Lecture 1: Measuring Disease Occurrence: Prevalence, incidence, incidence density

Dankmar Böhning

Southampton Statistical Sciences Reserch Institute
University of Southampton, UK

2 - 4 March 2015

Purpose

The purpose of this material is to provide an overview on the most important measures of disease occurrence:

- prevalence
- incidence (cumulative incidence or risk)
- incidence density

Examples

The concepts will be illustrated with examples and practicals.

Measuring Disease Occurrence: Prevalence

Measuring Disease Occurrence: Incidence

Measuring Disease Occurrence: Incidence Density

Measuring Disease Occurrence: Prevalence

Prevalence:

is the **proportion** (denoted as p) of a specific population having a particular disease. p is a number between 0 and 1. If multiplied by 100 it is **percentage**.

Examples

In a population of 1000 there are two cases of malaria:
$p = 2/1000 = 0.002$ or 0.2%.
In a population of 10,000 there are 4 cases of skin cancer:
$p = 4/10,000 = 0.0004$ or 0.04%.

Measuring Disease Occurrence: Prevalence

epidemiological terminology

In epidemiology, disease occurrence is frequently small relative to the population size. Therefore, the proportion figures are multiplied by an appropriate number such as 10,000. In the above second example, we have a prevalence of 4 per 10,000 persons.

Exercise

In a county with 2300 inhabitant there have occurred 2 cases of leukemia. Prevalence?

Quantitative Aspects:

What is Variance and Confidence Interval for the Prevalence!

sample:

sample (population survey) of size n provides for disease status for each unit of the sample:

$$X_i = 1, \text{ disease present}$$

$$X_i = 0, \text{ disease not present}$$

consequently,

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n}$$

plausible **estimator of prevalence**.

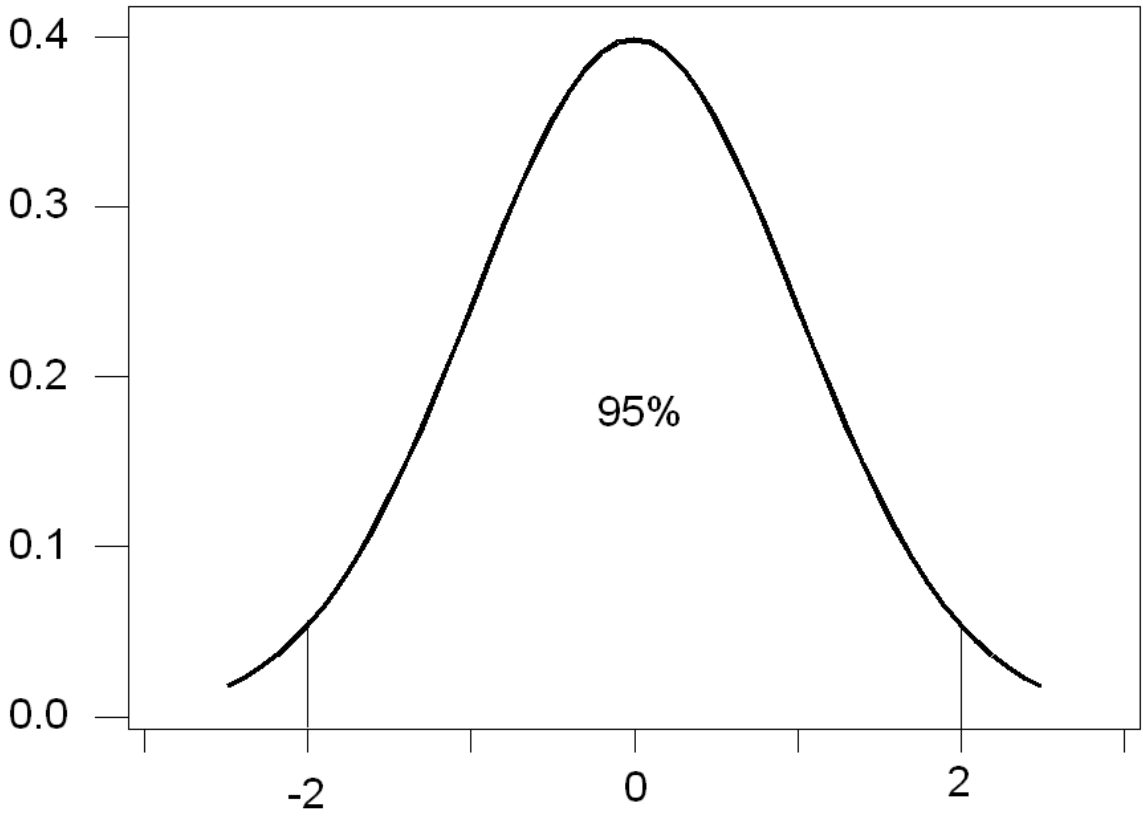
Computing Variance of Prevalence of $\hat{p}$:

consequently,

$$Var\,(\hat{p}) = \frac{p(1-p)}{n}$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

$\hat{p}$ is approx. normal



using the normal distribution for $\hat{p}$:

with 95% probability

$$-2 \leq \frac{\hat{p} - p}{SD(\hat{p})} \leq +2$$

$\Leftrightarrow$

$$\hat{p} - 2SD(\hat{p}) \leq p \leq \hat{p} + 2SD(\hat{p})$$

$\Leftrightarrow$

$$95\%CI : \hat{p} \pm 2SD(\hat{p})$$
$$= \hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}$$

Examples

In a population of 1000 there are two cases of malaria: $p = 2/1000 = 0.002$ or 0.2%.

$$Var(\hat{p}) = 0.002(1 - 0.002)/1000 = (0.00141280)^2,$$

$$SD(\hat{p}) = 0.00141280$$

$$95\%CI : \hat{p} \pm 2\sqrt{\hat{p}(1 - \hat{p})}/\sqrt{n}$$

$$= 0.002 \pm 2 \times 0.0014 = (0 - 0.0048)$$

Exercise

In a county with 2300 inhabitants there have occurred 2 cases of leukemia. Prevalence with CI?

Measuring Disease Occurrence: Incidence

Incidence:

is the proportion (denoted as I) of a specific, **disease-free** population **developing** a particular disease **in a specific study period**. I is a number between 0 and 1. If multiplied by 100 it is percentage.

Examples

In a malaria-free population of 1000 there are four new cases of malaria within one year : $I = 4/1000 = 0.004$ or 0.4%.
In a skin-cancer free population of 10,000 there are 11 new cases of skin cancer: $I = 11/10,000 = 0.0011$ or 0.11%.

Measuring Disease Occurrence: Incidence

Exercise

In a rural county with 2000 children within pre-school age there have occurred 15 new cases of leukemia within 10 years. Incidence?

Quantitative Aspects: How to determine Variance and Confidence Interval for the Incidence?

sample (population cohort - longitudinal) of size n, which is **initially disease-free**, provides the disease status for each unit of the sample **at the end of study period**:

$$X_i = 1, \text{ new case}$$

$$X_i = 0, \text{ disease not present}$$

consequently,

$$\hat{I} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

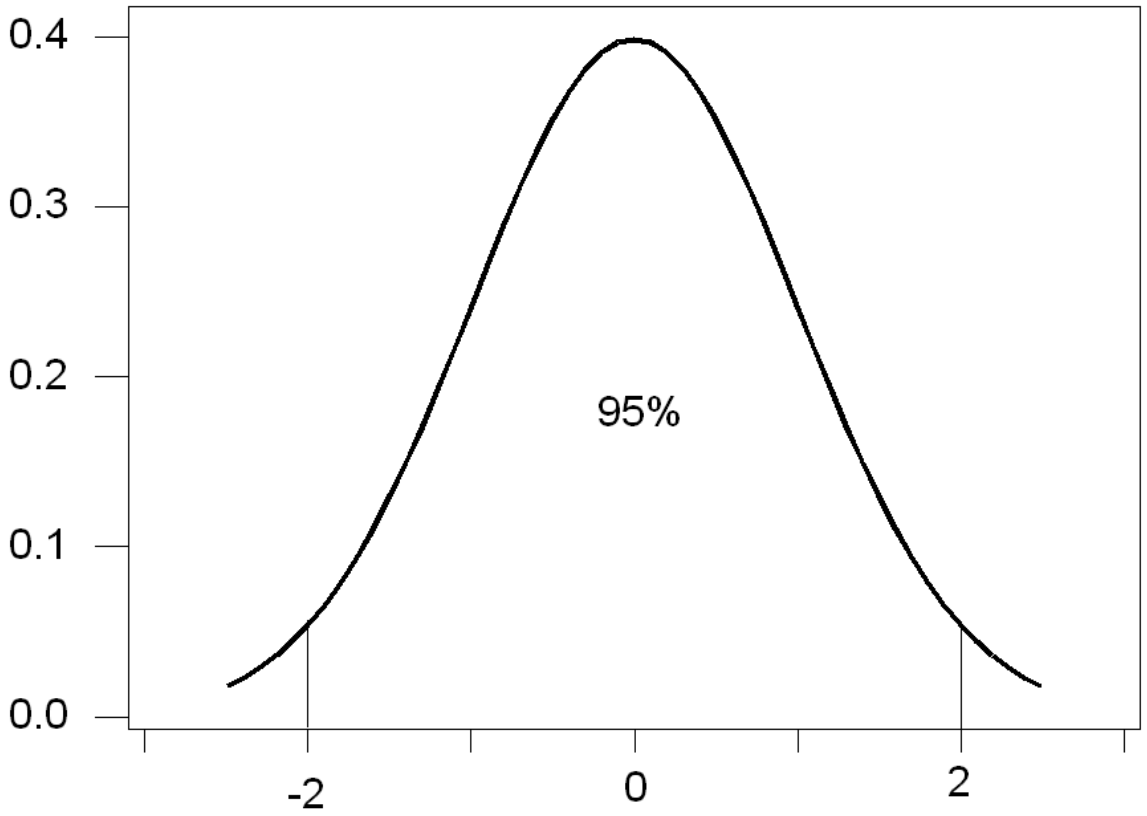
plausible **estimator of incidence**.

Computing Variance of Incidence

$$Var\left(\hat{I}\right) = \frac{I(1-I)}{n}$$

$$SD(\hat{I}) = \sqrt{\frac{I(1-I)}{n}}$$

$\hat{I}$ is approx. normal



95% confidence interval for the incidence density

with 95% probability

$$-2 \leq \frac{\hat{I} - I}{SD(\hat{I})} \leq +2$$

$\Leftrightarrow$

$$\hat{I} - 2SD(\hat{I}) \leq I \leq \hat{I} + 2SD(\hat{I})$$

$\Leftrightarrow$

$$95\%CI : \hat{I} \pm 2SD(\hat{I})$$

$$= \hat{I} \pm 2\sqrt{\hat{I}(1-\hat{I})}/\sqrt{n}$$

Examples

In a malaria-free population of 1000 there are four new cases of malaria within one year : $I = 4/1000 = 0.004$ or .4%.

$$Var(\hat{I}) = 0.004(1 - 0.004)/1000 = (0.001996)^2,$$

$$SD(\hat{I}) = 0.001996$$

$$95\%CI : \hat{I} \pm 2\sqrt{\hat{I}(1 - \hat{I})}/\sqrt{n}$$

$$= 0.004 \pm 2 \times 0.001996 = (0.000008 - 0.0080)$$

Exercise

In a rural county with 2000 children within pre-school age there have occurred 15 new cases of leukemia within 10 years. Incidence with 95% CI?

Measuring Disease Occurrence: Incidence Density

Incidence Density:

is the rate (denoted as ID) of a specific, **disease-free** population **developing** a particular disease **w. r. t. a specific study period of length** T. ID is a positive number, but not necessarily between 0 and 1.

estimating incidence density

suppose a disease-free population of size n is under risk for a time period T. Then a plausible estimator of ID is given as

$$\widehat{ID} = \frac{\sum_{i=1}^{n} X_i}{n \times T} = \frac{\text{count of events}}{\text{person–time}}$$

where $X_i = 1$ if for person i disease occurs and 0 otherwise.

Examples

A cohort study is conducted to evaluate the relationship between dietary fat intake and the development in prostate cancer in men. In the study, 100 men with high fat diet are compared with 100 men who are on low fat diet. Both groups start at age 65 and are followed for 10 years. During the follow-up period, 10 men in the high fat intake group are diagnosed with prostate cancer and 5 men in the low fat intake group develop prostate cancer.
The incidence density is $\widehat{ID} = 10/(1,000) = 0.01$ in the high fat intake group and $\widehat{ID} = 5/(1,000) = 0.005$ in the low fat intake group.

most useful generalization

occurs if persons are **different times under risk** and hence contributing differently to the person–time–denominator

estimating incidence density with different risk-times

suppose a disease-free population of size n is under risk for a time periods $T_1, T_2, ..., T_n$, respectively. Then a plausible estimator of ID is given as

$$\widehat{ID} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} T_i} = \frac{\text{count of events}}{\text{person–time}}$$

where $X_i = 1$ if for person i disease occurs and 0 otherwise, and T_i represents the person-time of person i in the study period.

Examples

Consider a population of $n = 5$ factory workers with $X_2 = 1$ and all other $X_i = 0$(here the disease incidence might be a lung disease). We have also $T_1 = 12,\ T_2 = 2, T_3 = 6, T_4 = 12, T_5 = 5$, so that

$$\widehat{ID} = \frac{1}{12 + 2 + 6 + 12 + 5} = 1/37.$$

interpretation of incidence density:

In the above example of diet-cancer study: $\widehat{ID} = 0.01$ means what? There is no longer the interpretation of 1 case per 100 men, **but** 1 case per 100 men-years!

The interpretation is now **number of events per person–time**!

Quantitative Aspects for the Incidence Density

sample (population cohort - longitudinal) of size n available:

$$\text{event indicators: } X_1, ..., X_n$$

$$\text{person times: } T_1, ..., T_n$$

estimate of incidence density

$$\widehat{ID} = \frac{X_1 + X_2 + ... + X_n}{T_1 + T_2 + ... + T_n} = \frac{X}{T}$$

a variance estimate can be found as

$$\widehat{Var}(\widehat{ID}) = \frac{\widehat{ID}}{T} = \frac{X}{T^2}$$

Quantitative Aspects for the Incidence Density

variance estimate can be found as

$$\widehat{Var}(\widehat{ID}) = \frac{\widehat{ID}}{T} = \frac{X}{T^2}$$

so that a 95% confidence interval is given as

$$\widehat{ID} \pm 2\sqrt{\frac{\widehat{ID}}{T}}$$

Example

Consider the population of $n = 5$ factory workers with $X_2 = 1$ and all other $X_i = 0$ (here the disease incidence might be a lung disease). We have $X = 1$ and $T = 37$, so that $\widehat{ID} = 1/37 = 0.027$. The variance is $\frac{\widehat{ID}}{T} = 0.0007$ and standard deviation 0.027. This leads to a 95% CI

$$\widehat{ID} \pm 2\sqrt{\frac{\widehat{ID}}{T}} = 0.027 \pm 2 \times 0.027 = (0, 0.081).$$

Exercise

We return to the cohort study mentioned before. It had been conducted to evaluate the relationship between dietary fat intake and the development in prostate cancer in men. In the study, 100 men with high fat diet are compared with 100 men who are on low fat diet. Both groups start at age 65 and are followed for 10 years. During the follow-up period, 10 men in the high fat intake group are diagnosed with prostate cancer and 5 men in the low fat intake group develop prostate cancer.

Compute 95% CI for incidence densities:

$$\text{high fat intake group: } \widehat{ID} = 10/(1,000) = 0.01$$

$$\text{low fat intake group: } \widehat{ID} = 5/(1,000) = 0.005$$