Biostatistics

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR Plano de Recuperação e Resiliência

GOVERNO DE PORTUGAL

Financiado pela União Europeia NextGenerationEU

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Introduction and general concepts

What is Epidemiology?

- The first part of the syllabus relates to epidemiological concepts and techniques.
- According to Woodward (2014, p 1), *epidemiology* is the study of the distribution and determinants of disease in human populations.
- Epidemiological information is used to plan and evaluate strategies to prevent illness and also as a way to guide the management of patients who already developed the disease.
- Still according to Woodward (2014, p 2), the term derives from the term ‘epidemic’, which appears to have been derived from *epidemeion*, a word used by Hippocrates when describing a disease that was ‘visiting the people’.
- Modern use of the term retains the restriction to human populations but has broadened the scope to include any type of disease, transient or not.

Measures of disease occurrence

- The fundamental observations in epidemiology are measures of the occurrence of disease.
- We will be considering that **disease outcomes are binary**: disease present or disease absent (or, more generally, health outcome of interest present or absent).
- Although fine levels of a disease outcome variable would enhance the understanding of the associations between disease and exposure to risk factors, quantifying a continuous level of disease may involve invasive methods and therefore may be impractical or unethical.

Measures of disease occurrence

↪ In what follows we will be using the term "**at risk**". To be 'at risk' means that an individual has been previously unaffected by the disease, or that susceptibility has been regained after previously contracting the disease and recovering (e.g., as with the cold/flue that no sufferer becomes fully immune).

↪ There are certain diseases that some subjects can never contract and therefore are never at risk for such diseases (e.g., women cannot develop prostate cancer).

Measures of disease occurrence

Prevalence and incidence

- The (point) **prevalence** of a disease is the proportion of a defined population at risk for the disease that is affected by it at a specified point in time.
- The prevalence at time, say t, is estimated as the ratio of the number of existing cases of the disease at time t to the size of the population at risk ~~at time t~~.
- The **incidence proportion** is the proportion of a defined population, all of whom are at risk for the disease at the beginning of a specified time interval, who become new cases of the disease before the end of the interval.
- Because the incidence proportion includes all individuals who contracted the disease over the entire interval, it is sometimes also referred to as the **cumulative incidence**.
- An incidence proportion can only be meaningful interpreted when the time interval is specified.

Measures of disease occurrence

Prevalence and incidence

↪ Toy example from Jewell (2003, p 12).

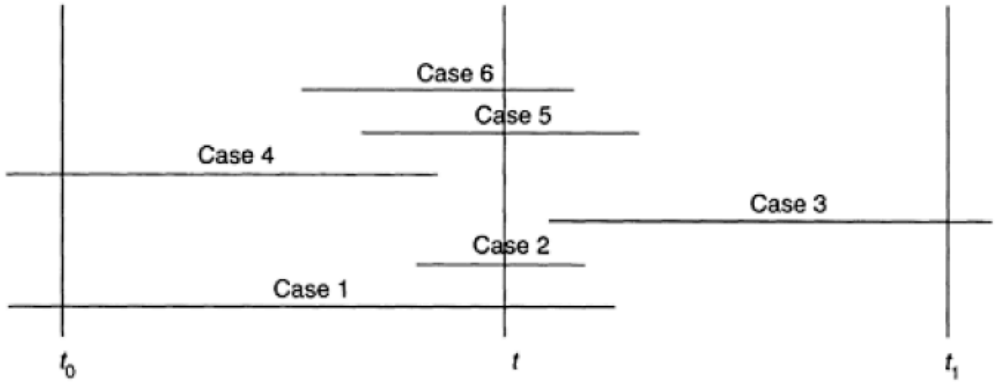

Figure 2.1 *Schematic illustrating calculation of an incidence proportion and point prevalence. Six cases of disease in a population of, say, 100 individuals are represented. Lines represent the duration of disease.*

↪ Assume that the 94 individuals not represented have not contracted the disease between time t_0 and t_1 and were all at risk for the disease at time t_0.

Measures of disease occurrence

Prevalence and incidence

- The prevalence at time t is either $4/100$ or $4/99$, depending on whether we consider case 4 to be at risk for the disease at time t or not, respectively (e.g., he/she could have became immune for the disease and then no longer at risk).

- The incidence proportion over the interval time $[t_0, t_1]$ is $4/98$ because cases 1 and 4 are not at risk for the disease at the beginning of the interval (they are already cases).

- Note that the prevalence and incidence proportion do not have units. They are simply proportions, sometimes expressed as percentages, that must lie on $[0, 1]$.

Measures of disease occurrence

Prevalence and incidence

- ↪ It is easy to understand that the prevalence is affected by the duration of the disease.
- ↪ Diseases of long duration tend to have a higher prevalence than short term illnesses (either because of recovery or death), even if the total number of affected individuals is about the same.
- ↪ That is, a population might have a low disease prevalence when: (1) the disease is rare, or (2) it occurs with higher frequency, but affected individuals stay diseased only for a short period of time (either because they recover or die).

Measures of disease occurrence

Prevalence and incidence

- Further, because the duration of the disease, and hence its prevalence, may be affected by medical treatment or other factors that are unrelated to those that caused the disease in first place, the prevalence is not as appropriate as the incidence proportion to investigate the causes of the disease (known in the epidemiological language as the etiology of the disease).
- The following example taken from Jewell (2003, p 13) illustrates the dangers of using prevalence data when trying to establish a causal association between an exposure and initiation of disease.
- The data are from the Framingham Heart Study (Friedman et al., 1966) and relates cholesterol levels and coronary heart disease (CHD) for men aged between 30 and 59 years old.

Table 2.1 *Prevalence and incidence data (proportions) on CHD in males*

	Incidence (10 year)		Prevalence	
Cholesterol	CHD	No CHD	CHD	No CHD
High	85 (75%)	462 (47%)	38 (54%)	371 (52%)
Low	28 (25%)	516 (53%)	33 (46%)	347 (48%)

Source: Friedman et al. (1966).

- To know more about CHD, please see `https://www.nhs.uk/conditions/coronary-heart-disease/`.

Measures of disease occurrence

Prevalence and incidence

- In this example, the incidence data refers to a group of men, initially free of CHD, whose cholesterol levels were measured at the beginning of a 10 year follow-up period during which the number of men who developed CHD was counted.
- On the other hand, for the prevalence data, both cholesterol and CHD were taken at the end of the 10-year monitoring period.
- Cholesterol levels were divided into four quartiles and ‘high’ and ‘low’ refer to the highest and lowest quartiles, respectively.
- It is clear from the incidence data that there is a considerable larger proportion of CHD cases in the high cholesterol group as compared with the low cholesterol group (75% versus 25%, respectively).
- However, this is not evident at all from the prevalence data (54% versus 46%), where both cholesterol and CHD measurement were taken at the end of the 10-year follow-up period.

Measures of disease occurrence

Prevalence and incidence

- ↪ This difference may arise if high cholesterol levels are associated only with those CHD cases who died rapidly (i.e., died before the end of the 10-year period) and thus were not included in the prevalence calculation.

- ↪ Another explanation might be that some surviving CHD patients modified their cholesterol levels (e.g. through changes in lifestyle or medication), so that their cholesterol levels were lower at the end of the follow-up period.

- ↪ According to Jewell (2013, p 13), this latter possibility was supported by a more detailed analysis of the Framingham data.

Measures of disease occurrence

Prevalence and incidence

↪ The prevalence is nonetheless extremely useful for measuring burden of long lasting diseases on a population, especially if those contracting the disease require specific medical attention.

↪ For example, the prevalent number of people in a population with end-stage renal disease predicts the need in that population for dialysis treatment.

Duv. 1 

Measures of disease occurrence

Prevalence and incidence

- The incidence proportion is also not free of drawbacks.
- The main drawback is a practical one and relates to the fact that subjects need to be followed up over a period of time.
- First, there are **competing risks** that may lead to some subjects dying before the observation period ends, making it impossible to ascertain whether they would have developed the disease of interest if they had not died early from a cause other than the outcome of the study.
- Secondly, it is also challenging to follow individuals for long periods of time, and subjects can become **lost to follow-up** (e.g., because they move away or choose not to participate further in the study) which also means that we do not know whether they would have developed the disease or not in case they have remained in the study.

Measures of disease occurrence

Prevalence and incidence

- The difficulty in computing the incidence proportion in studies in which there are competing risks and losses to follow-up is what to use in the denominator.
- If the denominator is comprised by the number of subjects who were initially being followed, then we would be underestimating the incidence proportion that would have been observed if there had been no competing risks or losses to follow-up.
- Generally, the incidence proportion is more useful when the follow-up time is relatively short and therefore one expect few losses to follow-up.
- However, when measuring the incidence of a rare disease, individuals need to be followed for long periods.
- Further, the definition of the incidence proportion implicitly assumes a closed population, i.e., no new individuals at risk are allowed to enter after the beginning of the risk period.

Measures of disease occurrence

Prevalence and incidence

- ↪ To address the aforementioned drawbacks, epidemiologists often resort to the **incidence rate** as a measure of disease occurrence.
- ↪ The incidence rate of a disease over a specified time interval is given by the number of new cases during the interval divided by the total amount of time at risk for the disease accumulated by the entire population over the same interval.
- ↪ The units of the incidence rate are thus $(\text{time})^{-1}$.
- ↪ It is worth mentioning that a **mortality rate** is an incidence rate in which the outcome under study is death.

Measures of disease occurrence

Prevalence and incidence

↪ Summing up, we calculate the **incidence proportion** over a time interval as the following ratio

$$\frac{\#\text{ individuals at risk for the disease at the beginning of the time interval who develop the disease during the interval}}{\#\text{ individuals at risk for the disease at the beginning of the time interval}},$$

whereas the **incidence rate** over a time interval is computed as the following ratio

$$\frac{\#\text{ individuals developing the disease over the time interval and who were at risk at some time during the interval}}{\text{total observation time accumulated by all those individuals at risk at some time during the interval}}.$$

Measures of disease occurrence

Prevalence and incidence

↪ The following example is a toy one from Jewell (2003, p 14).

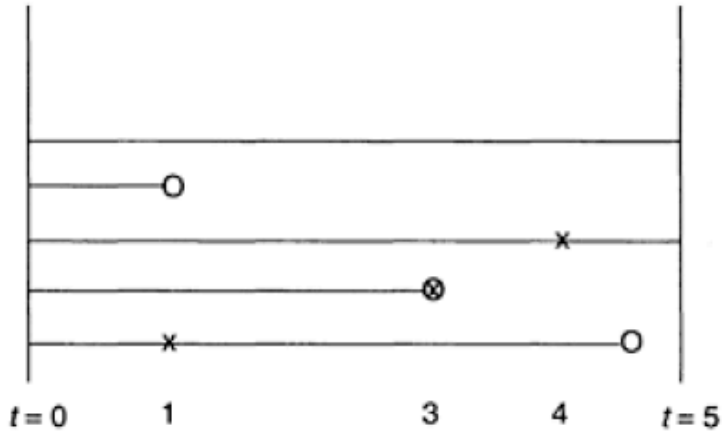

Figure 2.2 *Schematic illustrating calculations of an incidence proportion, point prevalence, and incidence rate. Population of 5; the symbols represent: O, death; x, incident case of disease. Here, lines represent time alive.*

Measures of disease occurrence

Prevalence and incidence

- In the following calculations we will first assume that the disease under study is chronic and that, as such, there is no recovery.
- Prevalence
  - At $t = 0$: $\frac{0}{5} = 0$.
  - At $t = 5$: $\frac{1}{2} = 0.5$.
- Incidence proportion over the time interval $[0, 5]$: $\frac{3}{5} = 0.6$.
- Incidence rate over the time interval $[0, 5]$: $\frac{3}{5+1+4+3+1} = \frac{3}{14} = 0.21$ cases per year.

Measures of disease occurrence

Prevalence and incidence

- ↪ Now, under the assumption that the disease is acute thus implying that individuals who recover immediately return to being at risk, then the denominator of the incidence rate would be $5 + 1 + \mathbf{5} + 3 + \mathbf{4.5}$, leading to an incidence rate of 0.16 cases per year.
- ↪ Unlike the prevalence and incidence proportion, the incidence rate does not (necessarily) lie on $[0, 1]$.
- ↪ It indeed has a lower bound of zero, but it can theoretically become as great as infinity.
- ↪ Because the denominator of an incidence rate is measured in time units, we can imagine that the time units can be smaller, making the rate larger.
- ↪ Indeed, the value of the incidence rate depends on what time unit is chosen.

Measures of disease occurrence

Prevalence and incidence

- As an example, suppose that we measure an incidence rate in a population as 47 cases occurring in 158 months.
- This would lead to $\frac{47}{158} = 0.3$ cases per month.
- We could restate this same incidence rate using cases per year instead of cases per month, thus leading to $\frac{47}{13.17} = 3.57$ cases per year.
- The above two expressions measure the same incidence rate; the only difference is the time unit chosen to express the denominator.
- The different time units affect the numerical values of the incidence rate.
- The situation is pretty much similar as expressing speed in different units of time or distance!