Data Interpretation for Public Health Professionals

Data Interpretation for Idaho Public Health Professionals

Welcome to Data Interpretation for Idaho Public Health Professionals. My name is Janet Baseman. I'm a faculty member at the Northwest Center for Public Health Practice at the School of Public Health and Community Medicine at the University of Washington in Seattle. I also work as an epidemiologist with the UW Center for Public Health Informatics and the Kitsap County Health District.

> **Data Interpretation for Idaho Public Health Professionals**
>
> Janet Baseman, MPH, PhD
>
> **Northwest Center for Public Health Practice** www.nwcphp.org

Introduction to Course

This hour-long course provides public health professionals with an introduction to data interpretation. The examples and interactive exercises in this module offer opportunities to increase your skills in presenting data to co-workers, community-based organizations, hospitals, public agencies, boards of health, and the general public. When available, Idaho-specific examples have been used. When not available, U.S. national examples are used.

> **Introduction to Course**
>
> - Hour-long introduction to data interpretation
> - Target audience: public health professionals
> - Examples and interactive exercises
> - Idaho-specific data used when available

Objectives

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

1) List at least three common data sources used to characterize the health or disease status of a community

2) Define and interpret basic epidemiology measures such as prevalence, incidence, mortality, and case-fatality

3) Define and interpret basic biostatistical measures such as the mean, median, mode, confidence interval, and p-value

4) Read and interpret tables and graphs, and

5) Determine the appropriate format for data presentation

> **Course Objectives**
>
> By the end of this module you should be able to:
>
> 1. List at least three common data sources used to characterize the health or disease status of a community
> 2. Define and interpret basic epidemiology measures such as prevalence, incidence, mortality, and case fatality
> 3. Define and interpret basic biostatistical measures such as the mean, median, mode, confidence interval, and p-value
> 4. Read and interpret tables and graphs
> 5. Determine the appropriate format for data presentation (table, line graph, pie chart)

Uses of Data in Public Health

Data, and the results of data analysis, have many uses in public health practice. These include, but are not limited to:

- Population or community health assessment
- Public health surveillance
- Disease investigation
- Prevention and control measures evaluation
- Program planning
- Future health problems and needs assessment, and
- Hypothesis generation for study design

Data Sources

Useful data for public health may come from national, state, and local sources, and are not limited to what you might typically identify as "public health data sources."

Sources include, but are not limited to: surveillance data, health related-surveys, administrative sources, vital statistics, outbreak investigations, research, and the US Census.
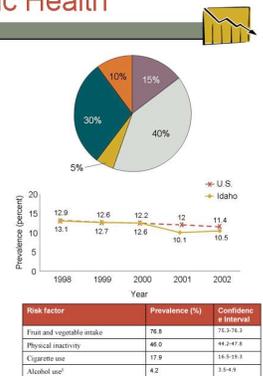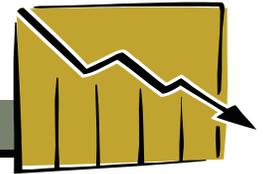

Data Attributes

To be useful, data should be of good quality. When we use high quality data, we are more likely to trust conclusions we draw from those data and act on our results to make improvements in whatever health condition is being studied.

The quality of data depends on several factors, including accuracy and completeness. For example, in the case of surveillance for notifiable conditions reporting, we hope that a reported diagnosis in our data correctly classifies the true diagnosis of an individual. This is called accuracy. As for completeness, if we get incomplete data, we might not have a good idea of what is truly going on specific to the health of our community. Even if our data are not 100% accurate and complete, the data will usually still be useful if we have an idea of how accurate and how complete the data are. Having some data is frequently better than having none at all.

When we use data to answer questions about the health of our communities, the data should be relevant to the populations and health conditions we are interested in, and the data should arrive in a timely enough fashion so that appropriate or necessary actions can be taken for control of a health-related event.

Data Limitations

But all data sets have limitations. Though we won't discuss them in detail, some limitations of public health data we might encounter include: inaccurate diagnoses or coding, poorly conducted data collection, data entry or data analysis, and issues that result in data not being representative of the population we'd like to draw conclusions about.

For those of us working in public health practice, it is good to know that reliable, published public health data sources are available at the national, state and frequently at the local level. Understanding the strengths and limitations of the data we work with is important. If you have questions about the quality of a data source you're interested in using, data experts within the state Departments of Health can often provide additional expertise and analysis.

Data Limitations

All data have limitations. Limitations include:
- Inaccurate diagnosis or coding
- Poorly conducted data collection, entry, or analysis
- Non-representative data

Check with data experts in the state department of health to confirm the quality of a data source you want to use.

What Is Descriptive Epidemiology?

In public health, we frequently look at descriptive data from sources that use the methods of descriptive epidemiology. But what is descriptive epidemiology? Basically, it is a systematic way to describe a health problem. And it answers the questions:

- Who is getting a disease or other health outcome of interest? We call this the person.
- Where is the health outcome occurring? We call this the place.
- When did the health outcome occur, and is the frequency of the health outcome changing over time?
- When we organize or analyze data by "person" we have several categories we can use. We may use inherent characteristics of people (for example, age, sex, or race), their acquired characteristics (such as their marital status), their activities (such as their occupation or leisure activities), or the conditions under which they live (such as socio-economic status).

What Is Descriptive Epidemiology?

A systematic methodology for describing a health problem
- Answers the questions:
  - **Who** has the greatest risk of getting a disease or other health outcome?
  - **Where** is the health outcome occurring?
  - **When** did the health outcome occur and is the frequency of the health outcome changing over time?
- Relies on an investigation of:
  - **Person**: age, race, sex, marital status, occupation, behavior, socioeconomic status
  - **Place**: residence, place of employment, geographic site (state, county, census tract, etc.), rural/urban, domestic /foreign
  - **Time**: onset, duration, seasonal occurrence, daily or hourly changes, annual trends

We describe a health event by place to gain insight into the geographical extent of the problem. We may use place of residence, birthplace, place of employment, school district, or hospital unit. Or, we may use geographic units such as country, state, county, census tract, street address, map coordinates, or some other standard geographical designation. Sometimes, we may find it useful to analyze data according to place categories such as urban or rural, or domestic or foreign.

To answer questions on when a health event has occurred, we use measures of time. We may consider the date of onset, duration, or seasonal occurrence of a disease, for example. We are often interested in how a health event changes over time, such as during an outbreak investigation when we are interested in daily or hourly changes. We may learn about annual trends by monitoring a health condition for expected or unexpected increases or decreases; or we may be interested in learning how a condition has changed between two time periods: for example, before and after an intervention.

Measures of Disease Frequency

Measuring events, such as disease or health events, is at the heart of public health surveillance and resource allocation. One of the simplest methods of measuring is just simply counting. However, as you will see in a later slide, simple counts often do not provide all of the information needed to understand the relationship of a health event to the population in which the event occurred. Counts alone are also insufficient for describing the characteristics of a population and for determining risk. The key is to relate the frequency of an event to an appropriate population. For this purpose we use ratios, proportions, and rates.

The next series of slides will introduce you to ratios, proportions, and rates, and then we'll provide an overview of four measures of disease frequency or severity—prevalence, incidence, mortality, and case-fatality – that are commonly used in public health.

Measures of Disease Frequency

A core public health activity is to count events in the source population.
Counts, alone, are insufficient to:
- Understand relationship of event to population
- Describe population characteristics
- Determine risk

We relate the frequency to the population with **ratios, proportions,** and **rates**.

Measures of disease frequency: prevalence, incidence, mortality, case-fatality

Numerator and Denominator

Ratios, proportions and rates are measures frequently used to define the health of our communities. Ratios, proportions, and rates all include both a numerator and a denominator. Let's start with the numerator. The numerator is the top part of a fraction. In the fraction ¾, 3 is the numerator. In public health, often the numerator is the number of health events of interest.

Now let's review the denominator. The denominator is the bottom part of a fraction. In the fraction ¾, 4 is the denominator. In public health, the denominator is often 'the population at risk'. The population at risk is the group of people, healthy or sick, who would be counted as cases if they had the health condition being studied.

Everyone in the population at risk must be eligible to be counted in the numerator if they have the event of interest. And you only count events in the numerator that occur within the population at risk, or the denominator. For example, in looking at the rate of colon cancer in women, we would not include men in the population at risk, because men with colon cancer would not be included in the number of events. Likewise, if we were interested in knowing the rate of prostate cancer in a certain population, women would not be included in the population at risk.

Numerator and Denominator

$$\frac{3}{4}$$

numerator
- number of events in a specified population
- count only events that occur among the population at risk

denominator
- number of people in a specified population, or **population at risk**, of experiencing the event
- must be eligible to be counted in the numerator if they have the event of interest

Ratios and Proportions

A ratio is obtained by dividing one quantity by another. For example, to find the ratio of Idaho women ages 65 and older to Idaho men ages 65 and older, we would divide the number of women by the number of men. (Here we're using the projected ratio of female to male adults ages 65 and older in Idaho in 2010.) Notice that the number of women is the numerator, and the number of men is the denominator [99,581 / 81,835, or 1.22].

A proportion is a ratio in which the numerator is included in the denominator.

The projected proportion of adults who will be 65 or older in Idaho in 2010 is 181,416. The total population is 1,517,291, which includes all ages of people. Notice that when you divide the numerator by the denominator, the value you get is 0.12. Proportions are often multiplied by 100 and reported as percentages, so we can say that the projected proportion of adults in Idaho who will be age 65 or older in 2010 is 12%.

Ratios and Proportions

- **Ratio:** any fraction obtained by dividing one quantity by another, such as x, the numerator, divided by y, the denominator, or x/y

- **Proportion:** a ratio in which x is included in y
  Proportions are often multiplied by 100 and reported as a percent

$\geq 65 = 181,417$

$$\frac{99,582}{81,835} = 1.22$$

$$\frac{181,417}{1,517,291} = 0.12 * 100 = 12\%$$

Source: www.census.gov/population/projections/13PyrmdID1.pdf

Proportion and Rates

We'll spend the next several slides talking about rates. A rate is often a proportion, with the added dimension of time.

Rates measure the frequency at which a health event occurs over a period of time. Later, we'll go through examples of how rates are calculated, but for now, let's just say that a rate is a numerator divided by a denominator times a standard unit of population size (like 100 or 1000 or 100,000 people). A rate represents the burden of disease or other health related outcome during a specific time period.

Proportions and rates are used in public health for quantifying morbidity and mortality

- Some measures of morbidity, or illness, include prevalence and incidence rate.
- Measures of mortality include mortality rates, and
- Measures of disease severity include case-fatality

Proportion and Rates

- **Rate:** *often a proportion*, with an added dimension of time
  - Measures the frequency at which a health event occurs over a period of time

$$\frac{x, \text{ number of events}}{y, \text{ population at risk}} \quad * \quad \begin{array}{c} \text{unit of population} \\ \text{(e.g., 100, 1,000, 100,000)} \end{array}$$

- Proportions and rates are used for quantifying morbidity and mortality
  - **Measures of morbidity**: prevalence; incidence rate
  - **Measures of mortality**: mortality rate
  - **Measures of disease severity:** case-fatality

May 28

Exercise on Proportions

Why Use Rates?

We use rates to describe the frequency of a health event or health status relative to the size of a population

When we use rates, the number of events is adjusted for the size of the source population in which they occurred.

Rates make it possible to compare disease frequency across different groups of people, places, and time periods.

Clinicians often report numbers of patients they see with certain diseases, but epidemiologists and public health agencies use proportions and rates, which allow us to describe the frequency of diseases relative to the size of the population in which they occurred.

Why Use Rates?

- Describe the frequency of a health event or health status relative to the size of a population
- Facilitate *comparison* of disease frequency across different:
  - **Groups of people** — account for change in size of community
  - **Places** — allow for comparison of communities of different sizes
  - **Time periods** — account for observations over different periods of time
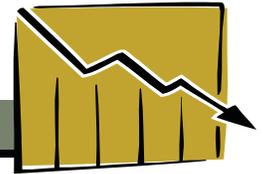
Rates allow us to make comparisons between groups of people, such as different age groups, or locations that have different population sizes, such as states versus cities or urban versus rural areas. Rates also allow us to make comparisons within the same population over time.

Rates are useful in many ways. With rates, the health department can identify groups in the community with an elevated risk of disease. With this information, risk factors can be examined and interventions targeted to high-risk groups.

Using Rates

In using rates to make comparisons, we need to account for the fact that the number of health events depends in part on the number of people in the community. For instance, we expect to find more cases in larger populations. To account for growth in a community or to compare communities of different sizes, we usually calculate rates to provide the number of events per population unit.

For example, in looking at this table of surveillance data, if you have 1000 people in your community and 20 cases of a disease, that could be a lot more significant than if you have 1 million people in your city and 20 cases of a disease.

Using Rates

	Cases		Total population	# cases / Total pop.	Crude rate
City A	20	÷	1,000	0.02 * 100,000	2,000 per 100,000
City B	20	÷	1,000,000	0.00002 * 100,000	2 per 100,000

Furthermore, if all 20 cases in either population occur within only one age group, race, or gender, this might influence your decision to investigate further.

When we divide the numerator by the denominator in each city, you can see that the rate in city A is much greater than the rate in city B – 1,000 times greater. Frequently, we will take these rates represented as decimals and multiply them by a multiple of 10 in order to convert them to whole numbers. Here, this is done by multiplying both crude rates by 100,000, and the result is in the far right column. As long as you multiply the rates for cities A and B by the same multiple of 10, you will still have a valid comparison. Also, this far right column makes the most sense in terms of interpretation of a rate. For example, in city B, we can say that the rate of disease in this population was 2/100,000, or for every 100,000 people in the population, two cases of disease were identified during the time period of interest.

Crude Rate

In the following slides we will introduce common rates that are used in public health. Each serves a different purpose.

Let's begin with crude rates: A crude rate is the rate calculated for the total population. Crude rates are recommended when a summary measure is needed and it is not necessary or desirable to take into account any other factors, such as the age of the population.

A crude rate is calculated by dividing the total number of events in a specified time period by the total number of individuals in the population who are at risk for these events and multiplying by a constant, such as 1,000 or 100,000 [in other

Crude Rate

- Rate calculated for the **total** population (all ages)

Between 2000 and 2004, the **crude rate for suicide deaths** in Idaho was 15.1 cases per 100,000 population **constant**

- **Advantages**
  - Easy to calculate
  - Broadly describes health outcome of interest
  - Captures time element
- **Do not account for**:
  - Underlying demographic differences between communities
  - Differences between time periods

words, (numerator/denominator) x constant]. We saw an example of how to calculate a crude rate on the previous slide, when we calculated rates of disease in city A versus city B.

An example of a crude rate would be: Between 2000 and 2004, the crude rate for suicide deaths in Idaho was 15.1 cases per 100,000 population.

The crude rate has the advantage that it is a simple, easily calculated measure that gives a broad depiction of the extent of a health outcome in a particular area in a particular time period. The crude rate presents the actual magnitude of an event within a population.

The problem with crude rates is that they do not account for the underlying demographic differences between communities or between time periods, and this can be a real problem when we're trying to compare rates of disease between two communities, for example, or between our county and the state or between two different time periods within our own county.

Category-Specific Rates

Category-specific rates are rates measured in a specific group, such as the rate of disease within one gender, ethnicity or age group. When the rate applies to a specific age group, such as those between the ages of 15–24, it is called the age-specific rate. Category-specific rates are used for comparisons when rates differ widely between groups. For example, if we know that the highest injury rates from unintentional falls are for children under 14 years and for people over 65 years, we might like to calculate injury rates in these age groups specifically instead of in the total population.

Category specific rates are recommended when specific causal or protective factors are different for different subgroups. They present the actual magnitude of an event within a designated group.

Category-Specific Rates

- Calculated for a **specific group** within a population
- Measures **frequency** of a health event in a certain group
- Used for **comparisons** when rates are strongly subgroup dependent
  - Recommended when factors are different for different subgroups
  - Presents the **actual magnitude** of a health event

Age-Adjusted Rates

Almost all diseases or health outcomes occur at different rates in different age groups. Many chronic diseases, including most cancers, occur more often among older people. Other outcomes, such as many types of injuries, may occur more often among younger people than in middle aged people. Therefore, the age distribution of a given population often determines what the most common health problems in a community will be. What I mean by this is that if a certain population mostly consists of older people, the burden of disease from cancer in that community will likely be greater that it would be in a community that mostly consists of younger people.

Age-Adjusted Rates

- Allow the comparison of rates of age-related health events between communities with different age structures
- Use a "standard" population
- Compare trends in a given population over time
- Can mask important trends
- Require knowledge of biostatistics

By convention, rates are frequently adjusted to the age distribution of the estimated U.S. population of the year 2000, commonly referred to as the standard population.
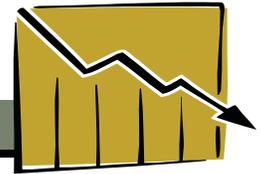
To summarize, age-adjusted rates are recommended when making comparisons in the rates of age-related health events between different populations or for comparing trends in a given popula-

tion over time. And age-adjusted rates are essential for events that vary with age (for example, cancer deaths), or when comparing populations with different age distributions. However, because age-adjusted rates are often calculated using a standard population, these rates can mask important trends, so it is also important to look at crude rates, as well as category-specific rates. Finally, age-adjustment requires training and a knowledge of biostatistics.

Effect of Age Adjustment

When comparing rates between communities, it is useful to calculate a rate that is not affected by differences in the age composition of the populations. Let's say that we have two populations. Not knowing anything about their age distributions, we can see that population one has a higher rate of disease than population two. There could be many reasons for this differ- ence. For example, there could be environmental, behavioral, or genetic differences between the two populations that result in different disease rates. But if you recall that many diseases occur more often in older people, another reason for this observed difference could be that population one has a higher proportion of older people than population two.

An age-adjusted rate mathematically removes the effect of the age composition of the underlying population, allowing comparisons between communities with different age distributions. By conven- tion, rates are frequently adjusted to the age distribution of the estimated U.S. population in the year 2000, which is commonly referred to as the "standard" population.

After the rates of disease in our different populations, and within our different age categories, are applied to the standard population, we are left with rates of disease that we can compare, having effectively removed, or adjusted for, the differences in age distributions of the two populations.

As you can see, the disease rate in population one is still higher than the disease rate in population two after adjusting for age. This implies that something other than age is causing the higher rate of disease in population one.
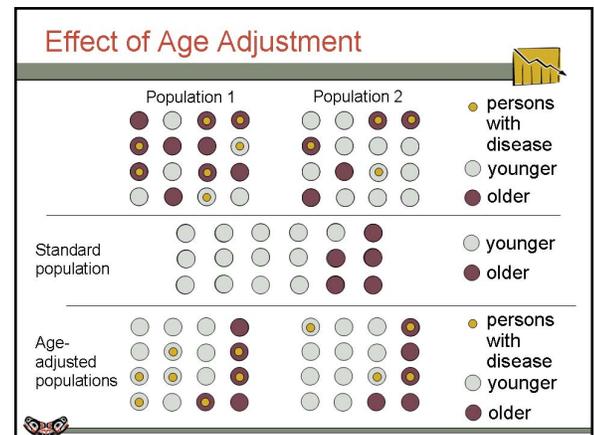
Why Compare Rates?

Comparing rates is useful for several reasons. For example, you might want to know whether the rate of some disease differs between your county and the rest of the state, or whether Idaho's rates are different from the rest of the U.S., or whether there have been changes in the rates of some disease between 1996 and 2006.

But it is important to remember that you should compare rates only when the events and population are defined the same way over time and place. Some things to consider when deter- mining whether two rates can be compared are:

- Consistency in definition of event
- Consistency in the methods used to collect the data, and
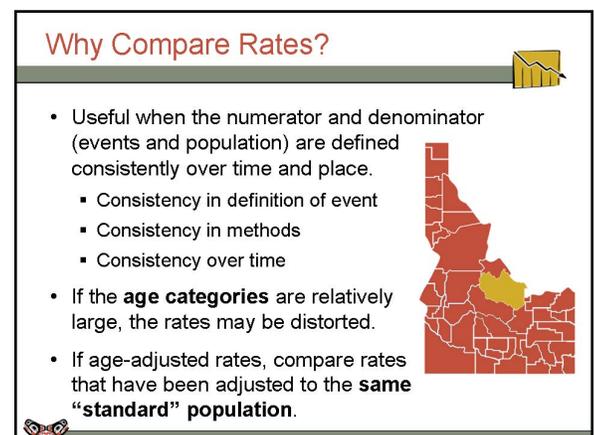
- Consistency over time. For example, you probably don't want to compare the rate of diabetes in Idaho in 2006 with the rate of diabetes in the U.S. in 1986. The exception is, of course, when you are purposefully comparing events between two time periods within the same population.

When comparing age-specific rates, if the age categories are relatively large, such as 15 to 29, the rates may be distorted, because this includes such a large group of people. Looking at smaller groupings (such as 15-19, or 20-24) might give you a better idea of what is going on in your population of interest. In the case of comparing age-adjusted rates, be sure to compare only rates that have been adjusted to the same "standard" population.

Exercise on Kinds of Rates

Prevalence

We are now going to turn our attention to specific types of measures frequently used in public health to evaluate the burden of disease in our communities. The first is prevalence.

Prevalence measures the number of cases (including both new and old cases) of a disease (or health-related condition or event) at a specific time point or period in time. Note that if prevalence is measured for a period of time, say three months, rather than at a point in time, the population denominator should represent the average population during that period.

Some important things to remember about prevalence are that:

- It is used to present a 'snapshot' view of the disease or health condition of interest.
- We frequently obtain prevalence data from surveys or surveillance databases, and that,
- Prevalence is a proportion (or percentage).

When we calculate prevalence, the numerator includes existing cases of a disease at a specified time and the denominator includes people in the defined population at that time. Because prevalence describes the burden of illness in a population, public health professionals use prevalence to assess the effect of a health event or disease on the resource needs of both the public health system and the health care delivery system.

Calculating Prevalence

Now let's work through an example of how prevalence is calculated.

In 2003, Idaho had a household population of 1,366,322. Of these people, 11 percent were 65 and older.

In 2003, Idaho reported 21,000 individuals 65 and older with diabetes.

In this example, the prevalence of diabetes among Idahoans

Prevalence

A measure used to evaluate the burden of disease in a community.

$$\frac{\text{Existing cases of a disease at a specified time}}{\text{People in the defined population at that time}}$$

- Used to present a **snapshot** of the disease or health condition of interest
- Data often obtained from surveys or surveillance databases
- Is a **proportion** (percentage)

Calculating Prevalence

2003 household population = 1,366,322
Idahoans ≥ 65 = 11% (150,295)
Idahoans ≥ 65 with diabetes = 21,000

$$\frac{21,000^{\dagger}}{150,295} = 0.140 * 100 = 14.0\%$$

140 cases of diabetes per 1,000 Idaho residents equal to or over 65 years of age

$$0.140 * 1,000 = 140$$

† BRFSS population estimates are weighted to account for sampling survey design. Source: Idaho Department of Health and Welfare

ages 65 and older is calculated by dividing the number of individuals 65 years and older with diabetes by the household population who were 65 years and older.

Here we see that calculation [21,000 / 150,295 = 0.140]. We multiply by 100, to report the prevalence as a percentage, and we see that in 2003, the prevalence of diabetes in Idaho residents equal to or over 65 years of age was 14%.

We can also express the prevalence as 140 cases per 1,000 Idaho residents equal to or over 65 years of age. We arrive at this value by multiplying 0.14 by 1000.

Exercise on Prevalence

Incidence (Rate)

Another measure of burden of disease is incidence. Incidence is defined as the number of new cases of a condition during a defined time interval divided by the number of persons at risk of developing the condition over that time interval.

Incidence rates provide a direct measure of the rate at which new illness occurs in the population, and therefore incidence rates—and the incident cases from which we derive them—can be used to study the causes of health events. Incidence rates are commonly expressed as the number of cases of disease or injury per 100,000 person-years of exposure to the risk.

Now let's look at an example of incidence.

Incidence (Rate)

New cases of a condition* during a defined time interval

Persons at risk of developing the condition over that time interval

- Direct measure of the rate at which new illness occurs in the population
- Used to help determine the causes of health events

Person-years describes the length of time of experience or exposure of a group of people who have been observed for varying periods of time.

- Expressed as the number of cases of disease or injury per 100,000 person years

*Condition may refer to disease, symptom, injury, behavior, or other event

Calculating Incidence

In 2004, 2784 new cases of chlamydia were reported in Idaho State.

The at-risk population in Idaho in 2004 was 1,393,262.

Incidence is equal to the new cases divided by the at-risk population [2784 / 1,393,262 = 0.001998].

In 2004, the incidence of chlamydia in Idaho State was 0.1998%.

We can also express this as 199.8 cases per 100,000 population.

Calculating Incidence

2004 new cases of chlamydia = 2784

At risk population = 1,393,262

$$\frac{2784}{1,393,262} = 0.001998 * 100 = 0.1998\%$$

or

$$0.001998 * 100,000 = 199.8 \text{ cases per } 100,000 \text{ people}$$

Source: Idaho Department of Health and Welfare

Exercise on Incidence

Mortality Rates

A mortality rate is a specific type of incidence rate. Mortality rates are used to describe the incidence of death in a population, rather than the incidence of disease. Mortality rates are frequently referred to as death rates. Mortality rates are calculated by dividing the number of deaths in the population during a stated time period by the number of persons at risk of dying during that period.

Because mortality is a rate, it can be expressed as number of deaths per 1000, per 10,000 or per 100,000, depending on the condition being described.

- Several useful types of mortality rates are frequently used in public health.
- The crude mortality rate for a population is the death rate from all causes.
- A cause-specific mortality rate is the mortality rate of a population associated with one disease or other cause.
- And age-specific mortality rate describes the rate of death in a certain age group.

Mortality can vary considerably by age, sex, race, and ethnicity. As a result age-specific death rates are often presented separately for the different genders or different ethnic groups.

Now we'll work through an example of how to calculate a mortality rate.

Mortality Rates

- Used to describe the incidence of death in a population, rather than the incidence of disease

$$\frac{\textbf{Deaths} \text{ in the population during a stated time period}}{\textbf{Persons at risk of dying} \text{ during that period}}$$

- Expressed as number per 1000, 10,000, or 100,000, depending on condition
- **Crude mortality rate** = deaths from all causes per 100,000 population (or other constant)
- **Cause-specific mortality rate** = death rate associated with one disease
- **Age-specific mortality rate** = death rate in a certain age group

Calculating Mortality Rates

Between the years 2000 and 2004, 1039 suicide deaths were reported among Idaho residents. If we are interested in knowing the annual suicide mortality rate during this time period, how would we calculate it?

Number of suicide deaths should go in the numerator. The number of suicide deaths we are given 1039, is for a 5 year time interval, so we can divide that by 5 to get the numerator for an average annual rate.

The mid-year, or 2002, population of Idaho was 1,341,131. This will be our denominator.

Dividing the average number of deaths per year by the 2002 Idaho population gives us the following value: 0.000155. This is the suicide mortality rate. Multiplying our rate by 100,000, we can say that the suicide mortality rate in Idaho between 2000 and 2004 was 15.5 per 100,000 people.

Calculating Mortality Rates

Suicide deaths 2000 – 2004

$$\frac{1,039}{5} = 207.8$$

$$\frac{207.8}{1,341,131} = 0.000155$$

$$0.000155 * 100,000$$
$$= 15.5 \text{ per } 100,000 \text{ Idaho residents}$$

Age-Specific Mortality Rate

Here's an age-specific mortality rate example. For annual rates, the age-specific rate is the number of new cases or deaths in given age group and year divided by the population in that age group.

The breast cancer mortality rate in 2003 among women between the ages of 20 and 44 in Idaho State was 4.2 per 100,000 women. This rate was calculated by dividing the number of breast cancer deaths in 2003 among women in this

Age-Specific Mortality Rate

Age-specific breast cancer mortality rate in 2003 for women in Idaho State

Ages	Number of deaths		Total population	* 100,000	Mortality rate per 100,000
20–44	10	÷	238,590	* 100,000	4.2
45–64	56		164,864	* 100,000	34.0
65+	95		87,991	* 100,000	108.0

Source: Idaho Department of Health and Welfare, Bureau of Health Policy and Vital Statistics, June 2006

age group by the total number of women in Idaho in that age group in 2003 and then multiplying that value by 100,000. The breast cancer mortality rate for women ages 45-64 was 34 per 100,000. Among women ages 65 and older, the breast cancer mortality rate was 108 per 100,000.

Case Fatality

Another type of measure we sometimes use in public health is the case-fatality. Case-fatality is calculated by dividing the number of deaths from a condition during a stated time period by the number of persons with the condition of interest. We call this case-fatality because in the denominator we're referring to those with the condition as cases.

The case-fatality provides us with a measure of the severity of the condition of interest. For example, among people older than age 70 with West Nile Virus meningoencephalitis, the case-fatality was 21% in the U.S. in the year 2002.

Now let's work through an example.

Case-Fatality

$$\frac{\text{Deaths from a condition during a stated time period}}{\text{Cases (persons with the condition)}}$$

- A measure of the **severity** of the condition
  - Among persons older than 70 years of age with West Nile Virus meningoencephalitis, the **case-fatality** was 21% among those with meningoencephalitis in the U.S. in 2002.

JAMA, 2003, Vol.290. No.4, West Nile Virus: Updates

Case Fatality Example

The National Highway Traffic Safety Administration reported that in the year 2000, 4,739 deaths occurred in the U.S. when a pedestrian was hit by a motor vehicle.

For calculation of case-fatality, these are the fatalities. Additionally, the report estimated that there were a total of 78,000 pedestrians injured after being struck by motor vehicles in the year 2000.

For calculation of case-fatality, these are the cases. Dividing the number of fatalities by the total number of cases of injury (including deaths) gives us a case-fatality of 6.1%, and we can say that in 2000, 6.1% of pedestrians injured by motor vehicles died as a result of their injuries.

Calculating Case-Fatality

$$\frac{4,739}{78,000} \frac{\text{fatalities}}{\text{cases}} = 0.061 * 100 = 6.1\%$$

Source: 2000 Motor vehicle traffic crashes, injury and fatality estimates, early assessment. Washington, D.C., U.S. Department of Transportation, 2001.

Comparing Mortality & Case-Fatality

Now let's compare mortality and case-fatality using an example. Assume you have a population of 150,000 people of whom 20 are sick with a certain disease X, and in one year, 18 people die from disease X. The mortality rate in that year from disease X is equal to 18/150,000.

Converting this to a mortality rate per 100,000 people, we can say that the mortality rate is 12 per 100,000 population. Note that this is a cause-specific mortality rate because we're only reporting the death rate due to disease X.

The case-fatality rate from disease X is equal to 18/20, or 90%.

Comparing Mortality & Case-Fatality

Population = 150,000   Deaths in
Sick = 20   one year = 18

Mortality rate

$$\frac{18}{150,000} = 0.00012 * 100,000 = 12 \text{ deaths per } 100,000 \text{ population}$$

Case-fatality rate

$$\frac{18}{20} = 0.9 * 100 = 90\%$$

From Gordis, Epidemiology, 2nd edition, page 45.

Exercise on Incidence and Prevalence

Descriptive Statistics and Tools

In public health, we frequently use statistics to gain a better understanding of our data. The next few slides will introduce you to several statistical techniques, often applied to public health data. We will discuss techniques used to summarize or describe a set of data, which allows us to simplify large amounts of data and to prepare those data for presentation. This process allows us to draw conclusions about the health of populations.

    The goal of using these statistical tools is ultimately to describe health conditions, events or behaviors in our communities in order to address key public health questions and ultimately improve health and reduce disease. So we use "descriptive statistics" to draw conclusions about large amounts of data. Some of these tools are referred to as summary measures.

    Summary measures not only include the average, or mean, but also the median and the mode. It is important to look at summary measures along with the entire data set in order to understand our data, because the same summary measures may be used to describe very different data sets.

    In addition to summary measures, we will also discuss two other commonly used statistical tools in public health practice: the confidence interval and p-value. Let's begin with summary measures.

Descriptive Statistics and Tools

- Techniques used to:
  - **Summarize**, or describe, a set of data
  - **Simplify** large amounts of data in a sensible way
- Tools
  - **Summary measures**: mean, median, mode
  - **Statistics**: confidence interval, p-value

Mean

The mean is the average value of a set of data. We calculate it by adding two or more quantities together and dividing by the number of quantities. For example, if we wanted to know the mean of two numbers, 6 and 7, we would add these two numbers and then divide by 2 (which is the number of quantities we have).

    The mean is a popular statistical measure because: it is familiar to most people; it provides useful summary information about our data, and it is easily used with other statistical measurements. The major disadvantage to the use of the statistical mean is that it can be affected by extreme values in the data set and therefore can be biased.

Mean

- The average obtained by adding two or more quantities and dividing by the **number** of quantities.

  $6 + 7 = 13 \div 2 = 6.5$ (the average or mean)

- **Advantages**: familiar; includes all data points; easy to use
- **Disadvantage**: affected by extreme values in the data set

0   mean = 25   50
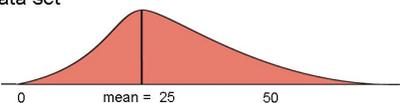
Median

The median is the midpoint, or "middle value," in a series of numbers arranged in order from small to large. Half the data values are above the median, and half are below. For example, in 2005 the median age of death in the U.S. was 75, meaning that half the people who died were older than 75 and half were younger.

If the list has an odd number of entries, the median is the middle entry in the list. If the list has an even number of entries, the median is equal to the sum of the two middle numbers divided by two.

The median, unlike the mean, is not affected by extreme data values.

Median

- The **middle value** of a set of numbers, arranged in order from small to large.
- Odd number of measurements
    x = [1 2 3 4 5 6 7]; median = 4
- Even number of measurements
    x = [1 2 3 4 5 6 7 8]; median = (4 + 5) ÷ 2 = 4.5
- The median is **not affected by extreme data values**.

0   mean = 25   50
        median = 25
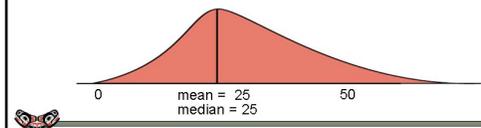
Mode

The final summary measure we will discuss is the mode.

In a list of numbers, the mode is the number that occurs most often, assuming at least one number, or data point, occurs more than once.

For example, in the following set of data [1, 3, 5, 5, 7, 9], 5 is the mode because it is the most frequently occurring value.

Some data are Unimodal, meaning they have only one mode; some data are bimodal meaning they have two modes. In this set of data [1, 3, 3, 5, 5, 7, 9], both 3 and 5 are the modes.

Mode

- In a list of numbers, the mode is the **number** that occurs most often, if there is one.
    - Unimodal = one mode
      1, 3, 5, 5, 7, 9 (5 is the mode)
    - Bimodal = two modes
      1, 3, 3, 5, 5, 7, 9 (3 and 5 are the modes)

Mean and Median Example

Let's compare the mean and median using the number of firearm-related deaths in Idaho in 2004. In the left column we see that there were a total of 177 deaths. We can also see that the range of ages of the firearm-related deaths was 6 to 93. The mean age of those who died was 46.7. The mean is the sum of all ages, divided by 177.

The median age of firearm-related deaths in Idaho in 2004 was 46. Therefore, half the cases were higher in age and half the cases were lower in age than the median of 46. In this example, the mean and median are very similar, but this is not always the case.

Mean and Median Example

Firearm-related deaths, Idaho, 2004 by age

Number of deaths	Minimum age	Maximum age	Mean age	Median age
177	6	93	46.7	46

The **mean** is the sum of all ages, divided by the total number cases.

The **median** is the AGE in which half the cases are HIGHER in age and half the cases are LOWER in age.

Source: Idaho Department of Health and Welfare, Bureau of Health Policy and Vital Statistics, June 2006

Exercises on Mean, Median, and Mode

95% Confidence Interval (CI)

Now let's look at two other statistical tools public health commonly uses. The first one is the 95% confidence interval.

A 95% confidence interval is a range of values that is used to describe how confident we are that a rate or proportion calculated from a sample of data represents the true underlying rate or proportion in the population from which the sample was drawn.

For each estimated rate, one would expect the rate to fluctuate somewhat, depending on several factors, but to remain within the confidence interval 95% of the time. So if we were to repeatedly calculate new rates from samples of our population using the same procedures each time, 95 times out of 100, we would expect the sample rates to fall within the 95% confidence interval.

The width of the confidence interval gives us some idea about how precise we are about the rate or proportion around which we've calculated the CI. A narrow confidence interval around a rate indicates that the population rate is probably quite close to the rate we observed in our sample. A wider CI indicates that our estimated rate might be further from the true population rate.

By convention, 95% confidence intervals are routinely calculated in public health, though you might also come across 99% or 90% confidence intervals.

95% CI Example

Let's take a look at an example of the 95% CI. In this example, we can look at the confidence interval and get a feel for how precise our estimates are of the annual rate of death in Idaho and in the different districts in Idaho. The annual rate of death per 100,000 people in the state is 7.2/100,000. The narrow confidence interval of 7.0 to 7.3 suggests that we can be 95% confident that the true death rate lies within this range.

We can also use confidence intervals to compare 2 rates to determine whether they are statistically different from each other. When comparing two rates, if the confidence intervals do not overlap, the difference in the rates is considered unlikely to be the result of chance. (We use the term "statistically significant" to say that something is unlikely to be the result of chance.) For example, when comparing the confidence intervals around the death rates for districts 2 and 3, we can see that they do not overlap. This suggests that the difference between the death rate of 9.0 per 100,000 for district 2 and 7.3 per 100,000 for district 3 is statistically significant.

It is worth noting that when comparing 2 rates, although non-overlapping CIs indicate a statistically significant difference between the 2 rates, the opposite is not true. In other words overlapping CIs do

95% CI Example

Residence	Annual death rate per 100,000	95% CI
Idaho State	7.2	7.0 – 7.3
District 1	8.4	8.0 – 8.8
District 2	9.0	8.4 – 9.6
District 3	7.3	7.0 – 7.7
District 4	5.9	5.7 – 6.2
District 5	8.0	7.6 – 8.4
District 6	7.5	7.1 – 7.9
District 7	6.3	5.9 – 6.7

not necessarily suggest that the 2 rates are statistically similar. In order to be sure, you would have to perform a statistical test to compare the 2 rates.

Exercise on Confidence Intervals

P-Value

Now let's take a look at the other useful statistical tool I mentioned at the beginning of the section: P-value. The p-value is frequently used in public health to determine whether observed differences between groups are 'real' differences. (Another way to say this is that the p-value is a measure of the statistical significance of a difference between rates or proportions.)

The p-value is a measure of how likely it is that the differences between two observed rates or proportions occurred by chance alone. We'll look at examples of p-values on the next slide, but for now let me just say that a very small p-value means that observed differences were very unlikely to have occurred by chance. For example, a p-value of 0.05 indicates that there was only a 5% chance that the observed differences between the two estimates you are comparing occurred by chance alone. This means that conversely, there was a 95% chance that the difference between the two estimates you observed resulted from something other than chance.

A p-value of less than .05 suggests that there was less than a 5% chance that the observed differences between the two estimates you are comparing occurred by chance alone.

It is common practice in publi health to use a cutoff of p less than .05 to establish that an oberved difference was unlikely to have occurred by chance alone.

P-Value Example

Here is an example of how to interpret p-values. In this slide, we are looking at a table of annual death rates per 100,000 people in Idaho by district. The far right column shows the p-values for the death rates for each district compared with the rest of the state.

A p-value that is less than .05 indicates that there is a statistically significant difference between the annual death rate in a certain district and the rest of the state. We can see that 5 of the p-values are less than .05. By looking at the rates themselves, we can see that in districts 1, 2 and 5, the annual death rates were statistically significantly higher than the rest of Idaho, and in districts 4 and 7, the death rates were statistically significantly lower than the rest of Idaho. Because the p-values were not less than .05 for the comparisons between district 3 and the rest of the state and district 6 and the rest of the state, we cannot say that there is a statistically significant difference between these death rates.

P-Value

- A measure of **how likely** it is that a difference in rates occurred by **chance alone**.
- A **very small p-value** means that the calculated difference in rates was **very unlikely** to have occurred **by chance**.
- If the p-value is **less than 0.05**, there is a **less than 5% chance** that the difference in rates occurred by chance alone.
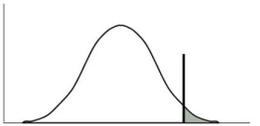
p-value = < 0.05

P-Value Example

Residence	Annual death rate per 100,000	95% CI	P-value
Idaho	7.2	7.0 – 7.3	
District 1	8.4	8.0 – 8.8	< .0001
District 2	9.0	8.4 – 9.6	< .0001
District 3	7.3	7.0 – 7.7	0.3507
District 4	5.9	5.7 – 6.2	< .0001
District 5	8.0	7.6 – 8.4	< .0001
District 6	7.5	7.1 – 7.9	0.1472
District 7	6.3	5.9 – 6.7	< .0001

Data Presentation

In the previous slides, we discussed ways to use data to measure the burden of disease in populations using disease frequency measures such as prevalence and incidence. Then we discussed ways to summarize data (such as with the mean or median) and we discussed ways to measure the precision of our estimates and make comparisons between groups (using confidence intervals and p-values). But once you have summarized your data, how do you know how to present those data in a clear and meaningful fashion?

   In the following slides we will discuss ways to present data and the strengths and limitations of each presentation format, and we will describe how to chose between different presentation options.

   Often it is difficult to determine the most appropriate way to visually display data. Although we may be comfortable with using one of these formats, the choice of graphic depends on what we want to emphasize, rather than simply trying to fit the data into a familiar framework. The next series of slides will review some common ways that data can be presented, and the strengths and limitations of each method.

   Examples of data presentation options that we will discuss are:
- Tables
- Line graphs
- Bar graphs
- Pie charts

Why Care About Data Presentation?

Why is data presentation so important?

   Data may be presented in several different ways but each presentation format has a similar goal: to organize and summarize data in a clear and accurate manner. Visual displays of data are often simpler and more understandable than standard writing, and as a result, they can make the interpretation of data easier, allowing the reader to identify and explore:
- Disease frequencies,
- Various comparisons between groups,
- Trends over time, and
- Other relationships in the data.

Tables

A table is a visual display of data arranged into rows and columns. One benefit of using tables is that they allow us to demonstrate a number of patterns or differences between groups, depending on what data are included in the table. Almost any quantitative information can be organized into a table.

Tables may take longer to read and understand than some other visual comparisons, such as graphs. A table should be as simple as possible. Because large complicated tables can be overwhelming for the reader, for clarity, sometimes it is better to create two or three small tables rather than one large table.

Although tables can be useful for presenting time trend data, sometimes other data presentation options might be preferable.

Tables

- Visual display of data **arranged in rows and columns**
- Show **patterns** between groups
- May take **longer to read** and understand than visual comparisons
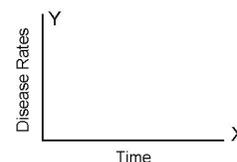- More **difficult to examine trends** or comparisons than in graphs

Table Example

This table illustrates the leading causes of death among Idaho residents below the age of 1 year in 2004. In the first column, causes of death are presented. In the second column, the number of deaths that occurred in each category is shown, and in the third column, the frequency of those cause-specific deaths is represented. In the bottom row of the table, total number of deaths is reported, and you can see that the frequencies of the cause-specific deaths add to 100%.

This table is simple and clear in its presentation. The fact that the causes of death are listed in order of how frequently they occurred makes it easy for the reader to identify the leading causes of death in this age group and perhaps to begin thinking about prevention strategies.

Table Example

Idaho resident deaths by cause of death, age less than 1 year, 2004

Cause of Death	Number	Percent
Congenital malformations	26	18.7
Short gestation/low birth weight	26	18.7
Sudden infant death syndrome	19	13.7
Complications of placenta, cord, and membrane	11	7.9
Maternal complications of pregnancy	8	5.8
Neonatal hemorrhage	6	4.3
All other causes	43	30.9
Total	139	100.0

Line Graphs

A line graph is a useful data presentation tool for showing a long series of data (such as disease trends over time). Line graphs are also useful for comparing several different series of data in the same graph. Line graphs display data in two dimensions. We call the dimensions the x-axis and the y-axis.

By convention the dependent or y variable is on the vertical axis and the x, or independent variable is on the horizontal axis. When reading a line graph, you'll notice that rises and falls in the line show how one variable is affected by another. Let's look at an example.
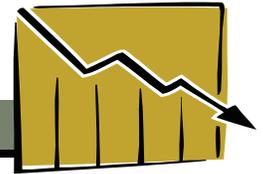
Line Graphs

- Useful for showing a long series of data and for comparing several series
  - Dependent variable: y or vertical axis
  - Independent variable: x or horizontal axis
- Rises and falls in the line show how one item is affected by another

# Data Interpretation for Public Health Professionals

Line Graph Example

Here's an example of a line graph that represents the prevalence of tobacco use during pregnancy among women in Idaho and in the U.S. between 1998 and 2002. Along the y-axis, we have the percent (or prevalence) of women who reported having used tobacco during pregnancy, and along the x-axis, we have the year. Two trends are represented here: one is the change in prevalence of tobacco use during pregnancy in Idaho, and the second is the change in prevalence of tobacco use during pregnancy in the U.S. This allows us to evaluate the changes in prevalence separately for the U.S. and Idaho but also allows us to visually compare the two trends with each other.

Line Graph Example

Tobacco Use During Pregnancy:
Idaho Resident Live Births (1998 – 2002)

For example, we can see that in 2002, the prevalence of tobacco use during pregnancy was lower for Idaho than it was for the U.S. Note that if we wanted to determine whether this difference of 11.4% in the U.S. versus 10.5% in Idaho was statistically significant, we would have to see a p-value comparing the two.
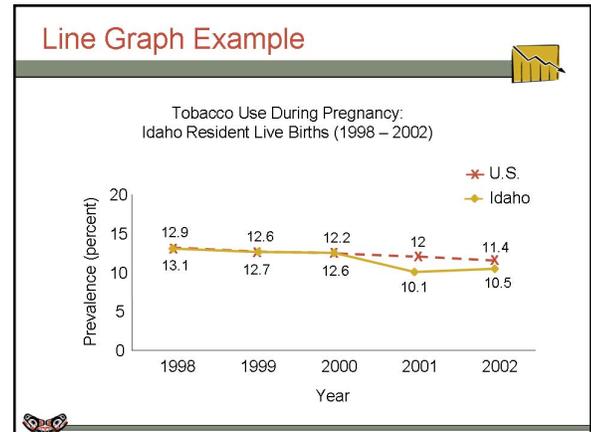
Exercise on Line Graphs

Bar Graphs

Bar graphs are also used to compare data and show relationships between two or more variables (or groups or items).

Each independent variable is discrete, such as race or gender (which only has two categories: male and female). If you wanted to display data comparing, for example, the prevalence of smoking in people of different ages, using a bar graph, you would group the age variable into categories (such as ages 15-19 or 20-24) for clarity of presentation.

Bar graphs are a quick and intuitive way to show big differences in data.

Bar Graphs

- Used to compare data and show relationships between groups or items (variables).
  - Each variable is discrete, such as race or gender.
- Bars represent categories of the variable.
- Quick way to show big differences in data.

Caution: A change in the interval or scale may alter the visual perception of data, which may persuade or trick the reader.

Vertical Bar Graph

Here we see a simple bar graph representing the proportion (or percent) of Idaho adults who had ever been told by a health care professional that they have diabetes. This is a measure of prevalence. We can see that the data have been presented for the state and also separately for each health district. What can we say about these data?

One thing we can say is that the prevalence was higher in District 6 than in all the other districts and that the prevalence was lowest in District 4. But keep in mind that we don't know whether any observed differences are real resulting from true

Vertical Bar Graph

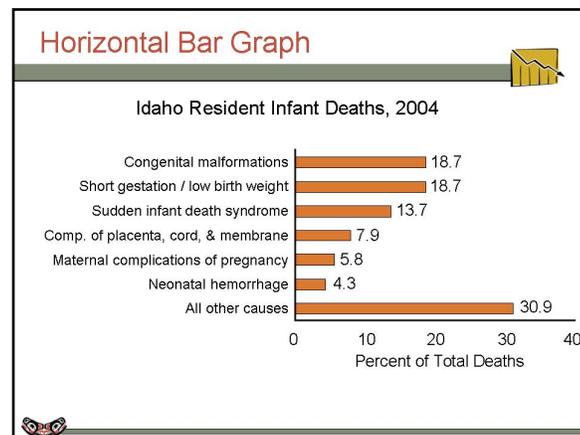Percent of Idaho adults who had ever been told they have diabetes, 2004
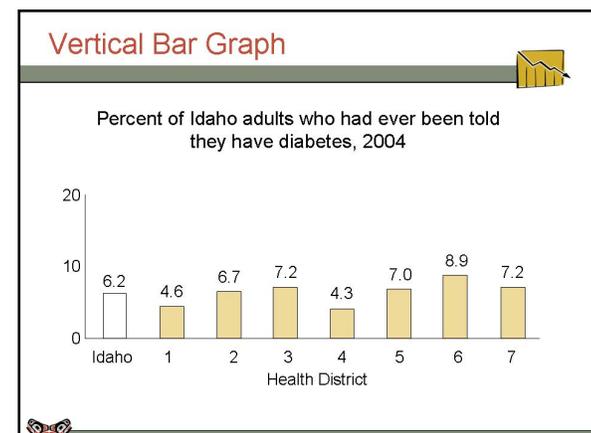
differences in the people who live in district 6 vs. district 4 for example, or whether differences in prevalence results from random error or bias in our sample of respondents.

Again, in order to determine whether differences in prevalence between two or more districts are statistically significant, we would have to see a p-value. Recall that the p-value is obtained from performing a statistical test.

Horizontal Bar Graph

Bar graphs can be displayed either vertically or horizontally. Here we see an example of a horizontal bar graph, displaying infant deaths in Idaho in 2004. Along the y axis infant deaths are displayed. Along the x axis, the percent of total infant deaths is displayed. Thus, the bars represent the percent of total deaths attributable to each cause of death.

These data could also be displayed using a vertical bar graph. Your choice of which to use can depend on personal preference, but it might be useful to construct a bar graph both vertically and horizontally to determine which presentation format is clearest given the data you are presenting.

Horizontal Bar Graph

Idaho Resident Infant Deaths, 2004

Cause	Percent
Congenital malformations	18.7
Short gestation / low birth weight	18.7
Sudden infant death syndrome	13.7
Comp. of placenta, cord, & membrane	7.9
Maternal complications of pregnancy	5.8
Neonatal hemorrhage	4.3
All other causes	30.9

Percent of Total Deaths

Pie Charts

Pie charts are frequently used to show how part of something relates to the whole. Pie charts are useful for showing the component parts of a single group or variable. The basic design is a circle, the shape of a pie, and the components, or slices of the pie, are usually percentages of the different categories of the variable.

Pie charts are a way to effectively present percentages in which the "slices" of the pie add up to 100%.

Pie Charts

• Used to show how a part of something relates to the whole

• Slices represent component parts of a single group or variable

• The "slices" of the pie add up to 100%

(Pie chart with slices: 10%, 15%, 40%, 5%, 30%)
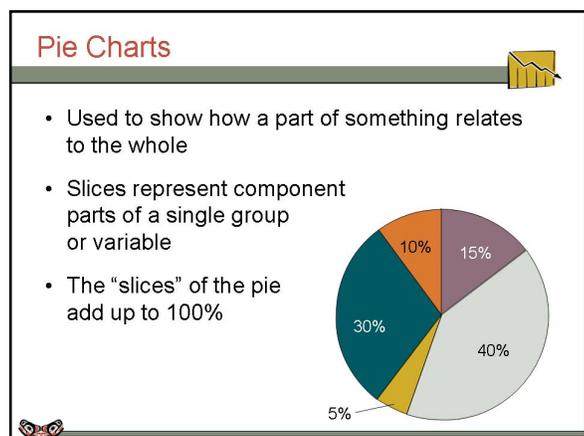
Pie Chart Example

Let's look at a pie chart. The whole pie represents the total number of deaths from unintentional injuries among children in Idaho between the ages of 1 and 4 for the times period 2000 to 2002. From this pie chart we can see that drowning or submersion- related injuries were the leading cause of unintentional injury deaths among 1–4 year olds in Idaho during that time period.

Although these data could certainly be presented in a table or using a bar graph, note how easy it is to make comparisons between the causes of death using the pie chart in this example. When a single variable is being presented (such as deaths due

Pie Chart Example

Unintentional injury child mortality aged 1 – 4: Idaho resident deaths (2000-2002) N=32

(Pie chart with slices: Drowning / submersion 34.4%, Firearm 3.1%, Other unintentional injuries 25.0%, Motor vehicle accidents 28.1%, Suffocation / strangulation 9.4%)
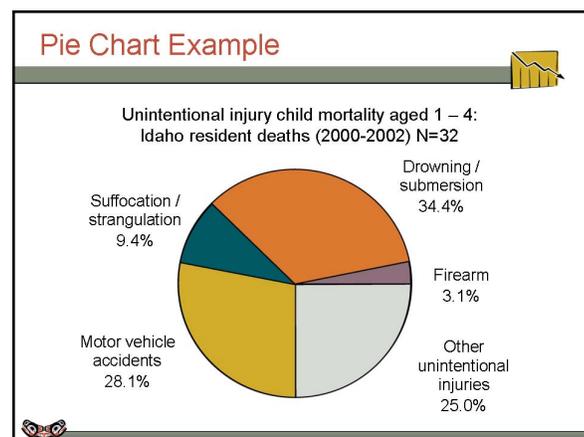
to unintentional injuries), and the information you want to convey is how parts relate to the whole (like how the causes of deaths due to unintentional injuries relate to the total number of injuries), you should consider using a pie chart to display your data.

Exercise on Display of Data

Summary

In this course, we covered some basic concepts you will need to understand and talk about public health data. These concepts include:

- Measures of disease frequency, such as prevalence, incidence, mortality, and case-fatality
- Biostatistical tools, such as the mean, median, mode, confidence interval, and p-value
- Graphical forms of displaying data, such as tables, line graphs, bar graphs and pie charts

Remember, knowing how to read understand, and interpret data specific to your community will help you better understand your community's health needs.

Resources

Here is a list of useful resources that provide further information about this topic. You can also print out the list of these resources by clicking the resources link in the attachments drop-down box located at the top of the screen.

Online Resources

CDC WONDER, http://wonder.cdc.gov/. Provides a single point of access to a wide variety of reports and numeric public health data.

E is for EPI, North Carolina Center for Public Health Preparedness. http://www.sph.unc.edu/nccphp/training/training_list/t_e_epi.htm.

Principles of Epidemiology, CDC, Second Edition, 1992. http://www.phppo.cdc.gov/phtn/catalog/pdf.

Books

Basic and Clinical Biostatistics, Beth Dawson and Robert G. Trapp. McGraw Hill, 2004.

A Cartoon Guide to Statistics, Larry Gonick and Woollcott Smith, Harper Collins, 1994.

Epidemiology, Leon Gordis, W.B. Saunders Company, 2000.

Epidemiologic Methods, Thomas Koepsell and Noel White, Oxford University Press, 2003.

Epidemiology for Public Health Practice. Robert H. Friis and Thomas A Sellers. Jones and Parlett Publishers, 2004.

Intuitive Biostatistics, Harvey Motulsky, Oxford University Press, 1995.