower response rates than from other types of control groups and from the case group, differing quality of information if the interview setting differs for the cases and the controls, and higher costs for obtaining information.

The hospital patient control group is selected from patients at the same hospital or clinic that the cases attended. This control group may share the selective factors that influenced the cases to come to a particular hospital or clinic, such as residence, ethnicity, or income. These patients (the controls) are readily available, often have the time to accommodate study interviewers, and can be more cooperative. The disadvantage of the hospital control group is that it is composed of people with an illness who may differ from the general population with regard to factors

often associated with disease, such as smoking habits and/or drug use. In addition, the factors that cause patients to attend a particular hospital may not be the same for all diseases. For example, a hospital with a national reputation for treating Hodgkin disease may have patients with this disease from all over the country, whereas its population of coronary disease patients may come only from the region surrounding the hospital; thus the two patient groups may differ greatly. Similarly, healthy people attending a hospital screening clinic may differ markedly in ethnic, socioeconomic, or other factors from the inpatient population of that hospital. One consideration in selecting controls is whether to draw them from the hospital's entire patient population or to exclude patients who have diseases related to exposure factors under study. For example, in a case-control study of the relationship between lung cancer and smoking, it would seem logical to exclude from the control group persons who have emphysema because emphysema is related to smoking, the exposure factor under study. There also may be the problem of a lack of knowledge of whether factors being studied are related to diseases present in hospital controls. Selecting controls with differing types of diseases would minimize this problem.

Spouses and siblings are the relatives used most commonly as controls because of similarity in ethnicity and environment with the case group. Moreover, sibling controls genetically are similar to the cases. Spousal controls are appropriate if there is an approximately equal number of male and female cases, and the age range of cases is such that a high proportion of spouses are likely to be alive. When siblings are the controls, one sibling should be selected per case; using all available siblings would result in the control group having many characteristics related to family size, which may confound any observed associations between the exposure factor and the disease. In contrast, cases with no siblings would have to be excluded from the study (for lack of an equivalent control), which may result in biased study results.

A control group of associates of cases such as neighbors, coworkers, friends, or schoolmates has the advantage of being composed of generally healthy individuals who are similar to the case group with regard to lifestyle characteristics; for example, neighborhood controls are usually of the same SES as the cases. However, such associates might be more similar to cases than members of the general population with respect to risk factors under investigation, thus impairing the ability of the study to detect true differences in exposure between people with and without disease. Other disadvantages of associates as controls are the effort necessary to identify them, a response rate different from that of cases, and probable variations in the quality of information obtained from cases and controls.

Sampling

Once the source of the control group has been determined, one must decide on the method of selecting the controls. Either all eligible individuals are selected from a specific group—although this is usually not required—or a sample is selected. Whenever sampling is employed, its protocol should be defined and adhered to throughout the sampling period. Examples of common sampling strategies are (1) random sampling, (2) systematic sampling, and (3) paired sampling.

In random sampling, each member of the source group has an equal chance of being represented in the control group. For example, all individuals might be assigned a number, and the sample would be selected using a table of random numbers.

In systematic sampling, the source group for controls is assumed to have an ordered sequence, and every *n*th individual is selected. As long as the sequence of the source group is not related to an important study variable (e.g., age), the resulting characteristics of a systematic sample are similar to those of a random sample.

In addition to random or systematic sampling, a popular method of selecting controls is paired sampling. In paired sampling, one or several controls are selected for each case based on a predefined relationship to the case. For example, if hospital controls are used, the person who was admitted immediately before or after the case might be chosen for the control group. The investigator may choose to select for each case one or more controls who are individually matched with the case on characteristics such as gender, age, or SES—which, if not controlled, might lead to spurious associations in the final results. For example, as a neighborhood control, the resident of the nearest dwelling to the right of the case's house who is of the same gender and age (±5 years) as the case might be selected. Such matching at the outset of the study is one way of taking into account any variables known to be associated with both the disease and the exposure of interest.

Sources of Bias

Bias must be acknowledged as a potential issue for nearly every type of epidemiologic study design. It is defined as a systematic error in the design, execution, or analysis of a study that results in an erroneous estimate of the effect of an exposure of interest to the risk of an outcome or disease.

While bias is more common in case-control studies, it also may occur in cohort studies; for example, information about outcome measures may be obtained dif-

ferently in exposed and unexposed subjects. However, the underlying principle is the same: Any difference in the way information is obtained from the study groups may bias the results of the study.

There are two main categories of bias to be aware of: selection bias and information bias (or measurement error).

A. Selection Bias

The appropriate control group should be chosen judiciously because when a systematic error is made in the selection of one or more study groups, selection bias may result. Under the null hypothesis, cases and controls have been equally "exposed" to the study factor. Therefore, selection of the cases and controls must use similar eligibility criteria to ensure that both groups are comparable and therefore more likely to be representative of the same underlying population so that if we reject the null hypothesis and determine that cases differ from controls on the study factor, it is not because we selected them to be different by using a biased procedure. Because the case group usually is chosen first, selection bias is avoided by a careful choice of the appropriate control group.

As an example of how selection bias can occur, suppose that the study is about the relationship between Alzheimer disease and previous exposure to lead. The case group is chosen from the inpatient population of a private hospital and the control group from the outpatient clinic of the same hospital. Once the cases and controls are selected, it is discovered that they differ dramatically with respect to SES—the inpatient population being predominantly upper middle class and the clinic population predominantly lower class. Thus, if the study finds that the cases and controls differ in terms of prior lead exposure, it would not be known whether this is a true difference or whether the difference is a consequence of other factors related to SES.

Selection bias also can occur if the control group is composed of people who volunteer for the study, because volunteers differ in significant ways from nonvolunteers; for example, they may be more educated, more active in community affairs, or less likely to be smokers.

B. Information Bias

In interviewing study subjects about past exposures or events, the interviewer who knows the disease status of the individual (case or control) may pose questions unconsciously or probe for answers in a different manner, commonly referred to as *interviewer bias.* For example, in a case-control study of factors related to lung cancer, an interviewer might pursue in greater depth questions concerning asbestos exposure when obtaining work or environmental histories from cases than from controls.

To avoid this bias, the procedure used to collect information should be identical for cases and controls. Ideally, the data collector is unaware of the hypotheses being tested and whether the subject is a case or control; however, in collecting information of a medical or personal nature, it is often difficult to avoid learning of the person's disease status. Every effort therefore must be made to keep interviews as comparable as possible (e.g., place, length, and format of questionnaire; attempts to gain cooperation and accurate information; and other aspects of the interview), and each interviewer should see an equal number of cases and controls.

Another source of information bias can occur when a study subject is asked to recall past exposures or events because recall might depend on the person's current disease status. For example, a person with lymphoma is more likely to recall remote exposure to pesticides than a control subject without cancer. To minimize recall bias in this instance, one might try to obtain independent verification of previous exposure. It is also advantageous to use information recorded before the time of diagnosis wherever possible. In using data from interviews in which the case has a serious illness and the control has not, the items on which cases and controls can be compared with the greatest confidence are those least subject to recall bias. For example, prior surgery is a more objectively reported event than prior drug use.

Misclassification of study subjects also can bias study results owing to inaccuracies in the methods by which data are gathered from study subjects or methods by which information is abstracted from various sources. Misclassification bias comes in two forms—differential and nondifferential. Differential misclassification that is related to disease or exposure status can lead to the appearance of a relationship between exposure and disease where one does not truly exist, or perhaps more unsettling, it can mask a true association. Nondifferential misclassification is not related to exposure or disease status and tends to attenuate any association between exposure and disease.

Confounding

The phenomenon of *confounding* is another explanation for an apparent association between an exposure and a disease and also may cause no association to be observed when a true association exists. As with bias, confounding may occur in any type of analytic epidemiologic study. By definition, a factor that is associated with the exposure of interest and is also an independent cause of the disease being studied is a confounder. When confounding occurs, an observed association between an exposure and a disease is in fact due wholly or in part to the association of the exposure with the confounding factor, which, in turn, is itself a cause of the disease. If the suspected con-

founder is not differentially associated with the exposed subjects or is not a cause of the disease, it cannot be considered a confounding factor.

An example of a confounding factor is cigarette smoking in a study of an occupational exposure and lung cancer. Cigarette smoking is a known cause of lung cancer. If the cigarette smoking prevalence were greater (or less) in the population exposed to the occupational exposure agent, failure to control for smoking in the study design or analysis would lead to an apparently greater (or lesser) association between the occupational exposure and lung cancer.

Analysis of Case-Control Studies

Data from the case-control study are conventionally arrayed so that cases and controls can be compared on exposure to a hypothesized etiologic factor:

		Disease Status	
		Cases	Controls
Exposure	Yes	a	b
	No	c	d
		a + c	b + d

The incidence of disease among the exposed and nonexposed cannot be calculated by using case-control data because the cases and controls in the study rarely reflect the true proportions of diseased and nondiseased persons in the population. [The investigator usually selects roughly equal numbers of cases ($a + c$) and controls ($b + d$) in the study, whereas there are likely to be many more nondiseased than diseased people in the general population.] Therefore, the relative risk (RR) of disease associated with exposure cannot be calculated directly in a case-control study, as it was for the cohort study. However, an estimate of the RR, known as the *odds ratio* (OR), can be calculated if the proportion of diseased people in the general population is small compared with the proportion of nondiseased (almost always true). Recall that the true RR using data from a cohort or incidence study is as follows:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

where a is the number of cases among the exposed group in a cohort study, b is the number of noncases among the exposed group, c is the number of cases among the nonexposed group, and d is the number of noncases among the nonexposed group.

In a cohort study, as in the general population, a is very small relative to b. Similarly, c is very small relative to d. Thus, in the general population (and the usual cohort study), $a/(a + b) \approx a/b$ and $c/(c + d) \approx c/d$. Consequently, the formula for relative risk reduces to

$$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc} = \text{odds ratio (estimated relative risk)}$$

Example: *One hundred men with lung cancer and 100 controls are interviewed regarding smoking history:*

	Cases	Controls
Smokers	80	30
Nonsmokers	20	70
	100	100

$$\text{Odds ratio} = \frac{ad}{bc} = \frac{80 \times 70}{30 \times 20} = \frac{5600}{600} = 9.3$$

Because the OR is an estimate of RR, one can conclude that these data show a ninefold increased risk of lung cancer in smokers compared to nonsmokers.

PAR (i.e., the proportion of all instances of the disease in the population that can be attributed to the exposure of interest) can be estimated from case-control studies by using the following equation:

$$PAR = \frac{p\,(OR-1)}{p\,(OR-1)+1}$$

where p is the proportion of the population with exposure of interest [estimated from controls as $b \div (b + d)$], and OR is the estimated RR (OR) associated with the characteristic.

Matched Case-Control Studies

Controls frequently are selected in a case-control study so as to be individually matched to the cases as to characteristics such as age, gender, race, or SES that are known to be related to the disease. Matching helps to make the two groups similar with respect to factors other than the exposure of interest in the study and thus serves to reduce the likelihood of spurious associations. The investigator must be careful not to overmatch, that is, to match cases and controls on factors related to the exposure of interest; overmatching can artificially reduce—and may even eliminate—true exposure differences between diseased and nondiseased individuals in the study. It should be obvious that cases and controls cannot be compared in the analysis with respect to any characteristics on which they have been matched.

The data in a matched-pairs analysis are organized as shown below:

		Controls	
		Exposed	Not exposed
Cases	Exposed	r	s
	Not exposed	t	u

where r is the number of pairs in which both case and control are exposed to the factor (concordant), s is the number of pairs in which the case but not the control is exposed to the factor (discordant), t is the number of pairs in which the control but not the case is exposed to the factor (discordant), and u is the number of pairs in which both case and control are not exposed to the factor (concordant).

To compute the OR (estimated RR) for a matched-pairs study, only the discordant pairs enter into the calculation:

$$\text{Odds ratio} = \frac{s}{t}$$

$$\text{where } t \neq 0$$

Example: *One hundred seventy-five children age 5–15 years admitted to hospital in 1968 with acute asthma were matched on age, gender, race, and date of admission to 175 controls. All children in the study or their parents were interviewed regarding personal habits and home characteristics during the month preceding admission. The results regarding environmental tobacco smoke (ETS) exposure were as follows:*

		Controls		
		Yes ETS	No ETS	Totals
Cases	Yes ETS	10	57	67
	No ETS	25	95	108
		35	152	187

$$\text{Odds ratio} = \frac{s}{t} = \frac{57}{25} = 2.3$$

These data show that children who have asthma have a 2.3 times greater odds of environmental tobacco smoke exposure than do children without an acute asthma admission. Show the calculation for how you arrived at the answer by way of example.

3. The Experimental Study

The experimental study is the type of design most familiar to clinical investigators, but it is rarely encountered in occupational epidemiology. Unlike the cohort and case-control studies, which are observational in nature—that is, the investigator observes exposed individuals for the development of disease or diseased individuals for past exposures—in an experimental study, the investigator manipulates exposures and studies the impact on disease. The intervention can occur at different points in the natural course of the disease. Subjects are normally randomly assigned to the different interventions in an experimental study. Ideally, study outcomes also should be determined by individuals blind to the exposure status of the subjects.

Experimental clinical trials often are undertaken among individuals with the same disease who are assigned to different treatment groups. An example is the Carotene and Retinol Efficacy Trial (CARET) study, in which men with asbestos exposure, who are at increased risk of lung cancer, were randomly assigned to receive beta-carotene or a placebo. The study was undertaken to determine whether beta-carotene decreases the risk of developing lung cancer.

Alternatively, intervention might occur in the form of a screening program offered to one group of people at risk of disease and not to another similar group. An example of this type of intervention study is the National Cancer Institute's Cooperative Screening for Early Lung Cancer Program. Men aged 45 years and older with a history of heavy cigarette smoking were assigned to a dual-screened group receiving chest radiographs and sputum cytologic testing or to a group receiving only chest radiographs. The objective was to determine whether the addition of sputum cytologic testing to regular chest radiography resulted in earlier detection and improved lung cancer survival.

CAUSAL ASSOCIATION

An epidemiologic study may demonstrate an association that is not valid because of chance, bias, or confounding, as discussed previously. If the association is believed to be valid—that is, the disease occurrence is in fact not equal among the exposed and unexposed subjects—and the observed association cannot be explained by chance, bias, or confounding, the investigator must consider whether the data support a cause-and-effect association.

This process involves consideration of the study itself and all existing data on the subject. Factors that should be considered in evaluating whether an association is causal include (1) the strength of the association, (2) whether dose-response relationships are present, (3) consonance with existing knowledge (i.e., other studies demonstrating the same finding), (4) biologic plausibility (i.e., whether there is a proposed biologic mechanism), and (5) the temporal sequence of events (i.e., cause precedes effect).

While uncertainties always will exist following an epidemiologic study, action on the findings of a study

will depend in part on how strongly the data support a causal association and on the need for action versus the consequences of obtaining more data.

REFERENCES

Epidemiology Software

Epi Info 6, Epi Map 2, www.cdc.gov/epiinfo/Epi6/EI6.htm (Epi Info is designed for public health professionals. It is easy to use and comes free from CDC. The package includes word processing, data management, and epidemiologic analysis programs. Epi Map is a program for displaying counts or rates on geographic maps. For PC only.).

EpiCalc 2000,:www.brixtonhealth.com/ (An easy-to-use statistical calculator; can be customized for languages other than English. For PC. Available for free with other epidemiologic and statistical programs.).

Vitalnet, www.chdp.com/vitalnet/index.htm/ (A data-analysis program for analysis of mortality and population data. It provides the data analysis/data dissemination infrastructure for a national, state, or city. Runs locally or over the Internet.).

The R Project for Statistical Computing, R 2.2.1, www.r-project.org/ (R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows, and Mac OS.)

Share your own customer images

Search inside this book

Are You an Author or Publisher?
Find out how to publish your own Kindle Books

Current Occupational & Environmental Medicine (Lange Medical Books) (Paperback)

by Joseph LaDou (Author)

Key Phrases: evaluating chronic effects, involuntary smoke exposure, nary function testing, United States, Occup Environ Med, Essentials of Diagnosis (more...)

☆☆☆☆☆ (1 customer review)

List Price: ~~$66.95~~

Price: $58.25 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $8.70 (13%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Only 3 left in stock--order soon (more on the way).

Want it delivered Tuesday, May 13? Order it in the next 2 hours and 8 minutes, and choose **One-Day Shipping** at checkout. See details

40 used & new available from $41.47

Also Available in: List Price: Our Price: Other Offers:
Paperback (3) 7 used & new from $18.92

Better Together

Buy this book with A Practical Approach to Occupational and Environmental Medicine by Robert J McCunney today!

Buy Together Today: $126.25

Add both to Cart

Customers Who Bought This Item Also Bought

Page 1 of 13

Back Next

Occupational and Environmental Health: Rec... by Barry S Levy
(1) $71.96

Occupational Medicine Secrets by Rosemarie M. Bowler
(1)

Textbook of Clinical Occupational and Envi... by Linda Rosenstock
$189.00

No image available

Environmental and Occupational Medicine by William N Rom
(1) $224.10

Any Category Biology Biostatistics Environmental Science Family Practice Hospital Administration Medicine Nursing Orthopedics Physical Examination Preventive Medicine Public Health Safety & Health Test Preparation & Review Toxicology

Editorial Reviews

Product Description

Up-to-the-minute, thorough, clinical coverage of common and important occupational and environmental diseases, injuries, and exposures

Complete, yet concise, this clinically focused guide offers the definitive overview of common occupational and environmental illnesses, covering their diagnosis and treatment-plus preventive and remedial measures in the workplace and community. With its practical format and emphasis on fundamental topics, CURRENT Occupational and Environmental Medicine is just as essential for students and residents as it is for practicing physicians. You can count on the new fourth edition to deliver the bottom-line answers you need to stay on track in this complex, fast-breaking field.

Features

- The latest OSHA/NIOSH guidelines for occupational exposure standards
- Detailed diagnostic checklist for major diseases, injuries, and exposure that help expedite diagnosis and treatment
- The most clinically relevant perspectives on disability prevention-required reading for the occupational physician
- Skill-building insights on the importance of ergonomics in the workplace
- A step-by-step review of how to effectively manage an occupational health and safety program
- Details on substance abuse and employee assistance programs, health risk analysis, and the legal aspects of occupational and environmental medicine
- Preventive approaches to terrorist attacks on industry
- Information-packed primer on epidemiology and biostatistics for the occupational and environmental health specialist
- Up-to-date references with PMID numbers and peer-reviewed websites

Book Info
Univ. of California, San Francisco. Clinically focused text, for professionals, students, and residents, emphasizes basic concepts. Presents the latest OSHA/NIOSH guidelines, and discusses disability prevention, surveillance of occupational and environmental diseases, and health risk assessment. Previous edition: c1997. Softcover. *--This text refers to an out of print or unavailable edition of this title.*

See all Editorial Reviews

Product Details

Paperback: 846 pages
Publisher: McGraw-Hill Medical; 4 edition (October 23, 2006)
Language: English
ISBN-10: 0071443134
ISBN-13: 978-0071443135
Product Dimensions: 9 x 7.3 x 1.3 inches
Shipping Weight: 2.7 pounds (View shipping rates and policies)
Average Customer Review: (1 customer review)
Amazon.com Sales Rank: #109,802 in Books (See Bestsellers in Books)