Biostatistics - STAT-241

Weeks 1-2

Descriptive methods for categorical data

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics

Qatar University

Learning Objectives

- Basic concepts in statistics and probability.
- Compute and interpret the proportions.
- Comparative studies in-control study and screening tests.
- Compute different types of rates and changes.
- Rates: Measures for morbidity and mortality.
- Differences between proportions and rates.
- Ratios, relative risk, odds ratio, Mantel-Haenszel and standardized mortality ratio.

What is Biostatistics

- **Biostatistics** is a sub topic of statistics using the applied statistical methodologies to a wide range of topics in biology.

- **Biostatistics** implements some basic methods of statistics in order to answer questions that arise in biomedical and health sciences, especially in medicine, pharmacy, agriculture and fishery.

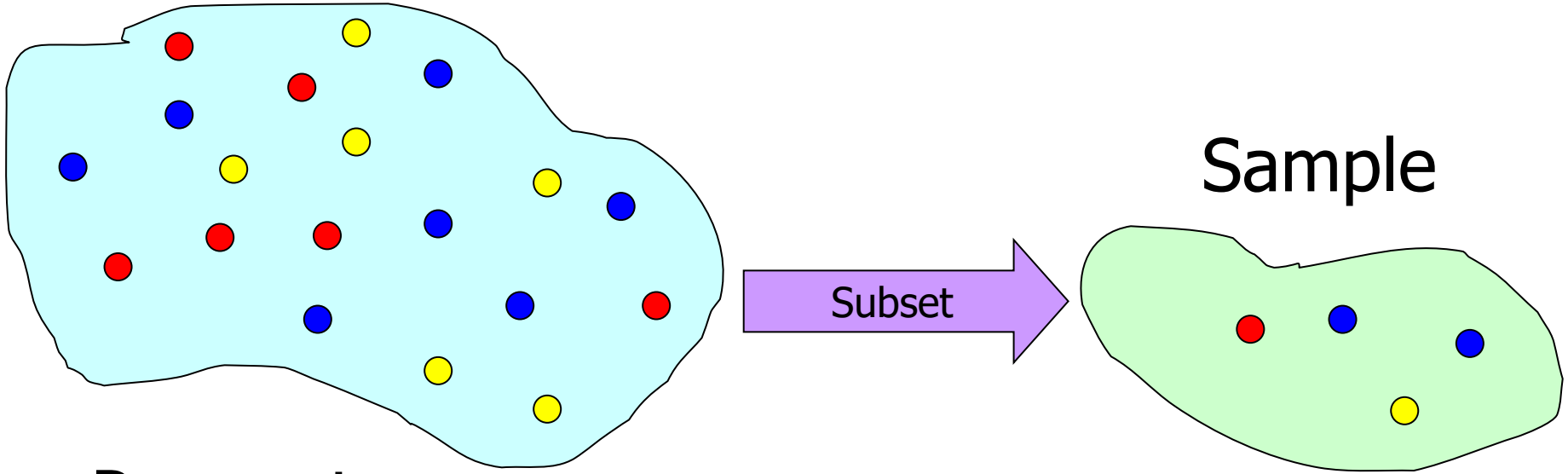
- **Biostatistics** focus on data related to biology while **statistics** tend to be more general. However, the underlying theory are similar.

Variables and Data

- **A variable** is a characteristic that changes or varies over time and/or for different individuals or objects under consideration.
- **Examples:** Hair color, white blood cell count, time to failure of a computer component.
- **An experimental unit** is the individual or object on which a variable is measured.
- **A measurement** results when a variable is actually measured on an experimental unit.
- A set of measurements, called **data**, can be either a **sample** or a **population.**
- The term **Statistics** means a summarized figure from observed data.

Key Statistical Concepts

Population



Sample

Parameter

Statistic

Populations have Parameters
(usually unknown)

Samples have Statistics
(always known)

Parameters and Statistics

	Population	Sample
Size	N	n
Mean	μ (mu)	$\bar{x}$ (x-bar)
Variance	σ^2 (sigma squared)	S^2 (s squared)
Standard Deviation	σ	S
Proportion	π (pi)	P
Covariance	σ_{xy}	S_{xy}
Coefficient of Correlation	ρ (rho)	r

Definitions

- Many outcomes can be classified in two outcome categories as positive (+) and negative (-).
- An outcome is positive if the primary category under investigation (e.g., people suffered from diabetes) is observed and is negative if the other category is observed (e.g., people do not suffered from diabetes).
- **Proportion** is a number, p, used to describe a group of people according to a dichotomous (binary) characteristic under investigation.

$$p = \frac{x}{n} = \frac{number\ of\ positive\ outcome}{the\ group\ size}$$

- This proportion p is sometimes expressed as a **percentage** and is calculated as follows:

$$\text{percent } (\%) = \frac{x}{n}(100)$$

Disease prevalence $= \dfrac{\textit{number of diseased persons at the time of investigation}}{\textit{total number of persons examined}}$

Example 1.1

Example 1.1 A study published by the Urban Coalition of Minneapolis and the University of Minnesota Adolescent Health Program surveyed 12,915 students in grades 7 through 12 in Minneapolis and St. Paul public schools. The report stated that minority students, about one-third of the group, were much less likely to have had a recent routine physical checkup. Among Asian students, 25.4% said that they had not seen a doctor or a dentist in the last two years, followed by 17.7% of Native Americans, 16.1% of blacks, and 10% of Hispanics. Among whites, it was 6.5%.

Comparative Studies

- Comparative studies are intended to show possible differences between two or more groups.
- **Cross-sectional studies** describe the relationship between diseases and other factors at one point in time in a defined population.
  - The survey cited in Example 1.1 also provided the following figures concerning boys in the group who use tobacco at least weekly. Among Asians, it was 9.7%, followed by 11.6% of blacks, 20.6% of Hispanics, 25.4% of whites, and 38.3% of Native Americans.
- **Case-control studies**
  - **Cases** of a specific disease are ascertained as they arise from population-based registers or lists of hospital admissions,
  - **Controls** are sampled either as disease-free persons from the population at risk or as hospitalized patients having a diagnosis other than the one under study.
  - Data for comparative studies may come from different sources; the two fundamental designs being retrospective and prospective.
  - **Retrospective studies** gather past data from selected cases and controls to determine differences, if any, in exposure to a suspected risk factor.
  - A **prospective studies** watch for outcomes, such as the development of a disease, during the study period and relates this to other risk or protection factor(s).

Example 1.2

- Example 1.2 A case-control study was undertaken to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia. Cases were identified from these sources:
  a) Diagnoses since 1970 at the single large hospital in Brunswick.
  b) Diagnoses during 1975-1976 at three major hospitals in Savannah.
  c) Death certificates for the period 1970-1974 in the area.

- The ***exposure*** under investigation, **"shipbuilding"**, refers to employment in shipyards during World War II is a suspected ***risk factor***.
- A *confounder* is a factor (in this case, smoking), an exposure by itself, not under investigation but related to the disease (lung cancer) and the exposure (shipbuilding).
- Previous studies have linked smoking to lung cancer, and construction workers are more likely to be smokers.

Example 1.2 (Cont.)

TABLE 1.1

Smoking	Shipbuilding	Cases	Controls
No	Yes	11	35
	No	50	203
Yes	Yes	84	45
	No	313	270

- **Smokers**

1. For the controls,

$$\text{proportion of exposure} = \frac{45}{315} = 0.143 \quad \text{or} \quad 14.3\%$$

2. For the cases,

$$\text{proportion of exposure} = \frac{84}{397} = 0.212 \quad \text{or} \quad 21.2\%$$

- **Nonsmokers**

1. For the controls,

$$\text{proportion of exposure} = \frac{35}{238} = 0.147 \quad \text{or} \quad 14.7\%$$

2. For the cases,

$$\text{proportion of exposure} = \frac{11}{61} = 0.180 \quad \text{or} \quad 18.0\%$$

- The results reveal different exposure histories:
  - The proportion among cases was higher than that among controls.
  - It is not in any way conclusive proof, but it is a good clue, indicating a possible relationship between the disease (lung cancer) and the exposure (shipbuilding).

- The results also reveal different exposure histories:
  - The proportion among cases was higher than that among controls

Example 1.2 (Cont.)

- The analyses above also show that the difference between proportions of exposure among smokers, that is,

$$21.2 - 14.3 = 6.9\%$$

  is different from the difference between proportions of exposure among nonsmokers, which is

$$18.0 - 14.7 = 3.3\%$$

- The differences, 6.9% and 3.3%, are measures of the strength of the relationship between the disease and the exposure, one for each of the two strata: the two groups of smokers and nonsmokers, respectively.
- The calculation above shows that the possible effects of employment in shipyards (as a suspected risk factor) are different for smokers and nonsmokers.
- This difference of differences, if confirmed, is called *a three-term interaction* or *effect modification*, where smoking alters the effect of employment in shipyards as a risk for lung cancer.
- In that case, smoking is not only a confounder, it is an *effect modifier*, which modifies the effects of shipbuilding (on the possibility of having lung cancer).

Example 1.3

- Data for persons registered blind from glaucoma are listed in Table 1.2.

TABLE 1.2

	Population	Cases	Cases per 100,000
White	32,930,233	2832	8.6
Nonwhite	3,933,333	3227	82.0

- Disease prevalence $= \dfrac{\text{number of diseased persons at the time of investigation}}{\text{total number of persons examined}}$
- The number ‘‘100,000’’ was selected arbitrarily; any power of 10 would be suitable so as to obtain a result between 1 and 100,
- sometimes between 1 and 1000; it is easier to state the result ‘‘82 cases per 100,000’’ than to say that the prevalence is 0.00082.

Screening Tests

- **Screening Tests** are diagnostic procedures (clinical observations, or laboratory techniques). Following these procedures, people are classified as healthy or as falling into one of a number of disease categories.
- Almost all such tests are imperfect.

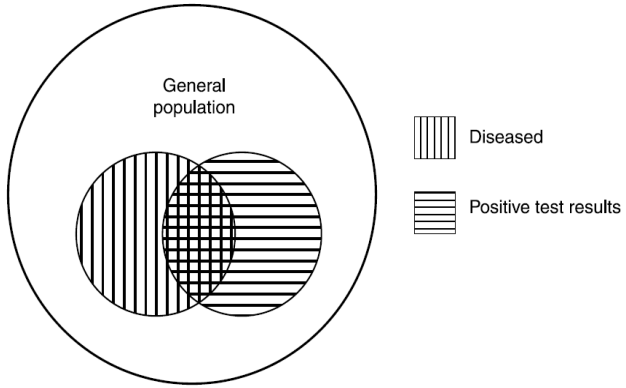


Figure 1.1 Graphical display of a screening test.

- **Sensitivity** is the proportion of diseased people detected as positive by the test:

$$\text{Sensitivity} = \frac{\text{number of diseased persons who screen positive}}{\text{total number of diseased person}}$$

  - The corresponding errors are *false negatives*.

- **Specificity** is the proportion of healthy people detected as negative by the test:

$$\text{Specificity} = \frac{\text{number of healthy persons who screen negative}}{\text{total number of healthy persons}}$$

  - The corresponding errors are *false positives*.

Example 1.4

- A cytological test was undertaken to screen women for cervical cancer. Consider a group of 24,103 women consisting of 379 women whose cervices are abnormal (to an extent sufficient to justify concern with respect to possible cancer) and 23,724 women whose cervices are acceptably healthy. A test was applied and results are tabulated in Table 1.3. (This study was performed with a rather old test and is used here only for illustration.)

TABLE 1.3

	Test		
True	−	+	Total
−	23,362	362	23,724
+	225	154	379

- The calculations

$$\text{sensitivity} = \frac{154}{379} = 0.406 \quad \text{or} \quad 40.6\%$$

$$\text{specificity} = \frac{23,362}{23,724} = 0.985 \quad \text{or} \quad 98.5\%$$

show that the test is highly specific (98.5%) but not very sensitive (40.6%); there were more than half (59.4%) false negatives.

- The implications of the use of this test are:
  - If a woman without cervical cancer is tested, the result would almost surely be negative.
  - If a woman with cervical cancer is tested, the chance is that the disease would go undetected because 59.4% of these cases would lead to false negatives.

Summary

- A test result is a **true positive** if it is positive and the individual has the disease.
- A test result is a **true negative** if it is negative and the individual does not have the disease.
- A test result is a **false positive** if it is positive and the individual does not have the disease.
- A test result is a **false negative** if it is negative and the individual does have the disease.
- The **sensitivity** of a test is the conditional probability that the test is positive, given that the individual has the disease.
- The **specificity** of a test is the conditional probability that the test is negative, given that the individual does not have the disease.
- The **predictive value of a positive test** is the conditional probability that an individual has the disease, given that the test is positive.
- The **predictive value of a negative test** is the conditional probability that an individual does not have the disease, given that the test is negative.

Displaying Proportions

- Bar Charts
- Pie Charts
- Line Graphs
- We will see more later!

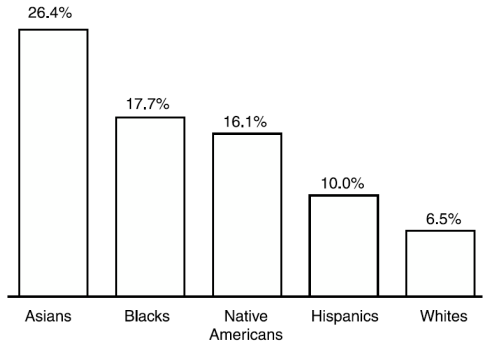


Figure 1.2 Children without a recent physical checkup.

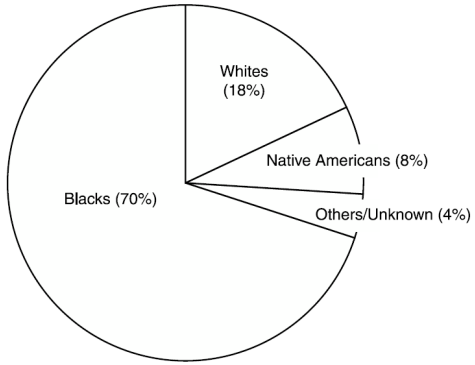


Figure 1.3 Children living in crack households.

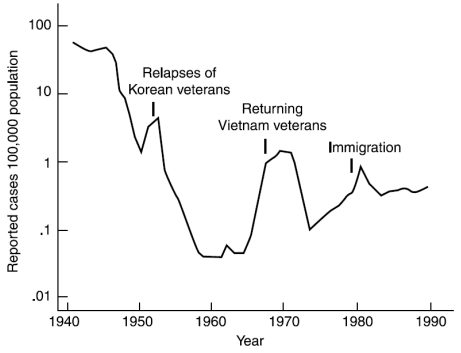


Figure 1.6 Malaria rates in the United States, 1940–1989.

TABLE 1.5

Year	Crude Death Rate per 100,000
1984	792.7
1985	806.6
1986	809.3
1987	813.1

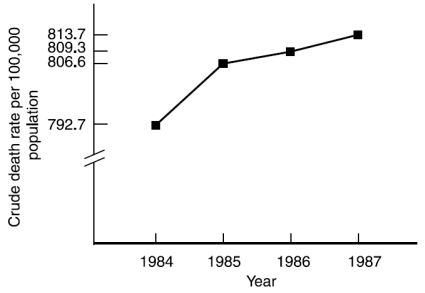
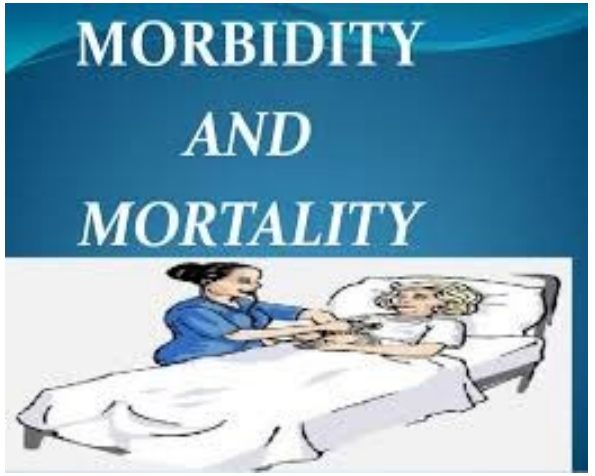


Figure 1.5 Death rates for U.S. women, 1984–1987.

Rates

- **Rates** used interchangeably with proportions as measures of morbidity and mortality.
- In contrast to the static nature of proportions, rates are aimed at measuring the occurrences of events during or after a certain time period.
- **Change Rate** is used primarily for description and are not involved in common statistical analysis. It is defined by

$$\text{change rate } (\%) = \frac{\text{new value} - \text{old value}}{\text{old value}} \times 100$$



- In general, change rates could exceed 100%. They are not proportions (a proportion is a number between 0 and 1 or between 0 and 100% in the case of percentage).
- **Note:** In many cases proportions and rates are defined very similarly, and the terms proportions and rates may even be used interchangeably.

Example 1.11

- A total of 35,238 new AIDS cases was reported in 1989 by the Centers for Disease Control (CDC), compared to 32,196 reported during 1988. The 9.4% increase is the smallest since the spread of AIDS began in the early 1980s. For example, new AIDS cases were up 34% in 1988 and 60% in 1987. In 1989, 547 cases of AIDS transmissions from mothers to newborns were reported, up 17% from 1988; while females made up just 3971 of the 35,238 new cases reported in 1989; that was an increase of 11% over 1988.

1. Find the change rate for new AIDS?
2. Find the new AIDS transmitted from mothers to newborns?
3. In 1989 new AIDS cases, find proportions of females and males?

1. The change rate for new AIDS cases was calculated as

$$\frac{35{,}238 - 32{,}196}{32{,}196} \times 100 = 9.4\%$$

2. For the new AIDS cases transmitted from mothers to newborns,

$$17\% = \frac{547 - (1988\ \text{cases})}{1988\ \text{cases}} \times 100$$

leading to

$$1988\ \text{cases} = \frac{547}{1.17} = 468$$

1. Among the 1989 new AIDS cases, the proportion of females is

$$\frac{3971}{35{,}238} = 0.113 \text{ or } 11.3\%$$

and the proportion of males is

$$\frac{35{,}238 - 3971}{35{,}238} = 0.887 \text{ or } 88.7\%$$

The proportions of females and males add up to 1.0 or 100%.

Measures of Morbidity and Mortality

- **Mortality rate** is a measure of the number of deaths in a particular population, scaled to the size of that population, per unit of time.
- **Morbidity rate** is either the prevalence or incidence of a disease per unit of time.
- There are three common measures of mortality rate: crude, specific, and adjusted (or standardized).
- **Crude death rates (CDR):** is defined as the number of deaths in a calendar year divided by the population on July 1 of that year.

$$\text{CDR} = \frac{\text{d}}{\text{n}} = \frac{\text{Number of deaths during the year}}{\text{The size of the population during the year}} \times 1000$$

- **Example:** The 1980 population of California was 23,000,000 (as estimated by July 1) and there were 190,237 deaths during 1980, leading to

$$\text{CDR} = \frac{190{,}237}{23{,}000{,}000} \times 1000 = 8.3 \text{ deaths per 1000 persons per year}$$

- **Specific death rates** consider these differences among subgroups or categories of diseases.
- **Adjusted (standardized) death rates:** Used to make valid summary comparisons between two or more groups possessing different age distributions (see slide 24).
  - Risks of death change by age, so CDR is affected by population age structure.
  - Aging populations can have rising CDRs, even as the health conditions are improving.

Measures of Morbidity and Mortality

- As for morbidity, the disease prevalence, as defined in Section 1.1, is the actual number (proportion) of cases within a period of time, whereas incidence is a rate used in connection with new cases of the disease occurring within a period of time

$$\text{incidence rate} = \frac{\text{number of persons who developed the disease over a defined period of time (a year; say)}}{\text{number of persons initially without the disease who were followed for the defined period of time}}$$

- In other words, the prevalence presents a snapshot of the population's morbidity widespread at a certain time point, whereas the incidence is aimed to investigate possible time trends.

For example, consider a sample of 450 people, and want to determine the incidence rate of developing HIV over a 5-year period:

  - At the beginning of the study we find 50 cases of existing HIV. These people are not counted as they cannot develop HIV a second time.
  - A follow-up at 5 years we find 40 new cases of HIV.

- The prevalence is $\frac{50+40}{450} \times 100 = 20\%$.
- The incidence after 5 years is $\frac{40}{450} \times 100 = 8.9\%$.

Measures of Morbidity and Mortality

- Another interesting use of rates is in connection with **cohort studies.**
- **The cohort study design** focuses on a particular exposure rather than a particular disease as in case-control studies.
- Each member of a cohort belongs to one of three types of termination:

1. Subjects still alive on the analysis date.
2. Subjects who died on a known date within the study period.
3. Subjects who are lost to **follow-up** after a certain date (these cases are a potential source of bias; effort should be expended on reducing the number of subjects in this category)

- The contribution of each member is the length of follow-up time from enrollment to his or her termination. The quotient, defined as the number of deaths observed for the cohort, divided by the total follow-up times (in person years, say) is the rate to characterize the mortality experience of the cohort:

$$\text{follow-up death rate} = \frac{\text{number of deaths in cohort study}}{\text{total person-years}}$$

- Rates may be calculated for total deaths and for separate causes of interest.
- Follow-up death rates may be used to measure the effectiveness of medical treatment programs.

Standardization of Rates

- Crude rates, as measures of morbidity or mortality, can be used for population description and may be suitable for investigations of their variations over time; however, comparisons of crude rates are often invalid because the populations may be different with respect to an important characteristic such as age, gender, or race (these are potential confounders).
- To overcome this difficulty, an adjusted (or standardized) rate is used in the comparison; the adjustment removes the difference in composition with respect to a confounder.
- Age-standardized rates are constructed by applying the rates for each age group to a standard population.

- Infant mortality rate, the number of infant deaths in a year divided by the number of live births in the year is an important indicator.
- Child mortality: the probability of a child dying before the age of 5- is also used as a proxy measure of the health status of the population.

Ratio

- In many cases, such as disease prevalence and disease incidence, proportions and rates are defined very similarly, and the terms proportions and rates may even be used interchangeably.
- A **Ratio** is positive but may exceed 1. It is a computation of the form:

$$ratio = \frac{a}{b}$$

where a and b are similar quantities measured from different groups or under different circumstances.

- **Example:** The $\frac{male}{female}$ ratio of smoking rates; such a ratio is positive but may exceed 1.

Relative Risk

- **Relative Risk** is a concept, most often used in epidemiological studies used to measure seriousness, or the magnitude of the harmful effect of suspected risk factors.
- Relative Risk used for the comparison of two groups or populations with respect to a certain unwanted event (e.g., disease or death).

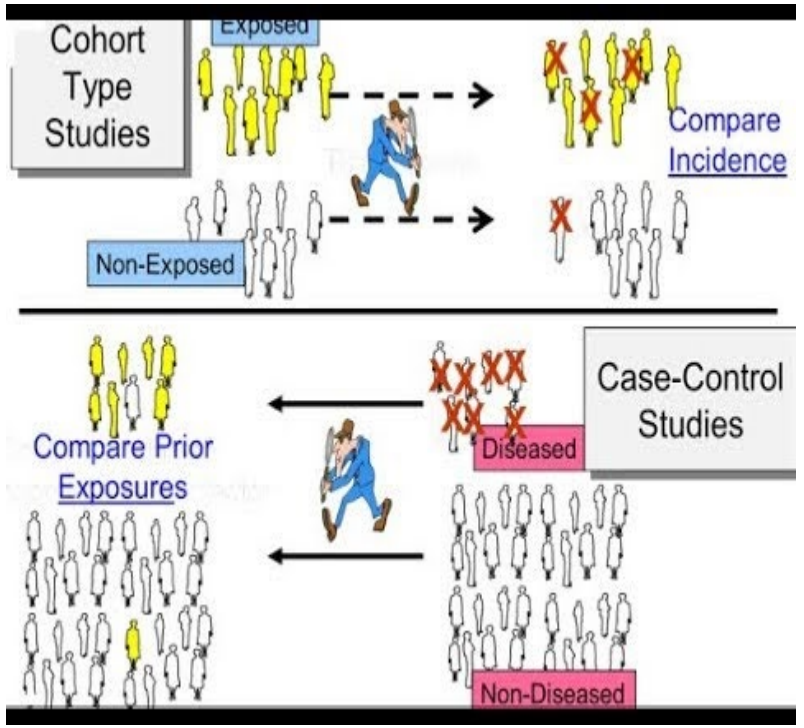
$$relative\ risk = RR = \frac{disease\ incidence\ in\ group\ 1}{disease\ incidence\ in\ group\ 2} = \frac{probability\ \ that\ an\ exposed\ gets\ disease}{probability\ \ that\ an\ unexposed\ gets\ disease}$$

- Usually, group 2 is under standard conditions (such as nonexposure to a certain risk factor) against which group 1 (exposed) is measured.
  - A relative risk equals 1.0 indicates no effects.
  - A relative risk greater than 1.0 indicates harmful effects.
  - A relative risk below 1.0 indicates beneficial effects.

- Example: If group 1 consists of smokers people and group 2 of nonsmokers, where RR = 3, we have a relative risk due to smoking. That is, smokers have a risk of contracting the disease that is approximately three times the risk of those nonsmokers. Note that, in this example, smoking is the suspected risk factor.

Odds and Odds Ratio

- **In Cohort studies:** The relative risk, also called the risk ratio, is an important index in epidemiological studies because in such studies it is often useful to measure the increased risk (if any) of incurring a particular disease if a certain risk factor is present.
- **In a case-control studies:** the data do not present an immediate answer to this type of question (because the case subjects already have the disease), and we now consider how to obtain a useful shortcut solution.
- **Odds ratios:** are the only valid measure of relative risk for case-control studies.
- **Odds ratio:** is the percent increase or decrease in the odds of the outcome.
- Odds ratios = $\frac{p}{1-p}$ are calculated from logistic regression as we will see later.
- $$\text{Odds ratios} = \frac{\text{odds that a case } (with\ a\ disease) \text{ was exposed}}{\text{odds that a control } (without\ a\ disease) \text{ was exposed}}$$



Risk versus Odds

If the risk = p/p	The odds = p/(1-p)
1/2 (50%)	1 (1:1)
6/10 (60%)	1.5 (3:2)
3/4 (75%)	3 (3:1)
1/10 (10%)	0.1 (1:9)

- Odds ratios can be interpreted as risk ratios for **rare** outcomes.
- Rule of thumb for defining **"rare"**: outcome occurs in <10% of the reference/control group.
- But, when the outcome is **common**, odds ratio distort the effect size and need to be interpreted cautiously.

TABLE 1.11

	Disease		
Factor	+	−	Total
+	A	B	$A+B$
−	C	D	$C+D$
Total	$A+C$	$B+D$	$N=A+B+C+D$

- The entries A, B, C and D in the table are sizes of the four combinations of disease presence/ absence and factor presence/absence, and the number N at the lower right corner of the table is the total population size.
- The Relative Risk (RR) is $RR = \frac{A}{A+B} \div \frac{C}{C+D} = \frac{A(C+D)}{C(A+B)}$
- Note that $C + D \approx D$ and $A + B \approx B$. Thus the approximated Relative Risk (RR) is $RR = \frac{A/C}{B/D} = \frac{AD}{BC}$.
- The resulting ratio, $\frac{AD}{BC}$, is an approximate relative risk, but it is often referred to as an odds ratio.

✓ A/B and C/D are the odds in favor of having disease from groups with or without the factor.

✓ A/C and B/D are the odds in favor of having been exposed to the factors from groups with or without the disease.

Example 1.14

- The role of smoking in the etiology of pancreatitis has been recognized for many years. To provide estimates of the quantitative significance of these factors, a hospital-based study was carried out in eastern Massachusetts and Rhode Island between 1975 and 1979.
- Ninety-eight patients who had a hospital discharge diagnosis of pancreatitis were included in this unmatched case-control study.
- The control group consisted of 217 patients admitted for diseases other than those of the pancreas and biliary tract.
- Risk factor information was obtained from a standardized interview with each subject, conducted by a trained interviewer.
- Some data for the males are given in Table 1.12.

Example 1.14 (Cont.)

- The approximate relative risks or odds ratios are calculated as follows:
  - **(RR_e) for ex-smokers:**

$$RR_e \simeq \frac{13/2}{80/56} = \frac{(13)(56)}{(80)(2)} = 4.55$$

  - **(RR_c) for current smokers:**

$$RR_c \simeq \frac{38/2}{81/56} = \frac{(38)(56)}{(81)(2)} = 13.14$$

TABLE 1.12

Use of Cigarettes	Cases	Controls
Never	2	56
Ex-smokers	13	80
Current smokers	38	81
Total	53	217

- In these calculations, the nonsmokers (who never smoke) are used as references.
- These values indicate that the risk of having pancreatitis for current smokers is approximately 13.14 times the risk for people who never smoke.
- The effect for ex-smokers is smaller (4.55 times) but is still very high (compared to 1.0, the no-effect baseline for relative risks and odds ratios).

Generalized Odds for Ordered $2 \times k$ Tables

- **Example:** Let us consider an example concerning the use of seat belts in automobiles. Each accident in this example is classified according to whether a seat belt was used and to the severity of injuries received: none, minor, major, or death (Table 1.13).

TABLE 1.13

	Extent of Injury Received			
Seat Belt	None	Minor	Major	Death
Yes	75	160	100	15
No	65	175	135	25

- To compare the extent of injury from those who used seat belts with those who did not, we can calculate the percent of seat belt users in each injury group that decreases from level ‘‘none’’ to level ‘‘death,’’ and the results are:

None:	$\frac{75}{75+65} = 54\%$	Minor:	$\frac{160}{160+175} = 48\%$
Major:	$\frac{100}{100+135} = 43\%$	Death:	$\frac{15}{15+25} = 38\%$

- What we are seeing here is a trend or an association indicating that the lower the percentage of seat belt users, the more severe the injury.

Generalized Odds for Ordered $2 \times k$ Tables

- We now present the concept of generalized odds, a special statistic specifically formulated to measure the strength of such a trend.
- In general, consider an ordered $2 \times k$ table with the frequencies shown in Table 1.14.

TABLE 1.14

	Column Level				
Row	1	2	$\cdots$	k	Total
1	a_1	a_2	$\cdots$	a_k	A
2	b_1	b_2	$\cdots$	b_k	B
Total	n_1	n_2	$\cdots$	n_k	N

- The number of concordances is calculated by

$$C = a_1(b_2 + \cdots + b_k) + a_2(b_3 + \cdots + b_k) + \cdots + a_{k-1}b_k$$

- (The term concordance pair corresponds to a less severe injury for the seat belt user.)
- The number of discordances is

$$D = b_1(a_2 + \cdots + a_k) + b_2(a_3 + \cdots + a_k) + \cdots + b_{k-1}a_k$$

- To measure the degree of association, we use the index $\theta = \mathrm{C}/\mathrm{D}$ and call it the generalized odds; if there are only two levels of injury, this new index is reduced to the familiar odds ratio.

Example 1.15

- For the study above on the use of seat belts in automobiles, we have from the data shown in Table 1.13,

$$C = (75)(175 + 135 + 25) + (160)(135 + 25) + (100)(25)$$
$$= 53{,}225$$
$$D = (65)(160 + 100 + 15) + (175)(100 + 15) + (135)(15)$$
$$= 40{,}025$$

leading to generalized odds of

$$\theta = \frac{C}{D} = \frac{53{,}225}{40{,}025} = 1.33$$

TABLE 1.13

	Extent of Injury Received			
Seat Belt	None	Minor	Major	Death
Yes	75	160	100	15
No	65	175	135	25

- That is, given two people with different levels of injury, the (generalized) odds that the more severely injured person did not wear a seat belt is 1.33. In other words, the people with the more severe injuries would be more likely than the people with less severe injuries to be those who did not use a seat belt.

Example 1.16

- **Example 1.16** A case-control study of the epidemiology of preterm delivery, defined as one with less than 37 weeks of gestation, was undertaken at Yale-New Haven Hospital in Connecticut during 1977. The study population consisted of 175 mothers of singleton preterm infants and 303 mothers of singleton full-term infants.

- Table 1.15 gives the distribution of mother's age. We have

$$
\begin{aligned}
C &= (15)(25+62+122+78) + (22)(62+122+78) \\
&\quad + (47)(122+78) + (56)(78) \\
&= 23{,}837 \\
D &= (16)(22+47+56+35) + (25)(47+56+35) \\
&\quad + (62)(56+35) + (122)(35) \\
&= 15{,}922
\end{aligned}
$$

leading to generalized odds of

$$\theta = \frac{C}{D} = \frac{23{,}837}{15{,}922} = 1.50$$

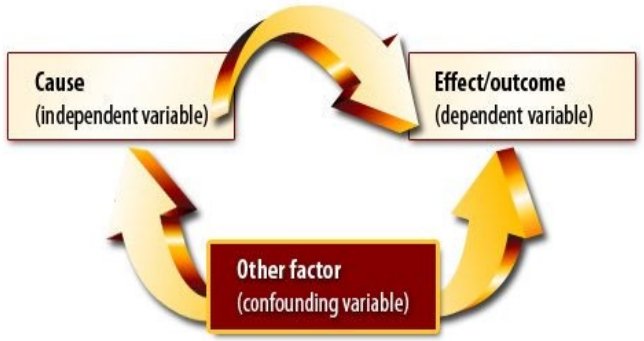
TABLE 1.15

Age	Cases	Controls
14–17	15	16
18–19	22	25
20–24	47	62
25–29	56	122
≥30	35	78

- This means that the odds that the younger mother has a preterm delivery is 1.5. In other words, the younger mothers would be more likely to have a preterm delivery.

Mantel-Haenszel Method

- **Mantel-Haenszel ($\mathrm{OR_{MH}}$) method:** Assess association between disease and exposure after controlling for one or more confounding variables.
- A **confounder**, or **confounding variable**, is a variable that may be associated with either the disease or exposure or both.



- When both the disease and the exposure are binary, the Mantel-Haenszel ($\mathrm{OR_{MH}}$) method is usually used to asses the association between the disease and the exposure.
- $\mathrm{OR_{MH}}$ provides one single estimate for the common odds ratio and can be summarized as follows:
  - We form 2×2 tables, one at each level of the confounder.
  - At a level of the confounder, we have the data listed in Table 1.17.

- The odds ratio at each level is estimated by ad/bc
- The Mantel–Haenszel procedure pools data across levels of the confounder to obtain a combined estimate (some kind of weighted average of level-specific odds ratios):

$$OR_{MH} = \frac{\sum ad/n}{\sum bc/n}$$

TABLE 1.17

	Disease Classification		
Exposure	+	−	Total
+	a	b	r_1
−	c	d	r_2
Total	c_1	c_2	n

Example 1.18

- A case-control study was conducted to identify reasons for the exceptionally high rate of lung cancer among male residents of coastal Georgia as first presented in Example 1.2. The primary risk factor under investigation was employment in shipyards during World War II, and data are tabulated separately for three levels of smoking (Table 1.18). There are three 2×2 tables, one for each level of smoking.

TABLE 1.18

Smoking	Shipbuilding	Cases	Controls
No	Yes	11	35
	No	50	203
Moderate	Yes	70	42
	No	217	220
Heavy	Yes	14	3
	No	96	50

Example 1.18 (Cont.)

- For the nonsmokers:

Shipbuilding	Cases	Controls	Total
Yes	11 (a)	35 (b)	46 (r_1)
No	50 (c)	203 (d)	253 (r_2)
Total	61 (c_1)	238 (c_2)	299 (n)

$$\frac{ad}{n} = \frac{(11)(203)}{299} = 7.47$$

$$\frac{bc}{n} = \frac{(35)(50)}{299} = 5.85$$

- For the For moderate:

Shipbuilding	Cases	Controls	Total
Yes	70 (a)	42 (b)	112 (r_1)
No	217 (c)	220 (d)	437 (r_2)
Total	287 (c_1)	262 (c_2)	549 (n)

$$\frac{ad}{n} = \frac{(70)(220)}{549} = 28.05$$

$$\frac{bc}{n} = \frac{(42)(217)}{549} = 16.60$$

- For heavy smokers

Shipbuilding	Cases	Controls	Total
Yes	14 (a)	3 (b)	17 (r_1)
No	96 (c)	50 (d)	146 (r_2)
Total	110 (c_1)	53 (c_2)	163 (n)

$$\frac{ad}{n} = \frac{(14)(50)}{163} = 4.29$$

$$\frac{bc}{n} = \frac{(3)(96)}{163} = 1.77$$

- These results are combined to obtain a combined estimate for the common odds ratio:

$$\text{OR}_{\text{MH}} = \frac{7.47 + 28.05 + 4.28}{5.85 + 16.60 + 1.77} = 1.64$$

- This combined estimate of the odds ratio, 1.64, represents an approximate increase of 64% in lung cancer risk for those employed in the shipbuilding industry.

Example 1.19

- A case–control study was conducted to investigate the relationship between myocardial infarction (MI) and oral contraceptive use (OC). The data, stratified by cigarette smoking, are listed in Table 1.20.

TABLE 1.20

Smoking	OC Use	Cases	Controls
No	Yes	4	52
	No	34	754
Yes	Yes	25	83
	No	171	853

TABLE 1.21

	Smoking	
	No	Yes
ad/n	3.57	18.84
bc/n	2.09	12.54

Application of the Mantel–Haenszel procedure yields the results shown in Table 1.21. The combined odds ratio estimate is

$$\mathrm{OR}_{\mathrm{MH}} = \frac{3.57 + 18.84}{2.09 + 12.54} = 1.53$$

representing an approximate increase of 53% in myocardial infarction risk for oral contraceptive users.

Standardized Mortality Ratio

- In a cohort study, the follow-up death rates are calculated and used to describe the mortality experience of the cohort under investigation. However, the observed mortality of the cohort is often compared with that expected from the death rates of the national population (used as *standard or baseline*).
- **Standardized (adjusted) rate:** A rate which differs from a crude rate in having been standardized to a different population (usually to a standard population) to remove the influence of some extraneous variable, such as age.
- **Standardization (Adjustment)** is a weighted average derived from a standard population. Two forms of standardization are commonly used: direct and indirect. Here is the Indirect one:
  - **Standardized Mortality Ratio (SMR)**: Divide the actual deaths (d) by the expected deaths (e):

$$SMR = \frac{d}{e}$$

  - The expected number of deaths is calculated using published national life tables, and the calculation can be approximated as follows:

$$e \approx \lambda T$$

  - T is the total follow-up time (person-years) from the cohort.
  - λ is the annual death rate (per person) from the referenced population.
- Indirect standardization yields an expected number of deaths, which can then be compared to the number of actual deaths, as in the SMR, or to the expected number of deaths in another population.

Example 1.20

- Some 7000 British workers exposed to vinyl chloride monomer were followed for several years to determine whether their mortality experience differed from those of the general population.
- The data in Table 1.22 are for deaths from cancers and are tabulated separately for four groups based on years since entering the industry.
  - Taking the ratio of two standardized mortality ratios is another way of expressing relative risk. For example, the relative risk of the 15+ years group is 1.58 times the risk of the risk of the 5-9 years group, since the ratio of the two corresponding mortality ratios is

$$\frac{111.8}{70.6} = 1.58$$

  - Similarly, the risk of the 15+ years group is 2.51 times the risk of the 1-4 years group because the ratio of the two corresponding mortality ratios is $\frac{111.8}{44.5} = 2.51$

TABLE 1.22

Deaths from Cancers	Years Since Entering the Industry 1–4	5–9	10–14	15+	Total
Observed	9	15	23	68	115
Expected	20.3	21.3	24.5	60.8	126.8
SMR (%)	44.5	70.6	94.0	111.8	90.7