Frequency Measures used in Epidemiology

Prof. Giuseppe Verlato

Unit of Epidemiology & Medical Statistics

Further information can be obtained from the following Web sites:
http://biometria.univr.it
http://www2a.cdc.gov/phtn/catalog/pdf-file/LESSON2.pdf
http://www.pitt.edu/~super1/

Epidemiologic data come in many forms and sizes.

One of the most common forms is a **rectangular database** made up of rows and columns.

Each row contains information about one individual, and it is called a "record" or "**observation**."

Each column contains information about one characteristic such as sex, date of birth or disease, and it is called a "**variable**".

The first column of an epidemiologic database usually contains the individual's initials, or identification number which allows us to identify who is who. We can also use the name if this does not infringe the individual's right to privacy.

SEX	AGE years	Height (m)	Weight Kg	SMOKE	INFARCTION
M	42	1,70	58	F	N
M	48	1,84	90	N	I
M	51	1,66	70	F	I
M	54	1,78	76	F	I
M	58	1,74	72	N	N
M	60	1,76	85	N	I
M	62	1,64	62	F	I
M	64	1,90	88	F	I
M	65	1,72	69	N	N
M	70	1,77	77	N	N
M	75	1,68	73	F	I
M	81	1,74	75	F	I
F	45	1,68	59	F	N
F	49	1,58	55	N	N
F	51	1,62	68	N	N
F	53	1,65	64	F	I
F	60	1,72	70	N	N
F	63	1,69	65	F	I
F	68	1,70	73	N	I
F	75	1,66	52	N	N

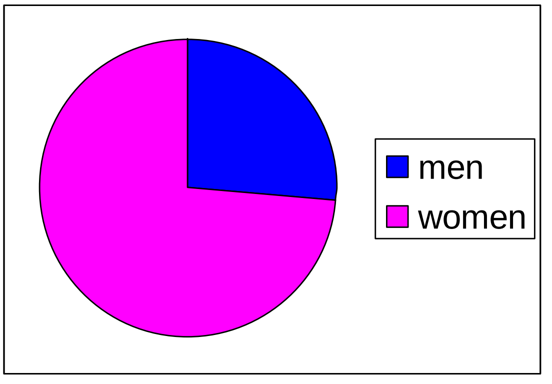
With large databases, it is very difficult to pick out the information needed at a glance. Instead, it is more convenient to summarize variables into tables called "**frequency distributions**."

A frequency distribution shows **the values** a variable can take, and the **number of people or records with each value**.

For example, suppose we want to describe the parity of a group of women, i.e. the number of children each woman has given birth to. To construct a frequency distribution showing these data, we first list, from the lowest observed value to the highest, all the values that the variable parity can take. For each parity value, we then enter the number of women who had given birth to that number of children.

Frequency distribution of a categorical variable (sex)

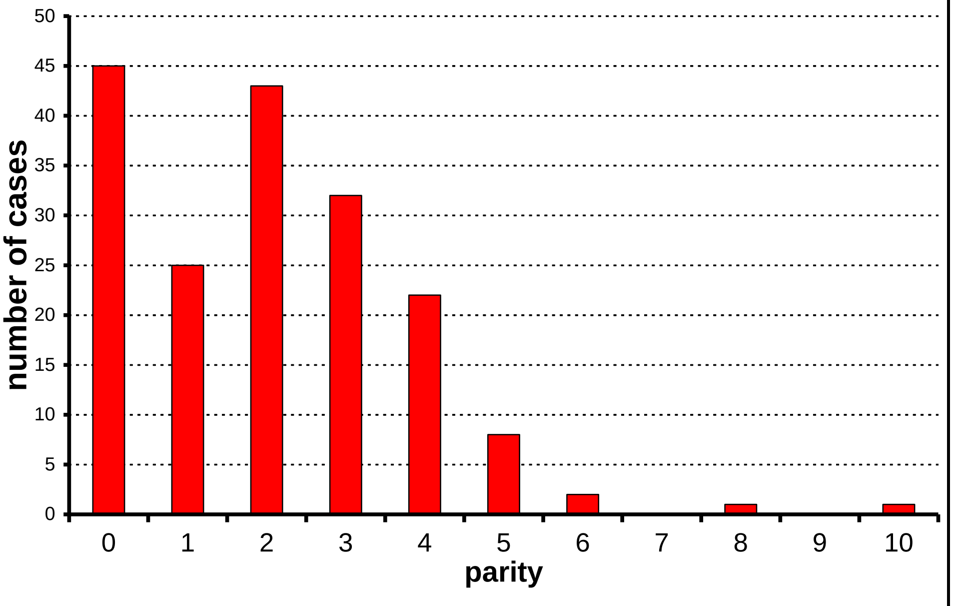
Sex	Number (absolute frequency)	Percent frequency
Men	33	26.4%
Women	92	73.6%
Total	125	100%



Frequency distribution of a quantitative variable (parity)

The table shows the resulting frequency distribution. Notice that we listed *all* values of parity between the lowest and highest observed, even though there were no cases for some values. Notice also that each column is properly labeled, and that the total is given in the bottom row.

parity	n° of cases	% frequency	cumulative freq.	cum. % freq.
0	45	25,1%	45	25,1%
1	25	14,0%	70	39,1%
2	43	24,0%	113	63,1%
3	32	17,9%	145	81,0%
4	22	12,3%	167	93,3%
5	8	4,5%	175	97,8%
6	2	1,1%	177	98,9%
7	0	0,0%	177	98,9%
8	1	0,6%	178	99,4%
9	0	0,0%	178	99,4%
10	1	0,6%	179	100,0%
total	179	100,0%		



Introduction to Frequency Measures

In epidemiology, many nominal variables have only two possible categories: alive or dead; case or control; exposed or unexposed; and so forth. Such variables are called **dichotomous variables**.

The frequency measures used with dichotomous variables are **ratios, proportions,** and **rates**. All these three measures are based on the same formula:

$$\textit{Ratio, proportion, rate} = (y/x) \times 10^n$$

In this formula, *x* and *y* are the two quantities that are being compared. The formula shows that *x* is divided by *y*. 10^n is a constant used to transform the result of the division into a uniform quantity. 10^n is read as "10 to the nth power." The size of 10^n may equal 1, 10, 100, 1000 and so on depending upon the value of *n*. For example,

$10^2 = 10 \times 10 = 100$

$10^3 = 10 \times 10 \times 10 = 1000$

$10^4 = 10 \times 10 \times 10 \times 10 = 10{,}000$

RATIO

PROPORTION	NOT a proportion
people with a given characteristic / WHOLE population	people with a given characteristic / people with another characteristic
Proportion of males = $\frac{\text{males}}{\text{all people}}$	males to females ratio = $\frac{\text{males}}{\text{females}}$
	Ageing index = ≥65yy / 0-14yy (165.3% in Italy in 2017)
	Total dependency ratio = (0-14 + ≥65) / 15-64yy (55.8% in Italy in 2017)
	Old-age dependency ratio = ≥65 / 15-64yy (34.8% in Italy in 2017)
range: 0 ⟷ 1	range: 0 ⟷ ∞

Example: During the first 9 months of national surveillance for eosinophilia-myalgia syndrome (EMS), CDC received 1,068 case reports which specified sex; 893 cases were in females, 175 in males. Calculate the female-to-male ratio for EMS.

1. Define *x* and *y*: *x* = cases in females, *y* = cases in males

2. Identify *x* and *y*: *x* = 893, *y* = 175

3. Set up the ratio *x/y*: 893/175

4. Reduce the fraction so that either *x* or *y* equals 1:
893/175 = 5.1 to 1

Thus, there were just over 5 female EMS patients for each male EMS patient reported to CDC.

Swygert LA, Maes EF, Sewell LE, et al. Eosinophiliamyalgia syndrome: Results of national surveillance. JAMA 1990;264:1698-1703.

Example - 2

Based on the data in the example above, we will demonstrate how to calculate the proportion of EMS cases that are male.

1. Define x and y: x = cases in males, y = all cases

2. Identify x and y: x = 175, y = 1,068

3. Set up the ratio x/y: 175/1,068

4. Reduce the fraction so that either x or y equals 1:
175/1,068 = 0.16/1 = 1/6.10

Thus, about one out of every 6 reported EMS cases were in males.

In the first example, we calculated the female-to-male ratio. In the second, we calculated the proportion of cases that were male.

The female-to-male ratio is not a proportion, since the numerator (females) is not included in the denominator (males), i.e., it is a ratio, but not a proportion.

The terms "ratio", "proportion", "rate" are often misused.

For instance, prevalence is a proportion, not a rate; however, the phrase "prevalence rates" appears 3,789 times in the Titles of abstracts of the current literature, according to the ISI Web of Knowledge (1990-October 2009).

Ratios, Proportions, and Rates Compared - 2

The third type of frequency measure used with dichotomous variables, **rate**, *is often a proportion*, with an added dimension: it measures the occurrence of an event in a population over time. The basic formula for a rate is as follows:

$$Rate = \frac{\text{number of cases occurring during a given time period}}{\text{population at risk during the same time period}} \times 10^n$$

This formula has three important aspects.

1) The persons in the denominator must reflect the population from which the cases in the numerator arose.
2) The counts in the numerator and denominator should cover the same time period.
3) In theory, the persons in the denominator must be "at risk" for the event, that is, it should have been possible for them to experience the event.

Frequency measures used in epidemiology (measures of occurrence)

The disease is a STATUS (condition)

Frequency of "being ill", defined at a particular point in time.

We take a "snapshot" of the studied population with regard to the disease

↓

Prevalence (P)

Proportions of persons in a population who have a particular disease at a specified point in time.

The disease is an EVENT:

Frequency of occurrence of NEW CASES over a SPECIFIED PERIOD of time.

The population has to be followed-up over time.

↓

Incidence (I)

Frequency with which NEW CASES occur in a population over a SPECIFIED PERIOD of time.

Proportion (0-1, 0-100%)
no time dimension
(no measure units) ← **DIMENSION** → time $^{-1}$
n / (person*years)

The disease as an event (incidence)

Fixed population (cohort)

A group of individuals:

1) identified as they experienced a common event at time zero (t_0, beginning of the study)
2) Followed-up over time

Example:

The 600 students attending the Nursery School at the University of Verona in the academic year 2004/05 are followed-up till December 2014 to evaluate the incidence of occupational diseases.

Dynamic population

A group of individuals belonging to a same community.

In this population there is a turn-over, as people can both enter and exit the population.

Example:

The incidence of occupational diseases is evaluated from 2004/05 to 2014/15 among students attending the Nursery School of the University of Verona. Students are followed-up when attending the University but not thereafter.

Incidence Rates

Incidence rates are the most common way of measuring and comparing the frequency of disease in populations. Incidence rates are used instead of raw numbers for comparing disease occurrence in different populations because rates adjust for differences in population sizes.

Incidence is a measure of the frequency with which an event, such as a new case of illness, occurs in a population over a period of time.

Since incidence is a measure of risk, when one population has a higher incidence of disease than another, the first population is said to be at a higher risk of developing disease than the second, all other factors being equal.

CUMULATIVE INCIDENCE (CI):
probability (risk) that an healthy subject has to develop the disease during a specified period of time.

In t_0 : number of subjects under observation = 9
number of healthy subjects = 6

Between t_0 and t_1: number of subjects that developed the disease = 4
$\Rightarrow$ Cumulative Incidence = 4/6 = 0.67 between t_0 and t_1

$$\text{Cum.Incidence} = \frac{\text{new cases occurring during a given time period}}{\text{population at risk during the same time period}} \times 10^n$$

Example:
In a study on the relationship between oral contraceptives and development of bacteriuria, 2390 healthy women aged between 16 and 45 years, were followed-up for 3 years. 486 of these women used oral contraceptive on the 1st of January 1973. Between 1973 and 1976, 27 of these women developed the disease.

$$\text{Cumulative Incidence} = \frac{27}{486} = 0.056 = 5.6\%$$

Probability that a woman (aged 16-45 years) taking oral contraceptives develops an urinary tract infection during a three-years period.

***NB:* 5.6% in 3 years ≠ 5.6% in 3 months ≠ 5.6% in 10 years**

BUT…

- **subjects can join the study at different points in time**
- **some subjects are lost to follow-up**

SINCE ...

- **a subject is no longer at risk to develop the disease when he actually develops the disease**

PERSON-TIME:
Sum of all the observation times of people at risk.

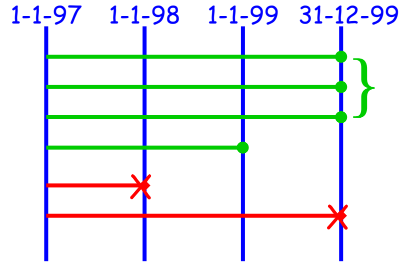
Person-time

Person-time is the denominator of incidence rate. Typically, each person is observed from a specified beginning point to an established end point (onset of disease, death, migration out of the study, or end of the study). The denominator is the sum of the time each person is observed, totaled for all persons.

Incidence rate (Mortality Rate) - 1

Population = a cohort (population) of 6 diabetic patients, followed-up between 1-1-1997 and 31-12-1999.

Event = death.



3 patients are live on 31-12-99

1 emigrates to Brazil on 31-12-98

1 patient dies on 31-12-97

1 patient dies on il 31-12-99

● = ~~event~~ (alive) → withdrawn alive at the end of the study
→ lost during the follow-up

X = event (death)

$$\text{Incidence} = \frac{\text{Number of events}}{\text{Sum of observation times}} = \frac{2}{3+3+3+2+1+3 \text{ yrs}} = \frac{2}{15 \text{ yrs}} =$$

$$= \frac{0.133}{1 \text{ year}} = \frac{133 \text{ deceased}}{1000 \text{ person*years}}$$

An incidence of 133 deceased / 1000 person·years is equivalent to:

133 deceased per 1000 persons in 1 year

13 deceased per 100 persons in 1 year

1,333 deceased per 10,000 persons in 1 year

133 deceased per 10,000 persons in 1/10 of year

Incidence Rate (Mortality rate) - 2

Population = citizens of a town in the Far-West (Tombstone) between l'1-1-1880 and l'1-1-1890 (dynamic population).

Event = death.

1-1-1880: 1000 inhabitants — 1000 deaths — 1-1-1890: 5000 inhabitants → time

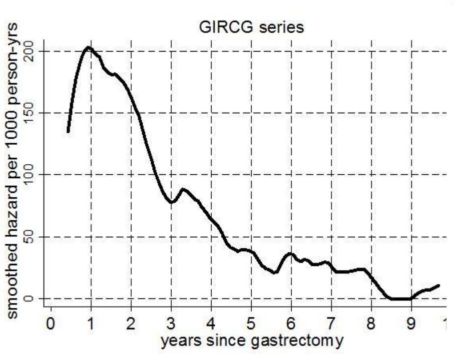
$$\text{Incidence} = \frac{\text{Number of events}}{\text{average population* observation time}}$$

$$\text{average population} = \frac{\text{(starting population) + (final population)}}{2} = \frac{1000 + 5000}{2} = 3000$$

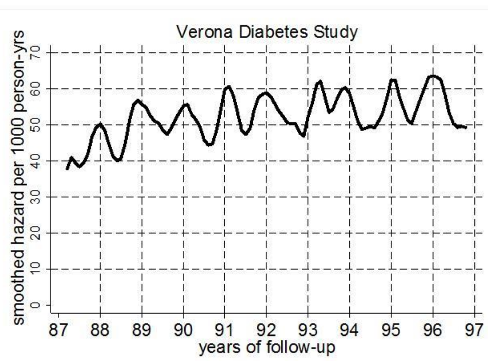
$$\text{Incidence} = \frac{\text{1000 events}}{\text{(3000) * (10 years)}} = \frac{\text{1 event}}{\text{30 person·years}} = \frac{\text{33.3 deceased}}{\text{1000 person·years}}$$

Crude mortality rates

Mortality in gastric cancer



Mortality in type 2 diabetes



Verlato et al, World J Gastroenterol 2014

Pay attention to the denominator !

• the **unity of time is arbitrary**: the rate could be expressed in days^{-1}, weeks^{-1}, month^{-1}, years^{-1}, ...

100 cases / 1000 persons • years =

= 10 000 cases / 1000 persons • centuries =

= 8.33 cases / 1000 persons • months =

= 1.92 cases / 1000 persons • weeks =

= 0.27 cases / 1000 persons • days

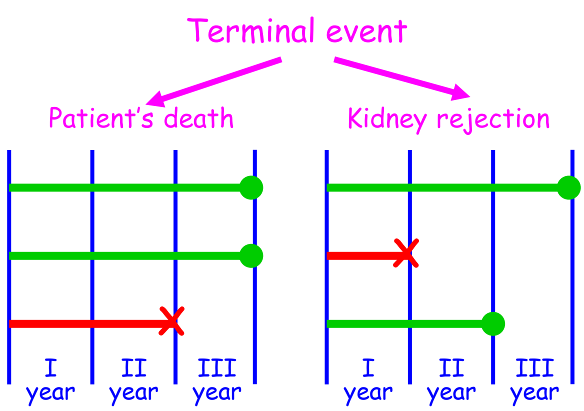
Pay attention to the terminal event !

subjects: a cohort of 3 patients receiving a kidney transplantation

everything OK after 3 years

Kidney rejected after 1 year, patient now on dialysis

deceased by car accident after 2 years, no kidney rejection



Incidence

$$\frac{1 \text{ event}}{8 \text{ person}\cdot\text{years}} \qquad \frac{1 \text{ event}}{6 \text{ person}\cdot\text{years}}$$

Example of incidence rate (fixed cohort)

1000 students enter a three-year University Course .
In those three years 200 students retire, while 800 graduate.
Which is the incidence of the event "retirement" in this population?

$$\text{Incidence} = \frac{\text{Event number}}{\text{Average population * observation period}}$$

$$\text{Incidence} = \frac{\text{200 events}}{\text{(1000 students) * (3 yrs)}} = \frac{\text{0.0667 events}}{\text{1 person·year}} = \frac{\text{66.7 events}}{\text{1000 person·years}}$$

This is a rough calculation: it doesn't consider that the retired students remain in the study less than 3 years.

We assume that the retired students have an average period of observation of 1.5 years.

$$\text{Incidence} = \frac{\text{200 events}}{800*3 + 200*1.5} = \frac{\text{200 events}}{\text{2700 person·years}} = \frac{\text{74.1 events}}{\text{1000 person·years}}$$

Comparison of cumulative incidence and incidence rate

When the denominator is the size of the population at the start of the time period, the measure is called **cumulative incidence**. This measure is a proportion, because all persons in the numerator are also in the denominator. It is a measure of the **probability** or **risk** of disease, i.e., what proportion of the population will develop illness during the specified time period.

In contrast, the **incidence rate** is like velocity or speed measured in miles per hour. It indicates *how quickly* people become ill measured in people per year.

Depending on the circumstances, the most appropriate denominator will be one of the following:

- average size of the population over the time period
- average of the population size at the start and end of the time period
- size of the population at the middle of the time period
- size of the population at the start of the time period

For 10^n, any value of *n* can be used. For most nationally notifiable diseases, a value of 100,000 or 10^5 is used for 10^n. Otherwise, we usually select a value for 10^n so that the smallest rate calculated in a series yields a small whole number (for example, 4.2/100, not 0.42/1,000; 9.6/100,000, not 0.96/1,000,000).

Since any value of *n* is possible, the investigator should clearly indicate which value is being used.

PREVALENCE

PROPORTION of persons in a population, who have a particular disease or attribute at a specified POINT in time or over a specified PERIOD of time.

$$P = \frac{\text{All (new and pre-existing) cases at a specified point in time}}{\text{Total population (healthy + diseased)}}$$

$$P = \frac{\text{Persons having a particular attribute at a specified point in time}}{\text{Total population (healthy + diseased)}}$$

The value of 10^n is usually 1 or 100 for common attributes. The value of 10^n may be 1,000, 100,000, or even 1,000,000 for rare traits and for most diseases.

PREVALENCE - EXAMPLES

2477 individuals aged 52-85 years
310 suffering from cataract
Which is the prevalence of cataract in this population ?

$$P = \frac{310}{2477} = 0.125 = 12.5\%$$

In 1986 in Verona there were 7488 diabetic patients out of an overall populaton of 301,519 inhabitants. Which is the prevalence of diabetes in this population?

$$P = \frac{7488}{301519} = 0.0248 = 2.48\%$$

Muggeo M, Verlato G, et al (1995) The Verona Diabetes Study: a population-based survey on known diabetes mellitus prevalence and 5-year all-cause mortality. *Diabetologia*, 38: 318-325

Prevalence at a given point in time = point-prevalence

Point prevalence is perfect from a theoretical point of view, but it is rather difficult to compute in practice.

Hence prevalence is usually computed over a given time period.

Prevalence computed over a specified period of time	Time unit
one-day prevalence	**one day**
one-week prevalence	**one week**
one-month prevalence	**one month**
one-year prevalence	**one year**
life prevalence	**the entire life**

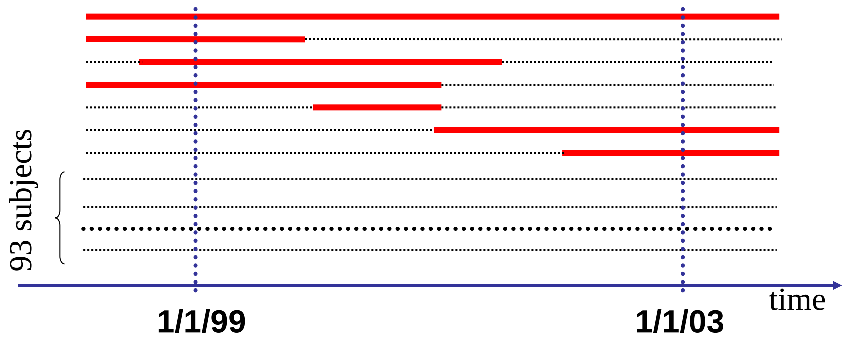
Life-prevalence: the numerator includes all the subjects who suffered from the disease at least once in their life.

Exercise 1:

1/1/1999: 4 asthma cases in a population of 100 subjects

1/1/1999 - 1/1/2003: 3 subjects recovered
1 healthy subject developed the disease and then recovered
2 healthy subjects developed the disease without recovering

A) What is asthma prevalence on the 1st of January 1999? 4 / 100 = 4%
B) What is asthma prevalence on the 1st of January 2003? 3 / 100 = 3%
C) Which is the prevalence between 1/1/1999 and 1/1/2003? 7 / 100 = 7%



Comparison of prevalence and incidence

The prevalence and incidence of disease differ both in the **numerator** and **denominator**.

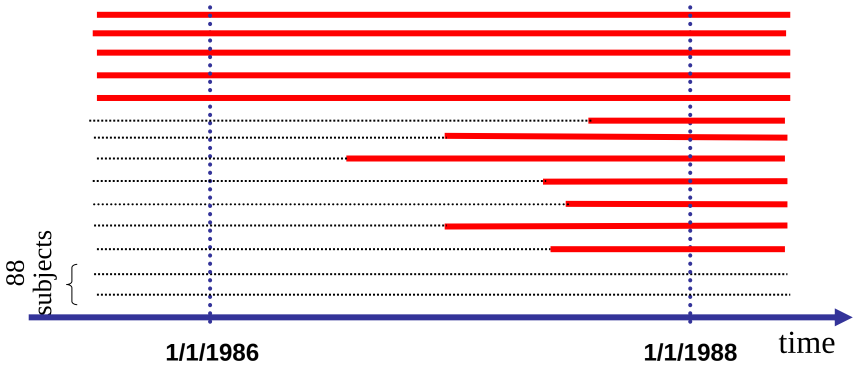
Numerator of Incidence = new cases occurring during a given time period
Numerator of Prevalence = all cases present during a given time period

The numerator of an incidence rate consists only of persons whose illness began during a specified interval. The numerator for prevalence includes **all** persons ill from a specified cause during a specified interval (or at a specified point in time) **regardless of when the illness began**. It includes not only new cases, but also old cases representing persons who remained ill during some portion of the specified interval. A case is counted in prevalence until death or recovery occurs.

The **denominator** of incidence includes only subjects "at risk" for a disease, hence subjects who already have the disease at the beginning of the study are excluded. The denominator for prevalence includes all subjects, either with or without the disease.

Exercise 2:
1/1/1986: 5 cases of angina pectoris in a population of 100 subjects
1/1/1986-1/1/1988: 7 new cases of angina pectoris

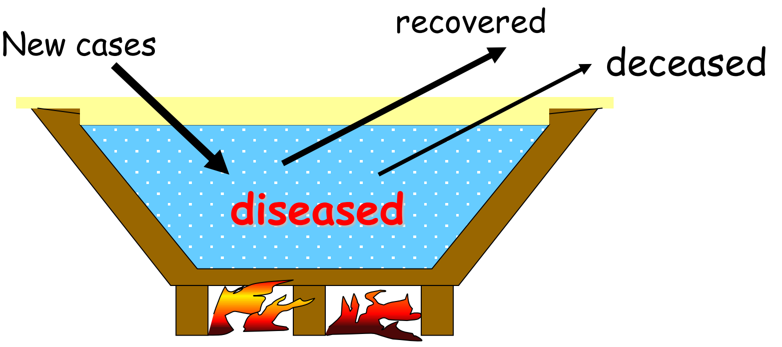
A) Which is the two-year prevalence of angina pectoris? 12 / 100
B) Which is the cumulative incidence in 2 years? 7 / 95 = 7.4%



Prevalence is based on both incidence (risk) and duration of disease. High prevalence of a disease within a population may reflect high risk, or it may reflect prolonged survival without recovery. Conversely, low prevalence may indicate low incidence, a rapidly fatal process, or rapid recovery.

Prevalence is often used rather than incidence to measure the occurrence of chronic diseases such as diabetes or chronic obstructive pulmonary disease, which have long duration and dates of onset which are difficult to pinpoint.

Relation between incidence and prevalence



Prevalence ≈ incidence * duration

with P<0.1

Prevalence = incidence * duration

(2 / 100 000 persons*year) * 5 years

10 / 100 000

Duration ?

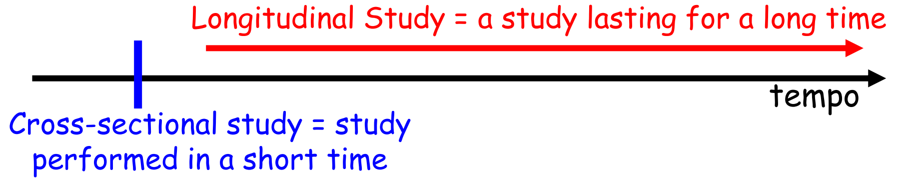
Duration = prevalence / incidence

= (10 / 100 000) / (2 / 100 000 years)

= (10 / 100 000) * (100 000 years / 2)

10 years / 2 = 5 years

Disease	Prevalence	Most suited studies
Acute disease (infectious disease)	Large variability: High during epidemics, otherwise close to null	**Longitudinal studies (mandatory reports to health authorities)**
Diseases with high fatality rate (cancer)	Low prevalence	**Longitudinal studies (cancer registries)**
Chronic-degenerative diseases	Long duration, hence high prevalence	**Cross-sectional studies**



Chronic-degenerative diseases = coronary heart diseases, cerebrovascular diseases, chronic obstructive pulmonary diseases, diabetes mellitus, osteoarthrosis

Prevalence of gastric cancer in Italy

In Italy on the January the 1° 2006:

58,751,711 inhabitants

2,243,953 (3.82%) people with cancer diagnosis

- 69,225 (0.12%) with diagnosed gastric cancer
- 0.11% in Veneto region, 0.31% in Romagna

AIRTUM Working Group. Italian cancer figures – Report 2010. Cancer prevalence in Italy. Epidemiol Prev 2010; 34 (5-6) suppl 2.

Number needed to follow during the lifespan (from 0 to 84 years) in order to observe one cancer case, as a function of sex and tumore site. Pool AIRTUM 2008-2013.

	Males	Females
Prostate	**8**	-----
Lung	**10**	36
Colon/rectum	**11**	18
Bladder	**14**	77
Stomach	**32**	65
Liver	**33**	89
Kidney/pelvis/ureter	**39**	90
Mouth/Pharynx/Larynx	**41**	182
All tumours	2	3

	Males	Females
Breast	598	**8**
Colon/rectum	11	**18**
Lung	10	**36**
Uterus (body)	-----	**47**
Thyroid	130	**49**
Non-Hodgkin lymphoma	44	**62**
Stomach	32	**65**
Pancreas	49	**65**
All tumours	2	3

I numeri del cancro in Italia 2017, a cura di AIOM, AIRTUM, Fondazione AIOM. Il Pensiero Scientifico Editore, Roma, 2017.

Case-Fatality Rate (cumulative)

$$\text{Case-fatality rate} = \frac{\text{Number of deaths due to a disease}}{\text{Number of people with the same disease}}$$

Example:

- 600 people have the disease
- 9 of them die from the disease
- 1 dies from a traffic accident

$$\text{Case-fatality rate} = \frac{9}{600} = 1.5\%$$

Case-Fatality Rate

$$\text{Incidence} = \frac{\text{Number of events}}{\text{Average population * observation time}}$$

$$\text{Case-fatality rate} = \frac{\text{Number of deaths from a disease}}{\text{Number of patients with that disease * observation time}}$$

Example: during a specified year there are on the average 30,000 diabetics and 100 patients with pancreatic cancer in a population of 1,000,000 inhabitants. During that year 600 persons died from diabetes and 80 from pancreatic cancer.

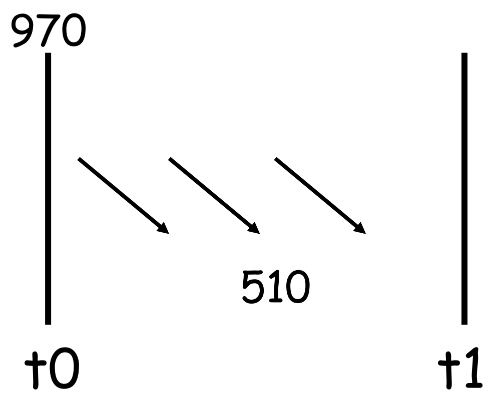
	Average prevalence	Cause-specific mortality	Case-fatality rate
diabetes	30,000 / 1,000,000 = 3%	600 / 1,000,000 = 6 / 10,000 p*yrs	600 / 30,000 = 20 / 1,000 p*yrs
Pancreatic cancer	100 / 1,000,000 = 0,01%	80 / 1,000,000 = 0,8 / 10,000 p*yrs	80 / 100 = 800 / 1000 p*yrs

IMPORTANT DEMOGRAPHIC INDEXES

		in Italy
birth rate =	$\frac{\text{number of live births per year}}{\text{average population (in person*years)}}$	7.8 / 1000 p.y. in 2016
total fertility rate =	average number of children born to a woman over her lifetime according to current age-specific fertility rates	1.34 children per woman in 2016
life expectancy = (at birth)	Average number of years a newborn would survive if he/she experienced current age-specific mortality rates throughout his/her life	42 in M, 43 in F in 1899 80.6 in M, 85.1 in F in 2016 (*female survival advantage*)
mortality rate =	$\frac{\text{number of deaths per year}}{\text{average population (in person*years)}}$	10.1 / 1000 p.y. in 2016
infant mortality rate =	$\frac{\text{number of deaths of children <1 year of age}}{\text{total live births}}$	3.0 / 1000 in 2016
perinatal mortality rate =	$\frac{\text{number of stillbirths after 22 weeks of gestation and deaths in the first week of life}}{\text{total births (stillbirths + life births)}}$	

SOLUTIONS TO TEST 1

Computing prevalence



Prevalence at t0

$$\frac{970}{100\ 000} = 0.97\%$$

One-year prevalence

$$\frac{970+510}{100\ 000} = 1.48\%$$

Computing incidence

Population at risk

99030 (t0) — 510 — 98547 (t1)

100,000-970 = 99030

970+510-27=1453
100000-1453 = 98547

Average population at risk = (99030+98547) / 2 = 98778.5

$$\frac{510}{98778.5 \text{ persons} * 1 \text{ year}} = \frac{5.16}{1000 \text{ person*years}}$$

MULTIPLE SCLEROSIS

$$\frac{55}{\cancel{100\,000}} \times \frac{\cancel{100\,000}\ yy}{5} = 11 \text{ years}$$

MOTONEURON DISEASE

$$\frac{7}{\cancel{100\,000}} \times \frac{\cancel{100\,000}\ yy}{1.7} = 4.12 \text{ years}$$

	Old town centre	Suburb
≤65 years	5/440 = 1.2%	70/6311= 1.1%
>65 years	310/3617 = 8.6%	72/717 = 10.0 %
Total	315/4057 = 7.8%	142/7028 = 2.0 %