## Chapter 2

# Principles of Statistics

The theoretical development of modern epidemiology was largely influenced by statisticians. Because of the primary role that statistics plays in modern epidemiology, this chapter will provide a brief review of several selected statistical concepts. This chapter will cover a general view of statistics in epidemiology, including basic statistical concepts such as data, data description, probability, sampling, estimation of statistics, hypothesis testing, decision errors, and estimation.

## STATISTICS IN EPIDEMIOLOGY

Depending on the type of problem to be solved, statistics can be divided into four areas: descriptive, probability, inferential, and statistical techniques. *Descriptive statistics* involves methods of organizing, summarizing, and describing numerical data. In epidemiology, we use descriptive statistics to study the distribution (frequency and pattern) of health-related states or events; that is, statistical methods are used in epidemiology to provide a description of the who, what, when, and where aspects of health-related states or events in selected populations.

*Probability* is used when discussing the chance or likelihood that a given event will occur. Probability is used extensively in epidemiology to assess the likelihood of experiencing an outcome, based on exposure information.

Methods of sampling and for assessing the validity of screening tests are based on probability. Probability also provides a basis for assessing the reliability of the conclusions we reach.

*Inferential statistics* involves making inferences about a population's characteristics from information in the sample. Epidemiologic studies often rely on sample data, with study findings used to draw inferences about what happened in the sample and in the world beyond the sample.

Finally, *statistical techniques* are analytic approaches that utilize statistical methods to investigate a range of problems. Epidemiology relies on a number of statistical techniques in its overall study of the distribution and determinants of health-related states or events and in evaluating public health interventions. A summary of basic statistical notation that will be used throughout this book is presented in Appendix A.

## BASIC STATISTICAL CONCEPTS

### Data

Statistics is the science of data, where *data* are pieces of information such as observations or measurements of phenomena of interest. *Statistics* involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. We obtain data by observing or measuring some characteristic or property of the population of interest. An *experimental unit* is an object (person or thing) upon which we collect data.

Data and variables are either quantitative or qualitative. *Quantitative data* are observations or measurements measured on a numerical scale (e.g., biometric scores, number of injuries, or dose of radiation). These data have a fixed interval or ratio scale. *Qualitative data* can only be classified into a group of categories and provide a general description of properties that cannot be described numerically (e.g., appearance, feelings, and tastes). Qualitative data have a nominal or ordinal scale.

*Scales of measurement* are the ways that variables are defined and categorized. *Nominal scale* refers to placing data into categories, where there is no logical order or structure (e.g., Yes or No). *Ordinal scale* refers to placing data into categories where the gross order of the categories is informative, but the relative positional distances are not quantitatively meaningful (e.g., a ranking). *Interval scale* refers to a measurement where the difference between the intervals is meaningful, but there is no true definition of zero (e.g., temperature, since zero on Fahrenheit or Celsius scales does not mean "no temperature"). *Ratio scale* has the same properties as interval scale, but a true zero is involved where if the variable is zero, there is none of that variable (e.g., height, weight, dosage, and Kelvin scale of temperature). Ratio scale data uses the same statistical techniques as interval scale data.

*Parametric statistics* is a branch of statistics concerned with data measurable on interval or ratio scales. The usual central measure is the mean. The assumed distribution is normal (a theoretical frequency distribution for a random variable that has a bell-shaped curve and is symmetric about its mean), and the assumed variance is homogeneous (constant). On the other hand, *nonparametric statistics* involves data measurable on nominal or ordinal scales. The usual central measure is the mode. Nonparametric statistics does not assume a specific distribution or variance.

## Describing Data

There are many approaches for evaluating and describing data (**Table 2.1**). In addition to frequency distribution tables and summary statistics for describing data, a number of graphs are listed in the table. Deciding when to use these graphs depends on the type of data and statistical measures being evaluated. A list and examples of several types of graphs are presented in **Table 2.2**. In general, tables and graphs are used to help clarify the public health problem. They identify patterns, trends, aberrations, similarities, and differences in the data (numbers, ratios, proportions, or rates). They are important for communicating public health information according to person, place, and time factors.

## Probability

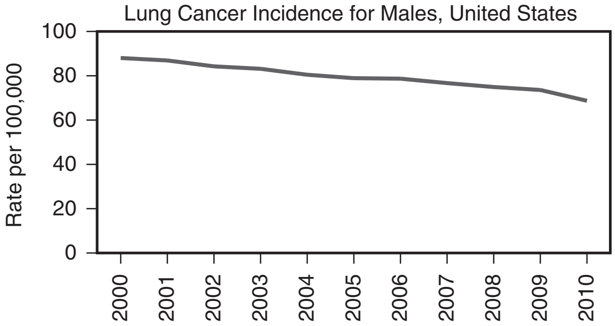
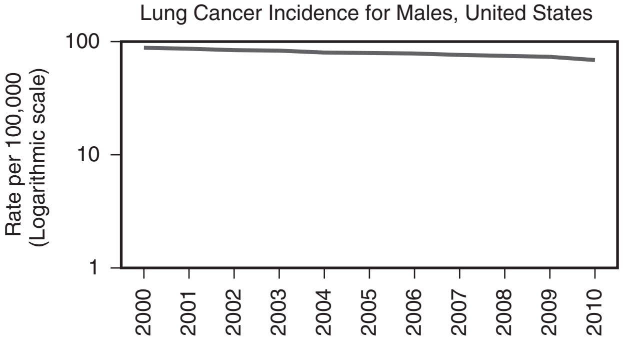
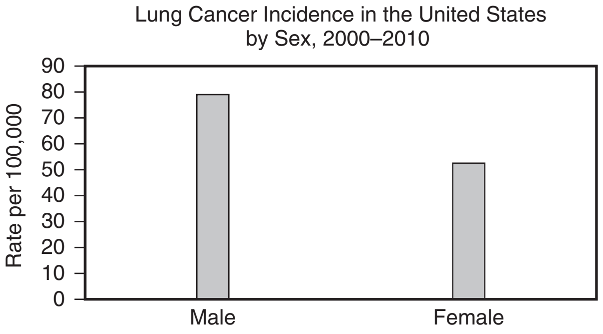
Probability provides a basis for assessing the reliability of the conclusions we make under conditions of uncertainty. Probability theory is applied

**Table 2.1 Scales of Measurement and Corresponding Statistics and Graphs**

| Scale | Description | Example | Statistics | Graphs |
|---|---|---|---|---|
| Nominal | Categorical observations | Exposed (Yes/No)<br>Diseased (Yes/No)<br>Sex, Race/Ethnicity, Marital Status, Education Status | Frequency<br>Relative frequency | Line graphs<br>Bar chart/pie chart<br>Spot map<br>Area map |
| Ordinal | Categorical observations<br>Order of categories is informative | Preference rating (e.g., agree, neutral, disagree)<br>Likert scale | Frequency<br>Relative frequency | Bar chart/pie chart |
| Interval or Ratio | Quantitative observations where the difference between the intervals is meaningful, but there is no true definition of zero (Interval) or there is a true zero (Ratio) | Dose of ionizing radiation<br>Number of fractures | Geometric mean<br>Arithmetic mean<br>Median<br>Mode<br>Range<br>Variance<br>Standard deviation<br>Coefficient of variation | Bar chart/pie chart<br>Histogram or frequency polygon<br>Box plot<br>Scatter plot<br>Stem-and-leaf plot |

Modified from Merrill RM. *Fundamentals of Epidemiology and Biostatistics: Combining the Basics*. Burlington, MA: Jones & Bartlett Learning; 2013: 29.

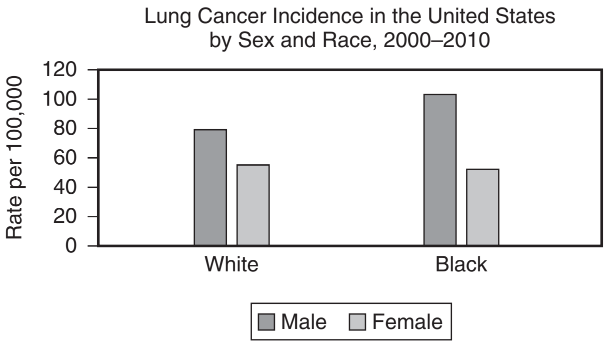
## Table 2.2 Graphs for Describing Data

| Type of Graph | Examples |
|---|---|
| Arithmetic-scale line graph | <br>Lung Cancer Incidence for Males, United States<br> |
| Logarithmic-scale line graph | <br>Lung Cancer Incidence for Males, United States<br> |
| Simple bar chart | <br>Lung Cancer Incidence in the United States by Sex, 2000–2010<br> |

(*continues*)

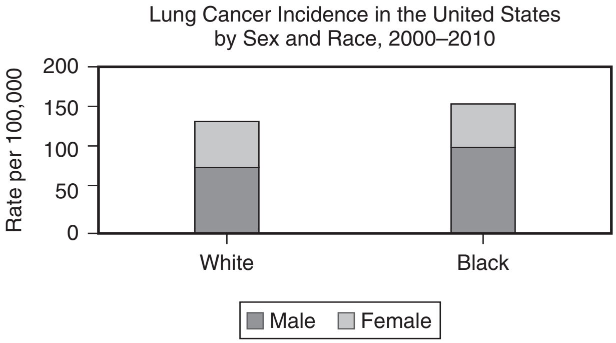
## Table 2.2 Graphs for Describing Data (Continued)

<table>
<tr><th>Type of Graph</th><th>Examples</th></tr>
<tr><td>Grouped bar chart</td><td>



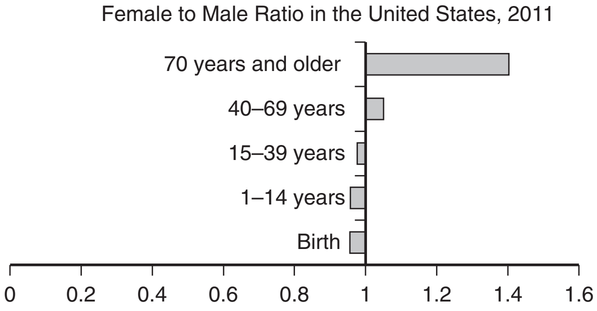
Lung Cancer Incidence in the United States by Sex and Race, 2000–2010

</td></tr>
<tr><td>Stacked bar chart</td><td>



Lung Cancer Incidence in the United States by Sex and Race, 2000–2010

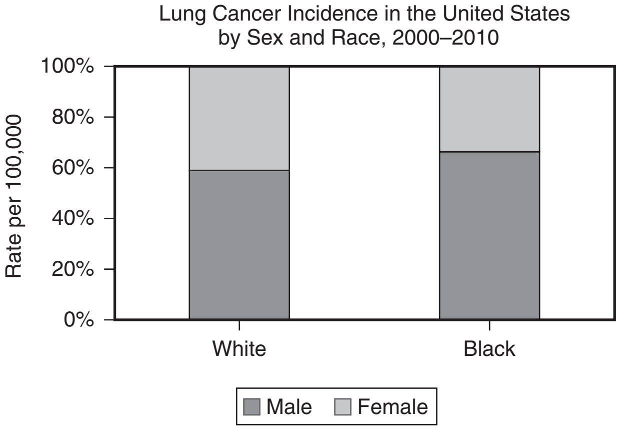
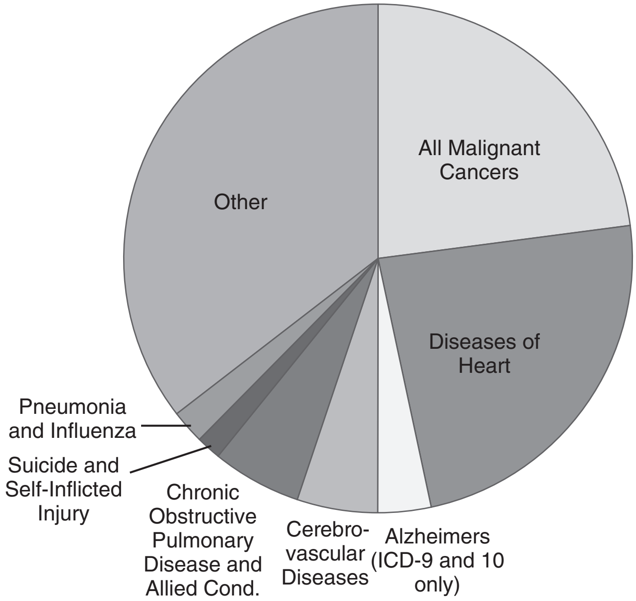
</td></tr>
<tr><td>Deviation bar chart</td><td>



Female to Male Ratio in the United States, 2011

</td></tr>
</table>

## Table 2.2 Graphs for Describing Data (Continued)

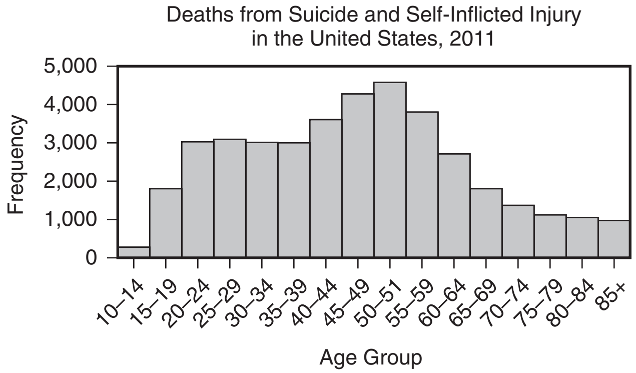
| Type of Graph | Examples |
|---|---|
| 100% component bar chart |  |
| Pie chart | Deaths in the United States, 2011<br> |

(*continues*)

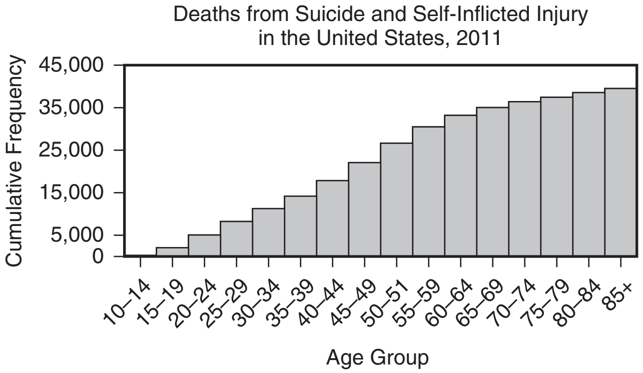
## Table 2.2 Graphs for Describing Data (Continued)

**Type of Graph** **Examples**

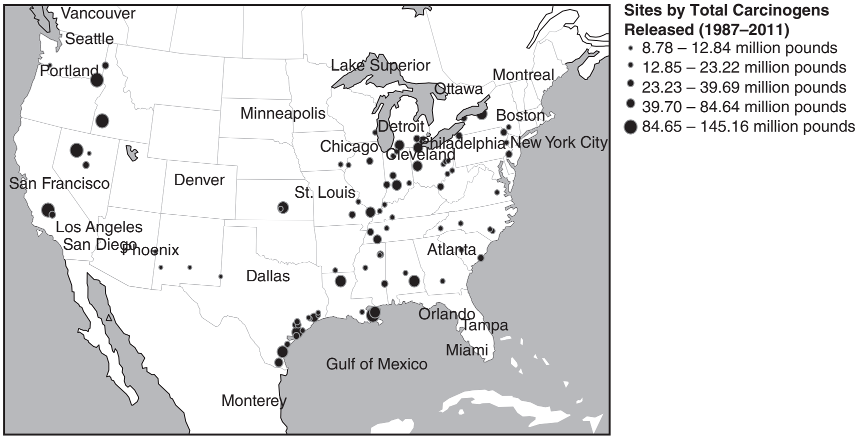
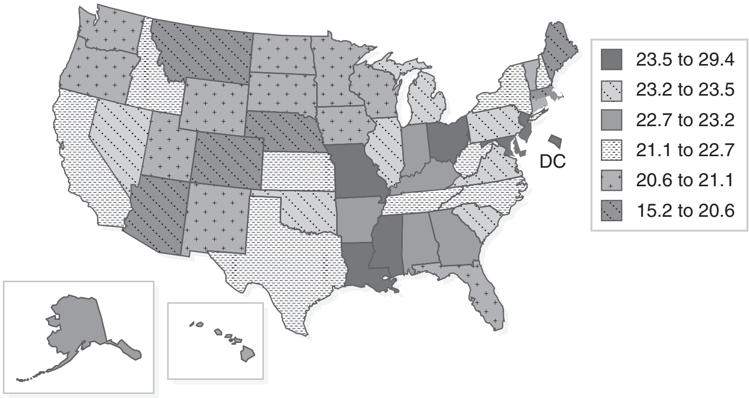
Histogram



Cumulative frequency



## Table 2.2 Graphs for Describing Data (Continued)

| Type of Graph | Examples |
|---|---|
| Spot map | Top 100 Toxic Sites: Carcinogens (Data courtesy of the U.S. EPA's Toxic Release Inventory), 1987–2011<br><br> Vancouver<br>Seattle<br>Portland<br>Lake Superior<br>Ottawa<br>Montreal<br>Minneapolis<br>Boston<br>Detroit<br>Chicago<br>Cleveland<br>Philadelphia<br>New York City<br>San Francisco<br>Denver<br>St. Louis<br>Los Angeles<br>San Diego<br>Phoenix<br>Atlanta<br>Dallas<br>Orlando<br>Tampa<br>Miami<br>Gulf of Mexico<br>Monterey<br>Sites by Total Carcinogens Released (1987–2011)<br>8.78 – 12.84 million pounds<br>12.85 – 23.22 million pounds<br>23.23 – 39.69 million pounds<br>39.70 – 84.64 million pounds<br>84.65 – 145.16 million pounds |
| Area map | Age-Adjusted Death Rates per 100,000 in the United States, 2007–2011<br>Female Breast Cancer, All Races (including Hispanic)<br><br> 23.5 to 29.4<br>23.2 to 23.5<br>22.7 to 23.2<br>21.1 to 22.7<br>20.6 to 21.1<br>15.2 to 20.6<br>DC |

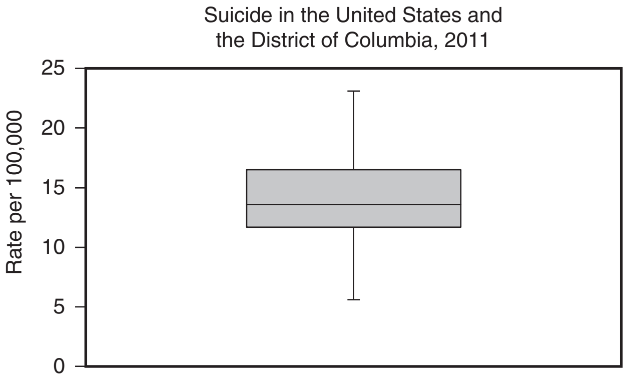
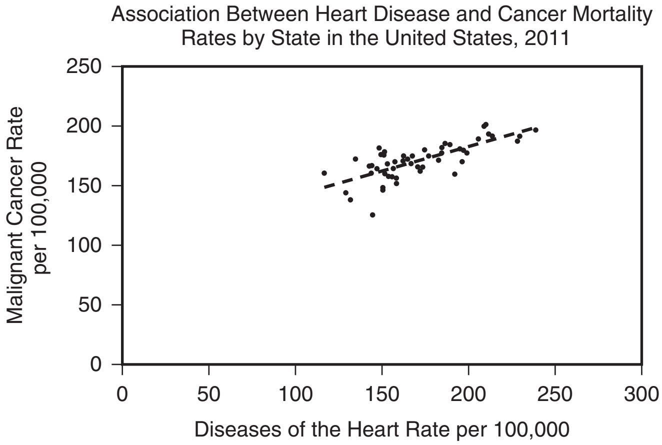
(*continues*)

## Table 2.2 Graphs for Describing Data (Continued)

| Type of Graph | Examples |
|---|---|
| Stem-and-leaf plot | Suicide rates according to states and the District of Columbia in the United States, 2011 |

| Stem | Leaf | No. |
|---|---|---|
| 23 | 1 | 1 |
| 22 | 5 | 1 |
| 21 | | |
| 20 | 4 | 1 |
| 19 | 44 | 2 |
| 18 | 245 | 3 |
| 17 | 489 | 3 |
| 16 | 567 | 3 |
| 15 | 12379 | 5 |
| 14 | 248 | 3 |
| 13 | 11266779 | 8 |
| 12 | 23445568 | 8 |
| 11 | 57 | 2 |
| 10 | 469 | 3 |
| 9 | 1389 | 4 |
| 8 | 13 | 2 |
| 7 | 4 | 1 |
| 6 | | |
| 5 | 6 | 1 |

## Table 2.2 Graphs for Describing Data (Continued)

| Type of Graph | Examples |
|---|---|
| Box plot |  Suicide in the United States and the District of Columbia, 2011 |
| Scatter plot |  Association Between Heart Disease and Cancer Mortality Rates by State in the United States, 2011 |

Note: Counts and rates for these graphs were calculated using data from the Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Mortality - All COD, Aggregated With State, Total U.S. (1969–2011) <Katrina/Rita Population Adjustment>, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released July 2014. Underlying mortality data provided by NCHS (www.cdc.gov/nchs).

extensively in epidemiology. For example, the epidemiologic measures of association between exposure status and disease status involve probability:

$$Odds\ Ratio = \frac{P(Exposed|Disease)\ /\ P(Unexposed\ |\ Disease)}{P(Exposed|No\ \ Disease)\ /\ P(Unexposed\ |\ No\ \ Disease)}$$

$$Risk\ \ Ratio = \frac{P(Disease|Exposed)}{P(Disease|Unexposed)}$$

Another example is Bayes' theorem, which is used to compute posterior probabilities from prior and observed probabilities. In epidemiology, the validity of screening and diagnostic tests can be assessed using this theorem. The study of probability can be extended to the concept of a random variable. The properties being observed or measured in a study are called variables. A *variable* is a characteristic that varies from one observation to the next and can be measured or categorized. A variable can take on a specified set of values. If the value of a variable is the result of a statistical experiment, it is a *random variable*. Random variables are represented using capital letters ($X$, $Y$, $Z$, etc.), and lowercase letters represent one of its values. A probability distribution is a table or an equation that represents each outcome of the random variable with its probability of occurrence.

If a variable can assume any value between two specified values, it is a continuous variable; if it cannot, it is a discrete variable. If a variable is continuous, its probability distribution is continuous. Likewise, if a variable is discrete, its probability distribution is discrete. There are several discrete and continuous probability distributions. Some of the more commonly used discrete and continuous probability distributions applied to epidemiologic data include the binomial and Poisson probability distributions and the normal and chi-square probability distributions, respectively (**Table 2.3**).

Table 2.3 makes reference to degrees of freedom (df). Degrees of freedom are central to the principle of estimating characteristics of populations based on sampled data. The degrees of freedom of an estimate are the number of independent pieces of information used to obtain the estimate. Suppose we know the mean body mass index (BMI) for a population of athletes to be 21.0. If a randomly sampled athlete has a mean BMI of 22.5,

**Table 2.3 Selected Probability Density Functions and Their Properties**

| Notation and Parameters | Probability Density Function $f(\mathbf{x})$ | Mean | Variance | Properties |
|---|---|---|---|---|
| Binomial<br>$X \sim BIN(n, p)$<br>$0 < \pi < 1$ | $P(X = x \mid n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$<br>$x = 0, 1, \ldots, n$ | $n\pi$ | $n\pi(1 - \pi)$ | 1. Fixed number of *n* trials of an experiment with two possible outcomes for each trial.<br>2. The number of successes is x, and the probability of success is $\pi$. The probability of failure is $1 - \pi$.<br>3. Trials are independent of one another. |
| Poisson<br>$X \sim POI(\mu)$<br>$0 < \mu$ | $P(X = x \mid \mu) = \frac{e^{-\mu} \mu^x}{x!}$<br>$x = 0, 1, \ldots$ | $\mu$ | $\mu$ | 1. Experimental outcomes classified as successes or failures. The actual number of successes is *x*.<br>2. We know the average number of successes ($\mu$) that occur in a given region.<br>3. The probability of a success is proportional to the size of the region (e.g., length, area, period of time, volume, etc.).<br>4. The probability of a success in a very small region is approximately zero. |

(*continues*)

**Table 2.3 Selected Probability Density Functions and Their Properties (Continued)**

| Notation and Parameters | Probability Density Function *f*(x) | Mean | Variance | Properties |
|---|---|---|---|---|
| Normal<br>$X \sim N(\mu, \sigma^2)$<br>$0 < \sigma^2$ | $\frac{1}{\sqrt{2\pi}\sigma} e^{-\left[\frac{x-\mu}{\sigma}\right]^2/2}$ | $\mu$ | $\sigma^2$ | 1. The total area under the normal curve is 1.<br>2. The probability that a normal random variable *X* equals any particular value is 0.<br>3. Each normal curve, regardless its mean and standard deviation, follows the empirical rule, which is that about 68% of the area under the curve falls within 1 standard deviation of the mean, about 95% of the area under the curve falls within 2 standard deviations of the mean, and about 99.7% of the area under the curve falls within 3 standard deviations of the mean. |
| Chi-Square<br>$X \sim X^2(v)$<br>$v = 1, 2, \ldots$ | $\frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2}$<br>$0 < x$ | $v$ | $2v$ | 1. The mean of the distribution is equal to the number of degrees of freedom: $\mu = v$.<br>2. The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2v$.<br>3. When the degrees of freedom are greater than or equal to 2, the maximum value for *X* occurs when $\chi^2 = v - 2$.<br>4. As the degrees of freedom increase, the chi-square curve approaches a normal distribution. |

then the variance is $(22.5 - 21.0)^2 = 2.25$. This estimate uses a single piece of information (df = 1) and estimates the population variance of athletes. If a second randomly sampled athlete has a mean BMI of 21.8, the variance estimate is 0.64. Averaging the two estimates gives 1.445, which is based on two independent pieces of information (df = 2). If the second athlete was chosen because he or she was a friend to the first, then these two estimates would not be independent.[1]

In reality, it is uncommon to know the population mean, so we estimate the mean using sampled data. This affects the degrees of freedom. In our example, the sample mean is $\frac{(22.5+21.8)}{2} = 22.15$. Now compute two estimates of variance as $(22.5 - 22.15)^2 = 0.1225$ and $(21.8 - 22.15)^2 = 0.1225$. However, these two estimates are not independent because both sampled BMIs contributed to the calculated mean. Therefore, we do not have two degrees of freedom. In general, the degrees of freedom for an estimate equals the number of observed values minus the number of parameters estimated in order to obtain the estimate in question. The formula for estimating the variance in a sample is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ . The denominator in this equation is the degrees of freedom.[1]

## Sampling

A *population* is a set or collection of items of interest in a study. In public health, where the focus is on human populations, a population refers to a collection of individuals who share one or more measurable personal or observational characteristics (e.g., a social group, an income level, a type of worker, geographic location) from which data may be collected and evaluated. A *sample* is a subset of items that have been selected from the population. The data set that represents the target of interest is called a population.

It is often not feasible to observe or measure an entire population, so we select a subset of values from the population. Random sampling is commonly employed when doing a questionnaire survey and is the best approach to achieve external validity. A *random sample* is a sample in which every element in the population has an equal chance of being selected. The method of random sampling is relatively easy to implement, but it becomes more difficult with larger populations, where there is a greater challenge in obtaining an accurate sampling frame. A *sampling frame* is the actual set of units from which a sample will be drawn. It is a list that contains every member of the population. Studies involving random selection often use statistical software packages to generate random numbers. For example, suppose we were interested in identifying the percentage of students at a given university who eat five or more servings of fruit and vegetables per day. If each student had a unique number assigned to them, say from 1 to 30,000 (assuming there are 30,000 students), we could select a random sample of 100 students by obtaining randomly generated numbers over the target population. Random numbers are easily generated in spreadsheets or statistical software. To illustrate, in an MS Excel spreadsheet, choose a cell and type "=RANDBETWEEN(1,30000)". This will generate a random number between 1 and 30,000. Copy this cell and paste it down the number of rows to match the number of random numbers desired. If we anticipate that 20% of the students contacted will not be interested in participating in the survey, then we will need $100 \times (1 \div [1 - 0.20]) = 100 \times 1.25 = 125$ unique randomly generated numbers. Representative sample data can then be analyzed and be used to make inferences about the population.

Descriptive statistics can now be applied to the data. If we are dealing with sample data, estimates of population characteristics are made with a corresponding measure of reliability for our estimate. *Parameters* are summary constants that measure characteristics of the population, such as the population mean and standard deviation. If the variable of interest is normally distributed, then these parameters nicely describe the population. Typically, we estimate parameters using information taken from sample data. Estimates of the population mean and standard deviation from sample data are examples of statistics. A *statistic* is a summary measure based on sample data. Our aim is to obtain unbiased estimates of the population parameters.

## Estimation in Statistics

*Estimation* is the process that we use to make inferences about a population, based on information that is obtained from a sample. Statistics are used to estimate parameters. We can estimate a parameter using a point estimate or an interval estimate. A point estimate of a parameter is a single value of a statistic (e.g., the sample mean is a point estimate of the population mean). Interval estimation is an interval of probable values of an unknown population parameter based on sample data.

An *estimator* is a random variable or statistic that is used to estimate an unknown parameter. An *estimate* is the actual numerical value obtained for an estimator. An estimator of a parameter is an unbiased estimator if its expected value equals the parameter. For example, if $x_1, \ldots, x_n$ denotes a random sample of size $n$ from $f(X)$ with $E(X) = \mu$ and $Var(X) = \sigma^2$, then

$$E(\overline{x}) = \mu,\ Var(\overline{x}) = \tfrac{\sigma^2}{n}, \text{and } E(s^2) = \sigma^2.$$

An interval estimate is two numbers between which a parameter is likely to be. A confidence interval is used to express the precision and uncertainty related to a given sampling method. Confidence intervals have a level of confidence threshold and a measurement error. Basically, the researcher is saying that the hypothesis will only be accepted if the scientist can have $(1 - \alpha) \times 100\%$ confidence that the results actually represent the truth. Typically, for exploratory studies, a popular level of confidence is 90%, while analytic studies generally require a 95% or 99% level of confidence.

The most common interpretation of a confidence interval is based on the relative frequency property of probability; that is, if a confidence interval is computed from several different samples, we would then expect over the long run that about $(1 - \alpha) \times 100\%$ of the intervals will include the true parameter. Thus, the confidence level represents the long-term frequency interpretation of probability. Usually, the confidence coefficient $\gamma$ is 0.95, and the confidence level is 95%. So if we are 95% confident, 100 different samples of the same population should include the true parameter 95 times. In practice, we only take one sample and one confidence interval, so we do not know whether the interval is one of the 95 or one of the 5. Hence, we are 95% confident.

If the sample is taken from a normal distribution, confidence limits (CL) have the general form

$$Estimator \pm t_{\alpha/2,df} Standard\ Error\ (ER)$$

If the sample size is 30 or greater, the $t$ distribution approximates the $z$ distribution, in which case

$$90\%\ \text{CL} = \mathit{Estimator} \pm 1.645 \times \mathit{SE}$$
$$95\%\ \text{CL} = \mathit{Estimator} \pm 1.960 \times \mathit{SE}$$
$$99\%\ \text{CL} = \mathit{Estimator} \pm 2.576 \times \mathit{SE}$$

If the sample size is less than 30 and the sample is not normally distributed, then a nonparametric approach is required (see Chapter 15).

The finite population correction factor is used to reduce the standard error by $\sqrt{\frac{N-n}{N-1}}$ if the sample size is large compared with the finite population. Otherwise, the finite population correction factor is close enough to 1 so that it can be ignored, which is generally the case. For example, suppose $N = 5000$ and $n = 100$; the finite population correction factor is 0.990. On the other hand, if $N = 500$ and $n = 100$, the finite population correction factor is 0.895. Use of the finite population correction factor when calculating confidence intervals will result in smaller confidence intervals as the sample approaches the finite population size.

## Hypothesis Testing

A *statistical hypothesis* is a belief about a population parameter. A *hypothesis* is a proposed explanation for a phenomenon in one or more populations. *Hypothesis testing* is a procedure based on sample information and probability that is used to test statements regarding a characteristic of one or more populations; it is a statement about the population parameter called the null hypothesis $H_o$. After formulating the null hypothesis, we then make a statement that contradicts $H_o$, called the alternative or research hypothesis, $H_a$. A set of six steps are used in hypothesis testing:

1. Formulate the null hypothesis in statistical terms. A parameter is used in expressing the null hypothesis.
2. Formulate the alternative hypothesis in statistical terms. A parameter is used in expressing the alternative hypothesis. Together, the null

and the alternative hypotheses cover all possible values of the population parameter in that one of the two statements is true.

3. Select the level of significance for the statistical test and the sample size. By convention, the level of significance is 0.05. However, if a more conservative test is desired, 0.01 may be used. On the other hand, in exploratory studies or if stepwise model selection procedures are used, then the level of significance may be 0.1 or higher. Note that stepwise model selection should not take the place of careful consideration of the underlying social or medical importance of selected variables.
4. Select the appropriate test statistic and identify the degrees of freedom and the critical value.
5. Collect the data and calculate the statistic.
6. Reject or fail to reject the null hypothesis. If we fail to reject, we are not saying the null hypothesis is true but that there is insufficient evidence from our sample to reject it. The alternative hypothesis may be true, but we simply do not have sufficient evidence to support it. This may occur if our sample size is too small or not representative of the truth.

## Decision Errors

Science is conducted with the knowledge that human measurement is imperfect, and the world of epidemiology is no exception. While science is the pursuit of truth, it can never actually prove truth without a shadow of a doubt because there is always the possibility of error. Hence, probability is employed in statistical inference to capture the chance of error.

*Statistical inference* is the process of drawing conclusions about the population based on a representative sample of the population. *Probability* is used to indicate the level of reliability in the conclusion. A test of the null hypothesis may result in two types of errors. *Type I error* refers to rejecting the null hypothesis given it is true. *Type II error* refers to failing to reject the null hypothesis given it is false. Because we do not know the actual values of the population from which we obtained our sample, we want to have studies that limit the chance of committing either type of error. Therefore, if our null hypothesis is in fact true, we will limit the probability of rejecting it to a value $\alpha$. This value is typically specified to be 0.01 or 0.05. If the null hypothesis is in fact false, we will limit the probability of accepting it to a value $\beta$, which is typically specified to be 0.1. The *power* of a test is $1- \beta$, or rejecting $H_0$ when $H_1$ is true.

The *p-value* is the probability that an effect as large or larger than that observed in a given study could have occurred by chance alone, given that there is truly no relationship between the exposure and the outcome. The *p*-value can also be thought of as a measure of chance. When analyses are based on sample data, it is possible to obtain a result that is due to the particular sample, which does not represent the overall population. The probability of a chance finding is decreased by increasing the sample size.

A confidence interval is similar to a *p*-value because it helps us to understand how confident we can be that our findings reflect the larger population. However, while a *p*-value is one number, a confidence interval is a range of values represented by the low and high on a range of possible values. Confidence intervals can be developed for ratios and proportions, and they are relatively easy to interpret.

Likelihood intervals are similar to confidence intervals in the sense that they show the potential range that a value could take along a reasonable distribution. However, likelihood intervals are more often used when the distribution of variables is abnormal. In one study, for example, likelihood intervals were used to look at incomplete paired binomial data because conventional confidence intervals were unusually wide.[2] In Bayesian statistics, a credible interval (or Bayesian confidence interval) is analogous to conventional confidence intervals.[3] However, the interested reader can refer elsewhere for how they are distinct.[4-6]

## Applications of Hypothesis Testing

Hypothesis testing can be applied to a number of types of statistical problems (**Table 2.4**).

### Table 2.4 Formulating and Testing Hypotheses for Selected Conditions

| | Conditions | Test Method |
|---|---|---|
| Proportions | Simple random sampling<br>Each sample point has two possible outcomes<br>The sample includes at least 10 successes and 10 failures<br>The population size is at least 10 times as big as the sample size | One-sample Z-test |
| Proportions from small samples | The sample does not include at least 10 successes and 10 failures | Binomial experiment<br>Hypergeometric experiment |
| Difference between proportions | Simple random sampling<br>Independent samples<br>The sample includes at least 10 successes and 10 failures<br>The population size is at least 10 times as big as the sample size | Two-proportion Z-test |
| Means | Simple random sampling<br>Sample drawn from a normal or near-normal population<br>Sampling distribution approximately normal if the population distribution is normal, and up to 15 (sampling distribution symmetric, unimodal, without outliers), 16–40 (sampling distribution is moderately skewed, unimodal, without outliers), or at least 40 (sampling distribution has outliers) | One-sample t-test |

(*continues*)

Table 2.4 Formulating and Testing Hypotheses for Selected Conditions (Continued)

| | Conditions | Test Method |
|---|---|---|
| Differences between means | Same as for means, but the samples are independent and each population is at least 10 times larger than its respective sample | Two-sample t-test |
| Differences between matched pairs | Same as difference between means, except the test is conducted on paired data (not independent) | Matched-pairs t-test |
| Goodness-of-fit | Used to determine if sample data are consistent with a hypothesized distribution<br>Simple random sampling<br>Categorical variable(s)<br>Sample size: 10 or more subjects in each group; 80% of the predicted counts are at least 5; other expected counts are greater than 2, with no 0 counts | Chi-square for goodness of fit |
| Homogeneity | Used to determine if frequency counts are distributed similarly across different populations<br>Simple random sampling<br>The population is at least 10 times as large as the sample<br>The variables being studied are categorical<br>The sample observations are displayed in a contingency table, and the expected count in each cell of the table is 5 or greater | Chi-square test for homogeneity |

| | | |
|---|---|---|
| Independence | Used to determine if there is a significant association between the variables<br>Simple random sampling<br>The population is at least 10 times as large as the sample<br>The variables being studied are categorical<br>The sample observations are displayed in a contingency table, and the expected count in each cell of the table is 5 or greater | Chi-square test for independence |

Note: Matched pairs refers to two measurements taken for each respondent (e.g., before-treatment measurement and after-treatment measurement), goodness of fit of a statistical model indicates how well it fits a set of observations, homogeneous refers to whether multiple populations are similar or equal in some characteristic, and independence refers to no association.

If $X$ is a random variable with mean $\mu$ and variance $\sigma^2$, then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

has mean 0 and variance 1. The normal approximation to the binomial can be used when $n$ is large. A rule of thumb is that the variance for the binomial must be 5 or greater (i.e., $n\pi(1 - \pi) > 5$). The standardized binomial variable is:

$$Z = \frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}}$$

If we are interested in the number of successes $X$ in $n$ trials:

$$Z = \frac{X / n - \pi}{\sqrt{\pi(1 - \pi) / n}}$$

To standardize a sample mean ($\overline{x}$):

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

The central limit theorem says that if $n$ is large, then the $Z$ distribution is a standard normal distribution; that is, $\overline{x}$ has a normal distribution with mean $\pi$ and a standard error of $\sigma / \sqrt{n}$; the standard error of the sample mean is referred to as the standard error (SE).

If the variance is unknown, then the $t$ statistic is appropriate in hypothesis testing. The equation for the $t$ statistic is:

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

The chi-square test is appropriate for assessing contingency table data. There are many forms of the chi-square test, depending on the study design and measure of association. These will be presented in later chapters. The Pearson chi-square statistic is:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

where $m_{ij} = \frac{R_i C_j}{n}$

This $\chi^2$ has (Row[R]–1)(Column[C]–1) degrees of freedom. That is, the chi-square is the sum of the observed frequencies minus the expected frequencies, squared, divided by the sum of the expected frequencies. If the sample size requirement is not satisfied, then some rows and/or columns may be combined.

## STATISTICAL TECHNIQUES

Several statistical techniques have been used in epidemiology to investigate a range of public health problems. For example, epidemiology has drawn heavily on statistical methods for analyzing proportions, rates, and time to failure. Regression methods have been used extensively for assessing proportions, rates, proportional hazards, and matched studies. Power and sample size estimation techniques are basic to many epidemiologic studies.

The number of statistical techniques currently available to epidemiologists is extensive, some requiring a fairly sophisticated understanding of statistics. Some of these techniques can be applied using a spreadsheet, but others require the use of computer software. Statistical Analysis System (SAS) procedure code will be presented in several of the chapters, with illustrations of an array of public health problems. Some SAS basics are presented in Appendix C.

It is important to recognize that statistical techniques are tools for addressing questions of scientific interest. Hence, the process begins with the research question. The question should directly correspond with the public health outcome of interest. The outcomes in epidemiologic research have historically involved disease, yet the increasing application of epidemiology means that the outcomes measure health-related states or events in general. The study design should be reflected in the methods of analysis. It is also important that the distribution assumptions about random mechanisms that generate a set of data be realistic. Violations of these assumptions will yield invalid results.[7]

## SUMMARY

1. This chapter covered four areas of statistics: descriptive statistics, which involves methods of organizing, summarizing, and describing numerical data; probability, which is used when discussing the chance or likelihood that a given event will occur; inferential statistics, which involves drawing a conclusion about a population's attributes from information in the sample; and statistical techniques, which are analytic approaches that draw on statistical methods to investigate a range of problems.

2. Statistics is the science of data, where data are pieces of information, such as observations or measurements of phenomena of interest. Quantitative data are observations or measurements measured on a numerical scale, whereas qualitative data cannot be described numerically.

3. Scales of measurement are the ways that variables are defined and categorized, such as nominal scale, ordinal scale, interval scale, and ratio scale.

4. Data are described using frequency distribution tables, summary statistics, and graphs. Tables and graphs are used to identify patterns, trends, aberrations, similarities, and differences in the data and for communicating public health information according to person, place, and time factors.

5. Probability theory provides a basis for assessing the reliability of the conclusions we make under conditions of uncertainty.

**6.** In public health, a population refers to a collection of individuals who share one or more measurable personal or observational characteristics. A sample is a subset of items that have been selected from the population.

**7.** Random sampling is the best approach for achieving external validity. A sampling frame is the actual set of units from which a sample will be drawn. Random numbers are generated and applied to the sampling frame to obtain a random sample.

**8.** Estimation in statistics is the process of making inferences about a population, based on information that is obtained from a sample. Statistics are used to estimate parameters.

**9.** Hypothesis testing is a procedure, based on sample information and probability, that is used to test statements regarding a characteristic of one or more populations.

**10.** Two types of decision errors in hypothesis testing are Type I error (rejecting the null hypothesis given it is true) and Type II error (failing to reject the null hypothesis given it is false). The probability of committing a Type I error is denoted as $\alpha$, and that of committing a Type II error is denoted as $\beta$. The power of a test is $1-\beta$ (rejecting the null hypothesis when the alternative hypothesis is true).

**11.** The *p*-value is the probability that an effect as large or larger than that observed in a given study could have occurred by chance alone, given that there is truly no relationship between the exposure and the outcome.

**12.** An estimator is a random variable or statistic that is used to estimate an unknown parameter; it is the actual numerical value obtained for an estimator. An estimator is unbiased if its expected value equals the parameter.

13. Statistical techniques are tools for addressing research questions of scientific interest. The question corresponds to a health problem and influences the selected study design, methods, and assumptions.

## EXERCISES

1. There are four general areas of statistics. List and describe each of these areas.

2. Match (A) nominal data, (B) ordinal data, or (C) interval/ratio with the following:

______ integers of counts that differ by fixed amounts, with no intermediate values possible

______ measurable quantities not restricted to taking on integer values

______ data that are ordered into categories or classes

______ data that are not ordered into categories or classes

3. What is a frequency distribution?

4. What is a random variable?

5. How is a probability distribution related to a random variable?

6. How does a parameter compare with a statistic?

7. What are three properties of the binomial probability distribution?

8. What are four characteristics of the Poisson probability distribution?

**9.** Under which two conditions will the binomial probability be approximately equal to the Poisson probability?

**10.** Describe the properties of the normal distribution.

**11.** What is the practical value of the standard normal distribution?

For questions 12–15, consider a study involving cancer-related claims reflecting type of service rendered; let *X* be a discrete random variable that represents the number of three specific combinations of services. The probability distribution for *X* appears as follows.[8]

**Cancer-Related Claims According to Type of Service Rendered***

| Physician Services | Nonphysician Professional Health Services | Hospital Services | No. | $P(X = x)$ |
|---|---|---|---|---|
| Yes | Yes | Yes | 698 | 0.310 |
| Yes | Yes | No | 151 | 0.067 |
| Yes | No | Yes | 788 | 0.349 |
| Yes | No | No | 526 | 0.233 |
| No | Yes | Yes | 10 | 0.004 |
| No | Yes | No | 13 | 0.006 |
| No | No | Yes | 69 | 0.031 |
| No | No | No | 0 | 0.000 |

Data source: DMBA enrollees during 1998–2006, aged 15–64.

*First ICD-9-CM cancer code assigned 140–208, excluding 172.0–173.9 (skin cancer).

**12.** Construct a graph of the probability distribution.

**13.** What is the probability that a cancer patient received all three types of services?

**14.** What is the probability that the cancer patient received both physician and hospital services only?

**15.** What is the probability of receiving nonphysician professional health services?

**16.** In the United States in 2012, the prevalence of asthma in adults was 8%.[9] If you selected repeated samples of size 10 from the U.S. population, what would the mean number of individuals per sample be of asthma?

**17.** Referring to question 16, what are the standard deviations that correspond to these expected values?

**18.** Referring to question 16, what is the probability of finding exactly two people who have asthma?

**19.** Referring to question 16, what is the probability of finding more than two people with asthma?

**20.** In 2011, the rate of female breast cancer among whites in New Mexico, ages 50–59, was 219.7 per 100,000.[10] In a sample of 100 women in this age range, what is the expected number of cases?

**21.** Referring to question 20, what is the probability that no women in this sample will have breast cancer?

**22.** Referring to question 20, what is the probability that exactly one woman will have breast cancer?

**23.** Referring to question 20, what is the probability that two or more women will have breast cancer?

**24.** Define the empirical rule and state when it may be used.

**25.** What are the properties of the central limit theorem?

**26.** Let $X$ be a random variable that represents the Beck Depression Inventory score. For a population of people aged 24–81, the mean Beck Depression Inventory score is approximately normally distributed with a mean of 4.35 and a standard deviation of 4.64.[11] What change in the curve results when this is transformed to the standard normal curve?

**27.** Referring to question 26, what is the area of the population curve above 13.63? Below –0.29?

**28.** Referring to question 26, what is the area of the population curve between –0.29 and 8.99?

**29.** Referring to question 26, what value in the population corresponds with $Z = -0.5$?

**30.** In 2012, within a sample of 1000 individuals aged 65 years and older in the United States, 60.1% had a flu shot within the past year. We may consider the selection of a person having had a flu shot within the past year as a "success." Find the probability that the number of $X$ successes will lie above 63%.

**31.** In the United States in 2012, the percentage of adults who were obese in each of the 50 states and 4 territories was normally distributed around a mean of $\mu = 28.1$. The standard deviation around the mean was $\sigma = 3.2$.[12] We took a random sample of $n = 10$ and obtained $\overline{X}$ for the sample of 29. Calculate the probability of getting a sample mean greater than this if the true mean is $\mu = 28.1$.

**32.** For the previous problem, assume that in addition to a sample mean of 29, the standard deviation is 3. Calculate a 95% confidence interval for the mean assuming a level of significance of 0.05.

**33.** In a breast cancer cohort study, a woman is considered to be exposed if she first gave birth at age 30 or older. In a sample of 4540 women who gave birth to their first child before age 30, 65 developed breast cancer. Of the 1628 women who first gave birth at 30 or older, 31 were diagnosed with breast cancer. Does having a first birth at age 30 or older increase the risk of breast cancer? Apply the six steps of hypothesis testing to this problem, assuming a level of significance of 0.05.

**34.** What is the probability of committing a Type 1 error in the previous problem?

**35.** What type of graph is appropriate for showing the minimum, maximum, first, second, and third quartiles of a distribution of discrete or continuous data?

**36.** Why might an area map be useful?

## REFERENCES

1. Lane DM. Degrees of freedom. In: *Online Statistics Education: An Interactive Multimedia Course of Study*. Developed by Rice University (Lead Developer), University of Houston Clear Lake, and Tufts University. http://onlinestatbook.com/2/estimation/df.html/.
2. Pradhan V, Menon S, Das U. Corrected profile likelihood confidence interval for binomial paired incomplete data. *Pharm Statist*. 2013;12(1):48–58. doi: 10.1002/pst.1551.
3. Lee PM. *Bayesian Statistics: An Introduction*, 2nd edition, London: Arnold; 1997.

4. O'Hagan A. *Kendall's Advanced Theory of Statistics, Vol 2B, Bayesian Inference*, Section 2.51. London: Arnold; 1994.
5. Antelman G. *Elementary Bayesian Statistics.* Madansky A, & McCulloch R, eds. Cheltenham, UK: Edward Elgar; 1997.
6. Pythonic Perambulations. Frequentism and Bayesianism III: confidence, credibility, and why frequentism and science do not mix. http://jakevdp.github.io/blog/2014/06/12/frequentism-and-bayesianism-3-confidence-credibility/. Accessed December 15, 2014.
7. Holford TR. *Multivariate Methods in Epidemiology*. New York, NY: Oxford; 2002.
8. Merrill RM, Baker RK, Lyon JL, Gren LH. Healthcare claims for identifying the level of diagnostic investigation and treatment of cancer. *Med Sci Monit*. 2009;15(5):PH25–31.
9. Centers for Disease Control and Prevention [CDC]. Prevalence and trend data. Nationwide (States, DC, and Territories). Current asthma prevalence, 2012. http://www.cdc.gov/asthma/asthmadata.htm. Accessed June 10, 2014.
10. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER* Stat Database: Incidence - SEER 13 Regs Research Data, Nov 2013 Sub (1992-2011) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2012 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the November 2013 submission.
11. Merrill RM, Aldana SG, Greenlaw RL, Diehl HA. The coronary health improvement project's impact on lowering eating, sleep, stress, and depressive disorders. *Am J Health Educ*. 2008;39(6):337–344.
12. Behavioral Risk Factor Surveillance System. Prevalence and trends data. Immunization – 2012. Available at: http://apps.nccd.cdc.gov/brfss/list.asp?cat=IM&yr=2012&qkey=8341&state=All. Accessed June 12, 2014.