Workshop on Health Analytics and Disease Modelling, Organized by National Disease Modelling Consortium

Statistics and Data Analysis in Public Health

Siuli Mukhopadhyay
Department of Mathematics, IIT Bombay

Topics

- Basic concepts in epidemiology
  - Incidence and Prevalence measures
  - Mortality measures
  - Epidemiological studies
  - Relative risk and odds-ratio
- Statistical concepts in Health
  - Probability distributions
  - Estimation
  - Hypothesis Testing

What is Epidemiology?

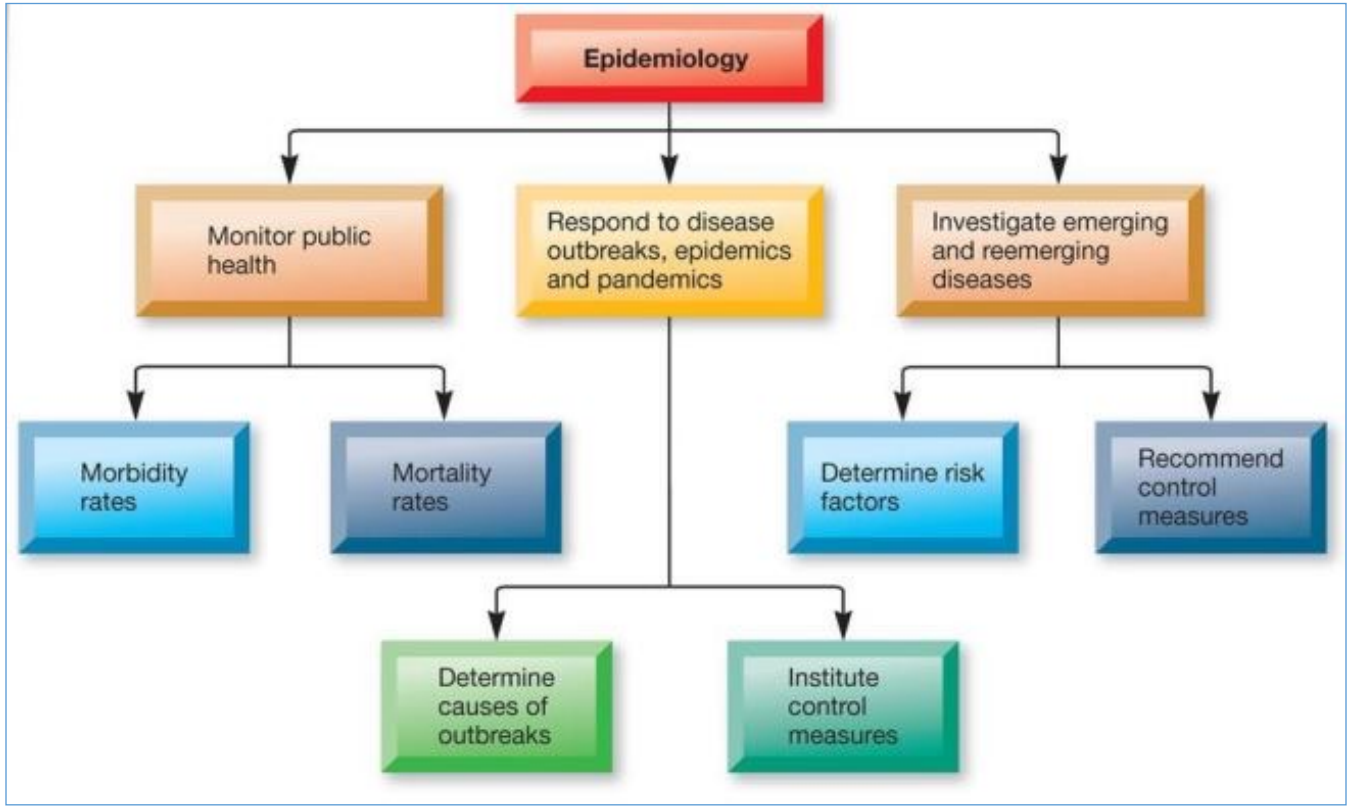


Fig. 1 Epidemiology is the investigation of the mass aspects of disease. An epidemiologist attempts to determine the various factors and remedies related to a disease and its transmission among populations. (Copyright © McGraw-Hill Education)

Disease Transmission

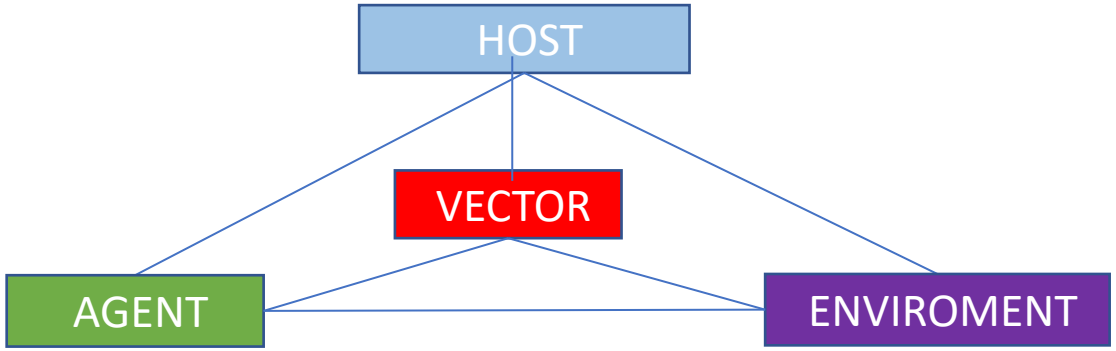


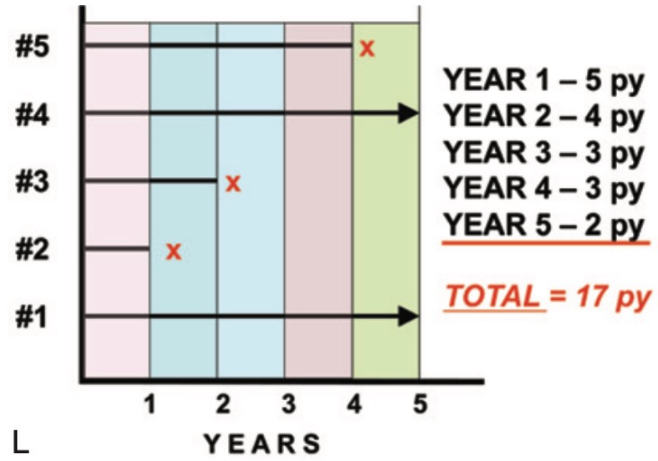
Fig. 2: The Epidemiologic triad of a disease

TABLE 1: Factors That May Be Associated With Increased Risk of Human Disease

Host characteristics	Types of Agents and Examples	Environmental Factors
Age Sex Race Religion Customs Occupation Genetic Profile Marital Status Family Background Previous Diseases Immune status	Biological: Bacteria, Viruses Chemical: Heavy metals, Alcohol, Smoke Physical: Trauma, Radiation, Fire Nutritional: Lack, Excess	Temperature Humidity Altitude Crowding Housing Neighborhood Water Milk Food Radiation Air Pollution Noise

Measures of Morbidity: INCIDENCE RATE

$$\text{Incidence rate per 1,000} = \frac{\text{No. of new cases of a disease occurring in the population during a specified period of time}}{\text{Total person-time (the sum of the time periods of observation of each person who has been observed for all or part of the entire time period)}} \times 1{,}000$$



In this rate, the result has been multiplied by 1,000 so that we can express the incidence Per 1,000 persons.

Important Facts about Incidence Rate

- *NEW* cases of disease.
- Incidence rate is a measure of events—the disease is identified in a person who develops the disease and did not have the disease previously.
- Incidence rate is a measure of risk since it is a measure of events (i.e., transition from a non-diseased to a diseased state)
- Two types of denominators: people at risk who are observed throughout a defined time period; or, when all people are not observed for the full time period, person-time (or units of time when each person is observed)
- This risk can be looked at in any population group, such as a particular age group, among males or females, in an occupational group, or a group that has been exposed to a certain environmental agent

PREVALENCE

- Prevalence is defined as the number of affected persons present in the population at a specific time divided by the number of persons in the population at that time;
- Prevalence is the proportion of the population is affected by the disease at that time

$$\text{Prevalence per 1,000} = \frac{\text{No. of cases of a disease present In the population at a specified time}}{\text{No. of persons in the population at that specified time}} \times 1{,}000$$

Two Types of Prevalence

- **Point prevalence:** Prevalence of the disease at a certain point in time
- **Period prevalence:** How many people have had the disease at any point during a certain time period?

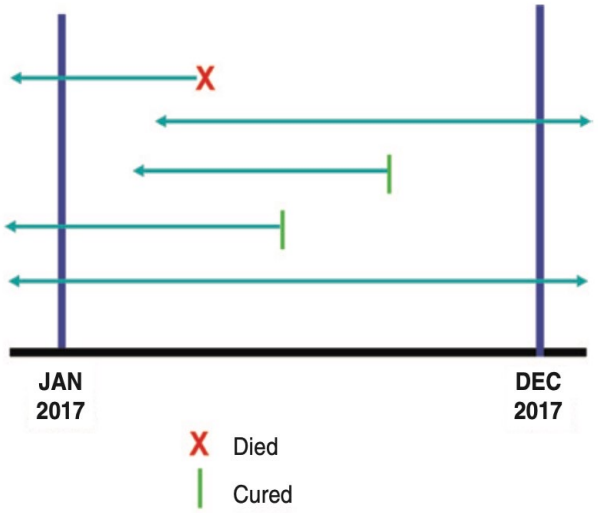


Fig 3.1 Example of incidence and point prevalence

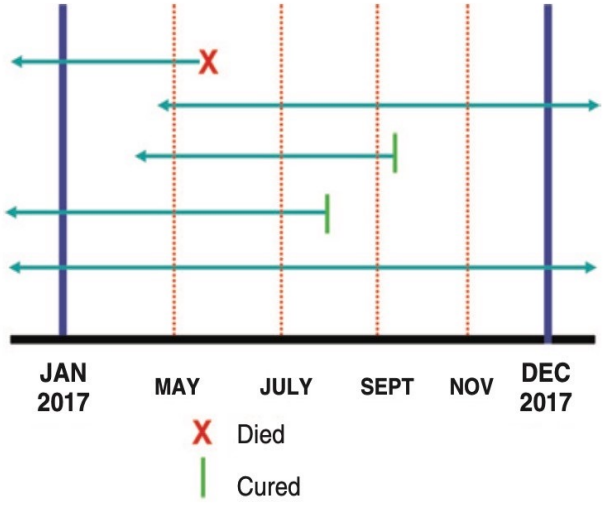


Fig 3.2 Example of incidence and period prevalence

RELATIONSHIP BETWEEN INCIDENCE AND PREVALENCE

- In a steady-state situation, in which the disease rates are not changing and in- migration equals out- migration, and when the prevalence is not too high, the following equation applies:

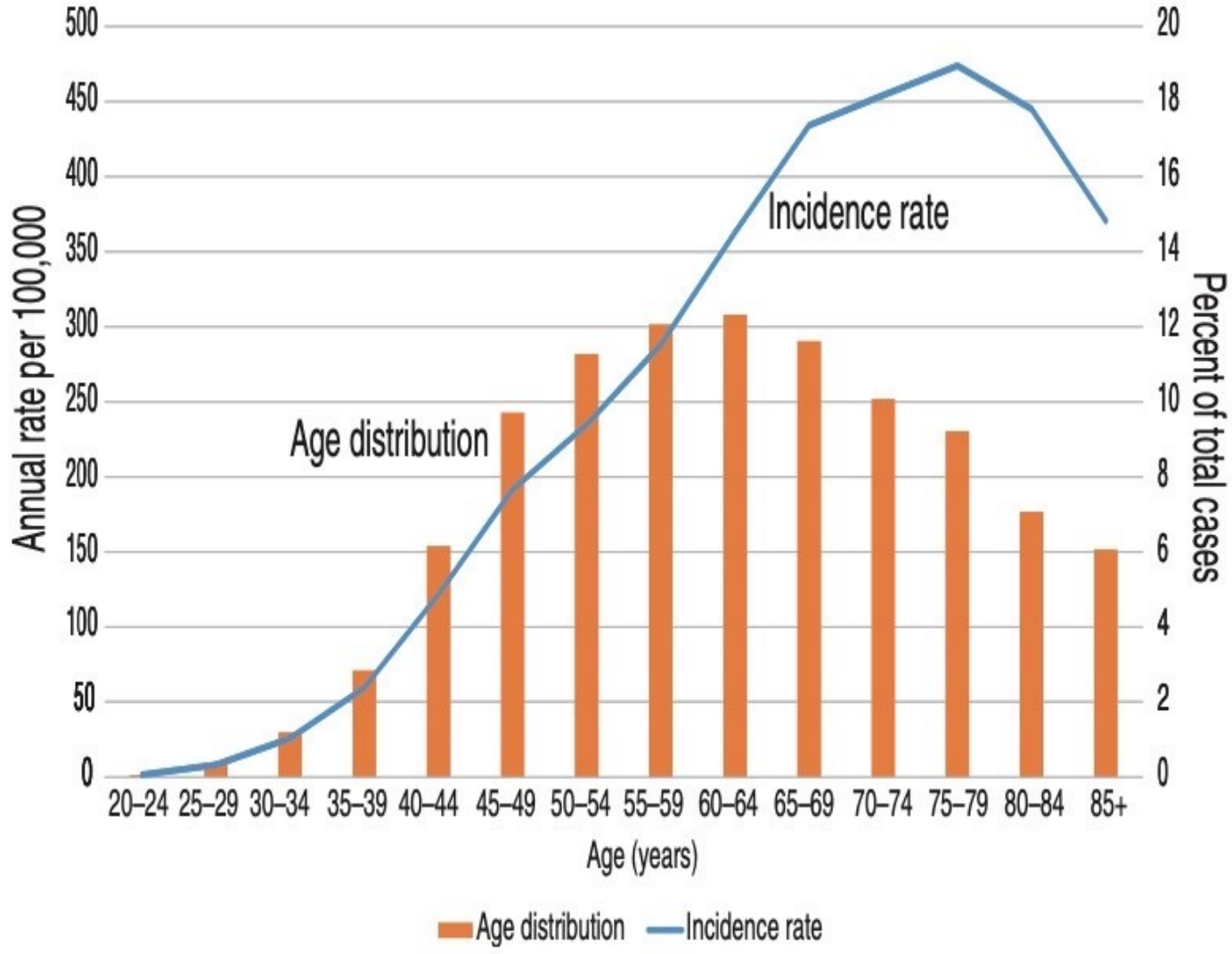
Prevalence = Incidence × Duration of Disease

TABLE 3.4 HYPOTHETICAL EXAMPLE OF CHEST X-RAY SCREEING: 3. PREVALENCE, INCIDENT, AND DURATION.

Screened population	Point Prevalence Per 1,000	Incidence (Occurrences/year)	Duration (year)
Hi-town	100	4	25
Lo-town	60	20	3

A proportion is not a rate

FIG. Breast cancer incidence rates in white women and distribution of cases by age, 2000-13.(Data from www.seer.cancer.gov)



Measures of Mortality

- The absolute *number* of people dying from cancer is seen increasing significantly through the year 2014
- Can we say that the *risk* of dying from cancer is increasing?
- If, for example, the size of the US population is also increasing at the same rate, then what happens to the risk of dying from cancer?

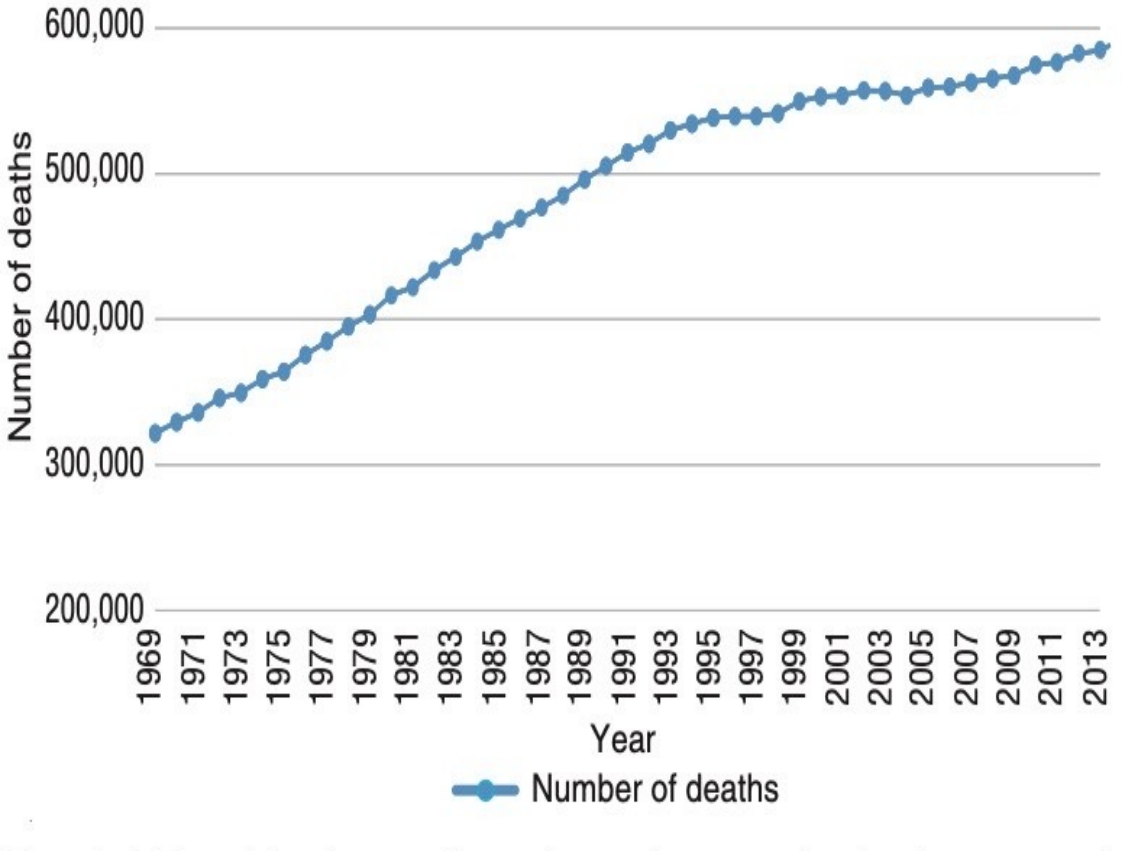


Fig 4.1 Trend in observed numbers of cancers deaths for in the United States 1969-2014.(Data from HK Anderson RN Coleman king SM, et al. (2016))

All Cause Mortality Rate

$$\frac{\text{Total no. of deaths from all causes in 1 year}}{\text{No. of persons in the population at midyear}} \times 100{,}000$$

We can have age-specific and cause-specific mortality rates

Case Fatality

Case-fatality (%) =

$$\frac{\text{No. of individuals dying during a specified period of time after disease onset or diagnosis}}{\text{No. of individuals with the specified disease}} \times 100$$

What is the difference between case-fatality and a mortality rate?

- Denominators
- Case-fatality is a measure of the *severity* of the disease.

Assume a population of 100,000 people of whom 20 are sick with disease X, and in 1 year, 18 of the 20 die from disease X

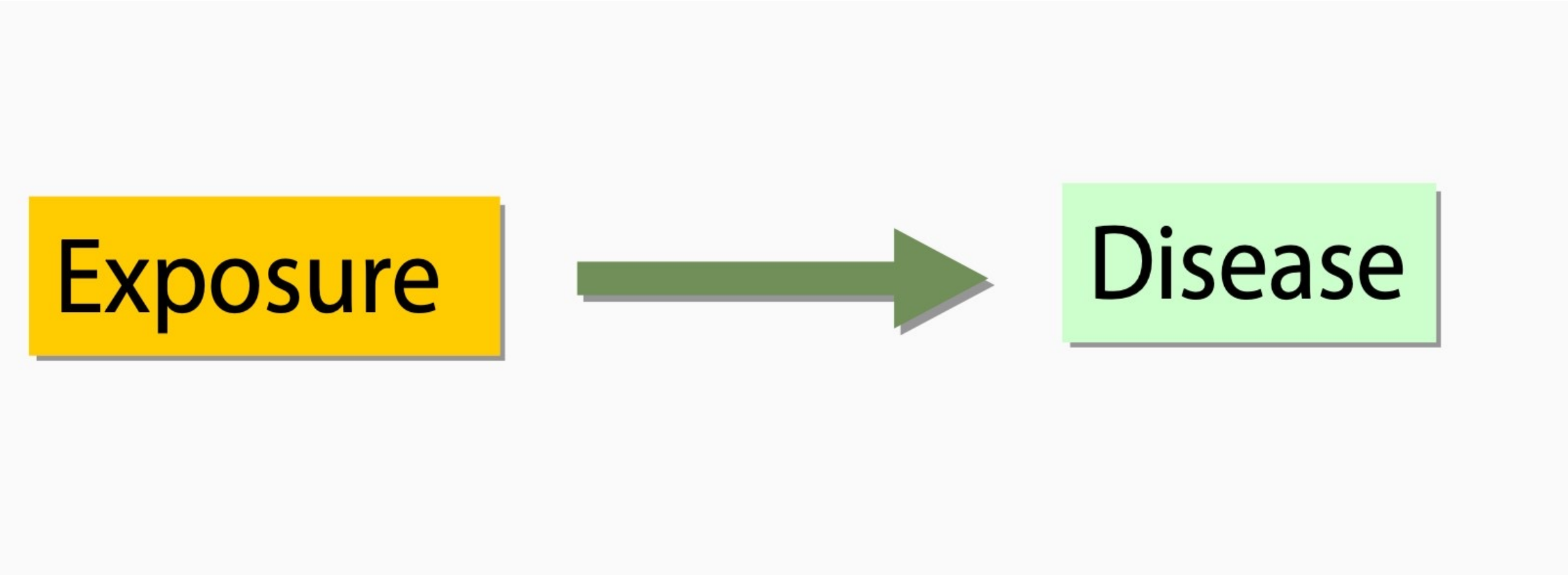
Mortality rate from disease $X = \frac{18}{100{,}000} = 0.00018$, or 0.018%

Case-fatality from disease $X = \frac{18}{20} = 0.9$, or 90%

Epidemiological Studies

- Cross-sectional studies
- Case-Control Studies
- Cohort Studies

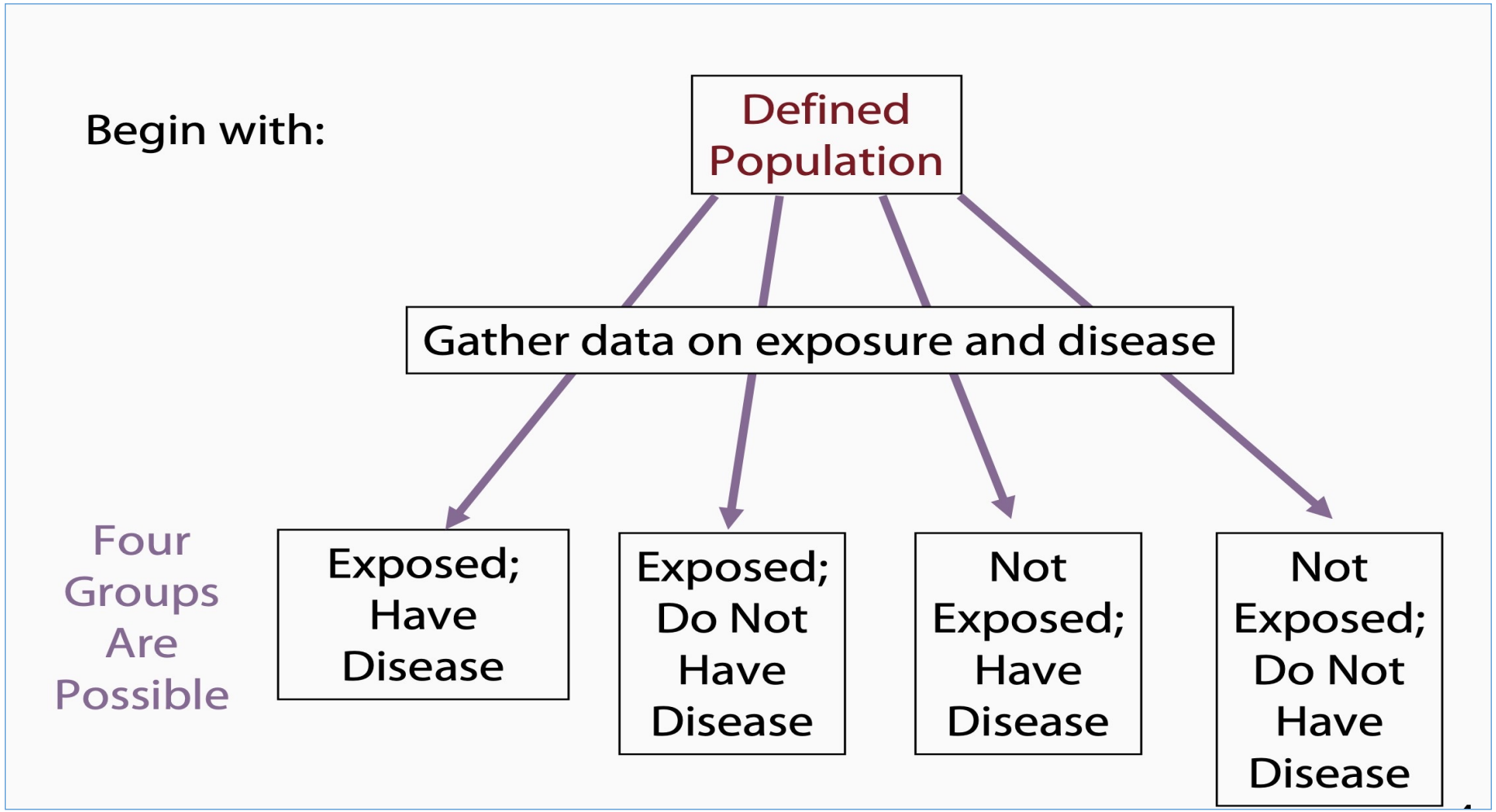
Primary Goals of Epidemiological Investigation



Cross sectional studies

- A cross-sectional study is an observational study in which exposure and disease are determined at the same point in time in a given population
- The temporal relationship between exposure and disease cannot be determined
- Example: Study the possible relationship of increased serum cholesterol level (exposure) to evidence of coronary heart disease (CHD, the disease).

Design of a Cross-Sectional Study

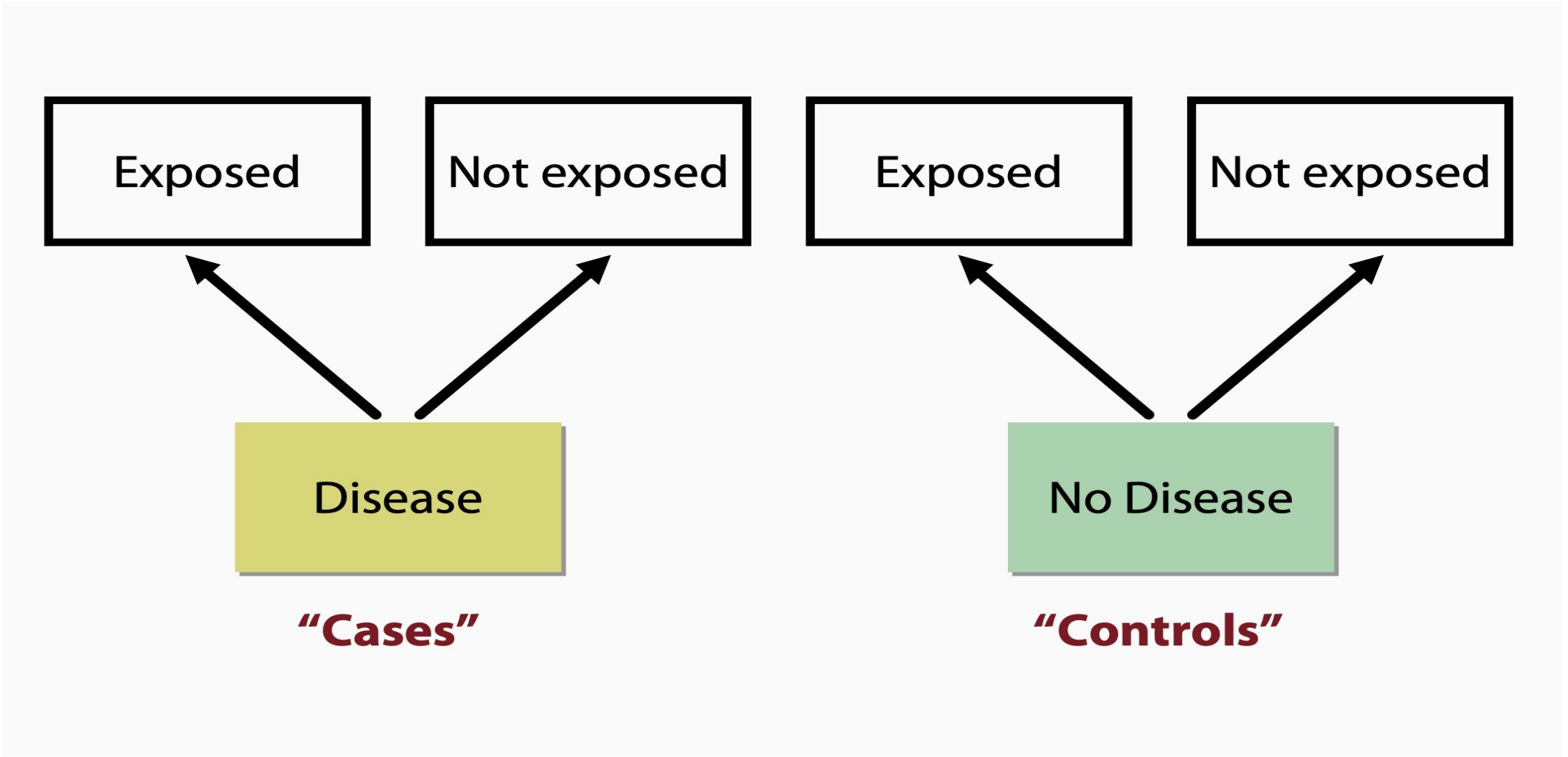


	Disease	No Disease
Exposed	a	b
Not Exposed	c	d

	Disease	No Disease
Exposed	a	b
Not Exposed	c	d

	Disease	No Disease
Exposed	a	b
Not Exposed	c	d

Designing a Case-Control Study

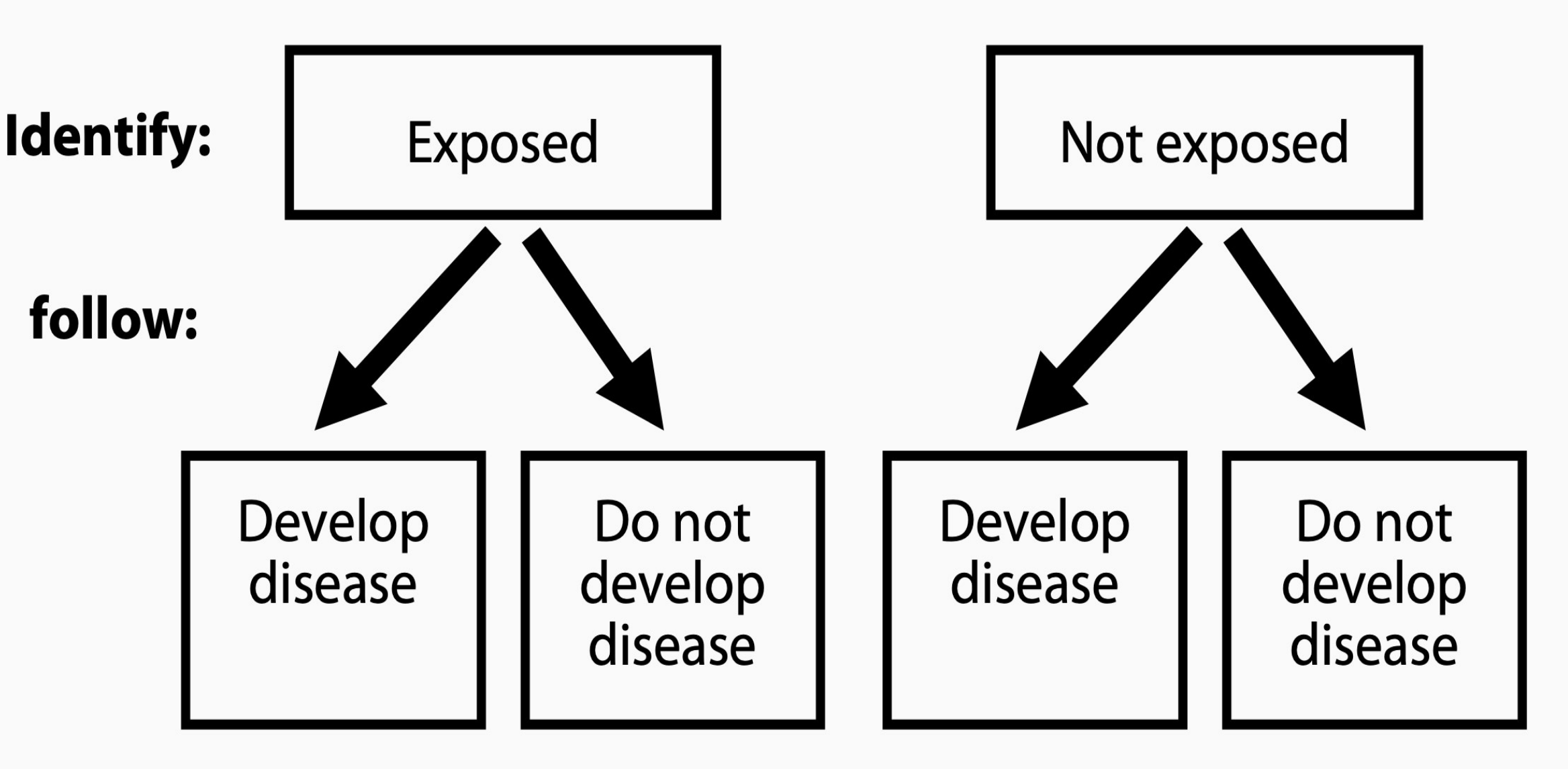


Example: Case-Control Study of Lung Cancer and Tobacco Smoking

		Lung Cancer	
		Case	Control
Tobacco Smoking	Yes	597	666
	No	8	114
		605	780

Source: Wynder and Graham.(1950).JAMA, 143:329-336.

Design of a Cohort Study



Then, follow to see whether

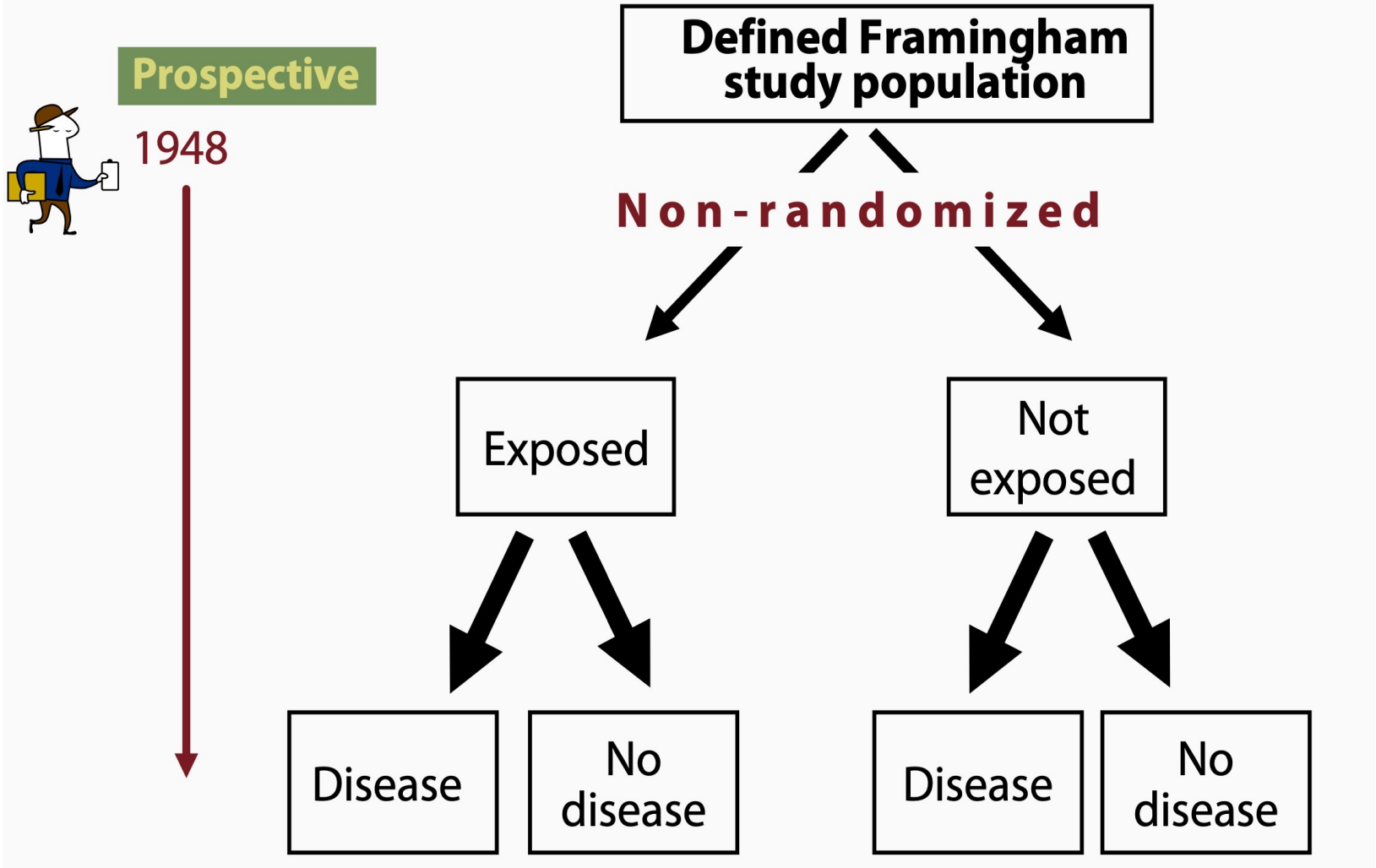
Calculate and compare

First, identify

	Disease develops	Disease does not develop	Totals	Incidence of disease
Exposed	a	b	a + b	$\frac{a}{a+b}$
Not exposed	c	d	c + d	$\frac{c}{c+d}$

$\frac{a}{a+b}$ = Incidence in exposed $\quad \frac{c}{c+d}$ = Incidence in not exposed

Framingham Study



Relative Risk or Risk Ratio

$$\text{Relative risk (RR)} = \frac{\text{Risk in exposed}}{\text{Risk in non-exposed}}$$

RR=1, risk in exposed equal to unexposed (no association)
RR>1, risk in exposed greater than unexposed (positive association)
RR<1, risk in exposed lesser than unexposed (negative association)

Cohort Study

Then follow to see whether

Calculate and compare

First, identify

	Disease develops	Disease does not develop	Totals	Incidence of disease
Exposed	a	b	a+b	$\frac{a}{a+b}$
Not exposed	c	d	c+d	$\frac{c}{c+d}$

$\frac{a}{a+b}$ = Incidence in exposed

$\frac{c}{c+d}$ = Incidence in not exposed

$$\text{Relative Risk} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Cohort Study

Then follow to see whether calculate

First select

	Develop CHD	Do not develop CHD	Totals	Incidence of disease
Smoke cigarettes	84	2916	3000	$\frac{84}{3000}$
Do not smoke cigarettes	87	4913	5000	$\frac{87}{5000}$

$$\text{Relative Risk} = \frac{\frac{84}{3000}}{\frac{87}{5000}} = \frac{28.0}{17.4} = 1.61$$

Odds Ratio (Relative Odds)

- "Odds" is often known as the ratio of money that may be won versus the amount of money bet
- In statistics, an **odds** of an event is the ratio of:
  - **The probability that the event WILL occur** *to* **the probability that the event will NOT occur**
    - For example, in 100 births, the probability of a delivery being a boy is 51% and being a girl is 49%
    - The odds of a delivery being a boy is **51/49 = 1.04**
- In simpler term, an odds of an event can be calculated as:
  - **Number of events divided by number of non-events**

Calculating Odds in a Cohort Study

	Develop Disease	Do Not Develop Disease
Exposed	a	b
Non-exposed	c	d

The odds that an exposed person develops disease $= \frac{a}{b}$

The odds that a non-exposed person develops disease $= \frac{c}{d}$

Calculating Odds in a Cohort Study

	Develop Disease	Do Not Develop Disease
Exposed	a	b
Non-exposed	c	d

Odds ratio is the ratio of the odds of disease in the exposed to the odds of disease in the non-exposed

$$OR = \frac{\text{odds that an exposed person develops the disease}}{\text{odds that a non - exposed person develops the disease}} = \frac{\frac{a}{b}}{\frac{c}{d}}$$

Calculating Odds Ratio in a Case-Control Study

	Case	Control
History of Exposure	a	b
No History of Exposure	c	d

The odds that a case was exposed $= \frac{a}{c}$

The odds that a control was exposed $= \frac{b}{d}$

Calculating Odds Ratio in a Case-Control Study

	Case	Control
History of Exposure	a	b
No History of Exposure	c	d

Odds ratio (OR) is the ratio of the odds that a case was exposed to the odds that a control was exposed

$$OR = \frac{\text{odds that a case was exposed}}{\text{odds that a control was exposed}} = \frac{\frac{a}{c}}{\frac{b}{d}}$$

Interpreting Odds Ratio of a Disease

- **If OR = 1**
  - Exposure is not related to disease
  - No association; independent
- **If OR > 1**
  - Exposure is positively related to disease
  - Positive association; ? causal
- **If OR < 1**
  - Exposure is negatively related to disease
  - Negative association; ? protective

Odds Ratio versus Relative Risk

- Odds ratio can be calculated in a cohort study and in a case-control study
  - The exposure odds ratio is equal to the disease odds ratio
- Relative risk can only be calculated in a cohort study

Random Variable and Distribution

- A **random variable** is a function that assigns numeric values to different events in a sample space.
- Two types of random variables: discrete and continuous.
- Otolaryngology Otitis media: Let *X* be the random variable that represents the number of episodes of otitis media in the first 2 years of life. Then *X* is a discrete random variable, which takes on the values 0, 1, 2, and so on.
- A **probability-mass function** is a mathematical relationship, or rule, that assigns to any possible value *r* of a discrete random variable *X* the probability $Pr(X = r)$. This assignment is made for all values *r* that have positive probability. The probability-mass function is sometimes also called a **probability distribution.**

Binomial Distribution

- The distribution of the number of successes in *n* statistically independent trials, where the probability of success on each trial is *p*, is known as the **binomial distribution** and has a probability-mass function given by

$$Pr(X=k)=\binom{n}{k}p^k q^{n-k}, \quad k=0,1,\ldots,n$$

What is the probability of obtaining 2 boys out of 5 children if the probability of a boy is .51 at each birth and the sexes of successive children are considered independent random variables?

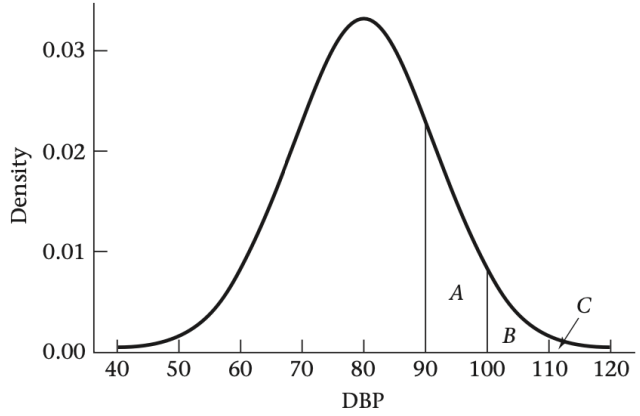
Use a binomial distribution with $n=5$, $p=.51$, $k=2$. Compute

$$Pr(X=2)=\binom{5}{2}(.51)^2(.49)^3=\frac{5\times4}{2\times1}(.51)^2(.49)^3$$

$$=10(.51)^2(.49)^3=.306$$

Continuous Random Variable

- The **probability-density function** of the random variable *X* is a function such that the area under the density-function curve between any two points *a* and *b* is equal to the probability that the random variable *X* falls between *a* and *b.* Thus, the total area under the density-function curve over the entire range of possible values for the random variable is 1.
- A pdf for DBP in 35- to 44-year-old men is shown below. Areas *A*, *B*, and *C* correspond to the probabilities of being mildly hypertensive, moderately hypertensive, and severely hypertensive, respectively. Furthermore, the most likely range of values for DBP occurs around 80 mm Hg, with the values becoming increasingly less likely as we move farther away from 80.



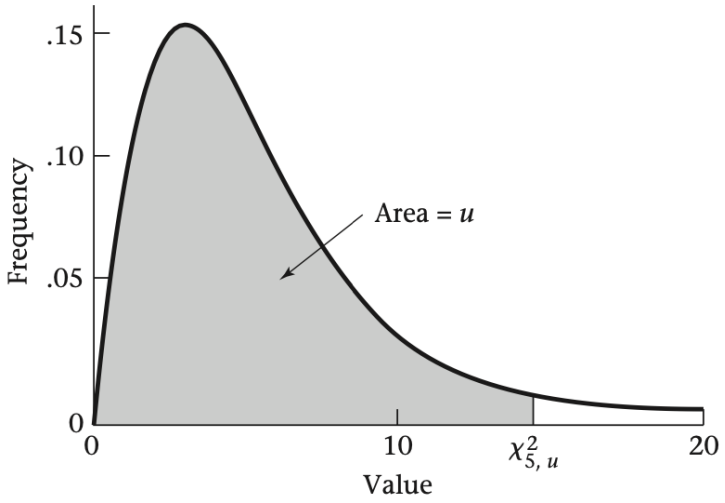
Chi-Square Distribution

If $G = \sum_{i=1}^{n} X_i^2$

where $X_1, \ldots, X_n \sim N(0,1)$

and the $X_i's$ are independent, then G is said to follow a **chi-square distribution with *n* degrees of freedom (*df*)**. The distribution is often denoted by χ_n^2.

Graphic display of the percentiles of a χ_5^2 distribution



Estimation

- Suppose we measure the systolic blood pressure (SBP) of a group of Indian villagers and we believe the underlying distribution is normal. How can the mean and variance of this distribution be estimated? How precise are our estimates?
- We can select a random sample and base our inferences on it.
- **Point Estimator**: A natural estimator to use for estimating the population mean is the sample mean
- Interval Estimator:

Confidence Interval for the Mean of a Normal Distribution

A 100% × (1 – α) CI for the mean μ of a normal distribution with unknown variance is given by

$$\left(\bar{x} - t_{n-1,1-\alpha/2}\, s/\sqrt{n}\,, \bar{x} + t_{n-1,1-\alpha/2}\, s/\sqrt{n}\right)$$

A shorthand notation for the CI is

$$\bar{x} \pm t_{n-1,1-\alpha/2}\, s/\sqrt{n}$$

Testing of Hypothesis

- Suppose the “average” cholesterol level in children is 175 mg/dL. A group of men who have died from heart disease within the past year are identified, and the cholesterol levels of their offspring are measured. Two hypotheses are considered:

The **null hypothesis,** denoted by H_0, is the hypothesis that is to be tested. The alternative hypothesis, denoted by H_1 is the hypothesis that in some sense contradicts the null hypothesis.

$$H_0: \mu = 175$$

$$H_1: \mu > 175$$

One-sample t-test

One-Samplet *t* Test for the Mean of a Normal Distribution with Unknown Variance (Alternative Mean < Null Mean)

To test the hypothesis $H_0: \mu = \mu_0, \sigma$ unknown vs. $H_1: \mu < \mu_0, \sigma$ unknown with a significance level of α, we compute

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If $t < t_{n-1,\alpha}$, then we reject H_0.

If $t \geq t_{n-1,\alpha}$, then we accept H_0.

***p*-value:** probability of obtaining a test statistic as extreme as or more extreme than the actual test statistic obtained, given that the null hypothesis is true.

Example: Longitudinal Study

(1) Identify a group of nonpregnant, premenopausal women of childbearing age (16–49 years) who are not currently OC users, and measure their blood pressure, which will be called the *baseline blood pressure*.

(2) Rescreen these women 1 year later to ascertain a subgroup who have remained nonpregnant throughout the year and have become OC users. This subgroup is the study population.

(3) Measure the blood pressure of the study population at the follow-up visit.

(4) Compare the baseline and follow-up blood pressure of the women in the study population to determine the difference between blood pressure levels of women when they *were* using the pill at follow-up and when they *were not* using the pill at baseline.

This is a **paired data** since each data point in the first sample is matched and is related to a unique data point in the second sample.

Paired t-test

SBP levels (mm Hg) in 10 women while not using (baseline) and while using (follow-up) OCs

i	SBP level while not using OCs (x_{i1})	SBP level while using OCs (x_{i2})	d_i^*
1	115	128	13
2	112	115	3
3	107	106	−1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	−2
10	115	117	2

*$d_i = x_{i2} - x_{i1}$

Assume that the SBP of the ith woman is normally distributed at baseline with mean μ_i and variance σ^2 and at follow-up with mean $\mu_i + \Delta$ and variance σ^2.

Paired *t* Test

Denote the test statistic $\bar{d}/\left(s_d/\sqrt{n}\right)$ by t, where s_d is the sample standard deviation of the observed differences:

$$s_d = \sqrt{\left[\sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2 \Big/ n\right] \Big/ (n-1)}$$

n = number of matched pairs

If $t > t_{n-1,1-\alpha/2}$ or $t < -t_{n-1,1-\alpha/2}$

then H_0 is rejected.

Two-sample t-test

- Suppose a sample of eight 35- to 39-year-old nonpregnant, premenopausal OC users who work in a company and have a mean systolic blood pressure of 132.86 mm Hg and sample standard deviation of 15.34 mm Hg are identified. A sample of 21 nonpregnant, premenopausal, non-OC users in the same age group are similarly identified who have mean SBP of 127.44 mm Hg and sample standard deviation of 18.23 mm Hg. What can be said about the underlying mean difference in blood pressure between the two groups?

Two-Sample *t* Test for Independent Samples with Equal Variances

Suppose we wish to test the hypothesis H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$ with a significance level of α for two normally distributed populations, where σ^2 is assumed to be the same for each population.

Compute the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s = \sqrt{\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right] / (n_1 + n_2 - 2)}$

If $t > t_{n_1+n_2-2,1-\alpha/2}$ or $t < -t_{n_1+n_2-2,1-\alpha/2}$

then H_0 is rejected.

Two-sample Binomial Test

- Suppose we are interested in the association between oral contraceptive (OC) use and the 5-year incidence of ovarian cancer from January 1, 2003 to January 1, 2008. Women who are disease-free on January 1, 2003 are classified into two OC-use categories as of that date: ever users and never users. We are interested in whether the 5-year incidence of ovarian cancer is different between ever users and never users.

Two-Sample Test for Binomial Proportions (Normal-Theory Test)

To test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$, where the proportions are obtained from two independent samples, use the following procedure:

(1) Compute the test statistic

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}, \hat{q} = 1 - \hat{p}$

and x_1, x_2 are the number of events in the first and second samples, respectively.

(2) For a two-sided level α test,

if $z > z_{1-\alpha/2}$

then reject H_0;

if $z \leq z_{1-\alpha/2}$

then accept H_0.

(3) The approximate p-value for this test is given by

$$p = 2[1 - \Phi(z)]$$

Example: Epidemiological Study

Cardiovascular Disease A study looked at the effects of OC use on heart disease in women 40 to 44 years of age. The researchers found that among 5000 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3-year period, whereas among 10,000 non-OC users, 7 developed an MI over a 3-year period. Assess the statistical significance of the results.

Note that $n_1 = 5000$, $\hat{p}_1 = 13/5000 = .0026$, $n_2 = 10{,}000$, $\hat{p}_2 = 7/10{,}000 = .0007$. We want to test the hypothesis H_0: $p_1 = p_2$ vs. H_1: $p_1 \neq p_2$. The best estimate of the common proportion p is given by

$$\hat{p} = \frac{13+7}{15{,}000} = \frac{20}{15{,}000} = .00133$$

Because $n_1\hat{p}\hat{q} = 5000(.00133)(.99867) = 6.7$, $n_2\hat{p}\hat{q} = 10{,}000(.00133)(.99867) = 13.3$, the normal-theory test in Equation 10.3 can be used. The test statistic is given by

$$z = \frac{|.0026 - .0007| - \left[\frac{1}{2(5000)} + \frac{1}{2(10{,}000)}\right]}{\sqrt{.00133(.99867)(1/5000 + 1/10{,}000)}} = \frac{.00175}{.00063} = 2.77$$

The p-value is given by $2 \times [1 - \Phi(2.77)] = .006$. Thus there is a highly significant difference between MI incidence rates for current OC users vs. non-OC users. In other words, OC use is significantly associated with MI incidence over a 3-year period.

Contingency Tables

	MI status over 3 years		
OC-use group	Yes	No	Total
OC users	13	4987	5000
Non-OC users	7	9993	10,000
Total	20	14,980	15,000

Chi-Square Test Statistic

- The test statistic is:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

- The degrees of freedom are:
  - (r–1)(c–1)
  - r = # of rows and c = # of columns
- Where:
  - O_i = the observed frequency in the i^{th} cell of the table
  - E_i = the expected frequency in the i^{th} cell of the table

Guidelines for Interpreting the X^2 Statistic

- The χ^2 statistic is calculated under the assumption of no association

- **Large value of χ^2 statistic** $\Rightarrow$ small probability of occurring by chance alone ($p < 0.05$) $\Rightarrow$ conclude that **association** exists between disease and exposure

- **Small value of χ^2 statistic** $\Rightarrow$ large probability of occurring by chance alone ($p > 0.05$) $\Rightarrow$ conclude that **no association** exists between disease and exposure

Example: Test of Association

	MI status over 3 years		
OC-use group	Yes	No	Total
Current OC users	6.7	4993.3	5000
Non-OC users	13.3	9986.7	10,000
Total	20	14,980	15,000

The test statistic is 7.67. From chi-square tables we see that the result implies a significant difference between the OC and non-OC groups.

Thank you