Measures of Disease Occurrence

- Population and period of observation
- Ratios, Proportions and Rates
- Prevalence and Incidence
- Standardized Rates

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Descriptive Epidemiology: *who*, *where*, and *when*

Person, **place**, and **time**, are the key components of **descriptive** epidemiology

The aim is to describe the **occurrence of disease** in **groups** of people, so **populations*** are at the heart of epidemiologists' measurements.

Population: group of people with a **common** characteristic; place of residence, religion, gender, age, use of hospital services, or life event (such as giving birth)…

Type of population	Key element	Example
Fixed/Closed	Membership is based on an *event* and is **permanent**	Japanese atomic bomb survivors
Dynamic/Open	Membership is based on a *condition* and is **transitory**	Residents of a city, hospital patients

* for this lesson we refer to ***population*** as if we were able to measure ***all*** the members

It is important here to make a **distinction** between :

- The «common» event/condition that define the **membership** to a population
- The event/condition that we aim to describe in the population (the «**outcome**»)
- The **time period** of observation (= data collection / **follow-up**)

Population 1: People living in Trieste as of January 1, 2024

Outcome 1: Incidence of influenza during the flu season

Time Period 1: January – March 2024.

Population 2 : Subject diagnosed with hypertension in Friuli-Venezia-Giulia [from 2020 to 2023]

Outcome 2: Hospitalization for stroke after the diagnosis

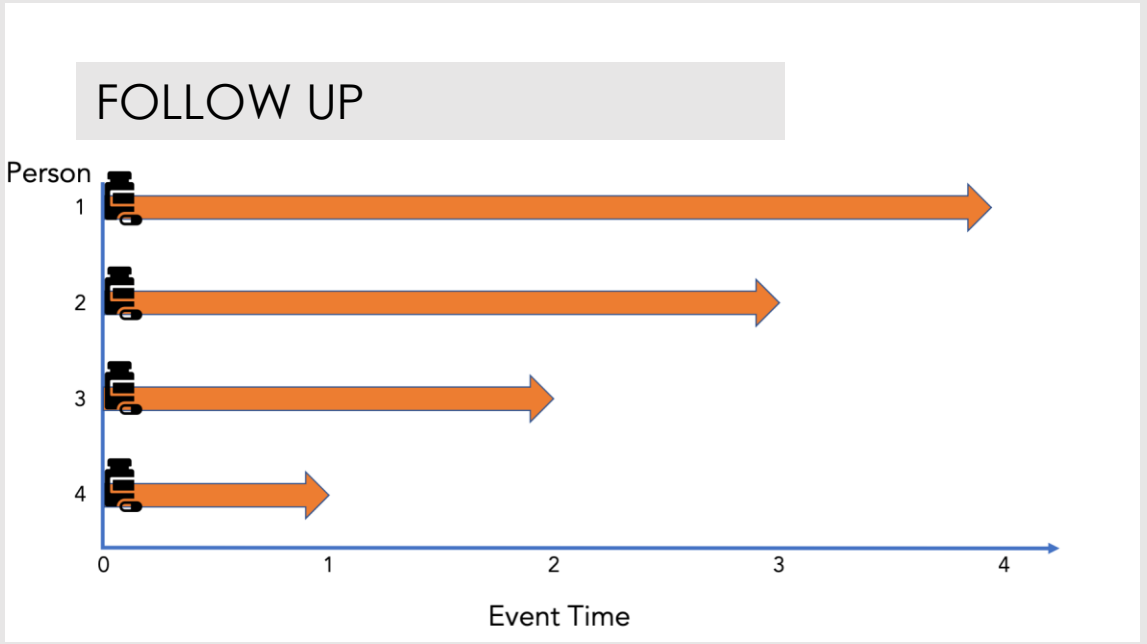
Time Period 2: January 2020 – December 2024.

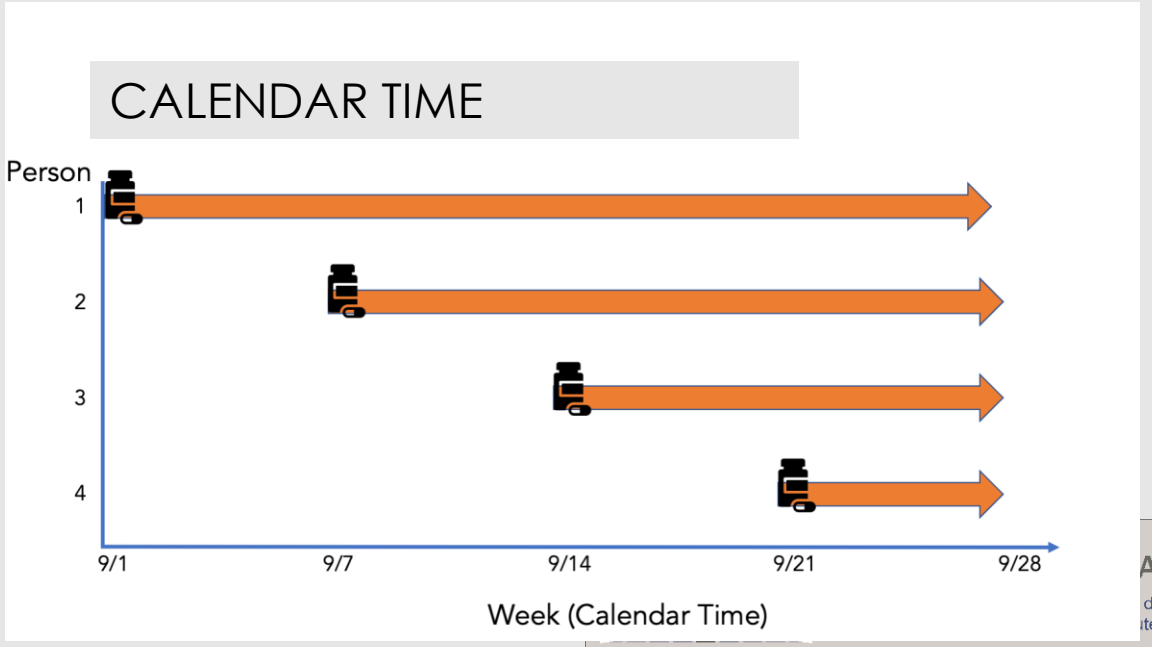
A **fixed/closed** population *adds no new members* over time (may *lose* someone to the observation)

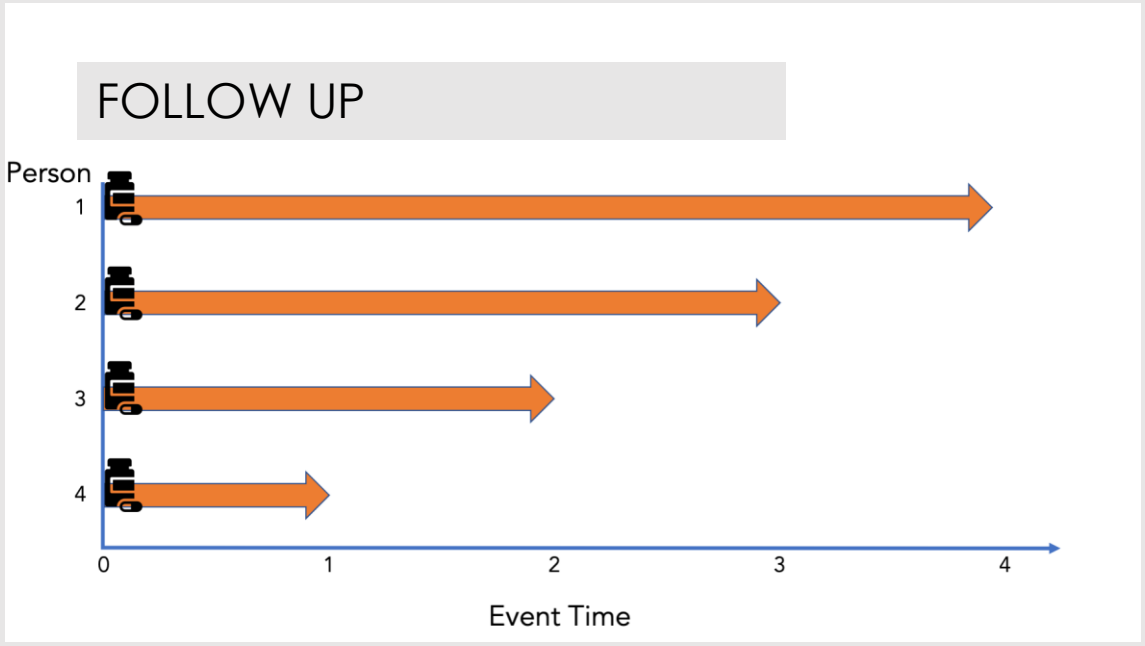
A **dynamic/open** population may *gain members over time* [immigration…birth] or *lose* members who are still alive […emigration].

The key distinction for fixed vs dynamic is on **how membership** is defined.

A more *subtle* concept relates also in how we express the **time** of observation:

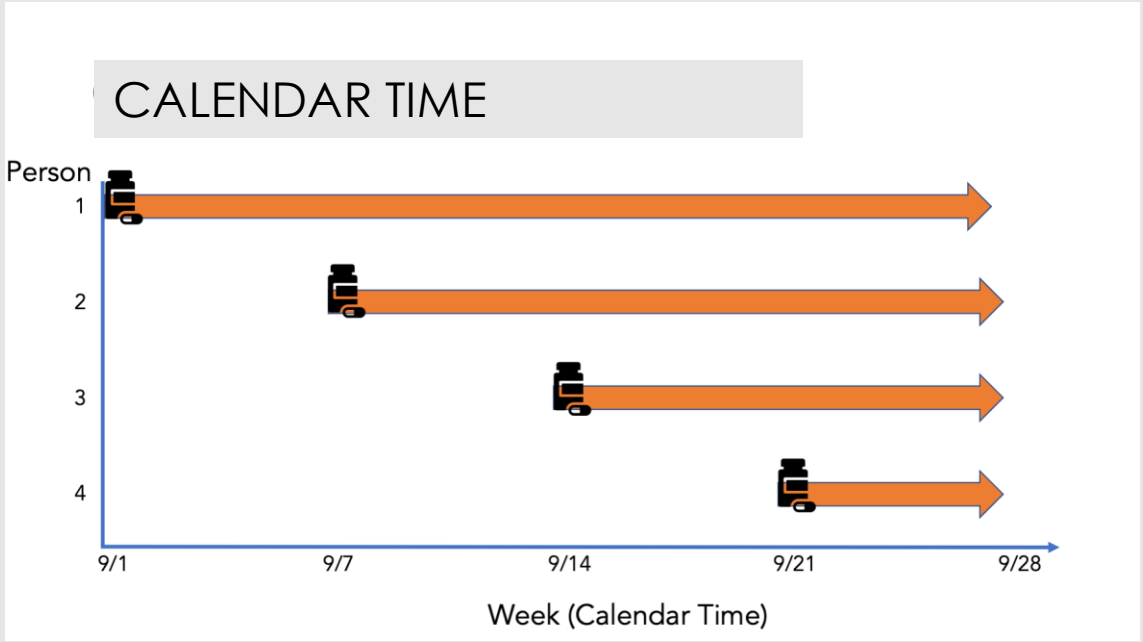






Time axis: follow up

All people who ***started* aspirin after a myocardial attack in the year 2020** would constitute a **closed** population: time is measured from ***start of their use*** of the drug (and observed thereafter for a time period, i.e. the successive month).



Time axis: calendar time

For 4 weeks in the month of September 2024 we observe **users** of aspirin. Membership is defined by ***taking Aspirin once a week***.

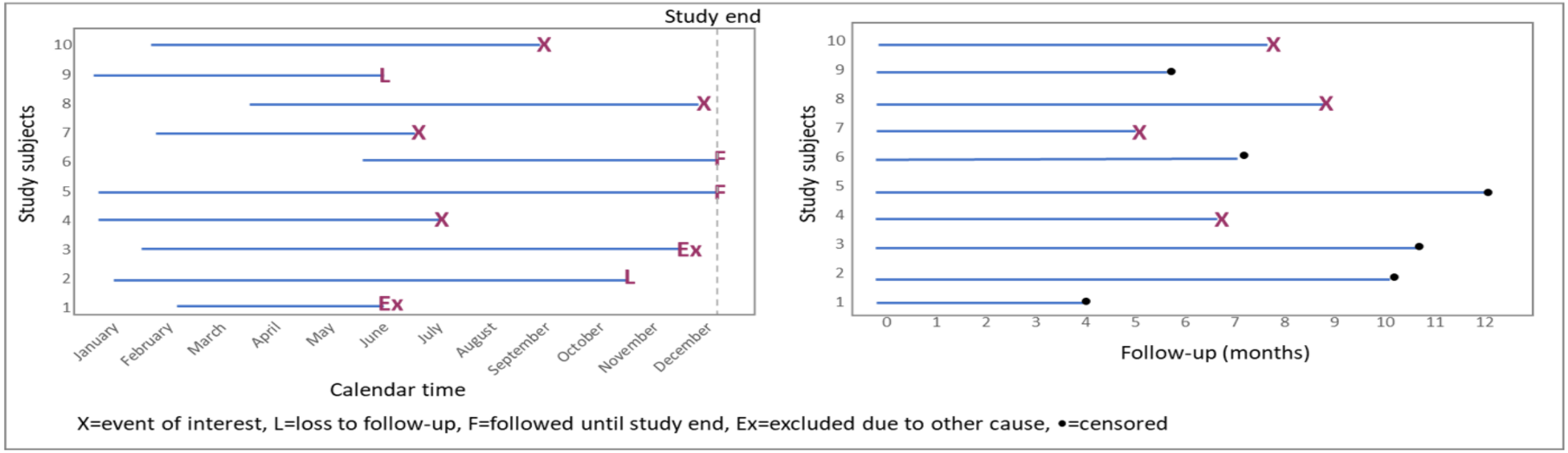
This is an **open** population in *calendar time*, because **new users** might accumulate during that month of observation.

POPULATION AS A ***COHORT***

The term **cohort** is also often used to indicate population in epidemiological studies

- ✓ members of the graduating class of a school in a given year
- ✓ A **birth cohort** is the cohort defined by being born at a particular time/place, e.g., all people born in Ethiopia in 1990 constitute the Ethiopian birth cohort for 1990.

CALENDAR TIME : «DATA COLLECTION» → **FOLLOW UP TIME: «DATA ANALYSIS»**



Once the **population** is defined, then:

3 factors to **quantitatively** measure how commonly a disease occurs in the population:

(1) the **number** of people who are affected by the disease
(2) the **size** of the population from which the cases of disease arise
(3) the **length of time** that the population is followed

County A:
population=50,000
100 new cases of breast cancer occurred over a 1-year period.

County B:
Population=5,000
75 new cases occurred over a 3-year period.

To compare, convert data into the **same** population size and time period:

County A:
200 cases/100,000 population/ 1 year.

County B:
75 cases/5,000 population/ 3 year=
25 cases/5,000 population/ 1 year=
500 cases/100,000 population / 1 year

Ratios, Proportions & Rates

Ratio : one number divided by another. The entities represented by the two numbers do not have to be related to one another. The individuals in the numerator can be **different** from those in the denominator.

Gender ratio is a ratio of two unrelated numbers: the number of males divided by the number of females, usually expressed as the number of males per 100 females.

Ex: according to 2016 U.S. Census estimates, the gender ratio among U.S. residents in Florida was 95.5 males per 100 females.

A **proportion** : one number divided by another, but the entities represented by these numbers are related to one another. The numerator of a proportion is always a **subset** of the denominator.

Ex: according to 2016 U.S. Census estimates, the proportion of Black U.S. residents was 0.141, or 14.1%.

Rate : one number divided by another, but **time** is an integral part of the denominator.

We are familiar with rates in our daily travels because a rate is a measure of how fast we travel.

The maximum speed or rate at which cars are permitted to travel is 90 kilometres per hour.

[This rate can also be written as 55 miles / 1 hour (≈ 90 km / h)].

The measure of **time** in the denominator is what makes this number a rate.

The measures of disease occurrence calculated previously for Counties A and B are also rates.

County A:
200 cases/100,000 population/ 1 year

County B:
500 cases/100,000 population / 1 year

Measures of Disease Occurrence

The **two** basic measures of disease frequency in epidemiology are **incidence** and **prevalence**.

Incidence measures the occurrence of **new** disease
Prevalence measures the existence of **current** disease

Incidence: the occurrence of **new** cases that develop in a **candidate population** over a specified **time** period.

New disease events: for diseases that can occur more than once, it usually measures the **first occurrence** of the disease

Candidate population: a population of people who are "**at risk**" of getting the disease.

Time period ➡ two different measures
cumulative *incidence*
incidence ***rate***

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Cumulative incidence: % of ***at risk*** population that becomes diseased ***over a specified period of time***.

$$\frac{\#\ New\ Cases}{\#\ At\ Risk}$$

[observed over a *certain period of time*]

Time is not an integral part of this proportion but rather is *expressed by the words* that accompany the numbers of the cumulative incidence measure.

Cumulative incidence can be thought of as the **average risk** of getting a disease over a certain period of time.

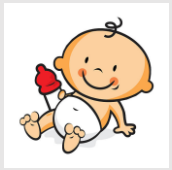
Example: **lifetime risk** of breast cancer among women

1 in 8 among U.S. women (about 12%) will develop breast cancer sometime during the course of their lives.

Cumulative Incidence: probability related to a *time interval*

Cumulative incidence (CI) is obviously influenced by the **length of time** to which it applies.

CI is mainly used in **fixed/closed** **populations** when there are no or *small* losses to follow-up (and *possibly* all subjects are potentially observed ***for the same amount*** of time).



Cumulative incidence of **pre-term birth** (< 37 weeks) is the number of pre-term births divided by the number of births over a (potential) follow up of 9 months (the same for all pregnant women).

Incidence rate : the occurrence of **new** cases of disease that arise during **person-time** of observation.

$$\frac{\#\ New\ Cases}{\#\ Person - Time}$$

The incidence rate's denominator **integrates time** and therefore is a true rate.

Person-time is only accrued ***among candidates*** for the disease. Could be applied to open/dynamic populations.

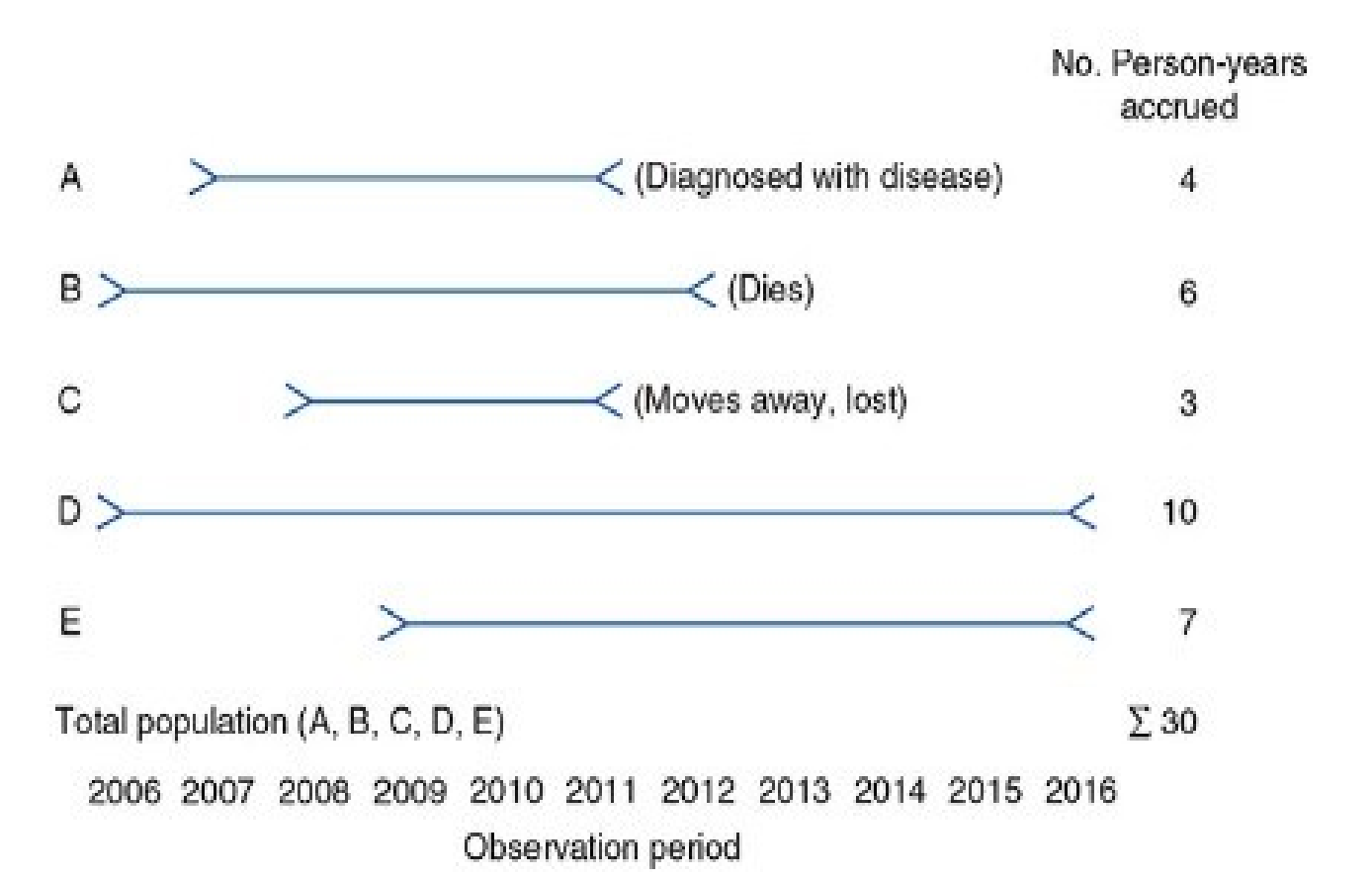
A person **contributes time to the denominator** of an incidence rate only up until he/she is diagnosed with the disease of interest.

Incidence rate (IR) is not based upon the assumption (as CI) that everyone in the candidate population has been followed for a *fixed pre-specified* time period.

Person-time is accrued only while the candidate ***is being followed***.

Accrual of person-time **stops** when the person dies or is lost to follow-up.

The incidence rate can be calculated for either a fixed or dynamic population.



IR : 1/30 person-years.

Regardless of how the person-time is accrued (e.g., from 5 x 6 years or 3 x 10 years), the **person-time units** are assumed to be *equivalent*.

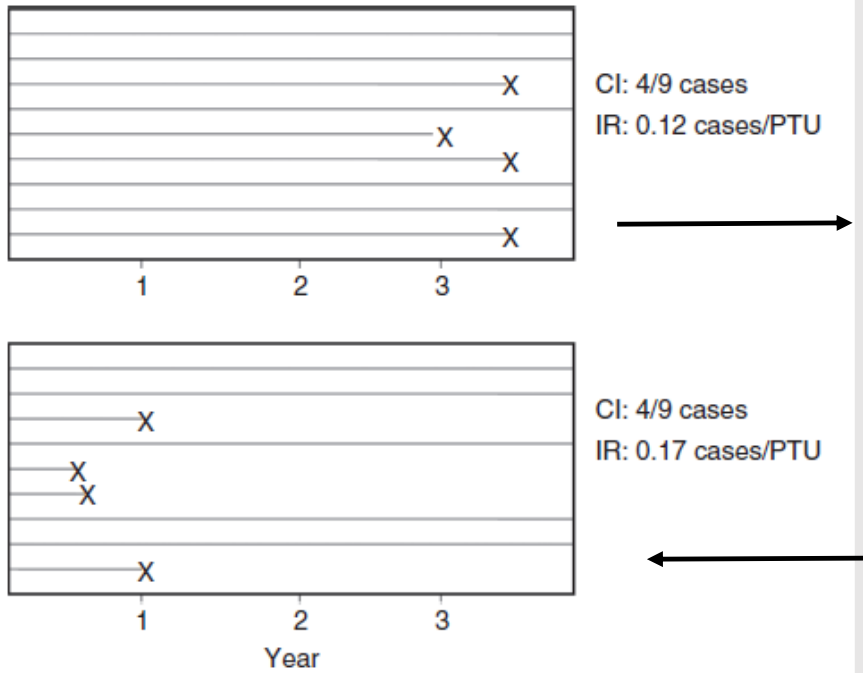
Because person-time is calculated for each subject, it can accommodate people coming into and leaving the study

CI and IR focus on measuring the ***transition*** from health to disease.

CI is easy to calculate and explain to the general audience

IR has greater accuracy, but its ***person-time*** denominator is more difficult to calculate and understand
[statisticians "prefer" survival analysis tools, block 4]

IR is more useful for **open/dynamic** populations, and CI is usually reserved for **fixed/closed** populations



$$PTU_1 = 4 + 4 + 3.5 + 4 + 3 + 3.5 + 4 + 4 + 3.5 = 33.5$$

$$IR_1 = \frac{4}{33.5} = 0.12$$

patients developed the disease *more slowly* than in the second example

$$PTU_2 = 4 + 4 + 1 + 4 + 0.5 + 0.5 + 4 + 4 + 1 = 23$$

$$IR_2 = \frac{4}{23} = 0.17$$

Prevalence:

Whereas incidence measures the frequency with which new disease develops, prevalence measures the frequency of ***existing*** disease.

It is simply defined as the **proportion** of the total population that is diseased.

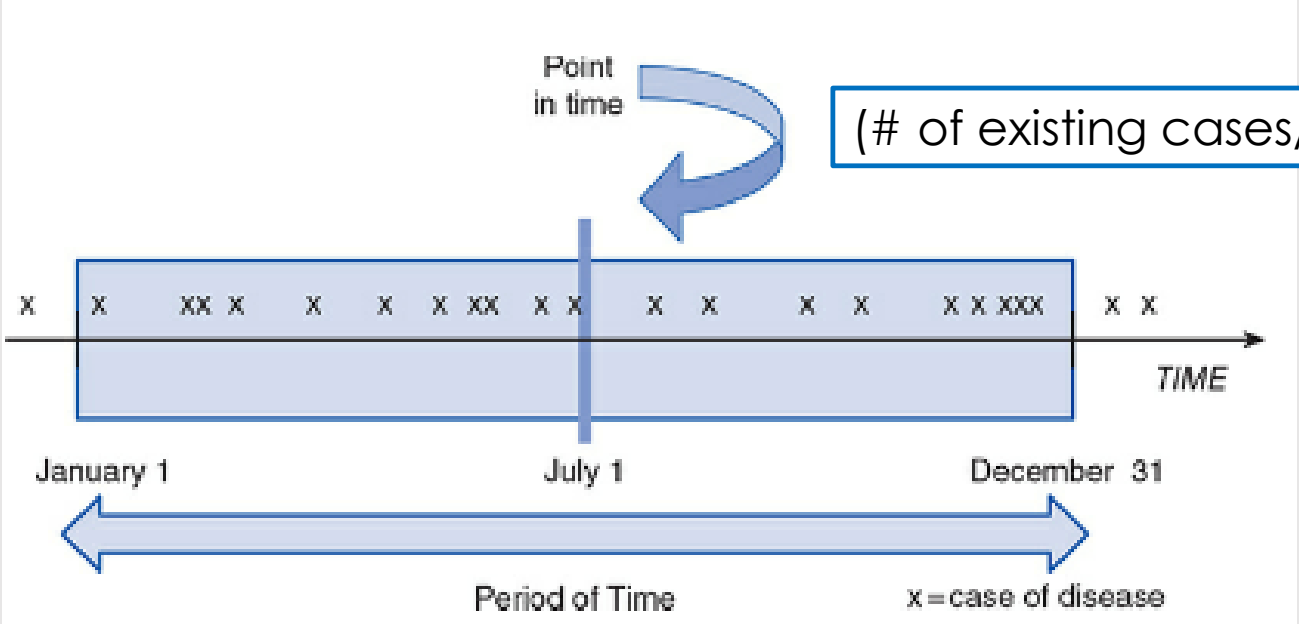
$$\frac{\#\ Cases}{\#\ People}$$

There are **two** types of prevalence measures - ***point*** prevalence and ***period*** prevalence - that relate prevalence to different amounts of time.

Point prevalence refers to the proportion of the population that is diseased at a **single point** in time and can be thought of as a **single snapshot** of the population.

The point can be either a particular calendar date such as July 1, 2017, or a point in someone's life, such as college graduation.

The **period** prevalence includes the number of cases that were present ***at any time*** over the course of a time-interval.



(# of existing cases/# in total population) *at point in time*

Ex: on July 1, 2017, there were 5 cases of pneumonia among the 500 nursing home residents.

The **point prevalence** of pneumonia was 5/500, or 1%.

(# of existing cases/# in total population) *during a period of time*

During the period January 1 through December 31, 2017, there were 45 cases of pneumonia among the 500 nursing home residents; the **period prevalence** was 45/500, or 9%, during the year.

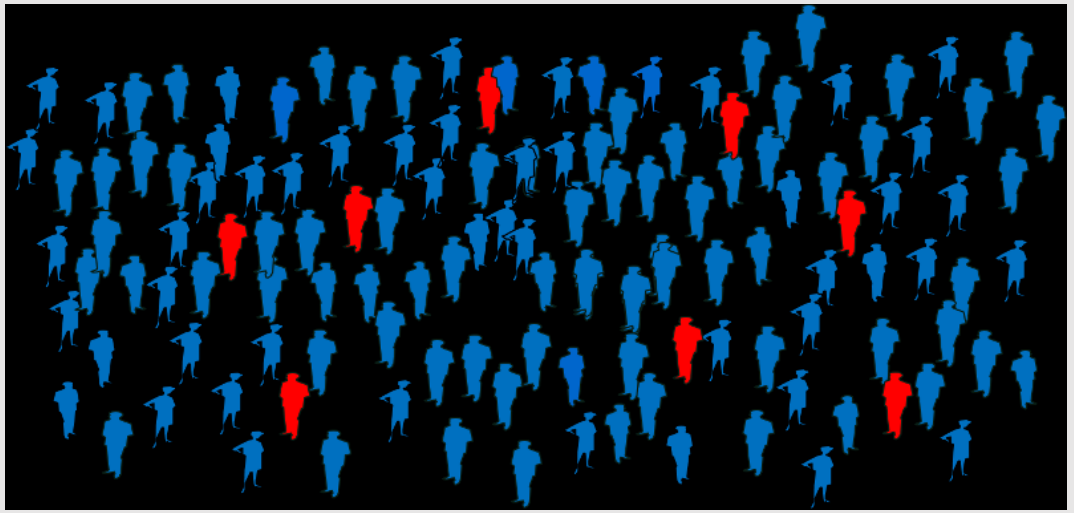
Of note: if the nursing home population had gained or lost members during the year, the ***average size*** of the nursing home population during 2017 would have been the appropriate denominator for the period prevalence measure (average or **mid-interval** population).

Note also that the numerator (existing cases) is a **subset** of the denominator (total population).

Unlike the numerator for the two incidence measures, the prevalence numerator includes **all currently living cases** regardless of when they first developed.

The denominator includes everyone in the population - sick, healthy, at risk, and not at risk.

Prevalence is a proportion, it is dimensionless, and its possible values range from 0 to 1, or 0% to 100%.



Just to fix :

Prevalence and incidence are sometimes **confused**.

Prevalence refers to proportion of people ***who have*** a condition at or during a particular time period, whereas incidence refers to the proportion or rate of people ***who develop*** a condition during a particular time period.

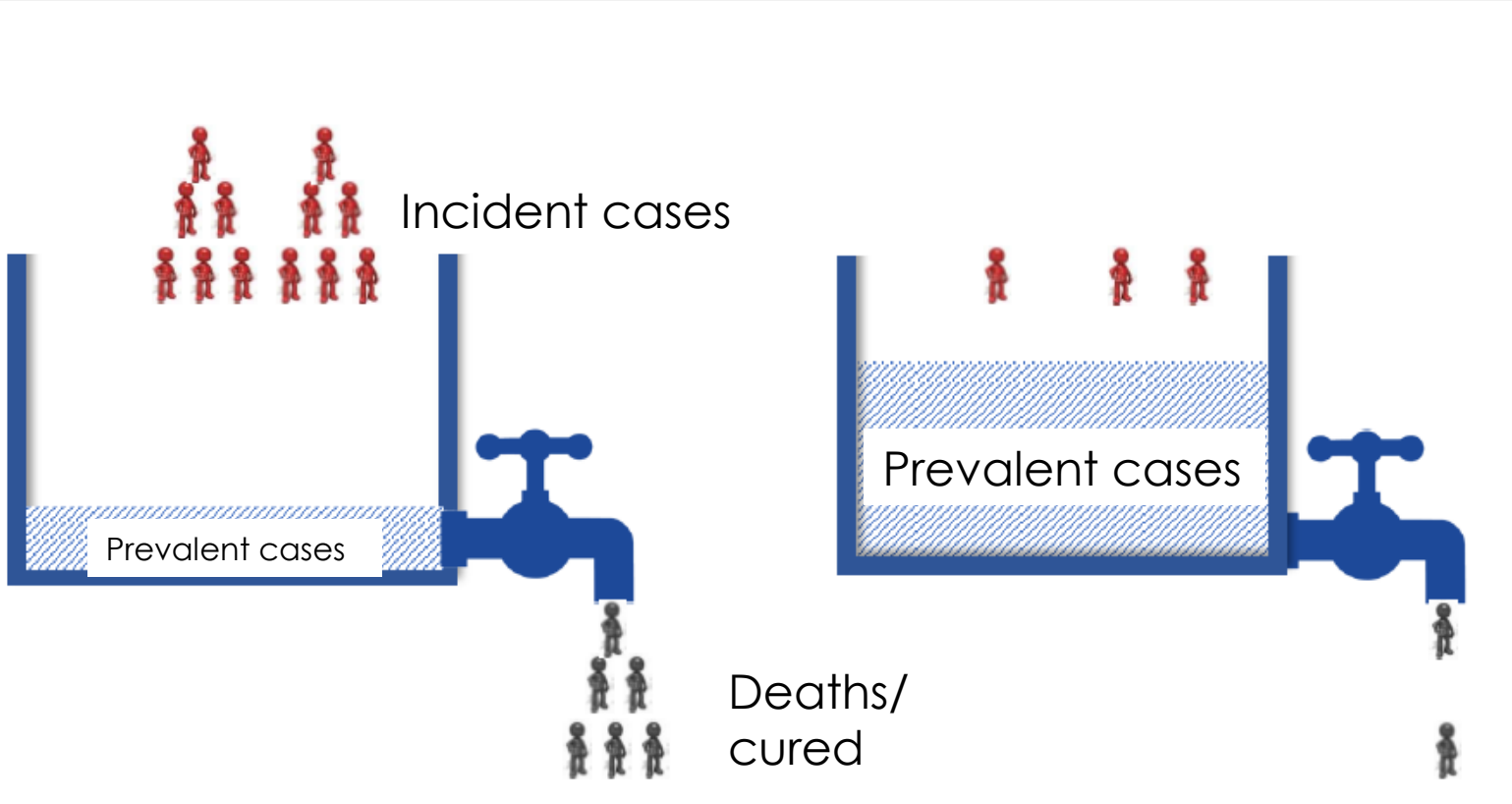
So prevalence and incidence are *similar*, but prevalence includes new and pre-existing cases whereas incidence includes new cases only.

The key difference is in their **numerators**.

Numerator of incidence = **new** cases that occurred during a given time period

Numerator of prevalence = **all** cases present during a given time period (regardless of when the illness began).

Prevalence (P) depends on the **rate** at which new cases of disease develop (IR) as well as the **duration** [average] or length of time that individuals have the disease (D).



$$P \approx IR * D$$

This equation assumes that the incidence rate and duration ***do not change*** over time.

Prevalence *obscures* causal relationships because it combines incidence and survival…

But, prevalence is useful for estimating the needs of medical facilities and allocating resources.

Moreover, for **chronic conditions** (whose beginnings are difficult to pinpoint) there is often no choice but to use prevalence.

Between 1973 and 1977 the incidence of lung cancer was 45.9 per 100,000, the *average* annual prevalence of 23 per 100,000. What was the *average duration* of the disease?

Duration = Prevalence/Incidence = 23/45,9 = 0.5 years = 6 months

High-incidence diseases can have a **low prevalence** if the average duration is **short**.

Example 2 :	
U.S. population at 1/7/1972	208.232.000
TBC cases at 1/1/1972	44.000
TBC cases «re-activated» during 1972	3500
New TBC cases in 1972	32882

$$CI_{72} = \frac{32882}{208232000 - 44000} = 0.000158$$

15.8 cases/100.000

$$Prev_{1-1-72} = \frac{44000}{208232000} = 0.00021$$

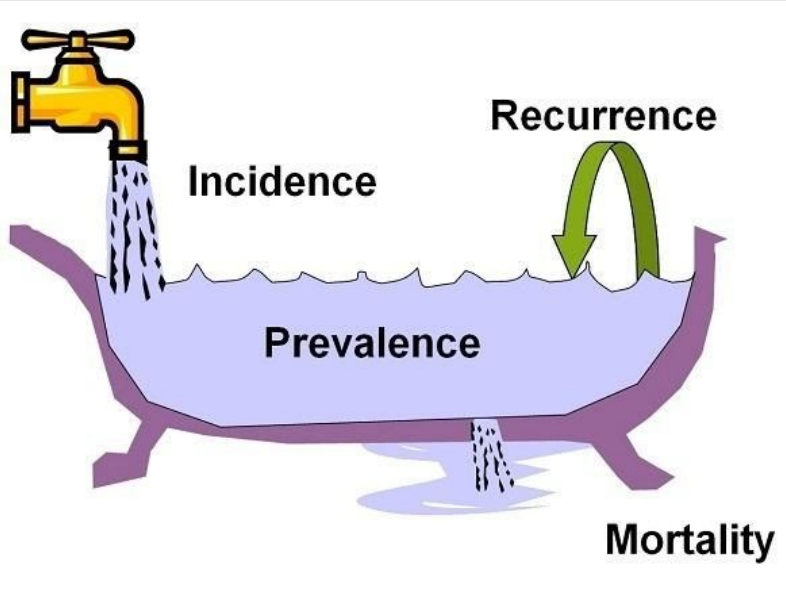
21 cases /100.000

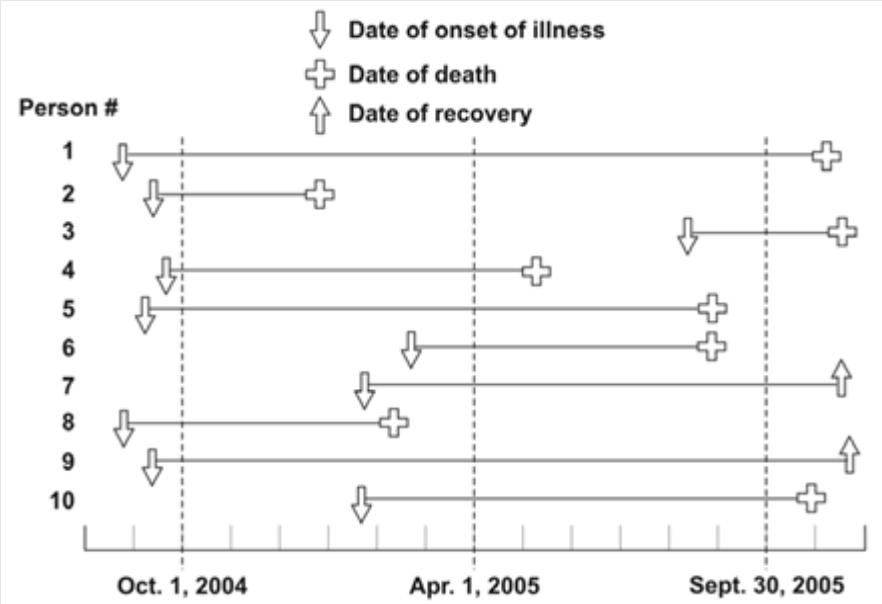
39 cases /100.000

$$Period\_Prev_{1972} = \frac{44000 + 3500 + 32882}{208232000} = 0.00039$$

Distinguishing Characteristics of Incidence and Prevalence

Measure	Type	Range	Num	Den	Major Use
Cumulative Incidence (CI)	Proportion	0 to 1	**New** cases	Population at risk	Causes, prevention, treatment
Incidence Rate (IR)	*True* rate	0 to infinity	**New** cases	Person-time at risk	Causes, prevention, treatment
Prevalence (P)	Proportion	0 to 1	**Existing** cases	Total population	Resource planning





It is possible to estimate (*approximately*) CI from IR:

Simplest situation : **closed** population with a **constant** IR in each time period (no **competing risk**, more on block 4)

$$CI = IR * t$$

This relationship would hold true if the population were infinitely large, but in a finite population this approximation becomes increasingly inaccurate over time, because the size of the population at risk **declines** over time.

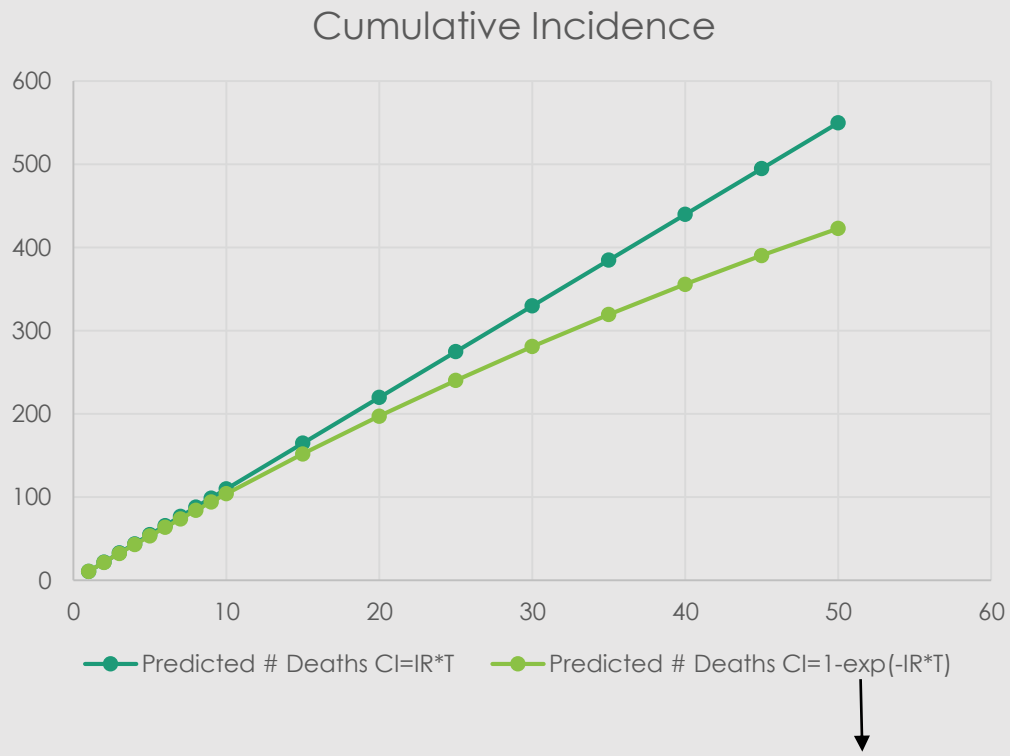
1000 people who experience a mortality rate of 11 deaths per 1000 person-years over a period of 50 years.

After 50 years the CI of death would be CI = IR X T = 11 X 50 = 550 deaths

In reality, there would only be 423 deaths after 50 years.

The **size** of the population ***at risk*** declines over time !!

After the first year there have been 11 deaths, and the population now has only 989 people, not 1000 (and so on..). **(block 4)**



exponential decay in the population at risk

When IR is not constant* it is necessary to take into account the different rates that prevail during each time period…

* Mortality rate among Hiroshima residents was much higher shortly after the atomic bomb explosion than during subsequent years

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

Standardized Rates

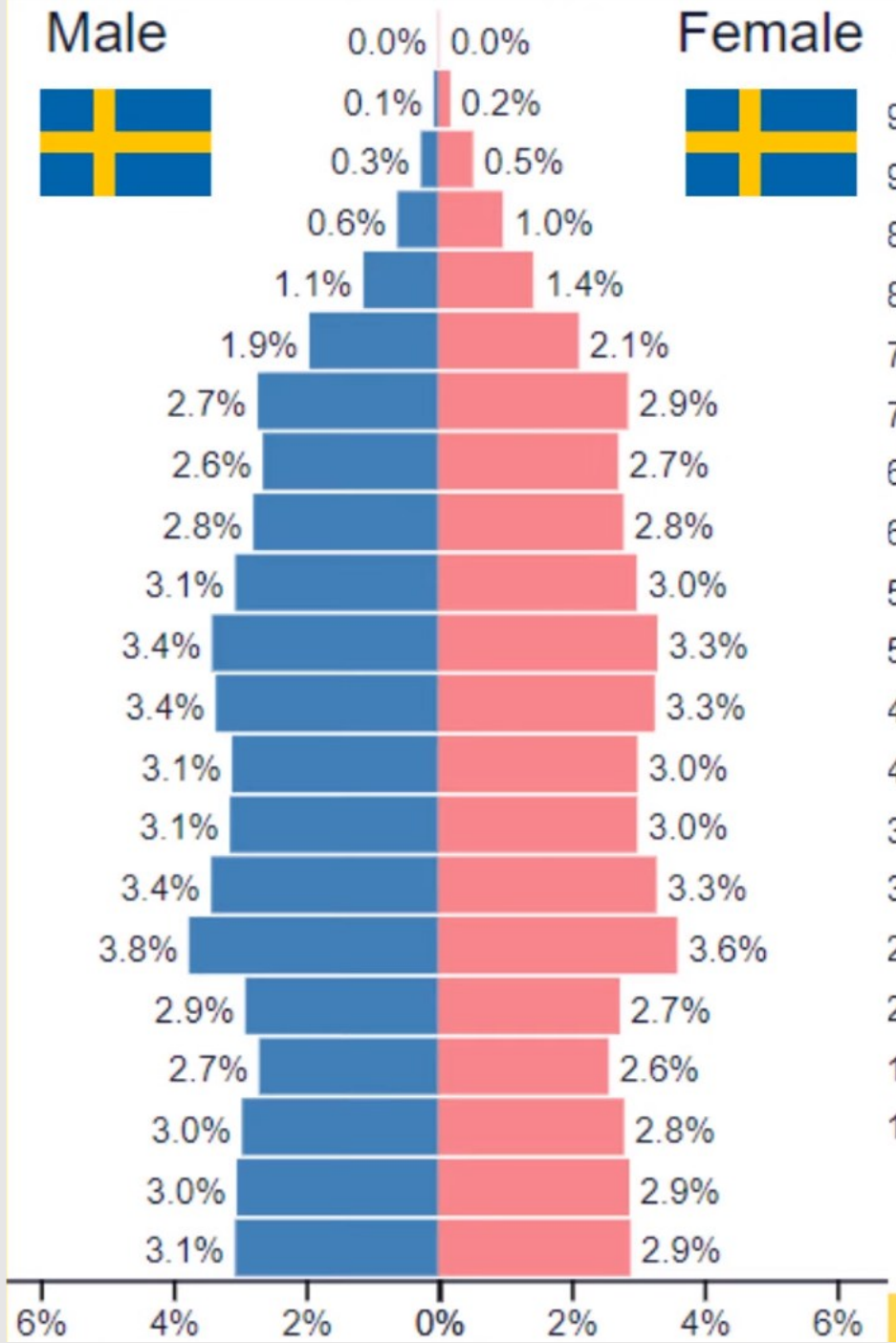
A principal role in **descriptive epidemiology** is to **compare** the incidence of disease or mortality between two or more populations.

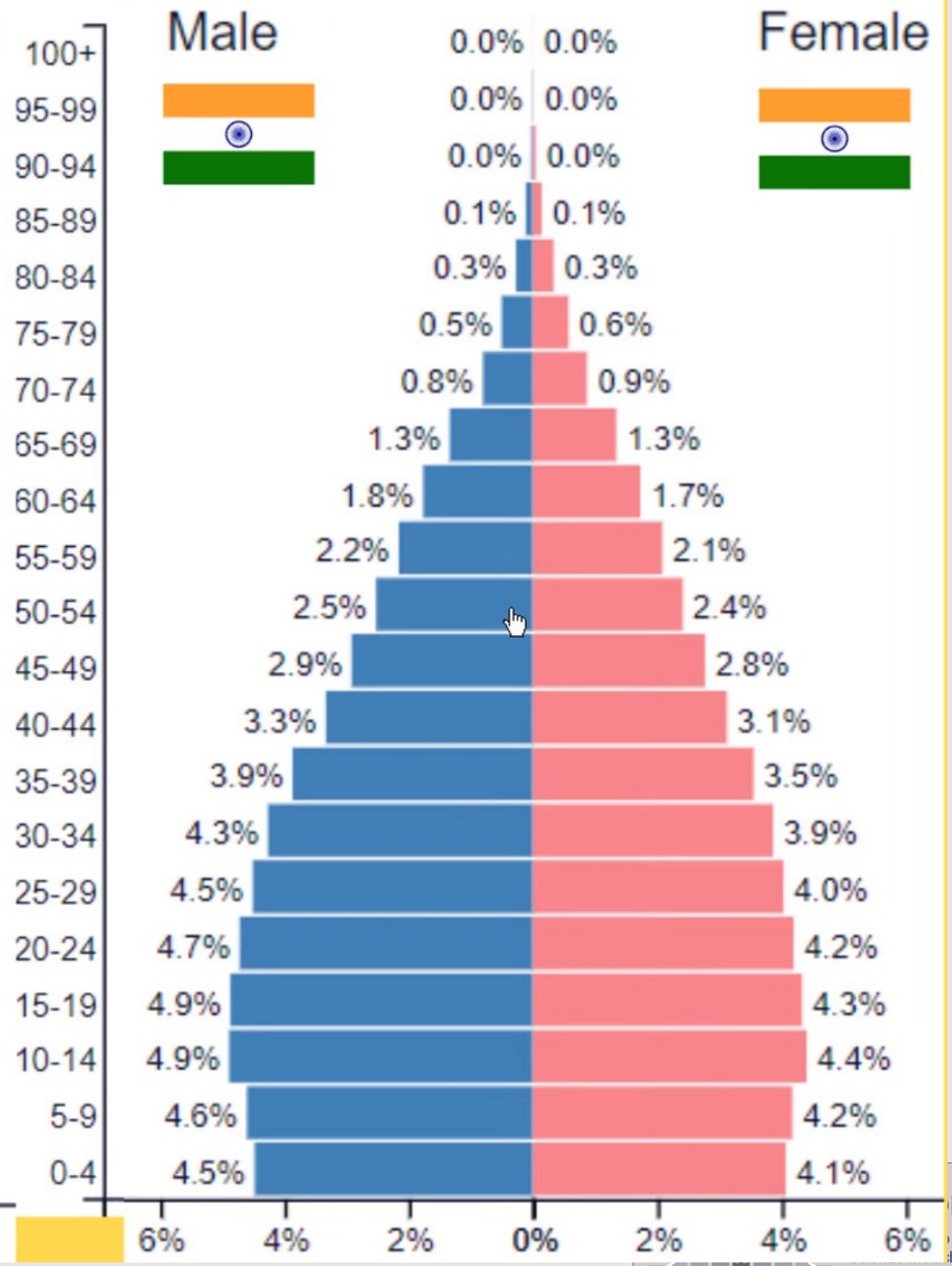
However, the comparison of **crude** mortality or morbidity rates is often **misleading** because the populations being compared ***may differ significantly*** with respect to certain underlying characteristics, such as age or sex, that will affect the overall rate of morbidity or mortality.

An older population will have a higher overall mortality rate than a younger population. As a result, variations in age will complicate any comparison between two or more populations that have different age structures.

To understand how a comparison of **crude rates** can be affected by differing population distributions, it should be recognized that a crude overall rate is simply a ***weighted* average** of the individual category specific rates, with the ***weights*** being the proportion of the population in each category.

SWEDEN





INDIA

For overcoming the effects of ***confounding*** variables such as age we could compare the **age specific rates**. While this allows for a more comprehensive comparison of mortality or morbidity rates between two or more populations, as the number of stratum specific rates being compared increases, the volume of data being examined may become unmanageable.

It is, therefore, more useful **to combine** category specific rates into a **single summary rate** that has been *adjusted* to take into account its age structure or other confounding factor. This is achieved in descriptive epidemiology by using the **methods of standardisation**.

DIRECT METHOD

the ***standard*** used is a *reference* population

1. Compute w_i from observed data
2. Apply w_i to standard Pop_i

$$\sum_{i=1}^{n} w_i * Pop_i$$

Both direct and indirect standardisation involves the calculation of numbers of ***expected*** events (e.g. deaths) which are compared to the number of ***observed*** events.

INDIRECT METHOD

The **standard** used is a set of specific rates

1. Use w_i from standard
2. Apply w_i to observed Pop_i

Direct method of standardisation

	Country A			Country B		
Age - group	No. of deaths	Population	Rate per 1,000 pyrs	No. of deaths	Population	Rate per 1,000 pyrs
0-29	7,000	6,000,000	1.2	6,300	1,500,000	4.2
30-59	20,000	5,500,000	3.6	3,000	550,000	5.5
60+	120,000	2,500,000	48	6,000	120,000	50
Total	**147,000**	**14,000,000**	**10.5**	**15,300**	**2,170,000**	**7**

The overall **crude** mortality rate is higher for country A (10.5 deaths per 1,000 person years) compared with country B (7 deaths per 1,000 person years), despite the **age-specific** mortality rates being always higher among all age-groups in country B.

Direct method of standardisation

Age-structure of a "**standard**" population

0-29	100,000
30-59	65,000
60+	20,000
Total	185,000

Number of **expected** deaths for A and B **applied to the standard** population.

	Country A	**Country B**
	Expected deaths	Expected deaths
0-29	0.0012 x 100,000 = 120	0.0042 x 100,000 = 420
30-59	0.0036 x 65,000 = 234	0.0055 x 65,000 = 357.5
60+	0.048 x 20,000 = 960	0.05 x 20,000 = 1,000
Total expected deaths	1,314	1,777.5
Age adjusted rate	**1,314/185,000 = 7.1 per 1,000 pyrs**	**1,777.5/185,000 =9.6 per 1,000 pyrs**
Age standard rate ratio (B:A)= 9.6/7.1 = 1.35		

**the rate is divided back by 1000 to give the basic rate

Controlling for age, mortality in B is **35% higher** than in A (CMR= *Comparative* Mortality Ratio).

The ***standard population*** may be the distribution of one of the populations being compared or an outside standard population such as the *European* or *World* standard population.

Indirect method of standardisation

Commonly used when age-specific rates are unavailable. For example if we did not know the age specific mortality rates for country B.

In this method, **a set of rates from a *standard population*** (country A) is applied to each of the populations being compared to calculate standardized morbidity/mortality ratios.

	Country A	Country B
	Expected deaths	Expected deaths
0-29	0.0012 x 6,000,000 = 7,200	0.0012 x 1,500,000 = 1,800
30-59	0.0036 x 5,500,000 = 19,800	0.0036 x 550,000 = 1,980
60+	0.048 x 2,500,000 = 120,000	0.048 x 120,000 = 5,760
Total expected deaths (E)	147,000	9,540
Total observed deaths (O)	147,000	15,300
Standardised Mortality Ratio O/E x 100	100	160

15300/9540=1.60

\# **expected deaths** if B had the same age-specific mortality rates as A.

\# **global observed deaths** in Country B is **60% higher** than the **#** we would expect if Country B had the same mortality experience as Country A

$$SMR = \frac{160}{100} = 1.6$$

UNITÀ DI BIOSTATISTICA
Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute