Calculation of rates

Gilles Guillot

Adapted from Chapter 3 of *Epidemiology and Biostatistics*
T. Zheng, P. Boffetta and P. Boyle, 2011

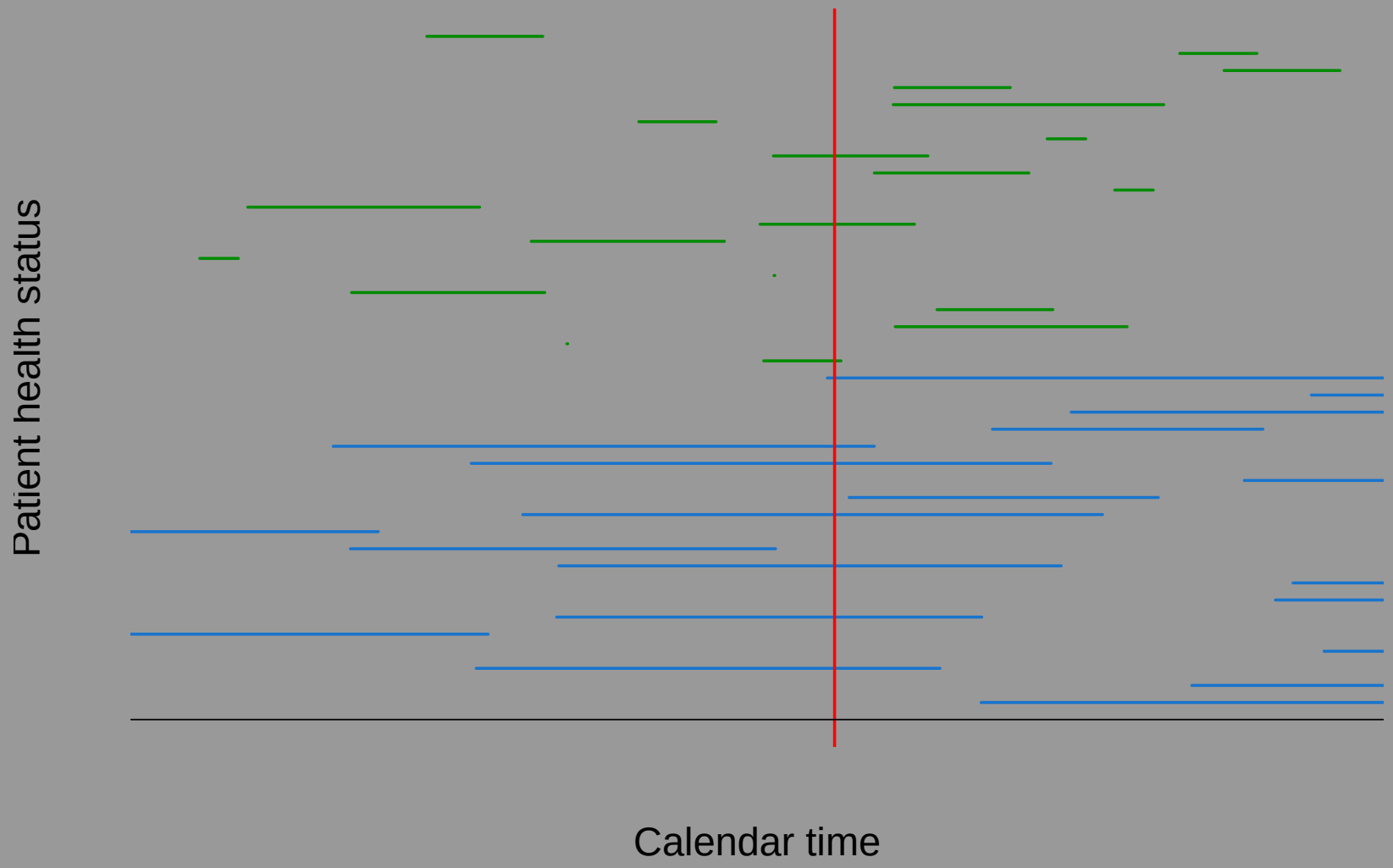
IPRI training in Epidemiology and Biostatistics
Baku - September 2019

Compilation date: 2019-08-01 16:12:56

Number of cases vs. rates

- Epidemiological analyses require numbers (statistics)
- At the lowest level: case occurence
- Number of cases easier to interpret if put in relation with the size of population

Prevalence rate: graphical example



Prevalence rate: definition

Proportion of individuals suffering from a disease at a certain point in time.

$$\text{Prev} = \frac{c}{P}$$

with c number of cases, P total population size.

Example: calculation of a prevalence rate

Prevalence of flu week 49 of year 2009 in France

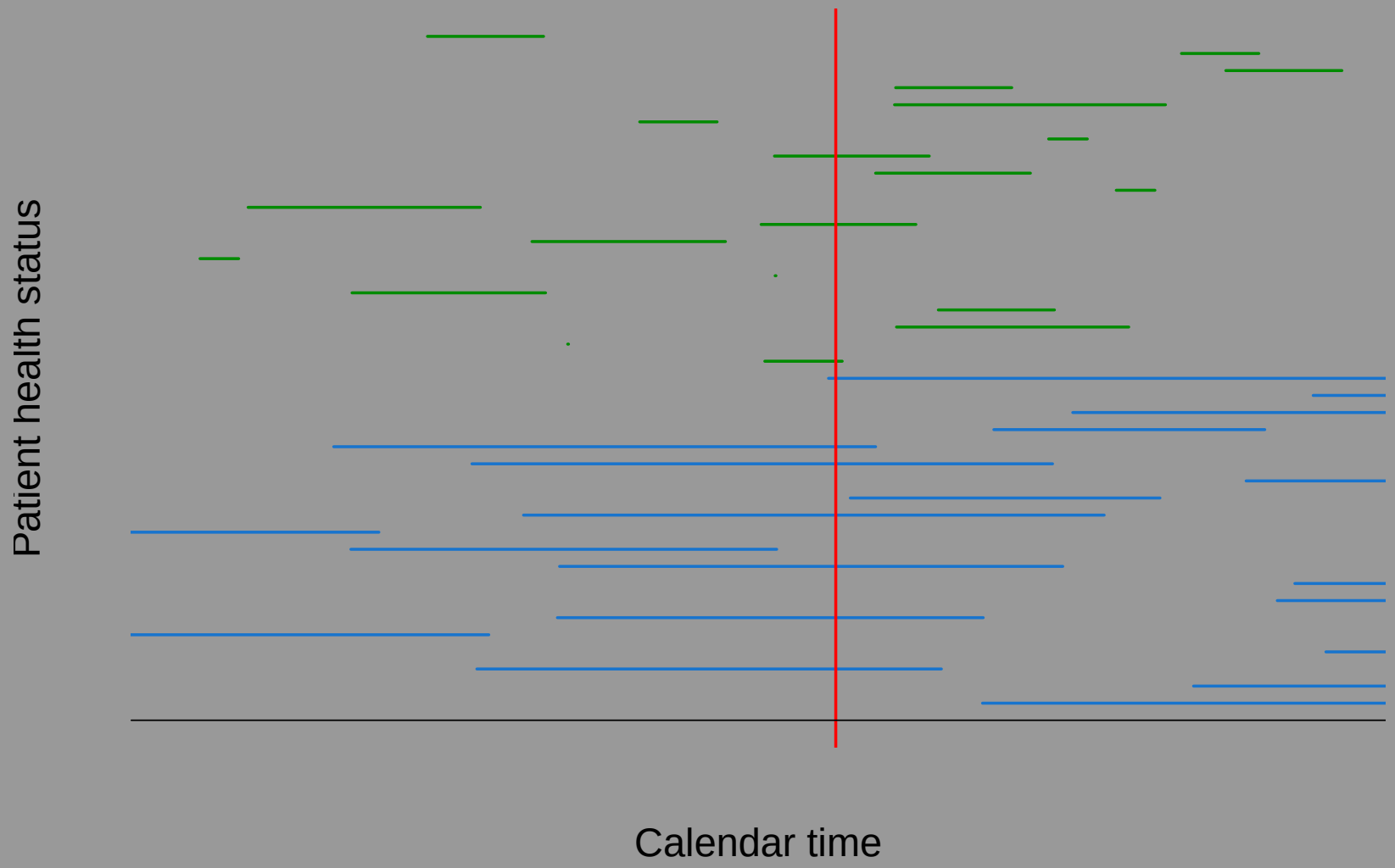
- $c = 800\,000$ cases of reported flu cases
- $P = 64\,660\,000$

$$\text{Prev} = \frac{c}{P} = \frac{800000}{64660000} \approx 0.012$$

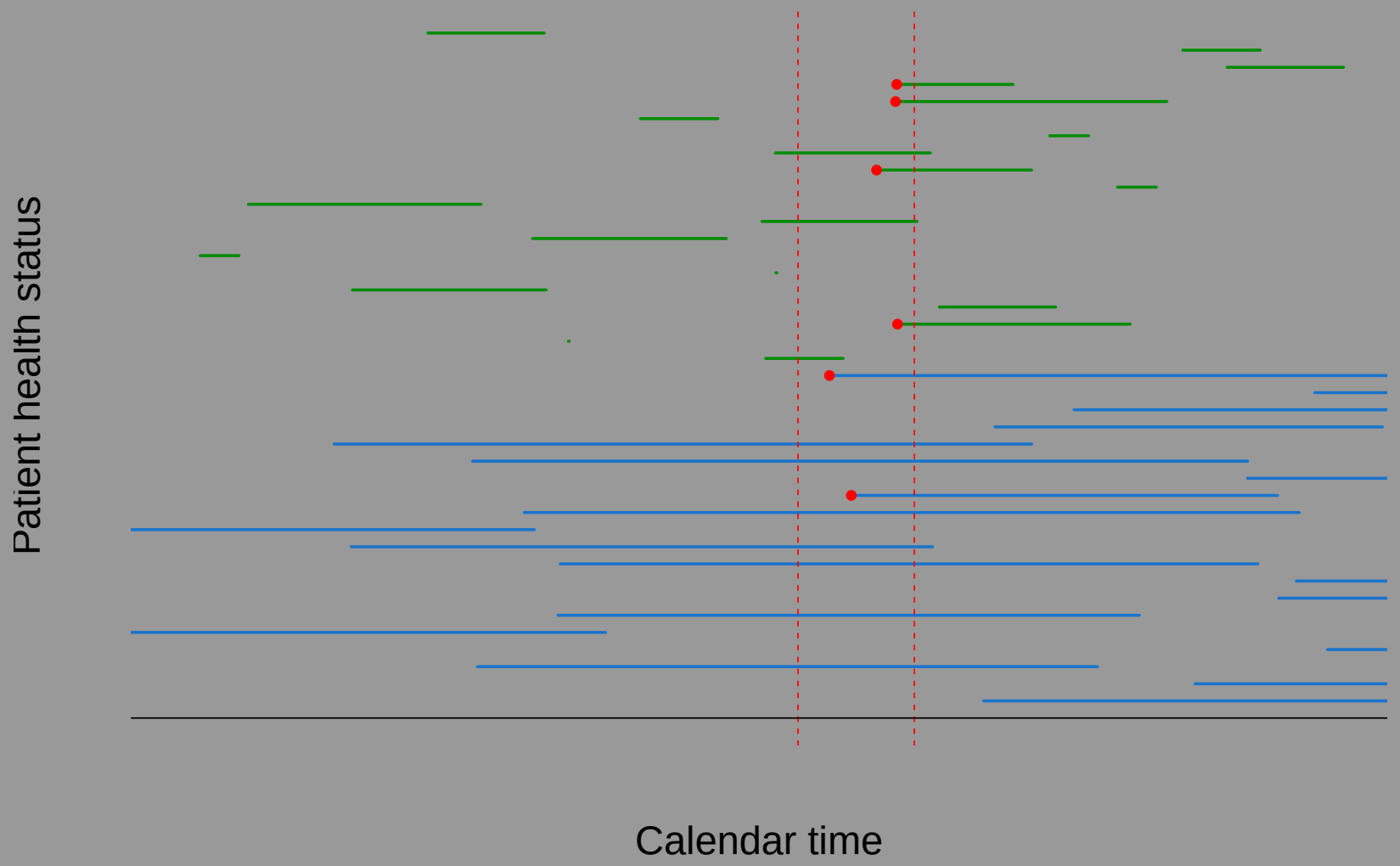
(per inhabitant)

More easily read as 1200 cases of flu per 100 000 inhabitants.

Prevalence depends on duration of illness



Incidence rate: graphical example



Incidence rate

Incidence rate $=$

$$\frac{\text{Number of persons getting sick under reference period}}{\text{Number of persons at risk}}$$

Can be contrasted with prevalence rate

$$\frac{\text{Number of persons being sick at a reference date}}{\text{Number of persons at risk}}$$

The denominator in the incidence

Denoting by N the number of persons **getting** sick under reference period and by P the number of persons at risk, the incidence rate is

$$R = \frac{N}{P} \tag{1}$$

- N is a count of occurence of new cases under the reference period
- P is the number of persons that could get sick under the reference period.

The denominator in the incidence (cont')

- If we change the reference period and consider a period twice longer, one should expect to count twice as many occurences of new cases (all things being equal).
- Example 1:
  - $P_1 = 100\ 000$ persons followed over one year
  - $N_1 = 6$ disease cases
  - Incidence $= 6/100\ 000 = 0.00006$ **cases per person x year** (reads person year)
  - More covenniently expressed as 6 **cases per 100 000 person x year**
- Example 2:
  - $P_2 = 100\ 000$ persons followed over ten years
  - $N_2 = 60$ disease cases
  - P_2 is equivalent to 1 000 000 persons followed for a year
  - Incidence $= 60/1\ 000\ 000$ (still 6 cases per 100 000 person x year)

The denominator in the incidence (cont')

The size of the population at risk P (denominator) can be decomposed as

$$P = p \times d$$

where p is the number of persons in the population and d is the duration of the reference period.

- If d is expressed in years, P will be expressed in person x years (most common for chronic diseases)
- If d is expressed in months, P will be expressed in person x months

Units when reporting incidence rates

- Official global health statistics data often provided at a 5 year resolution
- Rates expressed as number of cases per 100 000 person x year
- Example: cancer in the US
  - 320 M inhabitants, each followed over period 2015-2019
  - $N = 8$ M new cancer cases under period 2015-2019
  - Incidence $R = 8\,000\,000/(320\,000\,000 \times 5) = 8/1\,600$
    - $R = 0.005$ cases per person x year
    - $R = 500$ cases per 100 000 person x year

Exercise 1

A field study of an infectious disease in an emergency context reported 1389 new cases in a period of two months in an area inhabited by 452 000 persons. What are the incidence rates expressed in cases per person x year and in cases per 100 000 person x year?

Solution at the end

Interpretation of incidence rate

In a population where all individuals share the **same risk** to get sick, the incidence rate R can be interpreted as the **probability for a person to get sick during a period of one year**.

- The probability for a person to get sick during a period of 2 years is just $2R$.
- The assumption that all individuals share the **same risk** is often not correct, cf discussion on **age-specific incidence** and **cumulative incidence** below.

Variability (or precision) of incidence rate

Incidence rate: $R = N/P$, N and P known with certainty

- Number of cases N subject to variability
- Average/long-term/background value determined by demographic, socio-economic, environmental, genetic factors
- N varies even if factors are held constant

Variability can be quantified by variance:

$$Var[R] = \frac{R(1-R)}{P}$$

- The variability of R is
  - inversely proportional to the number of person x year P
  - small when R is close to 0 (most common situation) or 1
  - maximal when $R = 1/2$

Analytical derivation of $Var[R]$

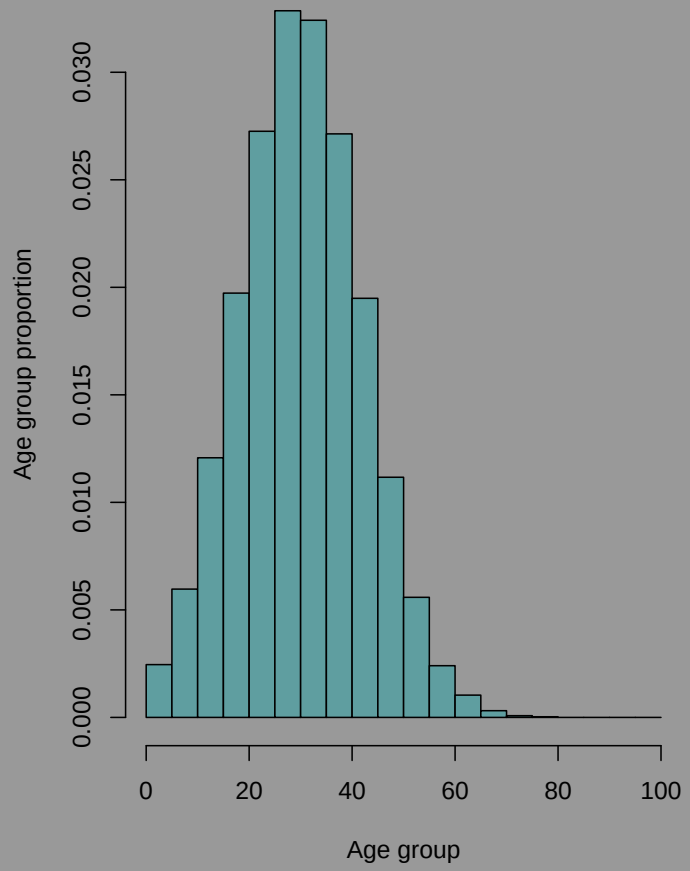
- In $R = N/P$, we consider that P is fixed and known with certainty while N is one observed value, among other potential values that could have been observed in the same population.
- Since N is a count arising from a population of size P, we assume it is binomial with size P and proportion R and $Var[N] = P\,R(1-R)$
- But $Var[R] = Var[N/P] = \frac{1}{P^2} Var[N]$ therefore,

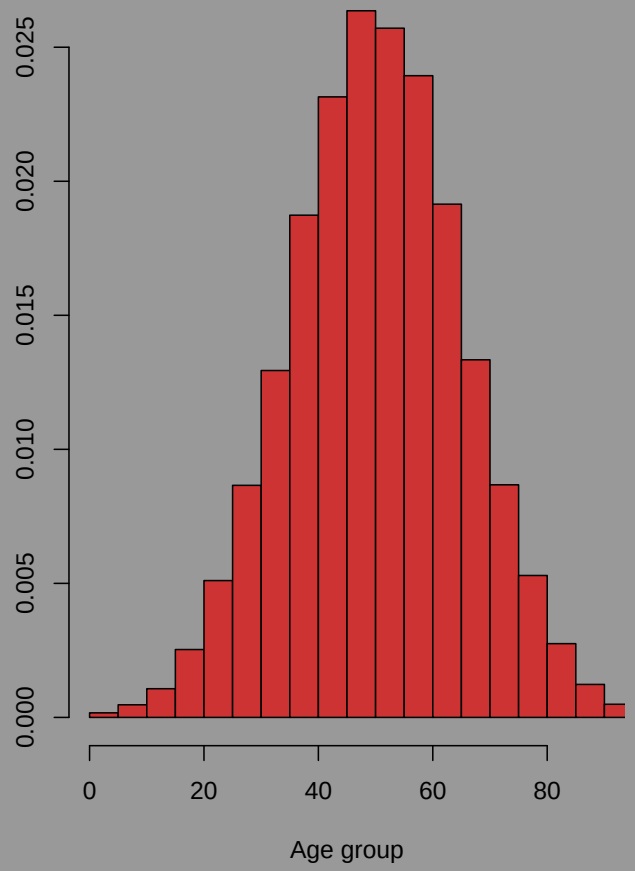
$$Var[R] = \frac{R(1-R)}{P}$$

Mortality rate

- From a purely statisticalp erspective, mortality is another health event.
- Computation follows the same rule as any other incidence rate
- A mortality is expressed in number of death events per 100 000 person x year
- The reference period has most often a duration of one year, therefore *number of death events per 100 000 person x year* is often shorten into *number of death events per 100 000 person*

Incidence rate and inter-country comparison





A global incidence rate can be misleading in presence of differences in demographic structure.

Age-specific incidence rate

- The (global) incidence rate that we defined earlier (Eq. 1) merges together all individual of the population as if they were exchangeable
- We know that many diseases affect differently the various age classes of a population
  - Prostate cancer are very rare among young men, more frequent after 50
- A more **fine grained** description of the incidence in a population is the **age-specific** incidence rate

Definition of age-specific incidence rate

For various age groups (e.g. 0-4, 5,9, 10-15,. . . indexed $i = 1, 2, 3...$), the incidence rate in age group i is defined as

$$r_i = \frac{n_i}{p_i} \tag{2}$$

where

- n_i is the number of cases among persons in age class i
- p_i is the number of persons at risk in age class i.

Cumulative risk up to age group k

The cumulative risk up to age group k is the proability to get sick at any age up to age group k.

It can be shown to be

$$\text{Cum Risk}_k = 1 - \prod_{\text{Age group} i \leq k} (1 - r_i)^{D_i} \tag{3}$$

Analytical approximation of cumulative risk up to age group k

It can be shown that

$$\text{Cum Risk}_k \approx 1 - \exp(-\text{Cum rate}_k) \tag{4}$$

where Cum Rate_k is the cumulative risk defined as

$$\text{Cum Rate}_k = \sum_{\text{Age group} i \leq k} D_i \times r_i \tag{5}$$

Analytical derivation of cumulative risk up to age group k

The event "getting sick in age group k" occurs with probability $r_i^{D_i}$.

The event "**not** getting sick in age group k" occurs with probability $(1 - r_i)^{D_i}$.

The event "getting sick in one of the age groups $1, ..., k$" occurs with probability

$$
\begin{aligned}
& 1 - P(\text{not getting sick in one of the age groups } 1, ..., k) \\
= \; & 1 - P(\text{not getting sick in age group 1 AND} \\
& \text{not getting sick in age group 2} \\
& \text{... AND not getting sick in age group } k) \\
= \; & 1 - \prod_{i \leq k} P(\text{not getting sick in age group } i) \\
= \; & 1 - \prod_{i \leq k} (1 - r_i)^{D_i}
\end{aligned}
$$

Analytical derivation of cumulative risk up to age group k (cont')

$$\ln(\prod_{i \leq k}(1 - r_i)^{D_i}) = \sum_{i \leq k} D_i \ln(1 - r_i)$$

By a Taylor expansion, $\ln(1 - r_i) \approx -r_i$ therefore

$$\ln(\prod_{i \leq k}(1 - r_i)^{D_i}) \approx \sum_{i \leq k} -D_i r_i \text{ and } \prod_{i \leq k}(1 - r_i)^{D_i} \approx \exp(\sum_{i \leq k} -D_i r_i)$$

Comparison of populations with different age structures

- Age is a crucial factor for many diseases.
- Comparison of health statistics from two populations with different age structures not straightforward
- We describe two methods for comparison
  - Age standardized rates
  - Standardized incidence ration (SIR)

Age standardization of global incidences

The global incidence defined earlier is $R = N/P$ which can rewritten with age-specific number of cases and sizes of population at risk:

$$R = \frac{\sum n_i}{\sum p_i} = \frac{\sum r_i p_i}{\sum p_i} \tag{6}$$

This writing shows that the global incidence is affected not only by the age-specific incidences but also by the age structure of the population.

Age standardization is an attempt to filter out the effect of the age structure of the population.

It answers to the question: *what would the global incidence be if the age structure was that of an arbitrary population with a fixed age structure?*

Standardization with the *world standard population*

The most commonly used population used to play the role of this arbitrary population is the world population in 1960[1]

Denoting by p'_i the age-specific world population sizes in the 60's, the age-standardized incidence is

$$ASR = \frac{\sum r_i p'_i}{\sum p'_i} \tag{7}$$

[1] often referred to as world standard population

Standardized incidence ratio (SIR)

- If we have two populations labelled A and B with respectively
  - number of cases n_i and n'_i in age group i
  - number of persons at risk p_i and p'_i in age group i
- The age specific incidences in age group i are $r_i = n_i/p_i$ and $r'_i = n'_i/p'_i$.
- $n_i \times r'_i$ represents the number of cases one **would expect** in a fictional population with age structure of A and incididences of B.

The standardized incidence ratio (SIR) is defined as

$$SIR = \frac{\text{number of observed cases}}{\text{number of expected cases}} = \frac{\sum n_i}{\sum p_i r'_i}$$

Standardized incidence ratio (SIR) (cont')

- SIR often reported as a percentage
- A value of 100 indicates that a change of rates in pop A from their current values to those of pop B (all things being equal) would not modify the total number of cases.

Solution to exercise 1

- Method 1: compute size of population at risk first then compute incidence rate
  - A population of 452 000 persons observed during two months is equivalent to a population 452 000 / 6 = 75333 persons observed during a year (hence to 75333 person x year)
  - $R = 1389/75333 = 0.018$ cases per person x year
  - this is equivalent to 1800 cases per 100 000 person x year
- Method 2: compute incidence rate first then convert into case per person x year
  - R = 1398 / 452 000 = 0.003 cases per person x 2months
  - This is equivalent to R = 0.003 x 6 = 0.018 cases per person x year