CHAPTER 2

Role of Epidemiology and Statistics in Advanced Nursing Practice

"Chance favors prepared minds."
—Louis Pasteur

Peter Wludyka

OBJECTIVES

- After completing this chapter the reader will understand that statistical methods and ideas as well as practical probability are essential to understanding epidemiology and its role in advanced nursing practice.
- The overall objective of the chapter is to familiarize the reader with the key elements of statistics (and statistical inference) including the role of level of measurement; methods for presenting, summarizing, and analyzing interval or categorical data; estimation; hypothesis testing; and modeling.
- The key epidemiological ideas of incidence and prevalence are defined, the distinction between rates and proportions is clearly made, and inferential methods associated with estimation of these parameters are presented.
- Statistical methods frequently used for examining the relationship between exposure and disease incidence, including survival analysis, the analysis of two by two tables, and regression modeling, are also discussed.

EASILY ACCESSIBLE RESOURCES IN EPIDEMIOLOGY AND STATISTICS

Many of the statistical ideas and methods in this chapter are discussed in more detail in van Bell, Fisher, Heagerty, and Lumley (2004). An excellent nonmathematical introduction to epidemiology can be found in Rothman (2002). For a more mathematical treatment of the statistical analysis of epidemiological data, Selvin (2004) is a solid resource. A comprehensive treatment of the analysis of means (ANOM) can be found in Nelson, Wludyka, and Copeland (2005). ANOM is a useful approach for analyzing stratified incidence (both proportions and rates) and prevalence data. Both SAS® and

PASW (formerly SPSS) were used to analyze data for examples in this chapter, and output (sometimes with the format modified) from these statistical packages appears throughout. OpenEpi, version 2 (Dean, Sullican, & Soe, 2009), which is available online, was also used and offers an easy way to statistically analyze epidemiological data. Cantor (1997) has an excellent discussion of survival analysis and explains how to use SAS to analyze survival data. Petrie and Sabin (2009) provide a very easy to read and understand general treatment of statistical methods useful in medical research, which also has an excellent glossary of statistical terms. There are several online glossaries of statistical and epidemiological terms, including Dorak (2010).

KEY IDEAS, DEFINITIONS, AND PRELIMINARIES

Epidemiology involves "the study of the distribution and determinants of disease frequency" (Rothman, 2002, p. 1970) and as such uses tools and intersects with many disciplines, including statistics, probability, demography, geography, and the biological sciences. Our interest in this chapter is statistics. The word "statistics" is used in two ways, illustrated by the following two sentences. "What is statistics?" "What are statistics?" The latter sense is the familiar one in which "statistics" are compilations or summarizations of data and includes the more technical use of the word "statistics" to describe characteristics (functions) of a sample. The former refers to the body of knowledge or discipline called "statistics." Both senses of the word are used in this chapter and the context determines which we use.

The "what is" aspect of statistics was first described to me by the late Professor Peter R. Nelson as "the art and science of dealing with variability." This captures the key idea: variability. When absent (which is almost never, and certainly never when one includes measurement error), there is no need for statistical methods. How much does a 5-pound bag of potatoes weigh? If one were studying 5-pound bags of potatoes, would statistical methods be appropriate? There are those who think of statistics as a science. Or even as an incarnation of the scientific method. But in practice art often seems to intervene. One of the most interesting (and distressing to some) aspects of statistics as a science is that two statistical analyses of the same phenomenon or process (even using the same data) can lead to different conclusions. Is this a result of ignorance? Mendacity? Honest disagreement?

The "what are" sense involves data, which is such a primitive idea that defining it might be impossible. Data can be thought of as "facts." These facts might be numbers or descriptors or almost anything. There are many sources for data, including government compilations (think Centers for Disease Control and Prevention, Census Bureau, etc.), private compilations such as those created by insurance companies and

hospitals, or data arising from experiments and studies, including clinical trials. Typically, in the modern sense we think of these as data sets or databases (the latter having some formal structure relating the objects in it as well as methods for querying the database).

Data sets contain one or more variables. For example, a data set arising from a study of bedsores at a nursing facility might consist of cases (patients, typically being rows in a spreadsheet) along with facts about each patient such as the patient's gender, age, date at which the bedsore was discovered (recorded), comorbidities, severity of the bedsore (perhaps a score), the duration of the bedsore (date of "cure" minus date of discovery), numbers of sores, and so on. Each fact (often columns in a spreadsheet) is a variable. Interest in these variables arises primarily from the potential they have to vary; that is, different patients (or subjects or cases) may have different ages or may or may not have bedsores.

It is useful to classify variables with regard to their level of measurement: nominal, ordinal, interval, and ratio. Nominal data merely names (e.g., the variable gender is nominal). Ordinal data require that the data be capable of being ranked (e.g., severity of the bedsore might be scored on a scale of 1 to 4 in which 4 is the most severe and 1 the least). Note that $4 > 3 > 2 > 1$. Data of this type are sometimes referred to as being measured on a Likert scale. There is no need for ordinal data to be represented by numerals. The numerals in this example might correspond to 1 = "mild," 2 = "moderate," 3 = "severe," 4 = "extremely severe" so that in the data set the numerals are recorded (or in some cases the text itself is recorded). It is quite possible that the labels are reversed (4 = "mild"). In many circumstances severity scores such as these have clear definitions often based on checklists.

Interval data have the property that differences between values have a clear interpretation; for example, for the interval variable age, a patient 21 years of age is 9 years older than a patient 12 years of age. Ratio data have the property that ratios of values have a clear interpretation; for example, a patient with four bedsores has twice as many as a patient with two bedsores. The key property that identifies ratio data is the existence of a true (interpretable) zero. Zero bedsores means none is present (the patient does not presently have the disease of interest). Compare this with a temperature of zero degrees Fahrenheit. Observe that ratio data are interval, interval data are ordinal, and ordinal data are nominal; hence, one usually uses the highest level as the descriptor for a particular variable. Also note that interval data can always be converted into strictly ordinal data (e.g., small, medium, and large) by creating definitions.

In epidemiological studies all these types of variables may occur. Of special interest are counts (or count data). The number of bedsores on patient number 77 is an example of this; however, a more common epidemiological usage of counts might be

the following: of 100 patients in the study, 26 were free of bedsores after 6 weeks of treatment. Or, using a slightly different type of counting, in the 100 patients in the study there were a total of 217 bedsores.

Why be concerned about level of measurement? The levels of measurement of the variables in a data set affect the method of data analysis. Most immediately, interval data can be analyzed using some of those things one learned in the fifth grade: adding, subtracting, and averaging. Or with nominal data one reverts to earlier grades in which one learned counting and perhaps later how to compute percentages (ratios).

So what is statistical data analysis? It almost always involves summarizing and organizing data. It frequently involves discovering and describing relationships between variables. In its more exotic forms it might involve creating models—something that may be going on unconsciously because a model is nothing more than a simplification of reality. A convenient definition of data analysis is drawing conclusions about some process or phenomenon using data. I prefer the more artistic characterization: determining and describing the story that a data set has to tell. This latter characterization allows different people (researchers) to tell different stories based on the same "set of facts" (data) collected about a process or phenomenon. Implied in the data analysis process are the notions of discovery and communication.

Data sets are of two general types: a census or a sample. The former is the totality of objects of interest (e.g., the USRDS database containing all dialysis patients in the United States for the year 2005). A sample is any subset of a population. Typically, there is little real interest in the sample itself. The interest lies in what the sample can tell us about the population. The activity of drawing conclusions about a population is called inference. If statistical methods are used, it is called statistical inference.

ROLE OF PROBABILITY IN EPIDEMIOLOGY

Probability is the study of the laws of chance. As a formal endeavor it arose in the study of gambling (games of chance) and as such it is naturally prospective. That is, it exists in reference to the future. This can be extended to ignorance (but that is often somewhat tenuous). In probability one deals with events. In the disease setting an event might be something like a patient with two bedsores. The probability of an event is a measure of the likelihood an event will occur. This measure is constrained to the interval [0, 1], in which a probability of 1 corresponds to absolute certainty and a probability of 0 is interpreted to mean that the event is impossible. This 0-1 characterization applies to all past or current events (but not necessarily to our knowledge about these events, which may be more problematic). That is, either they have oc-

curred or they have not occurred. So does it make sense to say the probability that a 77-year-old nursing home patient has bedsores is 0.15? Either the patient does or does not have bedsores. The more practical approach is a frequentist characterization: In the population of nursing home patients we are interested in the fact that 15% of those 77-years-olds have bedsores. One can clothe this probabilistically by saying that a randomly selected patient from this population has a probability of 0.15 of having a bedsore; however, the frequentist characterization is more useful. In epidemiology probability is typically described as risk.

Probabilities can be assigned in three ways: subjectively, a priori, or empirically. The first of these might arise from panels of experts. The a priori assignment of probabilities arises frequently in simple games of chance, such as rolling dice. These methods involve some version of counting all outcomes, counting all "favorable" outcomes and forming a ratio to describe the probability. The exercise is purely deductive (mathematical). Epidemiological assignments of risk (probabilities) are almost always arrived at empirically; that is, through observation. For example, suppose one wishes to know the risk of bacteremia in dialysis patients in 2005. One can query the USRDS database to count the number of patients with bacteremia (I am making this much more simple than it actually is) and divide that by the number of patients in the database. This is a census, so the result is (apart from misclassification errors) a parameter (characteristic of a population). If one were to select a sample of patients from a clinic in which dialysis is performed, then a similar ratio could be formed; however, this number is an estimate. Estimates are subject to error. The error arises from the fact that a sample seldom is an exact replica of the population.

STATISTICAL INFERENCE AND THE LANGUAGE OF STATISTICS

Drawing conclusions from samples using statistical theory (methods) is called statistical inference. The subject is vast. Often, this discussion begins with the distinction between descriptive statistics and inferential statistics. For the most part this is a distinction without a difference because descriptives (this word is intended to include measures as well as graphical objects) quite naturally lead to attempts to draw conclusions. The advantage of formal inference is that the inferences can be associated with certain "probability" measures that tell one something about the likelihood that certain errors might occur. Strictly speaking, this probabilistic interpretation has meaning only before collecting the data.

Beginning with descriptives is useful. These are results from summarizing and organizing data. These can be summary measures or graphical objects. Recall that

data analysis is discovering the story that a data set has to tell, so organizing and summarizing often make a good first step. No attempt will be made here to be complete. There are many standard statistics texts that are more complete (see, e.g., van Belle et al., 2004).

Summary Measures and Graphical Tools

A measure is a single number (or small set of numbers) used to describe a sample or population. With nominal data these are often percentages. It should be clear that using a single measure to describe a variable in a data set is futile if one's goal is completeness (recall that variability is the necessary condition for statistical methods to be useful). This reduction of a data set to a set of measures always involves (massive) loss of information. The benefit is conciseness and "understandability." The problem is addressed partly by using several measures that address aspects of the data. One approach is the LaSSO approach (useful for interval data): location, spread, shape, outliers. Addressing at least the first two of these is usually essential. Common measures of location are the mean and median. Common measures of spread are the standard deviation (SD; or its square, the variance), the range, and the average absolute deviation from the median (mean). The mean and SD have convenient mathematical properties, which is part of the reason they are so commonly used (as well as the fact that the normal model is completely described stochastically by these two parameters). Shape includes ideas such as skewness (the data have a long tail either left or right), bimodality (the data set consists of two piles or peaks), or more technical (and difficult to describe) notions such as kurtosis. Shape is usually best conveyed graphically using tools such as bar charts, histograms, and box plots. Outliers are "unusual" observations. Attempts at a formal definition are problematic. For univariate analyses graphs and plots (such as box plots) are useful for identifying potential outliers.

There are also summary measures useful for describing relationships among variables. The correlation coefficient (either Pearson or Spearman—along with others) measures the degree of linear association between two variables and is bounded on the interval negative 1 to positive 1 (where values near 0 indicate lack of correlation). The linear part of this is often suppressed or ignored by practitioners (and even more so by readers of research). That is, there is often a somewhat clear relationship between variables x and y but the correlation coefficient, r, is quite small in magnitude (near zero). Scatter plots often do a much better job of conveying relationships between interval variables, especially when curvature is present.

As an example consider a hypothetical nurse-based intervention study with diabetic patients. In the study the subjects in the treatment group were selected from an

urban clinic and the control group was selected from another (but similar) city's clinic. At end of study HbA_{1c} levels (an interval variable denoted HbA_{1c3}) were measured. **Figure 2-1** contains side-by-side box plots (sometimes called box and whiskers plots) of the HbA_{1c} data by treatment group (a categorical/nominal variable). Box plots are useful for assessing the relationship between a categorical variable (in this case treatment group) and an interval variable (in this case HbA_{1c}). The vertical spread of the boxes measures variability. The solid horizontal line in each box is the median. The boxes capture the middle 50% of the observations, with the top of the box corresponding to the 75th percentile and the bottom of the box the 25th percentile. The vertical spread of a box is the interquartile range, which is a measure of variability of the middle 50% of the observations). Do HbA_{1c} levels at end of study differ between the treatment and control groups? Because all the numbers used to construct these box plots are derived from samples, they are estimates so apparent differences could just be the result of random chance. This issue is treated later (in Inference, below). The treatment group has lower HbA_{1c} levels based on the box plots; for example, about 75% of the treatment group has levels less than the median level for the control subjects. Both groups exhibit a great deal of variability (look at the distance from

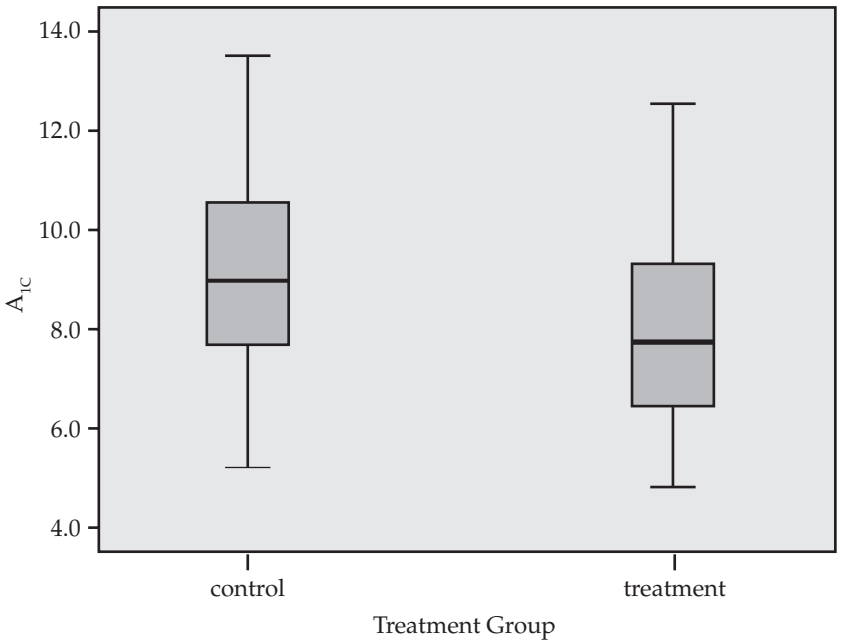


FIGURE 2-1 Box plot of variable HbA_{1C} of both control and treatment groups.

whisker tip to whisker tip, which is the range when no outliers are present, as well as the vertical height of the boxes).

The same data can be summarized with statistical measures. Several are displayed in **Table 2-1**. This data set is a sample so the measures are statistics (estimates); hence, for some of them standard errors are displayed. The standard errors are measures (themselves estimates) of sample-to-sample variability of the estimates. From this one sees that the average HbA_{1c} level for the treatment group (7.967) is lower than the average for the control subjects (8.976). Which group has greater variability? The SDs are about the same (recall that the SD roughly measures the average distance of the observations from the mean); technically, the SD in Table 2-1 is the square root of the average squared distance from the mean multiplied by the square root of $[n/(n-1)]$. The factor $[n/(n-1)]$ is close to 1 for reasonably large n. Definitions of the

Table 2-1 Descriptives of HbA_{1c}

Treatment Group	Statistic	Std. Error
Control		
Mean	8.976	0.2779
5% trimmed mean	8.950	
Median	8.900	
Variance	3.861	
Std. deviation	1.9650	
Minimum	5.3	
Maximum	13.5	
Range	8.2	
Interquartile range	2.8	
Skewness	0.122	0.337
Kurtosis	–0.350	0.662
Treatment		
Mean	7.967	0.2253
5% trimmed mean	7.907	
Median	7.900	
Variance	3.045	
Std. deviation	1.7450	
Minimum	5.0	
Maximum	12.4	
Range	7.4	
Interquartile range	2.6	
Skewness	0.460	0.309
Kurtosis	–0.010	0.608

measures in Table 2-1 can be found in any standard statistics text (e.g., van Belle et al., 2004).

One could convert the HbA_{1c} variable from interval to nominal by classifying patients as in control or not based on their HbA_{1c} level (e.g., less than 6.5 is in control). **Table 2-2** contains a cross-tabulation of HbA_{1c} by treatment group. Tables are often used to ferret out relationships between categorical variables. At end of study 12% of the control subjects are in control and 25% of the treatment group are in control.

The treatment groups are stratified by the presence or absence of hypertension (another categorical variable) in **Figure 2-2**. The dot above the whisker for treatment (without hypertension) is a potential outlier. What "story" does this figure have to tell?

When one wishes to examine the relationship between two interval variables, the scatter plot is a useful tool. In the nurse-based diabetes intervention study, one variable measured compliance (with good practices). **Figure 2-3** shows a scatter plot with HbA_{1c} on the vertical axis and of compliance on the horizontal axis. The data are stratified by treatment group. From the scatter plot one can see that as compliance increases, HbA_{1c} levels tend to decrease (for both treatment and control groups). This relationship can be measured with the correlation coefficient (r). For the treatment group $r = -0.44$, indicating that the two variables are inversely related (the negative sign for the correlation coefficient corresponds to the downward cast [negative slope] of the scatter plot as one goes from left to right).

Table 2-2 Cross-Tabulation of Treatment and Control Group of Variable HbA_{1c}

		A_{1c} In Control		
Treatment Group		**In Control**	**Not In Control**	**Total**
Control	Count	6	44	50
	% within treatment group	12.0%	88.0%	100.0%
Treatment	Count	15	45	60
	% within treatment group	25.0%	75.0%	100.0%
Total	Count	21	89	110
	% within treatment group	19.1%	80.9%	100.0%

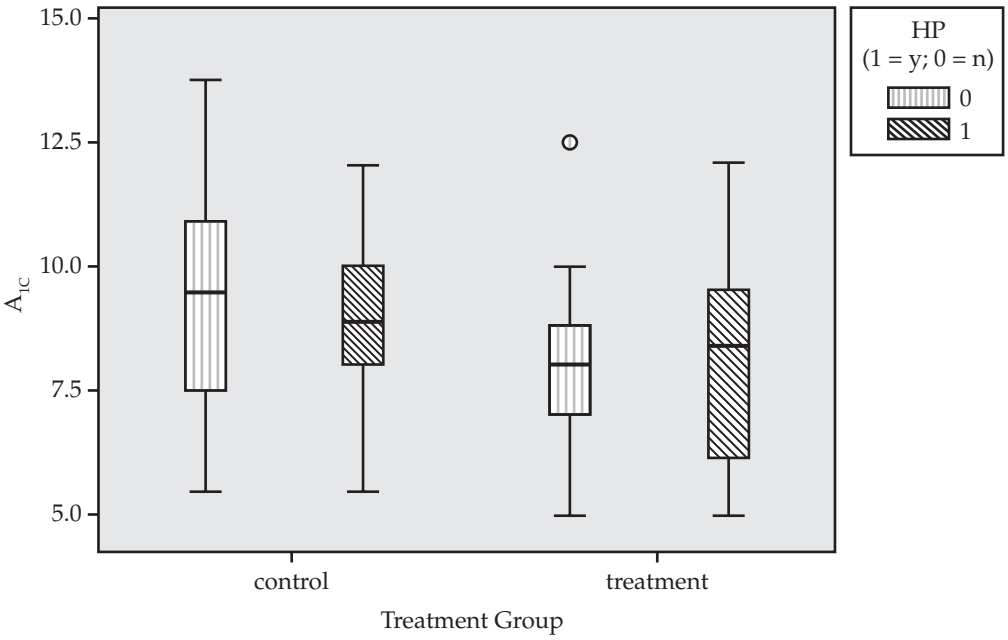


FIGURE 2-2 Box plot of variable HbA_{1c} with and without hypertension.

Inference

Recall that inference is about using samples and statistical theory to draw conclusions about populations. Inference can be divided into three activities: estimation, hypothesis testing, and model fitting (building). These are not strictly distinct activities (although the first two are often presented in that way). The last of these involves the first and often the second.

Estimation and Population Modeling

The underlying idea is that for some populations (collection of individuals, a process or phenomenon) there are parameters that describe this population (or that some probability model describes the population). The parameters might be simple measures such as a percentage, the mean, the SD, the 25th percentile (or any other quantile or percentile), or other characteristics of the population. For example, the population of interest might be home-birth babies in a particular county in 2009. One might wish to estimate the percentage of home-birth babies that are premature, the average weight in grams of home-birth babies, the SD of the weights of home-birth

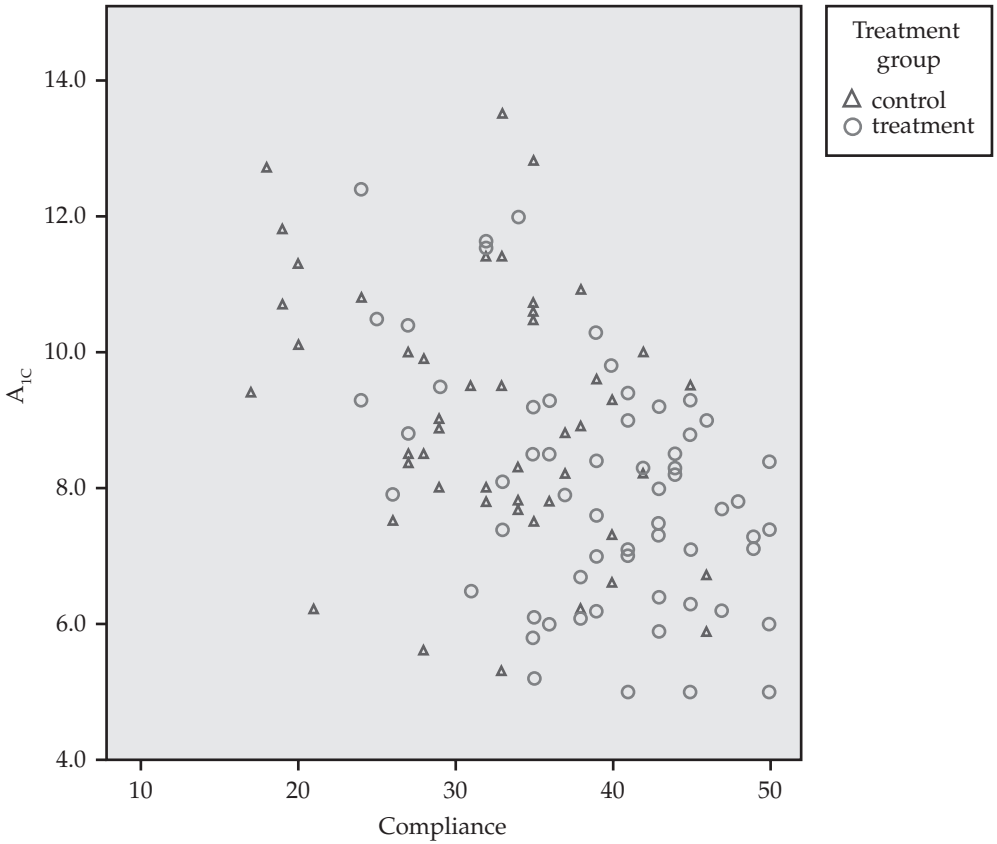


FIGURE 2-3 Scatter plot with HbA_{1C} on *y*-axis and compliance on *x*-axis.

babies, or the 25th percentile of the weights. Each of these requires particular methods that are not discussed here. We distinguish between two types of estimates: point estimates and interval estimates.

Suppose that in County X data on 99 home births have been collected by a health department nurse. Typically, before engaging in formal inferences, gaining an understanding of the nature of the population of interest is wise. Suppose the weights of the babies are of interest. This is an interval/continuous measure so methods appropriate for that level of measure are used. **Figure 2-4** is a histogram of the data (which includes summary statistics). Based on the sample mean and SD, a normal curve has been superimposed on the histogram. Based on the histogram, it appears that the "normal model" is an appropriate statistical model for the birth

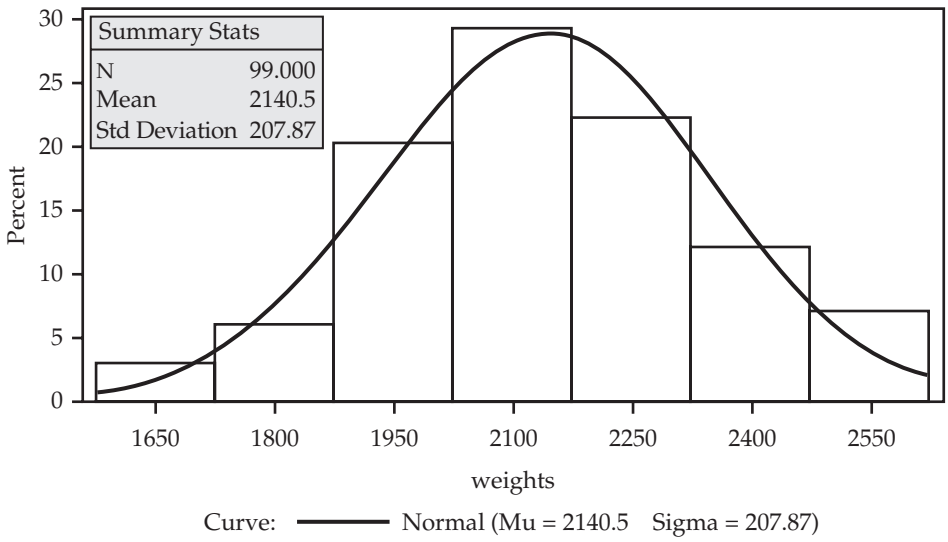


FIGURE 2-4 Histogram of baby weights superimposed with normal curve.

weight data. **Figure 2-5** shows a quantile-quantile plot of the data. This is a better tool for confirming normality (and can be used with any interval/continuous variable in fitting a particular distribution—probability model—provided the form of the model is completely specified). If the data conform to the distribution of interest, the plotted points fall along a straight line. This quantile-quantile plot offers no strong evidence against normality.

Researchers typically wish to estimate population parameters. Intervals of particular interest in this case are called confidence interval (CI) estimates. A point estimate is a single ("best number") guess. A CI is an interval with which there is associated a level of confidence. The level of confidence has a clear interpretation: If the "experiment" leading to construction of the CI is repeated many times, then the true population parameter will be trapped in the interval the percentage of times equal to the level of confidence. For example, in the home-birth study a 95% CI for mean weight is 2,099 to 2,182 g. This tells us nothing about whether the actual mean weight (pretend it is 2,150 g) is in the interval. It says only that if we use this method repeatedly, then 95% of the time the true mean will be captured in the interval; that is, if one constructed 100 intervals, then about 95 would contain 2,150. This is somewhat disappointing philosophically. Remedies have been suggested. In practice the CI approach leads to levels of belief in that there is stronger belief that the unknown

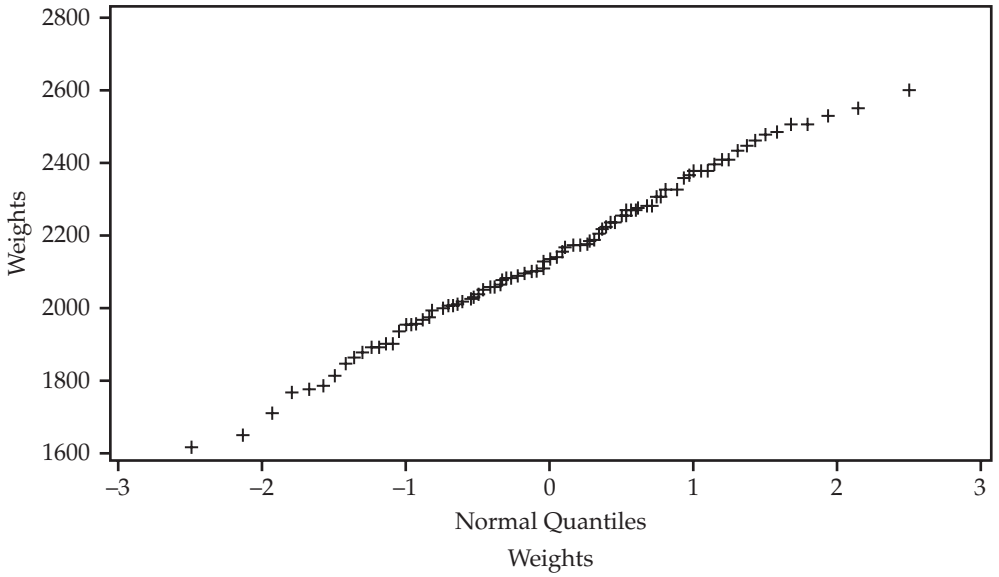


FIGURE 2-5 Quantile-quantile plot of baby weights.

parameter is in a 95% interval than a 90% interval. That is, the former interval is more likely to trap the parameter (which is achieved by widening the interval). Typically, interest focuses on the width of the interval, which is determined by the level of confidence (higher leads to wider), sample size (higher leads to narrower), and population variability (higher leads to wider). Hence, the typical interpretation of a 95% CI is a high level of belief that the parameter is in this interval. The researcher in the home-birth study is pretty sure that the average weight is in the interval 2,099 to 2,182 g.

A common misinterpretation of the CI is that 95% of babies weigh between 2,099 and 2,182 g. This is incorrect. To address this issue one needs a tolerance interval. A similar misuse would be to say that if a home birth is going to occur tomorrow, then there is a 95% chance the baby will weigh between 2,099 and 2,182 g. For this one needs a prediction interval. One could say that for the anticipated 200 home births next month the average weight of the babies will be between 2,099 and 2,182 with a high level of "confidence." Some other issues have crept in here and should be thought through. To continue this idea let's note that for epidemiological data any of the three types of intervals we have mentioned can be useful: CIs, prediction intervals, and tolerance intervals. **Table 2-3** contains interval estimates associated with birth weight data.

Table 2-3 Interval Estimates of Baby Birth Weights

Confidence	*k*	Prediction Limits	
Approximate Prediction Interval Containing All of k Future Observations			
95.00%	1	1,726	2,555
95.00%	2	1,665	2,616
95.00%	10	1,540	2,740
Approximate Prediction Interval Containing the Mean of k Future Observations			
95.00%	1	1,726	2,555
95.00%	2	1,846	2,435
95.00%	10	2,004	2,277
Prediction Interval Containing the Standard Deviation of k Future Observations			
95.00%	2	6.531	473.192
95.00%	10	112.654	311.573
Confidence	***p***	**Tolerance Limits**	
Approximate Tolerance Interval Containing at Least Proportion p of the Population			
95.00%	0.9	1,751	2,530

All the interval estimates in Table 2-3 are based on the assumption that the data are a sample from the normal distribution and were produced using SAS. The interval at the bottom of Table 2-3 has the following interpretation: With 95% confidence 90% of the home-born babies in County X weigh between 1,751 and 2,530 g. One might compare this with the observed quantiles (percentiles) of the baby weight sample that appear in **Table 2-4** (note that these quantiles in Table 2-4 are properties of the sample but could be used as estimates of the corresponding population parameters). Ninety percent of the sample weights are between 1,776 (the 5th percentile) and 2,506 (the 95th percentile). Tolerance intervals incorporate uncertainty with respect to estimation of the mean and the SD as well as population variability. The key idea is that the interval "traps" a certain proportion of the population with a given level of confidence. Notice that an interval of this type contains information of an entirely different nature than a CI estimate (which is a guess about a parameter—a fixed but unknown number). It is possible (but unlikely) that none of the babies has

Table 2-4 Birth Weight Descriptives (Percentiles)

Quantile	Estimate
100% max	2,601
99%	2,601
95%	2,506
90%	2,433
75% Q3	2,278
50% median	2,134
25% Q1	2,006
10%	1,875
5%	1,776
1%	1,621
0% min	1,621

weights in the CI for the mean. Often, what are called the natural tolerance limits are used to estimate a proportion of the population; that is, a natural 95% tolerance interval is given by mean ± 2SD. Be aware that infinitely many intervals contain 95% of the population but only one is symmetrical about the mean (and this symmetry is unlikely to produce the narrowest interval unless the population is symmetric).

Prediction intervals are used to answer questions about the future (but are based on the assumption that the future is stochastically identical with the period over which the data were collected). For example, suppose one wants to guess the weight on the next home-born baby in County X. A 95% prediction interval is 1,726 to 2,555 g (top of Table 2-3, $k = 1$). A prediction interval incorporates uncertainty (potential error) from three sources: estimation of the mean, estimation of the SD, and the inherent variability in the population (as measured by its SD). Observe that a prediction interval for the average weight of the next 10 babies to be born is narrower than the estimate for a single birth (2,004–2,277 g, with $k = 10$ in Table 2-3).

Hypothesis Testing

Hypothesis testing has to do with assessing the "truthfulness" of statistical statements (hypotheses) using samples and statistical theory (methods). In its most commonly

used and simplest form, it is a decision theoretic approach that leads to one of two mutually exclusive decisions regarding two mutually exclusive statements (the null hypothesis and the alternative hypothesis): One rejects the null hypothesis or one fails to reject the null hypothesis. The presumption is that there is a true state of nature (unknown to the researcher).

In testing a hypothesis one can commit two types of errors. A type I error is rejecting the null hypothesis when it is true. A type II error is failing to reject the null hypothesis when it is false. In repeated replications of an "experiment," these errors will occur with probability one (that is, with certainty). Why? Because the decision is based on a sample and sampling error will occur. The probability of a type I error is denoted alpha (α), called the level of significance of the test when alpha is expressed as a percentage. The probability of a type II error is a function of the value of the unknown parameter (in this example the true average birth weight). In most studies the level of significance is chosen before conducting the test (ideally before conducting the study). Typically, the mystical 5% level of significance is chosen because this has become something of an established scientific norm.

How do hypotheses arise? Typically, hypotheses arise from some sort of mind experiment. For example, a researcher wonders whether the average birth weight in this county is like the national average, which is 2,000 g. The null hypothesis is mean in County X = 2,000 g. The alternative can take many forms, but the typical two-sided alternative is mean in County X is not 2,000 g. Based on a sample, the researcher decides the hypothesis (this is equivalent to choosing between the null and alternative hypotheses). The basis for the decision is quite simple (but not one that is without critics): If the data are "inconsistent" with the null hypothesis, then one rejects the null hypothesis; otherwise, one fails to reject the null hypothesis. Now that statistical software packages are readily available, this decision is usually based on a *p* value, which is defined as the likelihood of observing a sample as different from the one hypothesized (in the null hypothesis) assuming that the null hypothesis is true. Along with this one must also make certain assumptions about the data to actually calculate a *p* value. One might ponder this for a moment to wonder why one supposes the null hypothesis is true to begin the formal analysis. For most parametric tests this is a very precise question. Let's continue the previous example in which a sample of 99 babies was selected. Having been to college, the researcher decides to perform a two-sided one-sample *t*-test. The test *t* is given by

$$t = \frac{(\text{sample mean}) - (\text{hypothesized mean})}{s / \sqrt{n}} = \frac{2{,}140.5 - 2{,}100}{207.87 / \sqrt{99}} = 1.939$$

The p value = $2P(t > 1.939) = 0.0555$, in which the probability is based in the t distribution with 98 degrees of freedom (this test is described in detail in nearly all methods texts, e.g., van Belle et al., 2004). The t-test is appropriate because we verified the sample was "approximately normal" via the quantile-quantile plot. Now how is one to interpret this p value? Using the classical decision theoretic approach is straightforward: $p >$ alpha, then one fails to reject the null hypothesis (mean weight = 2,100 g). That is, there is not sufficient evidence to conclude that the null is false. Does the researcher believe that the average weight of home-born babies in Country X is 2,100 g? Recall that the point estimate was 2,140.5 g and the CI for the mean was 2,099 to 2,182 g. The issue can be dodged by saying that there is not sufficient evidence to conclude that County X is different from the national average for home-born baby birth weights.

Why is rejecting the null hypothesis 1 time in 20 of particular interest? That is, why is the 5% level of significance so embedded in scientific practice? One might wonder what the effect on science would have been had the generally accepted level of significance been 2.5%. Or 10%. There is no doubt that the choice of 5% has had serious ramifications.

Power

The power of a test is a measure of the likelihood of rejecting false hypotheses. This is of some importance because the null hypothesis is usually thought to be false. The power can be assessed prospectively (at the time the study is planned) or post hoc (after data collection and test selection). The former is primarily aimed at avoiding futility and in sample size selection. The latter is frequently used for explaining away nonsignificant findings. Preplanned power calculations typically require a set of assumptions (often including a statistical model for the phenomenon of interest) and should be regarded with some healthy suspicion. These power calculations are based on two critical choices: the null value of the parameter and a second value corresponding to the minimum clinically significant difference between the null value and this clinically significant value. In the home-born baby example, one might say something like I want to be 90% sure the null hypothesis will be rejected if the average weight in County X differs from the national average by 25 or more grams. After deciding on the test statistic and making assumptions about the distribution of weights, one can calculate power estimates to "determine" the appropriate sample size. Ideally, a considerable amount of sensitivity analysis ("what if" analysis) accompanies this sample size decision.

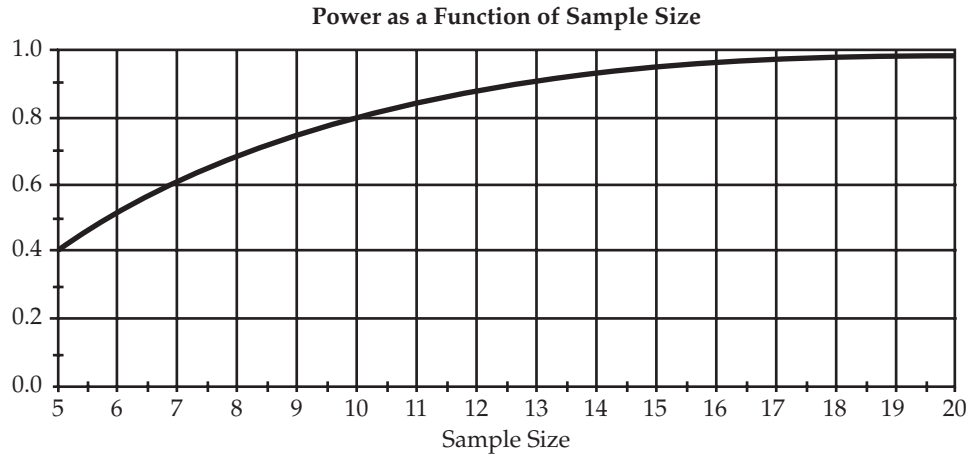


FIGURE 2-6 Power curve.

Graphs are an efficient method for doing what if analyses. The power curve in **Figure 2-6** is for a one-sample *t*-test for the mean with level of significance 5% and effect size $d = 1$. The effect size d is the magnitude of the difference between the null mean and the supposed mean measured in SDs. (See Cohen [1992] for a discussion of effect sizes.) For d (often applied to one-sample *t*-test, paired *t*-test, and two independent sample *t*-tests) an effect size of less than 0.2 is small, 0.8 is large, and 0.5 is medium. The effect size in Figure 2-6 corresponds, for example, to the null hypothesis in that the mean = 0 when the true mean is 1 and the population SD is 1. The power curve in Figure 2-6 has sample size on the horizontal axis and power on the vertical axis and applies to a specific test, effect size, and level of significance. The power increases as the sample size increases. If the effect size was smaller (e.g., 0.5), then the power would be lower (the curve would be below the one in Figure 2-6).

Power and Usefulness of Estimation

The key problem with classical hypothesis testing is that its primary use is in establishing what is not true when in fact scientific interest usually is in the truth. That is, for example, one wants to know something about the weights of home-born babies in County X. Suppose County X is in an affluent suburban area and researchers can reasonably guess that the birth weight distribution differs from the national one. In this

case the null hypothesis is essentially a straw man. This is not atypical. Does a company engage in a clinical trial of a drug believing that it is no better than placebo? One solution to this problem is to report an effect size as well as a *p* value. Strictly speaking, effect sizes are unitless measures. For example, the effect size for this one-sample *t*-test is (mean – sample mean)/SD so that the effect size for the birth weight example is $d = (2140.5 - 2100)/207.87) = 0.192$, which is a small effect using Cohen's suggestions. This makes the effect size universally interpretable but at a price that is addressed in the next paragraph.

The advantage of a CI is that it is in the units associated with the analysis and makes a direct statement about the parameter of interest in terms that are easy to understand clinically; for example, with 95% confidence the mean weight of home-born babies in County X is between 2,099 to 2,182 g. The estimated difference from the national average is between 2,099 – 2,100 = –1 g and 2,182 – 2,100 = 82 g. The CI for this discrepancy is (–1, 82), which contains zero and hence indicates nonsignificance at the 5% level of significance. However, with 95% confidence home-birth babies in County X could be as much as 82 g on average higher (or as much as 1 g lighter) than the national average.

One approach to the significance issue is to carefully separate statistical significance from clinical significance. Strictly speaking, one establishes statistical significance first, which is tantamount to saying that the observed "difference" is not (or is unlikely) to be explained by random chance. It is important to realize that statistical significance is conditioned on the effect size as well as the sample size (and the predetermined level of significance). In a sense the sample size is only important in its effect on accuracy; it is not an attribute of the population. Having established statistical significance, one looks at the difference in terms of its clinical significance. In the above example, using the classical approach the question is answered once the *p* value is produced. A more nuanced approach might ask if a difference of as much as 82 g is important with respect to its health/disease consequences. By the way one might wonder whether the mean is that informative. One could look at the tolerance interval or estimated quantiles of the weight distribution to assess risks.

MEASURES OF DISEASE OCCURRENCE AND RISK

Measuring disease risk and occurrence is more complicated than might appear on the surface. The idea of risk seems easy to understand, and most people believe they understand it.

Incidence and Prevalence

Epidemiological measures of risk and disease occurrence require that a distinction be made between proportions and rates (even though these words are often used interchangeably, e.g., one speaks of immunization rates but is actually referring to the proportion [or percentage] immunized in a particular population). Time is implicitly or explicitly involved in these measures. Also, one should distinguish between snapshots (pictures of a population that are more or less instantaneous) and measures over time. The former can be described as prevalence measures and the latter as incidence measures.

Incidence Proportion

The notion of risk is widespread. The simplest measure of risk is a pure number (unitless) that might be thought of as a probability. If there are x new cases of disease during a 6-month period and N subjects are followed for those 6 months, then risk is defined as

$$\text{Risk} = \frac{x}{N} = \frac{\text{number of new cases of disease in time period}}{\text{number of subjects followed for time period}}$$

The risk defined above is also referred to as an incidence proportion or fraction. Although the time period may or may not be reported, it is critical to the proper understanding of the measure produced. Consider a nursing facility with 120 patients who at the beginning of the measurement period do not have bedsores. Over the next 6 months 30 of these patients develop bedsores. Then the risk or incidence proportion is 30/120 = 0.25 (or 25%). The difficulty is that the actual phenomenon under study is seldom this simple and carefully defined. And one might wonder whether this adequately describes the bedsore "situation" at the nursing facility with respect to the risk of developing bedsores because the dynamics of the situation have been removed. What are some of these dynamics? The exposure time for the patients who complete the study is 6 months. For the 120 patients this is $6 \times 120 = 720$ person months. What if some of these 120 patients are discharged during the period of study? Then the total amount of exposure is less than 720 person months. In this case the risk is actually greater than the incidence proportion because not every patient was exposed for the full 6 months. One solution is to more carefully define incidence proportion:

$$\text{Incidence proportion} = \frac{x}{N} = \frac{\text{number of new cases of disease in time period}}{\text{number of subjects at beginning of time period}}$$

Notice that this does not completely solve the problem of how to measure risk. For that reason a more complicated measure is needed.

Incidence Rate

The incidence rate takes into account the total exposure of the subjects during the study period.

$$\text{Incidence rate} = \frac{x}{T} = \frac{\text{number of new cases of disease in time period}}{\text{person time at risk}}$$

Consider the bedsores example. Let's suppose that of the 120 patients beginning the study 110 patients were in the facility for the entire 6 months, 5 were discharged at the end of month 1 and another 5 at the end of month 4. Then the total exposure is $(110 \times 6) + (5 \times 1) + (5 \times 4) = 685$ person months. Hence, the incidence rate is $30/685 = 0.0438$ per person month. For 6 months this becomes 0.26. Note that this is larger than the incidence proportion of 0.25 (per 6 months). Why? The incidence proportion understated the exposure. Typically, incidence rates are multiplied by a time factor so that the time unit is something familiar, such as person years (or in this case 100 to yield 4.38 cases per 100 person months).

There are two ways (methods) for counting cases when the disease of interest is potentially recurrent. One way is to count only first cases. The other is to count all the cases, allowing a single individual to have multiple episodes. Of course, this distinction does not apply when only one episode can occur, such as one ending in death. In the bedsores example a patient could have no bedsores at the beginning of the observation period, develop bedsores (say after 1 month), have the bedsores disappear (cured after an additional 6 weeks), and then have the bedsores recur at month 5 of the study. If the first episode approach is used, then the numerator is easily defined (the number of new first cases), but what about the denominator? The solution is to divide by the "time at risk of disease," which means that for those 30 cases of bedsores the clock stops ticking when they got bedsores (the first time). That is, their period of exposure ends.

Prevalence Proportion

Prevalence measures are attempts to describe a situation (phenomenon) at an instant in time. Prevalence addresses the question of how widespread the disease is in the population. For example, at the beginning of the previous surveillance there were 120 patients without bedsores. Then one might say at beginning of the study that the

prevalence proportion or fraction was 0/120 = 0.0. Let's suppose that the facility at the beginning of the study had 140 patients (20 with bedsores and 120 without bedsores). Because in the previous section we were interested only in new cases, we used the denominator of 120 because these patients were disease free. That seems appropriate in the context of measuring risk in the sense that how likely is it that a disease-free patient will cease to be disease free during the period of study? However, if one is interested in a snapshot of the bedsores situation at the beginning of the study, then one can reasonably argue that the denominator should be 140 and the numerator 20. That is, the point prevalence at the beginning of the study (baseline) is 20/140 = 0.14 (14%).

Now, what about the prevalence at the end of the study? Well, one calculation might be (20 + 30)/140 = 0.36 (36%). This supposes that the 20 initial cases of bedsores and the 30 new cases that arose over the 6-month study period still have bedsores at the end of the 6 months (end of study). Notice that as a practical matter the nurse doing this study might take a much more direct approach (provided prevalence is the only measure of interest). At the beginning of the study the prevalence proportion is 20/140, which could be arrived at by simple enumeration. Then after 6 months the nurse counts the number of patients with bedsores, denoting this as y and forms the fraction $y/140$ (provided the head count at the facility is 140). Note that $y \leq (20 + 30)$—unless patients arrive at the facility with bedsores. The actual number of cases of bedsores could be reduced by some of the 50 cases of bedsores being cured.

Regardless of how the point prevalence fraction is calculated, it is unitless and is a snapshot of the disease situation at some moment in time. For that reason prevalence is a much easier measure (in a simple setting such as this nursing facility) to understand. When we calculated the incidence proportion, we implicitly took for granted that the head count at the facility and prevalence proportion depended in their calculation on key assumptions: There were 140 patients in the nursing facility at baseline and the same 140 patients were in the facility at end of study. Actual counting can be more complicated than that. Furthermore, implicit averaging is taking place when one uses units such as months when patients might be discharged (or die) at any time during the month.

Inferences for Incidence and Prevalence

Recall that inference is a process in which one uses data and statistical theory to draw conclusions about populations based on samples. Most of the theory is based on the notion that the data (sample) arose from some sort of random sample. The simple methods presented in this section are based on simple random sampling, which can

loosely be defined as a sampling method in which each subject in the population is equally likely to be chosen (to be in the sample). Very seldom do epidemiological data meet this standard. One usually satisfies this assumption by arguing that the sample is representative of the population of interest. This often implies some limitation or restriction on what can be called the scope of inference. That is, what is the population to which the inferences apply? Questions of this type are sometimes outside the scope of statistics because they require subject matter knowledge to be properly answered.

Inferences About Incidence and Prevalence Proportions

Because the incidence proportion and the prevalence proportion are each proportions, inference in this case reduces to the study of proportions. If the underlying model for the data is assumed to be binomial (this is the typical assumption; see van Belle et al., 2004), then a $(1 - \alpha) \times 100\%$ (large sample) two-sided CI estimate for a proportion in which there are x "successes" out of N independent trials is

$$\frac{x}{N} \pm z\sqrt{\frac{\frac{x}{N}\quad 1-\frac{x}{N}}{N}}$$

Alpha (α) relates the CI to level of significance, the idea being that for a 95% CI, $\alpha = 0.05$. Success is equated with observation of the disease (or not, because symmetry prevails). The point estimate is for the proportion in x/N, z is a quantile from the standard normal distribution (the z with $\alpha/2$ in the upper right hand tail of the distribution), and the quantity to the right of z in the equation is the (asymptotic) standard error of the estimate (which is a measure of the sample to sample variation in x/N). These days CI estimates are almost always generated using statistical software. SAS 9.1 was used to generate **Table 2-5**, which contains the incidence estimate for the bedsores example. One would say with 95% confidence the incidence of bedsores is between 0.1725 and 0.3275. Or more likely, the researcher might believe it is highly likely that between 17.25% and 32.75% of patients will develop bedsores over a 6-month period.

There are several important features of the CI estimate: The width of the interval is determined by N, the level of confidence, and x/N. Larger Ns are associated with narrower intervals, the higher the level of confidence the larger z is, and hence the wider the CI. The closer x/N is to 0.5, the wider the interval. Now what is the proper interpretation of this interval? Previously, we formally defined CI. For practical purposes that definition is not useful. The phrase "between 17.25% and 32.75% of

Table 2-5 Incidence (Proportion) Estimate of Bedsores Along With Confidence Intervals

Proportion	0.2500
ASE (asymptotic standard error)	0.0395
95% Lower confidence limit	0.1725
95% Upper confidence limit	0.3275
Exact confidence limits	
95% Lower confidence limit	0.1755
95% Upper confidence limit	0.3373
Binomial proportion for disease status = bedsores	

patients" needs amplification. It is a fact that 25% of the disease-free patients in the nursing facility got bedsores during the observation period. What does the CI have to do with that? The answer is that the researcher was probably not interested (in a scientific sense) in this particular collection of 120 patients; rather, interest was in what these 120 patients tell us about some population of patients. Alternatively, the researcher might conduct a mind experiment that works something like this: Given another cohort of patients in the same facility, what can I expect to happen?

The second set of confidence limits is labeled "exact." These are also based on the binomial model but do not depend on the large sample properties of the binomial (asymptotic normality). There are also estimates that use a continuity correction factor as well as other methods designed to produce more accurate "coverage" probabilities (the likelihood before data collection that the parameter of interest is in the CI). Choices among these are partly a matter of taste. The key idea is that a CI is the preferred way to estimate incidence or prevalence proportions rather than a point estimate. A 90% CI (0.185, 0.315) for the incidence proportion is narrower.

Suppose that the researcher was motivated to conduct the bedsores study by an article in a nursing journal that stated the prevalence of bedsores for nursing facility patients was 22%. Based on the situation when the study began, the estimated prevalence was 20/140 = 0.14 (14%). Can the researcher conclude that the situation in her nursing facility is better that this national average? That is, can the discrepancy between the observed prevalence of 14% and the hypothesized prevalence of 22% be explained as being the result of random chance?

Table 2-6 Prevalence (Proportion) Estimates for Bedsores Along With Confidence Intervals

Proportion	0.1429
ASE (asymptotic standard error)	0.0296
95% Lower confidence limit	0.0849
95% Upper confidence limit	0.2008
Exact confidence limits	
95% Lower confidence limit	0.0895
95% Upper confidence limit	0.2120

Binomial proportion for disease status = bedsores.

One approach is to examine the 95% CI for the prevalence (0.085, 0.200) to see whether the hypothesized value of 0.22 is in the CI (**Table 2-6**). Because it is not, one can conclude that the prevalence of bedsores at the nursing facility is different from the national average (with level of significance 5%; this is where the $(1 - \alpha) \times 100\%$ comes into play). Alternatively, one can perform a formal test of the hypothesis that prevalence equals 0.22 versus the alternative that prevalence is not equal to 0.22. The *p* value for this test is 0.0276 (**Table 2-7**), which is less than 0.05; hence, one rejects the null hypothesis and concludes that the prevalence at the nursing facility is not 0.22. Occasionally, one might wish to test the one-sided alternative that the prevalence is less than 0.22 (one should be cautious about doing this). For the one-sided alternative the *p* value is 0.0138 (which of course is also less than 0.05). These two approaches to testing a hypothesis about a proportion (the CI approach and the *p* value

Table 2-7 Output for the Statistical Test of the Hypothesis That Proportion = 0.22 Including *p* Value

Test of H0: Proportion = 0.22	
ASE (asymptotic standard error) under H0	0.0350
Z	–2.2034
One-sided Pr < Z	0.0138
Two-sided Pr > |Z|	0.0276

approach) occasionally lead to different conclusions because the former is based on the observed proportion and the latter on the hypothesized proportion.

Inferences About Incidence Rates

The most commonly used statistical model for incidence rates is the Poisson model (Rothman, 2002, p. 133 or for a more technical description see Selvin, 2002, p. 80). The Poisson model has a single parameter called the average rate and is used to model counts data. A large sample CI for the incidence rate (based on the normal approximation) is

$$\frac{x}{T} \pm z\sqrt{\frac{x}{T^2}}$$

Recall that the incidence rate for the bedsores example was 30/685 = 0.0438 per person month. Hence, a 95% CI for the incidence rate is

$$\frac{30}{685} \pm 1.96\sqrt{\frac{30}{685^2}}$$

Hence, the lower confidence limit is 0.0438 – 1.96(0.0080) = 0.028 and the upper confidence limit is 0.059. That is, with 95% confidence the incidence rate is between 0.028 and 0.059 per person month. Using OpenEpi (Dean et al., 2009), one gets the results in **Table 2-8**, which show several of the many approaches to this problem. Note that the rates are in 100 person-months.

Table 2-8 Person-Time Rate and 95% Confidence Intervals: Per 100 Person-Time Units

	Lower CL	Rate	Upper CL
Mid-*P* exact test	3.009	4.38	6.173
Fisher's exact test	2.955		6.252
Normal approximation	2.812		5.947
Byar approx. Poisson	2.954		6.252
Rothman/Greenland	3.062		6.264

Number of cases, 30; person-time, 685. Results from OpenEpi, Version 2, open source calculator—PersonTime1.

Stratification

Often, it is useful to stratify epidemiological data. That can be planned or post hoc. If the post hoc analyses are done after data snooping (using data twice in an analysis such as examining the event rates in several districts and then comparing the district with the highest observed rate to the others in a formal statistical test is what is meant by "data snooping"), care should be taken to use the appropriate adjustments to the analysis. **Table 2-9** contains the bedsore data stratified by age. Both the incidence proportion and the incidence rate are shown for each of the three age groups.

Does the incidence proportion differ with respect to age group? The broader question is whether risk is related to age. There are several approaches to questions of this type. Two important factors are (1) the level of measurement one wishes to use for age (in Table 2-9 age is categorical) and (2) the measure of risk you wish to use. The case in which one treats age as interval/continuous is discussed below (see Regression Modeling; that section also includes logistic and Poisson regression modeling in which age is categorical). In this section the comparisons are done using the ANOM approach (for a comprehensive treatment see Nelson et al., 2005). This method has the advantage of producing a graphical solution to the problem (in the form of a decision chart).

Figure 2-7 contains the ANOM decision chart for the incidence proportion (produced using SAS). The null hypothesis is that the incidence proportions are the same for the three age groups. The center line has the overall incidence proportion (which is 0.25). The key idea is that if the three incidence proportions are "close" to one another, then they will be close to the center line. The observed incidence proportions for each group are plotted on the chart (top of the vertical lines). The dashed lines are the decision limits. If any one of the observed incidence proportions plots outside the decision limits, then one rejects the null hypothesis (at level of

Table 2-9 Data on Bedsores

Age	New Cases	*n*	6 Months	1 Month	4 Months	T	Incidence Proportion	Incidence Rate	Incidence Rate per 100
<60	5	30	27	2	1	168	0.167	0.030	2.976
60-70	10	50	44	3	3	279	0.200	0.036	3.584
70+	15	40	39	0	1	238	0.375	0.063	6.303
Total/All	30	120	110	5	5	685	0.250	0.044	4.380

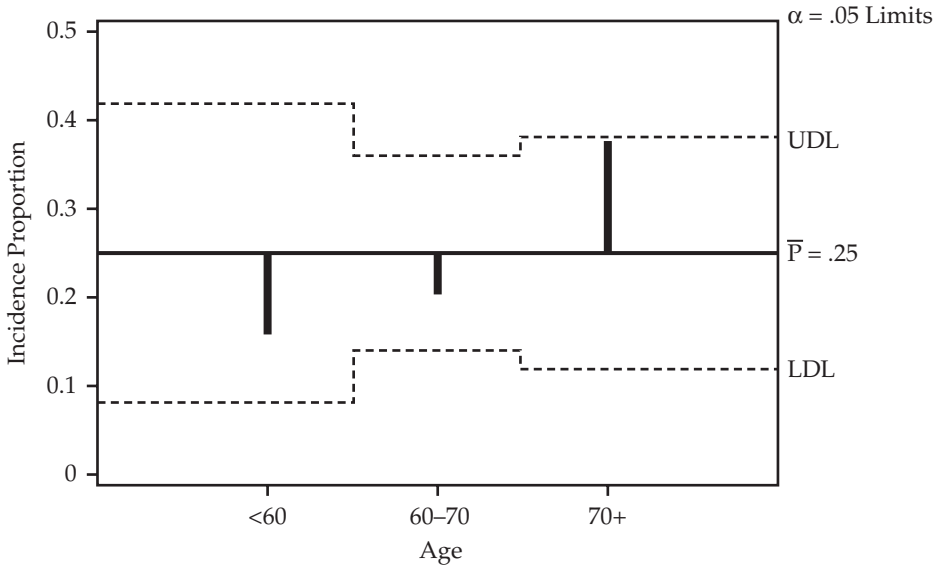


FIGURE 2-7 ANOM decision chart for incidence rate data.

significance = 5%). Because all three observed incidence proportions are within the decision limits, one cannot reject the null hypothesis. That is, there is not sufficient evidence to conclude that the incidence proportions are related to age.

There is some suspicion because the observed incidence proportions increase as one goes from left to right (perhaps this suggests an approach that takes this notion into account). If the level of significance increased (e.g., changed from 10% from 5%), the decision limits will move in toward the center line. The decision limits are different for each age group because the sample sizes are different. The vertical distance between the decision lines is inversely related to the sample sizes. The ANOM test for proportions is the analogue of the chi-squared test for the 3 by 2 table that can be constructed from the incidence proportion data. The underlying model is the normal approximation to the binomial.

ANOM analysis of the incidence rate data is based on the normal approximation to the Poisson model. The ANOM decision chart for the incidence rates (**Figure 2-8**) shows that the incidence rates differ among the age groups (at the 5% level of significance) because all three of the observed incidence rates plot outside the decision limits. The conclusions are that those younger than 60 and between 60 and 70 are at lower risk for bedsores than the overall average (the line labeled with $\overline{U} = 4.4$); those older than 70 years are at greater risk than the overall average. That is, one

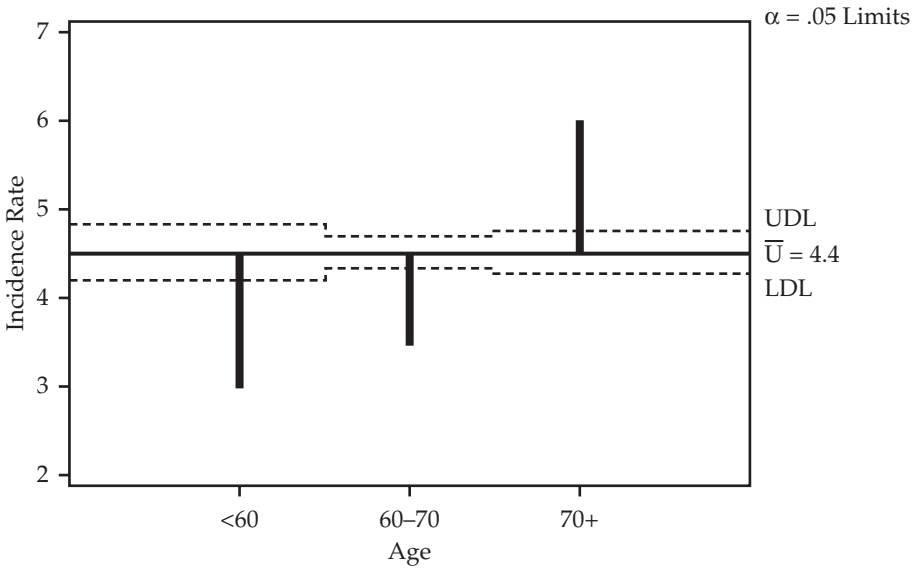


FIGURE 2-8 ANOM decision chart for the incidence rates.

rejects the null hypothesis that three incidence rates are the same; in addition one can see the nature of the differences.

Survival Analysis

Survival analysis and modeling has its origins in mortality studies; however, the time to any well-defined event can be studied using survival analysis models. Let the random variable Y be the time at which an event occurs. For example, in the bedsores example of the 120 subjects free of bedsores at the beginning of the study, 30 got bedsores by the end of the study (after 6 months or 24 weeks). Given that a patient gets bedsores, let Y denote the time in weeks at which the bedsores are cured. That is, we are interested in how long it takes for bedsores to be cured. In the most general setting there is a function called the survival function, denoted $S(t)$, which is the probability that the time Y is greater than t; i.e.,

$$S(t) = P(Y > t)$$

In our example, $S(t)$ is the probability that the bedsores are still present after t weeks. In practice the survival function must be estimated from data. There are two

general approaches: parametric estimation and nonparametric estimation. The former typically involves using models from the gamma family (e.g., exponential models), the Weibull, or others. In medical studies the nonparametric approach is the most commonly used (for a thorough discussion see Cantor, 1997). The most commonly used nonparametric estimate of the survival function is the Kaplan-Meier estimate.

Kaplan-Meier Estimation of the Survival Function

Several difficulties arise in estimating survival functions and survival distribution parameters. An example will help clarify these. Let's go back to the bedsores example. **Table 2-10** has data regarding the bedsores (of the 30 cases only a few are presented in Table 2-10 and the 30 are in no particular order) and contains something that in practice is not available—the actual duration of the disease in weeks. The patient identified as patient 1 had bedsores for 3 weeks. Patient 2 had bedsores for 25 weeks. But the study lasted for only 24 weeks, so this event was not actually observed during the course of the study. In fact, patient 2 did not acquire bedsores until week 10 of the study; hence, the maximum time that patient could be observed was 14 weeks.

So what do we know about that patient from observation? We know that the bedsores persisted for at least 14 weeks (see Table 2-10, Study Time). One approach to

Table 2-10 Data Tables Show Study Time

Patient	Week of Diagnosis	Maximum Study Time	Disease	Diabetic	Study Time	Duration
1	9	15	0	1	3	3
2	10	14	1	1	14	25
3	13	11	1	1	11	23
4	8	16	0	1	9	9
.	.	.	.	.	.	.
27	7	17	0	1	5	5
28	8	16	0	1	14	14
29	15	9	1	1	9	9
30	14	10	1	1	10	10

estimating the average time to cure for bedsores is to only count the complete cases, but this will lead to underestimating the average duration of the disease. In the jargon of survival analysis this patient whose cure was not observed is censored. In the Disease column patient 2 has a "1" recorded, indicating that at the end of the study bedsores were still present. Patient 1 has a "0" in the Disease column because the bedsores were acquired at week 9 and the duration was 3 weeks, meaning that the bedsores disappeared before the study ended. Study Time contains the time during which a patient was observed to have bedsores. That and the Disease column are all that is required to produce a Kaplan-Meier estimate of the survival function. **Figure 2-9** shows the survival function (SAS 9.1). Like all survival functions, the function is nonincreasing. That is, it drifts downward over time.

How is one to read (use) this estimated survival function? The survival probability is on the vertical axis and the time to cure (Study Time) is on the horizontal. What is the probability that a bedsore will still be present after 10 weeks? Find 10 on the horizontal axis and go up until intersecting the curve from about 0.4. That is, about

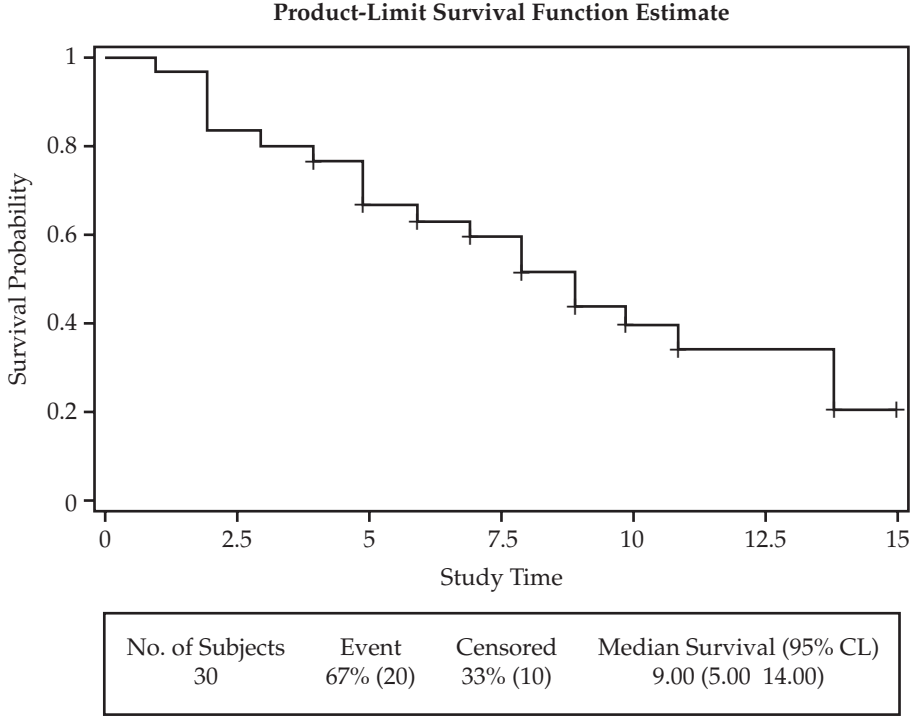


FIGURE 2-9 Survival function.

40% of the time a bedsore will take more than 10 weeks to be cured. What percent will be cured within 10 weeks? The relationship below tells us:

$$P(Y \leq t) = F(t) = 1 - P(Y > t) = 1 - S(t)$$

The probability that the bedsore is cured within 10 weeks is 1 minus the probability that it will take more than 10 weeks to cure. Hence, in about 60% of cases the cure will take place on or before week 10.

From Figure 2-9 one can see that of the 30 patients under study, 20 were cured and 10 were not (these censored times are indicated by a "+" in the graph). For example, there is a censored time at 4 weeks. The role of censoring in the construction of the curve is explained subsequently. The median survival time is 9.0; furthermore, with 95% confidence the median survival time is between 5 and 14 weeks. The median survival time is an estimate of the time by which 50% of cases of bedsores will be cured. More detail is available in **Table 2-11**, which has quartile estimates.

The estimated mean survival time (time to cure) is 8.62 weeks. The mean survival time was underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Details regarding the construction of Kaplan-Meier curves can be found in standard biostatistics texts (e.g., van Belle et al., 2004). For SAS details see Cantor (1997). The key idea is that the censored times do not alter the probability, but they reduce the number at risk at succeeding times. From the Kaplan-Meier (product limit) life table (**Table 2-12** is a partial listing), at time 1 week there was a cure. Because just before that there were 30 patients at risk (they had bedsores), the probability of a survival time greater than 1 week is 29/30 = 0.9667. There were four cures at week 2. The probability of surviving more than 2 weeks is (25/29) × 0.9667 = 0.8333 (note

Table 2-11 Quartile Estimates of Survival Time

		95% CI	
Percent	**Point Estimate**	**Lower**	**Upper**
75	14.0	10.0	
50	9.0	5.0	14.0
25	5.0	2.0	8.0

Table 2-12 Kaplan-Meier Life Table

Product-Limit Survival Estimates					
Study Time	**Survival**	**Failure**	**Survival Standard Error**	**Number Failed**	**Number Left**
0.0000	1.0000	0	0	0	30
1.0000	0.9667	0.0333	0.0328	1	29
2.0000	.	.	.	2	28
2.0000	.	.	.	3	27
2.0000	.	.	.	4	26
2.0000	0.8333	0.1667	0.0680	5	25
3.0000	0.8000	0.2000	0.0730	6	24
4.0000	0.7667	0.2333	0.0772	7	23
4.0000	.	.	.	7	22
5.0000	.	.	.	8	21
5.0000	.	.	.	9	20
5.0000	0.6621	0.3379	0.0871	10	19

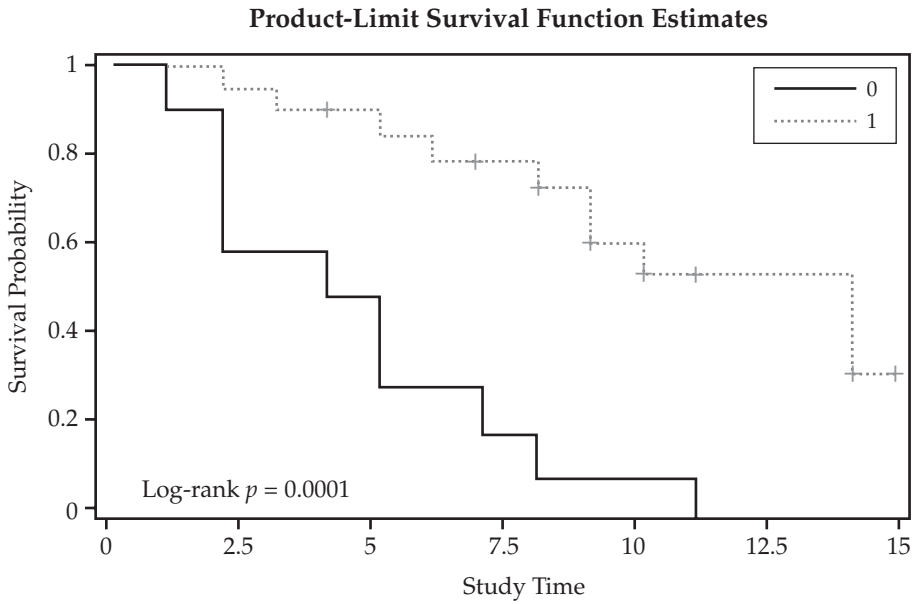
that the survival probability is not updated until the row corresponding to the fourth cure for that time period and hence the preceding three rows are filled with periods to indicate that no calculations were made). When censored times occur (the first one is at 4 weeks), the denominator (those at risk) is reduced, but the probability is not affected.

Comparing Survival Functions and the Log-Rank Test

An interesting application of survival analysis is its value in comparing groups. This arises frequently when comparing treatments (e.g., interventions and drugs) or groups (these can be demographic or prognostic, e.g., disease severity for breast cancer). In the bedsores study the nurse also recorded comorbidities, one of which was diabetes (see the Diabetic column in Table 2-10 in which "1" corresponds to diabetic and "0"

to not diabetic). The two survival functions are shown in **Figure 2-10**. From visual inspection one can see that the time to cure is longer for diabetic patients (the survival curve for diabetics is above the curve for nondiabetics). The median time to cure for diabetics is 14.0 weeks compared with 4.5 weeks for nondiabetic patients.

Is this apparent (observed) difference statistically significant? One commonly used test for homogeneity of survival curves is the log-rank test. In this example p = 0.0001 (chi-squared = 14.88 with 1 degree of freedom), so that the null hypothesis of homogeneous survival functions is rejected (e.g., 0.0001 < 0.05). The formal hypothesis being tested is that the survival functions are the same for all time points t. It is worth noting that the Kaplan-Meier estimate of the survival function is unbiased (it equals the true survival function on average) and $S(t)$ is asymptotically normal so CIs can be constructed at any time t.



	No. of Subjects	Event	Censored	Median Survival (95% CL)
0	10	100% (10)	0% (0)	4.50 (2.00 7.00)
1	20	50% (10)	50% (10)	14.00 (9.00 NA)

FIGURE 2-10 Two survival functions.

Analysis of 2×2 Tables

Relationships between categorical variables can be explored by means of tables. Fundamental to this type of analysis are 2×2 tables, which are tables with two rows and two columns. The focus here is in measuring risk. **Table 2-13** contains hypothetical data collected about entering students at a university. Data were collected by a nursing team at student health services. We assume that the data were derived from a random survey of incoming students (even though the convenient numbers such as 100 cigarette smokers are suggestive of a different design; see Chapter 3 for a discussion of designs such as case-control studies). The exposure of interest is cigarette smoking and the response is marijuana smoking (both self-reported).

The population about which these inferences apply are not discussed, but the scope of inference depends on the manner in which the data were collected and other factors related to the comparability of these students with the entire entering class or of this university to others. The study of tables can be extended to tables with more than two rows and two columns as well as to sets of tables. Sets of tables frequently

Table 2-13 Frequency of Cigarette and Marijuana Use

Table of Cigarettes by Marijuana

Cigarettes **Frequency** **Percent** **Row percent** **Col percent**	**Marijuana** **No**	**Yes**	**Total**
No	175 58.33 87.5 74.47	25 8.33 12.5 38.46	200 66.67
Yes	60 20 60 25.53	40 13.33 40 61.54	100 33.33
Total	235 78.33	65 21.67	300 100

arise in multicenter studies in which one controls for center effect or through stratification (e.g., by gender or major). Cochran-Maentel-Haenszel methods are often used for sets of tables.

An immediate measure of the relationship between cigarette smoking and marijuana use can be made by testing whether the two categorical variables are statistically independent. The standard test is the chi-squared test (in this example chi square = 29.7; $p < 0.0001$). Based on the p value, the hypothesis of independence is rejected. This is equivalent to rejecting the hypothesis that the percentage of cigarette smokers that use marijuana is the same as the percentage of non–cigarette smokers that use marijuana. This conclusion, although interesting, is not as informative (useful) as risk estimates, which are discussed in the next section.

Risks, Relative Risks, and Odds Ratios

Is the risk of marijuana smoking the same for those who smoke cigarettes and those who do not? The (estimated) risk of marijuana smoking by those who do not smoke is 25/200 = 0.125 (12.5%) compared with 40/100 = 0.4 (40%) for cigarette smokers. Another measure of risk uses the notion of odds. In general, the odds of an event are the risk of the event divided by the probability the event does not occur; that is, odds = risk/(1 – risk). The odds of marijuana use in smokers is 0.4/0.6 = 0.67. Note that the odds are less than 1.0, which means there is less than a 50% chance the event (marijuana use) occurs. For those not smoking cigarettes the odds (in favor of) marijuana use are 0.125/0.875 = 0.143.

Two commonly used measures useful for comparing risks are relative risk (RR) and the odds ratio (OR). The former is much more intuitive but can be meaningfully calculated only under certain data collection circumstances (which is addressed in Chapter 3). The RR = 0.4/0.125 = 3.2. That is, the risk of marijuana use among smokers is 3.2 times greater than that of nonsmokers. A 95% CI for the RR is (2.06, 4.96); that is, the risk is between 2.06 times greater and 4.96 times greater for cigarette smokers. Because this CI does not contain 1, the RR is statistically significant at 5% level of significance. The OR = 0.67/0.143 = 4.667 (95% CI: 2.61, 8.33); that is, the odds of marijuana use among cigarette smokers are between 2.61 and 8.33 times greater with 95% confidence. Again, this CI does not contain 1, which means the result is statistically significant at the 5% level of significance.

It is worth pointing out here that the measures of relative risk (RR and OR) are best understood in the context of absolute risk. If the exposed are 30 times more likely to be diseased than the unexposed but the absolute disease risk is 1 per million (per year) for the unexposed, then how daunting is the exposure? This shortcoming of RR

and OR as comparative measures is addressed in the next section by presenting additional measures.

Effect Measures Including Attributable Risk and Etiological Fraction

Does cigarette smoking cause marijuana smoking? This question makes a different impression than asking whether asbestos exposure causes cancer. Nevertheless, the first question raises the issue of causal effects. We have just finished comparing the risks of marijuana smoking for smokers and nonsmokers (compared the exposed with the nonexposed). Can we be certain that the differences in risk are attributable to exposure? "The ideal comparison would be of people with themselves in both an exposed and unexposed state" (Rothman, 2002, p. 45). This would be the counterfactual ideal. Instead, we settle with comparing the exposed and unexposed, believing that the two groups are identical apart from exposure (in some circumstances this can be investigated using statistical or other methods).

Define the risk difference as the difference between the risk for those with exposure and the risk for those without exposure (there is an analogous measure for rates). This requires that from data at hand one can estimate disease risk given exposure (more about this in Chapter 3 in which study design is discussed). Some epidemiologists call this quantity attributable risk. When discussing effects, there is not complete consistency among epidemiologists with regard to definitions. For the most part we either use or modify the terminology in Rothman (2002). The key idea is that attributable risk (risk difference) is the portion of the incidence of a disease in the exposed that is associated with (due to) the exposure. "It is the incidence of a disease *in the exposed* that would be eliminated if exposure were eliminated" (Kirch, 2008, p. 54). Attributable risk is sometimes referred to as excess risk. The attributable risk can be used to define the relative effect:

$$\text{Relative effect} = \frac{\text{RD}}{\text{risk in unexposed}} = \frac{\text{AR}}{\text{risk in unexposed}} = \text{RR} - 1$$

The right-hand part of the equation is the result of elementary algebra. In the cigarette exposure example the risk difference is 0.40 – 0.125 = 0.275 and the relative effect is 0.275/0.125 = 2.2, which is RR – 1 = 3.2 – 1 = 2.2. The rationale for this is that when there is no exposure effect the relative risk is 1; hence, subtracting the 1 makes it an incremental measure. Note that the above risk difference is a point estimate (that is, 0.275 is based on a sample); hence, one could construct a 95% CI estimate, which is 0.169 to 0.381 (or 16.9 per 100 to 38.1 per 100). This indicates

considerable uncertainty regarding the true attributable risk based on this relatively small sample.

In addition to measuring the relative effect by dividing the risk difference by the risk in the nonexposed, one can calculate what is called the attributable risk fraction by dividing by the risk in the exposed. That is,

$$\text{Attributable fraction} = \frac{\text{RD}}{\text{risk in unexposed}} = \frac{\text{AR}}{\text{risk in exposed}} = \frac{\text{RR} - 1}{\text{RR}}$$

Again the right-hand expression is the result of elementary algebra. This fraction can be expressed as a percentage. In the cigarette exposure example the attributable fraction is 0.275/0.40 = 0.687, or 68.7%. Pushing this to its logical limits (achieved partly by assuming that all bias has been removed from differences between the exposed and unexposed groups), one can say that were cigarette smoking to end, the risk of marijuana smoking would be reduced from 40 per 100 to 12.5 per hundred in the exposed group, which is a 68.7% reduction. Be aware that there are a number of ways to express percentages. In our example the RR was 3.2 (the relative effect was 2.2), which corresponds to a 220% increase in risk (from 12.5 cases per 100 to 40 cases per 100, which is [(40 – 12.5)/12.5] × 100% = 220%).

The population attributable risk is the portion of the incidence of a disease in the population (exposed plus unexposed) that is a result of exposure. It is the incidence of a disease in the population that would be eliminated if exposure were eliminated. Hence, population attributable risk equals the disease incidence in the population minus the disease incidence in the unexposed. Note that the comparison is being made to the existing pattern of exposure. For the cigarette example the incidence of marijuana smoking in the population is 65/300 = 0.217 (21.7%). Hence, population attributable risk = 0.217 – 0.125 = 0.092. That is, were cigarette smoking to end, marijuana smoking would be reduced by 9.2 per 100 students. The population attributable risk fraction is found by dividing population attributable risk by the population incidence, which in this example is 0.092/0.217 = 0.423. This can be expressed as a percentage; that is, cessation of cigarette smoking would lead to a 42.3% reduction in the incidence of marijuana smoking in the student population (from 21.7 per 100 to 12.5 per 100). The comparison (denominator) is to the existing exposure–disease relationship.

Let's return to the question: Does cigarette smoking cause marijuana smoking in university students or some appropriate population? Until now we have seen that it appears to play a role. One might ask whether it preceded marijuana use. One might even have some theory or even biological evidence to back up such an assertion. Instead, we rely on a purely epidemiological approach by defining the etiological frac-

tion (EF) as the proportion of the cases in which the exposure played a causal role in its development. That is,

$$\text{EF} = \frac{N_{\text{ED}} - N_{\text{UD}}}{N_{\text{ED}}}$$

In the above equation N_{ED} is the number of exposed persons in the population that are diseased and N_{UD} is the number of unexposed persons in the population that are diseased. If one were estimating EF based on the data in Table 2-13, then

$$\text{EF} = \frac{N_{\text{ED}} - N_{\text{UD}}}{N_{\text{ED}}} = \frac{40 - 25}{40} = 0.375$$

In the preceding equation, EF is an estimate and the *N*s are observed counts.

Regression Modeling

Regression modeling has to do with building models to describe a phenomenon. In this chapter we consider only models for which there is a single response variable (sometimes called the dependent variable); models of this type are called univariate (whereas models with more than one response variable are multivariate). Typically, the modeler "fits" a function that explains the response. This function is typically a mathematical expression involving one or more explanatory (predictor) variables. One way to look at this is to think of the response as the left-hand side of an equation and the predictors as being on the right-hand side. In the case in which there is a single explanatory variable, the model is called "simple," and when there is more than one explanatory variable, the model is referred to as a multiple regression model. The function can take many forms, but the linear model is the most commonly used, where "linear" allows for various transformations of the explanatory variables to appear (such as quadratics or square roots).

Linear regression has been extensively studied. What distinguishes statistical models from mathematical models is the inclusion of stochastic properties in the model relationship. The nature of the stochastic component in the model influences how the model is fit, where fitting usually refers to estimating the parameters in the model. Often, these parameters include model coefficients in the function form.

The level of measurement of the response variable is typically used to describe the form of regression model one is using. In the case in which *y* is an interval continuous variable (or approximately so), standard regression modeling is used. For

example, the response might be systolic blood pressure (SBP) in obese patients. In the case in which y is dichotomous (or ordinal), one can adopt logistic regression methods. For example, the response might be A1c in control versus A1c not in control. When dealing with counts data, one can use Poisson regression (although there are other choices), which can be described as various generalized linear models. For example, the response might be emergency department visits during the last 6 months by asthmatic children.

Linear Models With an Interval Continuous Response

A linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

The response is y, the explanatory variables are x (of which there are k), the betas are the coefficients, and epsilon is the random error term. The random error term is intended to capture the effects of variables not in the model—these often include idiosyncratic characteristics of patients as well as variables either not considered by the investigator or not available to the investigator. The nature of the random error term affects inference and model fitting (this refers to estimating the betas as well as estimating parameters associated with the random error such as its variance). The mechanics and theory of fitting regression models is not our concern here. Computer packages are now used to fit models regardless of the method used. Once the model has been fit, the betas are replaced with numerical estimates and one may "predict" y by "plugging in" values for x. The betas are slope parameters and their estimates determine the linear change in y associated with changes in an x. For example, a fitted model might be

$$y = 140 + 3x_1 - 8x_2$$

Such an equation is often referred to as a regression equation. Suppose this model is used to predict SBP in obese individuals. The first predictor is the amount by which body mass index (BMI) exceeds 30 and the second predictor is gender (male = 0 and female = 1). The latter is called a dummy or indicator variable. Then the predicted SBP for a male with a BMI of 30 is 140. For each one unit increase in BMI, the expected SBP increases by three units. On average SBP is eight units lower for females. In this model the effect of BMI is independent of gender. In a model in which these factors interact, the slope coefficient for BMI would be different for males and fe-

males. Geometrically, this regression equation can be represented as two parallel lines (one for males and one for females). If there was an interaction, the lines would not be parallel.

One can construct CIs for the slope coefficients. If a CI does not contain 0, then the corresponding predictor is significant. In the final model typically only significant variables are included. The overall fit of the model can be assessed in several ways. The typical method for reporting effect size in regression is R^2. Higher values (as one approaches from below 1, the maximum value for R^2) indicate a greater effect size. Given a fitted model, inferences about a particular coefficient depend on the other variables in the model. A regression model can have several uses, including prediction, establishment of association (e.g., for the population under study gender and BMI are related to SBP), and assertions about the exact nature of relationship (for each unit increase in BMI one can expect a three-unit increase in SBP).

With respect to prediction there are two general types: One can estimate the average value for y given values for the predictors, or one can predict the value of y given values for the predictors. The point estimates are the same in either case, but the interval estimates differ. Prediction intervals (the second case) are wider than CIs for the average value of y. Consider the SBP model. Given a female with a BMI of 36, the predicted SBP is $140 + 3(6) - 8(1) = 150$. If a patient enters a clinic and one wishes to guess this patient's SBP, then one needs a prediction interval (note that this is not an interval for the actual SBP reading but the patient's true SBP). If one wishes to guess the average SBP in a large group of female persons with BMIs of 136, then one should use a CI.

Logistic Regression

Only dichotomous logistic regression is discussed. The basic idea is that the response can take on only two values, such as lived/died or blood sugar under 110 versus not under 110. Usually, interest is in estimating probabilities and odds ratios as well as identifying variables associated with occurrence of the event of interest. A dichotomous outcome can be coded (e.g., 0 and 1), but fitting a standard regression model to the data with the hope of using the regression equation to predict probabilities of particular outcomes in unlikely to be satisfactory because predicted values outside the interval [0, 1] may occur. There are also other problems associated with this approach.

The previous example concerning the relationship between cigarette smoking and marijuana use was based on a 2×2 table. Data of that type (and any $2\times k$ table) can be analyzed using logistic regression; however, the main benefits of logistic modeling

are that one can conveniently include more than one predictor and in addition one can use interval/continuous predictors without explicitly converting them to categorical variables.

The idea behind dichotomous logistic regression is to model the log odds of the event of interest. Let p be the probability of the event of interest (e.g., patient has a stroke). Hence, one fits a model of the form

$$\log \text{odds(event)} = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + b_k x_k$$

Having fit the model, one can recover estimates for p by exponentiating the result (raising the log odds to the power e), which produces an estimate of the odds of the event. Then $p = 1/(1 + \text{odds})$. Fortunately, computer packages can perform these computations.

Suppose the data for the study relating cigarette smoking and marijuana use can be stratified by gender (**Table 2-14**). Fitting a main effects logistic regression model with gender and cigarette smoking as predictors (these are main effects) showed that gender was not significant; however, fitting a model with the interaction between gender and cigarette smoking showed that the interaction was significant. **Table 2-15** contains parameter estimates for the logistic model. Cigarettes ($p < 0.0001$) and the cigarettes gender (cigarettes × gender) interaction ($p < 0.0001$) are significant. Gender alone was not ($p = 0.2379$) but was left in the model because the interaction was in-

Table 2-14 Frequency of Cigarette and Marijuana Use Based on Gender

Cigarettes	Marijuana	Gender	Count
Yes	Yes	M	10
Yes	Yes	F	30
Yes	No	M	30
Yes	No	F	30
No	Yes	F	12
No	Yes	M	13
No	No	M	75
No	No	F	100

Table 2-15 Logistic Regression: Analysis of Maximum Likelihood Estimates

Parameter		df	Estimate	Standard Error	Wald Chi-Square	$p > \chi^2$
Intercept		1	–1.2428	0.1548	64.4298	<0.0001
Cigarettes	No	1	–0.6935	0.1548	20.0627	<0.0001
Gender	F	1	0.1827	0.1548	1.3929	0.2379
Cigarettes × gender	No × F	1	–0.3666	0.1548	5.6049	0.0179

cluded (this is a matter of taste). **Figure 2-11** shows the predicted probabilities (risks of marijuana use on the vertical axis) from which one can see the nature of the interaction. Cigarette smoking in females is associated with a much higher marijuana use risk.

With logistic modeling one can use any mixture of interval and categorical predictors. Consider a variant of the bedsore analysis based on data from another nursing facility in which through a retrospective chart review data were collected on the incidence of bedsores during a 1-year period. Included in the data collected were age of the patient at diagnosis and whether the patient was diabetic. Data on 50 patients were collected (19 with bedsores) of which 22 were diabetic (**Table 2-16**). To investi-

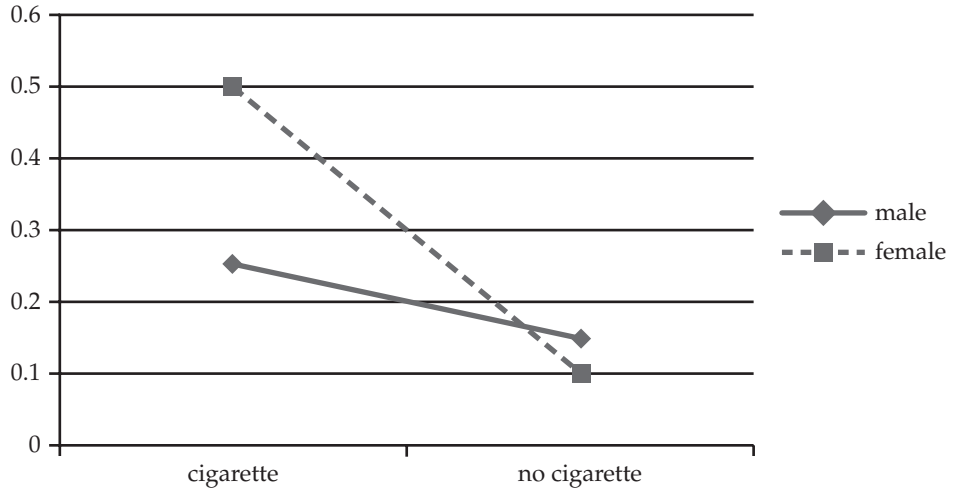


FIGURE 2-11 Predicted probabilities of marijuana and cigarette use.

Table 2-16 Frequency of Bedsores Based on Diabetes

	Bedsores		
Diagnosis/Statistic	**Present**	**Not Present**	**Total**
Diabetic			
Frequency	16	6	22
Percent	32	12	44
Row percent	72.73	27.27	
Column percent	84.21	19.35	
Not diabetic			
Frequency	3	25	28
Percent	6	50	56
Row percent	10.71	89.29	
Column percent	15.79	80.65	
Total	19	31	50
Percent of total	38	62	100

gate the relationship between bedsores and age as well as a diagnosis of diabetes, a logistic regression model was fitted. **Table 2-17** shows the coefficient estimates, from which one can see that both age ($p = 0.0056$) and diabetes ($p = 0.0007$) are significant. **Table 2-18** contains estimated ORs as well as CIs for them. Neither CI contains 1 (which would indicate no relationship), and the intervals are positive. The odds of bedsores are 263 times greater for diabetics (while accounting for age). Compare this with the OR of $22.2 = (16 \times 25)/(3 \times 6)$ from Table 2-16. The total picture is most easily portrayed in **Figure 2-12**, which contains the graphs of the estimated probability of bedsores for the two disease groups.

Table 2-17 Coefficient Estimates: Analysis of Maximum Likelihood Estimates

Parameter	**Diabetes Present**	***df***	**Estimate**	**Standard Error**	**Wald Chi-Square**	$p > \chi^2$
Intercept		1	−17.5	6.2318	7.9341	0.0049
Age		1	0.273	0.0986	7.666	0.0056
Diabetic diagnosis	Diabetic	1	2.768	0.812	11.6198	0.0007

Table 2-18 Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.314	1.083	1.594
Diabetic vs. not diabetic	253.68	10.517	>999.9

SUMMARY OF EPIDEMIOLOGICAL MEASURES OF DISEASE RISK

Table 2-19 is a general 2×2 table that can be used to estimate epidemiological parameters. The body of the table contains observed counts (e.g., n_{ED} is the number of exposed persons for whom the event [disease] occurred and n_{Ed} is the number of exposed persons for whom the event did not occur). The edges contain sums, all of which are sample values (e.g., N_E is the number of person in the sample that suffered exposure). **Table 2-20** shows via formulas the estimators (and in effect the definitions of the risk measures being described). Underlying this is some population from which this sample was collected. The implicit assumption is that the design (see Chapter 3) was such that the parameters in the table are estimable. To avoid unnecessary complications, symbols for the parameters are not given.

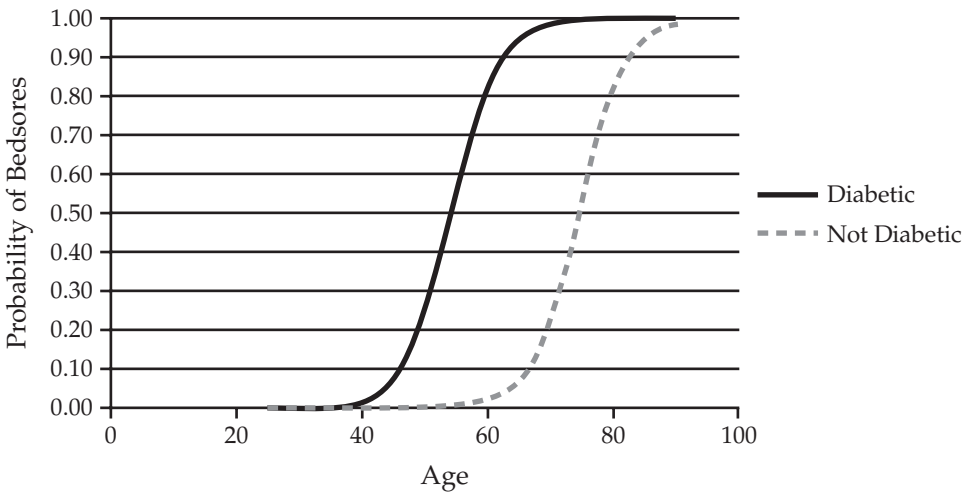


FIGURE 2-12 Estimated probability of bedsores for two disease groups.

Table 2-19 Exposed and Nonexposed Calculations

	Event (D)	No Event (d)	
Exposed	n_{ED}	n_{Ed}	$N_E = n_{ED} + n_{Ed}$
Unexposed	n_{UD}	n_{Ud}	$N_U = n_{UD} + n_{Ud}$
	$N_D = n_{ED} + n_{UD}$	$N_d = n_{Ed} + n_{UD}$	N

Table 2-20 Formulas

Parameter	Parameter Description	Estimator (formula)
r_E	Incidence (risk) of disease in exposed	n_{ED}/N_E
r_U	Incidence (risk) of disease in unexposed	n_{UD}/N_U
RR	Relative risk	r_E/r_U
OR	Odds ratio	$\frac{r_E}{1-r_E} \Big/ \frac{r_U}{1-r_U}$
AR	Attributable risk	$r_E - r_U$
ARF	Attributable risk fraction	$(r_E - r_U)/r_E$
PPE	Proportion of population exposed	N_E/N
r_P	Incidence (risk) of disease in population	N_D/N
PAR	Population attributable risk	$r_P - r_U$
PAF	Population attributable risk fraction	$(r_P - r_U)/r_P$
EF	Etiological fraction	$(n_{ED} - n_{UD})/n_{ED}$

Vital Statistics

Statistics of birth, death, marriage, and divorce rates in a community can be calculated by using the following formulas.

Crude birth rate is defined as number of live births per 1,000 population.

$$\text{Crude birth rate} = \frac{\text{number of births}}{\text{estimated midyear population}} \times 1{,}000$$

Crude death rate is defined as number of deaths per 1,000 population.

$$\text{Crude death rate} = \frac{\text{number of deaths}}{\text{estimated midyear population}} \times 1{,}000$$

Crude divorce rate is defined as number of divorces per 1,000 population.

$$\text{Crude divorce rate} = \frac{\text{number of divorces}}{\text{estimated midyear population}} \times 1{,}000$$

Crude marriage rate is defined as number of marriages per 1,000 population.

$$\text{Crude marriage rate} = \frac{\text{number of marriages}}{\text{estimated midyear population}} \times 1{,}000$$

Fetal death rate is defined as number of fetal deaths (20 weeks or more gestation) per 1,000 live births plus fetal deaths.

$$\text{Fetal death rate} = \frac{\text{number of fetal deaths (20 + Weeks Gestation)}}{\text{number of live births + number of fetal deaths}} \times 1{,}000$$

Fertility rate is defined as the total number of births in a year per 1,000 female population aged between 15 and 44 years

$$\text{Fertility rate} = \frac{\text{number of live births}}{\text{estimated midyear female population aged between 15–44 yrs}} \times 1{,}000$$

Infant mortality rate is defined as deaths to individuals less than one year old per 1,000 live births.

$$\text{Infant mortality rate} = \frac{\text{number of infants deaths}}{\text{number of live births}} \times 1,000$$

Maternal mortality rate is defined as the number of deaths as a result of complications of pregnancy, childbirth, abortion, or the puerperium per 10,000 live births.

$$\text{Maternal mortality rate} = \frac{\text{number of maternal deaths}}{\text{number of live births}} \times 10,000$$

Neonatal mortality rate is defined as deaths to individuals less than 28 days old per 1,000 live births

$$\text{Neonatal mortality rate} = \frac{\text{number of deaths} < 28 \text{ days}}{\text{number of live births}} \times 1,000$$

Perinatal mortality rate is defined as fetal deaths (20 weeks or more gestation) plus neonatal deaths (occurring in the first 27 days of life) per 1,000 live births plus fetal deaths.

$$\text{Perinatal mortality rate} = \frac{\text{number of fetal deaths} + \text{number of neonatal deaths}}{\text{number of fetal deaths} + \text{number of live births}} \times 1,000$$

Graphical Representation of Data

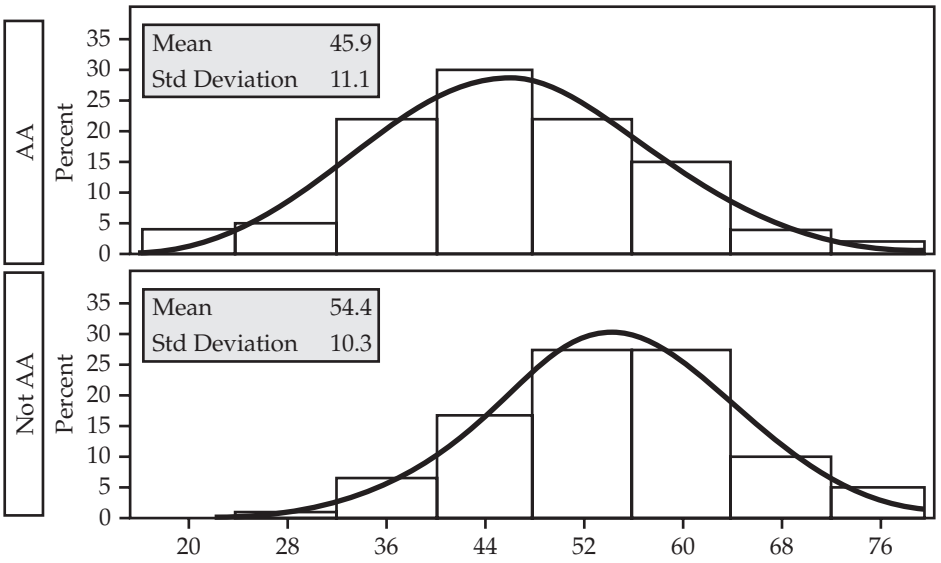
The data can be organized in the form of two-dimensional graphs. These visuals feed information faster and conserve readers' time and energy. To create these data graphs one can use Microsoft excel, PASW (SPSS), SAS, and many other statistical packages available in the market.

The following are examples:

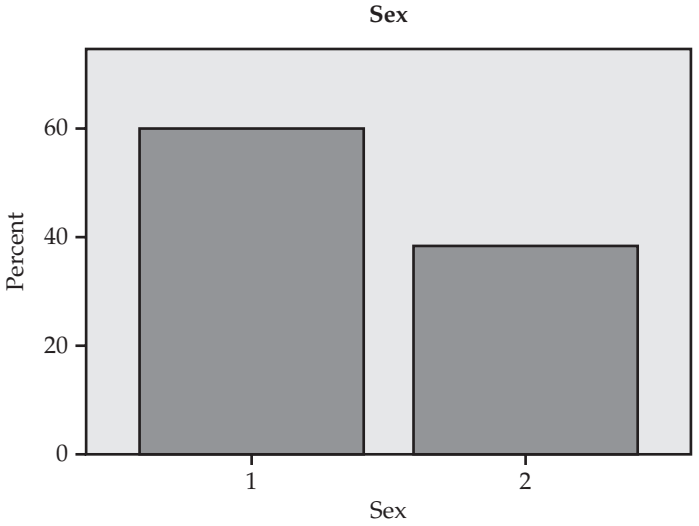
- The following are comparative histograms with superimposed normal curves. The histogram was generated using SAS, version 9.1 The vertical axis in each case represents the relative frequency (percent) for age at diagnosis. The data have been stratified by race (AA = African American; Not AA = not African American). Summary statistics are in the upper left-hand corner. Note that the

age at diagnosis is lower for African Americans; however, the variability is about the same.

Comparative Analysis of Age at Diagnosis With Superimposed Normal Curve

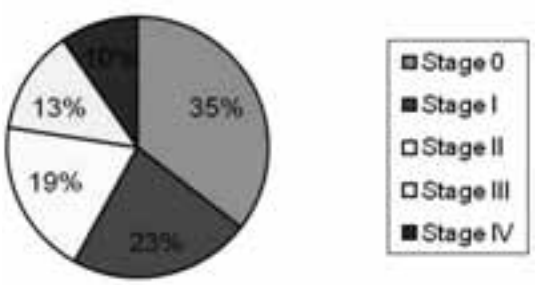


- The following is a bar chart created from gender (1 = male; 2 = female) percentage values involved in a study.

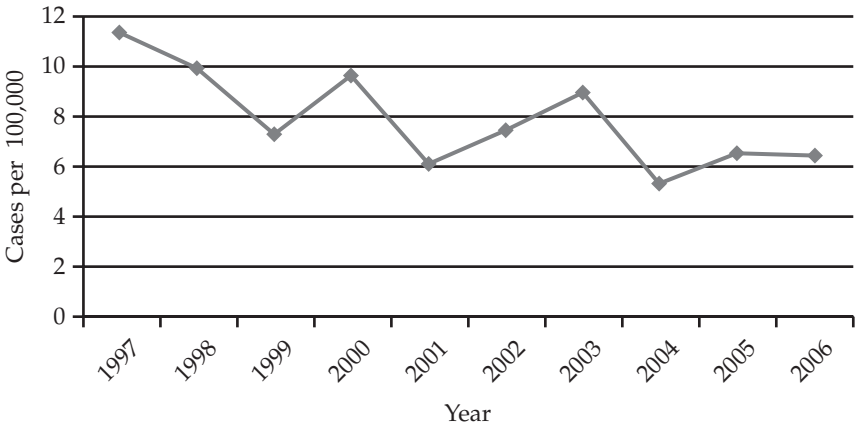


- These are cases at an urban clinic for a one month period. The table has cases and the pie chart has percentages.

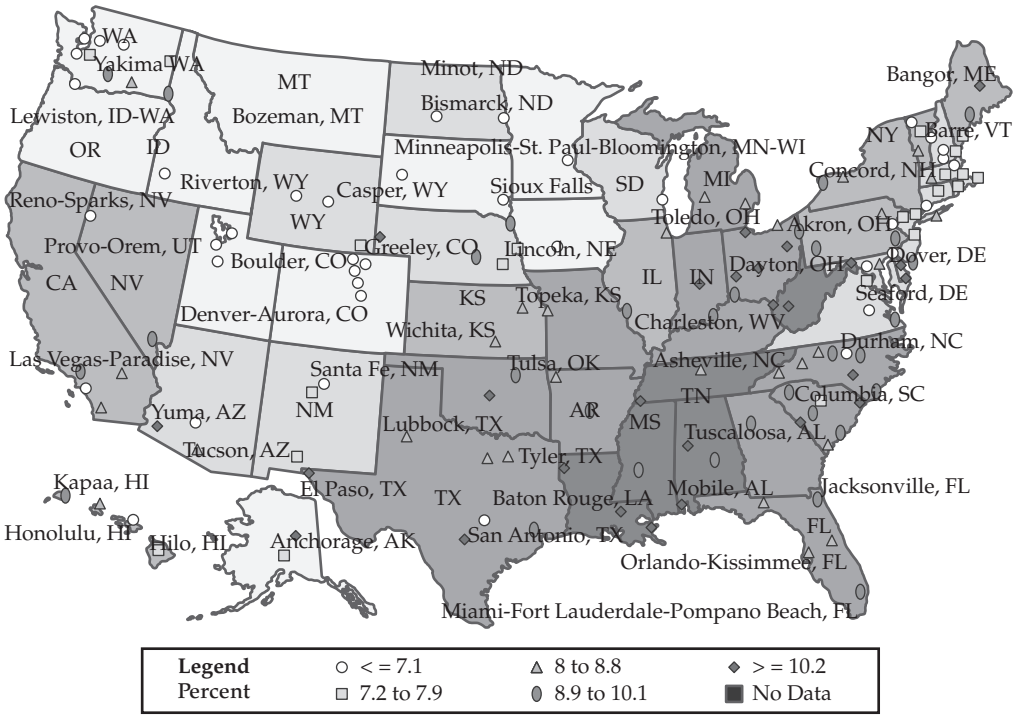
Stage	Cases
Stage 0	22
Stage I	14
Stage II	12
Stage III	8
Stage IV	6



- The following line graph was created based on 10-year incidence rates of campylobacteriosis.



- The following is an example of a geographical information system software output file. Health data based on zip codes, counties, and states can be shown in maps.



Source: CDC, http://www.cdc.gov/brfss/index.htm; Accessed on June 10th, 2010.

- P-P plots, Q-Q plots, and scatter plots already shown in the chapter are graphical tools for analyzing and presenting data.

KEY POINTS

- Epidemiology is concerned with disease. A common focus is on the relationship between exposure and disease where exposure can be taken to have quite a broad meaning.
- Almost all epidemiological data, especially data arising in clinical epidemiology, are samples.
- Samples can be described using graphical as well as summary measures. The level of measurement of data influences methods of presentation and analysis.
- Inference, drawing conclusions about populations based on samples and statistical theory, consists of estimation, hypothesis testing, and modeling. Estimation is fundamental, and interval estimates are almost always more useful than tests.

- All inference is subject to error associated with sampling. Regarding a particular decision concerning a statistical test, one never knows whether an error has occurred.
- Key epidemiological measures of disease risk are prevalence, incidence fraction, and incidence rate. Rates are measured in units of person-time exposure. Although rates are frequently thought of as more accurate measures of risk, they are more difficult to interpret and effectively communicate than risk measured by a proportion.
- Relationships between exposure and disease can be statistically analyzed (modeled) using an assortment of tools including (but certainly not limited to) CIs, prediction intervals, tolerance intervals, parametric tests (such as *t*-tests, chi-squared tests, ANOM, and analysis of variance), survival curves, 2×2 tables, and regression modeling (including logistic regression).

CRITICAL QUESTIONS

1. Why and in what manner are CI estimates superior to point estimates?
2. Suppose that a 95% CI for the odds ratio for a disease event given exposure compared with nonexposure is (0.95, 3.53). What does this tell you about the risk associated with exposure? What does this tell you about a formal statistical test of the hypothesis that the OR is 1.0 (at what level of significance?). Suppose a 90% CI is (1.05, 3.17).
3. Suppose a researcher is interested in SBP in a population of individuals in high stress jobs and takes SBP readings from 100 subjects selected from the population. What would be more useful: a 95% CI estimate for the mean SBP or a 95% tolerance interval estimate for 90% of the population? In the first case, what would the lower end point of the interval tell you? In the second case, what would the lower end point of the interval tell you?
4. In what circumstance is logistic regression modeling most useful in assessing the relationship between exposure and disease? Consider the level of measurement of the predictors and the number of predictors in your answer.
5. What are the consequences of ignoring (or being unaware of) interaction between exposure and some demographic characteristic (or between two exposures) in modeling the relationship between exposure and disease (or behavior)? Examine this question by constructing an example on which you can perform mind experiments associated with different scenarios such as synergistic interaction, antagonistic interaction, and no interaction.

6. In judging risky behavior, which is more informative: relative risk or attributable risk? Think of this in terms of making a decision yourself or advising someone considering engaging in risky behavior.

REFERENCES

Cantor, A. (1997). *Extending SAS® survival analysis techniques for medical research*. Cary, NC: SAS Institute.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Dean, A. G., Sullican, K. M., & Soe, M. M. (updates 2009/2005). *OpenEpi: Open source epidemiologic statistics for public health*. Retrieved from http://www.OpenEpi.com

Dorak, M. T. (updated April 2010). *Common concepts in statistics*. Retrieved from http://www.Dorak.info/mtd/glosstat.html

Kirch, W. (Ed.). (2008). *Encyclopedia of public health*. Danvers, MA: Springer.

Nelson, P. R., Wludyka, P. S., & Copeland, A. F. (2005). *The analysis of means: A graphical method for comparing means, rates, and proportions*. SIAM, Philadelphia, ASA, Alexandria, VA: ASA-SIAM Series on Statistics and Applied Probability.

Petrie, A., & Sabin, C. (2009). *Medical statistics at a glance*. Chichester, UK: John Wiley & Sons.

Rothman, K. J. (2002). *Epidemiology: An introduction*. New York: Oxford University Press.

Selvin, S. (2004). *Statistical analysis of epidemiological data*. New York: Oxford University Press.

van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: A methodology for the health sciences*. Hoboken, NJ: John Wiley & Sons.