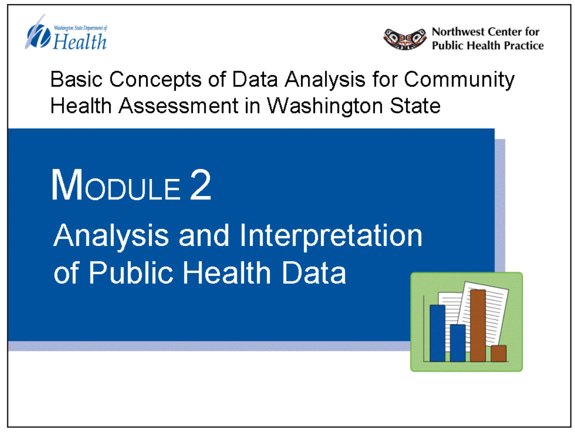
Basic Concepts of Data Analysis for Community Health Assessment

Module 2: Analysis and Interpretation of Public Health Data: Part 1

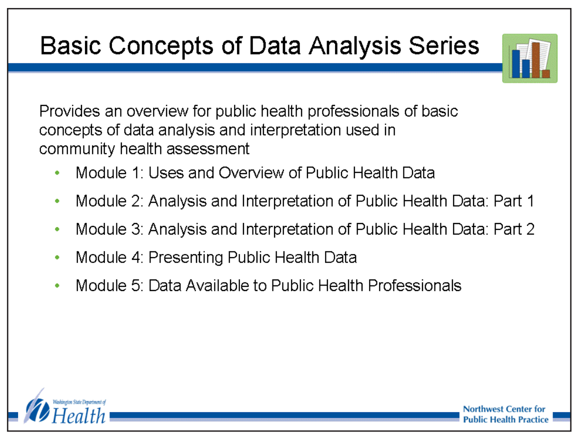
Transcript

Analysis and Interpretation of Public Health Data Part 1



Welcome to Analysis and Interpretation of Public Health Data, the second module in the series on Basic Concepts of Data Analysis for Community Health Assessment in Washington State. I'm Jane Ballard. I helped create this course, along with a team of other community health assessment experts at both local and state health departments in Washington and Oregon. I have a doctorate in epidemiology and manage the Health Statistics and Assessment Program at Snohomish Health District. I have worked at Snohomish Health District since 2000.

Basic Concepts of Data Analysis Series



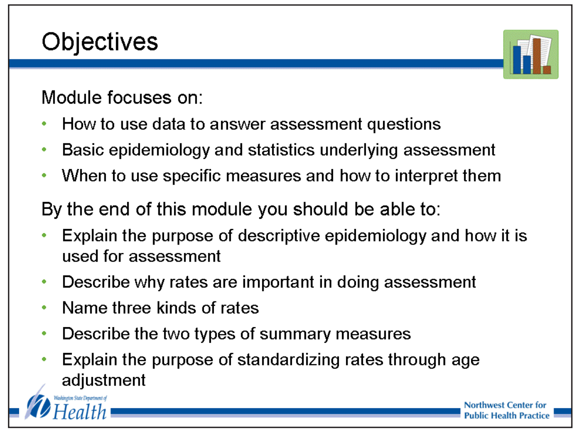
This series provides an overview for public health professionals of the basic concepts of data analysis and interpretation used in community health assessment. The training is intended to help professionals who work in public health practice at state and local agencies hone their assessment skills.

Module 1 provides an overview of public health data sources and uses. Module 2 introduces the analysis and interpretation of public health data. Module 3 continues the discussion of the analysis and interpretation of the public health data. Module 4 provides information on how to present public health data, and module 5 describes data available to public health professionals.

This series was developed by Washington State Department of Health in partnership with the Northwest Center for Public Health Practice. Many of the examples use Washington State or county level data, but the concepts they illustrate are relevant to public health professionals in any location.

Objectives

In this module we'll focus on how to use data to answer assessment questions. We'll also review some of the basic epidemiology and statistics underlying assessment and introduce you to the concepts of some frequently used measures in assessment.

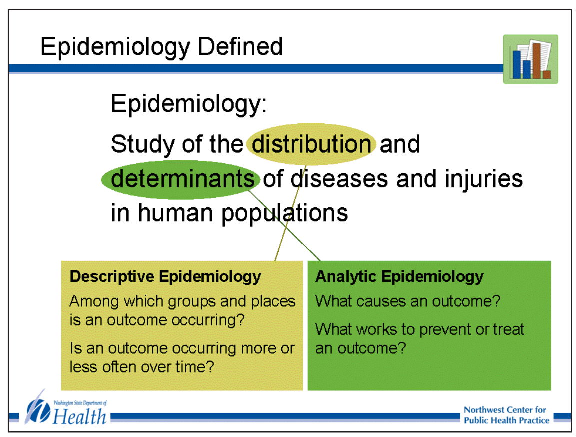


Although we'll describe how to calculate these measures, the main goal of the module is to give you information to help you determine when and how the measures should be used and how to interpret their results.

By the end of this module, you should be able to:

- Explain the purpose of descriptive epidemiology and how it is used for assessment
- Describe why rates are important in doing assessment
- Name three kinds of rates
- Describe the two types of summary measures, and
- Explain the purpose of standardizing rates through age-adjustment.

Epidemiology Defined



Let's start with a basic definition. Epidemiology is the study of the distribution and determinants, or causes, of diseases and injuries (and, in fact, any health event) in human populations. Epidemiology is commonly divided into two branches: descriptive and analytic. These two branches answer different kinds of questions.

Descriptive epidemiology answers such questions as:

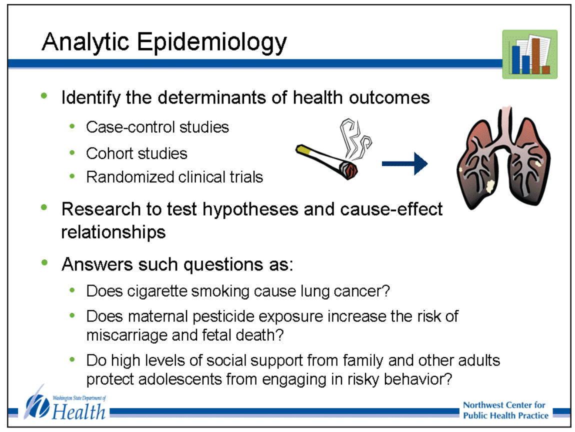
- Among which groups and places is an outcome occurring? and
- Is an outcome occurring more or less often over time?

Analytic epidemiology answers such questions as:

- What are the causes or associations of exposure and outcome? and
- What works to prevent or treat an outcome?

Analytic Epidemiology

Generally in the field of public health, researchers use analytic epidemiology since their focus is on identifying the causes, determinants, or relationships of exposures and health outcomes. Using case-control and cohort studies and randomized clinical trials to test hypotheses about cause and effect relationships, researchers can answer such questions as:

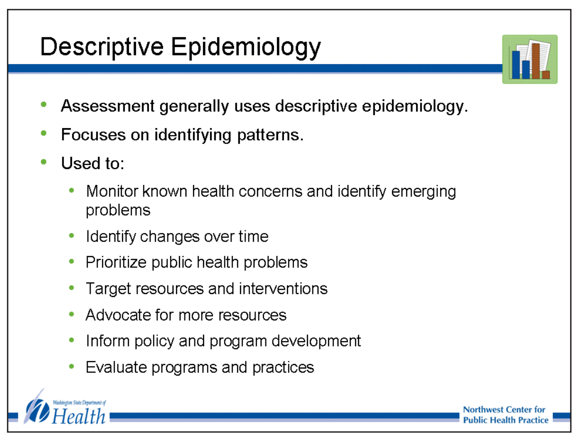


- Does cigarette smoking cause lung cancer?
- Does maternal pesticide exposure increase the risk of miscarriage and fetal death?
- Do high levels of social support from family and other adults protect adolescents from engaging in risky behavior?

The work of community health assessment usually does not involve analytic epidemiology, so this module will not cover the measures used in analytic epidemiology studies.

You can find links to information about analytic epidemiology measures in the Resources at the end of this module.

Descriptive Epidemiology



As I said, the assessment work done in public health generally uses descriptive epidemiology, which focuses on patterns rather than on questions of cause and effect.

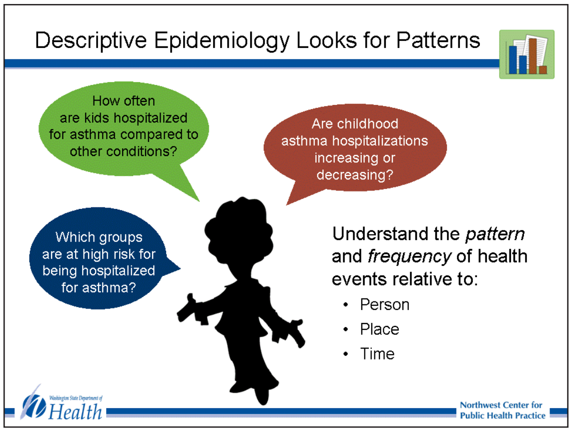
Descriptive epidemiology is used to:

- Monitor known health concerns and identify emerging problems so we can intervene as early as possible
- Identify changes over time
- Prioritize public health problems so we can focus our efforts where they'll have the greatest effect
- Target resources and interventions
- Advocate for more resources
- Inform policy and program development, and
- Evaluate the effectiveness of our programs and practices.

Descriptive Epidemiology Looks for Patterns

Descriptive epidemiology looks for patterns and the frequency of health events. Let's say you work in a program area concerned with childhood asthma. Some of the things you might want to know are:

- How often are children hospitalized for asthma compared to other conditions?



- Which groups are at highest risk for being hospitalized for asthma and where do they live? and
- Are childhood asthma hospitalizations increasing or decreasing?

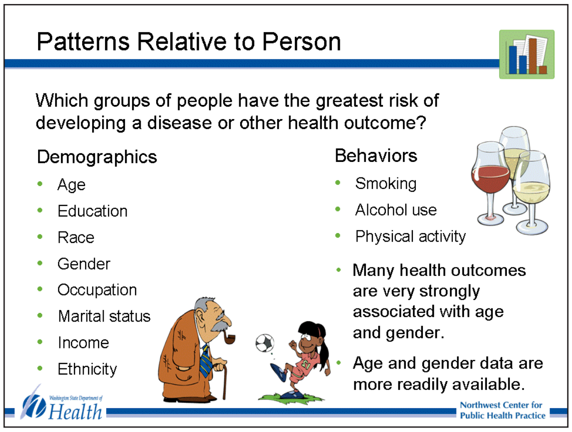
To answer these questions you would use descriptive epidemiology to look for the pattern and frequency of childhood asthma relative to person, place, and time:

- Characteristics of the children who are hospitalized
- ZIP code concentrations of asthmatic children, and
- Is the rate of hospitalization increasing or decreasing or is there a seasonal variation?

Let's pause now so you can check your understanding of the different types of epidemiology. Please note that the exercises may take a moment to load.

Practice: Types of Epidemiology

Patterns Relative to Person



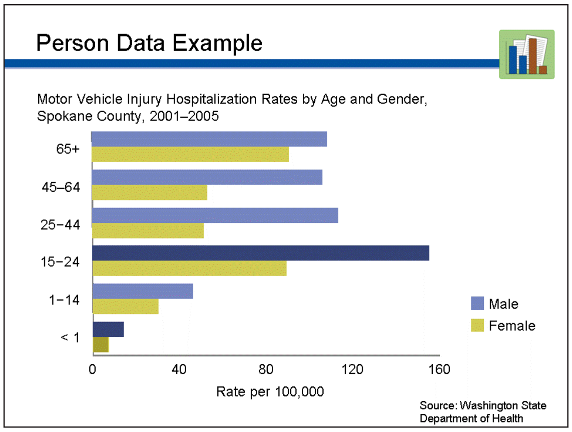
When we look at patterns relative to persons, we're asking, "Which groups of people have the greatest risk of developing a disease or experiencing a health outcome?" Some of the common characteristics used to group people are: age, education, race, gender, occupation, marital status, income or poverty, and ethnicity.

Most of these characteristics are considered to be demographic information, but we can also look at patterns in people's behavior—such as smoking, alcohol use, or even physical activity.

We categorize people most commonly by age and gender for two reasons. First, many health outcomes are very strongly associated with age and gender. And second, this kind of demographic information tends to be more readily available than other kinds of characteristics.

For many of these characteristics, we use different strata, or categories, depending on the purpose of our analysis. For example, age strata for children will be defined differently if the topic is teen pregnancy than if asthma or immunization rates are being studied.

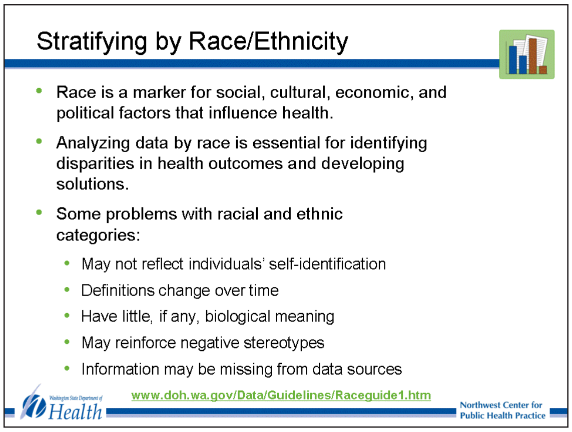
Person Data Example



Let's look at an issue to be aware of when considering data on persons. Here we see data on motor vehicle injury hospitalization rates by age in Spokane County, averaged over a five-year period. The pattern shows two age groups that have high hospitalization rates—the 15- to 24-year-old age group and the 65 and older age group.

Here are the same data stratified by both age and gender. This graph reveals an even more pronounced pattern with higher rates for males in every age group and, in particular, the 15- to 24-year-old group.

Stratifying by Race/Ethnicity



Another characteristic commonly used to stratify people is race or ethnicity. We use this category for a number of reasons.

Race is a marker for social, cultural, economic, and political factors that influence health, and analyzing data by race is essential when we want to identify racial or ethnic disparities in health outcomes and develop strategies for addressing the social and economic inequities that underlie them.

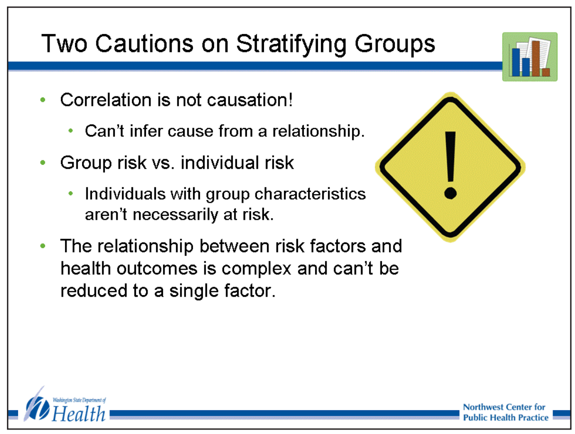
As with age and gender, caution should also be used when stratifying by race and ethnicity. For one thing, racial categories that people are forced to choose when filling out a form may not reflect an individual's self-identification. And for another thing, the categories change over time. For example, it wasn't until the 2000 U.S. Census that people were offered the option for choosing multiple races.

Also, keep in mind that racial categories are social and historical concepts with little—if any—biological meaning. By presenting health data by race—especially data on individual risk behavior—we may reinforce negative stereotypes and stigmatize communities of color.

Finally, information on race and ethnicity is frequently missing or incomplete in some data sources. Small numbers may also be a concern when analyzing data by race.

For these reasons, investigators should clearly understand their rationale for using race and ethnicity as factors in their analysis. Public health professionals in Washington State can refer to the guidelines on using race and ethnicity posted on the Department of Health Web site.

Two Cautions on Stratifying Groups

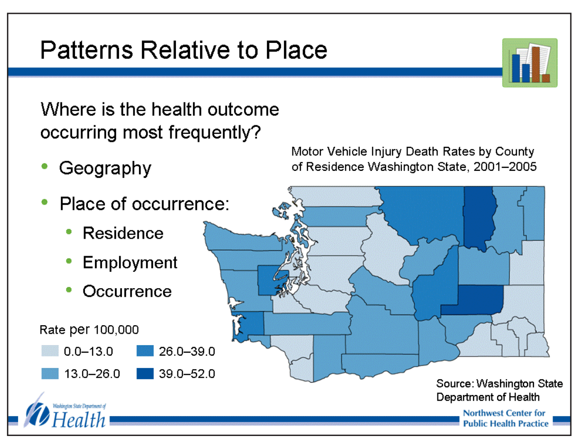


I'd like to emphasize two cautions about patterns relative to person. You've probably heard that "correlation, or association, is not causation." Remember that we're looking for patterns—not for causes. So, for example, even though you may see a pattern in obesity rates by income that suggests a relationship between poverty and obesity, you cannot infer that being obese causes poverty, nor that being poor causes obesity.

Also, keep in mind the difference between group risk and individual risk. You may identify a particular group of people with certain characteristics who have a much higher rate of disease. But that doesn't mean that each individual who has those characteristics is at higher risk of experiencing that disease.

For example, within a group of people who smoke, not all of them will develop lung cancer. This is because the relationship between risk factors and health outcomes is complex and can't be reduced to a single factor.

Patterns Relative to Place



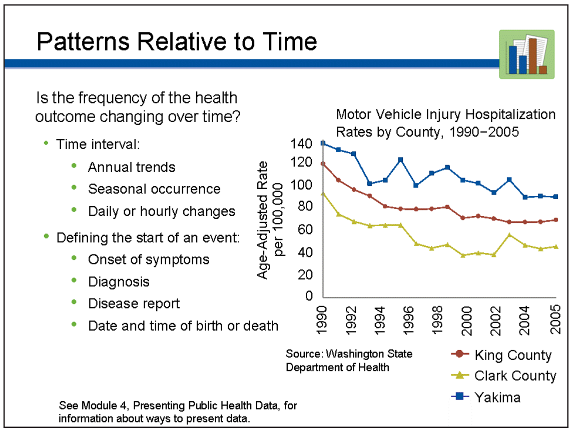
Patterns relative to place indicate the geographic or spatial distribution of disease. They answer the question "Where is the health outcome occurring most frequently?" Different geographic levels are used to answer that question: state, county, and sub-county areas such as ZIP code areas or census tracts. Geographic areas can also be grouped by specific characteristics, such as rural and urban.

Location of occurrence can mean several different things. Usually with health outcome data, location means the residence of the person experiencing the health event, because we're interested in the health status of our county or state population. However, it could mean place of employment if the topic is work-related injuries. Or, it could mean the location where something happened, such as a car accident or a food-borne disease outbreak.

Often, place is defined in terms of traditional political boundaries such as state, county, or municipality. Although communities may not have specific spatial boundaries, they can be loosely defined by groups of census tracts or ZIP codes. Mapping your data can provide a quick visual overview, as in this map of motor vehicle deaths over a five-year period by county. One note of caution, though: the State Center for Health Statistics provides tables for injuries by county of occurrence and county of residence. It is important that you know which table you are using and which location the data represent.

In this map, the relative difference in motor vehicle injury death rate is shown across the state, but we don't know if it represents where the injury happened or where the driver lived, based on the graphic alone. You need to understand your data and know what it represents. In this case the data represent motor vehicle deaths by county of residence of the person who died.

Patterns Relative to Time



We look at patterns in time to answer the question "Is the frequency of the health outcome changing—either, increasing or decreasing—over time?" Usually, frequency is examined by annual or even seasonal trends. But during an outbreak, daily or hourly changes in the occurrence of disease may be important.

Patterns according to time are best visualized by trend charts. Here is a trend chart that shows seasonal variations in the incidence of *E. coli* cases in Washington State. And here is another chart showing motor vehicle injury hospitalization rates in three counties from 1990 to 2005. The rates appear to be decreasing in all three locations.

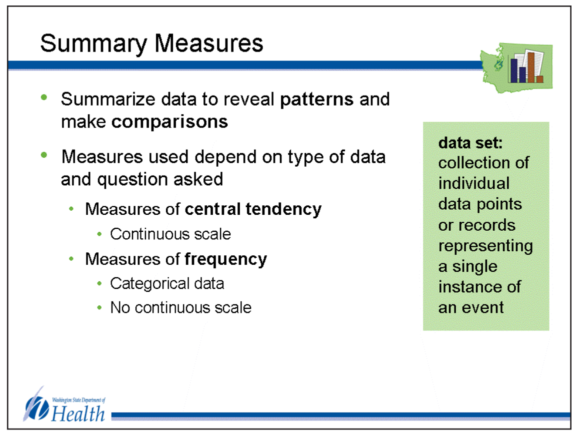
Different criteria can be used to determine when an event occurs. For example, a communicable disease can be dated according to when symptoms first presented, when the disease was diagnosed, or when it was reported to public health authorities. Events such as birth and death have both a date and time of occurrence.

By the way, module four of this series looks in more detail at various ways to present data.

Let's pause now so you can answer a question about the material we've just covered.

Practice: Person, Place, Time

Summary Measures



So far, we've looked at data related to person, place, and time. But a data set is a just collection of individual data points or records, each of which represents a single person or occurrence of some event. We need to summarize those data points in order to see their patterns and make comparisons.

We summarize the data points by calculating what we broadly refer to as summary measures. Which measures we use depend on the type of data we're using and the purpose of the questions we're asking.

There are two types of summary measures—measures of "central tendency" and measures of "frequency." Central tendency refers to the data that can be measured on a continuous scale, such as age in years or the Injury Severity Score.

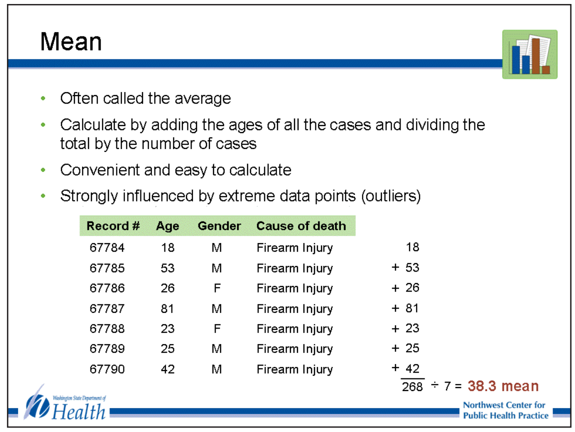
Measures of frequency usually summarize data that are categorical in nature and cannot be measured on a continuous scale, for example, births, deaths, or disease. However, continuous data can be also aggregated for summarization, for example, the age group "0–17 years" or the income category "less than $50,000 a year."

Let's pause so you can check your understanding of these concepts and then we'll look at two measures of central tendency (in other words, continuous data), the mean and the median.

Practice: Summary Measures

Mean

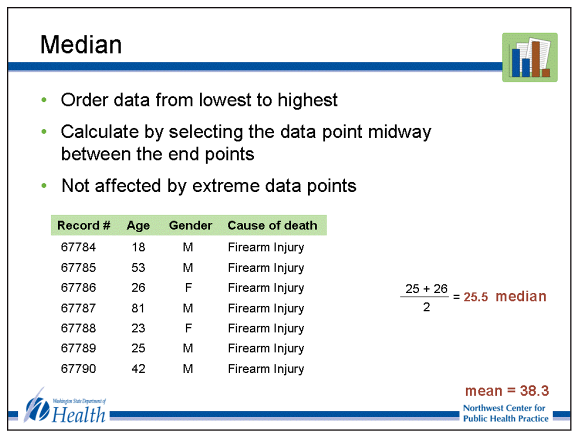
The mean is often called the average. In this example, we're interested in the mean or average age at death from firearm injury. We calculate the mean by adding the ages of all the cases and dividing the total by the number of cases. We get a mean age of victims from firearm injury deaths of 38.3.The mean is convenient and easy to calculate but it is strongly influenced by extreme data points, or in technical terms, outliers. In this example, the 81 year old is an outlier and affects the mean by inflating it. You can see that if we remove the 81 year old, the mean, or average, is quite a bit lower than when we include it.



Median

Because the mean is affected by extreme points, it's often preferable to use the median.

First we must order the data from lowest to highest. Then we calculate the median by selecting the data point, or case, that lies midway between the end points so that there are an equal number of cases on either side of the number. When you have an even number of cases, the median is the average of the two middle cases. In this example we want to know the median age of the people who died by firearms. As you can see, the median age is much lower than the mean age and isn't affected by extreme data points. In fact, if we remove the outlier, the median isn't particularly different.



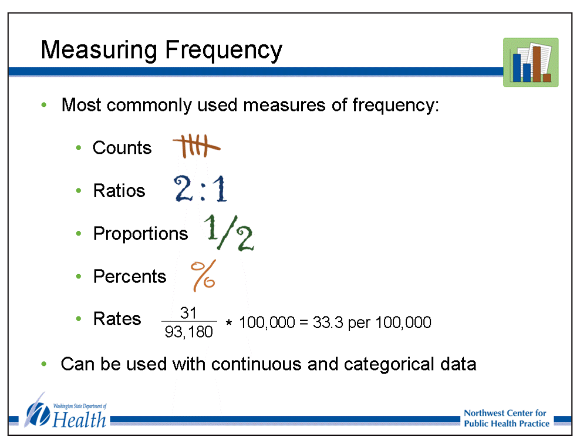
Let's pause so you can answer a question about what we've just covered.

Practice: Median

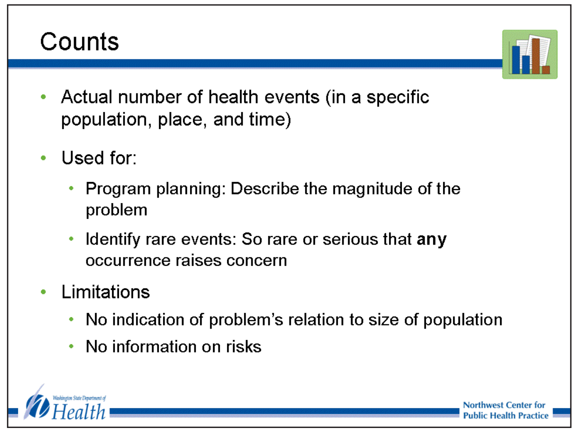
Measuring Frequency

As I mentioned earlier, the two types of summary measures are measures of central tendency and measures of frequency.

In community health assessment practice, we're usually interested in looking at the frequency of health events. The most commonly used measures for assessing frequency are counts, ratios, proportions, percents, and rates. These measures can be used with both continuous and categorical data.



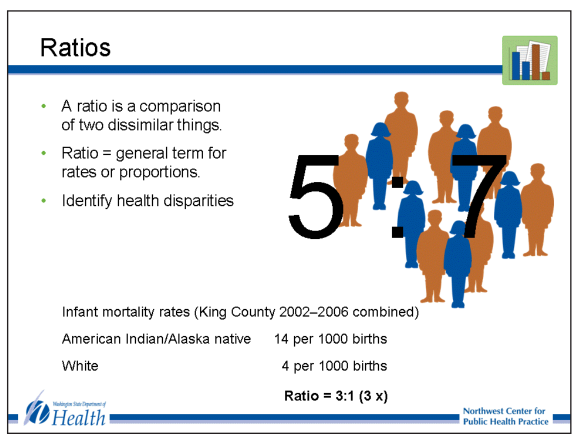
Counts



Counts are the simplest measure of frequency. A count is just the actual number of health events in the population of interest. Counts are commonly used for program planning (to gauge the magnitude of the problem) and for the identification of rare (or in technical terms, sentinel) and severe events. These rare or severe events should trigger action even though the number of cases is small (for example, the first case, in a specific area, of West Nile virus or measles).

The use of only counts has several limitations, however. Counting events or cases does not give any indication of the problem in relation to the size of the population. For example, 10 cases of influenza in a school population of 50 students is a greater public health concern than 10 similar cases in a student population of 500. And also, the use of counts alone does not provide any information on the risks associated with having the event or disease.

Ratios

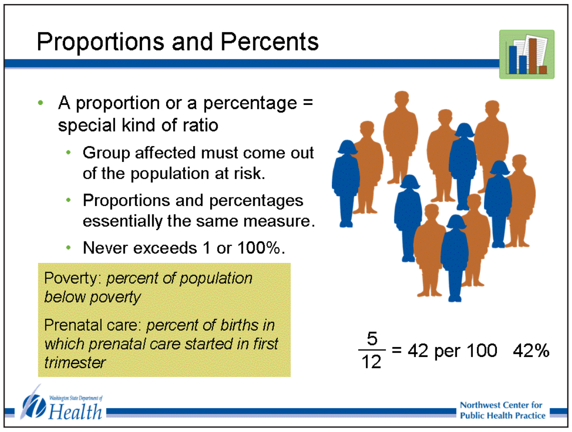


A ratio compares two dissimilar things by dividing one quantity by another. For example, say a group of people has 5 women and 7 men. The ratio of women to men is 5 to 7. A ratio is also a general term for rates or proportions.

Ratios are used in public health to identify disparities. An example of this use would be the analysis of infant mortality by race. In King County for the combined years 2002 through 2006 infant mortality among American Indians and Alaskan Natives was 14 per 1000 births whereas it was 4 per 1000 births for Whites. This is a ratio of 3:1 or 3 times the infant mortality for American Indians and Alaskan Natives compared to Whites in King County.

Proportions and Percents

In our group of men and women, we can also express the relationship of women to the entire group as a proportion, or a percentage, of the total number of people. A proportion or a percentage is a kind of ratio in which the part of the group affected must come out of the population at risk. In mathematical terms, we would say that



the numerator must be drawn from the denominator. For example, our group contains 12 people. The proportion of women in the group is 5 divided by 12. We can also express this fraction as 42 per 100, or 42%.

As you can see, proportions and percents are essentially the same measure. By the way, the value of a proportion never exceeds 1, or 100 percent.

An example of percents or proportions is a poverty measure such as the percent of population below the designated level of poverty income. Another example is a prenatal care measure, for example, the percent of births in which prenatal care started in the first trimester.

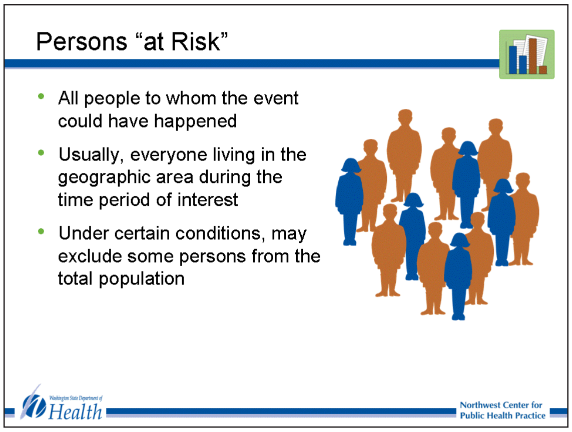
Let's pause so you can answer some questions on what we've just covered.

Practice: Counts, Proportions, Ratios

Practice: Percents

Practice: Limitations of Counts

Persons "at Risk"

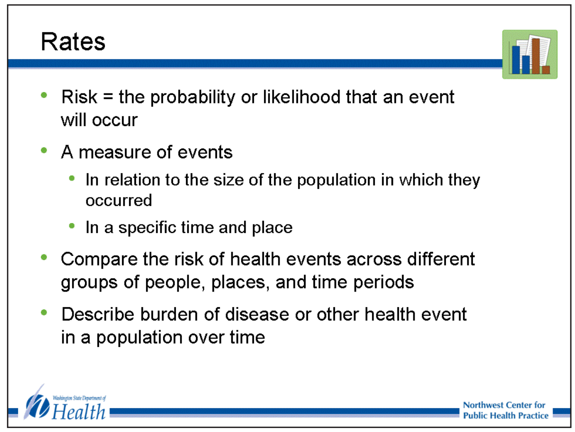


Before we go on to look at rates, I want to discuss a term you'll often hear in assessing community health: "Persons at risk." Persons at risk are the people in the total population, or in other words, the denominator, to whom the event or health condition being studied could have happened. Usually this is everyone living in a specific geographic area during the time period of interest. Under certain conditions, however, we exclude some persons from the total population. For example, when looking at prostate cancer, we would exclude women.

Take a moment to check your understanding of this concept.

Practice: Persons at Risk

Rates

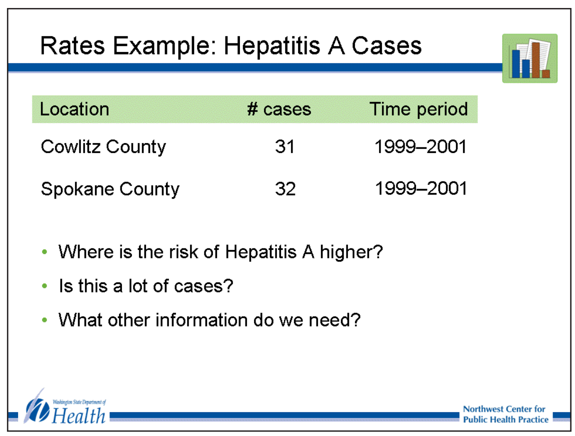


We generally use rates to estimate risk. Risk, by the way, as used by epidemiologists, is the probability or likelihood that an event will occur, for example, the likelihood that an individual will become ill or die, within a stated period of time.

Rates measure events in relation to the size of the population in which they occurred in a specific time and place, or in other words, the population at risk. Rates allow us to compare the risk of health events across different groups of people, places, and times. And rates help us to describe the burden of disease or other health events in a population over time.

Let's look at an example to help clarify the concept of rates.

Rates Example: Hepatitis A Cases



Here's a table showing Hepatitis A cases in two Washington counties from 1999 through 2001. There are 31 cases in Cowlitz County and 32 cases in Spokane County.

Some of the questions we might ask are:

- Where is the risk of Hepatitis A higher?
- Is this a lot of cases?
- What other information do we need?

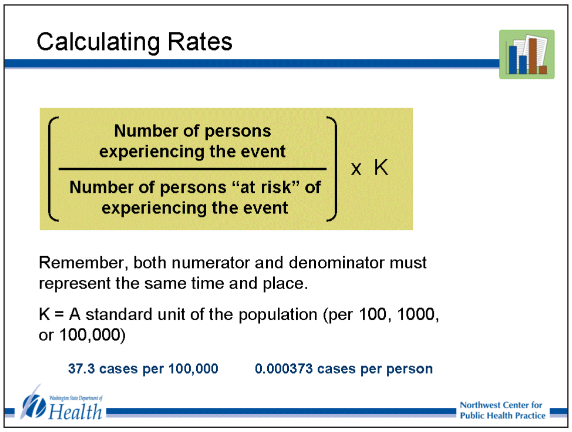
Although the absolute count of Hepatitis A cases is higher in Spokane, what we really want to know is if there are disproportionately more cases in Spokane County than in Cowlitz County. To answer that question, we need to determine the total population at risk and then calculate the rates for each county.

Calculating Rates

You calculate a rate by taking the number of persons experiencing an event—such as Hepatitis, hospitalization, death, or birth—and dividing that number by the number of persons at risk of experiencing that condition. Remember, "persons at risk" refers to the people in the total population to whom the event or health condition we're studying could have happened.

When calculating a rate, both the numerator (the persons experiencing an event) and the denominator (the persons at risk of experiencing that condition) must represent the same time period and place.

We express rates in terms of standard units of population to make comparisons easier. For example, 37.3 per 100,000 is easier to read and discuss than 0.000373. The standard unit of population that you use when you calculate a rate is based on convention.

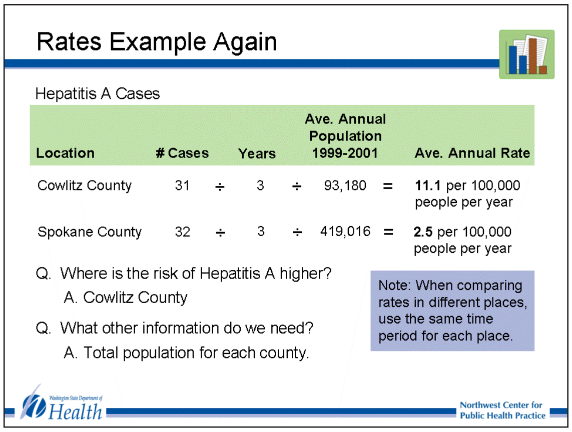


Rates Example Again

Let's look again at our Hepatitis A cases in Cowlitz and Spokane counties and the questions we asked.

- Where is the risk of Hepatitis A higher?
- What other information do we need?

To answer those questions, we've added some additional information about the total population in those counties. Notice that although we're now using two different sources of data, case counts and population data, we still must use the same time period for each.



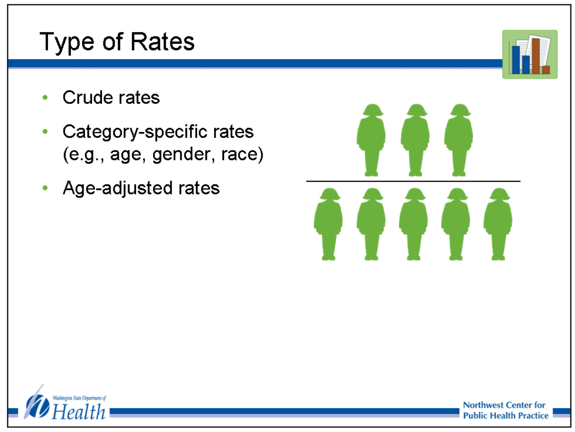
By using our formula to calculate a rate in each county, we can see that the rate of Hepatitis A was higher in Cowlitz County than in Spokane County. So we can say that in this time period, the risk of hepatitis A was higher in Cowlitz County.

Let's pause so you can check your understanding of what we've just covered.

Practice: Counts, Ratios, Rates

Type of Rates

Several common types of rates are used in descriptive epidemiology. In this section we'll look at crude rates, category-specific rates (such as age, gender, and race), and age-adjusted rates, and when to use each one.



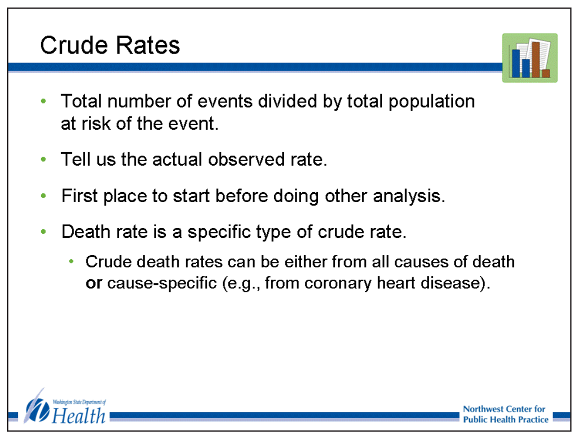
Crude Rates

Crude rates are calculated by dividing the total number of events by the total population at risk of the event. In our Hepatitis A example, we calculated the crude Hepatitis A rates for Spokane and Cowlitz counties.

We use crude rates when we want to express the actual observed rate in the population. Crude rates should be the first place to start before we do any other analysis.

A specific kind of a crude rate is a crude death rate. When you calculate a crude death rate, depending on what question you want to answer, you can include deaths from all causes or just deaths from a specific cause, such as coronary heart disease.

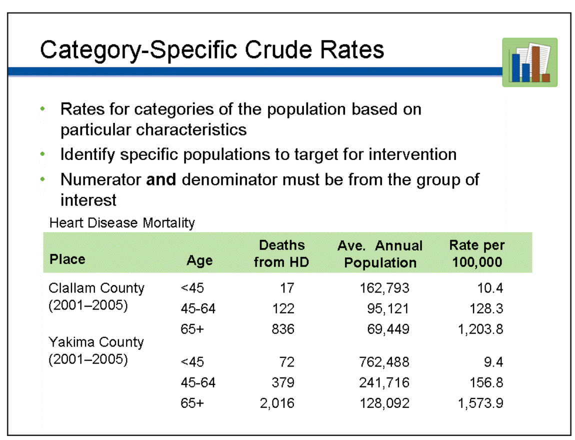
Remember that as with all rates, the numerator and denominator must be a part of the group of interest, in other words they must share the same time and place.



Practice Crude Rates

Category-Specific Crude Rates

We can also calculate crude rates for categories of the population based on particular characteristics, such as age, gender, and race. These rates are sometimes referred to as age- or gender-specific rates and can help us identify specific populations to target for intervention.



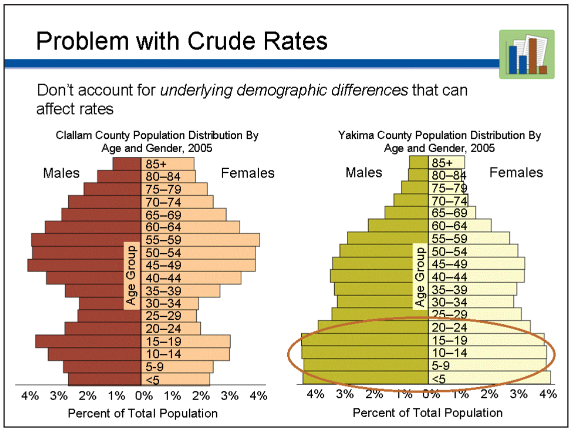
In category-specific rates, the numerator and denominator must both be drawn from the group of interest, rather than the total population.

Let's look at the crude rates of heart disease deaths in two counties—Clallam and Yakima. It appears that Clallam County has a higher rate than Yakima County.

But if we break these data into three age groups, we see a slightly different story. Now you can see that although the rate is still slightly higher for those under 45 in Clallam County, it's actually lower in the two older age groups.

Let's pause so you can answer some questions on what we just covered.

Problem with Crude Rates



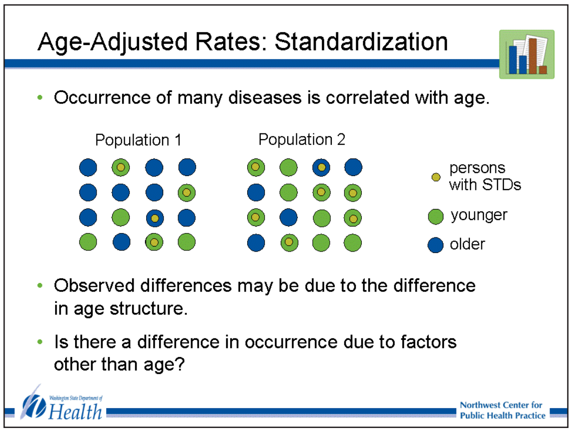
Although crude rates are the foundation of other rates, we rarely use them alone. The problem with the crude rates we calculated is that they don't account for underlying demographic differences between communities (or between time periods) that can affect rates.

To apply this concept to our heart disease death data in Clallam and Yakima counties, let's take a look at what the populations in those counties look like. This is a population pyramid for Clallam County and here's one for Yakima. You can see that the population in Yakima is much younger.

So we have to ask: Is the higher rate of heart disease that we observed in Clallam simply attributable to the fact that a higher proportion of older people live there?

We can't answer this question with crude rates. Instead we need to standardize the rates through a process called age adjustment.

Age-Adjusted Rates: Standardization

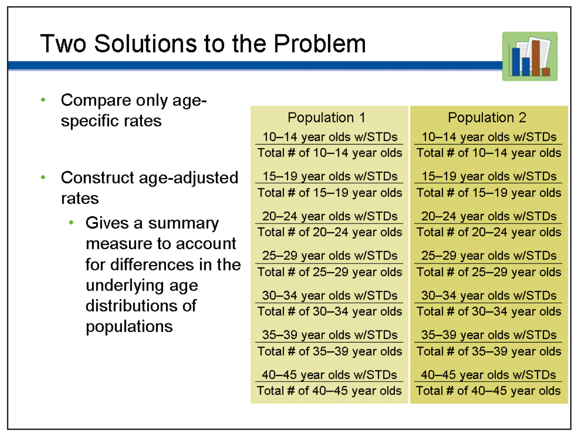


The occurrence of many diseases is correlated with age. The incidence of chronic diseases (such as coronary heart disease) naturally increases with age or, as with cancer, often takes a long time to develop. Other diseases, such as STDs, are found more often in younger age groups because they're related to behaviors that are more prevalent in younger people.

So why is this important? We would expect to find more chronic diseases in areas with older populations—and areas with younger populations are likely to have higher STD rates. This means that any observed differences between areas with different age distributions may simply be attributable to the difference in age structure rather than to factors that can be affected by public health interventions.

What we really want to know is whether there is a difference in the occurrence of the disease due to factors other than the age of the populations. In other words, once we remove age from the equation, is there any remaining difference in rates?

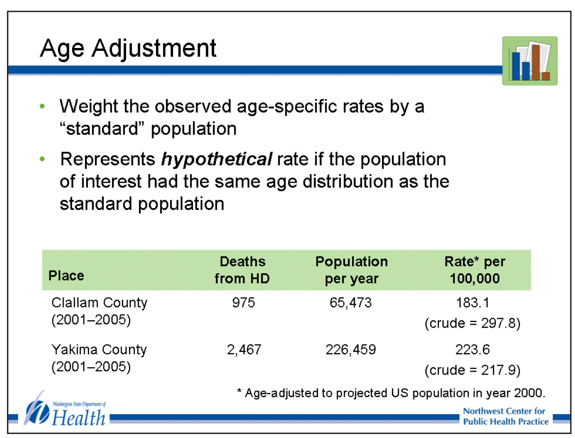
Two Solutions to the Problem



We have two solutions for comparing populations with different ages. We could make comparisons between two populations using age-specific rates. However, it's cumbersome to make so many comparisons. And in the end, it doesn't capture what's happening in the total population. Usually we want a summary measure.

The way to get a summary measure is to construct age-adjusted rates. Age-adjusted rates will give us a summary measure that takes into account the differences in the underlying age distributions of different populations.

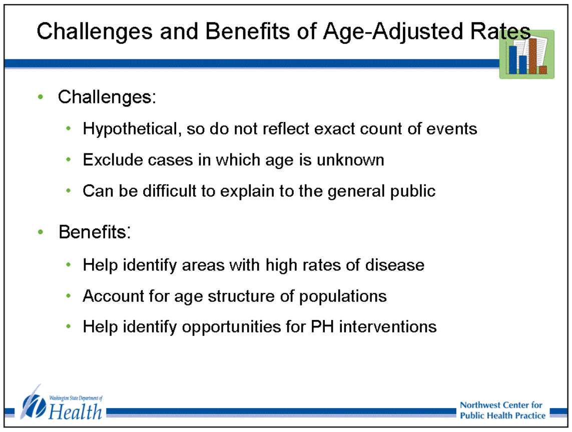
Age Adjustment



Age-adjustment means that we weight the observed age-specific rates by a "standard" population. In the United States, the standard population is usually the most recent U.S. Census population. The resulting age-adjusted rate represents the hypothetical rate that would be observed if the populations of interest had the same age distribution as the standard population.

After we calculate age-adjusted rates for Yakima and Clallam counties, we see that Yakima actually has a higher rate of heart disease. This age-adjusted rate might point to modifiable risk factors for heart disease in Yakima (such as obesity) because age is no longer the reason for the high rates. Whenever we want to compare rates between two different geographical areas and we do not want age to be a factor, we need to do age adjustment.

Challenges and Benefits of Age-Adjusted Rates



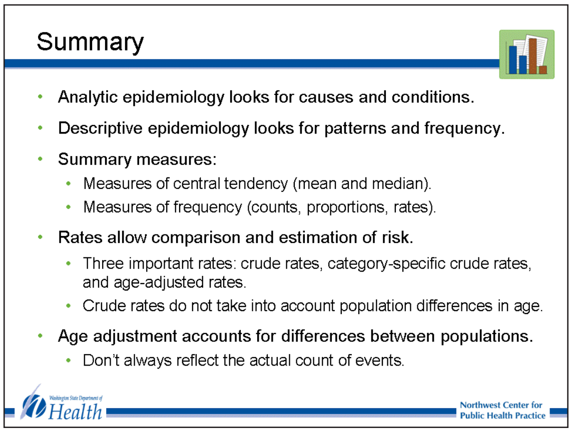
There are three problems with age-adjusted rates. Because they're hypothetical constructions, they do not reflect the exact count of events. Also, if you don't have information on age, those cases without age information will be excluded. And finally, age adjustment can be difficult to explain to the general public.

In spite of these challenges, age-adjusted rates are very useful because they help us identify areas with high rates of disease by taking into account the age structure of their populations. In other words, they help us identify opportunities for public health interventions.

Let's pause so you can answer some questions on what we just covered.

Practice: Age-Adjusted Rates

Summary



To summarize, in this module we looked briefly at the difference between analytic and descriptive epidemiology.

Analytic epidemiology conducts studies to look for causes or associations of exposure and diseases or conditions. In contrast, descriptive epidemiology looks for patterns and frequency of health events. Descriptive epidemiology is most frequently used in public health assessment.

We use a number of different measures to describe health events. One category of summary measures are the measures of central tendency, which measures continuous data and includes the mean and the median.

Another type of summary measure are the measures of frequency, which are used to measure categorical data and include counts, proportions, and rates.

Rates are used for estimating risk. They allow us to compare data from different times and places and from different sizes of populations. Keep in mind that it is always important to look at the data to determine if they make sense or if small numbers are creating unreliable results.

Basic Concepts of Data Analysis for Community Health Assessment

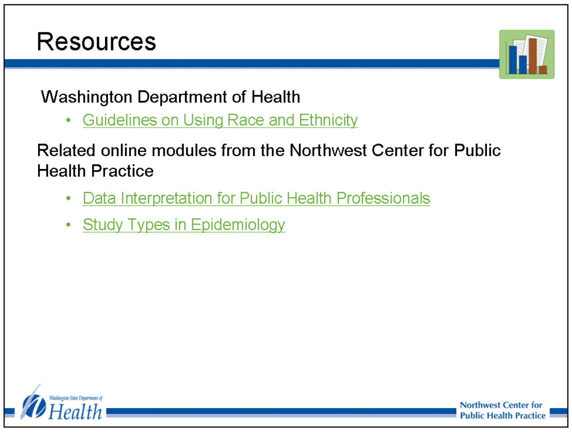
Module 2: Analysis and Interpretation of Public Health Data: Part 1

Transcript

Three important rates for doing assessment are crude rates, category-specific crude rates, and age-adjusted rates.

A major problem with crude rates is they do not take into account population differences in age. Different populations can be compared only after we perform an age-adjustment on the data in order to account for the differences that may be simply a result of differences in age groups in the populations. It's important to keep in mind, however, that age-adjusted rates don't reflect the actual counts of events.

Resources



If you would like to learn more about the concepts in this module, you might want to explore some of the resources listed here.

Washington Department of Health

Guidelines on Using Race and Ethnicity
www.doh.wa.gov/Data/Guidelines/Raceguide1.htm

Related online modules from the Northwest Center for Public Health Practice

Data Interpretation for Public Health Professionals
www.nwcphp.org/data
Study Types in Epidemiology
www.nwcphp.org/study-types

Now, if you're ready, please go on to the final test.