Unit I – Problem 10 – Handout (3): Statistics in Epidemiology

- **Epidemiology**: is the study of the distribution, determinants and dynamics of health and disease in groups of people in relation to their environment and ways of living. The basic statistical measures in epidemiology are rates and measures of risk.
- **Rates:**
  - All rates consist of a numerator (usually the number of people with a particular condition) and a denominator (usually the number of people at risk) and they usually specify a unit of time. The most important rates are incidence and prevalence (which are both measures of morbidity), mortality and case-fatality.
- **Incidence:**
  - The incidence of a disease is the number of new cases occurring in a particular time period, such as 1 year. The incidence rate is therefore the ratio of new cases of the disease to the total number of people at risk.

$$\textbf{Incidence rate} = \frac{\textit{number of new cases of the disease}}{\textit{total number of people at risk}}\ \textit{per unit of time}$$

  - The incidence rate is often stated per 100,000 of the population at risk, or as a percentage. Incidence rates are found by the use of cohort studies, which are therefore sometimes also known as incidence studies. For example, if the incidence of shingles in a community is 2000 per 100,000 per annum, this tells us that in 1 year, 2% of the population experiences an episode of shingles.
- **Prevalence:**
  - The prevalence of a disease is the number of people affected by it at a particular moment in time. The prevalence rate is therefore the ratio of the number of people with the disease to the total number of people at risk:

$$\textbf{Prevalence rate} = \frac{\textit{number of people with the disease}}{\textit{total number of people at risk}}\ \textit{at a particular time}$$

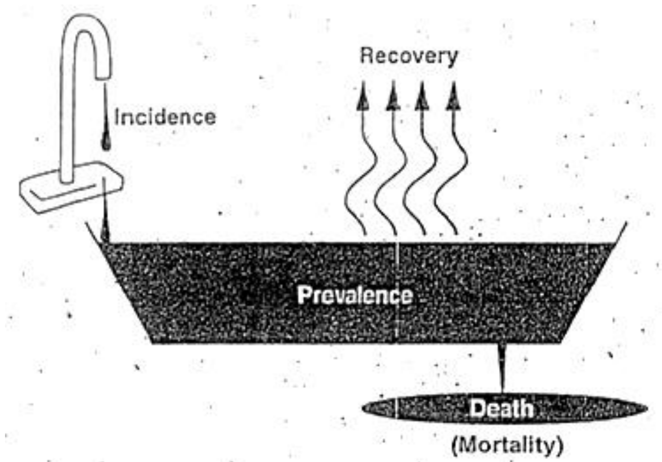
  - Like incidence rates, prevalence rates are often stated per 100,000 people, or as a percentage. They are generally found by prevalence surveys. For example, at a given time 170 of every 100,000 people (0.17%) in a community might be suffering from shingles.
  - Prevalence is an appropriate measure of the burden of a relatively stable chronic condition (such as hypertension or diabetes). However, it is not generally appropriate for acute illnesses as it depends on the average duration disease – it is of little value to speak of the prevalence of pulmonary emboli or myocardial infarctions.
  - Prevalence is equal to the incidence multiplied by the average duration of the disease, so an increased prevalence rate may merely reflect increased duration of an acute illness, rather than suggesting that members of the population are at greater risk of contracting the disease.
  - The incidence and prevalence rates of shingles given in the above examples suggest that the average episode of this illness lasts approximately 1 month, as the prevalence is one-twelfth of the annual incidence. If a new treatment cut the duration of an episode of shingles in half, to 2 weeks, but did nothing to prevent shingles from occurring, the incidence would not change but the prevalence at any given time would be cut in half.
    - ✓ Before new treatment:
      Prevalence = annual incidence x average duration (in years)
      = 2% x 1/12
      = 0.17%
    - ✓ After new treatment:
      Prevalence = annual incidence x average duration (in years)
      = 2% x 1/24
      = 0.085%

  Incidence and prevalence are both measures of morbidity, or the rate of illness.

- **Mortality:**
  - Mortality is the number of deaths. The mortality rate is the ratio of the number of people dying (whether of a specific disease or of all causes) to the total number of people at risk:

$$\textbf{Mortality rate} = \frac{\textit{total number of deaths}}{\textit{total number of people at risk}}\ \textit{per unit of time}$$

  - Like incidence and prevalence, mortality rates may be expressed as a percentage, or the number of deaths per 1000 or 100,000 people, typically per annum. Mortality is actually a special form of incidence in which the event in question is death rather than contracting a disease. Mortality figures are likely to be more accurate than incidence figures, because deaths are always recorded whereas episodes of illness are not. However, accurate records of causes of death are often unavailable, and mortality rates will not reflect the total burden of illness except in the case of diseases that are always fatal.
- **The epidemiologist’s bathtub:**
  - The relationships between incidence, prevalence and mortality in any disease can be visualized with the aid of the “epidemiologist’s bathtub” shown in the figure.
    - ✓ The flow of water through the faucet into the bathtub is analogous to incidence, representing the arrival of new cases of the disease.
    - ✓ The level of water in the tub represents the prevalence, or number of cases of the disease existing at any given point in time.
    - ✓ The flow of water out through the drain represents mortality.
    - ✓ The evaporation of water represents either cure or natural progression to recovery.



- **Case fatality:**
  - The case fatality rate (CFR) is the ratio of the number of people dying in a particular episode of a disease to the total number of episodes of the disease, expressed as a percentage:

$$\textbf{CFR} = \frac{\textit{total number of people dying in an episode of the disease}}{\textit{total number of episodes of the disease}}\ x\ 100$$

  - The CFR is a measure of the prognosis, in terms of life or death, for an episode of a given disease, because it shows the likelihood of one episode or occurrence of it resulting in death. It is used to follow the effectiveness of treatments over time or in different places (e.g. what is the CFR of acute myocardial infarction or pulmonary embolism in a given hospital or community).

- **Attack rates:**
  - The attack rate is the ratio of the number of people contracting a particular disease to the total number of people at risk, expressed as a percentage:

$$\textbf{Attack rate} = \frac{\textit{number of people contracting a disease}}{\textit{total number of people at risk}}\ \boldsymbol{x\ 100}$$

  - For example, if 1000 people eat at a barbecue at which contaminated food is served and 300 of these people become sick, the attack rate is (300/1000) x 100 = 30%
  - Attack rates are useful in attempting to deduce the source of an epidemic. To take the barbecue example, different people will have eaten different combinations of foods. The table below shows attack rates for each food:

Food	Number who ate	Number who got sick	Attack rate
Chicken only	100	25	25%
Ribs only	80	10	12.5%
Cole slaw only	20	7	35%
Chicken & ribs only	200	18	9%
Chicken & ribs & cole slaw	600	240	40%
Overall	1000	300	30%

  - The source of the illness can be deduced by inspecting the table for the maximum difference between any two attack rates. The largest difference between any two attack rates is 31%, which is the difference between the lowest rate, 9% (in those who ate the chicken and ribs only), and the highest rate, 40% (in those who ate the chicken, ribs and cole slaw). The implication is therefore that the cole slaw is the source.
- **Adjustment of rates:**
  - Researchers may want to compare rates across different populations (e.g. to compare the incidence of a disease in two cities or countries). However, if the populations differ significantly on one or more factors that are relevant to the illness in question, the comparison will be biased.
    - ✓ For example, a researcher wants to compare the prevalence of AIDS in two cities of equal size. City A has a large proportion of elderly people, whereas city B does not – so it would not be surprising if city A had a lower prevalence of AIDS than city B. However, due to the confounding effects of the different age structures of the two cities populations, this prevalence rate alone tells the researcher nothing about any real underlying difference in the prevalence of AIDS in the two cities.
  - This biasing influence of a confounding variable such as age can be removed by the technique of adjustment (or standardization) of rates. This involves calculating rates for the two populations as if they were both the same in terms of the factors (such as age) that are relevant to the disease, so that their rates of the disease can be compared. This kind of process of adjustment (or standardization) can be done not only for age, but also for any other relevant factor that differs substantially between two populations that are being compared. Adjustment can be made for two or more factors simultaneously, for example, if city A had many more IV drug abusers than city B, this difference could also be adjusted for. Mortality rates are commonly standardized, producing a statistic called standardized mortality ratio (SMR).
- **Measurement of risk:**
  - Information about the risk of contracting a disease is of great value in medicine. The knowledge that something is a risk factor for a disease can be used to help in:
    - ✓ Preventing the disease.
    - ✓ Predicting its future incidence and prevalence.
    - ✓ Diagnosing it (diagnostic suspicious will be aroused if it is known that a patient was exposed to the risk factor).
    - ✓ Establishing the cause of a disease of unknown etiology.

- **Absolute risk:**
  - The fundamental measure of risk is incidence. The incidence of a disease is, in fact, the absolute risk of contracting it. For example, if the incidence of a disease is 10 per 1000 people per annum, the absolute risk of a person actually contracting it is also 10 per 1000 per annum or1% per annum.
  - It is useful to go beyond absolute risk and to compare the incidence of a disease in different groups of people to find out if exposure to a suspected risk factor (such as smoking cigarettes) increases the risk of contracting a certain disease (such as lung cancer). A number of different comparison of risk can be made, including relative risk, attributable risk and the odds ratio. All these are called measures of effect – they measure the effect of being exposed to a risk factor on the risk of contracting a disease.
  - The ideal way of determining the effect of a risk factor is by a controlled experiment, but this is rarely ethical. The best alternative is the cohort (prospective) study, in which the incidence of disease in exposed and non-exposed people can be observed directly.
  - One of the main goals of these studies is to find the extent to which the risk of contracting the disease is increased by exposure to the risk factor. The two measures that show this are relative risk and attributable risk.
- **Relative risk:**
  - Relative risk states by how many times exposure to the risk factor increases the risk of contracting the disease. It is therefore the ratio of the incidence of the disease among exposed persons to the incidence of the disease among unexposed persons:

$$\textbf{Relative risk} = \frac{\textit{incidence of the disease among persons exposed to the risk factor}}{\textit{incidence of the disease among persons not exposed to the risk factor}}$$

  - As an example, the table below reports the results of a hypothetical cohort study of lung cancer in which 1008 heavy smokers and 1074 non-smokers were followed for a number of years. The incidence of lung cancer over the total time period of the study among people exposed to the risk factor (heavy cigarette smoking) is 283/1008 or 0.28 (28%) while the incidence among those not exposed is 64/1074 or 0.06 (6%).

	Disease outcome		
Risk	**Lung cancer**	**No lung cancer**	**Total**
Exposed (heavy smokers)	283	725	1008
Non-exposed (non-smokers)	64	1010	1074
Total	347	1735	2082

  - The relative risk is therefore 0.28/0.06 or 4.67, showing that people who smoked cigarettes heavily were 4.67 times more likely to contract lung cancer than were non-smokers. (note that this not a measure of absolute risk-it states nothing about the likelihood of heavy smokers contracting cancer in absolute terms).
  - Because relative risk is a ratio of risks, it is sometimes called the risk ratio, or morbidity ratio. In the case of outcomes involving death, rather than just disease, it may also be called the mortality ratio.
  - Many clinical trials report relative risk reductions due to the use of a drug; relative risk reduction is equal to 1 – relative risk:
    Relative risk reduction = 1 – relative risk
  - Relative risk reduction figures may be misleading if not understood properly. This can be illustrated by the well-known West of Scotland Coronary Prevention (WOSCOPS) study (Shepherd et al., 1995):
    - ✓ This study was a double-blind randomized clinical trial in which approximately 6000 men with elevated cholesterol levels were randomly assigned to groups taking either a placebo or the cholesterol-lowering drug pravastatin for an average of 4.9 years.
  - There were 73 deaths from cardiovascular causes in the placebo group (3293 men); the cardiovascular mortality rate was therefore 73/3293 = 0.022 (2.2%) in this group.

In the pravastatin group (3302 men), there 50 deaths from cardiovascular causes, giving a mortality rate of 50/3302 = 0.015, or 1.5%. The relative risk of death in those given the drug id 1.5/2.2 – 0.68, so the relative risk reduction is 1 – 0.68 – 0.32, or 32% - showing that an impressive 32% of cardiovascular deaths were prevented by the drug.

- However, the absolute risk reduction is 2.2% - 1.5% = 0.7% - a far less impressive sounding figure, showing that of all men given the drug for 4.9 years, 0.7% of them were saved from cardiovascular death.
- Absolute risk reduction allows circulation of another statistic that is of clinical importance. If 0.7% of patients were saved by the drug, this implies that (100/0.7) = 143 patients would have to be treated to save 1 life. This figure is called the number needed to treat, or NNT, and it allows the effectiveness of different treatments to be compared:

Number needed to treat = 1/absolute risk reduction

- The NNT allows a further calculation – the cost of saving one life with the treatment. WOSCOPS showed that 143 men needed to be treated for 4.9 years (58 months) to save 1 life; as the dose of pravastatin used in the study costs approximately $100 per month, it would cost $100 x 58 = $5800 to treat one man for this length of time. It would therefore cost $5800 x 143 = $829,400 to prevent one cardiovascular death over this period. This gives a very different perspective on the value of a treatment that the statement that it reduces the risk of death by 32%. similar analyses can be performed easily for almost any clinical trial in the literature.

- **Attributable risk:**
  - The attributable risk is the additional incidence of a disease that is attributable to the risk factor in question. It is equal to the incidence of the disease in exposed minus the incidence of the disease in non-exposed persons.
  - In the example of lung cancer and smokers, the attributable risk is 0.28 – 0.06, or 0.22 (22%). In other words, of the 28% incidence of lung cancer among the heavy smokers in this study, 22% is attributable to smoking. The other 6% is the "background" incidence of the disease – its incidence among those not exposed to this particular risk factor. Attributable risk is sometimes called risk difference, because it is the difference in the risks or incidences of the disease between the two groups of people.
- **Odds ratio:**
  - Relative risk and attributable risk both require the use of cohort (prospective) studies, as shown previously. Notice that cohort studies are generally expensive and time-consuming and are therefore often impractical.
  - A common alternative is to use a case-control (retrospective) study, which compares people with the disease (case) with otherwise similar people without the disease (controls), attempting to look back into the past to see if a possible risk factor is found more frequently among the cases than the controls. If the proportion of people who were exposed to the possible risk factor is greater among the cases, then the risk factor is implicated as a cause of the disease. The odds ratio (or relative odds) is a measure of these relative proportions – it is the ratio of the odds that a case was exposed to the odds that a control was exposed:

$$\textbf{Odd ratio} = \frac{\textit{odds that a case was exposed to the risk factor}}{\textit{odds that a control was exposed to the risk factor}}$$

  - Because the proportion of people in the study who do have the disease is determined by the researcher's choice, and not by the actual proportion in the population case-control studies cannot determine the incidence or prevalence of a disease, so they cannot determine the risk of contracting a disease. The odds ratio must therefore be used instead of relative risk when analyzing case-control data instead of cohort data.
    - ✓ For example, the hypothetical data that were used to illustrate the calculation of relative risk in the previous table can be used, but now it will be assumed that these were generated by a case-control study in which history of prior

exposure to the risk factor (cigarette smoking) was compared between 347 cases (with lung cancer) and 1735 controls (without lung cancer).

- As defined previously, the odds ratio is the ratio of the odds that a case was exposed to the odds that a control was exposed, it can be shown that this is equal to:

$$\frac{\boldsymbol{number\ of\ cases\ exposed\ to\ risk\ factor\ (A) x\ number\ of\ controls\ not\ exposed\ (D)}}{\boldsymbol{number\ of\ controls\ exposed\ to\ risk\ factor\ (B) x\ number\ of\ cases\ not\ exposed\ (C)}}$$

The table below shows that:

  - ✓ 283 of the cases were exposed to the risk factor (A)
  - ✓ 725 of the controls were exposed to the risk factor (B).
  - ✓ 64 of the cases were not exposed to the risk factor (C).
  - ✓ 1010 of the controls were not exposed to the risk factor (D).

The odds ratio, therefore, $\frac{283\ x\ 1010}{725\ x\ 64} = \frac{285830}{46400} = 6.16$

	Disease outcome		
Risk	**Lung cancer**	**No lung cancer**	**Total**
Exposed (heavy smokers)	283 (A)	725 (B)	1008
Non-exposed (non-smokers)	64 (C)	1010 (D)	1074
Total	347	1735	2082

In other words, among the people studied, a person with lung cancer was 6.16 times more likely to have been to the risk factor (cigarette smoking) than was a person without lung cancer.

- An odds ratio of 1 indicates that a person with the disease is no more likely to have been exposed to the risk factor than is a person without the disease, suggesting that the risk factor is not related to the disease. An odds ratio of less than 1 indicates that a person with the disease is less likely to have been exposed to the risk factor than is a person without the disease, implying that the risk factor may actually be a protective factor against the disease.
- In some ways the odds ratio is similar to the relative risk: both figures demonstrate the strength of the association between the risk factor and the disease, albeit in different ways. As a result of their similarities, the odds ratio is sometimes called estimated relative risk –it provides a reasonably good estimate of relative risk provided that the incidence of the disease is low (which is usually true of chronic diseases), and that the cases and controls examined in the study are representative of people with and without the disease in the population.