5. Measuring Disease and Exposure

Descriptive statistics; measuring occurrence and extent of disease; prevalence, incidence (as a proportion and as a rate), and survivorship; weighted averages, exponents, and logarithms.

> "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of Science, whatever the matter may be."
>
> Lord Kelvin (quoted in Kenneth Rothman, *Modern Perspectives in Epidemiology*, 1 ed. Boston, Little Brown, 1986, pg 23)

At the beginning of this text we noted four key aspects of epidemiology: its multidisciplinary nature, and its concern with populations, measurement, and comparison. As all empirical scientistis, epidemiologists devote a great deal of attention to issues of measurement – the application of numbers to phenomena. Every object of study – a disease, an exposure, an event, a condition – must be defined and measured. Since epidemiology deals with populations, epidemiologists need methods to describe and summarize across populations. This chapter discusses various aspects of measurement, including the definition, computation, and interpretation of key measures of health events and states in populations. The next chapter deals with comparisons between these measures.

Numeracy: applying numbers to phenomena

Numeracy is the concept of summarizing phenomena quantitatively. Faced with an infinitely detailed and complex reality, the researcher attempts to identify and quantify the meaningful aspects. Two of the innumerable examples of this process in epidemiology are:

Atherosclerosis score: David Freedman, an epidemiologist who received his doctoral degree from UNC, conducted his dissertation research on the relationship of atherosclerosis in patients undergoing coronary angiography to plasma levels of homocysteine. A basic question he had to address was how to measure atherosclerosis in coronary angiograms. Should he classify patients as having a clinically significant obstruction, count the number of obstructions, or attempt to score the extent of atherosclerosis? An atherosclerosis score would capture the most information and could provide a better representation of the phenomenon as it might be affected by homocysteine levels. But should an atherosclerosis score measure surface area of involvement or extent of narrowing? How should it treat lesions distal to an occlusion, which have no effect on blood flow? These and other decisions would need to depend upon his conceptual model of how homocysteine would affect the endothelium. For example, would homocysteine be involved primarily in causing initial damage, in which case the total surface area involved would be relevant, or would it be involved in the progression of atherosclerosis, in which case the extent of narrowing would

be relevant. Compromises might be forced by limitations in what measurements could be made from the angiograms.

Measuring smoking cessation: at first glance, smoking cessation, in a study of the effects of smoking cessation or of the effectiveness of a smoking cessation program, would seem to be straightforward to define and measure. Even here, though, various questions arise. The health benefits from cessation may require abstinence for an extended period (e.g., years). However, biochemical validation techniques, considered necessary when participants would have a reason to exaggerate their quitting success, can detect smoking during a limited period of time (e.g., about seven days for salivary cotinine). Should cessation be defined as no tobacco use for 7 days, to facilitate validation, or for at least a year, when the relapse rate is much lower?

Conceptual models underlie measures

In general, how we apply numbers and what type of measures we construct depend upon:

1. the purpose of the measure
2. the nature of the data available to us.
3. our conceptualization of the phenomenon

These three factors will pervade the types of measures to be covered.

Ideally we would like to watch phenomena unfold over time. In practice we must often take a few measurements and infer the rest of the process. Conceptual models pervade both the process of applying numbers to phenomena and the process of statistically analyzing the resulting data in order to identify patterns and relationships. Not being able to record all aspects of phenomena of interest, we must identify those aspects that are biologically, psychologically, or otherwise epidemiologically important. These aspects are embodied in operational definitions and classifications. The method by which we apply numbers and analyze them must preserve the important features while not overburdening us with superfluous information. This basic concept holds for data on individuals (the usual unit of observation in epidemiology) and on populations. Although we employ mathematical and statistical models as frameworks for organizing the resulting numbers, for estimating key measures and parameters, and for examining relationships, conceptual models guide all of these actions.

Levels of measurement

One area where objectives, availability of data, and conceptual models come to bear is the **level of measurement** for a specific phenomenon or construct. Consider the construct of educational attainment, a variable that is ubiquitous in epidemiologic research. We can (1) classify people as being or not being high school graduates; (2) classify them into multiple categories (less than high school, high school graduate, GED, trade school, technical school, college, professional degree, graduate degree); (3) record the highest grade in school they have completed; or (4) record their scores on standardized tests, which we may need to administer.

The first alternative listed illustrates the most basic "measurement" we can make: a **dichotomous** (two category) classification. People can be classified as "cases" or "noncases", "exposed" or "unexposed", male or female, etc. Communities can be classified as having a mandatory seat-belt law or not, as having a needle exchange program or not, etc.

Potentially more informative is a **polytomous** (more than two categories) classification, such as country of origin, religious preference, ABO blood group, or tumor histology (e.g., squamous cell, oat cell, adenocarcinoma). A polytomous classification can be **nominal** – naming categories but not rank ordering them, as is the case for the four examples just given – or **ordinal**, where the values or categories can be rank-ordered along some dimension. For example, we might classify patients as "non-cases", "possible cases" "definite cases" or injuries as minimal, moderate, severe, and fatal.

The values of the different levels of a nominal variable provide no information beyond identifying that level, and so they can be interchanged without constraint. We can code squamous cell "1", oat cell "2", and adenocarcinoma "3"; or instead, squamous cell "2" and oat cell "1" or even "5"). The numbers simply serve as names. The values of the different levels of an ordinal variable signify the ranking of the levels. The values can be changed, but generally not interchanged. We can use "1", "2", "3", respectively, for non-case, possible case, and definite case, or we can use "1" "3" "8", but we can not use "1" "3" "2", since this coding would not preserve the ordering.

When the values themselves, or at least the size of the intervals between them, convey information, then the phenomenon has been measured at the **interval** level. Temperature measured on the Fahrenheit scale is an interval scale, since although we can say that 80°F is twice 40°F, the ratio is not meaningful in terms of the underlying phenomenon. Psychological scales are often regarded as being interval scales. What differentiates an interval scale from most of the measures we use in physical sciences is the absence of a fixed **zero point**. Since only the intervals convey meaning, the scale can be shifted up or down without changing its meaning. An interval scale with values "1", "1.5", "2", "3", "4" could just as well be coded "24", "24.5", "25", "26", "27".

A **ratio** scale, however, has a non-arbitrary zero point, so that both intervals and ratios have meaning. Most physical measurements (height, blood pressure) are ratio scales. The values of an ratio scale can be multiplied or divided by a constant, as in a change of units, since comparisons of intervals and ratios are not distorted. If value B is twice value A before multiplication, it will still be twice value A afterwards. A ratio scale with values "1", "1.5", "2", "3", "4" can be transformed to "2", "3", "4", "6", "8" (with appropriate substitution of units), but not as "2", "2.5", "3", "4", "5", since only intervals but not ratios will be preserved.

One type of ratio scale is a **count**, such as birth order or parity. A count is a **discrete** variable, because its possible values can be enumerated. A **continuous** variable, as defined in mathematics, can take on any value within the possible range, and an infinitude of values between any two values. Measurements in epidemiology are no where nearly as precise as in the physical sciences, but many measurements used in epidemiology have a large enough number of possible values to be treated as if they were continuous (e.g., height, weight, or blood pressure).

Whether continuous or discrete, however, both interval and ratio scales generally imply a linear relationship between the numerical values and the construct being measured. Thus, if we measure

educational attainment by the number of years of school completed, we are implying that the increase from 10th grade to 11th grade is the same as the increase from 11th grade to 12th grade, even though the latter usually conveys a high school diploma. We are also implying that completing 12th grade with three advance-placement or honors classes in a high-achievement school is the same as completing 12th grade with remedial courses in a low-achievement school, or as completing 12th grade but reading at only a 9th grade level, or completing 12th grade but without taking any mathematics beyond elementary algebra, etc., not to mention ignoring the educational aspects of travel, speaking multiple languages, or having learned a trade. Even chronological age may not be an interval or ratio scale when certain ages have special meaning (e.g., 16 years, 18 years, 21 years, 40 years, 65 years). Many measures that appear to be interval or ratio scales may not really behave as such, due to **threshold effects** (differences among low values have no real significance), **saturation effects** (differences among high values have no real significance), and other **nonlinearities**.

Absolute and relative measures — the importance of a denominator

While the absolute values of age, educational attainment, blood pressure, and cigarettes/day are meaningful, other measures are expressed as concentrations (e.g., 20 μg of lead per deciliter of blood, 500 T-cells per cubic centimeter of blood, 1.3 persons per room, 392 persons/square kilometer) or **relative** to some other dimension (e.g., body mass index [weight/height^2], percent of calories from fat, ratio of total cholesterol to HDL cholesterol). Most population-level measures are not meaningful unless they are relative to the size and characteristics of a population and/or to expected values, even if only implicitly. Other than a report of cases of small pox, since the disease has now been eradicated world wide, how else can we assess whether a number of cases represents an outbreak or even an epidemic? For this reason epidemiologists often refer disparagingly to absolute numbers of cases or deaths as "numerator data". Exceptions illustrate the general principle. A handful of cases of angiosarcoma of the liver in one manufacturing plant led to an investigation that uncovered this hazard from vinyl chloride. A handful of cases of adenocarcinoma of the vagina in teenage women in one hospital led to the identification of the effect of diethylstilbesterol (DES) on this disease. A handful of cases of acquired immunodeficiency syndrome (AIDS) alerted public health to the start of this pandemic. Since these were very rare or previously unobserved conditions, an expectation was already defined.

Types of ratios

As illustrated with several of the above examples, we express a quantity relative to another by forming a ratio, which is simply the quotient of two numbers, a numerator divided by a denominator. Ratios are ubiquitous in epidemiology, since they enable the number of cases to be expressed relative to their source population.

Two special classes of ratios in epidemiology are proportions and rates. **Proportions** are ratios in which the numerator is "contained in" or "part of" the denominator. The statement that 12% of the population is age 65 or above expresses a proportion, since people age 65 and above are a fractional component of the population. Because the numerator is a fractional component of the denominator, a proportion can range only between 0 and 1, inclusive. Proportions are often expressed as percentages, but any **scaling factor** can be used to yield a number that is easier to express. For example, the proportion 0.00055 would often be expressed as 5.5 per 10,000 or 55 per

100,000. Note that the ratio of abortions to live births, although of the same order of magnitude, is *not* a proportion, since the numerator is not contained in the denominator.

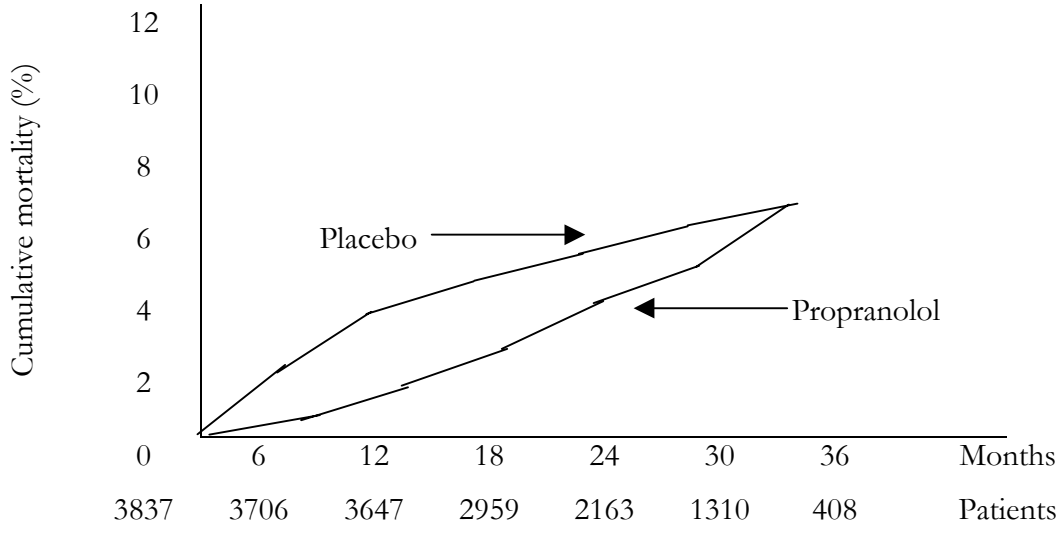
Although many types of ratios (including proportions) are frequently referred to as "rates", in its precise usage a **rate** is the ratio of a change in one quantity to a change in another quantity, with the denominator quantity often being time (Elandt-Johnson, 1975). A classic example of a rate is velocity, which is a change in location divided by a change in time. Birth rates, death rates, and disease rates are examples if we consider events — births, deaths, newly diagnosed cases — as representing a "change" in a "quantity". Rates can be absolute or relative, according to whether the numerator is itself a ratio that expresses the change relative to some denominator. Most rates in epidemiology are relative rates, since as discussed above the number of cases or events must generally be related to the size of the source population.

"Capturing the phenomenon"

All measures, of course, are summaries or indicators of a complex reality. The question always is, "does the measure capture what is important about the phenomenon given our objective?". This principle applies at both the individual level (for example, when can a person's constantly-varying blood pressure and heart rate be meaningfully represented by single numbers) and population level.

For example, although the proportion of a group of patients who survive for 5 years is a measure of treatment effectiveness, if the proportion is low then when deaths occur is especially important. The statement that the "five-year survival rate following coronary bypass surgery was 60%" does not tell us whether the 40% who died did so during the procede, soon afterward, gradually during the period, or not until at least three years following surgery. When the **time-to-occurrence** of an event is important, then survivorship analysis is employed, such as in the following figure similar to that reported from the Beta-blocker Heart Attack Trial (BHAT), a double-blinded, randomized trial of propranolol to treat patients experiencing an acute myocardial infarctions.

Life table cumulative mortality in the Beta Blocker Heart Attack Trial (BHAT)

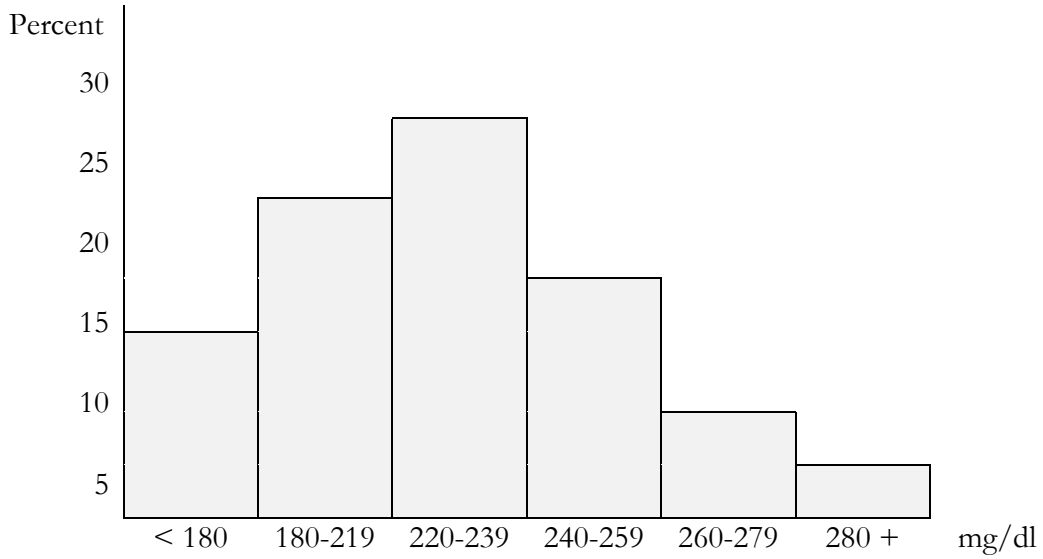


[Source: *JAMA*, March 26, 1982; 247:1707]

Distributions – the fuller picture

More generally, when the object of study involves not merely "presence" or "occurrence" but rather a **polytomous** or **measurement** variable, we should examine the full distribution, e.g.

Serum cholesterol levels - Distribution



Although distributions are informative, they are cumbersome to work with and to present. Therefore we try to "capture" the essential information about the distribution by using summary statistics, such as the mean, median, or quartiles, and the standard deviation or interquartile range (see below). While it is often essential to compress a distribution, curve, or more complex picture into a number or two, care must be taken that the necessary simplification does not distort the resulting computation, presentation, and interpretation. Indeed, it may be the persons at one end of the distribution who are most important or informative in respect to health consequences.

If the data are distributed in a familiar fashion, we can adequately characterize the entire distribution by its parameters (e.g., the mean and standard deviation for a "normal" [Gaussian] distribution). But it can be hazardous to assume that the data conform to any particular distribution without verifying that assumption by examining a histogram (e.g., see *Statistics for Clinicians*, Figure 7-7, for several distributions with identical mean and standard deviation but dramatically different appearance).

Common summary statistics for description and comparison

Mean – The "average" value of the variable

Median – The middle of the distribution of the variable – half of the values lie below and half lie above

Quartiles – The values that demarcate the 1st, 2nd, and 3rd quarter of the distribution of the variable [the median is the 2nd quartile]

Percentiles – The values that demarcate a percentage of the distribution, e.g., the 20th percentile (also called the second decile) is the value below which the lowest 20% of the observations fall.

Standard deviation – Roughly speaking, the distance of a typical observation from the mean of the distribution (more precisely, the square root of the average of the squared distances of observations from the mean) [Not to be confused with the **standard error**, which is a measure of the imprecision of an estimate.]

Interquartile range – The distance between the 1st and 3rd quartiles.

Skewedness – The degree of asymmetry about the mean value of a distribution. Positively skewed or right-skewed means that the distribution extends to the right; in a positively-skewed distribution, the mean (overall average) lies to the right of the median, due to the influence of the outlying values.

Kurtosis – The degree of peakedness of the distribution relative to the length and size of its tails. A highly peaked distribution is "leptokurtic"; a flat one is "platykurtic".

When interpreting summary statistics, it is important to consider whether the summary statistics represent the most relevant features of the distributions that underlie them. Several examples:

Community health promotion:

Suppose that surveys before and after a community alcohol control program find a reduction in mean alcohol consumption of 1 drink/day in the target population. That reduction could reflect either:

- a 5 drink/day reduction for each person in the highest consumption 20 percent of the population

or

- a 1.25 drink/day reduction for all people but those in the highest consumption 20%,

with very different implications for health.

Black-white differences in birth weight:

The distribution of birth weight has an approximate Gaussian ("normal") shape, with a range from about 500 grams (the lower limit of viability) to about 5,000 grams and a mean of about 3,000 grams. Statistically the distribution is smooth and reasonably symmetrical. However, the biological implications vary greatly across the distribution, since the majority of infant deaths occur for babies weighing less than 2,500 grams. For babies weighing 1,000-2,000 grams, the mortality rate is 33%; for babies weighing less than 1,000 grams, the mortality rate is 75%.

The birth weight distributions for Black and White Americans are generally similar, with that for Blacks shifted slightly to the left. But that slight shift to the left translates into a substantially greater proportion below 2,500g, where mortality rates are much higher.

Per capita income:

> Should health care resources for poor people be allocated on the basis of per capita income of counties? At least one study has found that the barriers to health care experienced by the poor in the U.S. appear to be similar in wealthy counties and in other counties, so that per capita income (i.e., mean income per person) is not as good a criterion for determining the need for public health care programs as is the number of poor persons in the area (Berk M, Cunningham P, Beauregard K. The health care of poor persons living in wealthy areas. *Social Science in Medicine* 1991;32(10):1097-1103).

The moral: in order to interpret a change or difference in a summary measure it is necessary to know something about the shape of the distribution and the relationship between the variable and the relevant health outcome.

Heterogeneity and distributions of unknown factors – any summary is a weighted average

Since populations differ in characteristics which affect health, an overall number, such as a proportion or mean, often conceals subgroups that differ meaningfully from the overall picture. Even when we cannot identify these subgroups, we should be mindful of their likely existence. Because most diseases vary across subgroups, epidemiologic measures are more interpretable with knowledge of the composition of the group they refer to, at least in terms of basic demographic characteristics (notably age, sex, geographical area, socioeconomic status, employment status, marital status, ethnicity) and important exposures (e.g., smoking).

> E.g., a workforce experiences 90 lung cancer deaths per l00,000 per year: To know what to make of this it is essential to know the age distribution of the workforce and if possible the distribution of smoking rates.

Virtually any measure in epidemiology can be thought of as a weighted average of the measures for component subgroups. We can use "specific" measures (e.g., "age-specific rates," "age-sex-specific rates") where the overall ("crude") measure is not sufficiently informative. Also, we can produce "adjusted" or "standardized" measures in which some standard weighting is used to facilitate comparisons across groups. Adjusted measures are typically weighted averages – the weights are key. The concept of weighted averages is fundamental and will resurface for various topics in epidemiology. (Rusty on weighted averages? See the Appendix on weighted averages.)

Types of epidemiologic measures

Purpose of the measure:

There are three major classes of epidemiologic measures according to the question or purpose. We use **measures of frequency or extent** to address questions such as "How much?", "How many?", "How often?", "How likely?", or "How risky?". We use **measures of association** to address questions about the strength of the relationship among different factors. We use **measures of impact** to address questions of "How important?".

Availability of data:

We can also categorize epidemiologic measures according to the type of data necessary to obtain them:

1. Measures derived from routine data collection systems, e.g., vital events registration, cancer registries, reporting of communicable diseases.

2. Measures derived from data collected in epidemiologic studies or for related purposes (e.g., clinical studies, health insurance records).

3. Measures derived from theoretical work in biometry - no data necessary! e.g., Risk of disease in exposed = Pr[D | E]

$$\text{Incidence density} = -\frac{d(N_t)}{N_t\, dt}$$

The usefulness of the third class of measures is in refining measurement concepts and in advancing understanding. Measures in the first two classes generally involve compromises between the theoretical ideal and practical reality. Epidemiology is fundamentally a practical field. In the rest of the chapter we will touch on the first class and then dwell on the second.

Measures derived from routinely collected data

In this area come the vital statistics data compiled by health authorities and statistical agencies, such as the World Health Organization, the U.S. National Center for Health Statistics, state health departments, and their counterparts in other countries. Examples of measures published from such data are:

- total death rates
- cause-specific death rates
- birth rates (births per 1,000 population)
- infant mortality rates
- abortion/live birth ratio
- maternal mortality rate

[See Mausner and Kramer, ch 5; Remington and Schork, ch 13.]

The denominator for vital statistics and other population-based rates (e.g., death rates, birth rates, marriage rates) is generally taken from population estimates from the national census or from other vital events data, as in the case of the infant mortality rate:

$$\text{Infant mortality rate} = \frac{\text{Deaths of children < 1 year of age in one year}}{\text{Total live births in one year}}$$

Results are usually scaled so that they can be expressed without decimals (e.g., 40 deaths per 1,000 or 4,000 deaths per 100,000).

Optional aside – Assessing precision of an estimated rate, difference in rates, or ratio of vital statistics rates

If r is a rate (e.g., an infant mortality rate) and n is the denominator for that rate (e.g., number of live births), then a 95% confidence interval for r can be constructed using the formula:

$$r \pm 1.96 \times \sqrt{(r/n)}$$

E.g., in an area with 30 infant deaths and 1,000 live births, r = 30/1,000 = 30 per 1,000 or 0.03. The 95% confidence interval for r is:

$$0.03 \pm 1.96 \times \sqrt{(0.03/1{,}000)} = 0.03 \pm 0.0107 = (0.0193, 0.0407),$$

or between 19.3 and 40.7 per thousand

The 95% confidence interval for the difference, D, between two rates, r_1 and r_2, based, respectively, on number of deaths d_1 and d_2, and denominators n_1 and n_2, is:

$$(r_1 - r_2) \pm 1.96 \times \sqrt{(r_1/n_1 + r_2/n_2)}$$

The 95% confidence interval for the ratio, R, of r_1 and r_2 is:

$$R \pm R \times 1.96 \times \sqrt{(1/d_1 + 1/d_2)}$$

where d_2 (the number of deaths for the denominator rate) is at least 100.

Source: Joel C. Kleinman. Infant mortality. Centers for Disease Control. National Center for Health Statistics *Statistical Notes*, Winter 1991;1(2):1-11.

The basis for the above can be stated as follows. The number of rare events in a large population can often be described by the Poisson distribution, which has the notable feature that its mean is the same as its variance. For a Poisson distribution with mean d (and variance d), if the number of events is sufficiently large (e.g., 30), then 95% of the distribution will lie within the interval $d \pm 1.96\sqrt{d}$. If we divide this expression by the population size (n), we obtain the 95% confidence interval for the rate as:

$$d/n \pm (\sqrt{d})/n = r \pm \sqrt{(r/n)}$$

Reporting systems and registries for specific diseases, hospital admissions, and ambulatory care visits provide data on incidence or health care utilization for some conditions. Communicable diseases have long been reportable, though the completeness of reporting is quite variable. Major investments in state cancer registries are creating the basis for a national cancer registry system in the

U.S. Several states have reporting systems for automobile collisions. For the most part, however, data on non-fatal disease events are less available and complete than mortality data.

Remember: All rates, ratios, and other measures can be:

Specific to a group defined by age, sex, and/or other factors.

Adjusted for age, sex, or other relevant variable(s);

Crude (i.e., neither specific nor adjusted).

These terms apply with respect to particular variable(s) and are therefore not mutually exclusive. For example, a rate can be adjusted with respect to age, specific with respect to gender, and crude with respect to ethnicity, geographical region, etc. (e.g., an age-adjusted rate for women of all ethnicities and all geographical regions).

The basic concept underlying adjustment procedures is that of the **weighted average**. The limitations of adjusted measures derive from this aspect – validity of comparison depends upon the similarity of the component weights; validity of interpretation depends upon the numerical and conceptual homogeneity of the component specific measures.

Measures derived from data collected in epidemiologic studies

For most epidemiologic studies, routinely collected data are not adequate, so data must be collected specifically for the study purposes. The reward for the time, effort, and expense is a greater opportunity to estimate measures that are more suited for etiologic and other inferences. Three principal such measures are prevalence, incidence, and case fatality.

Prevalence – the proportion of cases within a population
$\text{Prevalence} = \dfrac{\text{Cases}}{\text{Population-at-risk}}$

Prevalence – a kind of "still life" picture – is the most basic of epidemiologic measures. It is defined as the number of cases divided by the population-at-risk. Note that:

- Prevalence is a proportion, so must lie between 0 and 1, inclusive.
- Population at risk (PAR) means "eligible to have the condition".
- Prevalence can be used to estimate the probability that a person selected at random from the PAR has the disease [Pr(D)]

Example:

$$\text{Prevalence} = \frac{\text{No. of persons with senile dementia at a given time}}{\text{No. in study population at risk for senile dementia}}$$

$$= \frac{175}{1{,}750} = 0.10 = 10\%$$

Optional aside – Assessing precision of an estimated prevalence.

Since prevalence is a proportion, a confidence interval can be obtained using the binomial distribution or, where there are at least 5 cases, the normal approximation to the binomial distribution. The variance of a point binomial random variable is pq (where p is the probability of a "success" and q=1—p), so the standard error for the estimated probability is $\sqrt{(pq/n)}$. Thus, the 95% confidence interval for a prevalence estimate p is: $p \pm 1.96 \sqrt{[p(1—p)/n]}$. For the preceding example, the 95% confidence limits are $0.10 \pm 1.96\sqrt{[(0.10)(0.90)/1750]} = (0.086, 0.114)$. When there are fewer than 5 cases, an exact procedure is required.

Prevalence has three components:

1. Existing cases
2. Population "at risk" to have the condition
3. Point (or sometimes a period) in time to which the prevalence applies

Incidence – the occurrence of new cases

$$\text{Incidence} = \frac{\textit{New} \text{ cases}}{\text{Population-at-risk } \textit{over time}}$$

Incidence – a "motion picture" – describes what is happening in a population. Incidence is defined as the number of new cases divided by the population at risk over time. Incidence therefore includes three components:

1. New cases
2. Population at risk.
3. Interval of time.

Note that:

- Incidence involves the passage of time.

- Incidence may be expressed as a proportion or as a rate.
- Incidence can be used to estimate the risk of an event during a stated period of time.

Example:

$$\text{e.g., Cumulative incidence} = \frac{\text{New cases of senile dementia in 5 years}}{\text{No. of persons at risk}}$$

In infectious disease epidemiology, this measure is often termed the **attack rate** or **secondary attack rate**, especially when referring to the proportion of new cases among contacts of a primary case.

Case fatality is a measure of the severity of a disease. Though often called the case fatality "rate", the measure is generally computed as a proportion:

Case fatality – proportion of cases who die
$\text{5-year case fatality} = \frac{\text{Deaths from a condition}}{\text{Number of persons with the condition}}$

If the time period under discussion does not encompass the entire period of risk of death from the condition, then the time period must be stated explicitly or the statistic is uninterpretable. The case fatality rate for AIDS increases with every year following diagnosis, but that for an episode of influenza or for a surgical procedure does not change after a month or so.

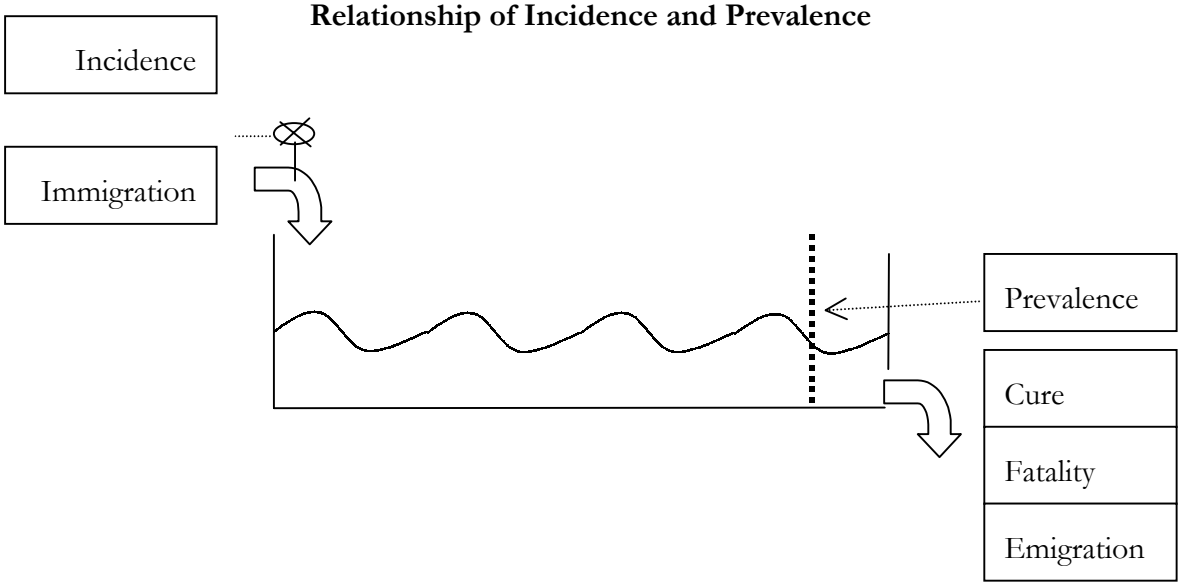
Example:

$$\text{Case fatality rate} = \frac{\text{Deaths from senile dementia in 5 years}}{\text{Number of persons diagnosed with senile dementia}}$$

Relationship of incidence and prevalence

Incidence, mortality, and prevalence are intimately related, of course, just as are births, deaths and population size. Demographers study the latter phenomena, and their techniques are used in epidemiology (under other names, naturally, to "protect the innocent"!).

Relationship of Incidence and Prevalence



In a stationary population, in which there is no migration of cases or noncases, if the incidence, prevalence, and duration of a condition remain constant then the number of new cases that occur must be balanced by the number of existing cases that leave the population through death or cure. In such a situation, the prevalence is a function of incidence and the average duration of being a case. For a rare disease, prevalence ≈ incidence × duration (see "Incidence and prevalence in a population", below).

Influences on the relation of incidence and prevalence

The relationships among incidence, mortality, and prevalence are affected by such factors as:

Virulence of the disease - Is it rapidly fatal?

Health care - When do cases come to medical attention?

Can cases be cured?

Does earlier detection alter prognosis?

Behavior - Do people recognize and act promptly on symptoms?

Do patients comply with treatment?

Competing causes of death - Are people with the disease likely to die of other causes?

Migration - Are people with the disease likely to leave the area?

Are people with the disease like to migrate to the area?

Because prevalence is affected by factors (e.g., duration and migration) that do not affect the development or detection of a disease or condition, measures of incidence are generally preferred over measures of prevalence for studying etiology and/or prevention. Both incidence and prevalence are useful for various other purposes (surveillance and disease control, health care

planning). Also, prevalence may be more readily estimated than incidence and may be looked to for etiologic inferences despite its limitations.

It is important to note, however, that although incidence itself is not affected by factors unrelated to etiology, observed incidence reflects the influence of a variety of nonetiologic factors (how quickly the disease produces symptoms that prompt a health care visit, access to health care, whether the health care provider selects the correct diagnostic maneuver, accuracy of the exam result and its interpretation, and accuracy and promptness of reporting). There are, accordingly, great difficulties in interpreting reported incidence of many diseases and conditions (e.g., Alzheimer's disease, AIDS, HIV, other sexually transmitted infections, Lyme disease, and prostate cancer, to name but a few).

An example of how disease natural history distorted trends in observed incidence comes from the early years of the AIDS epidemic, when AIDS case reporting was the primary means of tracking the HIV epidemic. Due to the considerable variability in the time between HIV infection and development of opportunistic infections signaling the onset of AIDS, the upward trend in AIDS cases exaggerated the upward trend in HIV infections. The mechanism for this effect can be illustrated as follows. Suppose that the numbers of new HIV infections during the first four years of the epidemic were 500, 1,000, 1,500, 2000, respectively, indicating a linear increase of 500/year. Suppose that 5% of HIV infections progress to AIDS during each year following infection, for a median time-to-AIDS of 10 years. During the first year 25 cases of AIDS will occur (5% of 500 infections). During the second year 75 cases of AIDS will occur (5% of 500 plus 5% of 1,000). During the third year 150 cases of AIDS will occur (5% of 500 plus 5% of 1,000 plus 5% of 1,500). During the fourth year 250 cases of AIDS will occur, so the trend in AIDS (25, 75, 150, 250) will initially appear to increase more steeply than the trend in HIV (HIV infections double in year 2, but AIDS cases triple) and then will appear to level off despite no change in the HIV incidence trend. There will also be a change in the ratio of AIDS to HIV, as also occurred during the early years of the epidemic. (The phenomenon was described in an article in the *American Journal of Epidemiology* in about 1987; I am looking for the citation.)

Prevalence versus incidence

	Prevalence	Incidence
Cases	Entities	Events
Source population (PAR)	At risk to be a case	At risk to become a case
Time	Static (point)	Dynamic (interval)
Uses	Planning	Etiologic research

Considerations relevant for both prevalence and incidence

Cases

1. **Case definition** – What is a case?

Examples: arthritis, cholelithiasis, cardiovascular disease, diabetes, psychiatric disorder, epidemiologic treatment of syphilis or gonorrhea, prostate cancer

2. **Case development** – When is a case?

   Issues: induction, latency, progression, reversibility

   Examples: atherosclerosis, cancer, cholelithiasis, diabetes, hypertension, AIDS

3. **Case detection** – When is a case a "case"?

   Issues: Detectability is a function of technology and feasibility. What can be detected is not the same as what is detected.

   Examples: Atherosclerosis, breast cancer, cholelithiasis, osteoporosis, asymptomatic infections, prostate cancer

Source population [Population at risk (PAR)]

1. What is the relevant population — who is really "at risk"?
   E.g., age (most diseases), sex (breast cancer), STD's and sexual activity, uterine cancer and hysterectomy, gallbladder cancer and cholecystectomy, genotypes?
2. What about previous manifestations?

   Of the same disease? (influenza, tumors, injuries)

   Of a related disease (stroke after CHD, cancer at a different site)
3. What about death from other causes? (competing risks)
   E.g., deaths for diabetes reduce the rate of death from coronary artery disease, heart disease deaths reduce the rate of death from lung cancer to the extent that smokers are at excess risk for both

Choosing the right denominator

The choice of the most appropriate denominator can be complex. For example, what is the most appropriate denominator for motor vehicular injuries or deaths?

Total population?

Population age 16 years and above?

Licensed drivers?

Registered vehicles?

Vehicle miles?

Passenger miles?

Which one to choose depends upon whether the question of interest concerns:

Injury risk by age and/or sex (population denominator?)

Effect on risk of seat-belt use (passenger-miles?)

Effect on deaths of 55 mph limit (passenger-miles?)

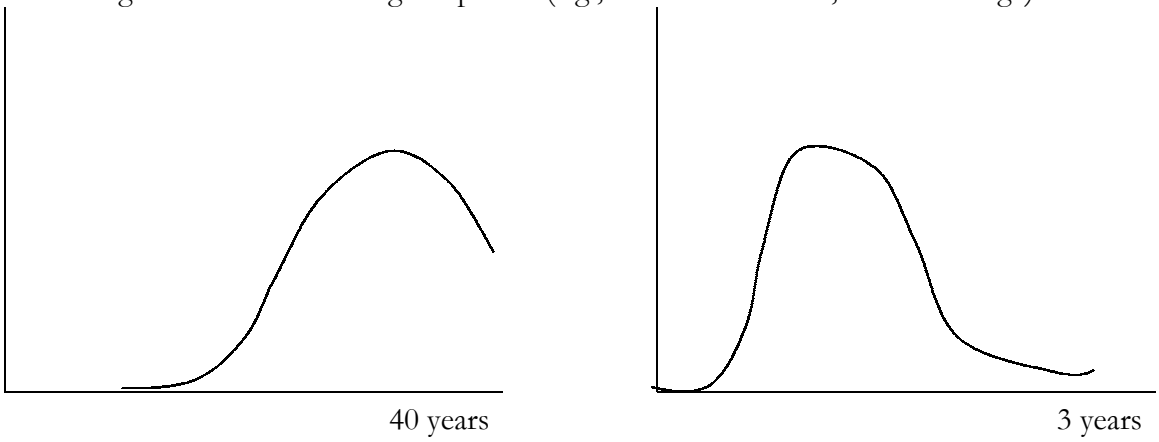
Role of alcohol in motor vehicular fatalities

Evaluation of alternate transportation policies

For example, older drivers have a higher crash rate per 100 million vehicle miles traveled than teen drivers do. But the rate of crashes per licensed driver is no higher for older drivers, because older drivers limit their driving.

Passage of time [incidence only] – what period of observation?

1. Natural history of the disease - period of risk versus period of observation
   E.g., atom bomb survivors and solid tumors, motor vehicle injury, congenital malformations
2. Different periods of observation for different subjects (does 1 person observed for 2 years = 2 people observed 1 year?)
3. Changes in incidence during the period (e.g., seasonal variation, secular change)



Cancer in atomic bomb survivors

Congenital malformations

Types of source populations for incidence

Source populations can be defined in various ways, including residence in a geographical area, employment in a company or industry, attendance in a school or university, membership in an organization, seeking health care from a given set of providers, or explicit recruitment into a study. Incidence involves the passage of time and therefore implies some type of follow-up of population. A key characteristics of a source population is in what ways its membership can change over time. Rothman and Greenland (1998) present a detailed discussion of types of populations and terminology that has been used to describe these. The primary distinction we will make here is that between a **fixed cohort**, whose membership changes only through attrition, and a **dynamic population** (Rothman and Greenland call this an **open cohort**), whose membership can change in various ways. (The fixed cohort versus dynamic population terminology come from Ollie Miettinen by way of Kleinbaum, Kupper, and Morgenstern.)

Cohort – entrance into the population is defined on the basis of some aspect or event in the lives of members of the study population (e.g., living in a geographical area when a major environmental event occurred, start of employment in a worksite or industry, receipt of a medical or surgical treatment, onset of a condition, start of an exposure, or simply enrollment into a study). Exits from the cohort (from death, out-migration, dropout) are problematic; entrances into the cohort are permitted only in relation to the qualifying event that defines the start of follow-up for that person. Note that once recruitment has been completed a cohort will become smaller over time due to attrition, and the entire age distribution will become older.

Variants:

Retrospective or historical cohort - the population is defined at some time in the past (e.g., based on employment records) and then followed forward in time towards the present by the use of available records.

"Dynamic cohort" – follow-up time is counted from the time of entrance into the study or in relation to some event that occurs at different times for different people (e.g., a medical procedure), so that accrual to the cohort continues over a period of time. In a classic cohort study, follow-up time for each subject and calendar time are identical; in a dynamic cohort, each participant's follow-up time may take place over a different interval of calendar time (this does not appear to be a widely-used term).

Dynamic population – a population is defined over a period of time and their experience is monitored during that period. The study population may be defined in the same way (e.g., geographical residence, employment, membership, etc.). In a dynamic population, however, both entrances and exits are expected and accommodated. For example, the population of a geographical area will experience births, deaths, and possibly substantial migration. Over time, a dynamic population can increase or decrease in size, and its age distribution can change or remain the same.

Special case:

A dynamic population is said to be **stable** or **stationary** when its size and age distribution do not change over time. The assumption of stationarity is often made, since it greatly simplifies analysis. (See Rothman and Greenland, 1998 for more on this.)

Types of incidence measures: cumulative incidence (incidence proportion) and incidence density (incidence rate)

There are two major types of incidence measures, differing primarily in the way in which they construct the denominator: **cumulative incidence** and **incidence density** (again, this is Olli Miettinen's terminology, adopted by Kleinbaum, Kupper, and Morgenstern; Rothman and Greenland use **incidence proportion** and **incidence rate**, respectively). Cumulative incidence (CI) is simply the proportion of a population that experience an event or develop a condition during a stated period of time. Incidence density (ID) is the rate at which new cases develop in a population, relative to the size of that population.

Cumulative incidence (incidence proportion)

$$CI = \frac{\text{New cases during stated period}}{\text{Number of persons at risk}}$$

Incidence density (Incidence rate)

$$ID = \frac{\text{New cases during stated period}}{\text{Population-time}}$$

Cumulative incidence (CI), a.k.a. Incidence proportion (IP)

The definition of CI is based on the following "ideal" scenario:

1. A population known to be free of the outcome is identified at a point in time (a cohort);
2. All members of the cohort are at risk of experiencing the event or outcome (at least once) for the entire period of time;
3. All first events or outcomes for each person are detected.

For example, consider a study of the risk that a rookie police officer will suffer a handgun injury during his first six months on patrol duties. Data are collected for a cohort of 1,000 newly-trained police officers entering patrol duties with the San Francisco Police Department (SFPD). During their first six months with the SFPD, 33 of the officers suffer a handgun injury. The other 967 officers have carried out patrol duties during the six-month period with no handgun injuries. The 6-months CI of handgun injury is 33/1,000 = 0.033. We use this observed CI to estimate the six-month risk of handgun injury to new patrol officers in San Francisco.

This example conforms to the ideal scenario for CI: there is a population "at risk" and "in view" for the entire period, and all first events were known. For the moment we assume away all of the reasons that might result in a member of the cohort not remaining "at risk" (e.g., transfer to a desk job, extended sick leave, quitting the force) and "in view" (e.g., hired by another police department).

Some things to note about CI:

1. The period of time must be stated (e.g., "5-year CI") or be clear from the context (e.g., acute illness following exposure to contaminated food source);
2. Since CI is a proportion, logically each person can be counted as a case only once, even if she or he experiences more than one event;
3. As a proportion, CI can range only between 0 and 1 (inclusive), which is one reason it can be used to directly estimate risk (the probability of an event).

Sample calculation:

200 people free of chronic disease X observed over 3 years

10 cases of X develop

3-year CI = 10 cases / 200 people = 10/200 = .05

Thus, the 3-year risk for one of the 200 people to develop disease X, conditional on not dying from another cause, is estimated as 0.05 or 5%.

> **Optional aside** – Assessing precision of an estimated cumulative incidence
>
> Since cumulative incidence is a proportion, a confidence interval can be obtained in the same manner as for prevalence (see above).

Risk and odds

In epidemiology, the term "risk" is generally taken to mean the probability that an event will occur in a given stated or implicit time interval (be alert for other uses, though). In its epidemiologic usage, risk is a conditional probability, because it is the probability of experiencing an event or becoming a case conditional on remaining "at risk" (eligible to become a case) and "in view" (available for the event to be detected).

Any probability can be transformed into a related measure, the "odds". **Odds** are defined as the ratio of the probability of an outcome to the probability of another outcome. When the only outcomes are (case, non-case), then the odds are the ratio of the probability of becoming a case to the probability of not becoming a case. If the risk or probability of becoming a case [Pr(D)] is p, then the odds of becoming a case are p/(1-p). If the risk, or probability, of developing disease X is 0.05 (5%), then the odds of developing disease X are .05/.95 = 0.0526 (the odds always exceed the risk, especially for large risks).

The mathematical properties of odds make them advantageous for various uses. Whereas probabilities are restricted to the 0 – 1 interval, odds can be any nonnegative number. Odds = 1.0 ("fifty-fifty") corresponds to probability = 0.5, the middle of the set of possible values. The logarithm of the odds can therefore be any real number, with log(odds) = 0 corresponding to the middle of the set of possible values. The natural (Naperian) logarithm of the odds (called the "logit", for "logarithmic transformation") is widely used in biostatistics and epidemiology. For the above example, with risk = 5%, odds = 0.0526, the ln(odds), or logit = -2.944; since the ln(odds) is zero when the risk is .5, a risk smaller than 0.5 yields a negative logit. [Rusty on logarithms? See the Appendix on logarithms and exponents.]

Cumulative incidence when there is loss to follow-up

In the example above, all 200 people who were originally free of disease X were observed over all 3 years. What if instead 20 of the people had died of other causes before developing X? Then not all 200 would have been "at risk" for the entire 3 years.

There are four principal alternatives to estimating the 3-year CI:

1. Ignore the deaths:

   3-year CI = 10/200 = .05

2. Ignore the people who died (analyze only the people followed for all 3 years):

   3-year CI = 10/(200-20) = .056

3. Compromise by counting the 20 people who died as if they were 10 people who were at risk for the full 3 years:

   3-year CI = 10/(200-20/2) = .053

4. Use a lifetable, in which (a) CI is computed for each segment of the period (e.g., annually) to estimate the risk during that segment; (b) risks are converted to survival probabilities (1-risk); and (c) risks are multiplied to obtain the 3-year survival probability and therefore the 3-year risk (1 - survival probability).
5. Take the inverse of the Kaplan-Meier estimated survival proportion. This method is the same as the previous one except that the segments are made so short that only a single case occurs in any one segment. Segments with no cases have 100% survival, so the K-M survival estimate is the product of the proportion surviving during each interval when a case occurs.

Each of these methods makes certain assumptions about when the disease occurs during the three-year period, whether it will be detected when it occurs, and whether the people who die of other causes were more or less likely to develop X had they lived.

Incidence density (ID)

$$ID = \frac{\text{New cases during stated period}}{\text{Number of person-years of observation}} \qquad \text{(person months, etc.)}$$

Note that:

- ID is a relative rate, not a proportion.

- The units of time must be stated, since otherwise the numeric value is ambiguous (e.g., 15 cases/100,000 person-years = 15 cases/1,200,000 person-months).*
- Ideally, incidence density is the instantaneous rate of disease occurrence at each moment in time. In practice, epidemiologists generally compute average ID during one or more periods.

Interpretation:

ID addresses the question "How rapidly is the disease occurring in the population, relative to its size?", or "What is the intensity with which the disease is occurring?". It has been argued that ID has no interpretation at the individual level (see Morgenstern H, Kleinbaum, DG, Kupper LL, 1980). However, it is possible that ID can be thought of as at least indirectly addressing the question, "How soon might this happen to me?").

Sample calculation:

In our original example for CI, we had 10 cases of chronic disease X develop in 200 people initially free of X and observed over 3 years with no loss to follow-up. Here are the values of CI and ID for this example:

3-year CI = 10 cases / 200 people = 10/200 = .05

ID ≈ 10 cases / (200 people × 3 years) = 10 / 600 person-years

≈ 0.167 cases per person-year (py) = 0.167 / py = 167 / 1000py

The reason for the approximation is that, as we shall see, people stop contributing person-time when they develop the disease so the denominator must be reduced accordingly. The more nearly correct calculation is 10 / (200 × 3 years – 10 × 1.5 years) = 10/585 = 0.17/py, assuming that cases occurred uniformly during the 3 years.

Calculating ID

In calculating ID, we use the same cases as for CI except that we may want to allow multiple events per person. If we regard the recurrences as independent of one another, then we can simply count

---

* The importance of stating units can perhaps be appreciated from the following:
"On Sept. 23, 1999, NASA fired rockets intended to nudge its Mars Climate Orbiter into a stable low-altitude orbit. But after the rockets fired, NASA never heard from its expensive spacecraft again, and scientists later concluded that it had either crashed on the Martian surface or had bounded away, escaping the planet completely.
"The reason for the debacle, scientists concluded months later, was that the manufacturer, the Lockheed Martin Corporation, had specified the rocket thrust in pounds, while NASA assumed that the thrust had been specified in metric-system newtons." Browne, Malcom W. Refining the art of measurement. Science Times, New York Times, 3/20/2001, page D6¶.

them as new cases. If not, we can define the disease as the first occurrence. Other considerations can also affect the choice.

There are several methods used to compute population-time.

1) If individuals are being followed over time, so that the period of disease-free observation is known for each person, we simply add up the disease-free time for all persons:

   population-time = Σ (disease-free time for each person)

2) If a fixed cohort is being followed, but not in sufficient detail to know the period of disease-free time for each individual, we can estimate population time as follows:

   population-time = average population size during the period
   × length of the period of observation

   If there are N_0 disease-free people at the beginning of the period, and during the period there are "C" cases, "D" deaths from causes other than the disease of interest, and "W" persons whose disease status is unknown ("withdrawals"), then the number of disease-free persons at the end of the period is $(N_0 - C - D - W)$. The average number of disease-free people, assuming that cases, deaths, and withdrawals occur uniformly during the period, is:

$$\frac{N_0 + (N_0 - C - D - W)}{2} = (N_0 - C/2 - D/2 - W/2)$$

   and the population-time at risk can be estimated as:

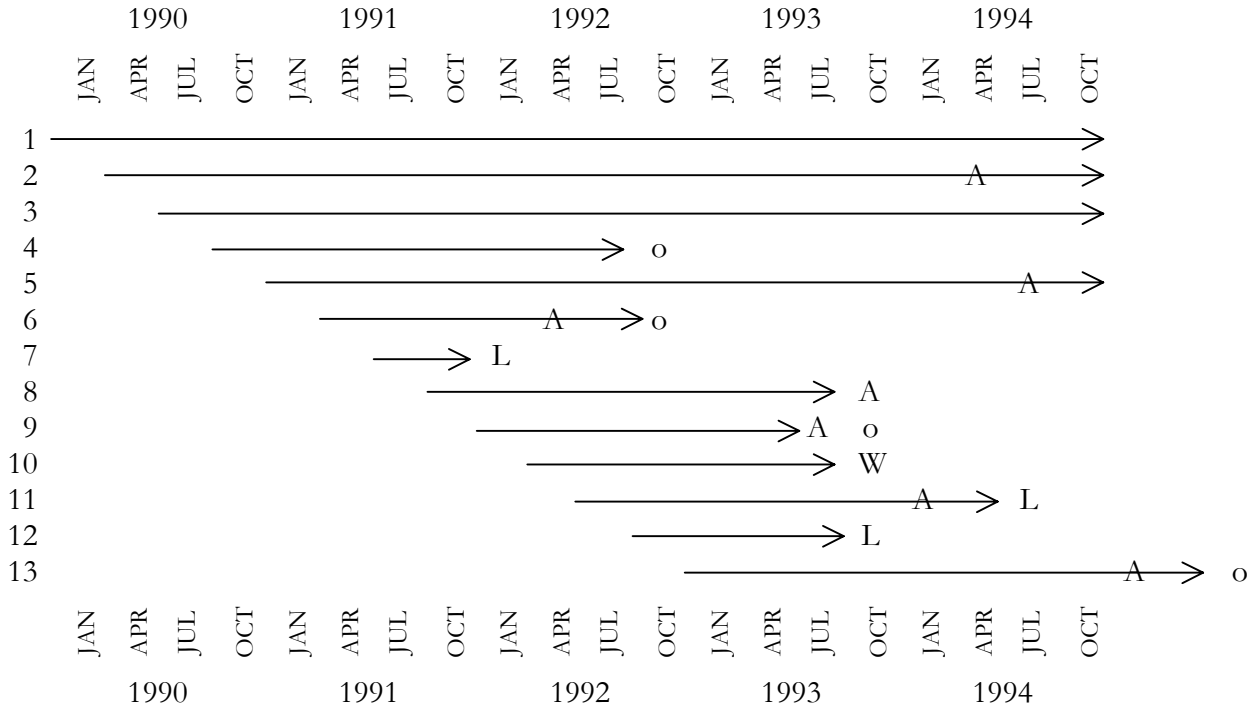
$$(N_0 - C/2 - D/2 - W/2) \times (\text{time interval})$$

3) If we are following a dynamic population (a.k.a. "open cohort") instead of a fixed cohort, we can use the same strategy of multiplying the average size of the disease-free population by the time period. It may be possible to estimate the average number of disease-free people by taking the average of the number of disease-free people at the beginning and end of the period. If we can assume that the population is "stable" (the number of disease-free people who are lost to the population through out-migration, death, and developing the disease of interest is balanced by in-migration), then the number of disease-free people is approximately constant. If we have any usable estimate of the average number of disease-free persons (N_0), then we estimate population time as $N_0 \times$ (time interval)

   If the disease is rare, then the number of disease-free persons (N_0) will be approximately equal to the total number of persons (N), which is more likely to be known. In that case, we can estimate population time as $N \times$ (time interval), where N is the average population size without regard to disease status. Annual death rates and other annual vital statistics rates are typically computed using the estimated mid-year (July 1) population as the denominator, which is approximately the average size of the population on any day in the year if the population is approximately constant or changing in a monotonic fashion.

Calculation of person-time in a cohort
when individual follow-up times are known

Graph of hypothetical follow-up experience for 13 advanced Alzheimer's patients being cared for at home during January 1990 - December 1993 and followed until December 31, 1994 for admittance to a nursing home, in order by study entrance date (after Kleinbaum, Kupper, and Morgenstern, 1982).



Key:

A = admitted to nursing home care

L = lost to follow-up

W = withdrew

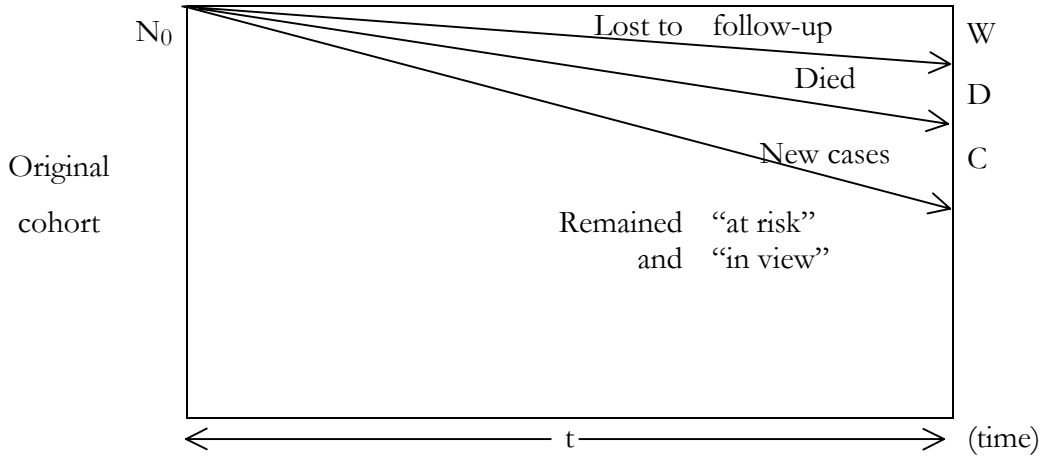
o = died

$$ID = \frac{\text{Cases}}{\text{Sum of disease-free follow-up over all individuals}}$$

Subject	Cases	Follow-up
1		5.0
2	1	4.0
3		4.5
4		2.0
5	1	3.5
6	1	1.0
7		0.5
8	1	2.0
9	1	1.5
10		1.5
11	1	1.5
12		1.0
13		2.0
Total	6	30.0

$$ID = \frac{6}{30 \text{ person-years}} = 0.20 \text{ patients admitted per year}$$

Calculation of person-time in a cohort
when individual follow-up times are not known

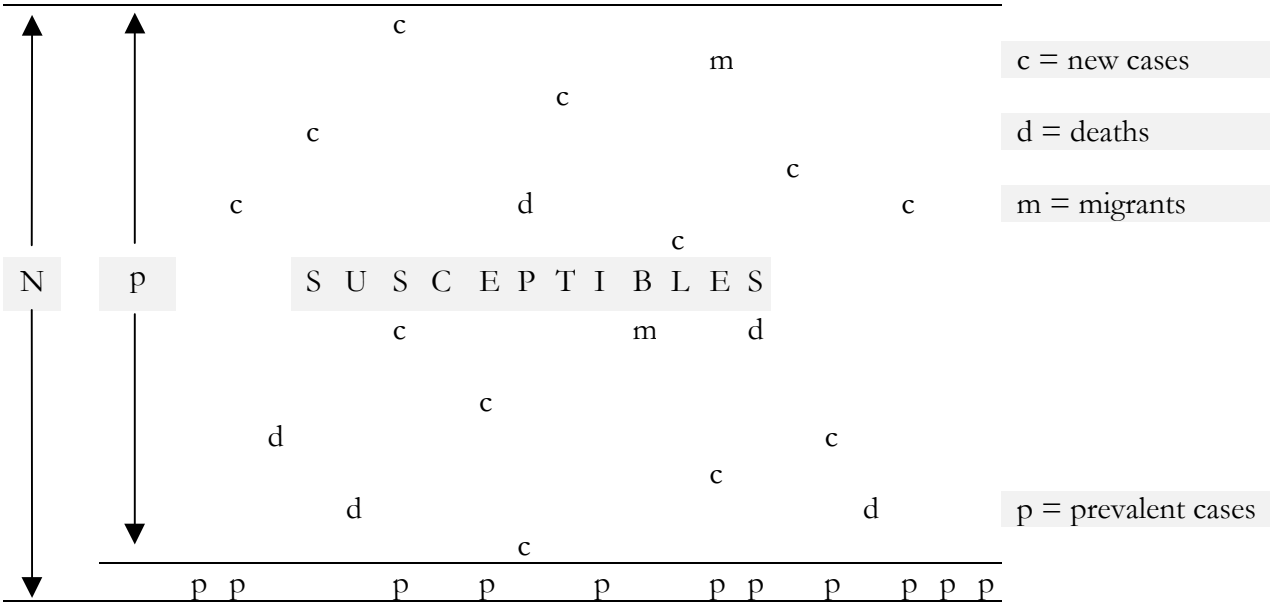


$$ID = \frac{C}{(N_0 - W/2 - D/2 - C/2)\, t}$$

(t = time interval)

(Since the area of a triangle = base × height/2, the person-time lost to follow-up can be estimated by one half times the number of withdrawals [the base of the triangle] times the length of the time interval [the height]. The procedure is the same for follow-up time lost due to deaths and to incident cases. These estimates assume that cases are detected as they occur and that only the first case per subject is counted.)

Calculation of person-time in a stable, dynamic population



Processes at work: Immigration of cases, noncases
Out-migration of cases, noncases
Death of cases, noncases
Development of new cases

$$ID = \frac{\text{cases}}{N_0 t} \quad \text{or} \quad ID = \frac{\text{cases}}{Nt}$$

(t = time interval)

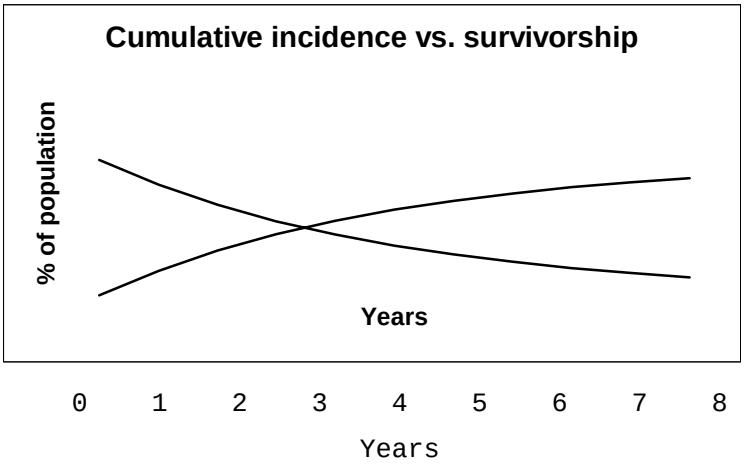
Relationship of CI and ID

Both ID and CI are actually old acquaintances who have changed their outfits. When we calculated life expectancy in the first topic, we used the terms death rate, hazard, cumulative mortality, cumulative survival. ID is essentially the hazard, now applied to events other than death. CI is essentially the cumulative mortality proportion, now applied to events of any variety. Both represent different summary statistics from survivorship analysis (known in engineering as failure-time analysis).

ID is the rate at which the size of the unaffected population is changing, relative to the size of the unaffected population; CI is the proportion of the original population that has been affected by time t. CI is a cumulative measure from a baseline time to a specific later point in time. CI estimates the

average risk for a member of the cohort. In principle, ID can apply to an instant in time, though it can be computed only as an average over some interval. ID is sometimes referred to as the "force of morbidity", in analogy to the hazard function (the "force of mortality").

The following figure shows the relationship between CI and its inverse, the proportion unaffected (survivorship). ID is the relative rate of decline in the survivorship curve.

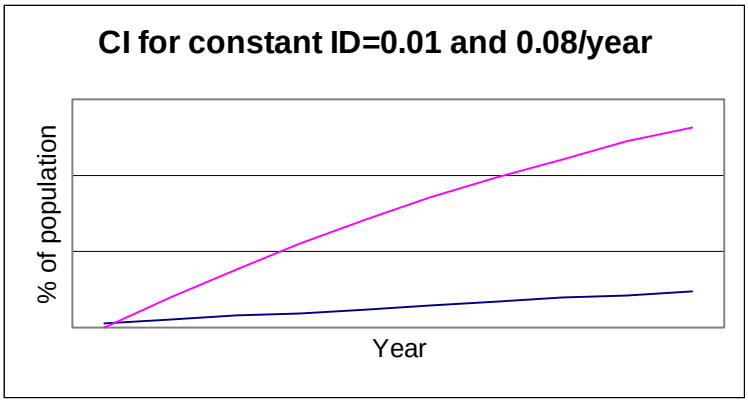


The incidence of AIDS in gay men in San Francisco from 1984 might look something like the left half of this graph.

The mathematical relationship between CI and ID over time can be seen by considering an incurable disease in a hypothetical fixed cohort defined at a point in time and with no entrances or exits other than from the disease in question. Assuming that ID_t (the force of morbidity) is constant over time, cases will develop throughout the follow-up period. However, since the number of unaffected (at risk) cohort members is diminishing, the number of new cases will be smaller in each successive time interval. Because the number of cases is smaller in each interval, the slope of the curve for CI will tend to flatten out as it approaches 1.0 (its maximum value), at which time the entire cohort has developed the disease. The proportion unaffected (the inverse of CI: 1-CI) also becomes less steep. ID_t, of course, we have assumed to be constant. In this situation, the mathematical relationship between CI and ID is:

$$CI = 1 - \exp[-\int(ID_t dt)] = 1 - \exp(-ID\ \Delta t)$$

For a rare disease with a constant ID (or during a sufficiently short time interval): $CI \approx ID \times \Delta t$ (where Δt is the time interval), because since the cohort does not become depleted, the number of new cases in each time interval remains about the same.



Example:

- ID = 0.01/year (1 case per 100 person-years)
- In 5 years, CI will be 0.049, or about the same as ID×5 (=0.05); 95% of the cohort remains disease free and therefore exposed to the 0.01/year ID.
- In 10 years, CI will be .096, only slightly below ID×t (=0.10); 90% of the cohort remains disease free.
- ID = 0.05/year (5 cases per 100 person-years)
- In 5 years, CI will be 0.226, slightly smaller than ID×5 (=0.25); 77% of the cohort remains disease free.
- In 10 years, CI will be 0.40, while ID×t (=0.50); only 60% of the cohort remains disease free.

CI vs. ID - a real-life example

(courtesy of Savitz DA, Greenland S, Stolley PD, Kelsey JL. Scientific standards of criticism: a reaction to "Scientific standards in epidemiologic studies of the menace of daily life", by A.R. Feinstein. *Epidemiology* 1990;1:78-83; it was actually Charles Poole who spotted this *faux pas* [Poole C, Lanes SF, Davis F, *et al*. "Occurrence rates" for disease (letter). *Am J Public Health* 1990; 80:662]; the specific issue being discussed is the effect of alcohol on breast cancer risk)

> ". . . substantially different occurrence rates of breast cancer: about **6.7 per thousand** (601/89,538) in the nurses cohort and about **18.2 per thousand** (131/7,188) in the NHANES cohort." (Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 1988;242:1259 quoted in Savitz DA et al., p.79, emphasis added)
>
> Implication:
>
> (1) Different rates suggest errors in ascertainment of breast cancer
>
> (2) With under/overascertainment, there may be biased ascertainment
>
> (3) The bias may produce more complete or overdiagnosis among drinkers

However:

Nurses: 601 cases/89,538 women over 4 years

CI = 6.7 per thousand (4 years)

ID = 1.68 per 1,000 women-years

NHANES: 121 cases/7,188 women over 10 years (10 cases should have been excluded by Feinstein)

CI = 16.8 per thousand (10 years)

ID = 1.68 per 1,000 women-years

This example illustrates the importance of stating the follow-up period for a CI and the problem that can arise in comparing CI's for different amounts of follow-up.

Two complementary measures of incidence: CI and ID

Cumulative incidence (CI)

1. increases with period of observation (i.e., it is "cumulative")
2. has problems with:
   - multiple events in one subject
   - differing follow-up times for subjects

But

3. it is not necessary to know exact time of onset of the disease
4. directly estimates risk

Incidence density (ID)

1. suggests ability to extrapolate over time - "duration free";
2. accommodates:
   - multiple events in one subject
   - different follow-up times for subjects
3. does not require a cohort to estimate or interpret
4. may be more appropriate for etiologic inference

Choosing between CI and ID

A. Objective

Estimate rate or risk

B. Natural history

Does the period of interest fit within the period of observation? (restricted versus extended risk period)?

E.g., If one wanted to analyze the relative longevity of men and women, the lifetime risk (CI) of death would be useless.

C. Availability of data, e.g.

Fixed cohort, dynamic cohort, dynamic population

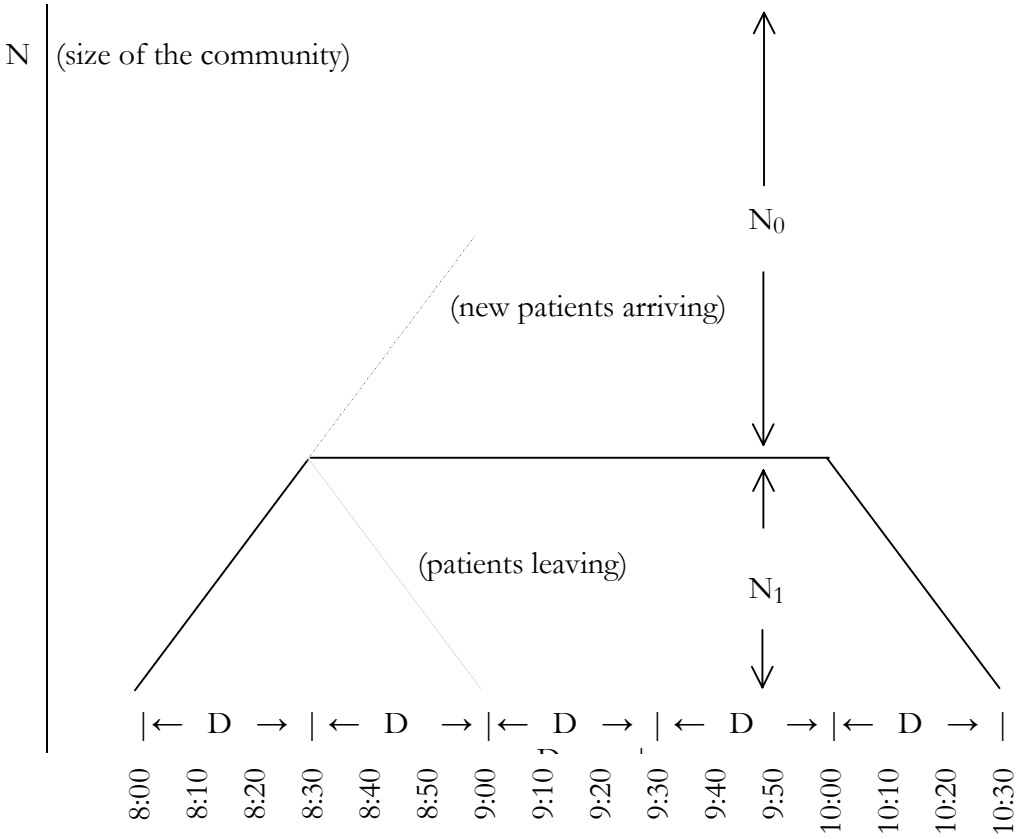
Different follow-up times

Knowing when events occur may favor one method or the other.

Incidence and prevalence in a population

The relationship between incidence and prevalence is the population-level analog for many familiar situations, such as the number of people on line at the grocery store check-out, the number of patients in a waiting room or a hospital, or the number of simultaneous log-ins for an internet service provider.

Incidence, prevalence, and duration: patient flow in a community-based clinic



If a clinic opens at 8:00am, a patient arrives every 10 minutes (6/hour), and it takes 30 minutes for a patient to be seen and treated, then the number of patients in the clinic will rise for the first 30 minutes and then remain constant at 3 patients until the clinic closes and the last 3 patients are

treated. If the rate at which patients arrive were to increase to 10/hour, then in the half-hour it takes to treat the first patient 5 more will arrive, so the number of patients in the clinic will stabilize at 5, instead of 3. Similarly, lengthening the treatment time from 30 to 60 minutes would cause the number in the clinic to increase for the first hour, for a total of 6 patients in the clinic at any time until closing.

With the original assumptions, 6 patients arrive at the clinic every hour during 8:00am-10:00am, and 6 patients leave the clinic each hour during 8:30am-10:30am. During 8:30am-10:00am the clinic is in equilibrium, with 3 patients there at any given time. This equilibrium number, N_1, equals the arrival rate (6/hour) times the average time a patient remains (0.5 hours):

$$N_1 = \text{arrival rate} \times D$$

where D is average duration of a clinic visit.

If the clinic is the only one in a community of size N (or is the approved source of care for N people), then we can express the arrival rate as a function of the size of the community:

$$\text{Arrival rate (patients/hour)} = I \times N_0$$

where I is the incidence of visiting the clinic and N_0 is the number of people available to go to the clinic (N minus the N_1 people already in the clinic, which assumes that people can return to the clinic as soon as they leave or that they immediately leave the community and are replaced by other people eligible to go to the clinic). We can also express the number of patients in the clinic, N_1, as a function of the size of the community, using P as the population "prevalence" of clinic attendance.

$$N_1 = P \times N$$

Making use of these three equations, we can write:

$$\begin{aligned} N_1 &= \text{arrival rate} \times D \\ &= (I \times N_0) \times D \\ P \times N &= (I \times N_0) \times D \end{aligned}$$

$$P = \frac{N_0}{N} \; I \times D$$

Prevalence odds = incidence × average duration

If the number of visitors to the clinic is small in relation to the size of the community, then $N_0/N \approx 1$, and we have the approximation **prevalence = incidence × average duration**.

Otherwise the relationship can be written as **prevalence odds = incidence × average duration**, since:

$$P = \frac{N_0}{N} \; I \times D = \frac{N - N_1}{N} \; I \times D$$

$$P = (1 - P) \times I \times D \quad \text{and} \quad \frac{P}{(1 - P)} = I \times D$$

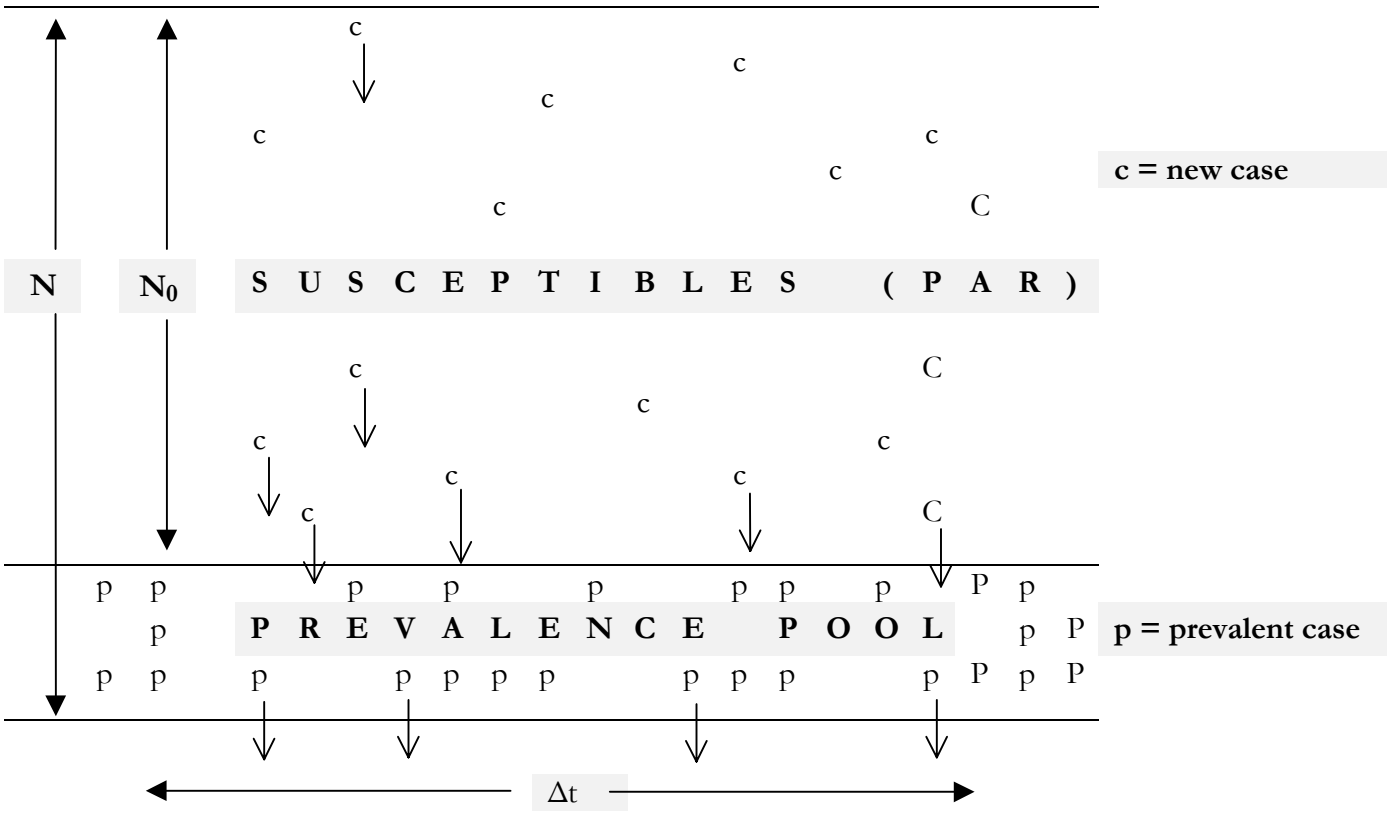
Odds are defined as the ratio of two probabilities, most often the ratio of a probability divided by its inverse (probability for/probability against). The prevalence of a condition is an estimate of the probability that a randomly selected member of the population is a case [Pr(case)]. If the prevalence is p, then the **prevalence odds** are **p/(1-p)**. So the prevalence odds, i.e., the odds that a randomly selected person in the population has the disease (i.e., is a prevalent case) are:

$$\begin{aligned} \text{prevalence odds} &= \text{prevalence} \; / \; (1 - \text{prevalence}) \\ &= (N \times \text{prevalence}) \; / \; (N - N \times \text{prevalence}) \\ &= (N \times \text{prevalence}) \; / \; N_0 \;\; = \;\; (N/N_0) \times \text{prevalence} \end{aligned}$$

Incidence, prevalence, and duration in a stationary population

The following diagram displays the above process as it might appear for cases of a disease occurring in a population followed during an interval of time, in equilibrium with respect to disease incidence, duration, and entrances and exits from the population. An alternate derivation of the relation prevalence odds = incidence × duration follows. (See Rothman and Greenland, 1998 for more on this topic.)

Incidence and prevalence in a population of size N observed for a time interval Δt



c's are incident (new) cases

p's are prevalent (existing) cases

Δt indicates the time interval

↓ indicates exits from unaffected population or from prevalence pool

Size of the population = N = disease-free persons + existing cases = N_0 + prevalence pool

The assumption that incidence and prevalence are constant means that:

$$\text{New cases} = \text{Terminations}$$

$$(\text{Incidence} \times N_0) \times \Delta t = (\text{Prevalence} \times N \times \text{Termination rate}) \times \Delta t$$

$$\text{Prevalence} \times \frac{N}{N_0} = \frac{\text{Incidence}}{\text{Termination rate}}$$

Since the termination rate is the rate at which existing cases leave the prevalence pool, this rate is the reciprocal of the average duration of a case. To see this, consider the termination rate for a single case:

$$\text{Termination rate} = \frac{\text{Terminations}}{\text{No. of cases} \times \Delta t}$$

For a single case,

$$\text{Termination rate} = \frac{1}{1 \times \Delta t} = \frac{1}{\Delta t}$$

$$\text{Average duration (i.e., } \Delta t) = 1 \text{ / Termination rate}$$

Thus, in the above relationship between incidence and prevalence, we can substitute Duration (D) for 1 / Termination rate:

$$\text{Prevalence} \times \frac{N}{N_0} = \text{Incidence} \times \text{Duration}$$

So in a population that is in a steady state with respect to a given condition, the prevalence odds of that condition equals the incidence times the average duration (the prevalence does too, if it is sufficiently small). Conversely, if we observe that the prevalence odds of a condition remains constant (and can assume a stable population with no net migration of cases), then the incidence must balance the loss of cases due to death or cure. Since prevalence is often easier to ascertain than is incidence, we can make use of this relationship to draw inferences about incidence.

Estimating incidence from prevalence data

This relation has been used as the basis for estimating HIV seroincidence from seroprevalence data, using a seroassay procedure designed to identify recently-infected persons (Janssen et al., 1998). This technique makes use of the fact that ELISA tests for HIV antibody have become considerably more sensitive since they were first developed. People who test HIV-positive with a current (highly sensitive) HIV antibody test are then re-tested with a "detuned" version of an older, less-sensitive test. Since it takes time for the anti-HIV antibody titer to increase to the level that it can be detected with the less sensitive test, there is a period of time (about four months) during which the less sensitive test will be negative. The discordant results of the two HIV antibody tests defines a short-lived (average duration 129 days) "condition" whose prevalence can be used to estimate occurrence of new HIV infections. Solving the relation Prev odds = I × D yields I = Prev odds / D ≈ P/D for

small prevalence. So if the seroprevalence of recent infection in a stable population is 2%, the incidence of new HIV infections is approximately 0.02/129 days = 0.057/year. = 5.7/100py.

However, the stable population assumption is often not met in practice, and the above model is also grossly simplified in that it treats the entire population as a homogenous entity, ignoring the influence of age (see Rothman and Greenland, 1998). When we examine the relationship between incidence and prevalence within a specific age-band, we need to consider the effect of entrances and exits due to aging into or from the age band of interest. For example, the U.S. armed forces have conducted serologic testing for HIV antibody of all recruits, active duty military, and reserves since the antibody test became available. Within each age group, the seroprevalence of HIV antibody has been approximately constant over a period of years. If we could ignore the effect of age, then using the relationship prevalence odds = incidence × average duration, we could conclude that HIV incidence should equal the (small) proportion of the population who leave the prevalence pool each year due to discharge or death. However, another manner of exiting from the prevalence pool of a given age group is to age out of it into the next one. Since HIV seroprevalence increases with age (up to about age 35 years), it can be inferred that infections (incident cases) are occurring more rapidly than necessary to balance deaths and discharges among cases. The reason is that each year, some of the persons in each age group are replaced by persons from the next younger age group, a group with lower seroprevalence. If infections were not occurring at a rate sufficient to balance this outflow of prevalent cases, then the prevalence in each age group would decrease over time, as the lower prevalence groups move up in age (see David Sokol and John Brundage, Surveillance methods for the AIDS epidemic, *NYS J Medicine* May 1988).

Furthermore a meaningful incidence measure still requires identification of a cohort or source population. Although the detuned serologic assay for recent HIV infection has been used to estimate HIV "incidence" among clinic patients, the interpretation of those estimates is highly problematic (Schoenbach, Poole, and Miller, 2001).

Bibliography

Bailar, John C., III; Elaine M. Smith. Progress against cancer? *N Engl J Med* 1986; 314:1226-32. (included in an assignment)

Elandt-Johnson, Regina C. Definition of rates: some remarks on their use and misuse. *Am J Epidemiol* 1975;102:267-271.

Gable, Carol Brignoli. A compendium of public health data sources. *Am J Epidemiol* 1990; 131Z:381-394.

Gaffey, WR. A critique of the standardized mortality ratio. *J Occupational Medicine* 1976;18:157-160.

Glantz, Stanton H. *Primer of biostatistics*. NY, McGraw-Hill, 1981.

Hook, Ernest B. Incidence and prevalence as measures of the frequency of birth defects. *Am J Epidemiol* 1983;116:743-7

Janssen RS, Satten GA, Stramer SL, et al. New testing strategy to detect early HIV-1 infection for use in incidence estimates and for clinical and prevention purposes. JAMA 1998; 280:42-48.

Morgenstern H, Kleinbaum, DG, Kupper LL. Measures of disease incidence used in epidemiologic research. *Int J Epidemiology* 9:97-104, 1980.

Remington, Richard D. and M. Anthony Schork. *Statistics with applications to the biological and health sciences*. Englewood Cliffs, NJ, Prentice-Hall, 1970.

Rothman, Kenneth J. Clustering of disease. Editorial. *Am J Public Health* 1987; 77:13-15.

Victor J. Schoenbach, Charles Poole, William C. Miller. Should we estimate incidence in undefined populations? *Am J Epidemiol* 2001;153(10).935:937.

Smouse, Evan Paul; Martin Alva Hamilton. Estimating proportionate changes in rates. *Am J Epidemiol* 1983; 117:235-43.

Zeighami EA, Morris MD. The measurement and interpretation of proportionate mortality. *Am J Epidemiol* 1983; 117:90-7.

Appendix on weighted averages

Because epidemiology studies populations, and populations contain various subgroups, weighted averages figure prominently in epidemiology. Nearly any population-based measure can be regarded as a weighted average of the value of that measure across the subgroups that comprise the population. Weighted averages are used to standardize or adjust crude measures to make them more comparable across populations with different subgroup proportions. Both the concept and the mathematics are fundamental.

A weighted average is like an ordinary mean except that the components being averaged can have more or less influence (weight) on the resulting average. For example, suppose we measure systolic blood pressure on 10 occasions and obtain the following values (mmHg): 95, 100, 100, 105, 105, 105, 110, 110, 115, 120. If we want the mean (average) systolic blood pressure, we simply sum the individual measurements and divide by the number of readings: 1,065/10 = 106.5 mmHg. Since some of the readings occur more than once, we could achieve the same result by using a weighted average:

Number of readings	Value	Weighted sum
1	95	95
2	100	200
3	105	315
2	110	220
1	115	115
1	120	120
10		1,065

Average = 1,065 / 10 = 106.5 mmHg.

A small business might use a layout like this to compute the average price paid for some commodity over some time period. In that situation, the first column might show the number of sacks purchased, the second column the price per sack, and the third column the total dollar amount.

With a little generalization (to permit the "number of readings" to be a fractional number), we have the procedure for creating a weighted average. Familiar examples are grade-point averages (course grades weighted by credit hours), average cost per share of a stock purchased in multiple buys, and average price per gallon for gasoline purchased on vacation.

Mathematically, a weighted average is a linear combination where the coefficients (p_i) are proportions whose sum is 1.0. Several equivalent formulations are:

$$\frac{w_1a_1 + w_2a_2 + \ldots + w_na_n}{w_1 + w_2 + \ldots + w_n} = \frac{w_1a_1 + w_2a_2 + \ldots + w_na_n}{W}$$

$$= \frac{w_1a_1}{W} + \frac{w_2a_2}{W} + \ldots + \frac{w_na_n}{W} = \sum\left(\frac{w_ia_i}{w_i}\right)$$

$$= p_1a_1 + p_2a_2 + \ldots + p_na_n = \sum(p_ia_i)$$

where $W = w_1 + w_2 + \ldots + w_n$ and $p_1 + p_2 + \ldots + p_n = 1$

For the gasoline price example, the w_i represent the amount purchased at each stop and the a_i represent the price of each purchase.

Appendix on exponents and logarithms

(Adapted from Defares JG and Sneddon IN. *An introduction to the mathematics of medicine and biology*. The Netherlands, North-Holland, 1960)

Some simple facts:

$2^2 = 2 \times 2 = 4$

$2^3 = 2 \times 2 \times 2 = 8$

Square root of 4 = 2

Cube root of 8 = 2

Exponents:

b^x means b raised to the x^{th} power; x is referred to as an exponent.

If x is 2, then $b^x = b^2 = b \times b$. If x is 3, then $b^x = b^3 = b \times b \times b$. From this we can reason that:

1) $b^m \times b^n$ must be equal to $b^{(m+n)}$

(The product of a number raised to the m-th power multiplied by the same number raised to the n-th power equals that number raised to the sum of the powers.)

2) b^m / b^n must be equal to $b^{(m-n)}$

(The quotient of a number raised to the m-th power divided by the same number raised to the n-th power equals that number raised to the difference of the powers (numerator power minus denominator power.)

3) $(b^m)^n$ must be equal to $b^{(m \times n)}$

(The m-th power of a number raised to the n-th power equals that number raised to the (m × n)-th power.)

For exponents that are not positive integers, we define b^x in order to preserve the above three rules. So $b^0=1$ and $b^{-x} = 1 / b^x$.

When the base number (b in the above examples) is e, a transcendental number that is approximately 2.7183, then we write e^x or (for typographical convenience) exp(x). e and Naperian logarithms have special properties that recommend them for use in mathematics and statistics.

Logarithms:

If for a number b (greater than 1.0), it is possible to find a number x such that:

$$y = b^x$$

then we say that x is the logarithm of y to the base b:

$$x = \log_b y$$

Taking the logarithm is the inverse of exponentiation, so that if $y=b^x$:

$$\log_b(y) = \log_b(b^x) = x \qquad \text{and}$$

$$b^x = b^{(\log(y))} = y$$

To preserve consistency with the rules for exponents, above, we see that:

1) $\log_b(xy) = \log_b x + \log_b y$

(the logarithm of a product is the sum of the logs)

2) $\log_b(x/y) = \log_b x - \log_b y$

(the logarithm of a quotient equals the logarithm of the numerator minus the logarithm of the denominator), and

3) $\log_b(x^n) = n \log_b x$

Logarithms are defined so that these rules generalize to the case of fractional and negative exponents. The log of a negative number, however, is undefined.

The base b must be a positive number greater than 1.0, and is usually 10 (for "common logarithms") or e (for "natural" or Naperian logarithms). The latter are most often seen in mathematics, statistics,

and epidemiology. The notation ln(x) or simply log(x) is often used when Naperian logarithms are understood.

Note that for base e (= 2.7183), exp(x) (a) must be greater than zero, (b) will equal 1 when x=0, and (c) will increase very rapidly for large x. In contrast, ln(x) (a) will be negative for x<1, (b) will equal 0 when x=1, and (c) will be positive for x>1. So if x is a positive ratio whose null value is 1.0, ln(x) will represent a transformation of x with the null value at 0 and other values distributed symmetrically around it. These properties of logarithms are useful for transforming variable distributions and for the analysis of ratios.