Definition of Epidemiology

A commonly cited definition for Epidemiology is "*the study of the distribution and determinants of disease frequency in man*" (MacMahon and Pugh. Epidemiology: Principles & Methods. Second Edition. Little, Brown and Company; Boston. 1970).This definition implies two quantitative aspect of epidemiology:

1. Measuring **disease distribution** in regards to person, place, and time (**descriptive epidemiology**), and

2. Measuring the association between a **disease and it determinants** (**analytic epidemiology**).

The first aspect of this definition refers to measuring disease (outcomes). Common examples include measuring the existence of disease in a population (**prevalence**) and the occurrence of disease in a population when it is followed over time (**incidence**). The second aspect of this definition refers to measuring associations between determinants (risk factors) and disease. The primary concern of analytic epidemiology is etiology by identifying the causes (risk factors) of a disease and quantifying the magnitude of their effects on the disease. Such a process begins with a **measure of association** between a risk factor and a disease, followed by an argument that this measure reflects the causal effect of that risk factor on the disease. In practice, there are often many explanations for an observed association between a risk factor and a disease. Arguing that an observed measure of association reflects a causal effect involves eliminating other possible explanations for the observed association. Hennekens and Buring (Epidemiology in Medicine. Little, Brown and Co; Boston: 1987) state that these alternative explanations fall under the broad headings of

1. **Confounding**
2. **Bias**
3. **Chance**

Clinical Epidemiology often focuses on the (causal) outcomes of an intervention (e.g. treatment). It also involves the prediction of disease (diagnosis and prognosis). The Final Report of the HSPH Department of Epidemiology Committee on the Status of Clinical Epidemiology (Singer at al, Fall, 2009 – personal communication) defined **Clinical Epidemiology** as follows:

> *Clinical epidemiology applies the concepts and techniques of epidemiology, statistics, and decision analysis to clinical problems. Clinical epidemiology emphasizes the study of patients, physicians, and systems of health care.* ***The focus is on diagnosis, prognosis, and treatment of disease but it also studies the etiology of disease.*** *Exposures and outcomes are informed by clinical and biologic knowledge as well as broader environmental and societal determinants.* ***Distinctive areas of clinical epidemiology are the development of diagnostic and prognostic models.*** *The gold standard for the study of medical treatments is the*

randomized clinical trial. Observational analyses of treatments are necessary to gain estimates of their impact in real world clinical practice and to screen for rare effects. These observational analyses of medical interventions face the distinctive hurdle of confounding by indication and/or contraindication

Historical Development of Epidemiology

The methods used in epidemiology research have evolved over time and newer methods are being developed. We owe much to key people in our past, who developed and refined these methods. Much of the following was taken from the textbook by Aschengreau and Seage *(Essentials of Epidemiology in Public Health, Second Edition. Jones and Bartlett Publishers. Sudbury 2008)* but can be found in other texts on epidemiology.

The oldest record of an epidemiology study that I could find is from the Bible and the **Book of Daniel** (1:8 -16). Daniel was one of those young Jewish men who were taken off to Babylon during the Babylonian Captivity, about 600 BC. The plan was to train Daniel and some of his companions for the king's service. As part of this training program, Daniel and his colleagues were offered the same food and wine that was served at the King's table. For religious reasons, Daniel and some of his colleagues resisted eating that particular diet, and requested a more traditional diet. The steward for the chief chamberlain resisted this request for fear of punishment by the king if Daniel and his colleagues looked "wretched by comparison with other young men of your age". In response Daniel requested a trial:

"*Please test your servants for 10 days. Give us vegetables to eat and water to drink. Then see we how we look in comparison with the other young men who eat from the royal table*."

By no means does this example imply that the Bible should be read as a text in epidemiology, but this passage contains many of the elements of a classical epidemiology study.

First, a clear research question is specified: does a diet on vegetables and water lead to better outcomes than a traditional diet. The exposure of interest is the diet of the young men put into the king's service. The exposure category of interest is the diet of vegetables and water and the non-exposure category is a standard diet of food and wine from the king's table. The outcome of interest is the appearance of the study participants after 10 days on either of these two diets. The study is limited to young men in the king's service. This limitation increases the baseline comparability of the two groups under study (but also limits the generalizability of the results). No mention is given on how the two groups might differ in other baseline characteristics that might influence the outcome (potential confounders). In addition, the outcome appears is a subjective evaluation of the overall appearance of all study subjects by a single person (steward of the chief chamberlain). No mention is given about the criteria used for this assessment or if the

steward is “blinded” to the diet of a study participant when making an outcome assessment. It is possible that the knowledge of the diet of a participant might influence the outcome classification, introducing a potential for a measurement bias in this study. The topics of confounding and bias will be discussed in future lectures of this course.

Despite the short duration of follow-up, the results from this study were striking:

“*after 10 days, they looked healthier and better-fed than any of the young men who ate from the royal table*”

About 200 years later, the teachings of the Greek physician **Hippocrates**, the father of modern medicine, (http://en.wikipedia.org/wiki/Hippocrates) were recorded and included the following quotation:

> *"Whoever wishes to investigate medicine properly should process thus-- in the first place, consider the seasons of the year, and the effects of each of them. Then the winds, the hot and the cold. One should consider most attentively the waters in which the inhabitants use, and the mode in which the inhabitants live, and what are their pursuits, whether they are fond of drinking, and eating to excess, and given to indolence, and are fond of exercise and labor." (Hippocrates. On Air, Water, and Places. Translated and republished in Medical Classics 3:19-42, 1938)*

In this passage Hippocrates points out that there are characteristics of people and their environment that influence their risks of developing disease. Disease development is not a totally random event and the goal of epidemiology is to identify those risk factors that influence the risk of disease and quantify their effects.

Roughly 2000 years, the British investigator **John Graunt** (http://en.wikipedia.org/wiki/John_Graunt), published the "Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality". The bills of mortality were routinely collected and reported mortality data, describing who was dying, and from what cause. John Graunt used those weekly counts of death to look at associations between characteristics of people and their causes of death. He reported yearly and seasonal mortality trends. He reported the common and the uncommon causes of disease. He reported other findings, including that men were at higher risk of dying than women. An important feature of Graunt’s work was the use of routinely collected data to identify associations between characteristics of individuals and mortality.

About 100 years later, a Scottish physician, **James Lind**, performed one of the first clinical experiments (clinical trial). He performed this experiment to identify the appropriate treatment for the disease scurvy. We know today that scurvy is caused by a deficiency of vitamin C in a person's diet, but in the 1700s, when Lind was treating such people on British ships, he decided to do a study to test possible treatments for scurvy. He enrolled 12 people who had been diagnosed and were sickened with this disease. He

divided them into six treatment groups, each containing two people. All 12 participants received a common diet, but each group received something in addition. One group received a quart of cider. One group received the elixir vitriol (sulfuric acid). One group received vinegar. One group received seawater. One group received a spicy paste made of mustard and garlic, along with some barley water. The last group received two oranges and one lemon. Lind followed all groups and reported that recovery was best for the group that was given the two oranges and the lemon.

A key point of Lind's study is that it is an experimental study. Unlike Daniel and his colleagues who self-selected to take the diet of vegetables and water, the participants in Lind's study were assigned by Lind to the treatment groups. In an experiment, the investigator decides which treatment a person receives, not the individual themselves. This often raises ethical issues. In addition, unlike Lind's trial, assignment in experimental studies is often determined in a random fashion. Randomization, ethics and other features of experimental studies will be discussed in future lectures.

Two major figures in the development of the epidemiology methods, **James Farr and John Snow**, lived in the next century and identified the cause of cholera, a major health problem at that time. Information about these investigators can be found at

http://www.ph.ucla.edu/epi/snow.html

http://en.wikipedia.org/wiki/William_Farr

William Farr was essentially the chief statistician for Great Britain. He was the statistical superintendent of the general register office of England and Wales in the 1800s. He was responsible for collecting and routinely distributing information about the health of the population, for example reporting causes of mortality for different regions of the country. During a cholera outbreak in London he published weekly reports, describing the number of deaths from cholera by such factors as age, sex, air temperature, wind, rainfall, day of the week, elevation, crowding, and property value.

The reported relationship between elevation and mortality is particularly noteworthy. At the time, when Farr was making these reports, the common belief was that the primary cause of cholera was essentially polluted air, miasma. The mechanism for contracting cholera was by breathing harmful particles into your body. Farr believed in the miasma theory. He supported this belief by showing people who lived at low elevations where the air might be denser, had higher mortality rates compared to people who lived at higher elevations, where the air was less dense.

At the same time Farr was reporting these results, John Snow, an anesthesiologist, was developing evidence to support an alternative hypothesis about the cause of cholera. As an anesthesiologist, Snow had knowledge about gases. He didn't believe that it was the bad air that was causing people to contract and die cholera. As a physician, he cared for people who contracted cholera. He noticed their signs and symptoms, in particular their complaints of belly pains. He believed that the cause of disease was not breathing

polluted air, but ingesting polluted water. Snow was able to use prior knowledge and observations based on data to argue his hypothesis. Using data from Farr's reports, he was able to perform two important landmark studies, to convince people, including Farr, to disprove the miasma theory and successfully argue for polluted water as the cause of cholera.

One of the studies was "The Great Experiment," was initially based on data published by Farr showing the mortality rates of cholera for areas of London served by different water companies. Some of these water companies took their water from the polluted portions of the Thames River. Others took it from a less polluted area upstream from London. It was Snow who noticed that people who were receiving water from the less polluted sources had a lower mortality rate from cholera than those who were receiving their water from the polluted areas of the Thames River

In another study Snow investigated an outbreak of cholera in a particular neighborhood of London, the Soho District, during 1854. Reading the weekly results that were reported by Farr, Snow noticed that this region of London suddenly had a huge outbreak of cholera, whereas the mortality from cholera during prior weeks was very low. Snow gathered additional data on the inhabitants of this region (primary data collection), and noticed a clustering of deaths near a public water pump in that area. He was able to build a convincing argument that the likely source of cholera in that region was polluted water from the well served by that pump.

Snow and Farr show the combination that's needed for epidemiology and clinical epidemiology research. On one hand, we need hypotheses suggesting potential risk factors for causing disease. Often these hypotheses are based on routine reports, other investigations, or personal experience. Snow’s background as a physician and an anesthesiologist provided the background for his hypothesis. Snow believed that polluted water was the potential culprit for cholera, but he needed data, partially furnished by Farr, to support hypothesis.

A more recent contributor to epidemiology methods is **Austin Bradford Hill**. In the 1940s, Hill was the statistician on a clinical trial investigating a new treatment for tuberculosis, streptomycin. Hill’s study enrolled homogeneous cases of tuberculosis. One group of patients received the standard treatment of bed rest. Another group received bed rest plus a new potential treatment, streptomycin. Hill used a random device to assign patients to these two groups so that roughly half the people got streptomycin, and half did not. He had only a limited supply of streptomycin available, enough for approximately half the participants in the study, so he could argue that random allocation to patients is an ethical way to decide who should get a new treatment. Hill also introduced the notion of blinding when assessing outcomes. The outcomes were determined by looking at x-ray films, and readers of these ray films, the investigators interpreting these outcomes, were blinded on whether or not a study participant received streptomycin or didn't receive streptomycin. Therefore, the knowledge of the assigned treatment could not influence the outcome decision. We'll discuss the role of randomization, blinding, and ethics in experimental studies in a future lecture.

A colleague of Bradford Hill, **Richard Doll**, did a series of landmark epidemiology studies in the 1950's. Doll examined the association between smoking and lung cancer in two important studies. The first was a very large **case control study**, where Doll enrolled compared past smoking habits for people who developed lung cancer (cases) and for people who didn't develop lung cancer (controls). The second study was a **prospective cohort study**, where he enrolled people who were free of lung cancer, smokers and non-smokers, follow them forwards in time, to see which group developed lung cancer more rapidly, more commonly. These study designs will be discussed in detail in later lectures in this course.

A final example of an important contribution to the development of epidemiology methods is not a person but a study, the **Framingham Heart Study**. This study will be briefly discussed in these series of lectures and discussed in more detail later in this course.

The Framingham Heart Study

The Framingham Heart Study started in 1948. Framingham, Massachusetts is located about 10 miles west of Boston. In 1948 it had both urban and rural aspects. It had been involved in previous tuberculosis study, so the people might be willing to be interviewed, and be involved in another research study to identify the risk factors for developing cardiovascular disease (CVD). The phase "**risk factor**" was coined by the investigator in the Framingham Heart Study. The study enrolled 5,209 men and women, aged 30 to 62. The selected age range of the participants reflects the "dark side" of epidemiology. Epidemiologists typically observe disease occurring among people and identify the risk factors that cause these outcomes. This requires enrolling enough participants into a study to observe a sufficient number of disease cases for finding associations between risk factors and disease. The plan for the Framingham heart study was to enroll people in 1948, follow them for 20 years, and observe who developed cardiovascular disease (CVD). To observe a sufficient number of cases of CVD during follow-up may require a large number of low-risk study subjects, or a smaller number of higher risk subjects. Subjects under 20 years of age are expected to be at low risk for developing CVD and therefore may provide few cases of disease if followed for 20 years. Somewhat ironically, in 1948 many older people were considered to already have early signs of CVD. Following only elderly subjects might show few of them <u>not</u> developing CVD, making it difficult to find associations between risk factors and CVD among such subjects.

The plan was to follow them for 20 years and record outcomes. Outcome and risk factor information were recorded during the follow-up by a series of biennial examination. Every two years, participants are asked to return to a testing center, where they are examined. Their risk factors are updated and their outcomes recorded. This method of follow-up provides the opportunity to collect detailed information on study subjects, but comes at an expensive price tag.

The Framingham Heart Study was initially planned to last for 20 years, but it is still going on today. Survivors from that original cohort of 5,209 people still retrun every two years to be examined. Unfortunately, many members of the original cohort have passed on, and the number of survivors is dwindling each year. There have been several "spin-off" studies of Framingham Heart Study. A second cohort (offspring cohort) of 5124 participants was developed in 1972 using children of the original cohort and their spouses. In 2002 a third cohort of 4095 participants was developed, using the grandchildren of the original cohort and their spouses. (Oppenheimer GM. Becoming the Framingham Heart Study: 1947-1950. *Am J Public Health* 2005;95:602-610.).

Much more information about the Framingham Heart Study can be found at:

Oppenheimer GM. Becoming the Framingham Heart Study: 1947-1950. *Am J Public Health* 2005;95:602-610..

http://www.framinghamheartstudy.org/

There is also an excellent video that was created by CBS news:

http://www.cbsnews.com/8301-3445_162-3358673.html

The investigators from the Framingham Heart Study identified the roles of smoking, blood pressure and total cholesterol, as risk factors for increasing your risk of CVD. They showed that physical activity lowers your risk. They showed the relationships between the components of total cholesterol, HDL and LDL, and the risk of heart disease. They also develop prediction rules that are available at their website to predict your risk for developing various CVD outcomes. We'll be talking about the Framingham Risk Model in a later lecture.

The Framingham Heart Study is a classic example (if not the classic example of a **cohort study**. This type of study design will be discussed in more detail in future lectures.

This course will also use a teaching data set that is available from the National Heart, Lung, and Blood Institute. The data set is derived from the actual Framingham Heart Study data, but it has been perturbed to protect the identity of the original cohort. The teaching data set contains 4,434 participants from the original 5,209 participants. The data pertains three examination cycles from the original Framingham cohort. The first exam cycle refers to the 1956 exam, and two follow-up exams, roughly six years apart are also provided. In addition, all subjects were followed for 24 years and multiple outcomes were recorded.

The teaching data set contains data on the age of each person at that exam, their sex, whether they were a smoker or not, and how much they smoked (in terms of cigarettes per day), their blood pressure (both on the systolic and the diastolic scale),

whether they were taking medications to treat hypertension at that time, their total cholesterol levels (and their HDL and LDL levels), whether they had diabetes , their glucose levels, their heart rate, and whether they had previously been diagnosed with coronary heart disease, stroke, or hypertension. Outcomes include death during those 24 years, evidence of developing coronary heart disease, stroke, or hypertension, and also the number of years from the 1956 exam until they developed these outcomes.

Role of Measurement in Epidemiology

As previously stated, this definition implies two quantitative aspect of epidemiology:

1. Measuring disease distribution in regards to person, place, and time (**descriptive epidemiology**), and

2. Measuring the association between a disease and it determinants (**analytic epidemiology**).

The first quantitative aspect of this definition reflects the domain of **descriptive epidemiology** and requires the specification of appropriate **outcome measures**. Proportions, rates, percentiles and means are examples of outcome measures. The second quantitative aspect of this definition reflects the domain of **analytic epidemiology** and requires the specification of a **measure of association** to compare the values for the outcome measures in different subgroups that are defined by the exposure. Usually these measures involve taking ratios or differences of the values for the outcome measures in the different subgroups. Odds Ratios, Risk Ratios and Mean Differences are examples of commonly used measures of association. For a valid epidemiologic study, these measures of association are used as estimates of the **causal effect of a risk factor on the outcome**.

The choice of an outcome measure depends on the properties of the numerical values that are assigned to the outcome. One common categorization of measurements (exposures and outcomes) is:

Possible Types of Measurements.

1. **Nominal**
2. **Ordinal**
3. **Interval**
4. **Ratio**

Nominal measurements are categorical with no intrinsic ordering. Examples of nominal measurements include integer values assigned to categories of race, religion, and marital status. The numerical values that are assigned to categories of a nominal variable are arbitrary labels and do not reflect the relative locations of the corresponding categories on any scale.

The simplest example of a nominal variable is the **binary indicator variable** (**dummy variable**) that uses integer values to indicate membership in one of two categories of a factor of interest. For example, the following is an example of an indicator variable (Male) representing the sex of a subject.

Male = 1 if a subject is a male
= 0 if a subject is a female

Ordinal variables have categorical responses with a well-defined order among categories. The numerical values assigned to the categories of an ordinal variable reflect the relative positions for these categories but may not necessarily reflect the distance between these categories on an underlying continuous scale. For example, the New York Heart Association Functional Classification Scale assigns patients to one of four categories of cardiac disability as described by the following table (Goldman L, et al. Comparative reproducibility and validity of systems for assessing cardiovascular function class: Advantages of a new specific activity scale. *Circulation* 1981;64:1227-1234:

New York Heart Association Functional Classification Scale.

Class	Description
I	Patients with cardiac disease but without resulting limitations of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea, or anginal pain.
II	Patients with cardiac disease resulting in slight limitation of physical activity. They are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or anginal pain.
III	Patients with cardiac disease resulting in marked limitations of physical activity. They are comfortable at rest. Less than ordinary physical activity causes fatigue, palpitation, dyspnea, or anginal pain.
IV	Patients with cardiac disease resulting in inability to carry on any physical activity without discomfort. Symptoms of cardiac insufficiency or of the anginal syndrome may be present at rest. If any physical activity is undertaken, discomfort is increased.

A subject who is classified as Class II is more disabled than a subject classified as Class I. However, this classification scheme does not assume that subjects in Class II are half-way on a disability scale between Class I and Class III subjects, or are twice as disabled as subjects in Class I, or are they half as disabled as subjects in Class IV. For this reason, averaging scores from an ordinal outcome across subjects may not be appropriate.

Values for **interval variables** reflect not only the order among categories/levels of the variable, but also reflect the distances between these categories/levels on a specified (but sometimes arbitrary) scale. Therefore calculations based on addition and subtraction are appropriate. However, unless the variable possesses a natural anchoring point (zero value), mean scores must be interpreted relative to the assigned range of values.

For example, different temperature scales possess different anchoring points. On a Celsius scale, a value of 0 represents the temperature at which water freezes on this scale. The same temperature translates to 32 degrees on a Fahrenheit scale and the value of 0 on the Fahrenheit scale represents a sub-freezing temperature. Therefore a ratio of the temperature on two days will be different on these scales. Doubling the temperature on a Celsius scale does not correspond to a doubling of the corresponding temperatures on a Fahrenheit scale.

Since many health status instruments used in clinical research are transformed to a 0 - 100 point scale, an average score of 50 represents a point that is half way between the two extreme scores. If the same scale is put on a 100 - 200 point scale, then the score of 50 on the original scale becomes a value of 150 on the new scale. This value is still half way between the two extremes but no longer is equal to 50% of the upper extreme.

Values for **ratio variables** not only reflect order and distances between categories/levels, but these variables also contain a natural zero value, representing the absence of the quantity that is being measured. This provides a natural anchoring point for the interpretation of values on this scale. Examples of ratio scales include temperature measured on the Kelvin Scale (where 0 = lack of molecular movement and the absence of temperature) and the age of individuals. Not only is a 50 year-old half way between the age of a 40 year-old and a 60 year-old, but also he/she is also twice as old as a 25-year old.

A **continuous variable** may be interval or ratio in scale, depending on the existence of a natural zero value.

The values for the categories of an ordinal variable are often represented by consecutive integers. These values reflect the order of the categories but may not reflect perceived distances between these response options. For example, a commonly used simple measure of health status is the following question:

How would you describe your health status?

Excellent	5
Very Good	4
Good	3
Fair	2
Poor	1

The consecutive integer values may not necessarily reflect distances between these health state categories, since the distance between fair and good health states may be perceived by many to be greater than the distance between very good and excellent health states. For example following values reflect the median locations of these 5 health states as reported by a group of students at HSPH taking EPI241 when asked to place the categories on a line with poor health assigned the value 1 (left end of the line) and excellent health assigned a value of 5 (right end of the line):

Median location of health states by HSPH students on 5-point scale

5.	Excellent	→	5
4.	Very Good	→	4.4
3.	Good	→	3.3
2.	Fair	→	2
1.	Poor	→	1

The students reported that very good health should be recorded with a value of 4.4 rather then 4.0, reflecting their view that very good health was closer to excellent health than to good health. Good health was recorded with a value of 3.3, reflecting their views that this state of health was closer to very good health than to fair health. Although the numbers on the right or the left both work fine for reflecting the order among the health states, only the numbers on the right (1, 2, 3.3, 4.4, 5) reflect both order and distance and could be used to calculate the average health status for a group of people.

Averages are often used to summarize continuous outcome in epidemiology (e.g. average blood pressure in a group of people). However, one should keep in mind the following quotation of a former baseball manger (Booby Bragen: http://sportsillustrated.cnn.com/vault/article/magazine/MAG1074702/index.htm)

> *"Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable*

If I have a group of people and half of them report having excellent health (5) and the other half report having poor health (1), then is it informative to report that the average health status of that group is 3? This value suggests slightly less than good health, when it underestimates the health status for half of the group and over estimates it for the other half?

Outcome Measures for Binary Variables

Since the values assigned to the categories of nominal variables are arbitrary labels, an outcome measure for a nominal outcome should be independent of the values that are assigned to the categories of the variable. Mean and percentile values are not appropriate outcome measures for nominal variables. However, percentages (proportions)

of subjects in categories provide a valid description of the distribution of a nominal variable.

Proportions relate the size of the population that have (or develop) the disease to the total size of the population of interest (**part/total measure**). **Odds** are an alternative measure of the likelihood of belonging to a particular category of a nominal variable. For example, the odds of disease measures the size of the population who have (or develop) the disease to the size of the population that does not have (or does not develop) the disease (a **part/non-part measure**). The value for an odds is easily calculated from the value for a proportion by the following transformation

odds = (proportion)/(1 - proportion)

Alternatively, the value for a proportion can be calculated from the value for an odds by the following formula:

proportion = (odds)/(1 + odds)

Values for proportions range from 0.0 to 1.0. Values for odds range from 0.0 to positive infinity. If the value for the proportion is small, then the corresponding odds will also be small. However, as the magnitude of the proportion increases, so does the difference between values for the odds and the proportion as shown in the following table

Relationship between Proportions and Odds.

Prevalence	Prevalence Odds
0.01	0.01
0.02	0.02
0.03	0.03
0.05	0.05
0.10	0.11
0.20	0.25
0.30	0.43
0.40	0.67
0.50	1.00
0.60	1.50
0.70	2.33
0.80	4.00
0.90	9.00
0.95	19.00
0.97	32.33
0.98	49.00
0.99	99.00

The previous table demonstrates that values for two proportions that are very close to one another (e.g. (0.01 and 0.02), or (0.98 and 0.99)) may have values for the corresponding odds that are also close to one another (0.01 and 0.02 for the first pair) or far apart (49.00 and 99.00 for the second pair). This implies that a measure of association that suggests large differences between values for two odds may not necessarily imply large differences between values for the corresponding proportions. Although a proportion may be a more intuitive outcome measure for a nominal variable, a particular analysis may require that the odds be the outcome measure of choice (e.g. logistic regression).

Prevalence

Prevalence measures how much outcome (disease) exists in the population **at a point in time**. Time can be measured in different dimensions. For example, one could think of a **chronologic date** and report the prevalence of AIDS in the United States today. Alternatively, time could refer to a person's **age** and you could consider the prevalence of low back pain among men at their 65^{th} birthday. Finally, time could be recorded in terms of **life time events** and consider the prevalence of cataracts at the time of retirement among men and women. In all cases, prevalence refers to a **snapshot** of a population at a point in time and quantifies the amount of a disease (or some other characteristic) that exists in the population at that time. It's involves examining people, perhaps taking a survey about their health, and reporting how much disease exist in that population at that one point in time.

Prevalence usually is defined as the proportion of people who have the disease at that point in time.

Prevalence = (# with disease)/(# examined)

If I had a camera and could take a picture of everyone taking this course, then I could see how many of you were wearing eyeglasses. I could report the prevalence of eye problems among people taking this course. If there were 100 people watching me right now, and 30 were wearing eyeglasses, then the prevalence of wearing eyeglasses (prevalence of having eye conditions) would be

30/100 = 30%

Alternatively, I could turn that prevalence into an odds and report the prevalence odds

30/70 = 43%

The following table is from the Framingham Heart Study teaching data set. It shows the prevalence of Coronary Heart Disease (CHD) at the 1956 exam, for groups defined by age and sex.

Sex	Age Group	Number at Exam	Number with CHD	Prevalence
Female	30-40	415	0	0/415= .00
	41-50	908	6	6/908= .01
	51-60	795	38	38/795=.05
	> 60	372	26	26/372=.07
Male	30-40	339	6	6/339=.02
	41-50	731	24	24/731=.03
	51-60	584	37	37/584=.06
	> 60	290	57	57/290=.20

The last column of this table reports the prevalence of CHD for the various groups. Not surprisingly, the prevalence of CHD increases with age for both men and women. In addition, for any age, the prevalence of CHD is higher for men than for women.

These conclusions reflect the curse of being an epidemiologist. We look at data and we report associations. For example, we observe that prevalence of CHD increases with age and is great for men than women. In addition to reporting associations, epidemiologists try to identify the reason for an observed association. For example, what are potential reasons for the higher prevalence of CHD existing for men, compared to women?

One possible explanation for the higher prevalence of CHD among men is that men might be at higher risk for developing coronary heart disease. Therefore a group of men would show a higher **incidence** for developing CHD than a comparable group of women. Perhaps, the reason we see a higher prevalence of CHD among men at the 1956 examination is that more new cases of CHD occurred among men in the previous weeks, months and years prior to the 1956 examination than among women. The topic of incidence will be discussed in the next sequence of lectures.

Another reason for the higher prevalence of CHD among men might be that men live longer with their heart disease. Suppose that the incidence of heart attacks is the same on both men and women, but men survive their heart attacks. When I measure the prevalence of CHD in a population I'll find men who developed heart disease a month ago, a year ago, two years ago, and are still alive. On the other hand I will not find the women who developed CHD in the past because they may not have survived. So another reason you might see higher prevalence of disease in one population than another has nothing to do with the incidence of developing disease but with the duration of disease.

Unfortunately, there are even more possible explanations, beyond just having higher incidence or having higher duration, which might explain the higher prevalence of CHD among men. Recall that we epidemiologists measure associations. We hope those associations represent the true, causal effect of a risk factor on an outcome. However,

before we can conclude that an association reflects a causal effect of a risk factor we have to exclude three possible alternative explanations: **bias, confounding, or chance.**

For example, it may be that men and women really do have the same prevalence of disease, but maybe men see their physicians more often and are tested more often for CHD. If so, the reason for the higher prevalence of CHD among men might be due to a measurement bias: the underreporting of CHD among women.

A fourth possible explanation for the higher prevalence of CHD is that men and women have the same risk of developing disease based on their sex, but men have more additional risk factors (e.g. smoking hypertension, …) that increases their risk of CHD and leads to a higher incidence of CHD (and their higher prevalence). The higher prevalence of CHD among men is not a reflection of the effect of sex on the incidence of CBD, but the effect of confounding factors (e.g. smoking, hypertension, …) that exist more commonly among men. The topic of confounding that we'll be talking about in future sessions.

Finally, suppose that in general, men and women have the same prevalence of CHD. However, your data contain only a sample of men and a sample of women. Because of sampling variability (chance), the observed prevalence of CHD among men in our data may be higher than what exists in general among all men. Suppose that the opposite is true for women: the observed prevalence of CHD among women in our data may be lower than what exists in general among all women. Therefore, the association that we see ion our data is an artifact of sampling variability.

In summary, there are many plausible explanations that could explain why the prevalence of CHD is higher among men than among women. The challenge for epidemiologists is to try to rule out the unlikely explanations, to conclude with what is the most likely explanation. In this manner, *epidemiologists* are often thought as detectives as suggested by the following quotation:

"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth? We know that he did not come through the door, the window, or the chimney. We also know that he could not have been concealed in the room, as there is no concealment possible. When, then, did he come?"

A. C. Doyle: *The Sign of the Four (1890)*

Unfortunately, there is still another possible explanation (reverse causation) for observing and association that occurs if some studies (but not this example) when using prevalence data. This will be discussed in a future lecture. However, for the present, the main take-home message for interpreting prevalence data is that prevalence is influenced by both **incidence**, the development of new cases of disease, and the **duration** of disease. This issue is demonstrated by the following example.

The following quotation is from the headline of a CNN report (CNN.Com Monday, June 13, 2005 Posted: 1:42 PM EDT (1742 GMT*):*

1 million living with HIV in U.S."

"Statistics provide good and bad news"

Why is it good news that one million people were living with HIV in the United States in 2005? The answer is found later in the article.

"for the first time since the height of the AIDS epidemic in the 1980s, more than a million Americans are believed to be living with the virus that causes AIDS, the government said on Monday."

"The latest estimate is both good news and bad news-- reflecting the success of the drugs that keep more people alive." The duration was increasing.

The good news implied by this quotation is that treatment for patients with HIV appeared to be working, in that is was extending the lifetime of patients with this condition. The bad news was the incidence of HIV was still increasing. There were two reasons why the prevalence was high: **higher incidence** of developing HIV (bad news), and **longer duration** (survival) with disease (good news).