

PRINCIPLES OF INFECTIOUS DISEASE EPIDEMIOLOGY

MODULE IV – STATISTICAL MEASURES

Much of this module was adapted from the Centers for Disease Control and Prevention (CDC) “Principles of Epidemiology, Second Edition, An Introduction to Applied Epidemiology and Biostatistics.”

Please note: Because of formatting constraints, the formulas in the outline may not appear in correct mathematical format, however, they do appear correctly in the course module.

I. INTRODUCTION

Module IV is designed to prepare public health workers to meet the following objectives:

- Define the most common statistical frequency measures used in infectious disease epidemiology
- Construct a frequency distribution
- Calculate and interpret the following statistical measures:
 - Ratios
 - Proportions
 - Incidence rates, including attack and secondary attack rates
 - Prevalence
 - Mortality rates
 - Relative risk and odds ratio
- Choose and apply the appropriate statistical measures

II. HOW DO WE USE STATISTICAL MEASURES IN EPIDEMIOLOGY?

- Statistics are used to summarize the data collected through disease surveillance or an outbreak investigation, so we can learn from the data. Calculating statistics helps us to:
 - describe risk
 - make comparisons among communities and smaller definable groups
 - identify high-risk groups
 - develop hypotheses about the cause(s) of disease
- Why do we need to describe and compare risk?
 - Differences in risk among different populations can provide clues for investigation of what caused one group to have a higher risk.
 - If causes can be identified, then perhaps prevention and control measures can be identified too.

- The most common statistical measures used in field epidemiology are “frequency measures,” which are simply ways of counting cases and comparing their characteristics. In contrast with statistics used in epidemiological research, frequency measures are relatively easy to calculate and use.

III. FREQUENCY DISTRIBUTIONS

- When we collect data about disease cases, we must put them in some kind of order. The most basic way to do this is to organize a “line listing”. Example: <http://www.dhss.mo.gov/CDManual/CDsec30.pdf> (scroll down to page 77).
- A line listing is actually a simple database, in which each row represents a case of the disease we are investigating. Each column contains information about one characteristic, called a “variable.”
- Look at the data in **Table 1**

Table 1
Neonatal Listeriosis, General Hospital A, Costa Rica, 1989

ID	Sex	Culture Date	Symptom Date	DOB	Delivery Type	Delivery Site	Outcome	Admitting Symptoms
CS	F	6/2	6/2	6/2	Vaginal	Del Rm	Lived	Dyspnea
CT	M	6/8	6/8	6/2	C-section	Oper Rm	Lived	Fever
WG	F	6/15	6/15	6/8	Vaginal	Emer Rm	Died	Dyspnea
PA	F	6/15	6/12	6/8	Vaginal	Del Rm	Lived	Fever
SA	F	6/15	6/15	6/11	C-section	Oper Rm	Lived	Pneumonia
HP	F	6/22	6/20	6/14	C-section	Oper Rm	Lived	Fever
SS	M	6/22	6/21	6/14	Vaginal	Del Rm	Lived	Fever
JB	F	6/22	6/18	6/15	C-section	Oper Rm	Lived	Fever
BS	M	6/22	6/20	6/15	C-section	Oper Rm	Lived	Pneumonia
JG	M	6/23	6/19	6/16	Forceps	Del Rm	Lived	Fever
NC	M	7/21	7/21	7/21	Vaginal	Del Rm	Died	Dyspnea

Abbreviations:

Vaginal = vaginal delivery
 Del Rm = delivery room
 Oper Rm = operating room
 Emer Rm = emergency room

- How many of the cases were male? We can easily pick out that information because there are only a few cases.
- But with a larger database, we cannot get that information at a glance. We must summarize variables into tables called “**frequency distributions.**”
- **A frequency distribution**
 - shows the values a variable can take, and
 - the number of people or records with each value.

- Example: suppose we are investigating an outbreak in an elementary school. We could construct a frequency table that shows how many of the ill children were in each classroom.

Table 2
Distribution of cases by classroom
Brown School, Missouri, June 2005

Classroom	Number of Cases
101	25
102	43
103	32
104	0
105	8
106	2
Total	110

Please notice some key points about the table format:

- All possible values of the variable, in this case all classrooms, are listed, even if there were no cases for some values.
- Each column is properly labeled.
- The total is given in the bottom row.

Now we can tell at a glance which rooms were most affected, and which were not affected at all.

- **Variables**

- The **values of a variable** may be
 - numbers (for example, number of tacos eaten);
 - an ordered numerical scale (for example, age); or,
 - categories (for example, ill or well), called a “nominal scale” because the categories are named.
- In epidemiology we often deal with variables that have only two categories, like alive or dead, ill or well, did or did not eat the potato salad.
- Any of these types of data may be summarized in a frequency distribution (See **Table 3**, which shows a variable with only two possible values).

Table 3
Influenza vaccination status among residents
Nursing Home A, Missouri, December 2005

Vaccinated?	Number
Yes	76
No	125
Total	201

IV. RATIOS, PROPORTIONS AND RATES

- Three kinds of frequency measures are used with two-category variables (also called dichotomous variables). These frequency measures are
 - Ratios
 - Proportions, and
 - Rates.
- Before you learn about specific measures, it is important to understand the relationship between the three types of measures and how they differ. All three measures are based on the same formula:

$$\text{Ratio, proportion, rate} = x/y \times 10^n$$

- In this formula, x and y are the two quantities that are being compared.
- The formula shows that x is divided by y .
- 10^n is a constant that we use to transform the result of the division into a uniform quantity.
- The size of 10^n may equal 1, 10, 100, 1000 and so on depending on the value of n .

Example:

$$10^2 = 10 \times 10 = 100$$

$$10^3 = 10 \times 10 \times 10 = 1000$$

$$10^5 = 10 \times 10 \times 10 \times 10 \times 10 = 100,000$$

A. Ratios:

- A ratio is used to compare the occurrence of a variable in two different groups.
- These may be two completely independent groups, or one may be included in the other.
- For example, we could compare the sex of children attending an immunization clinic in either of the following ways:

$$1) \quad \frac{\text{female}}{\text{male}} \qquad \text{or} \qquad 2) \quad \frac{\text{female}}{\text{all}}$$

In the first example, x (female) is completely independent of y (male). In the second example, x (female) is included in y (all). This second type of ratio, called a proportion, is examined in more detail in the next section of this module.

B. Proportions:

- The second type of ratio, in which x is part of y , is also called a proportion (as in the previous (female/all) example).
- Proportions are usually expressed as percentages.

Examples

➤ Independent x and y :

During the first 9 months of national surveillance for eosinophilia-myalgia syndrome (EMS), CDC received 1,068 case reports that specified sex; 893 cases were in females, 175 in males. Here is the method for calculating the **female-to-male ratio** for EMS.

1. Define x and y :
 x = cases in females
 y = cases in males
2. Identify x and y :
 x = 893
 y = 175
3. Set up the ratio x/y : 893/175
4. Reduce the fraction so that one value equals 1. Female to male =
 $893/175 = 5.1/1$
5. Express the ratio in one of the following ways: 5.1 to 1, or 5.1:1, or 5.1/1

Thus, there were just over 5 female EMS patients for each male EMS patient reported to CDC.

➤ x included in y :

Based on the same data, here is the method for calculating the **proportion** of EMS cases that were male.

1. Define x and y :
 x = cases in males
 y = all cases
2. Identify x and y :
 x = 175
 y = 1,068
3. Set up the ratio x/y : 175/1,068
4. Reduce the fraction so that one value equals 1. Divide the smaller number by the larger number: $175/1,068 = 0.16/1$

5. Proportions are usually expressed as percentages, so the value of the constant (10^n) = 10^2 = 100:
 $0.16 \times 100 = 16$ (16%)

Thus, 16% of the reported EMS cases were in males.

C. Rates:

- The third type of frequency measure used with two-category (dichotomous) variables is a rate.
- Rates have the added dimension of time. Rates measure the occurrence of an event in a population over time.
- The basic formula for a rate is:

$$\text{Rate} = \frac{\text{number of cases occurring during a given time period}}{\text{population at risk during the same time period}} \times 10^n$$

- Rates are always specific to a particular population. They reflect groupings of people based on time, place and person.
 - Time: a specific year, month, week, day or hour
 - Place: country, state, county, city, township, school, institution, area
 - Person: age, sex, membership in some group or class
- Rates take into account the size of the population, so comparisons can be made across different population groups.
 - By using rates instead of raw numbers, the occurrence of disease in one group can be fairly compared with another.
 - Example: males with females; one county with another; Missouri with Arkansas or the US.
- To calculate a rate, we must have an estimate of the population at risk during a specific time period for the denominator.
 - Ratios and proportions do not require this.
 - Earlier, we calculated ratios and proportions of EMS cases without knowing the number of people at risk of EMS.
 - Rates may be harder to get, because accurate denominator data may not be available for small, localized population groups.

To summarize:

- All three of these frequency measures are calculated in basically the same way. In practice, we use:
 - a **ratio** to compare two independent groups,
 - a **proportion** to compare one group with a larger one to which it belongs, and
 - a **rate** to measure an event in a population over time.
- Ratios, proportions, and rates are used in infectious disease epidemiology to describe **morbidity** (disease) and **mortality** (death).

V. MORBIDITY FREQUENCY MEASURES

- Several standard measures are used to measure and describe the frequency of disease.
- Each measure has its appropriate uses, depending on the situation and the information available to the epidemiologist.
- The two main types of rates are:
 - **incidence**
 - **prevalence**.

A. Incidence Rates

- Incidence rates are the most common way of measuring and comparing the frequency of disease in populations.
- Incidence is a measure of risk.
- When Population A has a higher incidence of a disease than Population B, we can say that Population A has a higher risk of developing the disease than Population B. If it is a lot higher, we could say that Population A is a **high-risk** group for that disease.
- **Table 4** shows the three types of incidence rates we will study, along with their formulas. We will discuss each of these in more detail.

Table 4
Frequently Used Measures of Morbidity

Measure	Numerator (x)	Denominator (y)	Expressed per Number at Risk (10^n)
Incidence Rate	# new cases of a specified disease reported during a given time interval	Average population during time interval	Varies: 10^n where $n = 2, 3, 4, 5, 6$
Attack Rate	# new cases of a specified disease reported during an epidemic period	Population at start of the epidemic period	Usually a percentage: 10^n where $n = 2$
Secondary Attack Rate	# new cases of a specified disease among contacts of known cases	Size of contact population at risk	Usually a percentage: 10^n where $n = 2$

We will discuss each of these in more detail.

1) Incidence

The basic **incidence rate** (sometimes called just **incidence**) is a measure of the frequency with which a disease occurs in a population over a period of time. The formula for calculating an incidence rate is:

$$\text{Incidence Rate} = \frac{\text{new cases occurring during a given time period}}{\text{population at risk during the same time period}} \times 10^n$$

- **The numerator (x)**
 - should include only **new** cases of the disease that occurred during the specified period.
 - should not include cases that occurred or were diagnosed earlier.
 - This is very important when working with chronic infectious diseases such as tuberculosis, malaria and HIV.

- **The denominator (y) is the population at risk.**
 - This means that the people included in the denominator should be able to develop the disease in question during the time period covered. In practice, we usually use census data for the denominator.
 - The denominator should also represent the population from which the cases in the numerator arose. The population may be defined by geographic area (e.g., St. Francois County) or by membership in a specific group (e.g., employee of Company X, student at School Y). If we are studying a specific group such as students in a school or residents in a long term care facility, we should use a census of that population for an exact denominator.

- **Any value of *n*** may be used in calculating incidence.
 - The epidemiologist should always make it clear what *n* value was used. National surveillance systems use a value of 10^5 or 100,000.
 - A good rule of thumb is to choose a value for 10^n so that the smallest rate calculated is a small whole number (for example, $4.2/100$, not $0.42/1,000$ or $0.042/10,000$). This is just easier for the reader to understand.
 - When comparing two incidence rates, always be sure that the same value of *n* was used in calculating both rates.

Disease incidence rates imply a change over time, from health to disease. So **the period of time must be specified**. For surveillance purposes this is usually one calendar year, but any time period may be used as long as it is stated.

Example

In 2003, 335,104 new cases of gonorrhea were reported among the US civilian population. The 2003 mid-year US civilian population was estimated to be 290,788,976. For these data we will use a value of 10^5 for 10^n . We will

calculate the 2003 gonorrhea incidence rate for the US civilian population using these data.

1. Define x and y : x = new cases of gonorrhea in US civilians during 2003
 y = US civilian population in 2003
2. Identify x , y , and 10^n : $x = 335,104$
 $y = 290,788,976$
 $10^n = 10^5 = 100,000$
3. Calculate $(x/y) \times 10^n$:

$$\frac{335,104}{290,788,976} \times 10^5 = .001152 \times 100,000 = 115.2 \text{ per } 100,000 \text{ or}$$

approximately 1 reported case per 1,000 population

The **numerator** of this incidence rate

- reflects new cases of gonorrhea that occurred or were diagnosed during the specified period.
- does **not** include cases that occurred or were diagnosed earlier.

The **denominator** is the population at risk.

- it represents the population from which the gonorrhea cases arose, in this case the US civilian population.

Notice that the numerator was limited to civilian cases. Therefore, we had to restrict the denominator to civilians as well.

2) Attack Rate

An **attack rate** is a specific type of incidence rate. It is calculated for a narrowly defined population observed for a limited time, such as during an outbreak. It is usually expressed as a percentage, so 10^n equals 100.

For a defined population (the population at risk), during a limited time period:

$$\text{Attack Rate} = \frac{\text{\# of new cases among the population during the period}}{\text{population at risk at the beginning of the period}} \times 100$$

- The attack rate is a measure of the **probability** or **risk** of becoming a case.
- Remember, attack rates have some special characteristics:
 - Highly specific by “person” variables
 - Limited by “place” variables
 - Time period is usually brief
 - Usually expressed as a percentage

Example

Of 75 persons who attended a church picnic, 46 subsequently developed a gastrointestinal illness. To calculate the attack rate of GI illness we first define the numerator and denominator:

x = Cases of GI illness occurring within the incubation period
 for GI illness among persons who attended the picnic = 46
 y = Number of persons at the picnic = 75
 Then, the attack rate for GI illness is $\frac{46}{75} \times 100 = 61\%$

In this example, we could say that among persons who attended the picnic, the probability of developing GI illness was 61%, or the risk of developing GI illness was 61%.

- Attack rates are usually calculated several times during the course of an outbreak investigation. The first time, early in the outbreak, the attack rate might be calculated as follows:

$$\text{Attack Rate} = \frac{\# \text{ of new cases in the community during the period}}{\text{the population of the community}} \times 100$$

- This would help us see whether the current situation is “unusual” compared with other time periods or communities.
- In the course of an investigation, attack rates will be recalculated as new cases are identified, diagnoses are confirmed, and other information comes to light.
- Attack rates may be needed for various subgroups within the community.
- Selecting the appropriate numerators and denominators is very important and can be a challenge.
 - Ideally, the denominator should include the smallest definable area or group that contains all the known cases.
 - In practice, however, accurate population counts may not be available.
 - For example, we may not be able to find out exactly how many people ate in a particular restaurant during a specified time.

3. Secondary Attack Rate

A **secondary attack rate** measures the frequency of new cases of a disease among the contacts of known cases. This can be very important for diseases that are spread from person to person, such as tuberculosis, measles, shigellosis, and varicella. The formula is:

$$\text{Secondary Attack Rate} = \frac{\# \text{ cases among contacts of primary cases during the period}}{\text{total number of contacts}} \times 100$$

- Secondary attack rates are often calculated for households.
 - To calculate the number of household contacts (denominator), subtract the number of primary cases from the total number living in the households.

- In some situations, contacts in other settings may be investigated (for example, residents of a homeless shelter, or people who work in a specific building). The calculation is done in the same way as for household contacts.

Example

Seven cases of hepatitis A occurred among 70 children attending a childcare center. Each infected child came from a different family. The total number of persons in the 7 affected families was 32. One incubation period later, 5 family members of the 7 infected children also developed hepatitis A. We will calculate the attack rate in the childcare center and the secondary attack rate among family contacts of those cases.

1. Attack rate in childcare center:

x = cases of hepatitis A among children in childcare center = 7
 y = number of children enrolled in the childcare center = 70

$$\text{Attack rate} = \frac{x}{y} \times 100 = \frac{7}{70} \times 100 = 10\%$$

2. Secondary attack rate (refer to Figure 1):

x = cases of hepatitis A among family contacts of children with hepatitis A = 5

y = number of persons at risk in the families (total number of family members – children already infected) = 32 – 7 = 25

$$\text{Secondary attack rate} = \frac{x}{y} \times 100 = \frac{5}{25} \times 100 = 20\%$$

B. Rate Ratio is another tool that is helpful for comparing rates between groups.

A **rate ratio** compares the rates of disease in two groups that differ by demographic characteristics or exposure history. The rate for the group of primary interest is divided by the rate for a comparison group.

$$\text{Rate Ratio} = \frac{\text{rate for group of primary interest}}{\text{rate for comparison group}} \times 1$$

Rate ratios may be calculated for incidence rates (including attack rates) or for mortality rates, discussed later.

Example

The Association of Interested Persons held their annual conference during the first week in June. There were two events: a dinner meeting on Wednesday evening (75 attendees), and a luncheon awards ceremony on Thursday at noon (60 attendees). Twenty (20) of the 75 Wednesday dinner participants subsequently developed signs and symptoms of gastrointestinal illness; 5 of the 60 luncheon participants became ill. Calculate the rate ratio to help determine which event may have been the source of the illness. The rate ratio is calculated as follows:

1. Calculate the attack rate for the dinner meeting:
 x = number of ill persons attending the dinner meeting
 y = number of persons attending the dinner meeting
 $\text{attack rate} = (x/y) \times 100 = (20/75) \times 100 = 27\%$
2. Calculate the attack rate for the luncheon:
 x = number of ill persons attending the luncheon
 y = number of persons attending the luncheon
 $\text{attack rate} = (x/y) \times 100 = (5/60) \times 100 = 8\%$
3. Calculate the rate ratio:
 $\text{Rate ratio} = \frac{\text{rate for group of primary interest}}{\text{rate for comparison group}} \times 1 = (27/8) \times 1 = 3.4$

The dinner meeting attendees were 3.4 times more likely to become ill than those who attended the luncheon.

Now that you know how to calculate and use each type of morbidity rate, you have mastered some important tools for investigating infectious diseases.

C. Prevalence (Prevalence Rates)

Prevalence is the proportion of people in a population who have a particular disease at a specified point in time, or over a specified period of time.

- The numerator includes not only new cases, but also old cases (people who remained ill during the specified point or period in time). A case is counted in prevalence until death or recovery occurs.
- This makes prevalence different from incidence, which includes only new cases in the numerator.
- Prevalence is most useful for measuring the burden of chronic diseases such as tuberculosis, malaria and HIV in a population. The formula for calculating prevalence is:

$$\text{Prevalence} = \frac{\text{all new and pre-existing cases during a time period}}{\text{population during the same time period}} \times 10^n$$

Point vs. Period Prevalence

The amount of disease present in a population obviously changes over time. Sometimes, we want to know how much of a particular disease is present in a population at a single point in time, a sort of snapshot view.

- **Point Prevalence:** For example, we may want to find out the prevalence of TB in Community A today. To do that, we need to calculate the **point prevalence** on a given date.
 - The numerator would include all known TB patients who live in Community A that day. That information could be determined from a TB case registry.
 - The denominator would be the population of Community A that day.

- Example: A review of patients reported to the tuberculosis registry in Midville revealed that as of July 1, 2005 there were 35 cases that had not yet completed therapy. The most recent population estimate for Midville was 57,763. The prevalence of TB in Midville on July 1, 2005 was:

$$\frac{35}{57,763} \times 10,000 = 6.1 \text{ per } 10,000 \text{ people}$$

- Point prevalence is useful in comparing different points in time to help determine whether an outbreak is occurring. In this case, we could also calculate point prevalence of TB for July 1, 2004, July 1, 1995 or other relevant points of comparison.

- **Period Prevalence:** At other times, we want to know how much of a particular disease is present in a population over a longer period. We would use **period prevalence** to do that.

- Period prevalence is calculated in exactly the same way as point prevalence, except the numerator is the number of people who had the disease at any time during a specified time period.
- Period prevalence can be calculated for a week, month, year, decade, or any other specified length of time.
- Example: Midville's TB registry indicates that during 2004, there were 89 new and pre-existing TB cases. The prevalence of TB in Midville in 2004 was:

$$\frac{89}{57,763} \times 10,000 = 15.4 \text{ per } 10,000 \text{ people}$$

- When comparing prevalence figures:
 - be sure that the length of the time period is the same. The prevalence of TB in Midville in 2004 should only be compared with the prevalence during other one-year time periods (2003, 1994, etc.).
 - It cannot be compared with the point prevalence on July 1, 2005, or with the period prevalence during a week or month.
 - Prevalence may be compared among different diseases or different populations, as long as the same length of time is used.

Example: Comparing Prevalence and Incidence

Two surveys were done of the same community 12 months apart. Of 5,000 people surveyed the first time, 25 had antibodies to histoplasmosis. Twelve months later, 35 had antibodies, including the original 25. We will calculate the prevalence at the second survey, and compare the prevalence with the 1-year incidence.

1. Prevalence at the second survey:
 - x = antibody positive at second survey = 35
 - y = population = 5,000
 - $(x/y) \times 10^n = 35/5,000 \times 1,000 = 7 \text{ per } 1,000$

2. Incidence during the 12-month period:

x = number of new positives during the 12-month period = $35 - 25 = 10$

y = population at risk = $5,000 - 25$ (already infected) = $4,975$

$(x/y) \times 10^n = 10/4,975 \times 1,000 = 2$ per 1,000

Prevalence is based on both incidence (risk) and duration of disease.

High prevalence of a disease within a population may reflect high risk, or it may reflect prolonged survival without cure. Conversely, low prevalence may indicate low incidence, a rapidly fatal process, or rapid recovery.

VI. MORTALITY FREQUENCY MEASURES

- Mortality rates measure the frequency of occurrence of death in a defined population during a specified interval.
- There are several specific kinds of mortality rates, but we will focus only on the ones that are used most often in infectious disease epidemiology.
- To calculate a simple mortality rate, we need to know the number of deaths in a given population during a specified time period, and the size of the population in which the deaths occurred. The basic formula is:

$$\text{Mortality rate} = \frac{\text{deaths occurring during a given time period}}{\text{size of the population in which the deaths occurred}} \times 10^n$$

The most commonly used values for 10^n are 1,000 and 100,000.

A. Crude Mortality Rate

- The crude mortality rate is the mortality rate from all causes of death for a population during a specified time period.
- The denominator is the population at the mid-point of the time period.
- For example, the crude mortality rate for Missouri in 2003 was 896 deaths per 100,000 people.

B. Cause-specific Mortality Rate

- This is the mortality rate from a specified cause for a population during a specified time period.
- The numerator is the number of deaths from that cause, and the denominator remains the size of the population at the mid-point of the time period.
- For example, the tuberculosis death rate for the US in 2002 was 0.3 per 100,000 (or 3 per 1,000,000).

C. Other Specific Mortality Rates

- Specific mortality rates may be calculated for population subgroups defined by age, sex, race, or other demographic factors.
- Combinations of factors are often used.

- For example, the mortality rate attributed to HIV among 25-to-44-year-olds in the US in 1987 was:

$$\text{HIV Mortality Rate} = \frac{9,280 \text{ deaths}}{77,600,000 \text{ aged 25-44 yrs}} \times 100,000 = 12 \text{ per } 100,000$$

This is an example of a cause- and age-specific mortality rate.

VII. RELATIVE RISK AND ODDS RATIO

The last two types of frequency measures we will study are **relative risk** (also called **risk ratio**) and **odds ratio**. These statistics are used in outbreak investigations and will be discussed again in the workshop portion of this course.

A. Relative Risk or Risk Ratio (RR)

- Compares the risk of disease or death in two groups.
 - The two groups may be defined by a demographic factor such as sex (for example, male vs. female).
 - More commonly, they are defined by a difference in their exposure to a suspected risk factor for disease (for example, ate the potato salad or didn't).
 - Often, the group of primary interest is labeled "exposed," and the comparison group is called "unexposed."
 - The group of primary interest goes into the numerator, and the comparison group is the denominator.

$$\text{Risk Ratio (Relative Risk)} = \frac{\text{Risk for group of primary interest}}{\text{Risk for comparison group}} \times 1$$

- "Risk" is defined as an incidence rate or attack rate of the disease in each group. To calculate RR, a two-by-two table is set up as shown:

Table 5
Number of Cases of Disease X by Sex, Smallville, 2004

	Disease X		Total
	Yes	No	
Female	a 46	b 1,438	1,484
Male	c 18	d 1,401	1,419

The term "two-by-two" refers to the two variables (sex and disease status), each with two categories. The outcome (illness or not) is shown at the top of the table and exposure or risk factor is shown along the left side. Note the letters assigned to each cell of the table (a-d). They are important in calculating the risk in each group.

Example

Using the data in the table above, we can calculate the relative risk of Disease X for females vs. males. First, we must calculate the risk of illness among females and among males:

$$\text{Risk among females (incidence)} = \frac{a}{a+b} = \frac{46}{1,484} = .031 \times 100 = 3.1\%$$

$$\text{Risk among males (incidence)} = \frac{c}{c+d} = \frac{18}{1,419} = .013 \times 100 = 1.3\%$$

To calculate the RR for females vs. males, females are considered the group of primary interest and males are the comparison group. The formula is:

$$\text{Risk ratio (Relative Risk)} = \frac{3.1\%}{1.3\%} = 2.4$$

So we can say that the risk of Disease X in females appears to be 2.4 times higher than the risk in males. If the RR is 1.0, that means the risk of disease is equal in the two groups. If the RR is greater than 1.0, then the group of interest has a higher risk of disease. If the RR is less than 1.0, then the group of interest has a lower risk of disease.

- However, before we can interpret RR figures, they must be subjected to a test of statistical significance such as the Chi square or some variation of it. This helps us judge the probability that the result could have occurred by chance alone. A probability of less than 5%, expressed as $p < .05$, is commonly used as a cutoff for statistical significance in field epidemiology.
- We will not teach more about statistical significance in this course, but the student should be aware that RR is affected by factors such as population size, and cannot stand alone. Statistical consultation is readily available from DHSS for the field epidemiologist.

B. Odds Ratio

The RR can only be calculated if incidence data are available. The **Odds Ratio (OR)** may be used in situations where we do not have denominator data to calculate incidence rates.

- The odds ratio is used frequently in case/control studies, which we will cover in more detail in the workshop portion of this course.
- In a case/control study, ill persons' characteristics and exposures are compared with those of well persons ("controls") selected from the same population in which the outbreak occurred.
 - Example: in an outbreak suspected to stem from exposure to contaminated food at a restaurant, the ill persons' food selections

could be compared with those of some well people who also ate at the restaurant the same day. This could be done even if we didn't know exactly how many people ate at the restaurant that day.

- A two-by-two table is constructed, just like the one used to calculate RR, with the same letters (a-d) used to label the cells. The OR is calculated by multiplying across the cells.

Example:

Table 7
Number of Cases of Disease X by Exposure History, Smallville, 2004

	Disease X		Total
	Yes	No	
Ate Tuna Casserole	a 46	b 25	71
Didn't Eat Tuna Casserole	c 18	d 40	58

The formula for OR is:

$$\text{Odds Ratio} = \frac{ad}{bc}$$

Where

- a* = number of persons with disease and with exposure of interest
- b* = number of persons without disease, but with exposure of interest
- c* = number of persons with disease, but without exposure of interest
- d* = number of persons without disease and without exposure of interest

a+c = total number of persons with disease ("cases")

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

The OR in this example is:

$$\text{Odds Ratio} = \frac{46 \times 40}{25 \times 18} = \frac{1840}{450} = 4.1$$

So those who became ill were 4.1 times as likely to have eaten the tuna casserole. We should probably look a little more deeply into the tuna casserole! We would still need to subject this result to a test of statistical significance (just like we do with the RR) to judge the probability that the result could have occurred by chance alone.