# BIOSTATISTICS / STATISTICS FOR HEALTH PROFESSIONALS

## Introduction

This module basically is designed to make us get exposed to statistics and understand statistics in depth including various types statistics and their importance's

Statistics is a set of tools used to organise and analyse data. Data must either be numeric in origin or transformed by researchers into numbers. For instance, statistics could be used to analyze percentage scores English students receive on a grammar test: the percentage scores ranging from 0 to 100 are already in numeric form. Statistics could also be used to analyze grades on an essay by assigning numeric values to the letter grades, e.g., A=4, B=3, C=2, D=1, and F=0.

Employing statistics serves two purposes,

(1) description and (2) prediction. Statistics are used to describe the characteristics of groups. These characteristics are referred to as variables. Data is gathered and recorded for each variable. Descriptive statistics can then be used to reveal the distribution of the data in each variable.

Statistics is also frequently used for purposes of prediction. Prediction is based on the concept of generalisability: if enough data is compiled about a particular context (e.g., students studying writing in a specific set of classrooms), the patterns revealed through analysis of the data collected about that context can be generalized (or predicted to occur in) similar contexts. The prediction of what will happen in a similar context is probabilistic. That is, the researcher is not certain that the same things will happen in other contexts; instead, the researcher can only reasonably expect that the same things will happen.

Prediction is a method employed by individuals throughout daily life. For instance, if writing students begin class every day for the first half of the semester with a five-minute free writing exercise, then they will likely come to class the first day of the second half of the semester prepared to again free write for the first five minutes of class. The students will have made a prediction about the class content based on their previous experiences in the class: Because they began all previous class sessions with free writing, it would be probable that their next class session will begin the same way. Statistics is used to perform the same function; the difference is that precise probabilities are determined in terms of the percentage chance that an outcome will occur, complete with a range of error. Prediction is a primary goal of inferential statistics.

## Descriptive Statistics

### *Revealing Patterns Using Descriptive Statistics*

Descriptive statistics, not surprisingly, "describe" data that have been collected. Commonly used descriptive statistics include frequency counts, ranges (high and low scores or values), means, modes, median scores, and standard deviations. Two concepts are essential to understanding descriptive statistics: *variables* and *distributions*. To read more about descriptive statistics, click on the items below:

- Variables

- Distributions

## Variables

Statistics are used to explore numerical data (Levin, 1991). Numerical data are observations which are recorded in the form of numbers (Runyon, 1976). Numbers are variable in nature, which means that quantities vary according to certain factors. For examples, when analyzing the grades on student essays, scores will vary for reasons such as the writing ability of the student, the students' knowledge of the subject, and so on. In statistics, these reasons are called variables. Variables are divided into three basic categories:

- Nominal Variables
- Ordinal Variables
- Interval Variables

## Distributions

A distribution is a graphic representation of data. The line formed by connecting data points is called a frequency distribution. This line may take many shapes. The single most important shape is that of the bell-shaped curve, which characterizes the distribution as "normal." A perfectly normal distribution is only a theoretical ideal. This ideal, however, is an essential ingredient in statistical decision-making (Levin, 1991). A perfectly normal distribution is a mathematical construct which carries with it certain mathematical properties helpful in describing the attributes of the distribution. Although frequency distribution based on actual data points seldom, if ever, completely matches a perfectly normal distribution, a frequency distribution often can approach such a normal curve.

The closer a frequency distribution resembles a normal curve, the more probable that the distribution maintains those same mathematical properties as the normal curve. This is an important factor in describing the characteristics of a frequency distribution. As a frequency distribution approaches a normal curve, generalizations about the data set from which the distribution was derived can be made with greater certainty. And it is this notion of generalisability upon which statistics is founded. It is important to remember that not all frequency distributions approach a normal curve. Some are skewed. When a frequency distribution is skewed, the characteristics inherent to a normal curve no longer apply.

## Inferential Statistics

### *Making Predictions Using Inferential Statistics*

Inferential statistics are used to draw conclusions and make predictions based on the descriptions of data. In this section, we explore inferential statistics by using an extended example of experimental studies. Key concepts used in our discussion are probability, populations, and sampling. To read more about inferential statistics, click on the items below:

- ExperimentsPopulation
- ProbabilitySampling
- Matching

## Experiments

A typical experimental study involves collecting data on the behaviours, attitudes, or actions of two or more groups and attempting to answer a research question (often called a hypothesis). Based on the analysis of the data, a researcher might then attempt to develop a causal model that can be populations.

A question that might be addressed through experimental research might be "Does grammar-based writing instruction produce better writers than process-based writing instruction?" Because it would be impossible and impractical to observe, interview, survey, etc. all first-year writing students and instructors in classes using one or the other of these instructional approaches, a researcher would study a sample or a subset of a population. Samplingor the creation of this subset of a population – is used by many researchers who desire to make sense of some phenomenon.

To analyse differences in the ability of student writers who are taught in each type of classroom, the researcher would compare the writing performance of the two groups of students. Two key concepts used to conduct the comparison are:

- Dependent Variables
- Independent Variables

**Probability**

Beginning researchers most often use the word *probability* to express a subjective judgment about the likelihood, or degree of certainty, that a particular event will occur. People say such things as: "It will probably rain tomorrow." "It is unlikely that we will win the ball game." It is possible to assign a number to the event being predicted, a number between 0 and 1, which represents *degree of confidence* that the event will occur. For example, a student might say that the likelihood an instructor will give an exam next week is about 90 percent, or .9. Where 100 percent, or 1.00, represents certainty, .9 would mean the student is almost certain the instructor will give an exam. If the student assigned the number .6, the likelihood of an exam would be just slightly greater than the likelihood of no exam. A rating of 0 would indicate complete certainty that *no* exam would be given(Shoeninger, 1971).

The probability of a particular outcome or set of outcomes is called a *p-value*. In our discussion, a p-value will be symbolized by a *p* followed by parentheses enclosing a symbol of the outcome or set of outcomes. For example, p(X) should be read, "the probability of a given X scores" (Shoeninger). Thus p(exam) should be read, "the probability an instructor will give an exam next week."

**Population**

A *population* is a group which is studied. In educational research, the population is usually a group of people. Researchers seldom are able to study every member of a population. Usually, they instead study a representative sample or subset of a population. Researchers then generalise their findings about the sample to the population as a whole.

**Sampling**Sampling *is* performed so that a population under study can be reduced to a manageable size. This can be accomplished via random sampling, discussed below, or via matching.

*Random sampling* is a procedure used by researchers in which all samples of a particular size have an equal chance to be chosen for an observation, experiment, etc (Runyon and Haber, 1976). There is no predetermination as to which members are chosen for the sample. This type of sampling is done in order to minimise scientific biases and offers the greatest likelihood that a sample will indeed be representative of the larger population. The aim here is to make the sample as representative of the population as possible. Note that the closer a sample distribution approximates the population distribution, the more generalised the results of the sample study are to the population. Notions of probability apply here. Random sampling provides the greatest probability that the distribution of scores in a sample will closely approximate the distribution of scores in the overall population.

### Matching

*Matching* is a method used by researchers to gain accurate and precise results of a study so that they may be applicable to a larger population. After a population has been examined and a sample has been chosen, a researcher must then consider variables, or extrinsic factors, that might affect the study. Matching methods apply when researchers are aware of extrinsic variables before conducting a study. Two methods used to match groups are:

- Precision Matching
- Frequency Distribution

Although, in theory, matching tends to produce valid conclusions, a rather obvious difficulty arises in finding subjects which are compatible. Researchers may even believe that experimental and control groups are identical when, in fact, a number of variables have been overlooked. For these reasons, researchers tend to reject matching methods in favor of random sampling.

### Methods

Statistics can be used to analyze individual variables, relationships among variables, and differences between groups. In this section, we explore a range of statistical methods for conducting these analyses.

Statistics can be used to analyze individual variables, relationships among variables, and differences between groups. To read more about statistical methods, click on the items below:

- Analysing Individual Variables
- Analysing Differences Between Groups
- Analysing Relationships Among Variables

### Analysing Individual Variables

The statistical procedures used to analyse a single variable describing a group (such as a population or representative sample) involve measures of *central tendency* and measures of *variation*. To explore these measures, a researcher first needs to consider the *distribution*, or range of values of a particular variable in a population or sample. *Normal distribution* occurs if the distribution of a population is completely normal. When graphed, this type of distribution will look like a bell curve; it is symmetrical and most of the scores cluster toward the middle. *Skewed Distribution* simply means the distribution of a population is not normal. The scores might cluster toward the right or the left side of the curve, for instance. Or there might be two or more clusters of scores, so that the distribution looks like a series of hills.

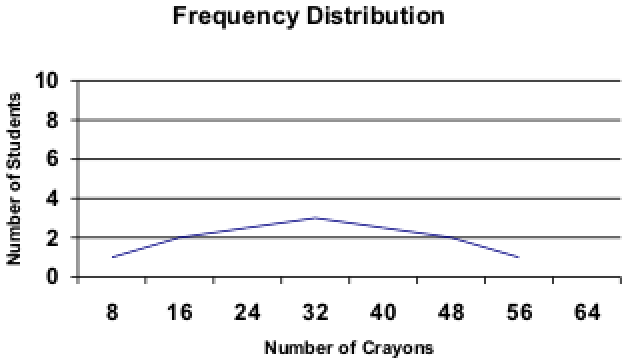
Once *frequency distributions* have been determined, researchers can calculate measures of central tendency and measures of variation. Measures of central tendency indicate averages of the distribution, and measures of variation indicate the spread, or range, of the distribution (Hinkle, Wiersma and Jurs 1988).

View More about Measures of Central Tendency
View More about Measures of Variation

**Measures of Central Tendency**

Central tendency is measured in three ways: *mean*,*median* and *mode*. The mean is simply the average score of a distribution. The median is the center, or middle score within a distribution. The mode is the most frequent score within a distribution. In a normal distribution, the mean, median and mode are identical.

| **Student** | **# of Crayons** |
|---|---|
| A | 8 |
| B | 16 |
| C | 16 |
| D | 32 |
| E | 32 |
| F | 32 |
| G | 48 |
| H | 48 |
| J | 56 |



**Measures of Variation**

Measures of variation determine the range of the distribution, relative to the measures of central tendency. Where the measures of central tendency are specific data points, measures of variation are lengths between various points within the distribution. Variation is measured in terms of range, mean deviation, variance, and standard deviation (Hinkle, Wiersma and Jurs 1988).

The *range* is the distance between the lowest data point and the highest data point. Deviation scores are the distances between each data point and the mean.

*Mean deviation* is the average of the absolute values of the deviation scores; that is, mean deviation is the average distance between the mean and the data points. Closely related to the measure of mean deviation is the measure of *variance*.

*Variance* also indicates a relationship between the mean of a distribution and the data points; it is determined by averaging the sum of the squared deviations. Squaring the differences instead of taking the absolute values allows for greater flexibility in calculating further algebraic manipulations of the data. Another measure of variation is the *standard deviation.*

*Standard deviation* is the square root of the variance. This calculation is useful because it allows for the same flexibility as variance regarding further calculations and yet also expresses variation in the same units as the original measurements (Hinkle, Wiersma and Jurs 1988).

**Analysing Differences between Groups**

Statistical tests can be used to analyze differences in the scores of two or more groups. The following statistical tests are commonly used to analyze differences between groups:

- T-Test
- Matched Pairs T-Test
- Analysis of Variance (ANOVA)

**T-Tests**

A t-test is used to determine if the scores of two groups differ on a single variable. A t-test is designed to test for the differences in mean scores. For instance, you could use a t-test to determine whether writing ability differs among students in two classrooms.

**Note:** A t-test is appropriate only when looking at *paired* data. It is useful in analyzing scores of two groups of participants on a particular variable or in analyzing scores of a single group of participants on two variables.

**Matched-Pairs T-Tests**

This type of t-test could be used to determine if the scores of the same participants in a study differ under different conditions. For instance, this sort of t-test could be used to determine if people write better essays *after* taking a writing class than they did *before* taking the writing class.

**Note:** A t-test is appropriate only when looking at *paired* data. It is useful in analyzing scores of two groups of participants on a particular variable or in analyzing scores of a single group of participants on two variables.

**Analysis of Variance**

The ANOVA (analysis of variance) is a statistical test which makes a single, overall decision as to whether a significant difference is present among three or more sample means (Levin 484). An ANOVA is similar to a t-test. However, the ANOVA can also test multiple groups to see if they differ on one or more variables. The ANOVA can be used to test between-groups and within-groups differences. There are two types of ANOVAs:

**One-Way ANOVA:** This tests a group or groups to determine if there are differences on a *single* set of scores. For instance, a one-way ANOVA could determine whether freshmen, sophomores, juniors, and seniors differed in their reading ability.

**Multiple ANOVA (MANOVA):** This tests a group or groups to determine if there are differences on *two or more* variables. For instance, a MANOVA could determine whether freshmen, sophomores, juniors, and seniors differed in reading ability and whether those differences were reflected by gender. In this case, a researcher could determine (1) whether reading ability differed across class levels, (2) whether reading ability differed across gender, and (3) whether there was an interaction between class level and gender.

### Analysing Relationships among Variables

Statistical relationships between variables rely on notions of correlation and regression. These two concepts aim to describe the ways in which variables relate to one another:

- Correlation
- Regression

### Correlation

Correlation tests are used to determine how strongly the scores of two variables are associated or correlated with each other. A researcher might want to know, for instance, whether a correlation exists between students' writing placement examination scores and their scores on a standardized test such as the ACT or SAT. Correlation is measured using values between +1.0 and -1.0. Correlations close to 0 indicate little or no relationship between two variables, while correlations close to +1.0 (or -1.0) indicate strong positive (or negative) relationships (Hayes et al. 554).

Correlation denotes positive or negative association between variables in a study. Two variables are *positively associated* when larger values of one tend to be accompanied by larger values of the other. The variables are *negatively associated* when larger values of one tend to be accompanied by smaller values of the other (Moore 208).

An example of a strong positive correlation would be the correlation between age and job experience. Typically, the longer people are alive, the more job experience they might have.

An example of a strong negative relationship might occur between the strength of people's party affiliations and their willingness to vote for a candidate from different parties. In many elections, Democrats are unlikely to vote for Republicans, and vice versa.

### Regression

Regression analysis attempts to determine the best "fit" between two or more variables. The independent variable in a regression analysis is a continuous variable, and thus allows you to determine how one or more independent variables predict the values of a dependent variable.

**Simple Linear Regression** is the simplest form of regression. Like a correlation, it determines the extent to which one independent variables predicts a dependent variable. You can think of a simple linear regression as a correlation line. Regression analysis provides you with more information than correlation does, however. It tells you how well the line "fits" the data. That is, it tells you how closely the line comes to all of your data points. The line in the figure indicates the regression line drawn to find the best fit among a set of data points. Each dot represents a person and the axes indicate the amount of job experience and the age of that person. The dotted lines indicate the distance from the regression line. A smaller total distance indicates a better fit. Some of the information provided in a regression analysis, as a result, indicates the slope of the regression line, the R value (or correlation), and the strength of the fit (an indication of the extent to which the line can account for variations among the data points).

**Multiple Linear Regression** allows one to determine how well multiple independent variables predict the value of a dependent variable. A researcher might examine, for instance, how well age and experience predict a person's salary. The interesting thing here is that one would no longer be dealing with a regression "line." Instead, since the study deals with three dimensions (age, experience, and salary), it would be dealing with a plane, that is, with a two-dimensional figure. If a fourth variable was added to the equations, one would be dealing with a three-dimensional figure, and so on.

## Commentary

### Misuses of Statistics

Statistics consists of tests used to analyze data. These tests provide an analytic framework within which researchers can pursue their research questions. This framework provides one way of working with observable information. Like other analytic frameworks, statistical tests can be misused, resulting in potential misinterpretation and misrepresentation. Researchers decide which research questions to ask, which groups to study, how those groups should be divided, which variables to focus upon, and how best to categorize and measure such variables. The point is that researchers retain the ability to manipulate any study even as they decide what to study and how to study it.

***Potential Misuses:***

- Manipulating scale to change the appearance of the distribution of data
- Eliminating high/low scores for more coherent presentation
- Inappropriately focusing on certain variables to the exclusion of other variables
- Presenting correlation as causation

***Measures against Potential Misuses:***

- Testing for reliability and validity
- Testing for statistical significance
- Critically reading statistics

## Key Terms

### Glossary of Key Terms

**Accuracy** A term used in survey research to refer to the match between the target population and the sample.

**ANCOVA (Analysis** Same method as ANOVA, but analyzes differences between dependent variables.

**of Co-Variance)**

**ANOVA (Analysis of Variance)** A method of statistical analysis broadly applicable to a number of research designs, used to determine differences among the means of two or more groups on a variable. The independent variables are usually nominal, and the dependent variable is usual an interval.

**Apparency** Clear, understandable representation of the data

**Bell curve** A frequency distribution statistics. Normal distribution is shaped like a bell.

**Case Study** The collection and presentation of detailed information about a particular participant or small group, frequently including the accounts of subjects themselves.

**Causal Model** A model which represents a causal relationship between two variables.

**Causal Relationship** The relationship established that shows that an independent variable, and nothing else, causes a change in a dependent variable. Establishes, also, how much of a change is shown in the dependent variable.

**Causality** The relation between cause and effect.

**Central Tendency** These measures indicate the middle or center of a distribution.

**Confirmability** Objectivity; the findings of the study could be confirmed by another person conducting the same study

**Confidence Interval** The range around a numeric statistical value obtained from a sample, within which the actual, corresponding value for the population is likely to fall, at a given level of probability (Alreck, 444).

**Confidence Level** The specific probability of obtaining some result from a sample if it did not exist in the population as a whole, at or below which the relationship will be regarded as statistically significant (Alreck, 444).

**Confidence Limits** (Same as confidence interval, but is terminology used by Lauer and Asher.) "The range of scores or percentages within which a population percentage is likely to be found on variables that describe that population" (Lauer and Asher, 58). Confidence limits are expressed in a "plus or minus" fashion according to sample size, then corrected according to formulas based on variables connected to population size in relation to sample size and the relationship of the variable to the population size--the larger the sample, the smaller the variability or confidence limits.

**Confounding Variable** An unforeseen, and unaccounted-for variable that jeopardizes reliability and validity of an experiment's outcome.

**Construct Validity** Seeks an agreement between a theoretical concept and a specific measuring device, such as observation.

**Content Validity** The extent to which a measurement reflects the specific intended domain of content (Carmines & Zeller, 1991, p.20).

**Context sensitivity** Awareness by a qualitative researcher of factors such as values and beliefs that influence cultural behaviors

**Continuous Variable** A variable that may have fractional values, e.g., height, weight and time.

**Control Group** A group in an experiment that receives not treatment in order to compare the treated group against a norm.

**Convergent Validity** The general agreement among ratings, gathered independently of one another, where measures should be theoretically related.

**Correlation** 1) A common statistical analysis, usually abbreviated as **r**, that measures the degree of relationship between pairs of interval variables in a sample. The range of correlation is from -1.00 to zero to +1.00. 2) A non-cause and effect relationship between two variables.

**Covariate** A product of the correlation of two related variables times their standard deviations. Used in true experiments to measure the difference of treatment between them.

**Credibility** A researcher's ability to demonstrate that the object of a study is accurately identified and described, based on the way in which the study was conducted

**Criterion Related Validity** Used to demonstrate the accuracy of a measuring procedure by comparing it with another procedure which has been demonstrated to be valid; also referred to as instrumental validity.

**Data** Recorded observations, usually in numeric or textual form

**Deductive** A form of reasoning in which conclusions are formulated about particulars from general or universal premises

**Dependability** Being able to account for changes in the design of the study and the changing conditions surrounding what was studied.

**Dependent Variable** A variable that receives stimulus and measured for the effect the treatment has had upon it.

**Design flexibility** A quality of an observational study that allows researchers to pursue inquiries on new topics or questions that emerge from initial research

**Deviation** The distance between the mean and a particular data point in a given distribution.

**Discourse Community** A community of scholars and researchers in a given field who respond to and communicate to each other through published articles in the community's journals and presentations at conventions. All members of the discourse community adhere to certain conventions for the presentation of their theories and research.

**Discrete Variable** A variable that is measured solely in whole units, e.g., gender and siblings

**Discriminate Validity** The lack of a relationship among measures which theoretically should not be related.

**Distribution** The range of values of a particular variable.

**Dynamic systems** Qualitative observational research is not concerned with having straight-forward, right or wrong answers. Change in a study is common because the researcher is not concerned with finding only one answer.

**Electronic Text** A "paper" or linear text that has been essentially "copied" into an electronic medium.

**Empathic neutrality** A quality of qualitative researchers who strive to be non-judgmental when compiling findings

**Empirical Research** "...the process of developing systematized knowledge gained from observations that are formulated to support insights and generalizations about the phenomena under study" (Lauer and Asher, 1988, p. 7)

**Equivalency Reliability** The extent to which two items measure identical concepts at an identical level of difficulty.

**Ethnography** Ethnographies study groups and/or cultures over a period of time. The goal of this type of research is to comprehend the particular group/culture through observer immersion into the culture or group. Research is completed through various methods, which are similar to those of case studies, but since the researcher is immersed within the group for an extended period of time more detailed information is usually collected during the research.

**Ethnomethodology** A form of ethnography that studies activities of group members to see how they make sense of their surroundings

**Existence or Frequency** This is a key question in the coding process. The researcher must decide if he/she is going to count a concept only once, for existence, no matter how many times it appears, or if he/she will count it each time it occurs. For example, "damn" could be counted once, even though it appears 50 times, or it could be counted all 50 times. The latter measurement may be interested in how many times it occurs and what that indicates, whereas the former may simply looking for existence, period.

**Experiment** Experimental Research A researcher working within this methodology creates an environment in which to observe and interpret the results of a research question. A key element in experimental research is that participants in a study are randomly assigned to groups. In an attempt to create a causal model (i.e., to discover the causal origin of a particular phenomenon), groups are treated differently and measurements are conducted to determine if different treatments appear to lead to different effects.

**External Validity** The extent to which the results of a study are generalisable or transferable. See also validity

**Face Validity** How a measure or procedure appears.

**Factor Analysis** A statistical test that explores relationships among data. The test explores which variables in a data set are most related to each other. In a carefully constructed survey, for example, factor analysis can yield information on patterns of responses, not simply data on a single response. Larger tendencies may then be interpreted, indicating behaviour trends rather than simply responses to specific questions.

**Generalisability** The extent to which research findings and conclusions from a study conducted on a sample population can be applied to the population at large.

**Grounded theory** Practice of developing other theories that emerge from observing a group. Theories are grounded in the group's observable experiences, but researchers add their own insight into why those experiences exist.

**Holistic** Taking almost every action or communication of the whole phenomenon of a certain community or culture into

**perspective** account in research

**Hypertext** A nonsequential text composed of links and nodes

**Hypothesis** A tentative explanation based on theory to predict a causal relationship between variables.

**Independent Variable** A variable that is part of the situation that exist from which originates the stimulus given to a dependent variable. Includes treatment, state of variable, such as age, size, weight, etc.

**Inductive** A form of reasoning in which a generalized conclusion is formulated from particular instances

**Inductive analysis** A form of analysis based on inductive reasoning; a researcher using inductive analysis starts with answers, but forms questions throughout the research process.

**Internal Consistency** The extent to which all questions or items assess the same characteristic, skill, or quality.

**Internal Validity** (1) The rigor with which the study was conducted (e.g., the study's design, the care taken to conduct measurements, and decisions concerning what was and wasn't measured) and (2) the extent to which the designers of a study have taken into account alternative explanations for any causal relationships they explore (Huitt, 1998). In studies that do not explore causal relationships, only the first of these definitions should be considered when assessing internal validity. See also validity.

**Interrater Reliability** The extent to which two or more individuals agree. It addresses the consistency of the implementation of a rating system.

**Interval Variable** A variable in which both order of data points and distance between data points can be determined, e.g., percentage scores and distances

**Interviews** A research tool in which a researcher asks questions of participants; interviews are often audio- or video-taped for later transcription and analysis.

**Irrelevant Information** One must decide what to do with the information in the text that is not coded. One's options include either deleting or skipping over unwanted material, or viewing all information as relevant and important and using it to re-examine, reassess and perhaps even alter the one's coding scheme.

**Kinesics** Kinesic analysis examines what is communicated through body movement

**Level of Analysis** Chosen by determining which word, set of words, or phrases will constitute a concept. According to Carley, 100-500 concepts is generally sufficient when coding for a specific topic, but this number of course varies on a case by case basis.

**Level of Generalization** A researcher must decide whether concepts are to be coded exactly as they appear, or if they can be recorded in some altered or collapsed form. Using Horton as an example again, she could code profanity individually and code "damn" and "dammit" as two separate concepts. Or, by generalizing their meaning, i.e. they both express the same idea, she could group them together as one item, i.e. "damn words."

**Level of Implication** One must determine whether to code simply for explicit appearances of concepts, or for implied concepts, as well. For example, consider a hypothetical piece of text about skiing, written by an expert. The expert might refer several times to "???," as well as various other kinds of turns. One must decide whether to code "???" as an entity in and of itself, or, if coding for "turn" references in general, to code "???" as implicitly meaning "turn." Thus, by determining that the meaning "turn" is implicit in the words "???," anytime the words "???" or "turn" appear in the text, they will be coded under the same category of "turn."

**Link** In hypertext, a pointer from one node to another

**Matched T-Test** A statistical test used to compare two sets of scores for the same subject. A matched pairs T-test can be used to determine if the scores of the same participants in a study differ under different conditions. For instance, this sort of t-test could be used to determine if people write better essays after taking a writing class than they did before taking the writing class.

**Matching** Process of corresponding variables in experimental groups equally feature for feature.

**Mean** The average score within a distribution.

**Mean Deviation** A measure of variation that indicates the average deviation of scores in a distribution from the mean: It is determined by averaging the absolute values of the deviations.

**Median** The centre score in a distribution.

**Mental Models** A group or network of interrelated concepts that reflect conscious or subconscious perceptions of reality. These internal mental networks of meaning are constructed as people draw inferences and gather information about the world.

**Mode** The most frequent score in a distribution.

**Multi-Modal Methods** A research approach that employs a variety of methods; see also triangulation

**Narrative Inquiry** A qualitative research approach based on a researcher's narrative account of the investigation, not to be confused with a narrative examined by the researcher as data

**Naturalistic Inquiry** Observational research of a group in its natural setting

**Node** In hypertext, each unit of information, connected by links

**Nominal Variable** A variable determined by categories which cannot be ordered, e.g., gender and colour

**Normal distribution** A normal frequency distribution representing the probability that a majority of randomly selected members of a population will fall within the middle of the distribution. Represented by the bell curve.

**Ordinal Variable** A variable in which the order of data points can be determined but not the distance between data points, e.g., letter grades

**Parameter** A coefficient or value for the population that corresponds to a particular statistic from a sample and is often inferred from the sample.

**Phenomenology** A qualitative research approach concerned with understanding certain group behaviours from that group's point of view

**Population** The target group under investigation, as in all students enrolled in first-year composition courses taught in traditional classrooms. The population is the entire set under consideration. Samples are drawn from populations.

**Precision** In survey research, the tightness of the confidence limits.

**Pre-defined or Interactive Concept Choice** One must determine whether to code only from a pre-defined set of concepts and categories, or if one will develop some or all of these during the coding process. For example, using a predefined set, Horton would code only for profane language. But, if Horton coded interactively, she may have decided to half-way through the process that the text warranted coding for profane gestures, as well.

**Probability** The chance that a phenomenon has a of occurring randomly. As a statistical measure, it shown as **p** (the "p" factor).

**Qualitative Research** Empirical research in which the researcher explores relationships using textual, rather than quantitative data. Case study, observation, and ethnography are considered forms of qualitative research. Results are not usually considered generalizable, but are often transferable.

**Quantitative Research** Empirical research in which the researcher explores relationships using numeric data. Survey is generally considered a form of quantitative research. Results can often be generalized, though this is not always the case.

**Quasi-experiment** Similar to true experiments. Have subjects, treatment, etc., but uses nonrandomized groups. Incorporates interpretation and transferability in order to compensate for lack of control of variables.

**Quixotic Reliability** Refers to the situation where a single manner of observation consistently, yet erroneously, yields the same result.

**Random sampling** Process used in research to draw a sample of a population strictly by chance, yielding no discernible pattern beyond chance. Random sampling can be accomplished by first numbering the population, then selecting the sample according to a table of random numbers or using a random-number computer generator. The sample is said to be random because there is no regular or discernible pattern or order. Random sample selection is used under the assumption that sufficiently large samples assigned randomly will exhibit a distribution comparable to that of the population from which the sample is drawn.

**Randomization** Used to allocate subjects to experimental and control groups. The subjects are initially considered not unequal because they were randomly selected.

**Range** The difference between the highest and lowest scores in a distribution.

**Reliability** The extent to which a measure, procedure or instrument yields the same result on repeated trials.

**Response Rate** In survey research, the actual percentage of questionnaires completed and returned.

**Rhetorical Inquiry** "entails…1) identifying a motivational concern, 2) posing questions, 3) engaging in a heuristic search (which in composition studies has often occurred by probing other fields), 4) creating a new theory or hypotheses, and 5) justifying the theory" (Lauer and Asher, 1988, p. 5)

**Rigor**
Degree to which research methods are scrupulously and meticulously carried out in order to recognize important influences occurring in a experiment.

**Sampling Error**
The degree to which the results from the sample deviate from those that would be obtained from the entire population, because of random error in the selection of respondent and the corresponding reduction in reliability (Alreck, 454).

**Sampling Frame**
A listing that should include all those in the population to be sampled and exclude all those who are not in the population (Alreck, 454).

**Sample**
The population researched in a particular study. Usually, attempts are made to select a "sample population" that is considered representative of groups of people to whom results will be generalized or transferred. In studies that use inferential statistics to analyze results or which are designed to be generalisable, sample size is critical--generally the larger the number in the sample, the higher the likelihood of a representative distribution of the population.

**Selective Reduction**
The central idea of content analysis. Text is reduced to categories consisting of a word, set of words or phrases, on which the researcher can focus. Specific words or patterns are indicative of the research question and determine levels of analysis and generalization.

**Serial Effect**
In survey research, a situation where questions may "lead" participant responses through establishing a certain tone early in the questionnaire. The serial effect may accrue as several questions establish a pattern of response in the participant, biasing results.

**Short-term observation**
Studies that list or present findings of short-term qualitative study based on recorded observation

**Skewed Distribution**
Any distribution which is not normal, that is not symmetrical along the x-axis

**Stability Reliability**
The agreement of measuring instruments over time.

**Standard Deviation**
A term used in statistical analysis. A measure of variation that indicates the typical distance between the scores of a distribution and the mean; it is determined by taking the

square root of the average of the squared deviations in a given distribution. It can be used to indicate the proportion of data within certain ranges of scale values when the distribution conforms closely to the normal curve.

**Standard Error (S.E.) of the Mean**
A term used in statistical analysis. A computed value based on the size of the sample and the standard deviation of the distribution, indicating the range within which the mean of the population is likely to be from the mean of the sample at a given level of probability (Alreck, 456).

**Survey**
A research tool that includes at least one question which is either open-ended or close-ended and employs an oral or written method for asking these questions. The goal of a survey is to gain specific information about either a specific group or a representative sample of a particular group. Results are typically used to understand the attitudes, beliefs, or knowledge of a particular group.

**Synchronic Reliability**
The similarity of observations within the same time frame; it is not about the similarity of things observed.

**T-Test**
A statistical test. A t-test is used to determine if the scores of two groups differ on a single variable. For instance, to determine whether writing ability differs among students in two classrooms, a t-test could be used.

**Thick Description**
A rich and extensive set of details concerning methodology and context provided in a research report.

**Transferability**
The ability to apply the results of research in one context to another similar context. Also, the extent to which a study invites readers to make connections between elements of the study and their own experiences.

**Translation Rules**
If one decides to generalize concepts during coding, then one must develop a set of rules by which less general concepts will be translated into more general ones. This doesn't involve simple generalization, for example, as with "damn" and "dammit," but requires one to determine, from a given set of concepts, what concepts are missing. When dealing with the idea of profanity, one must decide what to do with the concept "dang it," which is generally thought to imply "damn it." The researcher must make this distinction, i.e. make this implicit concept explicit, and then code for the frequency of its occurrence. This decision results in the construction of a translation rule, which instructs the researcher to code for the concept "dang it" in a certain way.

**Treatment**
The stimulus given to a dependent variable.

**Triangulation**

The use of a combination of research methods in a study. An example of triangulation would be a study that incorporated surveys, interviews, and observations. See also multi-modal methods

**Unique case orientation**

A perspective adopted by many researchers conducting qualitative observational studies; researchers adopting this orientation remember every study is special and deserves in-depth attention. This is especially necessary for doing cultural comparisons.

**Validity**

The degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure. A method can be reliable, consistently measuring the same thing, but not valid. See also internal validity and external validity

**Variable**

Observable characteristics that vary among individuals. See also ordinal variable, nominal variable, interval variable, continuous variable, discrete variable, dependent variable, independent variable.

**Variance**

A measure of variation within a distribution, determined by averaging the squared deviations from the mean of a distribution.

**Variation**

The dispersion of data points around the mean of a distribution.

**Verisimilitude**

Having the semblance of truth; in research, it refers to the probability that the research findings are consistent with occurrences in the "real world."

**Biostatistics**

Biostatistics involves the theory and application of statistical science to analyze public health problems and to further biomedical research. The faculty includes leaders in the development of statistical methods for clinical trials and observational studies, studies on the environment, and genomics/genetics. The department's research in statistical methods and interdisciplinary collaborations provide many opportunities for student participation.

Current departmental research on statistical and computing methods for observational studies and clinical trials includes survival analysis, missing-data problems, and causal inference. Other areas of investigation are environmental research (methods for longitudinal studies, analyses with incomplete data, and meta-analysis); statistical aspects of the study of AIDS and cancer; quantitative problems in health-risk analysis, technology assessment, and clinical decision making; statistical methodology in psychiatric research and in genetic studies; Bayesian statistics; statistical computing; statistical genetics and computational biology; and collaborative research activities with biomedical scientists in other Harvard-affiliated institutions.

**Measures of Risk**

This Lesson describes the measures of central location and spread, which are useful for summarizing continuous variables. However, many variables used by field epidemiologists are categorical variables, some of which have only two categories — exposed yes/no, test positive/negative, case/control, and so on. These variables have to be summarized with frequency measures such as ratios, proportions, and rates. Incidence, prevalence, and mortality rates are three frequency measures that are used to characterize the occurrence of health events in a population.

**Objectives**

*At the end of this module, you will be able to:*

*• Calculate and interpret the following epidemiologic measures:*

*– Ratio*

*– Proportion*

*– Incidence proportion (attack rate)*

*– Incidence rate*

*– Prevalence*

*– Mortality rate*

*• Choose and apply the appropriate measures of association and measures of public health impact*

**Frequency Measures**

A measure of central location provides a single value that summarizes an entire distribution of data. In contrast, a frequency measure characterizes only part of the distribution. Frequency measures compare one part of the distribution to another part of the distribution, or to the entire distribution. Common frequency measures are **ratios**, **proportions**, and **rates**. All three frequency measures have the same basic form: *numeratordenominator x* $10^n$

Recall that:

$10^0$ = 1 (anything raised to the 0 power equals 1)

$10^1$ = 10 (anything raised to the 1st power is the value itself)

$10^2$ = 10 x 10 = 100

$10^3$ = 10 x 10 x 10 = 1,000

So the fraction of (numerator/denominator) can be multiplied by 1, 10, 100, 1000, and so on. This multiplier varies by measure and will be addressed in each section.

**Ratio*:***

A ratio is the relative magnitude of two quantities or a comparison of any two values. It is calculated by dividing one interval- or ratio-scale variable by the other. The numerator and denominator need not be related. Therefore, one could compare apples with oranges or apples with number of physician visits.

***Method for calculating a ratio***

*Number or rate of events, items, persons, etc. in one group*

*Number or rate of events, items, persons, etc. in another group*

After the numerator is divided by the denominator, the result is often expressed as the result "to one" or written as the result ":1."

Note that in certain ratios, the numerator and denominator are different categories of the same variable, such as males and females, or persons 20–29 years and 30–39 years of age. In other ratios, the numerator and denominator are completely different variables, such as the number of hospitals in a city and the size of the population living in that city.

**EXAMPLE: Calculating a Ratio — Different Categories of Same Variable**

Between 1971 and 1975, as part of the National Health and Nutrition Examination Survey (NHANES), 7,381 persons ages 40–77 years were enrolled in a follow-up study.1 At the time of enrollment, each study participant was classified as having or not having diabetes. During 1982–1984, enrollees were documented either to have died or were still alive. The results are summarized as follows.

Original Enrollment Dead at Follow-Up

(1971–1975) (1982–1984)

Diabetic men 189 100

Nondiabetic men 3,151 811

Diabetic women 218 72

Nondiabetic women 3,823 511

Of the men enrolled in the NHANES follow-up study, 3,151 were nondiabetic and 189 were diabetic. Calculate the ratio of non-diabetic to diabetic men.

Ratio = 3,151 / 189 x 1 = 16.7:1

***Properties and uses of ratios:***

• Ratios are common descriptive measures, used in all fields. In epidemiology, ratios are used as both descriptive measures and as analytic tools. As a descriptive measure, ratios can describe the male-to-female ratio of participants in a study, or the ratio of controls to cases (e.g., two controls per case). As an analytic tool, ratios can be calculated for occurrence of illness, injury, or death between two groups. These ratio measures, including risk ratio (relative risk), rate ratio, and odds ratio, are described later in this lesson.

• As noted previously, the numerators and denominators of a ratio can be related or unrelated. In other words, you are free to use a ratio to compare the number of males in a population with the number of females, or to compare the number of residents in a population with the number of hospitals or dollars spent on over-the-counter medicines.

• Usually, the values of both the numerator and denominator of a ratio are divided by the value of one or the other so that either the numerator or the denominator equals 1.0. So the ratio of non-diabetics to diabetics cited in the previous example is more likely to be reported as 16.7:1 than 3,151:189.

**EXAMPLES: Calculating Ratios for Different Variables**

**Example A:** A city of 4,000,000 persons has 500 clinics. Calculate the ratio of clinics per person.

500 / 4,000,000 x 10n = 0.000125 clinics per person

To get a more easily understood result, you could set 10n = 104 = 10,000. Then the ratio becomes:

0.000125 x 10,000 = 1.25 clinics per 10,000 persons

You could also divide each value by 1.25, and express this ratio as 1 clinic for every 8,000 persons.

**Example B:** Delaware's infant mortality rate in 2001 was 10.7 per 1,000 live births.2 New Hampshire's infant mortality rate in 2001 was 3.8 per 1,000 live births. Calculate the ratio of the infant mortality rate in Delaware to that in New Hampshire.

10.7 / 3.8 x 1 = 2.8:1

Thus, Delaware's infant mortality rate was 2.8 times as high as New Hampshire's infant mortality rate in 2001.

***A commonly used epidemiologic ratio: death-to-case ratio:***

Death-to-case ratio is the number of deaths attributed to a particular disease during a specified period divided by the number of new cases of that disease identified during the same period. It is used as a measure of the severity of illness: the death-to-case ratio for rabies is close to 1 (that is, almost everyone who develops rabies dies from it), whereas the death-to-case ratio for the common cold is close to 0.

For example, in the United States in 2002, a total of 15,075 new cases of tuberculosis were reported.3 During the same year, 802 deaths were attributed to tuberculosis. The tuberculosis death-tocase ratio for 2002 can be calculated as 802 / 15,075. Dividing both numerator and denominator by the numerator yields 1 death per 18.8 new cases. Dividing both numerator and denominator by the denominator (and multiplying by 10n = 100) yields 5.3 deaths per 100 new cases. Both expressions are correct.

Note that, presumably, many of those who died had initially contracted tuberculosis years earlier. Thus many of the 802 in the numerator are not among the 15,075 in the denominator. Therefore, the death-to-case ratio is a ratio, but not a proportion.

**Proportion:**

A proportion is the comparison of a part to the whole. It is a type of ratio in which the numerator is included in the denominator. You might use a proportion to describe what fraction of clinic patients tested positive for HIV, or what percentage of the population is younger than 25 years of age. A proportion may be expressed as a decimal, a fraction, or a percentage.

***Method for calculating a proportion***

*Number of persons or events with a particular characteristic*

*Total number of persons or events, of which the numerator is a subset x 10n*

For a proportion, 10n is usually 100 (or n=2) and is often expressed as a percentage.

**EXAMPLE: Calculating a Proportion**

**Example A:** Calculate the proportion of men in the NHANES follow-up study who were diabetics.

Numerator = 189 diabetic men

Denominator = Total number of men = 189 + 3,151 = 3,340

Proportion = (189 / 3,340) x 100 = 5.66%

**Example B:** Calculate the proportion of deaths among men.

Numerator = deaths in men

= 100 deaths in diabetic men + 811 deaths in nondiabetic men

= 911 deaths in men

Notice that the numerator (911 deaths in men) is a subset of the denominator.

Denominator = all deaths

= 911 deaths in men + 72 deaths in diabetic women + 511 deaths in nondiabetic women

= 1,494 deaths

Proportion = 911 / 1,494 = 60.98% = 61%

**Your Turn:** What proportion of all study participants were men? (Answer = 45.25%)

***Properties and uses of proportions:***

• Proportions are common descriptive measures used in all fields. In epidemiology, proportions are used most often as descriptive measures. For example, one could calculate the proportion of persons enrolled in a study among all those eligible ("participation rate"), the proportion of children in a village vaccinated against measles, or the proportion of persons who developed illness among all passengers of a cruise ship.

• Proportions are also used to describe the amount of disease that can be attributed to a particular exposure. For example, on the basis of studies of smoking and lung cancer, public health officials have estimated that greater than 90% of the lung cancer cases that occur are attributable to cigarette smoking.

• In a proportion, the numerator must be included in the denominator. Thus, the number of apples divided by the number of oranges is not a proportion, but the number of apples divided by the total number of fruits of all kinds is a proportion. Remember, the numerator is always a subset of the denominator.

• A proportion can be expressed as a fraction, a decimal, or a percentage. The statements "one fifth of the residents became ill" and "twenty percent of the residents became ill" are equivalent.

- Proportions can easily be converted to ratios. If the numerator is the number of women (179) who attended a clinic and the denominator is all the clinic attendees (341), the proportion of clinic attendees who are women is 179 / 341, or 52% (a little more than half). To convert to a ratio, subtract the numerator from the denominator to get the number of clinic patients who are not women, i.e., the number of men (341 – 179 = 162 men.)Thus, ratio of women to men could be calculated from the proportion as:

Ratio = 179 / (341 – 179) x 1

= 179 / 162

= 1.1 to 1 female-to-male ratio

Conversely, if a ratio's numerator and denominator together make up a whole population, the ratio can be converted to a proportion. You would add the ratio's numerator and denominator to form the denominator of the proportion, as illustrated in the NHANES follow-up study examples (provided earlier in this lesson).

***A specific type of epidemiologic proportion: proportionate mortality:***

Proportionate mortalit*y* is the proportion of deaths in a specified population during a period of time that are attributable to different causes. Each cause is expressed as a percentage of all deaths, and the sum of the causes adds up to 100%. These proportions are not rates because the denominator is all deaths, not the size of the population in which the deaths occurred.

***Rate:***

In epidemiology, a rate is a measure of the frequency with which an event occurs in a defined population over a specified period of time. Because rates put disease frequency in the perspective of the size of the population, rates are particularly useful for comparing disease frequency in different locations, at different times, or among different groups of persons with potentially different sized populations; that is, a rate is a measure of risk.

To a non-epidemiologist, rate means how fast something is happening or going. The speedometer of a car indicates the car's speed or rate of travel in miles or kilometers per hour. This rate is always reported per some unit of time. Some epidemiologists restrict use of the term rate to similar measures that are expressed per unit of time. For these epidemiologists, a rate describes how quickly disease occurs in a population, for example, 70 new cases of breast cancer per 1,000 women per year. This measure conveys a sense of the speed with which disease occurs in a population, and seems to imply that this pattern has occurred and will continue to occur for the foreseeable future. This rate is an ***incidence rate***, described in the next section.

Other epidemiologists use the term rate more loosely, referring to proportions with case counts in the numerator and size of population in the denominator as rates. Thus, an **attack rate** is the proportion of the population that develops illness during an outbreak. For example, 20 of 130 persons developed diarrhea after attending a picnic. (An alternative and more accurate phrase for attack rate is **incidence proportion**.) A **prevalence rate** is the proportion of the population that has a health condition at a point in time. For example, 70 influenza case-patients in March 2005 reported in County A. A **case-fatality rate** is the proportion of persons with the disease who die from it. Forexample, one death due to meningitis among County A's population. All of these measures are proportions, and none is expressed per units of time. Therefore, these measures are not considered "true" rates by some, although use of the terminology is widespread.

**Incidence** refers to the occurrence of new cases of disease or injury in a population over a specified period of time. Although some epidemiologists use incidence to mean the number of new cases in a community, others use incidence to mean the number of new cases per unit of population. Two types of incidence are commonly used — **incidenceproportion** and **incidence rate**.

***Incidence proportion or risk:***

Incidence proportion is the proportion of an initially disease-free population that develops disease, becomes injured, or dies during a specified (usually limited) period of time. Synonyms include attack rate, risk, probability of getting disease, and cumulative incidence. Incidence proportion is a proportion because the persons in the numerator, those who develop disease, are all included in the denominator (the entire population).

***Method for calculating incidence proportion (risk)***

*Number of new cases of disease or injury during specified period*

*Size of population at start of period*

**EXAMPLES: Calculating Incidence Proportion (Risk)**

**Example A:** In the study of diabetics, 100 of the 189 diabetic men died during the 13-year follow-up period. Calculate the risk of death for these men.

Numerator = 100 deaths among the diabetic men

Denominator = 189 diabetic men

10n = 102 = 100

Risk = (100 / 189) x 100 = 52.9%

**Example B:** In an outbreak of gastroenteritis among attendees of a corporate picnic, 99 persons ate potato salad, 30 of whom developed gastroenteritis. Calculate the risk of illness among persons who ate potato salad.

Numerator = 30 persons who ate potato salad and developed gastroenteritis

Denominator = 99 persons who ate potato salad

10n = 102 = 100

Risk = “Food-specific attack rate” = (30 / 99) x 100 = 0.303 x 100 = 30.3%

***Properties and uses of incidence proportions:***

- Incidence proportion is a measure of the risk of disease or the probability of developing the disease during the specified period. As a measure of incidence, it includes only new cases of disease in the numerator. The denominator is the number of persons in the population at the start of the observation period. Because all of the persons with new cases of disease (numerator) are also represented in the denominator, a risk is also a proportion.

- In the outbreak setting, the term **attack rate** is often used as a synonym for risk. It is the risk of getting the disease during a specified period, such as the duration of an outbreak. A variety of attack rates can be calculated.

**Overall attack rate** is the total number of new cases divided by the total population.

A **food-specific attack rate** is the number of persons who ate a specified food and became ill divided by the total number of persons who ate that food, as illustrated in the previous potato salad example.

A **secondary attack rate** is sometimes calculated to document the difference between community transmission of illness versus transmission of illness in a household, barracks, or other closed population. It is calculated as:

*Number of cases among contacts of primary cases*

*Total number of contacts x 10n*

Often, the total number of contacts in the denominator is calculated as the total population in the households of the primary cases, minus the number of primary cases. For a secondary attack rate, 10n usually is 100%.

**EXAMPLE: Calculating Secondary Attack Rates**

Consider an outbreak of shigellosis in which 18 persons in 18 different households all became ill. If the population of the community was 1,000, then the overall attack rate was 18 / 1,000 x 100% = 1.8%. One incubation period later, 17 persons in the same households as these "primary" cases developed shigellosis. If the 18 households included 86 persons, calculate the secondary attack rate.

Secondary attack rate = (17 / (86 - 18)) x 100% = (17 / 68) x 100% = 25.0%

***Incidence rate or person-time rate:***

Incidence rate or person-time rate is a measure of incidence that incorporates time directly into the denominator. A person-time rate is generally calculated from a long-term cohort follow-up study, wherein enrollees are followed over time and the occurrence of new cases of disease is documented. Typically, each person is observed from an established starting time until one of four "end points" is reached: onset of disease, death, migration out of the study ("lost to follow-up"), or the end of the study. Similar to the incidence proportion, the numerator of the incidence rate is the number of new cases identified during the period of observation. However, the denominator differs. The denominator is the sum of the time each person was observed, totaled for all persons. This denominator represents the total time the population was at risk of and being watched for disease. Thus, the incidence rate is the ratio of the number of cases to the total time the population is at risk of disease.

***Method for calculating incidence rate***

*Number of new cases of disease or injury during specified period*

*Time each person was observed, totaled for all persons*

In a long-term follow-up study of morbidity, each study participant may be followed or observed for several years. One person followed for 5 years without developing disease is said to contribute 5 person-years of follow-up.

What about a person followed for one year before being lost to follow-up at year 2? Many researchers assume that persons lost to follow-up were, on average, disease-free for half the year, and thus contribute ½ year to the denominator. Therefore, the person followed for one year before being lost to follow-up contributes 1.5 person-years. The same assumption is made for participants diagnosed with the disease at the year 2 examination — some may have developed illness in month 1, and others in months 2 through 12. So, on average, they developed illness halfway through the year. As a result, persons diagnosed with the disease contribute ½ year of follow-up during the year of diagnosis.

The denominator of the person-time rate is the sum of all of the person-years for each study participant. So, someone lost to follow-up in year 3, and someone diagnosed with the disease in year 3, each contributes 2.5 years of disease-free follow-up to the denominator.

***Properties and uses of incidence rates:***

• An incidence rate describes how quickly disease occurs in a population. It is based on person-time, so it has some advantages over an incidence proportion. Because person-time is calculated for each subject, it can accommodate persons coming into and leaving the study. As noted in the previous example, the denominator accounts for study participants who are lost to follow-up or who die during the study period. In addition, it allows enrollees to enter the study at different times. In the NHANES follow-up study, some participants were enrolled in 1971, others in 1972, 1973, 1974, and 1975.

• Person-time has one important drawback. Person-time assumes that the probability of disease during the study period is constant, so that 10 persons followed for one year equals one person followed for 10 years. Because the risk of many chronic diseases increases with age, this assumption is often not valid.

• Long-term cohort studies of the type described here are not very common. However, epidemiologists far more commonly calculate incidence rates based on a numerator of cases observed or reported, and a denominator based on the mid-year population. This type of incident rate turns out to be comparable to a person-time rate.

• Finally, if you report the incidence rate of, say, the heart disease study as 2.5 per 1,000 person-years, epidemiologists might understand, but most others will not. Person-time is epidemiologic jargon. To convert this jargon to something understandable, simply replace "person-years" with "persons per year." Reporting the results as 2.5 new cases of heart disease per 1,000 persons per year sounds like English rather than jargon. It also conveys the sense of the incidence rate as a dynamic process, the speed at which new cases of disease occur in the population.

**EXAMPLES: Calculating Incidence Rates**

**Example A:** Investigators enrolled 2,100 women in a study and followed them annually for four years to determine the incidence rate of heart disease. After one year, none had a new diagnosis of heart disease, but 100 had been lost to follow-up. After two years, one had a new diagnosis of heart disease, and another 99 had been lost to follow-up. After three years, another seven had new diagnoses of heart disease, and 793 had been lost to follow-up. After four years, another 8 had new diagnoses with heart disease, and 392 more had been lost to follow-up. The study results could also be described as follows: No heart disease was diagnosed at the first year. Heart disease was diagnosed in one woman at the second year, in seven women at the third year, and in eight women at the fourth year of follow-up. One hundred women were lost to follow-up by the first year, another 99 were lost to followup after two years, another 793 were lost to follow-up after three years, and another 392 women were lost to followup after 4 years, leaving 700 women who were followed for four years and remained disease free.

Calculate the incidence rate of heart disease among this cohort. Assume that persons with new diagnoses of heart disease and those lost to follow-up were disease-free for half the year, and thus contribute ½ year to the denominator.

Numerator = number of new cases of heart disease

= 0 + 1 + 7 + 8 = 16

Denominator = person-years of observation

= (2,000 + ½ x 100) + (1,900 + ½ x 1 + ½ x 99) + (1,100 + ½ x 7 + ½ x 793) +

(700 + ½ x 8 + ½ x 392)

= 6,400 person-years of follow-up

or

Denominator = person-years of observation

= (1 x 1.5) + (7 x 2.5) + (8 x 3.5) + (100 x 0.5) + (99 x 1.5) + (793 x 2.5) +

(392 x 3.5) + (700 x 4)

= 6,400 person-years of follow-up

Person-time rate = Number of new cases of disease or injury during specified period

Time each person was observed, totaled for all persons

= 16 / 6,400

= .0025 cases per person-year

= 2.5 cases per 1,000 person-years

In contrast, the incidence proportion can be calculated as 16 / 2,100 = 7.6 cases per 1,000 population during the four-year period, or an average of 1.9 cases per 1,000 per year (7.6 divided by 4 years). The incidence proportion underestimates the true rate because it ignores persons lost to follow-up, and assumes that they remained diseasefree for all four years.

**Example B:** The diabetes follow-up study included 218 diabetic women and 3,823 nondiabetic women. By the end of the study, 72 of the diabetic women and 511 of the nondiabetic women had died. The diabetic women were observed for a total of 1,862 person-years; the nondiabetic women were observed for a total of 36,653 person-years.

Calculate the incidence rates of death for the diabetic and non-diabetic women.

For diabetic women, numerator = 72 and denominator = 1,862

Person-time rate = 72 / 1,862

= 0.0386 deaths per person-year

= 38.6 deaths per 1,000 person-years

For nondiabetic women, numerator = 511 and denominator = 36,653

Person-time rate = 511 / 36,653 = 0.0139 deaths per person-year

= 13.9 deaths per 1,000 person-years

***Prevalence:***

Prevalence, sometimes referred to as **prevalence rate**, is the proportion of persons in a population who have a particular disease or attribute at a specified point in time or over a specified period of time. Prevalence differs from incidence in that prevalence includes all cases, both new and preexisting, in the population at the specified time, whereas incidence is limited to new cases only.

**Point prevalence** refers to the prevalence measured at a particular point in time. It is the proportion of persons with a particular disease or attribute on a particular date.

**Period prevalence** refers to prevalence measured over an interval of time. It is the proportion of persons with a particular disease or attribute at any time during the interval.

***Method for calculating prevalence of disease***

*All new and pre-existing cases during a given time period*

*Population during the same time period x 10n*

***Method for calculating prevalence of an attribute***

*Persons having a particular attribute during a given time period*

*Population during the same time period x 10n*

The value of 10n is usually 1 or 100 for common attributes. The value of 10n might be 1,000, 100,000, or even 1,000,000 for rare attributes and for most diseases.

**EXAMPLE: Calculating Prevalence**

In a survey of 1,150 women who gave birth in Maine in 2000, a total of 468 reported taking a multivitamin at least 4 times a week during the month before becoming pregnant. Calculate the prevalence of frequent multivitamin use in this group.

Numerator = 468 multivitamin users

Denominator = 1,150 women

Prevalence = (468 / 1,150) x 100 = 0.407 x 100 = 40.7%

***Properties and uses of prevalence:***

• Prevalence and incidence are frequently confused. Prevalence refers to proportion of persons who *have* a condition at or during a particular time period, whereas incidence refers to the proportion or rate of persons who *develop* a condition during a particular time period. So prevalence and incidence are similar, but prevalence includes new and pre-existing cases whereas incidence includes new cases only. The key difference is in their numerators.

*Numerator of incidence = new cases that occurred during a given time period*

*Numerator of prevalence = all cases present during a given time period*

• The numerator of an incidence proportion or rate consists only of persons whose illness began during the specified interval.

The numerator for prevalence includes all persons ill from a specified cause during the specified interval **regardless ofwhen the illness began**. It includes not only new cases, but also preexisting cases representing persons who remained ill during some portion of the specified interval.

• Prevalence is based on both incidence and duration of illness. High prevalence of a disease within a population might reflect high incidence or prolonged survival without cure or both. Conversely, low prevalence might indicate low incidence, a rapidly fatal process, or rapid recovery.

• Prevalence rather than incidence is often measured for chronic diseases such as diabetes or osteoarthritis which have long duration and dates of onset that are difficult to pinpoint.

**EXAMPLES: Incidence versus Prevalence**

10 new cases of illness over about 15 months in a population of 20 persons whereby each horizontal line represents one person, The down arrow indicates the date of onset of illness. The solid line represents the duration of illness. The up arrow and the cross represent the date of recovery and date of death, respectively.

**Example A:** Calculate the incidence rate from October 1, 2004, to September 30, 2005, using the midpoint population (population alive on April 1, 2005) as the denominator. Express the rate per 100 population. Incidence rate numerator = number of new cases between October 1 and September 30

= 4 (the other 6 all had onsets before October 1, and are not included)

Incidence rate denominator = April 1 population

= 18 (persons 2 and 8 died before April 1)

Incidence rate = (4 / 18) x 100

= 22 new cases per 100 population

**Example B:** Calculate the point prevalence on April 1, 2005. Point prevalence is the number of persons ill on the date divided by the population on that date. On April 1, seven persons (persons 1, 4, 5, 7, 9, and 10) were ill.

Point prevalence = (7 / 18) x 100

= 38.89%

### Mortality Frequency Measures

***Mortality rate:***

A mortality rate is a measure of the frequency of occurrence of death in a defined population during a specified interval. Morbidity and mortality measures are often the same mathematically; it's just a matter of what you choose to measure, illness or death. The formula for the mortality of a defined population, over a specified period of time, is:

*Deaths occurring during a given time period*

*Size of the population among which the deaths occurred x 10n*

When mortality rates are based on vital statistics (e.g., counts of death certificates), the denominator most commonly used is the size of the population at the middle of the time period. In the United States, values of 1,000 and 100,000 are both used for 10n for most types of mortality rates. Table 3.4 summarizes the formulas of frequently used mortality measures.

***Crude mortality rate (crude death rate):***

The crude mortality rate is the mortality rate from all causes of death for a population. In the United States in 2003, a total of 2,419,921 deaths occurred. The estimated population was 290,809,777. The crude mortality rate in 2003 was, therefore, (2,419,921 / 290,809,777) x 100,000, or 832.1 deaths per 100,000 population.

### *Cause-specific mortality rate:*

The cause-specific mortality rate is the mortality rate from a specified cause for a population. The numerator is the number of  deaths attributed to a specific cause. The denominator remains the size of the population at the midpoint of the time period. The fraction is usually expressed per 100,000 population. In the United States in 2003, a total of 108,256 deaths were attributed to accidents (unintentional injuries), yielding a cause-specific mortality rate of 37.2 per 100,000 population.8

### *Age-specific mortality rate:*

An age-specific mortality rate is a mortality rate limited to a particular age group. The numerator is the number of deaths in that age group; the denominator is the number of persons in that age group in the population. In the United States in 2003, a total of

130,761 deaths occurred among persons aged 25-44 years, or an age-specific mortality rate of 153.0 per 100,000 25–44 year olds. Some specific types of age-specific mortality rates are neonatal,

postneonatal, and infant mortality rates, as described in the following sections.

### *Infant mortality rate:*

The infant mortality rate is perhaps the most commonly used measure for comparing health status among nations. It is calculated as follows:

*Number of deaths among children < 1 year of age reported during a given time period*

*Number of live births reported during the same time period x 1,000*

The infant mortality rate is generally calculated on an annual basis.

It is a widely used measure of health status because it reflects the health of the mother and infant during pregnancy and the year thereafter. The health of the mother and infant, in turn, reflects a wide variety of factors, including access to prenatal care, prevalence of prenatal maternal health behaviors (such as alcohol or tobacco use and proper nutrition during pregnancy, etc.), postnatal care and behaviors (including childhood immunizations and proper nutrition), sanitation, and infection control.

Is the infant mortality rate a ratio? Yes. Is it a proportion? No, because some of the deaths in the numerator were among children born the previous year. Consider the infant mortality rate in 2003.

That year, 28,025 infants died and 4,089,950 children were born, for an infant mortality rate of 6.951 per 1,000.8 Undoubtedly, some of the deaths in 2003 occurred among children born in 2002, butthe denominator includes only children born in 2003.

Is the infant mortality rate truly a rate? No, because the denominator is not the size of the mid-year population of children < 1 year of age in 2003. In fact, the age-specific death rate for children < 1 year of age for 2003 was 694.7 per 100,000.8

Obviously the infant mortality rate and the age-specific death rate for infants are very similar (695.1 versus 694.7 per 100,000) and close enough for most purposes. They are not exactly the same, however, because the estimated number of infants residing in the

United States on July 1, 2003 was slightly larger than the number of children born in the United States in 2002, presumably because of immigration.

***Neonatal mortality rate:***

The neonatal period covers birth up to but not including 28 days. The numerator of the neonatal mortality rate therefore is the number of deaths among children under 28 days of age during a given time period. The denominator of the neonatal mortality rate, like that of the infant mortality rate, is the number of live births reported during the same time period. The neonatal mortality rate is usually expressed per 1,000 live births. In 2003, the neonatal mortality rate in the United States was 4.7 per 1,000 live births.8

***Postneonatal mortality rate:***

The postneonatal period is defined as the period from 28 days of age up to but not including 1 year of age. The numerator of the postneonatal mortality rate therefore is the number of deaths among children from 28 days up to but not including 1 year of age during a given time period. The denominator is the number of live births reported during the same time period. The postneonatal mortality rate is usually expressed per 1,000 live births. In 2003, the postneonatal mortality rate in the United States was 2.3 per

1,000 live births

***Maternal mortality rate:***

The maternal mortality rate is really a ratio used to measure mortality associated with pregnancy. The numerator is the number of deaths during a given time period among women while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and the site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes. The denominator is the number of live births reported during the same time period. Maternal mortality rate is usually expressed per 100,000 live births. In 2003, the U.S. maternal mortality rate was 8.9 per 100,000 live births.

### *Sex-specific mortality rate:*

A sex-specific mortality rate is a mortality rate among either males or females. Both numerator and denominator are limited to the one sex.

### *Race-specific mortality rate:*

A race-specific mortality rate is a mortality rate related to a specified racial group. Both numerator and denominator are limited to the specified race.

### *Combinations of specific mortality rates:*

Mortality rates can be further stratified by combinations of cause, age, sex, and/or race. For example, in 2002, the death rate from diseases of the heart among women ages 45–54 years was 50.6 per

100,000.9 The death rate from diseases of the heart among men in the same age group was 138.4 per 100,000, or more than 2.5 times as high as the comparable rate for women. These rates are a cause-, age-, and sex-specific rates, because they refer to one cause (diseases of the heart), one age group (45–54 years), and one sex (female or male).

### EXAMPLE: Calculating Mortality Rates

The number of deaths from all causes and from accidents (unintentional injuries) by age group in the United States in 2002. Review the following rates. Determine what to call each one, then calculate it using the data provided. a. Unintentional-injury-specific mortality rate for the entire population

This is a cause-specific mortality rate.

Rate = number of unintentional injury deaths in the entire population x 100,000 estimated midyear population

= (106,742 / 288,357,000) x 100,000

= 37.0 unintentional-injury-related deaths per 100,000 population

b. All-cause mortality rate for 25–34 year olds

This is an age-specific mortality rate.

Rate = number of deaths from all causes among 25–34 year olds x 100,000 estimated midyear population of 25–34 year olds

= (41,355 / 39,928,000) x 100,000

= 103.6 deaths per 100,000 25–34 year olds

c. All-cause mortality among males

This is a sex-specific mortality rate.

Rate = number of deaths from all causes among males x 100,000 estimated midyear population of males

= (1,199,264 / 141,656,000) x 100,000

= 846.6 deaths per 100,000 males

d. Unintentional-injury-specific mortality among 25- to 34-year-old males

This is a cause-specific, age-specific, and sex-specific mortality rate

Rate = number of unintentional injury deaths among 25–34 year old males x 100,000 estimated midyear population of 25–34 year old males

= (9,635 / 20,203,000) x 100,000

= 47.7 unintentional-injury-related deaths per 100,000 25–34 year olds

**Age-adjusted mortality rate**: a mortality ratestatistically modified toeliminate the effect ofdifferent age distributionsin the differentpopulations.Mortality rates can be used to compare the rates in one area with the rates in another area, or to compare rates over time. However, because mortality rates obviously increase with age, a higher mortality rate among one population than among another might simply reflect the fact that the first population is older than the second.

Consider that the mortality rates in 2002 for the states of Alaska and Florida were 472.2 and 1,005.7 per 100,000, respectively. Should everyone from Florida move to Alaska to reduce their risk of death? No, the reason that Alaska’s mortality rate is so much lower than Florida’s is that Alaska’s population is considerably younger. Indeed, for seven age groups, the age-specific mortality rates in Alaska are actually higher than Florida’s.

To eliminate the distortion caused by different underlying age distributions in different populations, statistical techniques are used to adjust or standardize the rates among the populations to be compared. These techniques take a weighted average of the agespecific mortality rates, and eliminate the effect of different age distributions among the different populations. Mortality rates computed with these techniques are **age-adjusted** or **age-standardized mortality rates**. Alaska's 2002 age-adjusted mortality rate (794.1 per 100,000) was higher than Florida's (787.8 per 100,000), which is not surprising given that 7 of 13 agespecific mortality rates were higher in Alaska than Florida.

### ***Death-to-case ratio:***

The death-to-case ratio is the number of deaths attributed to a particular disease during a specified time period divided by the number of new cases of that disease identified during the same time period. The death-to-case ratio is a ratio but not necessarily a proportion, because some of the deaths that are counted in the numerator might have occurred among persons who developed disease in an earlier period, and are therefore not counted in the denominator.

### **EXAMPLE: Calculating Death-to-Case Ratios**

Between 1940 and 1949, a total of 143,497 incident cases of diphtheria were reported. During the same decade, 11,228 deaths were attributed to diphtheria. Calculate the death-to-case ratio.

Death-to-case ratio = 11,228 / 143,497 x 1 = 0.0783 or

= 11,228 / 143,497 x 100 = 7.83 per 100

### ***Case-fatality rate:***

The case-fatality rate is the proportion of persons with a particular condition (cases) who die from that condition. It is a measure of the severity of the condition. The formula is:

*Number of cause-specific deaths among the incident cases*

*Number of incident cases x 10n*

The case-fatality rate is a proportion, so the numerator is restricted to deaths among people included in the denominator. The time periods for the numerator and the denominator do not need to be the same; the denominator could be cases of HIV/AIDS diagnosed during the calendar year 1990, and the numerator, deaths among those diagnosed with HIV in 1990, could be from 1990 to the present.

### **EXAMPLE: Calculating Case-Fatality Rates**

In an epidemic of hepatitis A traced to green onions from a restaurant, 555 cases were identified. Three of the casepatients died as a result of their infections. Calculate the case-fatality rate.

Case-fatality rate = (3 / 555) x 100 = 0.5%

The case-fatality rate is a proportion, not a true rate. As a result, some epidemiologists prefer the term **case-fatality ratio**. The concept behind the case-fatality rate and the death-to-case ratio is similar, but the formulations are different. The death-tocase ratio is simply the number of cause-specific deaths that occurred during a specified time divided by the number of new cases of that disease that occurred during the same time. The deaths included in the numerator of the death-to-case ratio are not restricted to the new cases in the denominator; in fact, for many diseases, the deaths are among persons whose onset of disease was years earlier. In contrast, in the case-fatality rate, the deaths included in the numerator are restricted to the cases in the denominator.

### ***Proportionate mortality:***

Proportionate mortality describes the proportion of deaths in a specified population over a period of time attributable to different causes. Each cause is expressed as a percentage of all deaths, and the sum of the causes must add to 100%. These proportions are not mortality rates, because the denominator is all deaths rather than the population in which the deaths occurred.

***Method for calculating proportionate mortality***

For a specified population over a specified period,

*Deaths caused by a particular cause*

*Deaths from all causes x 100*

Sometimes, particularly in occupational epidemiology, proportionate mortality is used to compare deaths in a population of interest (say, a workplace) with the proportionate mortality in the broader population. This comparison of two proportionate mortalities is called a **proportionate mortality ratio**, or PMR for short. A PMR greater than 1.0 indicates that a particular cause accounts for a greater proportion of deaths in the population of interest than you might expect. For example, construction workers may be more likely to die of injuries than the general population. However, PMRs can be misleading, because they are not based on mortality rates. A low cause-specific mortality rate in the population of interest can elevate the proportionate mortalities for all of the other causes, because they must add up to 100%. Those workers with a high injury-related proportionate mortality very likely have lower proportionate mortalities for chronic or disabling conditions that keep people out of the workforce. In other words, people who work are more likely to be healthier than the population as a whole — this is known as the healthy worker effect.

### *Years of potential life lost:*

Years of potential life lost (YPLL) is one measure of the impact of premature mortality on a population. Additional measures incorporate disability and other measures of quality of life. YPLL is calculated as the sum of the differences between a predetermined end point and the ages of death for those who died before that end point. The two most commonly used end points are age 65 years and average life expectancy.

The use of YPLL is affected by this calculation, which implies a value system in which more weight is given to a death when it occurs at an earlier age. Thus, deaths at older ages are "devalued." However, the YPLL before age 65 (YPLL65) places much more emphasis on deaths at early ages than does YPLL based on remaining life expectancy (YPLLLE). In 2000, the remaining life expectancy was 21.6 years for a 60-year-old, 11.3 years for a 70- year-old, and 8.6 for an 80-year-old. YPLL65 is based on the fewer than 30% of deaths that occur among persons younger than 65. In contrast, YPLL for life expectancy (YPLLLE) is based on deaths among persons of all ages, so it more closely resembles crude mortality rates. YPLL rates can be used to compare YPLL among populations of different sizes. Because different populations may also have different age distributions, YPLL rates are usually age-adjusted to eliminate the effect of differing age distributions.

### *Method for calculating YPLL from a line listing:*

**Step 1.**Decide on end point (65 years, average life expectancy, or other).

**Step 2.**Exclude records of all persons who died at or after the end point.

**Step 3.**For each person who died before the end point, calculate that person's YPLL by subtracting the age at death from the end point. YPLLindividual = end point – age at death

**Step 4.**Sum the individual YPLLs. YPLL = .YPLLindividual

### *Method for calculating YPLL from a frequency:*

**Step 1.**Ensure that age groups break at the identified end point

(e.g., 65 years). Eliminate all age groups older than the endpoint.

**Step 2.**For each age group younger than the end point, identify the midpoint of the age group, where midpoint = age group's youngest age in years + oldest age + 12

**Step 3.**For each age group younger than the end point, identify that age group's YPLL by subtracting the midpoint from the end point.

**Step 4**. Calculate age-specific YPLL by multiplying the age group's YPLL times the number of persons in that age group.

**Step 5.**Sum the age-specific YPLL's.

The **YPLL rate** represents years of potential life lost per 1,000 population below the end-point age, such as 65 years. YPLL rates should be used to compare premature mortality in different populations, because YPLL does not take into account differences in population sizes.

### Natality (Birth) Measures

Natality measures are population-based measures of birth. These measures are used primarily by persons working in the field of maternal and child health. Table 3.11 includes some of the commonly used measures of natality.

### Measures of Association

The key to epidemiologic analysis is comparison. Occasionally you might observe an incidence rate among a population that seems high and wonder whether it is actually higher than what should be expected based on, say, the incidence rates in other communities. Or, you might observe that, among a group of casepatients in an outbreak, several report having eaten at a particular restaurant. Is the restaurant just a popular one, or have more casepatients eaten there than would be expected? The way to address that concern is by comparing the observed group with another group that represents the expected level.

A measure of association quantifies the relationship between exposure and disease among the two groups. Exposure is used loosely to mean not only exposure to foods, mosquitoes, a partner with a sexually transmissible disease, or a toxic waste dump, but also inherent characteristics of persons (for example, age, race, sex), biologic characteristics (immune status), acquired characteristics (marital status), activities (occupation, leisure activities), or conditions under which they live (socioeconomic status or access to medical care). The measures of association described in the following section compare disease occurrence among one group with disease occurrence in another group. Examples of measures of association include risk ratio (relative risk), rate ratio, odds ratio, and proportionate mortality ratio.

#### *Risk ratio:*

A risk ratio (RR), also called relative risk, compares the risk of a health event (disease, injury, risk factor, or death) among one group with the risk among another group. It does so by dividing the risk (incidence proportion, attack rate) in group 1 by the risk (incidence proportion, attack rate) in group 2 . The two groups are typically differentiated by such demographic factors as sex (e.g., males versus females) or by exposure to a suspected risk factor (e.g., did or did not eat potato salad). Often, the group of primary interest is labeled the exposed group, and the comparison group is labeled the unexposed group.

***Method for Calculating risk ratio:***

The formula for risk ratio (RR) is:

*Risk of disease (incidence proportion, attack rate) in group of primary interest*

*Risk of disease (incidence proportion, attack rate) in comparison group*

A risk ratio of 1.0 indicates identical risk among the two groups. A risk ratio greater than 1.0 indicates an increased risk for the group in the numerator, usually the exposed group. A risk ratio less than 1.0 indicates a decreased risk for the exposed group, indicating that perhaps exposure actually protects against disease occurrence.

**EXAMPLES: Calculating Risk Ratios**

**Example A:** In an outbreak of tuberculosis among prison inmates in South Carolina in 1999, 28 of 157 inmates residing on the East wing of the dormitory developed tuberculosis, compared with 4 of 137 inmates residing on the West wing. These data are summarized in the two-by-two table so called because it has two rows for the exposure and two columns for the outcome. Here is the general format and notation.

**Example B:** In an outbreak of varicella (chickenpox) in Oregon in 2002, varicella was diagnosed in 18 of 152 vaccinated children compared with 3 of 7 unvaccinated children. Calculate the risk ratio.

***Rate ratio:***

A rate ratio compares the incidence rates, person-time rates, or mortality rates of two groups. As with the risk ratio, the two groups are typically differentiated by demographic factors or by exposure to a suspected causative agent. The rate for the group of primary interest is divided by the rate for the comparison group.

*Rate for group of primary interest* Rate ratio = *Rate for comparison group*

The interpretation of the value of a rate ratio is similar to that of the risk ratio. That is, a rate ratio of 1.0 indicates equal rates in the two groups, a rate ratio greater than 1.0

indicates an increased risk for the group in the numerator, and a rate ratio less than 1.0 indicates a decreased risk for the group in the numerator.

**EXAMPLE: Calculating Rate Ratios**

Public health officials were called to investigate a perceived increase in visits to ships' infirmaries for acute respiratory illness (ARI) by passengers of cruise ships in Alaska in 1998.13 The officials compared passenger visits to ship infirmaries for ARI during May–August 1998 with the same period in 1997. They recorded 11.6 visits for ARI per

1,000 tourists per week in 1998, compared with 5.3 visits per 1,000 tourists per week in 1997. Calculate the rate ratio.

Rate ratio = 11.6 / 5.3 = 2.2

Passengers on cruise ships in Alaska during May–August 1998 were more than twice as likely to visit their ships' infirmaries for ARI than were passengers in 1997. (Note: Of 58 viral isolates identified from nasal cultures from passengers, most were influenza A, making this the largest summertime influenza outbreak in North America.)

***Odds ratio:***

An odds ratio (OR) is another measure of association that quantifies the relationship between an exposure with two categories and health outcome. The odds ratio is calculated as *a c Odds ratio* = ( *b*)( *d* ) = *ad / bc* where

a = number of persons exposed and with disease

b = number of persons exposed but without disease

c = number of persons unexposed but with disease

d = number of persons unexposed: and without disease

a+c = total number of persons with disease (case-patients)

b+d = total number of persons without disease (controls)

The odds ratio is sometimes called the **cross-product ratio** because the numerator is based on multiplying the value in cell "a"times the value in cell "d," whereas the denominator is the productof cell "b" and cell "c." A line from cell "a" to cell "d" (for thenumerator) and another from cell "b" to cell "c" (for thedenominator) creates an x or cross on the two-by-two table.

**EXAMPLE: Calculating Odds Ratios**

Use the data provided to calculate the risk and odds ratios.

1. Risk ratio

5.0 / 1.0 = 5.0

2. Odds ratio

(100 x 7,920) / (1,900 x 80) = 5.2

Notice that the odds ratio of 5.2 is close to the risk ratio of 5.0. That is one of the attractive features of the odds ratio — when the health outcome is uncommon, the odds ratio provides a reasonable approximation of the risk ratio.

Another attractive feature is that the odds ratio can be calculated with data from a case-control study, whereas neither a risk ratio nor a rate ratio can be calculated.

In a case-control study, investigators enroll a group of case-patients (distributed in cells a and c of the two-by-two table), and a group of non-cases or controls (distributed in cells b and d).

The odds ratio is the measure of choice in a case-control study.

A case-control study is based on enrolling a group of persons with disease ("case-patients") and a comparable group without disease ("controls"). The number of persons in the control group is usually decided by the investigator. Often, the size of the population from which the case-patients came is not known. As a result, risks, rates, risk ratios or rate ratios cannot be calculated from the typical case-control study. However, you can calculate an odds ratio and interpret it as an approximation of the risk ratio, particularly when the disease is uncommon in the population.

## Measures of Public Health Impact

A measure of public health impact is used to place the association between an exposure and an outcome into a meaningful public health context. Whereas a measure of association quantifies the relationship between exposure and disease, and thus begins to provide insight into causal relationships, measures of public health impact reflect the burden that an exposure contributes to the frequency of disease in the population. Two measures of public health impact often used are the attributable proportion and efficacy or effectiveness.

### *Attributable proportion:*

The attributable proportion, also known as the attributable risk percent, is a measure of the public health impact of a causative factor. The calculation of this measure assumes that the occurrence of disease in the unexposed group represents the baseline or expected risk for that disease. It further assumes that if the risk of disease in the exposed group is higher than the risk in the unexposed group, the difference can be attributed to the exposure. Thus, the attributable proportion is the amount of disease in the exposed group attributable to the exposure. It represents the expected reduction in disease if the exposure could be removed (or never existed). Appropriate use of attributable proportion depends on a single risk factor being responsible for a condition. When multiple risk factors may interact (e.g., physical activity and age or health status), this measure may not be appropriate.

### *Method for calculating attributable proportion*

Attributable proportion is calculated as follows:

*Risk for exposed group – risk for unexposed group*

*Risk for exposed group x 100%*

Attributable proportion can be calculated for rates in the same way.

### EXAMPLE: Calculating Attributable Proportion

In another study of smoking and lung cancer, the lung cancer mortality rate among nonsmokers was 0.07 per 1,000 persons per year. The lung cancer mortality rate among persons who smoked 1–14 cigarettes per day was 0.57 lung cancer deaths per 1,000 persons per year. Calculate the attributable proportion.

Attributable proportion = (0.57 – 0.07) / 0.57 x 100% = 87.7%

Given the proven causal relationship between cigarette smoking and lung cancer, and assuming that the groups are comparable in all other ways, one could say that about 88% of the lung cancer among smokers of 1-14 cigarettes per day might be attributable to their smoking. The remaining 12% of the lung cancer cases in this group would have occurred anyway.

### *Vaccine efficacy or vaccine effectiveness:*

Vaccine efficacy and vaccine effectiveness measure the proportionate reduction in cases among vaccinated persons. Vaccine efficacy is used when a study is carried out under ideal conditions, for example, during a clinical trial. Vaccine effectiveness is used when a study is carried out under typical field (that is, less than perfectly controlled) conditions. Vaccine efficacy/effectiveness (VE) is measured by calculating the risk of disease among vaccinated and unvaccinated persons and determining the percentage reduction in risk of disease among vaccinated persons relative to unvaccinated persons. The greater the percentage reduction of illness in the vaccinated group, the greater the vaccine efficacy/effectiveness. The basic formula is written as:

*Risk among unvaccinated group – risk among vaccinated group*

*Risk among unvaccinated group*

OR: *1 – risk ratio*

In the first formula, the numerator (risk among unvaccinated – risk among vaccinated) is sometimes called the risk difference or excess risk. Vaccine efficacy/effectiveness is interpreted as the proportionate reduction in disease among the vaccinated group. So a VE of 90% indicates a 90% reduction in disease occurrence among the vaccinated group, or a 90% reduction from the number of cases you would expect if they have not been vaccinated.

**EXAMPLE: Calculating Vaccine Effectiveness**

Calculate the vaccine effectiveness from the varicella data in Table 3.13.

VE = (42.9 – 11.8) / 42.9 = 31.1 / 42.9 = 72%

Alternatively, VE = 1 – RR = 1 – 0.28 = 72%

So, the vaccinated group experienced 72% fewer varicella cases than they would have if they had not been vaccinated.

**Summary**

Because many of the variables encountered in field epidemiology are nominal-scale variables, frequency measures are used quite commonly in epidemiology. Frequency measures include ratios, proportions, and rates. Ratios and proportions are useful for describing the characteristics of populations. Proportions and rates are used for quantifying morbidity and mortality. These measures allow epidemiologists to infer risk among different groups, detect groups at high risk, and develop hypotheses about causes — that is, why these groups might be at increased risk.

The two primary measures of morbidity are incidence and prevalence.

- **Incidence** rates reflect the occurrence of new disease in a population.

- **Prevalence** reflects the presence of disease in a population.

A variety of **mortality** rates describe deaths among specific groups, particularly by age or sex or by cause. The hallmark of epidemiologic analysis is comparison, such as comparison of observed amount of disease in a population with the expected amount of disease. The comparisons can be quantified by using such measures of association as risk ratios, rate ratios, and odds ratios. These measures provide evidence regarding causal relationships between exposures and disease. Measures of public health impact place the association between an exposure and a disease in a public health context. Two such measures are the attributable proportion and vaccine efficacy.

**Analyzing and Interpreting Data**

After morbidity, mortality, and other relevant data about a health problem have been gathered and compiled, the data should be analyzed by time, place, and person. Different types of data are used for surveillance, and different types of analyses might be needed for each. For example, data on individual cases of disease are analyzed differently than data aggregated from multiple records; data received as text must be sorted, categorized, and coded for statistical analysis; and data from surveys might need to be weighted to produce valid estimates for sampled populations.

For analysis of the majority of surveillance data, descriptive methods are usually appropriate. The display of frequencies (counts) or rates of the health problem in simple tables and graphs, as discussed in Lesson 4, is the most common method of analyzing data for surveillance. Rates are useful — and frequently preferred — for comparing occurrence of disease for different geographic areas or periods because they take into account the size of the opulation from which the cases arose. One critical step before calculating a rate is constructing a denominator from appropriate population data. For state- or countywide rates, general population data are used. These data are available from the U.S. Census Bureau or from a state planning agency.

For other calculations, the population at risk can dictate an alternative denominator. For example, an infant mortality rate uses the number of live-born infants; rates of surgical wound infections in a hospital nrequires the number of such procedures performed. In addition to calculating\ frequencies and rates, more sophisticated methods (e.g., space-time cluster analysis, time series analysis, or computer mapping) can be applied. To determine whether the incidence or prevalence of a health problem has increased, data must be compared either over time or across areas. The selection of data for comparison depends on the health problem under surveillance and what is known about its typical temporal and geographic patterns of occurrence.

For example, data for diseases that indicate a seasonal pattern (e.g., influenza and mosquito-borne diseases) are usually compared with data for the corresponding season from past years. Data for diseases without a seasonal pattern are commonly compared with data for previous weeks, months, or years, depending on the nature of the disease. Surveillance for chronic diseases typically requiresdata covering multiple years. Data for acute infectious diseases might only require data covering weeks or months, although data extending over multiple years can also be helpful in the analysis of the natural history of disease. Data from one geographic area are sometimes compared with data from another area. For example, data from a county might be compared with data from adjacent counties or with data from the state. We now describe common methods for, and provide examples of, the analysis of data by time, place, and person

## Bibliography

**Ewen, R.B. (1988).** ***The workbook for introductory statistics for the behavioral sciences.*****Orlando, FL: Harcourt Brace Jovanovich.**

**Glass, G. (1996, August 26). COE 502:** ***Introduction to quantitative methods.*** **Available:** http://seamonkey.ed.asu.edu/~gene/502/home.html

**Hartwig, F., Dearing, B.E. (1979).*****Exploratory data analysis.*** **Newberry Park, CA: Sage Publications, Inc.**

**Hinkle, Dennis E., Wiersma, W. and Jurs, S.G. (1988).*****Applied statistics for the behavioral sciences.*** **Boston: Houghton.**

**Kleinbaum, David G., Kupper, L.L. and Muller K.E.** ***Applied regression analysis and other multivariable methods2nd ed.*** **Boston: PWS-KENT Publishing Company.**

**Kolstoe, R.H. (1969).*****Introduction to statistics for the behavioral sciences.*** **Homewood, ILL: Dorsey.**

**Levin, J., and James, A.F. (1991).*****Elementary statistics in social research,5th ed.*** **New York: HarperCollins.**

**Liebetrau, A.M. (1983).** ***Measures of association.*** **Newberry Park, CA: Sage Publications, Inc.**

**Mendenhall, W.(1975).** ***Introduction to probability and statistics,4th ed.*** **North Scltuate, MA: Duxbury Press.**

**Moore, David S. (1979).** ***Statistics: Concepts and controversies, 2nd ed.*** **New York: W. H. Freeman and Company.**

**Mosier, C.T. (1997). MG284 Statistics I - notes. Available: http://phoenix.som.clarkson.edu/~cmosier/statistics/main/outline/index.html**

**Runyon, R.P., and Haber, A. (1976).*****Fundamentals of behavioral statistics, 3rd ed.*** **Reading, MA: Addison-Wesley Publishing Company.**

**Schoeninger, D.W. and Insko, C.A. (1971).*****Introductory statistics for the behavioral sciences.*** **Boston: Allyn and Bacon, Inc.**

**Stevens, J. (1986).** ***Applied multivariate statistics for the social sciences.*** **Hillsdale, NJ: Lawrence Erlbaum Associates.**

**Stockberger, D. W. (1996).** ***Introductory statistics: Concepts, models and applications.*** **Available:** http://www.psychstat.smsu.edu/ **[1997, December 8].**