


Statistics and Data Presentation

Valerie J. M. Watzlaf and Elaine Rubinstein

Student Study Guide activities for this chapter are available on the Evolve Learning Resources site for this textbook. Please visit <http://evolve.elsevier.com/Abdelhak>.

When you see the Evolve logo , go to the Evolve site and complete the corresponding activity, referenced by the page number in the text where the logo appears.

Overview of Statistics and Data Presentation

Role of the Health Information Management

Professional

Health Care Statistics

Vital Statistics

Rates, Ratios, Proportions, and Percentages

Mortality Rates

Gross Death Rate

Net Death Rate

Anesthesia Death Rate

Postoperative Death Rate

Maternal Death Rate

Neonatal, Infant, and Fetal Death Rates

Using and Examining Mortality Rates

Autopsy Rates

Morbidity Rates

Census Statistics

Key Words

Alternative hypothesis
Anesthesia death rate
Autopsy rate
Average daily inpatient census
Average length of stay
Bar graph
Bed turnover rate
Census statistics
Coefficient of variation
Community-acquired infection
Comorbidity
Confidence interval
Contingency table
Continuous data
Direct method of age adjustment
Discrete data
Dispersion
Fetal death rate
Frequency distribution
Frequency polygon
Gross death rate
Histogram
Hypothesis

Incidence
Incidence rate
Infant death rate
Infer
Inpatient bed occupancy rate
Interval data
Length of stay
Level of significance
Maternal death rate
Mean
Median
Mode
Morbidity rates
Mortality rates
Neonatal death rate
Net death rate
Nominal data
Nosocomial infection
Null hypothesis
Ordinal data
p value
Pearson correlation coefficient
Percentage

Percentage of occupancy
Pie chart
Postoperative death rate
Postoperative infection rate
Prevalence
Prevalence rate
Proportion
Random sample
Range
Rate
Ratio
Ratio data
Regression analysis
Sampling error
Standard deviation
Standardized mortality ratio
Stratified random sample
Systematic sampling
Test statistic
Tests of significance
Variance
Vital statistics
Weighted mean

Organizing and Displaying the Data

Types of Data

Nominal Data
Ordinal or Ranked Data
Interval Data
Ratio Data
Discrete Data
Continuous Data

Types of Data Display

Frequency Distribution
Bar Graph
Pie Chart
Histogram
Frequency Polygon

Statistical Measures and Tests

Descriptive Statistics

Measures of Central Tendency
Measures of Dispersion

Inferential Statistics

Tests of Significance
Interval Estimation
Sampling and Sample Size

Summary

Abbreviations

AIDS—Acquired Immunodeficiency Syndrome
ALOS—Average Length of Stay
ANOVA—Analysis of Variance
CV—Coefficient of Variation

dfb—Degrees of Freedom, Between Groups

dfw—Degrees of Freedom, Within Groups

DRG—Diagnosis-Related Group

HIM—Health Information Management

ICD—*International Classification of Diseases*

MSB—Mean Square Between Groups

MSW—Mean Square Within Groups

NCHS—National Center for Health Statistics

SMR—Standardized Mortality Ratio

SSB—Sum of Squares Between Groups

SSW—Sum of Squares Within Groups

Objectives

- Compute health care statistics, including mortality and morbidity rates, autopsy rates, measures of central tendency, and dispersion, and determine the most appropriate use of these health care statistics in health information management.
- Organize data generated from health care statistics into appropriate categories, including nominal, ordinal, discrete, and continuous.
- Display data generated from health care statistics using the most appropriate tables, graphs, and figures, including frequency tables, bar graphs, histograms, Pareto diagrams, pie charts, and frequency polygons.
- Determine which tests of significance should be used to test specific hypotheses and which are most appropriate for certain types of data.

OVERVIEW OF STATISTICS AND DATA PRESENTATION

Health care organizations continuously generate health care data. These data are used internally by patients, medical staff, nursing staff, and physical, occupational, and speech therapists and externally by state and federal regulatory agencies, the Joint Commission, and insurance companies, to name just a few. No matter who the user may be, statistics and data presentation focus on answering the user's questions while complying with the standards of the health care facility. To accomplish this goal, various methods are used to calculate specific types of statistics. Different rates, ratios, proportions, and percentages are used to evaluate mortality, autopsy, and morbidity rates and census and vital statistics.

Organizing and displaying health care data are necessary. To choose appropriate methods of displaying and

analyzing data, the health information management (HIM) professional must identify the level of measurement (nominal, ordinal, interval, or ratio) for variables and determine whether data are continuous or discrete. Measures of central tendency (mean, median, mode) and dispersion (variance and standard deviation) and tests of significance are used to describe and analyze data. It is also important for the HIM professional to understand basic principles of sample size determination and to be familiar with commonly used statistical tests such as analysis of variance (ANOVA), correlation, and regression.

This chapter explains basic and advanced health care statistics that are used in the health care field. Each statistic is defined and the formula for calculating each statistic is provided along with examples of how each statistic is used. Various methods of displaying data are described and illustrated.

Role of the Health Information Management Professional

Now more than ever, health care data are being collected to serve many purposes. One primary purpose is to establish health care statistics to compare trends in incidence of disease, quality and outcomes of care, and management of health information departments; another primary purpose is to conduct epidemiological research. The HIM professional's goal is to collect, organize, display, and interpret health care data properly to meet the needs of the users. Data can be manipulated in many ways to demonstrate one result or another. HIM professionals need a broad base of knowledge to determine which data elements should be used and when data are being analyzed appropriately or inappropriately. To do this, an understanding of health care and vital and public health statistics is necessary. Furthermore, knowledge of statistical analysis is necessary so that HIM professionals can be the forerunners in data analysis. Because HIM professionals oversee a vast array of health data, it is imperative that the interpretation of the analysis and results of health care data start with them.

The HIM professional should assume the lead in recommending and using statistical tests that promote improvement in the analysis, use, and dissemination of health care data. The HIM professional fills many diversified roles and responsibilities, such as clinical vocabulary manager, data miner, or clinical trials manager.

In each of these roles, understanding and applying the methods used to collect, analyze, display, interpret, and disseminate data are essential. Responsibilities undertaken in these roles may vary from person to person. For example, the clinical trials manager may play a clearly visible role in cancer research study analysis and interpretation of the data, the clinical vocabulary manager may play a key role in developing vocabularies and standards that can be effectively used in the design of the electronic health record, and the data miner may determine the appropriate databases to use when analyzing clinical and financial data.

The HIM professional may assume other managerial roles in which statistics are used to assess productivity in coding, transcription, correspondence, and record analysis. The HIM professional should have sufficient knowledge and skills to do the following:

- Collect quality health data
- Organize the data into databases
- Statistically analyze the data collected
- Develop, generate, and interpret health care statistical reports

HEALTH CARE STATISTICS

Vital Statistics

Vital statistics include data collected for vital events in our lives, such as births and adoptions, marriages and divorces, and deaths, including fetal deaths. Birth, death, and fetal

death certificates are familiar reports to HIM professionals. Although each state can determine the format and content of its certificates, the National Center for Health Statistics (NCHS) recommends standard forms that most states have adopted. The purpose of the NCHS standard forms is to have a national uniform reporting system of vital statistics. These standard forms are revised periodically. The attending physician is responsible for the completion of birth, death, and fetal death certificates. The accurate completion of these certificates is supervised by the HIM department, and a copy of the birth or death certificate is kept in the medical record. A copy of the fetal death certificate is kept in the mother's medical record.

When the certificate is complete, the original is sent to the local registrar, who keeps a copy and forwards the original to the state registrar. At each of these stages, the certificate is checked by the registrar to make sure it is complete. Individuals can obtain from the state registrar certified copies of birth, death, and fetal death certificates. Each state sends electronic files of birth and death statistics to the NCHS. The death statistics are then compiled in the National Death Index. The Death Index is a central computerized index of death record information used for research purposes by epidemiologists and other workers involved in health care research.¹ The natality, or birth, statistics are compiled in the monthly vital statistics reports, and the data files are also available for purposes of research.

Refer to your state health data center or division of vital statistics to receive state-specific information on preparing and registering vital records.

Rates, Ratios, Proportions, and Percentages

A **rate** is defined as the number of individuals with a specific characteristic divided by the total number of individuals or, alternatively, as the number of times an event did occur compared with the number of times it could have occurred.

A rate contains two major elements: a numerator and a denominator. The numerator is the number of times an event did occur. The number of events under study, or the numerator alone, conveys little information. However, when the numerator is compared with the denominator or the population of people in which the event could have occurred, a rate is determined. The results of a quality improvement study showed that 20 patients with diabetes had a stroke while taking a certain medication. What does this tell you? Should this medication be discontinued in this population? The data provided here include only the numerator. To compute a rate, the denominator is needed—for this example, total number of patients with diabetes who are taking the medication. This particular example included a sample size of 1000 patients. The rate is 20 in 1000 or 2 in 100. A rate is normally expressed in the following manner: 20 in 1000, 2 in 100, 1 in 100,000, 10 in 1,000,000, and so on.

However, rates are also commonly expressed as percentages by converting the rate into a decimal and then multiplying the decimal by 100. A **percentage** is based on a whole divided into 100 parts. In the preceding example, the rate could also be expressed as a percentage by taking $20/1000 = 0.02 \times 100 = 2\%$ or by taking $2/100 = 0.02 \times 100 = 2\%$. This tells us that 2% of the patients with diabetes (in the study) had a stroke while taking a certain medication. To express a fraction, such as $\frac{1}{5}$, as a percentage, the first step is converting the fraction into a decimal by dividing the numerator, 1, by the denominator, 5, to obtain 0.20. The decimal is then converted into a percentage by multiplying the decimal by 100, which can be accomplished by moving the decimal point two places to the right. The result of this process is 20%.

A **proportion** and a **ratio** are similar to a rate. A proportion, which is a part considered in relation to the whole, is normally expressed as a fraction— $\frac{20}{1000}$, $\frac{2}{100}$, $\frac{1}{100,000}$, $\frac{10}{1,000,000}$, and so on. A ratio is a comparison of one thing to another, such as births to deaths or marriages to divorces. A ratio is expressed as 20:1000, 2:100, 1:100,000, 10:1,000,000, and so on. The number of physicians relative to patients or teachers relative to students is normally expressed as a ratio. For example, if a physician group practice has 10 physicians and 1000 patients, the ratio is 10:1000, which reduces to 1:100.

Table 10-1 summarizes examples of rates, proportions, ratios, and percentages.^{2,3}

Once percentages are calculated, they can be compared across different subgroups, as seen in Table 10-2. This table concisely shows differences among geographic areas in the percentage of elderly people by age categories. It even allows a glance toward the future by projecting percentages for the years 2010 and 2025. Comparing percentages among areas shows that Europe has the highest percentage of population aged 65 years or older (13.7% in 1990) and that it should remain the world leader for at least the next 3 decades. North America and Oceania also have relatively high percentages of elderly people, which are projected to increase substantially from 1990 to 2025.⁴

Mortality Rates

Mortality rates are computed because they demonstrate an outcome that may be related to the quality of the health care provided. There are many types of mortality rates. Table 10-3 provides definitions and formulas for the most commonly used mortality rates.^{2,3}

Gross Death Rate

The **gross death rate** is a crude death rate for hospital inpatients because it does not consider such factors as age, gender, race, and severity of illness, which also play an important part in death rates. The use of the gross death rate as a measure of quality in health care has been questioned because it does not take these factors into account. As long as the HIM professional is aware that other factors influence this rate and that they have not been taken into account in the calculation, the gross death rate can be a quick, useful means of analyzing mortality in hospital inpatients (see the following example of gross death rate).

Example of Gross Death Rate

The discharge analysis report of Anywhere Health Care Facility shows 752 discharges (including deaths) for October 200X. Twelve deaths were also shown in the report.

Gross death rate:

$$= \frac{12 \text{ inpatient deaths}}{752 \text{ discharges (including deaths)}} \times 100$$

$$= 1.60\%$$

This means that 1.60% of total discharges from Anywhere Health Care Facility during October 200X ended in death or that the gross percentage of deaths, or the hospital death rate, for October was 1.60%.

Net Death Rate

The **net death rate** is different from the gross death rate because it does not include deaths that occurred less than 48 hours after admission to the health care facility. The net death rate is useful because it provides a more realistic account of patient deaths related to patient care provided by a specific health care facility. For example, a 90-year-old patient arrives at the emergency department with shortness of breath, chest pain, and arrhythmia. After being evaluated, the patient is admitted, and it is determined that he has had a severe myocardial infarction (see the example of net death rate).

Approximately 24 hours later, the patient has cardiac arrest and dies. This particular death would be included in the gross death rate but not the net death rate because it occurred less than 48 hours after admission. Reporting agencies sometimes request net death rates because they may provide a more realistic reflection of patient care provided than gross death rates.

Table 10-1 EXAMPLES OF RATIOS, PROPORTIONS, PERCENTAGES, AND RATES

Ratio	Proportion	Percentage	Rate (per 100,000)
1:100	$1/100 = 0.01$	1.0	1000 in 100,000
3:10,000	$3/10,000 = 0.0003$	0.03	30 in 100,000
250:100,000	$250/100,000 = 0.0025$	0.25	250 in 100,000

Table 10-2 PERCENTAGE OF ELDERLY BY AGE (YEARS): 1990–2025

Region	Year	Age 65 and Over	Age 75 and Over	Age 80 and Over
Europe*	1990	13.7	6.1	3.2
	2010	17.5	8.4	4.9
	2025	22.4	10.8	6.4
North America	1990	12.6	5.3	2.8
	2010	14.0	6.5	4.0
	2025	20.1	8.5	4.6
Oceania	1990	9.3	3.6	1.8
	2010	11.0	4.8	2.8
	2025	15.0	6.6	3.6
Asia	1990	4.8	1.5	0.6
	2010	6.8	2.5	1.2
	2025	10.0	3.6	1.8
Latin America, Caribbean	1990	4.6	1.6	0.8
	2010	6.4	2.6	1.2
	2025	9.4	3.6	1.8
Near East, North Africa	1990	3.8	1.2	0.5
	2010	4.6	1.6	0.8
	2025	6.4	2.2	1.1
Sub-Saharan Africa	1990	2.7	0.7	0.3
	2010	2.9	0.8	0.3
	2025	3.4	1.0	0.4

Source: U.S. Bureau of the Census: Center for International Research, International Data Base on Aging.

*Data exclude the former Soviet Union.

do. However, net death rates still do not take into consideration other risk factors that may also affect death, such as age, gender, race, and so forth. Therefore, an important note is that health care facilities are not necessarily responsible for deaths that occur more than 48 hours after patients are admitted; on the other hand, health care facilities are not necessarily free of responsibility for deaths that occur within 48 hours of admission. For this reason, some health care facilities do not make use of or report the net death rate.

Another consideration when computing any mortality rate is a health care facility must decide whether newborn inpatients will be included in these calculations. This decision is up to the health care facility; however, if a facility

decides that newborn inpatient deaths will be included in the numerator, all newborn discharges must also be included in the denominator.

Example of Net Death Rate

Inpatient deaths at Anywhere Health Care Facility for 200X totaled 50. Inpatient deaths that occurred less than 48 hours after admission to the facility totaled 15. Total discharges (including deaths) were 15,546.

Net death rate:

$$\begin{aligned}
 & \frac{50 \text{ inpatient deaths} - 15 \text{ inpatient deaths} < 48 \text{ hours}}{15,546 \text{ discharges (including deaths)}} \times 100 \\
 & = \frac{35}{15,546} \times 100 \\
 & = 0.23\%
 \end{aligned}$$

This means that 0.23% (fewer than 1%) of the deaths of discharges for 200X occurred more than 48 hours after admission to the health care facility, or that the net percentage of deaths or net death rate for 200X was 0.23%.

Anesthesia Death Rate

The **anesthesia death rate** can also be referred to as a cause-specific death rate because the death is determined by a physician or medical examiner to be due to a specific cause (e.g., an anesthetic agent). This rate indicates the number of deaths that are due to the administration of anesthetics for a specified period of time. If the recent anesthetic death rate is higher than the rate in previous periods, a focused evaluation may be necessary to determine why this is so (see the example of anesthesia death rate).

Example of Anesthesia Death Rate

Anywhere Health Care Facility performed 492 surgical procedures during November and administered 452 anesthetics. Deaths resulting from the administration of an anesthetic totaled two for the month.

Anesthesia death rate:

$$\begin{aligned}
 & = \frac{2 \text{ anesthetic deaths}}{452 \text{ anesthetics administered}} \times 100 \\
 & = 0.44\%
 \end{aligned}$$

This means that 0.44% (fewer than 1%) of anesthetics administered resulted in a patient's death, or that the anesthesia death rate for November was 0.44%.

Postoperative Death Rate

The **postoperative death rate** may be considered a cause-specific death rate as well. This death rate indicates the number of patients who die within 10 days of surgery divided

Table 10-3 MORTALITY RATES

Rate	Formula
Gross death rate (hospital death rate)	$\frac{\text{Total number of inpatient deaths}}{\text{Total number of discharges (including deaths)}} \times 100$
Net death rate	$\frac{\text{Total number of inpatient deaths} - \text{Inpatient deaths} < 48 \text{ hours}}{\text{Total discharges (including deaths)} - \text{Deaths} < 48 \text{ hours}} \times 100$
Anesthesia death rate	$\frac{\text{Total number of anesthetic deaths}}{\text{Total number of anesthetics administered}} \times 100$
Postoperative death rate	$\frac{\text{Total number of deaths (within 10 days of surgery)}}{\text{Total number of patients who received surgery}} \times 100$
Maternal mortality rate	$\frac{\text{Total number of direct maternal deaths}}{\text{Total number of obstetrical discharges (including deaths)}} \times 100$
Neonatal mortality rate	$\frac{\text{Total number of neonatal deaths}}{\text{Total number of neonatal discharges (including deaths)}} \times 100$
Infant mortality rate	$\frac{\text{Total number of infant deaths}}{\text{Total number of infants discharged (including deaths)}} \times 100$
Fetal Death Rates	
Early fetal death (abortion) rate	$\frac{\text{Total number of early fetal deaths}}{\text{Total number of births (including early fetal deaths)}} \times 100$
Intermediate fetal deaths	$\frac{\text{Total number of intermediate fetal deaths}}{\text{Total number of births (including intermediate fetal deaths)}} \times 100$
Late fetal (stillborn) deaths	$\frac{\text{Total number of late fetal deaths}}{\text{Total number of births (including late fetal deaths)}} \times 100$

Note: The numerator and denominator in each formula must be for the same time period.

by the number of patients who underwent surgery for the period; therefore, it expresses the number of deaths that may have resulted from surgical complications. In both the anesthesia and the postoperative death rates, other risk factors, such as age, gender, race, and severity of illness, are not considered. Therefore, if it is found that these rates are higher in certain periods than in others, specific evaluations are necessary to determine whether the increase is truly due to the anesthesia or surgery or to other risk factors (see the example of postoperative death rate).

Example of Postoperative Death Rate

Surgery was performed on 492 patients in the Anywhere Health Care Facility during November, and 27 of those patients died within 10 days of surgery.

Postoperative death rate:

$$\frac{27 \text{ postoperative deaths}}{492 \text{ patients having surgical procedures}} \times 100$$

= 5.49% or 5.5%

This means that 5.5% of the patients who underwent surgery died within 10 days of the procedure or that the postoperative death rate for November was 5.5%.

Maternal Death Rate

Death rates are further categorized according to the type of service or department, such as the maternal mortality or death rate. A maternal death results from causes associated with pregnancy or its management but not from accidental or incidental causes unrelated to the pregnancy. The **maternal death rate** is the number of maternal deaths divided by the number of obstetric discharges. Again, like all the rates described previously, the maternal death rate does not take into account any other risk factors. The maternal death rate is useful because maternal deaths are rare. Therefore, if there is even one maternal death in a period, it is necessary to examine the cause of death in more detail (see the example of maternal death rate).

Example of Maternal Death Rate

Anywhere Health Care Facility had a total of 752 discharges for October, including 120 obstetrical discharges (including deaths) and 1 maternal death.

Maternal death rate:

$$= \frac{1 \text{ maternal death}}{120 \text{ obstetrical discharges (including deaths)}} \times 100$$

$$= 0.83\% \text{ or } 0.8\%$$

This means that 0.8% (fewer than 1%) of obstetrical patients discharged during October died, or that the maternal death rate for October was 0.8%.

Neonatal, Infant, and Fetal Death Rates

The formulas for these rates are given in Table 10-3. **Neonatal and infant death rates** are computed to examine deaths of the neonate and infant at different stages. A neonatal death is the death of an infant within the first 27 days, 23 hours, and 59 minutes of life. An infant death is death of an infant at any time from the moment of birth through the first year of life. Both of these figures are compared with the number of neonates and infants, respectively, who were discharged and died during the same period.

Fetal death rates are computed to examine differences in the rates of early, intermediate, and late fetal deaths. The definition of early, intermediate, and late fetal deaths may vary from state to state. These deaths are distinguished by the length of gestation or the weight of the fetus.

- Early fetal death (abortion) = less than 20 weeks of gestation or weight 500 grams or less
- Intermediate fetal death = 20 completed weeks of gestation but less than 28 weeks of gestation or weight 501 to 1000 grams
- Late fetal death (stillborn) = 28 weeks of completed gestation and weight more than 1001 grams

See the example of neonatal, infant, and fetal death rates.

Example of Neonatal, Infant, and Fetal Death Rates

Anywhere Health Care Facility developed the following discharge analysis report for 200X. A segment of the report shows the following:

Live births 127

Neonatal discharges 115

Neonatal deaths (before 28 days) 2

Infant discharges 50

Infant deaths (before 1 year and at or after 28 days) 5

Intermediate fetal deaths 13

Neonatal mortality rate:

$$= \frac{2 \text{ neonatal deaths}}{115 \text{ neonatal discharges} + 2 \text{ neonatal deaths}} \times 100$$

$$= 1.71\% \text{ or } 1.7\%$$

This means that 1.7% of the neonates discharged/died or that the neonatal mortality rate for 200X was 1.7%.

Note: Because the intermediate fetal death rate is most commonly used, an example of that rate is given.

Intermediate fetal death rate:

$$= \frac{13 \text{ intermediate fetal deaths}}{127 \text{ live births} + 13 \text{ intermediate fetal deaths}} \times 100$$

$$= 9.29\% \text{ or } 9.3\%$$

This means that intermediate fetal deaths made up 9.3% of live births (excluding the live births at or before 20 weeks' gestation), or that the intermediate fetal death rate for 200X was 9.3%.

Infant mortality rate:

$$= \frac{5 \text{ infant deaths}}{50 \text{ infant discharges} + 5 \text{ infant deaths}} \times 100$$

$$= 0.09\% \text{ or } 9.1\%$$

This means that 9.1% of infants discharged died, or that the infant mortality rate for 200X was 9.1%.

Using and Examining Mortality Rates

Mortality statistics and trends are analyzed and used in many ways. When trends in mortality are examined, the possible reasons for differences in mortality rates should be considered. The influences can be grouped into three variables: time, place, and person. Changes over time include the following:

- Revisions in the rules for *International Classification of Diseases* (ICD) coding of death certificates
- Improvements in medical technology
- Earlier detection and diagnosis of disease
- In relation to place, the following factors influence mortality trends:
 - Changes in the environment
 - International and regional differences in medical technology
 - Diagnostic and treatment practices of physicians

Finally, the following characteristics of groups of people can also influence mortality:

- Age
- Gender
- Race
- Ethnicity
- Social habits (diet, smoking, alcohol intake)
- Genetic background
- Emotional and behavioral health characteristics

All these factors must be taken into consideration when mortality trends are examined within the health care facility or across health care facilities in relation to the quality of care provided.⁵

When examining mortality rates within a specific population as in the gross and net death rates, it is important to

show age-specific rates or to adjust for age. Mortality rates are routinely adjusted for age because it is the most important influence in relation to death. As a person ages, the likelihood that the person will die increases. Age-specific rates can be used, but it becomes difficult to make comparisons of data with four or more age levels or categories. Therefore, age adjustment is performed. Statistically, age adjustment removes the difference in composition with respect to age.¹

Two methods can be used to perform age adjustment. One is the direct method of age adjustment, and the other is the indirect method of age adjustment or standardized mortality ratio (SMR). The calculations for these two methods are shown in Table 10-4.

The direct method uses a standard population and applies the age-specific rates available for each population. The expected number of deaths in the standard population is then determined. To use the direct method of age adjustment, age-specific rates must be available for both populations and the number of deaths per age category should be at least five. The indirect method, or SMR, is used more often and can be used without age-specific rates and when the number of deaths per age category is small or fewer than five. Standard rates are then applied to the populations being compared to calculate the expected number of deaths, which is compared with the observed number of deaths.⁶

Because the SMR is used in most national and statewide mortality reports, it is explained in more detail here. For example, in Table 10-5, hospitals across a state are examined for death rates associated with the diagnosis-related group (DRG) 127—Heart Failure and Shock.

The actual or observed number of deaths in the hospital is compared with the expected number of deaths. The expected number of deaths is taken from a comparative national database adjusted for age and patient severity for each DRG. Table 10-5 shows a sample of the hospitals that treated patients included in DRG 127 and the actual and expected

number of deaths. An SMR of 1 means that the number of observed deaths and the number of expected deaths are equal, and therefore the mortality rate is equal to what is expected from national norms. An SMR less than 1 means that the observed deaths are lower than the expected deaths, and therefore the mortality rate is lower than expected from national norms. An SMR of greater than 1 means that the observed deaths are greater than the expected deaths, and therefore the mortality rate is higher than expected from national norms (see the examples of use of SMR).

Example Use of Standardized Mortality Ratio (SMR)

For hospital 1, an SMR of 1.09 means that the hospital had a 9% higher mortality rate for DRG 127 than is expected from national norms. This is calculated as follows:

$$\text{SMR} = \frac{23 \text{ actual deaths}}{21.03 \text{ expected deaths}} = 1.09$$

$$(1.09 - 1) \times 100 = 0.09 \times 100 = 9\%$$

Example Use of Standardized Mortality Ratio

For hospital 4, an SMR of 0.48 means that the hospital had a 52% lower mortality rate for DRG 127 than is expected from national norms. This is calculated as follows:

$$\text{SMR} = \frac{8 \text{ actual deaths}}{16.56 \text{ expected deaths}} = 0.48$$

$$(1 - 0.48) \times 100 = 0.52 \times 100 = 52\%$$

The statistical rating column displayed in Table 10-5 is covered later in this chapter in the discussion of tests of significance.

Table 10-4 AGE ADJUSTMENT METHODS

Direct Method	Formula
Age-adjusted death rate (A)	$\frac{\text{Total expected number of deaths at population A rates}}{\text{Total standard population}} \times \text{Constant}$
Age-adjusted death rate (B)	$\frac{\text{Total expected number of deaths at population B rates}}{\text{Total standard population}} \times \text{Constant}$
Compare ages = adjusted death rates for populations A and B	
Indirect Method	Formula
SMR for population A	$\frac{\text{Observed deaths in population A}}{\text{Expected deaths in population A at standard rates}}$
SMR for population B	$\frac{\text{Observed deaths in population B}}{\text{Expected deaths in population B at standard rates}}$
Compare the two SMRs for populations A and B.	

SMR, Standardized mortality ratio.

Table 10-5 DIAGNOSIS-RELATED GROUP 127 HEART FAILURE AND SHOCK

Hospital	Comments	Number of Patients	Average Admission Severity Score	Age 65 and over (%)	Deaths			Medically Unstable During First Week: Major Morbidity				
					Actual Number	Expected Number	Statistical Rating	Actual Number	Expected Number	Statistical Rating	Average Stay (Days)	Average Charge (\$)
1	√	268	2.5	85.1	23	21.03		35	25.12	—	7.9	15,420
2		412	2.4	87.1	30	33.08		61	36.66	—	9.2	8149
3		201	2.2	87.1	17	12.16		9	14.52		9.4	7645
4		208	2.6	64.4	8	16.56	+	24	21.52		7.7	15,669
5		471	2.5	89.0	40	40.31		40	44.18		8.3	8193
6		90	2.6	78.9	9	7.49		12	9.35		9.0	14,766
7	√	347	2.3	81.3	36	22.31	—	24	27.78		8.9	12,099
8		291	2.1	90.0	20	18.15		20	20.97		8.7	9180
9		255	2.3	82.0	11	17.04		18	21.09		6.4	6292
10		477	2.2	85.5	32	29.55		32	36.78		8.4	12,039

Hospital Effectiveness Report, Pennsylvania Health Care Cost Containment Council, Reporting Period January 1–December 31, 1991.

SELF-ASSESSMENT**Quiz**

- Statistics relating to vital events in our lives, such as births and adoptions, marriages, divorces, and deaths, including fetal deaths are called:
 - Vital statistics
 - Census statistics
 - Mortality statistics
 - Morbidity statistics
- 50/1000; 50/100; 4/100,000 are all examples of a:
 - Ratio
 - Proportion
 - Percentage
 - Measures of central tendency
- An early fetal death occurs at less than 20 weeks gestation or when fetal weight is 500 grams or less. This may also be referred to as a(n):
 - Abortion
 - Infant death
 - Neonatal death
 - Infant mortality ratio

Autopsy Rates

Autopsy rates are computed so that the health care facility can determine the proportion of deaths in which an autopsy was performed. This enables the facility to examine why a higher or lower autopsy rate may be seen from one month to another. Autopsies are performed to determine the cause of death, to better understand the disease process, or to collect tissue samples, as in patients with Alzheimer's disease. Autopsy rates can be further broken down to show the gross autopsy rate, or the rate of autopsies performed for total inpatient deaths; the net autopsy rate, or the rate of autopsies performed for inpatient deaths, excluding unautopsied coroner cases; and the adjusted hospital autopsy rate or the autopsy rate performed for all deaths of hospital patients whose bodies are available or

brought to the hospital for autopsy (those not removed by coroners, medical examiners, and so on). Autopsies may be performed after the deaths of inpatients, outpatients, home care patients, skilled nursing care residents, patients who died at home, previous patients, and so on (see example of hospital autopsy rates). Table 10-6 presents the most commonly used autopsy rates.³

Example of Hospital Autopsy Rates

Anywhere Health Care Facility developed the following report regarding discharges, deaths, and autopsies during January 200X.

Hospital Statistics

Discharges (including deaths) 1000

Total deaths 56

Inpatient (including two coroner cases) 52

Outpatient 2

Home care 2

Autopsies 13

Inpatient 10

Outpatient 1

Home care 2

Gross autopsy rate:

$$= \frac{10 \text{ autopsies on inpatients}}{52 \text{ inpatient deaths}} \times 100$$

$$= 19.23\% \text{ or } 19.2\%$$

This means that 19.2% of the hospital inpatients who died during January received an autopsy, or that the gross autopsy rate was 19.2%.

Net autopsy rate:

$$= \frac{10 \text{ autopsies on inpatients}}{52 \text{ inpatient deaths}} \times 100$$

$$- 2 \text{ unautopsied coroners' cases}$$

$$= 20\%$$

This means that 20% of the hospital inpatients who died during January received an autopsy within the hospital, or that the net autopsy rate was 20%.

Table 10-6 AUTOPSY RATES

Autopsy Rate	Formula
Gross autopsy rate (ratio of inpatient autopsies to inpatient deaths)	$\frac{\text{Total inpatient autopsies}}{\text{Total inpatient deaths}} \times 100$
Net autopsy rate	$\frac{\text{Total inpatient autopsies}}{\text{Total inpatient deaths—unautopsied coroners' cases}} \times 100$
Hospital autopsy rate (adjusted) Total hospital autopsies 100	$\frac{\text{Total hospital autopsies}}{\text{Number of deaths of hospital patients whose bodies are available for hospital autopsy}} \times 100$

Note: Numerators and denominators in each formula must be for the same time period.

$$\begin{aligned}
 &\text{Hospital autopsy rate (adjusted):} \\
 &= \frac{13 \text{ total hospital autopsies}}{52 \text{ inpatient deaths} - 2 \text{ coroner cases} + 2 \text{ outpatients} + 2 \text{ home health care patients}} \times 100 \\
 &= 24.07\% \text{ or } 24.1\%
 \end{aligned}$$

This means that 24.1% of all health care facility patients who died in January (inpatients, outpatients, and home care patients) received an autopsy within the hospital, or that the adjusted hospital autopsy rate was 24.1%.

Morbidity Rates

Morbidity rates can include complication rates, such as community-acquired, hospital-acquired or nosocomial, and postoperative infection rates. They can also include comorbidity rates and the prevalence and incidence rates of disease.

Hospitals use each of these rates to study the types of disease or conditions that are present within the health care facility and to examine the quality of care provided by the facility. These rates can aid health care facilities in planning specific health care services and programs. Table 10-7

provides a summary of the more common morbidity rates and the formulas used to compute them.³

Complications include infections, allergic reactions to medications, transfusion reactions, decubitus ulcers, falls, burns, and errors of medication administration. The complication rates for any of these complications can also be computed by using the formula for complication rates listed in Table 10-7.

One of the most common complications is infections. Infection rates are computed so that the health care facility can determine when infections developed and, therefore, how they may be prevented. A **nosocomial**, or facility-acquired, **infection rate** includes infections that occur more than 72 hours after admission.⁷ Health care facilities may be more interested in this rate because it may show infections that occur as a result of the care that is provided in the facility. Further analysis of the nosocomial infection rate may show that other risk factors, such as age, compromising conditions such as cancer, the use of chemotherapy treatment, and the overall severity of the disease, may make an individual patient more susceptible to infection. Therefore, as with several of the mortality rates, other factors play a part in the development of the nosocomial infection. The postoperative infection rate is normally calculated to pinpoint how the infection may have developed. **Postoperative infection rates** are important to examine because the health

Table 10-7 MORBIDITY RATES

Definition	Formula
Complication (condition that occurs during hospital stay that extends length of stay by at least 1 day in 75% of cases)*	$\frac{\text{Total number of complications}}{\text{Total number of discharges}} \times 100$
Nosocomial infection rate (infection that occurs >72 hours after admission to hospital)*	$\frac{\text{Total number of infections that occur > 72 hours after admission}}{\text{Total discharges}} \times 100$
Postoperative infection rate*	$\frac{\text{Total number of postoperative infections}}{\text{Total number of discharges}} \times 100$
Community-acquired infection rate (infection that occurs in community or <72 hours of admission)*	$\frac{\text{Total number of community-acquired infections that occur < 72 after admission}}{\text{Total number of discharges}} \times 100$
Total infection rate (includes both nosocomial and community-acquired infections)*	$\frac{\text{Total number of community-acquired and nosocomial infections}}{\text{Total number of discharges}} \times 100$
Comorbidity (preexisting condition that will, because of its presence with principal diagnosis, increase the length of stay by at least 1 day in 75% of cases)*	$\frac{\text{Total number of comorbidities}}{\text{Total number of discharges for a given period}} \times 100$
Prevalence (number of people with specific disease at specified period of time; number of existing cases of disease)	$\frac{\text{Number of cases of disease present in a population at specified time period}}{\text{Number of people in population at specified time period}} \times 100$
Incidence (number of people with disease during specified time period; number of new cases of disease)	$\frac{\text{Number of new cases of a disease occurring in population during a specified time period}}{\text{Number of people in population at specified time period}} \times 100$

*Numerators and denominators in each formula must be for the same time period.

care facility can determine which infections occur after surgery and are probably a result of the surgical procedure.

Distinguishing between nosocomial and community-acquired infections is important because **community-acquired infections** are typically present less than 72 hours before admission to the health care facility. Health care facilities may be interested in this rate because it demonstrates the infections that patients probably had before admission to the facility. If the facility finds that their community-acquired infection rate is high, they may need to develop community-wide prevention programs, such as administering a vaccine for pneumonia. Health care facilities can benefit from analysis of their total infection rate (both nosocomial and community-acquired infections) to determine the additional cost, length of stay, and overall effect the infections have on the quality of care provided to the patient.

Comorbidities are preexisting conditions, such as diabetes, hypertension, and osteoporosis. Analysis of the comorbidity rate is important because comorbidities can increase the length of stay and affect the outcome of care provided. Comorbidities include some of the other risk factors that affect mortality and morbidity rates.

Morbidity Data for Anywhere Health Care Facility During March 200X

Discharges (including deaths) 2000

Surgical operations 1543

Number of comorbidities 238

Number of complications 120

Nosocomial infections (includes postoperative infections) 22

Postoperative infections 8

Community-acquired infections 30

Complication rate:

$$= \frac{120 \text{ total complications}}{2000 \text{ discharges (including deaths)}} \times 100$$

$$= 6.0\%$$

This means that 6.0% of all discharges for March had at least one complication, or that the complication rate for March was 6.0%.

Nosocomial infection rate:

$$= \frac{22 \text{ nosocomial infections}}{2000 \text{ discharges (including deaths)}} \times 100$$

$$= 1.1\%$$

This means that 1.1% of all discharges for March had a nosocomial or hospital-acquired infection, or that the nosocomial or hospital-acquired infection rate for March was 1.1%.

Postoperative infection rate:

$$= \frac{8 \text{ postoperative infections}}{1543 \text{ surgical operations}} \times 100$$

$$= 0.52\%$$

This means that 0.5% of all those with surgical operations performed during March developed a postoperative infection, or that the postoperative infection rate for March was 0.5%.

Community-acquired infection rate:

$$= \frac{30 \text{ community-acquired infections}}{2000 \text{ discharges (including deaths)}} \times 100$$

$$= 1.5\%$$

This means that 1.5% of all discharges for March had a community-acquired infection, or that the community-acquired infection rate for March was 1.5%.

Total infection rate:

$$= \frac{52 \text{ total infections}}{2000 \text{ discharges (including deaths)}} \times 100$$

$$= 2.6\%$$

This means that 2.6% of all discharges for March had an infection, or that the total infection rate for March was 2.6%.

Comorbidity rate:

$$= \frac{238 \text{ total comorbidities}}{2000 \text{ discharges (including deaths)}} \times 100$$

$$= 11.9\%$$

This means that 11.9% of all discharges for March had at least one comorbidity, or that the comorbidity rate for March was 11.9%.

Prevalence and incidence rates are determined to examine the frequency of specific types of disease, such as cancer, acquired immunodeficiency syndrome (AIDS), and heart disease. **Prevalence** means the number of existing cases of disease, whereas **incidence** refers to the number of new cases of disease (see the examples of prevalence and incidence rate determinants).

The **prevalence rate** is the number of existing cases of a disease in a specified time period divided by the population at that time. The quotient is then multiplied by a constant, such as 1000 or 100,000.

Example of Prevalence Rate

In a community of elderly people, the number of women alive with osteoporosis in the year 200X is 3593. The population of women in this community is 100,000.

Prevalence rate:

$$= \frac{3593 \text{ women with osteoporosis}}{100,000 \text{ women in population}} \times 1000$$

$$= 35.93, \text{ or } 36 \text{ osteoporosis cases per } 100,000 \text{ women in this community}$$

The **incidence rate** is the number of newly reported cases of a disease in a specified time period divided by the population at that time. The quotient is then multiplied by a constant such as 1000 or 100,000.

Example of Incidence Rate

In the same elderly community, the number of new cases of osteoporosis reported in 200X is 1113, and the population of women in the community at that time is 100,000.

$$\begin{aligned} \text{Incidence rate:} \\ &= \frac{1113 \text{ new cases of osteoporosis}}{100,000 \text{ women in this community}} \times 1000 \\ &= 11.13, \text{ or } 11 \text{ new osteoporosis cases per } 100,000 \\ &\quad \text{women in population} \end{aligned}$$

To manage health care services effectively, the HIM professional should analyze prevalence and incidence rates of specific diseases that are prominent within a particular region or state. National sources of morbidity data include the National Health Care Survey. Originated in 1956, this survey is performed annually on a representative sample of 40,000 persons. Many subprograms are part of the National Health Care Survey, such as the National Hospital Discharge Survey, National Hospital Ambulatory and Medical Care Survey, and National Nursing Home Survey. Results of these surveys include incidence and prevalence rates of disease for specific geographic areas, length of hospital stays, cause of hospitalizations, and use of ambulatory care services.¹ The HIM professional should be aware that this information exists and can be used in conjunction with other morbidity rates to analyze further the distribution and effectiveness of health care services.

Characteristics similar to those that influence trends in mortality also influence trends in morbidity—time, place, person. For example, infectious diseases tend to occur more often at specific times of the year. The place of employment or geographic location can also increase susceptibility to disease. Age can influence the occurrence of infectious diseases; for example, diseases such as measles and chickenpox are more common in the young. Gender can influence morbidity trends, with differences in coronary artery disease for men and women. Race also influences morbidity trends, with hypertension being more prevalent in African Americans.⁵

Census Statistics

Ratios, percentages, and averages related to the length of stay, occupancy, bed turnover, and total number of patients present at a specified time within the institution can be useful

both to health care administrators and HIM professionals. Such data can be used for the following purposes:

- To evaluate the current status of the health care facility
- To plan for future health care events
- To compare utilization of various units within a health care organization

The **census statistics** are extremely useful in the overall analysis of how much, how long, and by whom the health care facility is being used. Table 10-8 provides the formulas for the common census statistics³ (see the example of census statistics).

Example of Census Statistics

A 500-bed health care facility, during the month of June (30 days), had a total of 3600 discharges (including deaths), a total of 14,647 inpatient service days, and a total of 15,567 discharge days.

$$\begin{aligned} \text{Inpatient bed occupancy rate (percentage of occupancy):} \\ &= \frac{14,647 \text{ inpatient services days}}{500 \text{ (beds)} \times 30 \text{ (Number of days in June)}} \times 100 \\ &= 97.6\% \end{aligned}$$

This means that 97.6% of the available beds were occupied, or that the inpatient bed occupancy ratio was about 14,647:15,000, or that the percentage of occupancy was 97.6%.

$$\begin{aligned} \text{Average daily inpatient census:} \\ &= \frac{14,647 \text{ inpatient service days}}{30 \text{ (Number of days in June)}} = 488 \end{aligned}$$

This means that the average number of inpatients during June was 488 or that the average daily inpatient census for June was 488.

$$\begin{aligned} \text{Average length of stay:} \\ &= \frac{15,567 \text{ discharge days}}{3600 \text{ discharges (including deaths)}} = 4.3 \text{ days} \end{aligned}$$

This means that patients stayed in the health care facility an average of 4.3 days during June, or that the average length of stay (ALOS) for June was 4.3.

Example of Census Statistics for Oncology Department

Discharges (including deaths)	1322
Hospital days for discharged Patients	10,576
Patient A	admitted June 18 and died the same day
Patient B	admitted June 18 and discharged on June 19

Table 10-8 CENSUS STATISTICS

Definition	Formula
Daily inpatient census (no. of inpatients present at census-taking time plus any inpatients who are both admitted and discharged after census-taking time the previous day).	Formula is presented as the definition.
Inpatient service day (unit of measure including services received by one inpatient in one 24-hour period).	Formula is presented as the definition.
Synonyms: patient day, inpatient day, census day, bed occupancy day.	
Inpatient bed count (no. of available inpatient beds [occupied and vacant] on any given day).	Formula is presented as the definition.
Note: Not all beds are included in the inpatient bed count. These include beds in examination rooms, therapy, labor rooms, and recovery rooms as well as bassinets. (Beds set up for temporary use are not included.)	
Average daily inpatient census (average number of inpatients in a facility for a given period of time).	$\frac{\text{Total number of service days for a period}}{\text{Total number of days in that period}}$
Length of stay (for an inpatient, number of calendar days from admission to discharge).	Duration of hospitalization for one inpatient. Day of admission is not counted unless it is the day of discharge or the day of discharge is not counted unless it is the day of admission. Either method is correct if done consistently.
Synonyms: discharge days, duration of inpatient hospitalization, days of stay.	
Total length of stay (for all inpatients: total days in facility of any group of inpatients discharged during specified period). Synonyms: discharge days, total inpatients days of stay.	
Average length of stay (average length of stay of inpatients discharged during specified period).	$\frac{\text{Total length of stay (discharge days)}}{\text{Total number of discharges (including deaths)}}$
Inpatient bed occupancy rates (proportion of inpatient bed occupied, defined as ratio of inpatient service days to inpatient bed count days in specified period).	$\frac{\text{Total service days for a period}}{\text{Total bed count days in the period}} \times 100$ (Bed count \times number of days in period)
Synonyms: percentage of occupancy, occupancy percentage, occupancy rate.	
Bed turnover rate (number of times a bed, on average, changes occupants during a given period of time).	Direct formula: $\frac{\text{Number of discharges (including deaths) for a period}}{\text{Average bed count during the period}}$ Indirect formula: $\frac{\text{Occupancy rate} \times \text{number of days in period}}{\text{Average length of stay}}$

Patient C admitted on June 19 and discharged June 25

Patient D admitted June 25 and discharged on August 8

Average length of stay for Oncology Department:

$$\frac{10,576 \text{ discharge days for oncology patients}}{1322 \text{ discharges from oncology department}} = 8.0 \text{ days}$$

Patients in the Oncology Department stayed an average of 8.0 days during June, or the average length of stay for the Oncology Department for June was 8 days. The

Oncology Department average length of stay (8 days) can then be compared with the overall facility length of stay (4 days) and determine why the Oncology Department length of stay is double that of the facility length of stay.

Length of Stay for individual patients

Patient A = 1 day because the patient was admitted and died on the same day

Patient B = 1 day because the patient was admitted one day and discharged the next

Patient C = 6 days by subtracting the date of admission from the date of discharge because the patient was admitted and discharged within the same month

Patient D = 44 days = 5 days in June + 31 days in July + 8 days in August

The individual patient lengths of stay can be compared with one another, especially if the patients received the same services or were from the same department.

Bed Turnover Rate:

Direct method :

$$= \frac{3600 \text{ discharges (including deaths)}}{500 \text{ beds}} = 7.2$$

Indirect method :

$$\begin{aligned} & 97.6\% \text{ (percentage of occupancy)} \\ & \times 30 \text{ days in June} \\ = & \frac{4.3 \text{ (average length of stay for June)}}{0.976 \times 30} = 6.8 \end{aligned}$$

This means that during June each of the hospital's 500 beds changed patients about 7.2 times according to the direct method and 6.8 times according to the indirect method—a small difference between the two methods.

SELF-ASSESSMENT

Quiz

- Fifty-five lymphedema cases per 1000 women diagnosed with breast cancer in San Diego is an example of what type of rate:
 - Prevalence rate
 - Incidence rate
 - Mortality rate
 - Morbidity rate
- Two-hundred Alzheimer's disease cases per 1000 men under age 50 in Minneapolis is an example of what type of rate:
 - Prevalence rate
 - Incidence rate
 - Nosocomial infection rate
 - Community-acquired infection rate
- Ratios, percentages, and averages related to the length of stay, occupancy, bed turnover, and total number of patients present at a specified time within the institution are statistics referred to as:
 - Infections rates
 - Mortality rates
 - Census statistics
 - Morbidity statistics

ORGANIZING AND DISPLAYING THE DATA⁸

Types of Data

Before it is decided how to display data, it is important to recognize that different methods of display are appropriate for different types of data. Variables or data can be grouped

into the following four categories on the basis of what kind of information or meaning the numbers convey:

- Nominal
- Ordinal
- Interval
- Ratio

Variables can also be classified as discrete or continuous based on how many possible values the variable can assume.

Nominal Data

The term **nominal data** is used to describe data collected on variables for which qualitative (what kind) rather than quantitative (how much or how many) differences exist between individuals. Nominal data are also called categorical, qualitative, or named data. Examples of nominal variables include the gender and race of subjects in a research study. To facilitate tabulating and analyzing data, numerical values are often assigned to the categories of nominal variables. Using the variable "gender" as an example, the female category could be coded as "0" and the male category could be coded as "1." The variable employment status could be coded "10" for employed and "9" for unemployed, and so on. It is important to realize the numerical values used to represent nominal data only serve as labels for categories. The values or codes convey no quantitative (how much or how many) information. Therefore, the choice of numerical values is arbitrary; gender could also be coded as 1 for female and 2 for male, and so forth. Table 10-9 shows nominal data of types of health insurance by gender.

Ordinal or Ranked Data

Ordinal data are data expressing rankings from lowest to highest according to some criterion. An example of the use of ordinal data is found in severity of illness scores used in assessing quality of care outcomes. Atlas, a severity of illness

Table 10-9 FREQUENCY TABLE—NOMINAL DATA: PRINCIPAL HEALTH INSURANCE COVERAGE BY GENDER

Health Insurance	Male (n = 50) Number (%)	Female (n = 50) Number (%)	Total (n = 100) Number (%)
Medicare	13 (26)	25 (50)	38 (38)
Medicaid	2 (4)	6 (12)	8 (8)
Blue Cross	25 (50)	10 (20)	35 (35)
Commercial	9 (18)	6 (12)	15 (15)
Other	1 (2)	3 (6)	4 (4)
Totals	50 (100)	50 (100)	100 (100)

From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from the American Health Information Management Association.

system, uses the following ordinal data to describe the severity of illness:

- 0 = no or minimal risk of vital organ failure
- 1 = low risk of vital organ failure
- 2 = moderate risk of vital organ failure
- 3 = high risk of vital organ failure
- 4 = presence of vital organ failure

Ordinal data can also include responses to questionnaires or interviews:

- 1 = strongly disagree
- 2 = disagree
- 3 = neutral
- 4 = agree
- 5 = strongly agree

Commonplace examples of ordinal data include class rank of graduating high school seniors and the ranking of sports teams within a league. A key feature of ordinal data is that the equal distances between ranks do not necessarily correspond to equal distances on the underlying criterion. For example the distance between class ranks of 1 and 2 is equal to the distance between class ranks of 3 and 4. However, when we consider the underlying criterion of grade point average, those of the students ranked as 1 and 2 may be closer to each other than are those of the students ranked and 3 and 4. See Table 10-10 for an example of ordinal data used in a frequency table explaining student perceptions of leadership characteristics.

Interval Data

Interval data convey more precise quantitative information than do ordinal data because it is assumed that equal differences between numbers correspond to equal differences in the trait or characteristic being measured. Examples of

interval data include scores on college examinations and scores on nationally administered tests such as the SAT (Scholastic Aptitude Test). If Student A received a score of 400 on the verbal section of the SAT, Student B received a score of 450, Student C received a score of 500, and Student D received a score of 550, it is assumed that the difference in verbal aptitude between Student A and Student B is equal to the difference in verbal aptitude between Student C and Student D.

Ratio Data

Ratio data share the property of equal differences with interval data. What is unique about ratio data is that the value of 0 represents the total absence of the trait or characteristic being measured. For example, if the speed of a car when it is stopped at a red light is 0, the value of 0 can be interpreted as the absence of speed; the car is not moving. Other examples of ratio data include height and weight of patients and length of stay at a hospital.

Discrete Data

Discrete data are data on quantitative variables that can only take on a limited number of values, typically only whole numbers. Examples of discrete data include the number of medications a person is taking, the number of children in a family, or the number of records that are coded. See Table 10-11 for an example of discrete data used in a frequency table on hospital admissions by residence.

Continuous Data

Continuous data are data on quantitative variables that can assume an infinite number of possible values. Examples include height, weight, temperature, and costs or charges. See Table 10-12 for an example of continuous interval data for total charges for inpatients in a large teaching hospital.

Table 10-10 FREQUENCY TABLE—RANKED ORDINAL DATA: STUDENT PERCEPTIONS OF LEADERSHIP CHARACTERISTICS

Leadership Characteristic Ranking	Management Clinical Internship (n = 35)	
	Before Number (%)	After Number (%)
1 Very weak	5 (14)	0 (0)
2 Weak	10 (29)	1 (3)
3 Moderate	15 (43)	5 (14)
4 Strong	2 (6)	12 (34)
5 Very strong	3 (9)	17 (49)

From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from the American Health Information Management Association.

Table 10-11 FREQUENCY TABLE—DISCRETE DATA: UNIVERSITY HEALTH CENTER HOSPITAL ADMISSIONS, BY RESIDENCE

	City (n = 77) Number (%)	Suburbs (n = 38) Number (%)	Rural (n = 29) Number (%)	Total (n = 144) Number (%)
Hospital A	20 (26)	8 (21)	12 (41)	40 (28)
Hospital B	30 (39)	4 (11)	9 (31)	43 (30)
Hospital C	10 (13)	12 (32)	6 (21)	28 (19)
Hospital D	17 (22)	14 (37)	2 (7)	33 (23)
Total	77 (100)	38 (100)	29 (100)	144 (100)

From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from the American Health Information Management Association.

Table 10-12 FREQUENCY TABLE—CONTINUOUS INTERVAL DATA: TOTAL CHARGES FOR 152 INPATIENTS IN A LARGE TEACHING HOSPITAL

Total Charges (\$)	Frequency	Relative Frequency (%)
0-4999	62	40.8
5000-9999	46	30.3
10,000-14,999	25	16.5
15,000-19,999	7	4.6
20,000-24,999	5	3.3
25,000-29,999	4	2.6
30,000-34,999	0	0
35,000-39,999	0	0
40,000-44,999	0	0
45,000-49,999	3	2.0

From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from the American Health Information Management Association.

Types of Data Display

Many methods are used to display data effectively (Box 10-1). Some methods of particular value to the HIM professional are frequency distribution tables, bar graphs, pie charts, histograms, frequency polygons, and Pareto diagrams.

Frequency Distribution

A **frequency distribution** table presents the number of times that each category of a qualitative variable or value of a quantitative variable is observed within a sample. When continuous variables that have a large number of possible values are represented, it is common for the frequencies to be reported in ranges or intervals of values, rather than the frequencies of individual values. For example, a frequency table representing the age of patients could show the

number of patients whose ages are between 20 and 29 years, 30 and 39 years, 40 and 49 years, and so forth. To allow comparisons to be made across samples of varying size, percentages that represent relative frequency are usually reported along with the frequency count. To compute these percentages, the number of observations within a category is divided by the total sample size, and the result is multiplied by 100. The frequency table should be self-explanatory and not show so much data that the table becomes uninterpretable. The table should be clearly labeled, the total sample size displayed, and units of measurement included. When intervals are used, the number of intervals should be not less than 5 and no more than 20 and of equal width, and the end points of the intervals are mutually exclusive and do not overlap. Tables 10-9 through 10-13 are examples of frequency distribution tables for nominal, ordinal, continuous, and discrete data, respectively.

Bar Graph

Bar graphs are normally used to illustrate nominal, ordinal, and discrete data. The discrete categories are shown on the horizontal, or *x*, axis and the frequency is shown on the vertical, or *y*, axis. The purpose of the bar graph is to show the frequency for each interval or category. The scale of the vertical axis must begin at zero so that the heights of the bars are proportional to the frequencies. By using different colors or patterns of bars to represent different samples, bar graphs can be used to compare the frequency of categories or intervals in two or more samples. Figure 10-1 is an example of a bar graph using the data in Table 10-9. Other types of bar graphs are those shown in Figures 10-2 and 10-3. These graphs are still considered bar graphs but incorporate a line over the bars to show the total number of cases of salmonellosis (see Figure 10-2). Also, Figure 10-3 shows the incidence of meningococcal disease by age with use of a stacked bar graph.

Pie Chart

A **pie chart** is effective for representing the relative frequency of categories or intervals within a sample. It is constructed by drawing a circle, 360 degrees, and dividing that circle into sections that correspond to the relative frequency in each category. For example, if the relative frequency is 15%, then the slice of pie should span $(0.15) \times (360 \text{ degrees})$, or 54 degrees. Figure 10-4 is an example of a pie chart with use of the data in Table 10-9.

Histogram

The **histogram** is usually used to present a frequency distribution for continuous data. It is similar to a bar graph, but the horizontal axis of the histogram usually represents intervals of a continuous variable rather than a discrete variable. Because theoretically there is no separation between the end point of one interval and the starting point of another, the bars of a histogram touch. The heights of the bars correspond to the frequency within each interval. Because intervals are usually of equal width,

Box 10-1 GUIDELINES FOR DISPLAYING DATA

- Ask yourself what is the main message you wish to convey to the reader and then choose the type of graph that is most appropriate for the information you want to communicate.
- Be careful to choose a type of graph that is consistent with the type of data shown in the graph. For example, do not choose a histogram for nominal data.
- Create an effective graph that is visually attractive and numerically accurate.
- Aim for simplicity; unnecessary detail in a graph can be distracting to the reader.
- When applicable, give careful consideration to the scaling of the vertical axis (the minimum and maximum values displayed, and the number of units between consecutive labeled values).

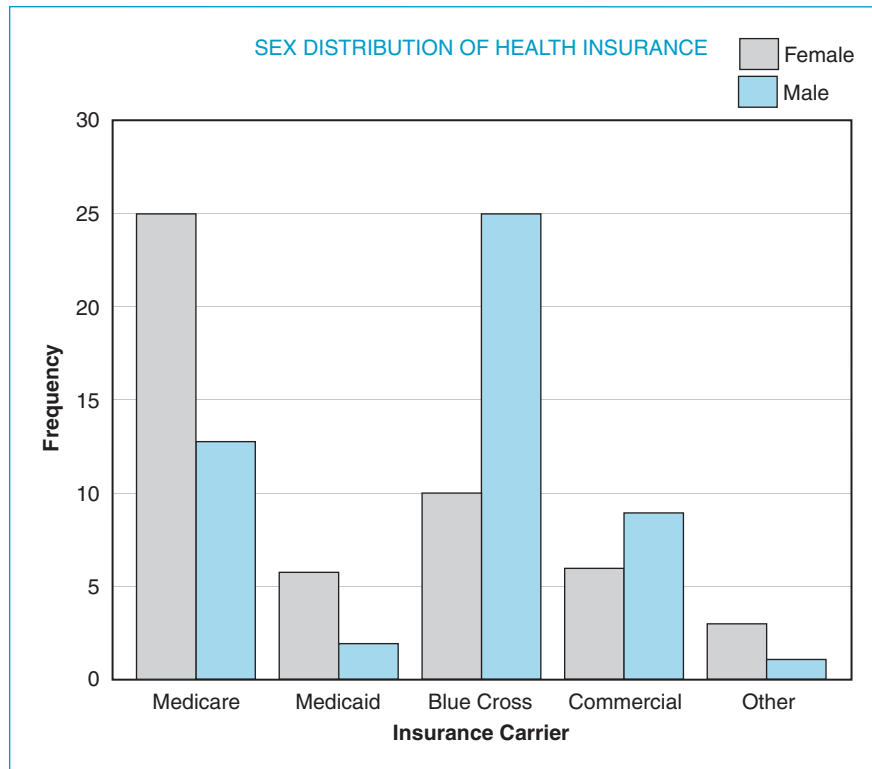


Figure 10-1 Bar graph. (From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from American Health Information Management Association.)

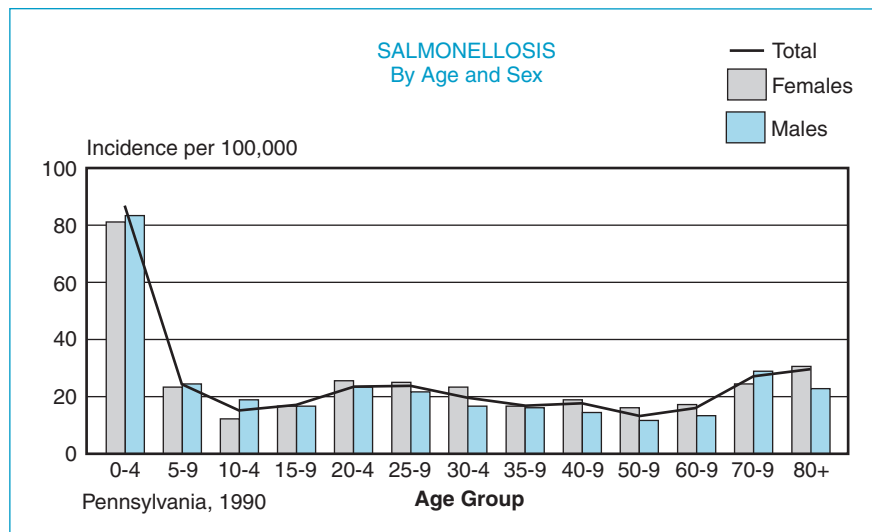


Figure 10-2 Bar graph and line graph. (From Infectious Disease Epidemiology Report, Pennsylvania Department of Health, Bureau of Epidemiology and Disease Prevention, 1990.)

Figure 10-3 Stacked bar graph. (From Infectious Disease Epidemiology Report, Pennsylvania Department of Health, Bureau of Epidemiology and Disease Prevention, 1990.)

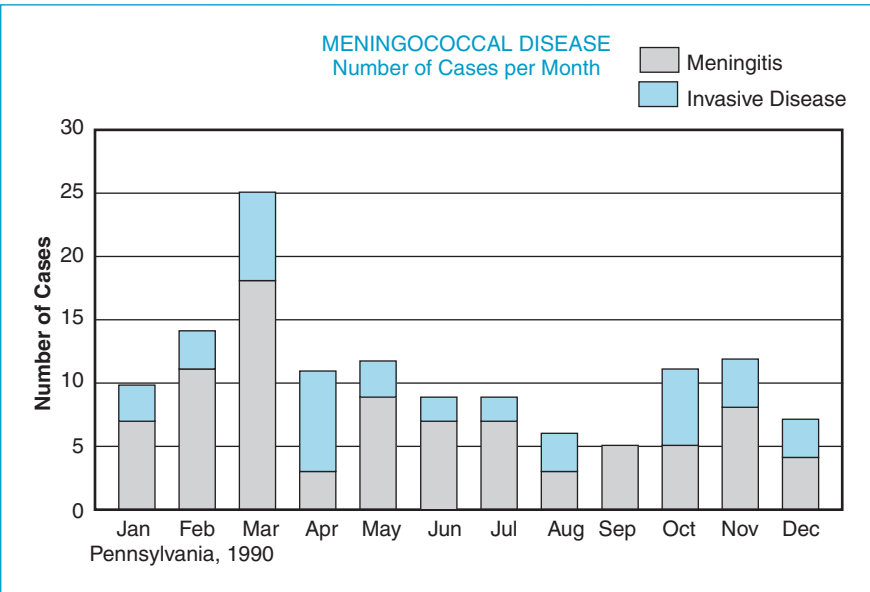


Figure 10-4 Pie chart. (From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from American Health Information Management Association.)

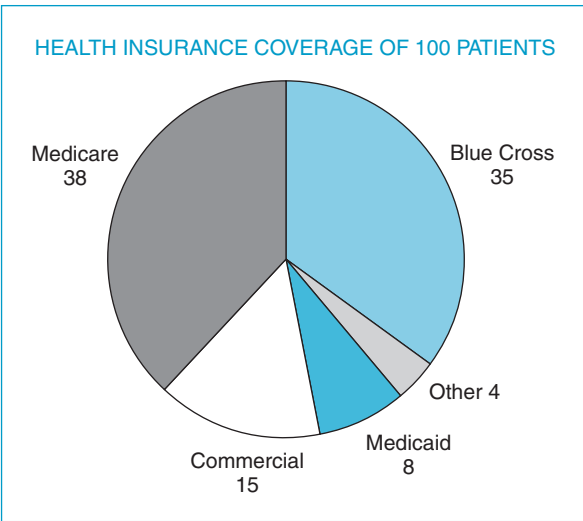
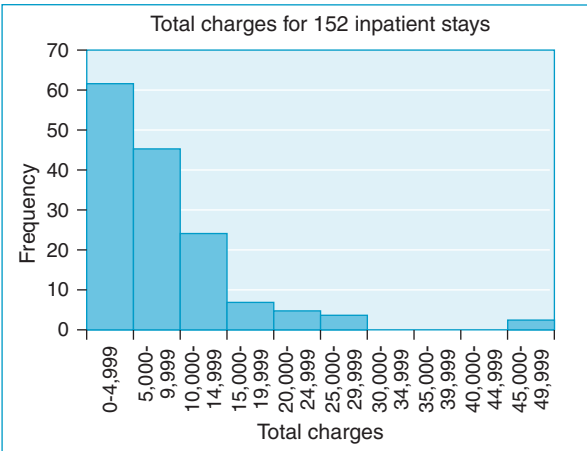


Figure 10-5 Histogram. (From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from American Health Information Management Association.)



the width of all bars of a histogram is usually the same. If the areas (interval width \times interval frequency) of all the bars are summed, the result should be equal to the total sample size multiplied by the common interval width. Figure 10-5 is an example of a histogram with use of data from Table 10-11.

Frequency Polygon

The **frequency polygon** is another method used to present a frequency distribution with continuous data. It is constructed by joining the midpoints of the tops of the bars of a histogram with a straight line. The total area under the polygon is equal to the sum of the areas of the bars in the histogram and therefore equal to total sample size multiplied by interval width. The frequency polygon is effective when comparing the distribution of a variable in two or more data samples. Figure 10-6 is an example of a frequency polygon using data from Table 10-11. Figure 10-7 is an example of a frequency polygon comparing two data sets.

The reader may find variations of these data presentation methods. For example, statistical process control, which is used in the quality improvement process, uses many of the graphs and figures displayed previously but may change them slightly or call them different things. For example, the Pareto diagram (Figure 10-8) is similar to the bar graph and the histogram and is used to order causes or problems from most to least significant.

graph has no space between bars and is normally used for continuous data.

- c. There are no differences between these two types of graphs.
- d. The histogram is used for discrete data only, and the bar graph is used for continuous data only.

STATISTICAL MEASURES AND TESTS

Descriptive Statistics

The following measures and methods are often referred to as descriptive statistics because the objective is to summarize and describe significant characteristics of a set of data.

Measures of Central Tendency

The common measures of central tendency are the following:

- Mean
- Median
- Mode

These measures are used to locate the middle, average, or typical value in a data set. The selection of which one of these measures is most suitable in a given situation depends on the type of data and the purpose for which the measure is being reported.

SELF-ASSESSMENT

Quiz

1. Gender, race, and insurance class are all examples of which type of data?
 - a. Nominal
 - b. Ordinal
 - c. Interval
 - d. Ratio
2. Scores on the Registered Health Information Administrator (RHIA) and Registered Health Information Technician (RHIT) examinations are examples of which type of data?
 - a. Nominal
 - b. Ordinal
 - c. Interval
 - d. Ratio
3. The major differences between the histogram and the bar graph are:
 - a. The bar graph has space between the bars and is normally used for discrete data, whereas the histogram has no space between bars and is normally used for continuous data.
 - b. The histogram has space between the bars and is normally used for discrete data, whereas the bar

Mean

The **mean** is the most common measure of central tendency because it is easily understood by audiences and also because determining the mean is a step toward deriving other statistics discussed later, such as the variance and standard deviation. The purpose of the mean is to summarize an entire set of data by means of a single representative value. The mean or average is calculated by adding up the values of all the observations and dividing the total by the number of observations. Sometimes it is necessary to calculate an overall mean for a total sample when separate means are reported for different

Example of Mean and Weighted Mean

The lengths of stay in the hospital for eight patients in a pediatric department are 6, 4, 2, 5, 20, 25, 18, and 4 days. The mean, or average, length of stay is calculated as follows:

Average length of stay (mean):

$$\begin{aligned}
 & 6 + 4 + 2 + 5 + 20 + 25 + 18 + 4 \\
 &= \frac{(\text{lengths of stays of each patient})}{8 \text{ patients}} \\
 &= \frac{84}{8} \\
 &= 10.5 \text{ days}
 \end{aligned}$$

Figure 10-6 Frequency polygon/line graph. (From Watzlaf VJM, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989. Reprinted with permission from American Health Information Management Association.)

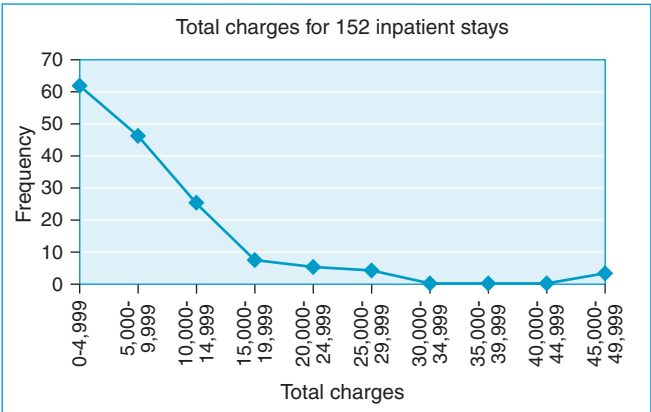


Figure 10-7 Frequency polygon comparing two types of data. (From Infectious Disease Epidemiology Report, Pennsylvania Department of Health, Bureau of Epidemiology and Disease Prevention, 1990.)

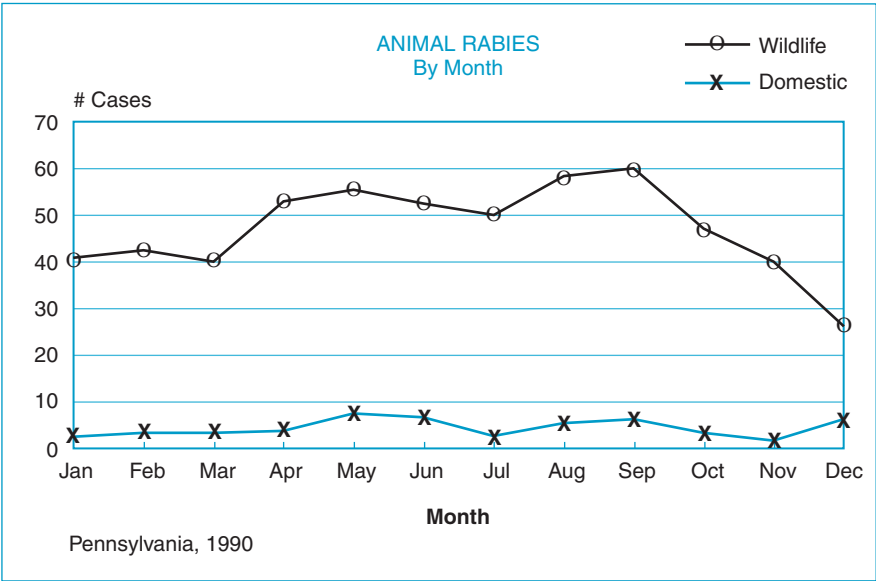
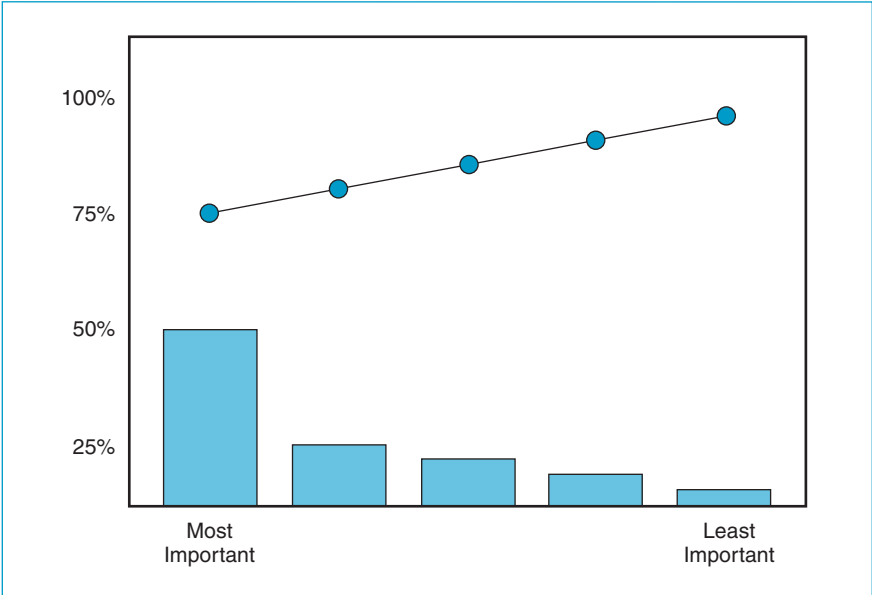


Figure 10-8 Pareto diagram.



Weighted Mean:

Department	Number of Patients	Average Length of Stay (days)
Internal medicine	40	6
General surgery	20	4
Pediatrics	5	2

If the average lengths of stay are 6, 4, and 2 days for three departments that have seen 40, 20, and 5 patients, respectively, the weighted mean, or average, length of stay is as follows:

$$\begin{aligned}
 &= \frac{(40 \times 6) + (20 \times 4) + (5 \times 2)}{40 + 20 + 5} \\
 &= \frac{330}{65} \\
 &= 5.1 \text{ days}
 \end{aligned}$$

subdivisions of the sample. In this situation, a **weighted mean** can be calculated, as illustrated in the example that follows.

Median

The **median** represents the middle value within a data set. When the values are arranged from lowest to highest, the number of values above the median is equal to the number of values below the median. For an odd number of observations, the median is the middle number in the ordered set of numbers; for an even number of observations, it is the mean of the middle two numbers.

The median is the most appropriate statistic to use for describing ordinal or ranked data because it allows for more meaningful descriptions of the data (e.g., the median response was between “strongly agree” and “agree”). It may also be useful to report the median for interval or ratio data when the data set contains extreme values, in other words, values that are much higher or much lower than the majority of other values. Although the mean can be strongly affected by extreme values, the median is not affected. Therefore, the median may provide a better representation of central tendency than the mean in some situations.

Example of Median (Odd Number)

Data: 1, 8, 6, 4, 2, 5, 9

Data after ordering: 1, 2, 4, 5, 6, 8, 9

Median = 5

Example of Median (Even Number)

Data: 16, 4, 21, 100, 7, 1

Data after ordering: 1, 4, 7, 16, 21, 100 =

$$\frac{7 + 16}{2} \text{ (two middle numbers)}$$

Median = 11.5

Mode

The **mode** is the value that occurs most frequently in a given set of values. Some distributions do not have a mode, whereas others may have two modes (bimodal). The mode is the only measure of central tendency that can be used with nominal data. In Table 10-9, the modal health insurance is Medicare because it is the one that occurs most often.

Measures of Dispersion

When interval or ratio data in a sample are summarized, determining the amount of dispersion or variability within the data set is important. **Dispersion** or variability refers to the extent to which scores within a set vary from each other. Measures of dispersion describe the extent to which scores in a set are spread out (or clustered together) around the mean.

Range

The **range** is one way to measure dispersion because it is the difference between the highest and lowest values. The major disadvantage of the range is that it is concerned only with the most extreme values and ignores all other values. When the range is reported, the highest and the lowest values should be included as well as the difference between them. The following statistics show length of stay for patients with pneumonia and how the range is computed by taking the highest length of stay minus the lowest length of stay for both community-acquired and nosocomial pneumonia.

Example of Range

Length of Stay: Patients with Pneumonia

Community-Acquired		Nosocomial	
Medical Record Number	Length of Stay (days)	Medical Record Number	Length of Stay (days)
207658	20	123579	15
214592	10	275816	22
221459	7	254137	18
158645	14	321096	10
129876	8	153992	8
Mean	11.8	Mean	14.6

The range for the length of stay for patients with community-acquired pneumonia was as follows:

$$20 \text{ (highest value)} - 7 \text{ (lowest value)} = 13$$

The range for the length of stay for patients with nosocomial pneumonia was as follows:

$$22 \text{ (highest value)} - 8 \text{ (lowest value)} = 14$$

Variance and Standard Deviation

The variance and standard deviation demonstrate how values are spread around the mean. The calculation of these measures is based on deviations (differences) between the

value of each score and the value of the mean. The **variance**, or s^2 , is computed by squaring each deviation from the mean, summing the deviations, and then dividing their sum by 1 less than n , the sample size.⁹ The **standard deviation**, represented by the symbol s , is the square root of the variance. The reason for taking the square root is to express dispersion in terms of the same units as values of the variable. (If height is measured in inches, the variance of height would express dispersion in units of square inches, but the standard deviation would express dispersion in units of inches). Because of this property, the standard deviation is the most commonly reported measure of dispersion.

Example of Computation of Variance and Standard Deviation of Length of Stay of Patients with Pneumonia

Community-Acquired Pneumonia Variance:

$$s^2 = \frac{(20 - 11.8)^2 + (10 - 11.8)^2 + (7 - 11.8)^2 + (14 - 11.8)^2 + (8 - 11.8)^2}{5 - 1}$$

$$= \frac{67.24 + 3.24 + 23.04 + 4.84 + 14.44}{4}$$

$$= \frac{112.8}{4} = 28.2$$

Standard Deviation:

$$s = \sqrt{28.2} = 5.3$$

Nosocomial Pneumonia Variance:

$$s^2 = \frac{(15 - 14.6)^2 + (22 - 14.6)^2 + (18 - 14.6)^2 + (10 - 14.6)^2 + (8 - 14.6)^2}{5 - 1}$$

$$= \frac{0.16 + 54.76 + 11.56 + 21.16 + 43.56}{4}$$

$$= \frac{131.2}{4} = 32.8$$

Standard Deviation:

$$s = \sqrt{32.8} = 5.7$$

The greater the deviations of the values from the mean, the greater the variance. Therefore compare the variance of the length of stay of the patients with community-acquired pneumonia (28.2) and the variance of the length of stay of the patients with nosocomial pneumonia (32.8). This shows that there is greater deviation from the mean of length of stay values in the nosocomial pneumonia group than in the community-acquired pneumonia group.

The standard deviation for length of stay for patients with community-acquired pneumonia was 5.3. This means that, on average, observed length of stay values fall 5.3 units from the mean. The standard deviation for the nosocomial pneumonia group was 5.7, which means that, on average, observed length of stay values fall 5.7 units from the mean. Therefore, there is greater variation or dispersion in the length of stay for patients with nosocomial pneumonia than in the length of stay for patients with community-acquired pneumonia. The health care facility may want to examine this further to determine whether there are more outliers in the nosocomial group and to examine the outlier cases in more detail.

Coefficient of Variation

When two samples or groups have very different means, the direct comparison of their standard deviations could be misleading. Therefore, when two groups have very different means, it is best to compare their standard deviations expressed as percentages of the mean. The coefficient of variation (CV) is used to do this. The coefficient of variation can also be used to compare dispersion in variables that are measured in different units.

$$CV = \frac{s}{|\bar{x}|} \times 100$$

where

s = standard deviation

$|\bar{x}|$ = absolute value of mean

Examples of Coefficient of Variation for Community-Acquired and Nosocomial Pneumonia

Coefficient of variation for community-acquired pneumonia:

$$= \frac{5.3}{|11.8|} \times 100 = 45\%$$

Coefficient of variation for nosocomial pneumonia:

$$= \frac{5.7}{|14.6|} \times 100 = 39\%$$

The CV was computed because it was important to determine whether the variation of length of stay values in the community-acquired pneumonia group was greater or less than the variation of length of stay values in the nosocomial pneumonia group. By using the CV, the variation can be computed exactly. The CV for the length of stay of the

patients with community-acquired pneumonia was 45%, and the CV for the length of stay of the patients with nosocomial pneumonia was 39%. This means that the variation in the length of stay of the patients with community-acquired pneumonia was somewhat greater than that of the patients with nosocomial pneumonia.

SELF-ASSESSMENT

Quiz

1. The most common measure of central tendency is the:
 - a. Mean
 - b. Median
 - c. Mode
 - d. Variance
2. The most common measure of dispersion is called:
 - a. Variance
 - b. Standard deviation
 - c. Weighted mean
 - d. Range
3. If a coefficient of variation for cholesterol levels is computed to be 42% for males and 56% for females, this means that the variation in cholesterol levels for females is somewhat _____ than that for males.
 - a. Greater
 - b. Less
 - c. Different
 - d. The same

Inferential Statistics

In almost all research studies, researchers can collect data from only a sample of the population of the cases about which they are interested in making conclusions. For example, a researcher interested in finding out whether the average salaries of entry-level HIM professionals differ across four geographic regions of the United States could not possibly collect data from every entry-level HIM professional in the country. However, through the use of a family of statistical tools known as inferential statistics, it is possible to make inferences (or generalizations) about a population based on data collected from a sample. With inferential statistics, the generalizations are made that go beyond the particular sample from which data are collected. The domain of inferential statistics can be subdivided into two main areas: tests of significance and estimation of population parameters.

Tests of Significance

The purpose of **tests of significance** is to determine whether observed differences between groups or relationships between variables in the sample being studied are likely to be due to

chance or **sampling error** or whether they reflect true differences or relationships in the population of interest. The term *sampling error* refers to the principle that the characteristics of a sample are not identical to the characteristics of the population from which the sample is drawn. Even if the average salaries of entry-level HIM professionals were equal across regions in the population, the sample means would still differ.

Although there are a number of tests of significance, all are based on the same underlying logic and all involve a similar series of steps. The first step in carrying out a test of significance is to state the null and alternative hypotheses. A **hypothesis** is a claim or statement about a property or characteristic of a population. It indicates the nature of the difference or relationship between characteristics. Each hypothesis can be expressed in two forms: as a null hypothesis or as an alternative hypothesis. The **null hypothesis** states that there is no difference or relationship in the population. If the null hypothesis is true, any differences or relationships that are observed in the sample are purely the result of chance. By contrast, the **alternative hypothesis** states that there is a true difference or relationship in the population. If the alternative hypothesis is true, the differences or relationships that are observed in the result are the true effect plus chance variation. In the example of the salaries of entry-level HIM professionals, here is how the null and alternative hypotheses could be stated:

Null hypothesis: There is no difference in the mean salaries of entry-level HIM professionals across geographic regions of the United States.

Alternative hypothesis: There is a difference in the mean salaries of entry-level HIM professionals across geographic regions of the United States.

The next step in carrying out a test of significance is computation of a statistic, sometimes referred to as the test statistic, which is based on the relevant data from the sample. The **test statistic** measures the size of the difference or relationship observed in the sample. In the example of the salaries of entry-level HIM professionals, the test statistic would reflect the size of the difference among sample means from the four geographic regions.

After the test statistic is computed, the probability that the observed value of the test statistic could occur in the event that the null hypothesis is true is determined. This probability is known as a ***p* value** and can range from 0 to 1. The *p* value provides an answer to the following question: How likely is it that we could observe a difference or relationship of this size in the sample if, in reality, the difference or relationship did not exist in the population? A second way of posing the same question is: How likely is it that the observed difference or relationship is due only to sampling error? A third way of phrasing the question is: What is the probability that the observed difference or relationship could occur through chance alone? The smaller the *p* value, or closer it is to 0, the less likely the null hypothesis is true and the smaller the probability that the observed difference or relationship could be due to chance or sampling error alone.

The availability of models of probability derived by statistical theorists makes it possible to determine the p value or probability associated with a given value of a test statistic. These models, known as theoretical sampling distributions, are discussed in greater depth in statistics textbooks. Commonly used theoretical sampling distributions include the normal, t , chi-square, and F distributions. Software programs that calculate test statistics determine the associated p values as well.

As an example of a model of probability that is familiar to you already, think about tossing a coin 10 times. It is well known that if a fair coin, that is, one that has not been tampered with, is tossed an infinite number of times, 50% of the tosses will result in a head and 50% of the tosses will result in a tail. Suppose you tossed a coin 10 times and every toss resulted in a head. You would suspect that the coin had been tampered with based on your knowledge of the model of probability. In fact, it can be shown that the probability of 10 heads when a fair coin is tossed 10 times is .001 or one in a thousand.

The last step in carrying out a test of significance is to decide whether the p value is small enough to support making the decision to reject the null hypothesis. How small is small enough? It is up to the researcher to choose the criterion or standard, which is known as the **level of significance**. The levels of significance that are most commonly chosen are .05 and .01. Setting the level of significance at .05 means that the decision will be made to reject the null hypothesis if the probability is smaller than 5 in 100 that the observed difference or relationship could be due to sampling error; setting it at .01 means the decision to reject will be made if the probability is smaller than 1 in 100. The recommended practice is for the researcher to choose the level of significance *before* carrying out the significance test.

When the p value is greater than the level of significance (usually either .05 or .01), the conclusion is stated as either “accept the null hypothesis” or “fail to reject the null hypothesis.” Whatever statement is used, this conclusion should not be interpreted as proof that the null hypothesis is true. It does mean that the sample evidence is not strong enough to warrant rejection of the null hypothesis.

When researchers make a decision to reject or accept the null hypothesis, can they be totally certain that they’ve made the correct decision? Unfortunately, the answer is no. Researchers make statistical decisions based on the statistical evidence available to them from the samples they study; they **infer** (but do not know) the true status of the null hypothesis in the population. The only way researchers could be certain about the actual status of the null hypothesis (in other words, whether there is a “real” effect) would be to collect data from the entire population; that avenue is almost never open to them. Therefore, there is always some degree of uncertainty whether statistical decisions are correct.

In introducing the concept of error in statistical decisions, many statistics textbooks make use of the analogy of a jury trial. The jury must make a decision about the innocence or guilt of the accused based on the evidence that is

presented during the trial. When the jury reaches a verdict, the jurors cannot be totally certain that their decision is correct, because they have no access to the absolute truth about what the accused did or did not do. It would require superhuman ability to know the absolute truth, because human judgment is fallible. If the jury reaches the decision either to convict a guilty person or acquit an innocent person, they have, of course, made a correct decision. If the jury reaches the decision to convict an innocent person or acquit a guilty person, they have made an incorrect decision or an error.

An important question is whether one kind of incorrect decision/error can be considered as “worse” than the other—in other words, whether it is more crucial to avoid or prevent one kind of incorrect decision than the other kind. Because the first kind of error is considered “worse” from the perspective of our legal system, safeguards against this error have been built into the system. The accused is presumed to be “innocent until proven guilty,” and the “burden of proof” rests on the prosecution rather than the defense.

Turning back to statistics, it is considered worse or more serious to reject the null hypothesis when it is true than to accept the null hypothesis when it is false. Therefore, the first type of error is known as *Type I error* and is denoted by the Greek letter alpha (α); the second type is known as *Type II error* and is denoted by the Greek letter beta (β). Safeguards are built into the system of hypothesis testing to control the probability of Type I error. The level of significance discussed earlier is equivalent to the probability of Type I error that the researcher is willing to tolerate.

Type I error: Reject the null hypothesis when it is true, alpha (α)

Type II error: Accept the null hypothesis when it is false, beta (β)

Whereas the probability of Type I error, α , is set by the researcher, the probability of Type II error, β , depends on several factors. One of these factors, the size of the sample, is within the researcher’s control. The larger the sample size, the smaller the probability of Type II error. Another important influence on the probability of Type II error is the size of the true difference or relationship in the population. This factor is not directly within the researcher’s control. The larger the difference or relationship in the population, the smaller the probability of Type II error.

Many different tests of significance are available to researchers. The choice of which test of significance to apply in a particular situation depends primarily on three factors:

1. What is the nature of the hypothesis? Does the hypothesis involve differences between groups, relationships between variables, or prediction?
2. What is the design of the study? How many groups are involved? Are the groups independent or matched on certain characteristics such as age or gender? Are data collected only at one time point or at two or more time points?

3. Which type of data (nominal, ordinal, or continuous) has been collected to measure each of the variables being studied?

Choosing the appropriate statistical method to test a particular hypothesis requires considerable knowledge of statistics. Therefore, it is recommended that HIM professionals consult a statistician as part of the process of planning a research study. The following sections describe several of the most commonly used tests of significance.

Independent Samples *t* Test

Researchers are often interested in finding out whether a significant difference exists between two or more groups with respect to a quantitative variable. Examples of questions of this type include: Is the length of stay shorter with a new protocol compared with an old protocol? Are health care costs different across regions? Is patient satisfaction greater in teaching hospitals or nonteaching hospitals? When there are two groups and the groups are independent (not matched), a significance test known as the independent samples *t* test is applied. This test examines the differences between the means of two groups and determines whether the difference is large enough to justify rejection of the null hypothesis. The decision of whether to reject the null hypothesis is based on the *p* value associated with a test statistic known as a *t* value.

The first step in performing an independent samples *t* test is to compute the mean and standard deviation for each group. Next, the standard deviations for the two groups are averaged, or pooled. Then a value of *t* is computed. Finally, it is determined whether the *p* value associated with the *t* value is smaller than the level of significance.

Pooled (averaged) standard deviation:

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

where

n_1 = number of cases within group 1

n_2 = number of cases within group 2

s_1^2 = variance of group 1

s_2^2 = variance of group 2

t value:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

\bar{x}_1 = mean of group 1

\bar{x}_2 = mean of group 2

n_1 = number of cases within group 1

n_2 = number of cases within group 2

s_p = pooled standard deviation

The data for the following example of an independent samples *t* test comes from a hypothetical data set (Chicago, Illinois, SPSS, 2008) about the benefits of emotional support for stroke patients suffering from depression. Female patients at a hospital who were diagnosed with depression following their stroke were randomly assigned to receive either physical therapy alone or both physical therapy and emotional support. Three months posttreatment, the patients' ability to carry out three common activities of daily life was evaluated. Scores could range from 0 to 12, with higher scores indicating greater impairment. The data are summarized below.

Example of Independent Samples *t* Test

Group	N	Mean	Standard Deviation
Control (PT only)	15	8.33	1.72
Experimental (PT + emotional support)	19	6.21	2.10

$$\begin{aligned}
 sp &= \sqrt{\frac{(15-1)(2.96) + (19-1)(4.41)}{(15-1) + (19-1)}} \\
 &= \sqrt{\frac{41.42 + 79.38}{14 + 18}} \\
 &= \sqrt{\frac{120.80}{32}} = 1.94 \\
 t &= \frac{8.33 - 6.21}{1.94 \sqrt{\frac{1}{15} + \frac{1}{19}}} \\
 &= \frac{2.12}{1.94 \sqrt{.067 + .053}} \\
 &= \frac{2.12}{(1.94)(.346)} = 3.16
 \end{aligned}$$

In this example, the *t* value is computed as 3.16, and the associated *p* value is found to be .003. At a significance level of .05, the null hypothesis is rejected. There is evidence that combining emotional support with physical therapy is beneficial to female stroke patients suffering from depression.

One-Way Analysis of Variance

When there are three or more groups, a significance test known as one-way analysis of variance (ANOVA) is applied to test for significant differences among group means. In this test, variability among subjects' scores is analyzed by

dividing it into two components: variability between groups as reflected in differences among the group means and variability within groups as reflected in differences among the scores of subjects who belong to the same group. The logical principle underlying ANOVA is that if true differences among group means exist, then between-group variability must be greater than within-group variability. As an application of this logical principle, the procedure for carrying out a one-way ANOVA involves three main steps.

Step 1 is to quantify the amount of between-groups variability—in other words, to evaluate how different the group means are from each other. Step 1 entails three substeps. In the first substep, a measure known as sum of squares between groups (SSB) is computed.

$$SSB = \sum n_j (\bar{X}_j - \bar{X})^2$$

where

Σ , the Greek letter sigma, stands for “the sum of”

n_j is the number of cases within the “jth” group

\bar{X}_j is the mean of the “jth” group, and

\bar{X} is the grand mean (the mean for the total sample)

In the second substep, the between-groups degrees of freedom (dfb) are computed.

$$dfb = J - 1$$

where

J = the number of groups

In the third substep, the mean square between groups (MSB) is computed:

$$MSB = \frac{SSB}{dfb}$$

The second main step is to quantify the amount of within-groups variability—in other words, to evaluate the size of differences among the scores of subjects that belong to the same group. The substeps of Step 2 parallel the substeps of Step 1. In the first substep, the sum of squares within groups (SSW) is computed.

$$SSW = \sum (n_j - 1) s_j^2$$

where

n_j is the size of the “jth” group and

s_j^2 is the variance of the “jth” group.

In the second substep, the within-groups degrees of freedom (dfw) is computed.

$$dfw = N - J$$

where

N = the total sample size

J = the number of groups

In the third substep, the mean square within groups (MSW) is computed. Like the pooled variance that is computed when carrying out an independent samples t test, the MSW represents the average amount of within-group variance. Notice the similarity between the formula for MSW and the formula for the pooled standard deviation.

$$MSW = \frac{SSW}{dfw}$$

The third main step is to compute the F ratio, the ratio of between-group variability to within-group variability. To reject the null hypothesis, the F ratio must be large enough; in other words, between-group variability must be sufficiently greater than within-group variability. The larger the F ratio, the smaller the associated p value will be.

$$F \text{ ratio} = \frac{MSB}{MSW}$$

The following example analyzes infant mortality data by state for 2005 made available by the National Center for Health Statistics. A researcher is interested in learning whether infant mortality rates differ significantly by geographical region. The 50 states (and the District of Columbia) have been grouped into four regions according to the grouping used by the U.S. Census Bureau. Summary data is shown below.

Region	n	Mean	Standard Deviation
Northeast states	9	6.04	0.78
Midwest states	12	6.84	1.07
South states	18	8.50	1.93
West states	12	6.11	0.73

Example of ANOVA Computation

The following example analyzes infant mortality data by state for the year 2005 made available by the National Center for Health Statistics. A researcher is interested in learning whether infant mortality rates differ significantly by geographical region. The 51 states (including the District of Columbia) have been grouped into four regions according to the grouping used by the US Census Bureau. Summary data is shown below.

Region	n	Mean	Standard Deviation
Northeast	9 states	6.04	0.78
Midwest	12 states	6.84	1.07
South	18 states	8.50	1.93
West	12 states	6.11	0.73

To compute the grand mean, we make use of the formula for a weighted mean provided earlier in the chapter.

$$\begin{aligned}\bar{X}(\text{grand mean}) &= \frac{(9)(6.04) + 12(6.83) + (18)(8.50) + (12)(6.11)}{9 + 12 + 18 + 12} \\ &= \frac{54.36 + 81.96 + 153.00 + 73.32}{51} \\ &= \frac{362.64}{51} = 7.11\end{aligned}$$

After computing the grand mean, we are ready to compute SSB and MSB.

$$\begin{aligned}\text{SSB} &= (9)(6.04 - 7.11)^2 + 12(6.83 - 7.11)^2 + \\ &18(8.50 - 7.11)^2 + 12(6.11 - 7.11)^2 \\ &= (9)(-1.07)^2 + (12)(-.28)^2 + (18)(1.39)^2 + 12(-1.00)^2 \\ &= (9)(1.15) + (12)(.08) + (18)(1.93) + (12)(1.00) \\ &= 58.05\end{aligned}$$

$$\text{dfb} = 4 - 1 = 3$$

$$\text{MSB} = = 19.35$$

Next, we need to compute SSW and MSW.

$$\begin{aligned}\text{SSW} &= (9-1)(.78)^2 + (12-1)(1.07)^2 + \\ &(18-1)(1.93)^2 + (12-1)(.73)^2 \\ &= (8)(.61) + (11)(1.14) + (17)(3.72) + (11)(.53) \\ &= 4.87 + 12.59 + 63.32 + 5.86 = 86.65\end{aligned}$$

$$\text{dfw} = 51 - 4 = 47$$

$$\text{MSW} = = 1.84$$

Finally, we can compute the F ratio.

$$F = \frac{19.35}{1.84} = 10.51$$

The p value associated with an F ratio of 10.51 when there are 3 between degrees of freedom and 47 within degrees of freedom is less than .001. Since the p value is less than .05, we reject the null hypothesis and conclude that infant mortality rate does vary across geographic regions.

The p value associated with an F ratio of 10.51 when there are three between-group degrees of freedom and 47 within-group degrees of freedom is less than .001. Because the p value is less than .05, we reject the null hypothesis and conclude that infant mortality rate does vary across geographic regions.

Pearson Correlation Coefficient

Many questions investigated in research studies are concerned with the relationships between variables. Different tests of significance are used to evaluate relationships depending on the type of data involved.

A statistic known as the Pearson correlation coefficient is used to assess the direction and degree of relationship between two continuous variables. The direction of the relationship between two continuous variables can be either positive or negative. In the following explanation of positive and negative relationships, the first variable is designated as X, and the second variable is designated as Y. When the relationship between X and Y is positive, high scores on X tend to be associated with high scores on Y; as

X increases, Y increases. When the relationship between X and Y is negative, high scores on X tend to be associated with low scores on Y; as X increases, Y tends to decrease. The relationship between HIM professionals' length of experience in the field and their current salaries would be expected to be positive; professionals with more experience would be expected to receive higher salaries. On the other hand, the relationship between the amount of stress experienced on the job and job satisfaction would be expected to be negative. As job-related stress increases, job satisfaction would be expected to decrease.

Values of the Pearson correlation coefficient for positive relationships can range from 0 to +1; values for negative relationships can range from 0 to -1. The closer the value is to +1 for positive relationships or to -1 for negative relationships, the stronger the relationship. The closer the value is to 0, the weaker the relationship. General guidelines for interpreting Pearson correlation coefficients suggest that values less than 0.30 may be considered to indicate weak relationships, values between 0.30 and 0.59 may be considered to indicate moderate relationships, and values of 0.6 or higher may be considered to indicate strong relationships. Because the procedure for calculating the Pearson correlation coefficient by hand is quite time-consuming, computation is almost always done by computer. Readers interested in learning about the computational procedure are encouraged to consult an elementary statistics textbook.

When the Pearson correlation coefficient is computed, a related significance test is performed. This test determines the probability that the observed value of the correlation coefficient could occur through sampling error alone. If the p value is smaller than the predetermined level of significance, there is evidence that a true relationship exists in the population. At this point, a word of caution is in order: a statistically significant relationship is not necessarily a strong relationship. Sample size has a great influence on the outcome of the test for the significance of a correlation coefficient. When sample size is very large, a very weak relationship can be significant. Therefore, in interpreting results it is essential to consider the value of the correlation coefficient as well as the p value.

As an example, we'll use data from a hypothetical data set (SPSS, 2008) concerning patients admitted to a hospital because of a suspected heart attack. A researcher was interested in learning if there was a relationship between patients' ages and their length of stay. The researcher found that for a sample of 154 patients the value of the Pearson correlation coefficient was .344 and the associated p value was less than .001. On the basis of these findings, the researcher concluded that there was a moderate positive relationship between patients' ages and their length of stay; as age increased, length of stay tended to increase.

Regression Analysis

Regression is a statistical method closely linked to correlation that has many useful applications in HIM. The purpose of regression analysis is to learn to what extent one or more

explanatory variables can predict an outcome variable. In the context of regression analysis predictor variables are denoted by X and outcome variables are denoted by Y . The single most important result of regression analysis is a statistic known as R -squared (also written as R^2), which can range from 0 to 1. R -squared represents the squared correlation between the explanatory variable(s) and the outcome variable. Although correlation coefficients can be negative, R -squared can never be negative because squaring a negative number generates a positive number. The value of R -squared indicates the proportion of variability in the outcome that is explained by the predictor variable(s). The closer the value of R -squared is to 1, the better or stronger the prediction. The p value associated with R -squared indicates the probability that the observed value of R -squared could occur through sampling error alone. When there is only one explanatory variable, R -squared may be computed by simply squaring the value of the Pearson correlation coefficient.

Regression analysis produces another important result known as a regression equation. The regression equation is a formula for calculating a case's predicted score on the outcome variable based on that case's score(s) on the predictor variable(s). Regression equations can be useful in making decisions in situations in which data have been collected on the predictor variable(s) but data on the outcome variable are not available. When there is only one explanatory variable, the regression equation takes the form of the formula for a straight line. As readers may recall from high school algebra, a straight line is determined by only two numerical values, the value of its slope and the value of its intercept. The slope is defined as the amount of change in Y per unit change in X and the intercept is defined as the value of Y when X is equal to 0. The general form of a regression equation when there is only one predictor is shown in the following formula:

$$\hat{y} = bx + c$$

where

\hat{y} = the predicted value of y

b = the slope

c = the intercept

The values of the slope and the intercept for a specific regression equation can be computed according to the following formulas:

$$b = r \frac{s_x}{s_y}$$

$$c = \bar{y} - b\bar{x}$$

r = value of the Pearson correlation coefficient

s_x = standard deviation of x

s_y = standard deviation of y

As an example of regression analysis, we return to the hypothetical data on patients admitted to a hospital because of a suspected heart attack. Suppose a researcher wanted to

learn if patients' length of stay predicted the cost of their treatment. The researcher collected the following data.

Variable	Mean	Standard Deviation
Length of stay in days (X)	5.45	1.39
Cost of treatment (Y)	\$36,028	\$8,660
Pearson correlation (r) = .72		

Substituting this data into the formulas for the slope and intercept, the researcher finds that:

$$b = \frac{(.72)8,660}{1.39}$$

$$b = \frac{(.72)8,660}{1.39}$$

$$= (.72)(6230.22)$$

$$= 4,485$$

$$c = 36,028 - (4485)(5.45)$$

$$= 36,028 - 24,309 = 11.72$$

Therefore, the prediction equation is:

$$\hat{y} = (4485)(x) + 11.72$$

Making use of the prediction equation, we would compute the predicted treatment cost for a patient who stayed 7 days as follows:

$$\hat{y} = (4485)(7) + 11.72$$

$$= 31,395 + 11.72$$

$$= 31,406.72$$

Chi-Square Test

The significance tests described up to this point are appropriate only for quantitative data. Many variables of interest to HIM professionals are nominal or qualitative rather than quantitative. The chi-square test is one of the most commonly used tests of significance that is appropriate for qualitative data. It can be applied to assess the degree of relationship or association between two qualitative variables or to determine whether there are differences between two or more groups with respect to a qualitative variable. The data used in computing a chi-square test are frequently displayed in the form of a **contingency table**, a table that displays the joint frequencies of the two variables. Table 10-13 displays

Table 10-13 FREQUENCIES FOR SUCCESS AND FAILURE OF SURGICAL PROCEDURE

Smoking Status	Surgical Outcomes		Total
	Success	Failure	
Nonsmoker	9 (75.0%)	3 (25.0%)	12
Smoker	4 (28.6%)	10 (71.4%)	14
Total	13	13	26

hypothetical data on the joint frequencies of smoking status and outcome of an experimental surgical procedure, coded as success or failure.

Certain terms are commonly used in referring to the parts of a contingency table. The variable for which categories define the rows is known as the row variable, and the variable for which categories define the columns is known as the column variable. In Table 10-13, therefore, smoking status is the row variable and surgical outcome is the column variable. The sum or total of frequencies within a row is known as the row total, and the sum or total of frequencies within a column is known as the column total. The sum or total of frequencies across all rows (or all columns) is known as the grand total. The grand total is equal to the total number of subjects or cases. The “square” within a contingency table that is formed by pairing a single category of the row variable with a single category of the column variable is known as a cell. Table 10-13, therefore, has four cells: non-smoker, success; non-smoker, failure; smoker, success; and smoker, failure.

From the data in Table 10-13, it can be seen that the overall success rate was 50% (13 of 26 patients). However, for non-smokers, the success rate was 75% (9 of 12 patients), whereas for smokers the success rate was 28.6% (4 of 14 patients). Is this difference significant? To answer this question, the chi-square test evaluates differences between observed and expected cell frequencies. Observed cell frequencies are the frequencies actually observed; expected frequencies are the frequencies that would be expected to occur if there was no association between the two variables. Subscripts are commonly used to identify cells within a contingency table. The first of the two numbers forming the subscript identifies the row and the second number identifies the column. For example, O_{11} stands for the observed frequency within the cell located in the first row, first column of the table; O_{12} stands for the observed frequency within the cell located in the first row, second column, and so forth.

The formula below is applied to compute expected frequencies for cells within a contingency table, and the example demonstrates the computation of expected cell frequencies for Table 10-13.

$$E_{ij} = \frac{(\text{Row total})(\text{Column total})}{\text{Grand total}}$$

Computation of expected frequencies:

$$E_{11} = \frac{(12)(13)}{(26)} = 6$$

$$E_{12} = \frac{(12)(13)}{(26)} = 6$$

$$E_{21} = \frac{(14)(13)}{(26)} = 7$$

$$E_{22} = \frac{(14)(13)}{(26)} = 7$$

The expected cell frequencies indicate that if the null hypothesis of no association between smoking status and surgical outcome were true, the success rates for nonsmokers and smokers would be equal to the overall success rate. In other words, the outcome of success would be expected for six nonsmokers (50% of the 12 nonsmokers) and seven smokers (50% of the 14 smokers). Likewise, the outcome of failure would be expected for six nonsmokers and seven smokers. After the expected cell frequencies have been computed, the chi-square statistic is then computed by applying the following formula:

$$\text{Chi-square} = \sum \frac{(O - E)^2}{E}$$

where

O = Observed cell frequency

E = Expected cell frequency

Computation of the chi-square statistic based on the observed cell frequencies shown in Table 10-13 and the associated expected frequencies is illustrated below.

$$\begin{aligned} \text{Chi-square} &= \frac{(9-6)^2}{6} + \frac{(3-6)^2}{6} + \frac{(4-7)^2}{7} + \frac{(10-7)^2}{7} \\ &= 1.50 + 1.50 + 1.29 + 1.29 \\ &= 5.58 \end{aligned}$$

The p value associated with the chi-square value of 5.58 is .018. Therefore, the null hypothesis is rejected at the .05 level of significance, and it is concluded that surgical outcome is associated with smoking status.

Interval Estimation

Significance tests constitute an important and useful set of data analysis techniques. However, they can be appropriately applied only in situations in which there are hypotheses to be tested. In some situations, the researcher's primary interest is in making use of data obtained from a sample to estimate the characteristics of a population, for example, in using the value of a sample mean to estimate the value of the population mean. It would be an extremely rare occurrence for the mean of a single sample to be exactly equal to the population mean. Furthermore, the means of multiple samples drawn from the same population would not all be the same but rather would vary from one sample to another. However, statistical theory has proved that when all possible samples of the same size are drawn from a given population and a mean is computed for each sample, the average of the sample means is equal to the population mean. Furthermore, statistical theory has demonstrated that as sample size increases, the average difference between the sample mean and the population mean decreases. In the long run, means of large samples will be closer to the population mean than means of small samples.

The application of statistical theory makes it possible to construct a confidence interval for the population mean based on the value of the sample mean. A researcher chooses a level of confidence, which is usually 95% but sometimes

90% or 99%. Like the choice of a level of significance, the choice of a level of confidence depends on the maximum probability of error a researcher is willing to tolerate. The probability of error is 100% minus the level of confidence. The interpretation of a 95% confidence interval is that there is a 95% probability that the population mean falls between the lower and upper limits of the interval (and consequently a 5% probability that the population mean does not fall within the interval).

The steps in constructing a confidence interval for the population mean include computing the mean and standard deviation of the sample and looking up the appropriate critical value (based on the level of confidence and sample size) in a table of the t distribution. The following formula is applied to compute the lower and upper limits of a confidence interval for the population mean:

$$\text{Lower limit} = \bar{X} - \left| \frac{(t_{\text{critical}})s}{\sqrt{n}} \right|$$

$$\text{Upper limit} = \bar{X} + \left| \frac{(t_{\text{critical}})s}{\sqrt{n}} \right|$$

The following example illustrates computation of a 95% confidence interval for the population mean. In the data used in the example of regression, the mean length of stay for 154 patients admitted to a hospital because of a suspected heart attack was 5.45 days, and the standard deviation was 1.39 days.

Example Computation of Confidence Interval

$$\bar{X} = 5.45$$

$$s = 1.39$$

$$n = 154$$

$$t_{\text{critical}} = 1.98$$

$$\text{Lower limit} = 5.45 - \left[(1.98) \frac{1.39}{\sqrt{154}} \right]$$

$$= 5.45 - \left[(1.98) \frac{1.39}{12.41} \right]$$

$$= 5.45 - .22$$

$$= 5.23$$

$$\text{Upper limit} = 5.45 + \left[(1.98) \frac{1.39}{\sqrt{154}} \right]$$

$$= 5.45 + \left[(1.98) \frac{1.39}{12.41} \right]$$

$$= 5.45 + .22$$

$$= 5.67$$

There is a 95% probability that the mean length of stay in the population for patients admitted because of a suspected heart attack is between 5.23 and 5.67. Since the sample size is fairly large, the distance between the lower and upper limits is small.

SELF-ASSESSMENT

Quiz

1. This hypothesis states that there is no difference or relationship in the population and the observations are the result of chance. It is called the:
 - a. Null hypothesis
 - b. Alternative hypothesis
 - c. Major hypothesis
 - d. Scientific hypothesis
2. Once a test statistic is computed, the probability that the observed value of the test statistic could occur in the event that the null hypothesis is true is determined. This probability is known as the _____ and can range from ____ to _____.
 - a. t value; 1 to 5
 - b. F statistic; 0 to infinity
 - c. p value; 0 to 1
 - d. chi-square value; 0 to 1
3. This inferential statistic is performed to learn to what extent one or more independent variables predict a dependent variable. It is called:
 - a. Correlation
 - b. Regression
 - c. Interval estimation
 - d. Analysis of variance

Sampling and Sample Size

Because most populations under study are fairly large, researchers usually choose to study samples of those populations. Results obtained by studying a sample can be generalized to the population from which the sample is drawn as long as the sample is representative of the population. When a sample is representative, the characteristics of the sample do not differ from the characteristics of the population in any systematic or consistent way. For example, if a researcher wished to select a sample of hospitals within a certain state and selected only urban hospitals, the sample would not be representative of the population of all hospitals in the state. The best way to ensure that a sample is representative is to apply random sampling. A random sample means that:

- every member of the population has the same chance of being included in the sample and
- the selection of one member has no effect on selection of another member-independent selection.¹⁰

Types of Random Sampling

In the current computer age, simple random sampling is usually carried out with randomization programs. Such programs are available either through statistical software packages or through stand-alone procedures available through the Internet. Simple random sampling can also be conducted with a

table of random numbers. For example, if the population consists of 50 patient records and you would like to obtain a random sample of 20 of those, you can assign each patient a number or use an existing medical record number. Then refer to a table of random numbers, randomly pick a page to start on, and go up, down, or across, looking at the first two digits. If the first two digits are in the number assigned to the patient record, then that patient record should be included in your sample. Continue this process until your sample of 20 is met.

A **stratified random sample** is obtained by dividing a population into groups or strata and taking random samples from each stratum. This is done to ensure that the sample is representative of the population. For example, when studying a group of college students that was 60% female and 40% male, the population would be divided into groups by gender, and the selected sample would reflect the same ratio as to total population. Another example is when studying coding accuracy, medical records could be grouped by the most common principal diagnoses and selected to reflect the percentage as would be found in the whole population of records coded.

Because both simple and the stratified random sampling are time-consuming without the aid of computer software, researchers have sometimes chosen another sampling method known as systematic sampling. With this method, the researcher must first decide what fraction or proportion of the population is to be sampled. If a researcher decided to sample one tenth of the population, the researcher would first randomly choose a starting point on a list of the population and then select every tenth record beginning with the designated starting point. Strictly speaking, systematic sampling can only be considered a random sampling method if the list itself is in random order.

Determining Sample Size

How do you know how many records or patients or subjects to include in the sample? This question is pondered by people completing performance improvement studies, epidemiologic research studies, or studies of any kind in which a sample is being taken. There is no easy answer to this question. Different approaches to sample size determination are taken depending on whether the researcher's purpose is interval estimation or hypothesis testing.

When the researcher's purpose is interval estimation, an important factor in sample size determination is the amount of error the researcher is willing to accept. Another way of thinking of the amount of error is in terms of how accurately or precisely the researcher wants to estimate the relevant population parameter, such as the population mean or proportion. The less error there is, the more accurate or precise the estimate is. The size of the sample increases the smaller the desired level of error and greater precision.

To illustrate this concept, suppose that a researcher wanted to estimate a population proportion. For example, a health care researcher wants to conduct a survey to determine the proportion of patients who were satisfied with the health care services they received. The population for this facility includes

1000 patients who were discharged within the past year, and interviewing all of them would take an unreasonably long time. Therefore, because no prior information is available to estimate p (the population proportion), p is set to be 0.5. (This is the best guess when no information is available.) Also, the researcher decides that an acceptable amount of error is 0.05. This means that the researcher wants the sample proportion to differ from the population proportion by no more than 0.05. N is the population size and n is the sample size. Therefore, the following formula can be used:

$$n = \frac{Npq}{(N - 1)D + pq}$$

where

p = estimated population proportion; for this example = 0.5

$q = 1 - p = 1 - 0.5$

$q = 0.5$

B = acceptable amount of error

N = the size of the population

n = sample size

$$D = \frac{B^2}{4}$$

$$= \frac{(0.05)^2}{4}$$

$$D = 0.000625$$

$$q = 1 - p = 1 - 0.5$$

$$q = 0.5$$

If we insert the information from the example, the formula is as follows:

$$n = \frac{(1000)(0.5)(0.5)}{(999)(0.000625) + (0.5)(0.5)}$$

$$= \frac{250}{0.874375}$$

$$= 285.9 \text{ or } 286$$

Therefore, for this example, the total sample needed in which only 5% error would be due to sampling variability is 286. Many other methods are used to estimate the appropriate sample size, depending on the sampling method chosen (simple random, stratified random, systematic). It is highly recommended that sample size selection be researched in more detail by using sampling books. *Elementary Survey Sampling* is an excellent book that clearly describes sample size and methods of selection.¹⁰

If the researcher's purpose is hypothesis testing, sample size determination is closely related to the concept of **power**, which is defined as the probability of correctly rejecting a false null hypothesis. Power is equal to 1 minus the probability of Type II error. The three factors that determine power are alpha (the level of significance set by the researcher), sample size, and effect size (the size of the difference between means or the strength of the relationship between variables in the population). Large effect sizes require smaller samples; small effect

sizes require larger samples. To find the appropriate sample size, it is necessary for researchers to arrive at an estimate of effect size. Effect size is defined differently, depending on the specific statistical method (i.e., *t* test, correlation, regression, etc.) to be used. Researchers often consult with statisticians regarding necessary sample size for hypothesis testing because of the number of factors that must be considered.

SUMMARY

Health information and informatics professionals are very involved in the use of health care statistics to evaluate a health care system, facility, research process, disease, or health outcome. The HIM professional should be knowledgeable about methods of data display, inferential statistics, hypothesis testing, and research methodology to fully evaluate health information and informatics topics of study.¹¹ HIM professionals who can plan studies and decide which statistics to apply to address particular research questions are essential for the field of health information management to move forward.

SELF-ASSESSMENT

Chapter Review


- Nosocomial infections are those infections:
 - occurring 72 hours after admission
 - occurring less than 72 hours before admission
 - occurring after surgery
 - both a and c
- If a *p* value of 0.001 was obtained, the researcher would most likely:
 - accept the null hypothesis
 - reject the null hypothesis
 - reject the alternative hypothesis
 - none of the above
- If a researcher accepts the null hypothesis when it is false, he has committed a Type II error.
 - True
 - False
- Which is true regarding comorbidities?
 - They are a preexisting condition.
 - They generally increase the length of stay.
 - They affect mortality and morbidity rates.
 - all of the above

- What was the median length of stay (LOS) for these seven psychiatric patients: 4, 11, 2, 1, 8, 22, 7 days?
 - 4 days
 - 7 days
 - 8 days
 - none of the above
- Referring to question 5, the range for these patients' length of stay was:
 - 7 days
 - 13 days
 - 21 days
 - none of the above

Referring to the table below, please answer questions 7–10:

- What was the mean LOS for patients having community-acquired viral pneumonia?
 - What was the mean LOS for patients with nosocomial viral pneumonia?
 - What was the median LOS for patients with community-acquired pneumonia?
 - What was the median LOS for patients with nosocomial pneumonia?
- Length of Stay: Patients with Pneumonia

Community-Acquired		Nosocomial	
Medical Record #	LOS	Medical Record#	LOS
207658	20	123579	15
214592	10	275816	22
221459	7	254137	18
158645	14	321096	10
129876	8	153992	8
Mean =	_____	Mean =	_____

Go to the Evolve site and complete the Chapter Review questions for this chapter. 

REFERENCES

- Kuzma J, Bohnenblust S: *Basic statistics for the health sciences*, ed 5, New York, 2005, McGraw-Hill.
- Skurka MF: Statistics. In *Health information management in hospitals*, Chicago, 1994, American Hospital Publishing, pp. 141-146.
- Hanken MA, Water K, editors: *Glossary of healthcare terms*, Chicago, 1994, American Health Information Management Association.
- U.S. Bureau of Census: International population reports. In *An aging world II*, Washington, DC, 1992, U.S. Government Printing Office, pp. 25, 92-93.

5. Lilienfeld D, Stolley PD: *Foundations of epidemiology*, ed 3, Oxford, 1994, Oxford University Press.
6. Slome C, Brogan D, Eyres S, et al: *Basic epidemiological methods and biostatistics: a workbook*, Belmont, CA, 1986, Jones & Bartlett.
7. Watzlaf VJM, Kuller LH, Ruben FL: The use of the medical record and financial data to examine the cost of infections in the elderly, *Topics Health Records Manage* 13:65-76, 1992.
8. Watzlaf V, Abdelhak M: Descriptive statistics, *J Am Med Record Assoc* 60:37-41, 1989.
9. Shott S: *Statistics for health professionals*, Philadelphia, 1990, WB Saunders.
10. Scheaffer R, Mendenhall W, Ott L: *Elementary survey sampling*, ed 3, Boston, 1986, Duxbury Press.
11. Layman E, Watzlaf V: *Health informatics research methods: principles and practice*, Chicago, 2009, American Health Information Association.