Grau d'Estadística

Estadística Mèdica

Francesc Miras

francesc.miras@upc.edu

Universitat de Barcelona – Universitat Politècnica de Catalunya

7th November, 2018

TODAY'S PROGRAMME

MEASURES OF DISEASE OCCURRENCE

- PREVALENCE
  - Definition
  - Examples
  - Estimation
  - Comments
- CUMULATIVE INCIDENCE
  - Definition
  - Comments
  - Exercises
- INCIDENCE RATE
  - Definition
  - Confidence interval for the incidence rate
  - Comparison of incidente rates
  - Relation between prevalence and incidente

Disease prevalence

Definition

The **prevalence** or **prevalence proportion** of a disease D is the proportion of individuals affected by D among the population of interest at a given time t:

$$P = \frac{X}{N},$$

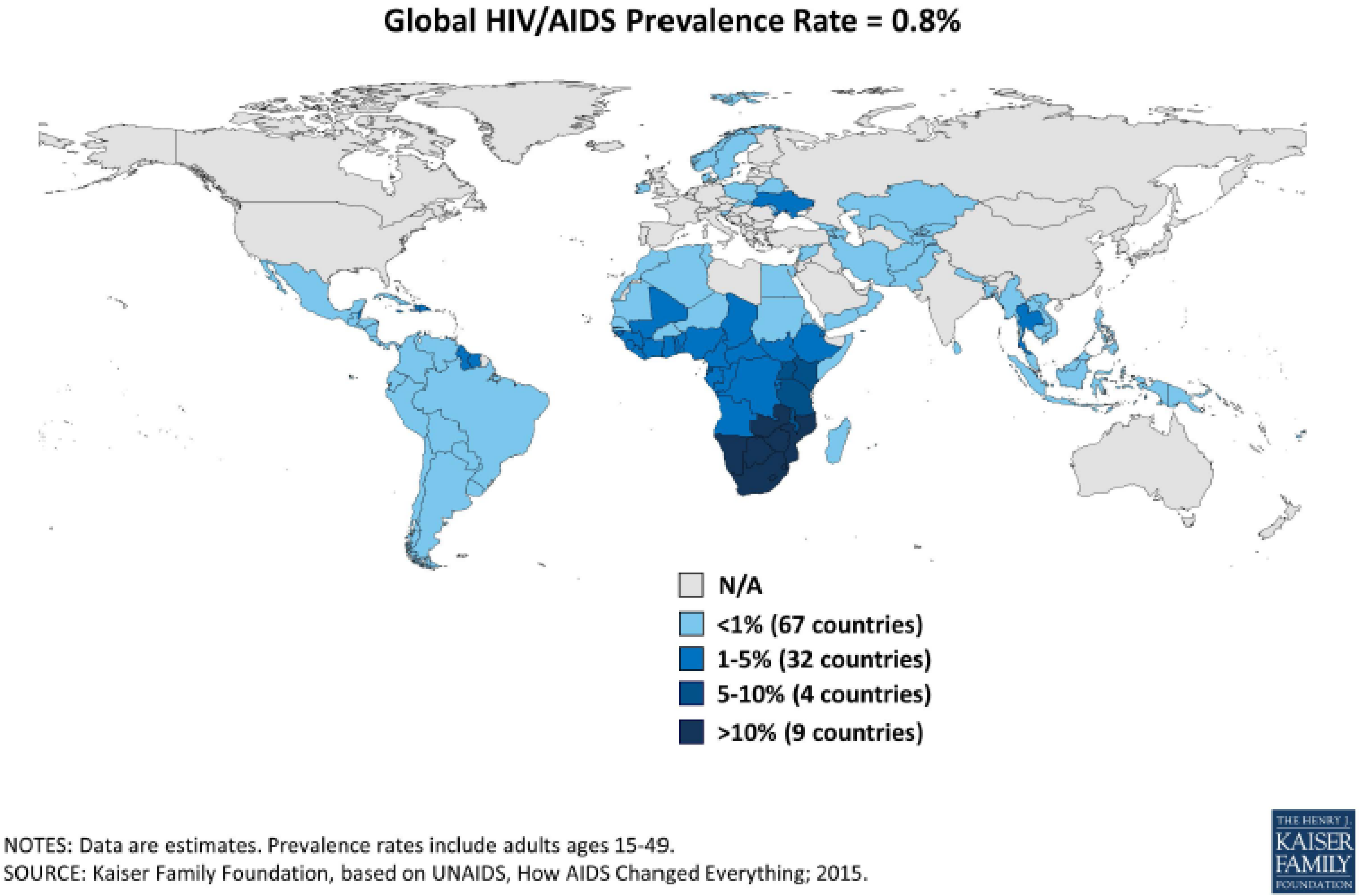
where X is the number of disease cases and N is the population size at time t.

The prevalence can be interpreted as the probability that a randomly selected subject from the population has the disease at time t.

Notes:

- Prevalence may also be given in absolute numbers (X).
- Time t may be a day, week, month, year, etc.

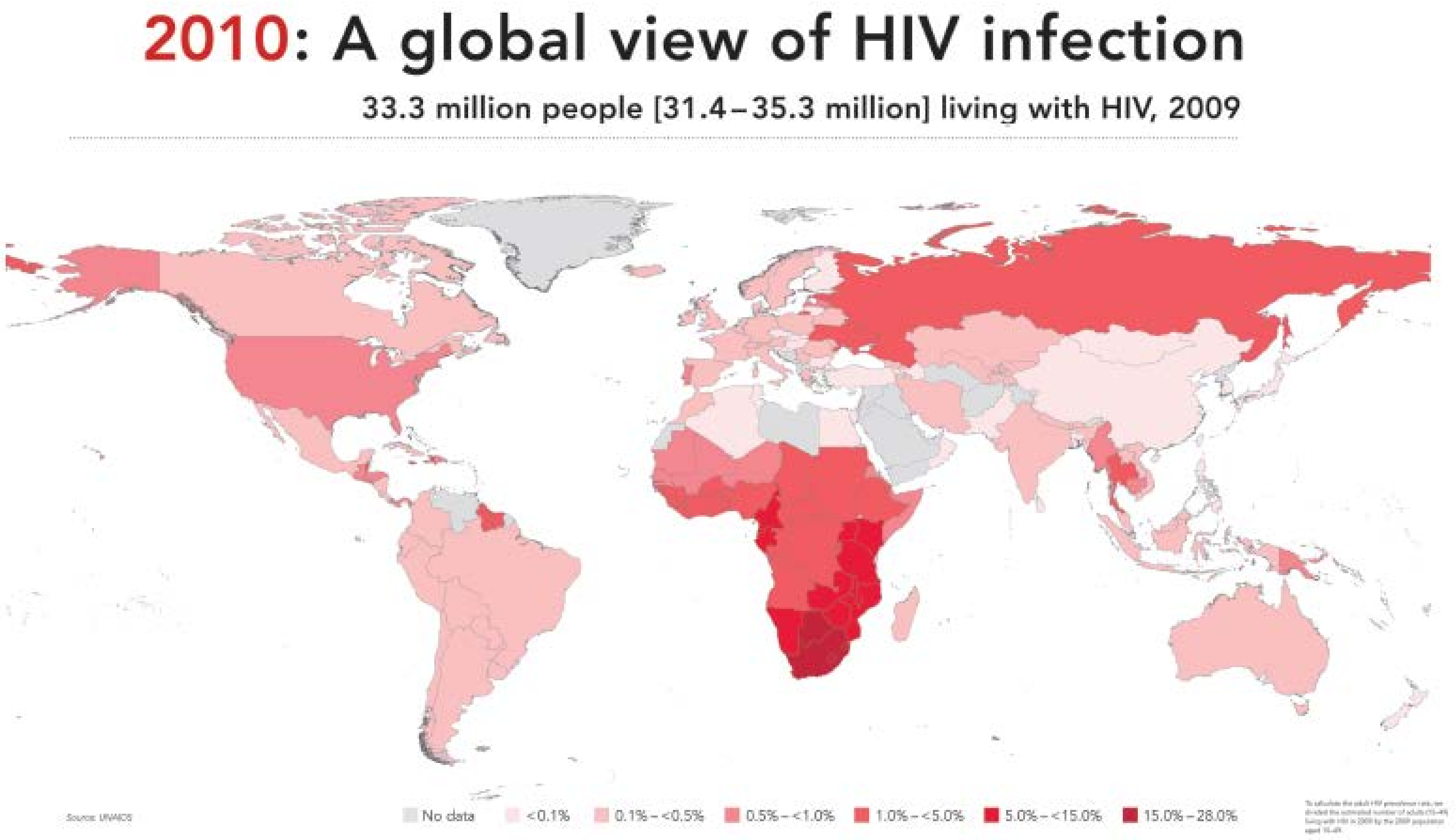
Disease prevalence: two examples



NOTES: Data are estimates. Prevalence rates include adults ages 15-49.
SOURCE: Kaiser Family Foundation, based on UNAIDS, How AIDS Changed Everything; 2015.

Figure 1: Adult HIV prevalence in 2014 (Source: UNAIDS 2015)

Disease prevalence: two examples (cont.)



2010: A global view of HIV infection

33.3 million people [31.4–35.3 million] living with HIV, 2009

Source: UNAIDS

To calculate the adult HIV prevalence rate, we divided the estimated number of adults (15–49) living with HIV in 2009 by the 2009 population aged 15–49.

Figure 2: Worldwide HIV prevalence in 2009 (Source: UNAIDS 2010)

Disease prevalence: two examples (cont.)

Muga *et al.* (2006): Significant reductions of HIV prevalence but not of hepatitis C virus infections in injection drug users from metropolitan Barcelona: 1987–2001.

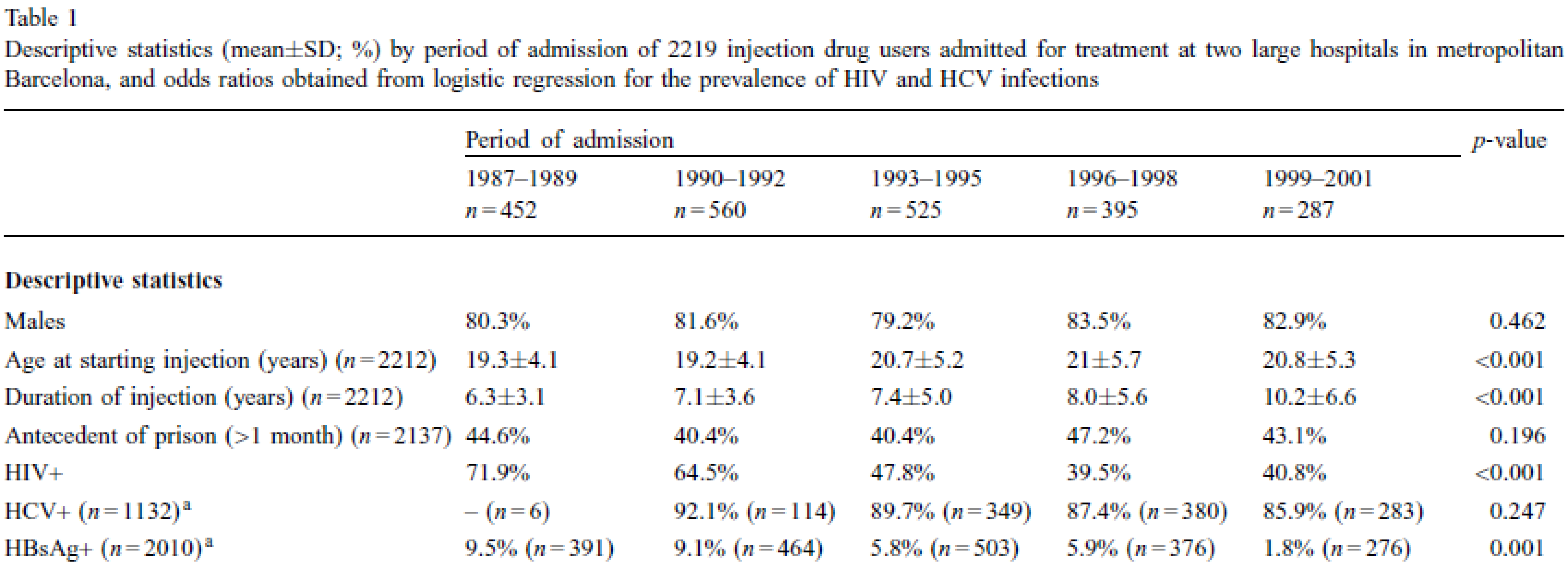


Table 1

Descriptive statistics (mean±SD; %) by period of admission of 2219 injection drug users admitted for treatment at two large hospitals in metropolitan Barcelona, and odds ratios obtained from logistic regression for the prevalence of HIV and HCV infections

	Period of admission					*p*-value
	1987–1989 $n=452$	1990–1992 $n=560$	1993–1995 $n=525$	1996–1998 $n=395$	1999–2001 $n=287$	
Descriptive statistics						
Males	80.3%	81.6%	79.2%	83.5%	82.9%	0.462
Age at starting injection (years) ($n=2212$)	19.3±4.1	19.2±4.1	20.7±5.2	21±5.7	20.8±5.3	<0.001
Duration of injection (years) ($n=2212$)	6.3±3.1	7.1±3.6	7.4±5.0	8.0±5.6	10.2±6.6	<0.001
Antecedent of prison (>1 month) ($n=2137$)	44.6%	40.4%	40.4%	47.2%	43.1%	0.196
HIV+	71.9%	64.5%	47.8%	39.5%	40.8%	<0.001
HCV+ ($n=1132$)[a]	– ($n=6$)	92.1% ($n=114$)	89.7% ($n=349$)	87.4% ($n=380$)	85.9% ($n=283$)	0.247
HBsAg+ ($n=2010$)[a]	9.5% ($n=391$)	9.1% ($n=464$)	5.8% ($n=503$)	5.9% ($n=376$)	1.8% ($n=276$)	0.001

Figure 3: HIV, HCV, and HBsAG prevalence among injection drug users in Barcelona (Source: Muga et al. (2006))

Disease prevalence (cont.)

Estimation of disease prevalence

Given a sample of size n, the prevalence can be estimated by

$$\hat{P} = \frac{X_n}{n},$$

being X_n the number of cases among the sample.

Given a sample of independent observations and assuming that X_n follows a **binomial distribution** with parameters n and P, a confidence interval for P can be computed in different ways:

- Computation of an exact interval using the binomial distribution (Clopper and Pearson 1934):
  $\leadsto$ R-function `binom.exact` (in package `epitools` (Aragon 2012)).

Disease prevalence (cont.)

Interval estimation of disease prevalence

- Approximation of the exact interval using the normal distribution:

  $$\mathrm{CI}(P; 1-\alpha) = \hat{P} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{P}(1-\hat{P})/n} \tag{1}$$

  or

  $$\mathrm{CI}(P; 1-\alpha) = \hat{P} \pm z_{1-\alpha/2} \cdot 0.5/\sqrt{n},$$

  where $1-\alpha$ is the confidence level and $z_{1-\alpha/2}$ the $(1-\alpha/2)$–quantile of the standard normal distribution.
  $\rightsquigarrow$ R-function `binom.approx` (`epitools`)

Note: Function `propCI` of the `prevalence` package (Devleesschauwer 2014) calculates three more types of confidence intervals for proportions.

Disease prevalence (cont.)

Comparison of formulas

The following table shows the coverage proportions of approximate 95%-confidence intervals (according to equation (1)) for several combinations of prevalence and sample size.

P	n	%[a]	P	n	%[a]	P	n	%[a]
0.05	100	87.7	0.25	100	94.5	0.5	100	94.3
	200	92.6		200	93.7		200	94.5
	500	93.2		500	94.4		500	94.6
	1000	94.2		1000	94.7		1000	94.7

[a] Based on 100 000 repetitions

Disease prevalence (cont.)

Comments:

- Disease prevalence can be estimated in cross-sectional studies, but not in cohort or case-control studies.
- Prevalence estimates of different populations or at different time points may be compared by means of statistical tests.
  $\leadsto$ R-functions `binom.test` and `prop.test`
- Prevalence depends on disease duration: the longer the duration, the higher the prevalence.
- Causal relation between disease and risk factors cannot be established by means of prevalence data.
- **Lifetime prevalence** is the number/proportion of individuals in a population that at some point in their life have had the disease.

Cumulative incidence

Definition

The **cumulative incidence** or **incidence proportion** (or just **risk**) of a disease is the proportion of new cases within a **period of time** of duration Δ among an initially disease-free population:

$$CI(\Delta) = \frac{I}{N_0},$$

where N_0 is the size of the initially disease-free population and I the number of new cases (**incident cases**) during the period of time.
The cumulative incidence can be interpreted as the probability that a disease-free individual comes down with the disease during the specified period of time.

Example: Martín-Santos *et al.* (2012). Research Letter: Is neuroticism a risk factor for postpartum depression? *Psychological Medicine*, 42(7), 1559–1565.

Cumulative incidence (cont.)

Comments:

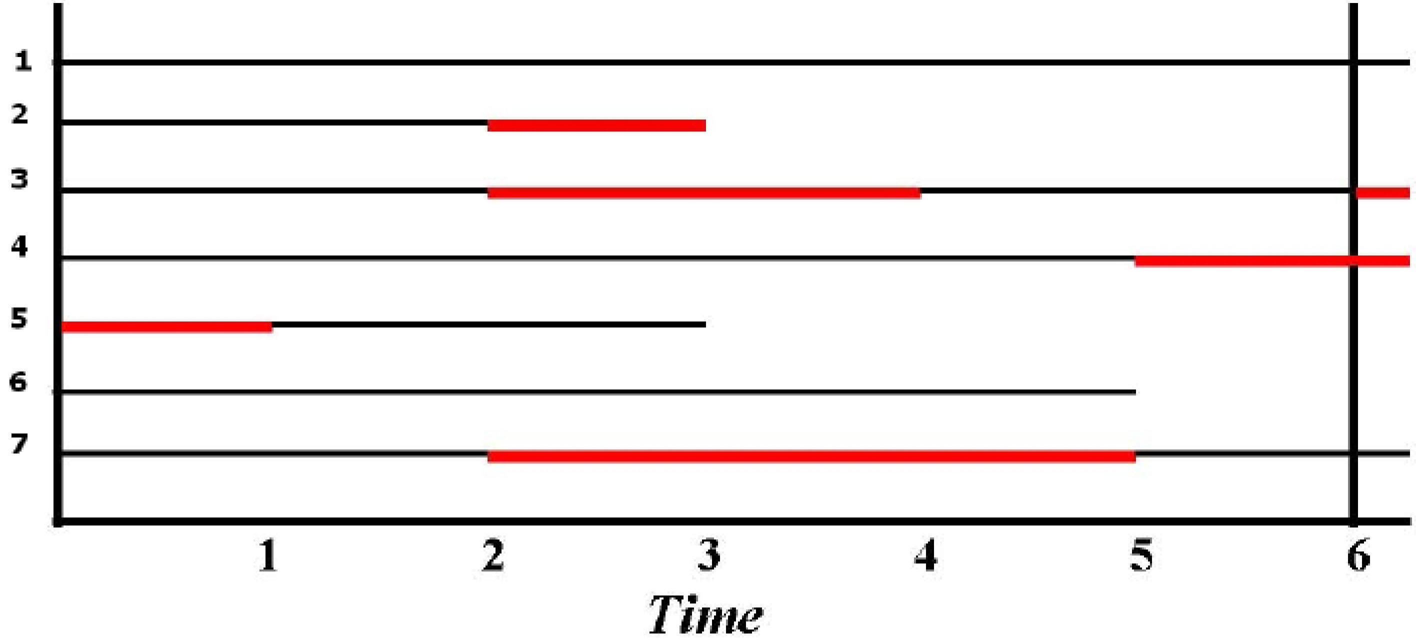
- The time scale of interest may be age, calendar time, or the elapsed time from a given time, e.g., the start of a therapy.
- The cumulative incidence can be estimated in cohort studies, but not in case-control nor in cross-sectional studies.
- Its estimator is the proportion of new cases among the cohort during follow-up. To calculate the confidence interval, given a sample of independent observations, one can use the binomial distribution or its approximation by the normal distribution.
- The **period prevalence** is the proportion of disease cases in a population at some point within a period of time of duration Δ:

  $$PP(\Delta) = \frac{X + I}{N},$$

  where X and I are the cases at period start and incident cases, resp..
- $CI(\Delta)$ may be difficult to estimate in dynamic populations.

Exercise 1

✍ In the following setting, each horizontal line represents an individual's time under observation illustrating in red the duration of a disease. Calculate the prevalence at times $t = 2$ and $t = 5.5$, as well as the cumulative incidence and the period prevalence within the time period from $t = 0$ until $t = 6$.



Exercise 2

The main objective of the Framingham Heart Study (`http://www.framinghamheartstudy.org`) is to identify the risk factors for cardiovascular disease.

The following table shows data on coronary heart disease (CHD) among men aged 30 to 59 according to cholesterol level reported by that study: in the left panel, the incidence of CHD after 10 years of follow-up; in the right panel, the prevalence by the end of follow-up.

✍ Compare prevalence and incidence data and discuss the findings. Are the prevalence data useful to study the possible causal relation between cholesterol level and CHD? If not, which might be the reasons?

Table: Prevalence and incidence of CHD (Jewell (2004); Friedman et al. (1966).)

	Incidence (10 years)			**Prevalence (after 10 y.)**		
Cholesterol	CHD	No CHD	Total	CHD	No CHD	Total
High	85	462	547	38	371	409
Low	28	516	544	33	347	380

Incidence Rate

Definition

The **incidence rate** (IR) or **incidence density** (or just incidence) of a disease D is the number of new cases (I) per unit of person-time at risk:

$$I_r = \frac{I}{\Delta T},$$

where ΔT is the total time **under risk** of the whole study population.

- The IR is "the rate at which new events occur in a population." (Porta 2008).
- To calculate the incidence rate, one needs to know how long each study subject is under risk for D.
- The IR is not a proportion and **cannot** be interpreted as a probability. Its unit is (time-unit)$^{-1}$.

Incidence rate (cont.)

An example

Among 10 persons under study, 4 come down with the disease of interest and the total time under risk is 20 years. Hence:

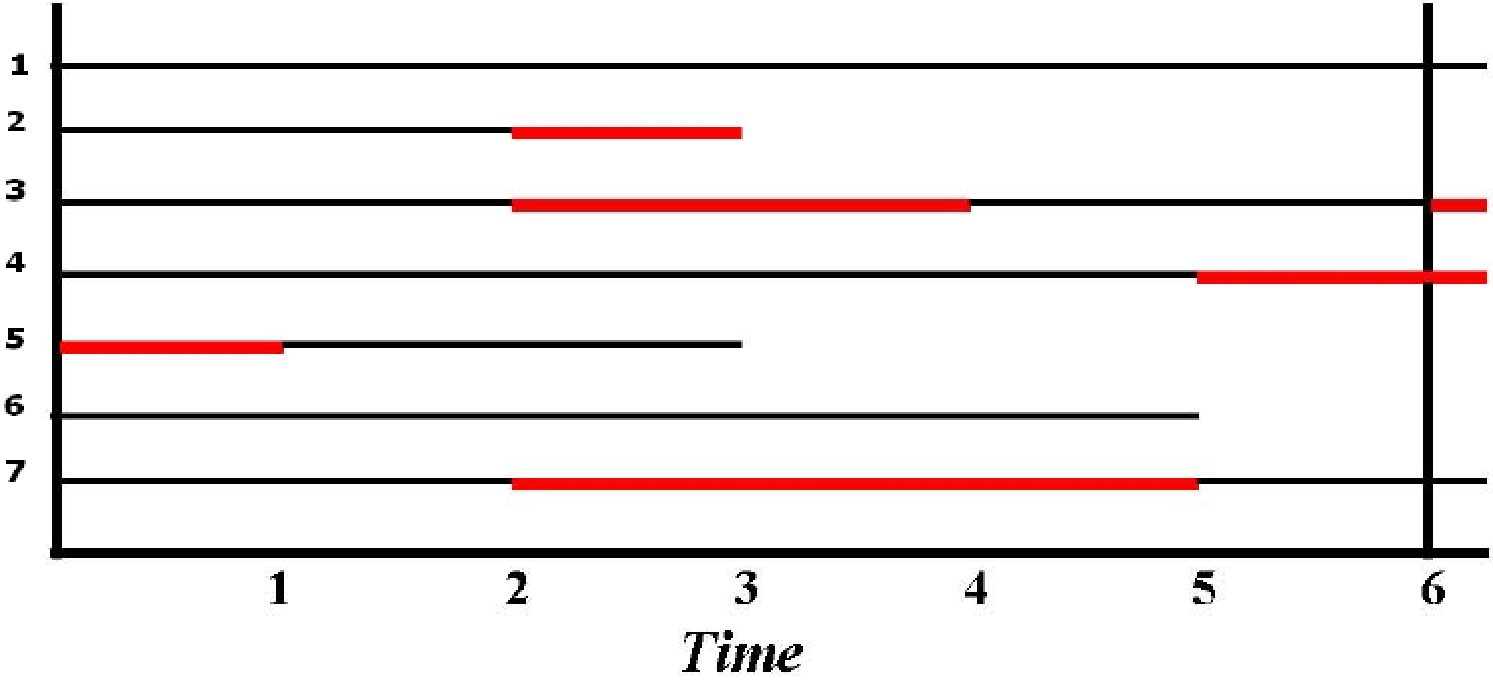
$$I_r = \frac{4 \text{ (cases)}}{20 \text{ years}} = 0.2 \quad \text{cases per year}$$
$$= 20 \text{ (cases) per 100 } \textbf{person-years}.$$

Comments:

- The IR can be estimated in cohort studies, but not in case-control nor in cross-sectional studies.
- If one is interested in the times until disease onset . . .
  $\rightsquigarrow$ Use survival analysis!

Exercise 1 (cont.)

✍ Calculate the incidence rate in the same setting as before:



Incidence rate (cont.)

Comments (cont.):

- The estimation of the IR yields the average IR for the time period under study. If it is assumed that the IR changes over time, it may be estimated separately for successive time periods:
  $\rightsquigarrow$ See, for example, Muga et al. (2007).
- The assumption of a constant incidence rate implies exponentially distributed times until disease onset, T. Hence:
  - The cumulative incidence in a time period of duration Δ can be estimated in the following manner:

$$CI(\Delta) = P(T \le \Delta) = 1 - \exp(-I_r \cdot \Delta) \overset{I_r\Delta < 0.1}{\approx} I_r \cdot \Delta.$$

  - The mean time until disease onset is the inverse of the IR: $\dfrac{1}{I_r}$.

Incidence rate (cont.)

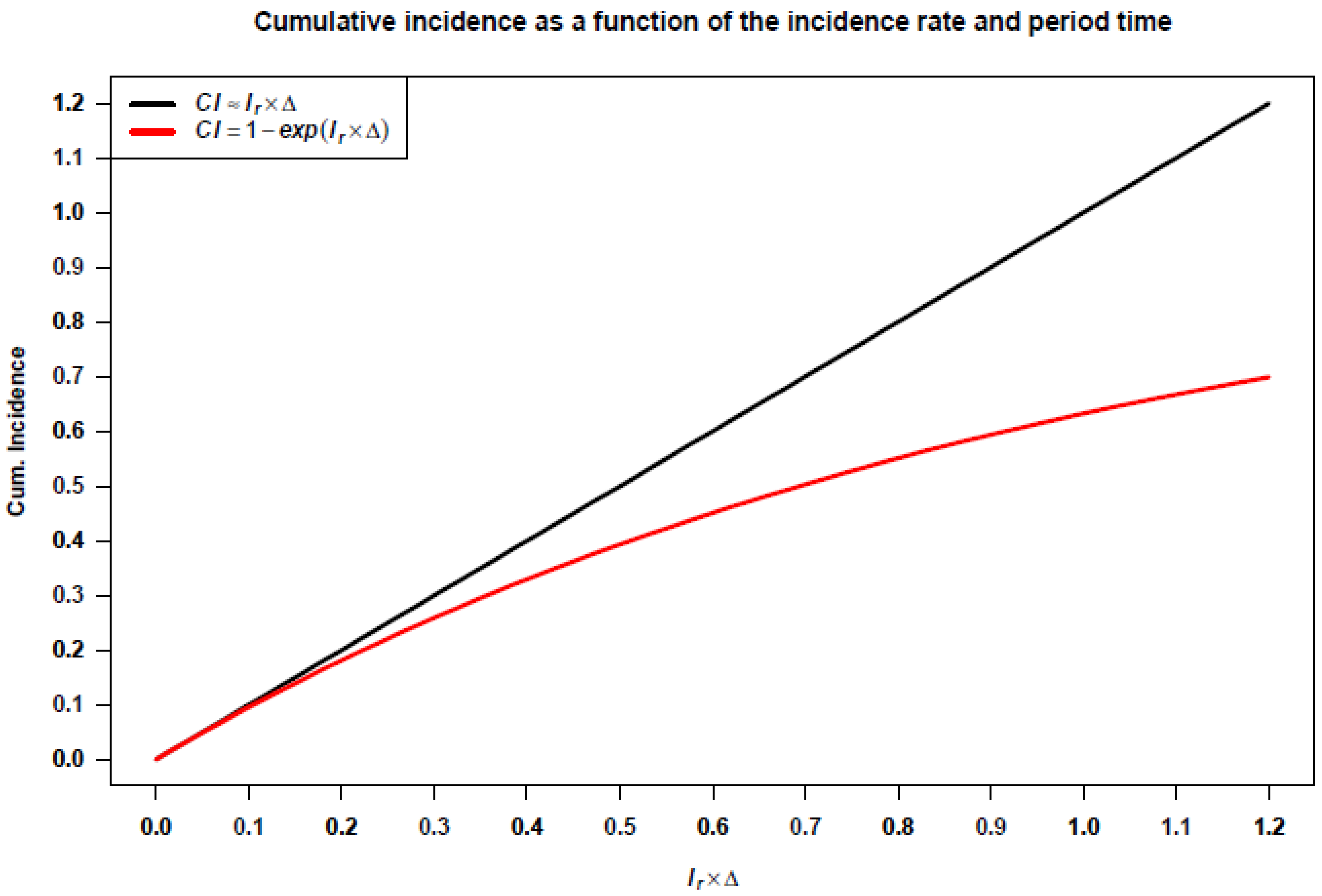


Figure 4: Comparison of exact and approximate formulas for cumulative incidence

Incidence rate (cont.)

Comments (cont.):

- In dynamic populations, the **annual incidence rate** of a disease is often estimated as follows:

  $$\hat{I}_r \overset{I \text{ small}}{\approx} \frac{I}{N^*},$$

  where I is the number of the incidence cases during the year and N^* the mid-year population, which is used as a proxy for ΔT.
  For further reading, see: Vandenbroucke, J. and N. Pearce (2012): Incidence rates in dynamic populations. *International Journal of Epidemiology* 41, 1472–1479.

- Assuming the number of incidence cases (I) follows a Poisson distribution, exact confidence intervals can be calculated based on this distribution or approximate ones based on the normal distribution:
  $\rightsquigarrow$ R-functions `pois.exact` and `pois.approx` (`epitools`).

Incidence rate (cont.)

Approximate confidence interval for the IR

Let I_r be the underlying true incidence rate, whose estimator is

$$\hat{I}_r = \frac{I}{\Delta T}.$$

It can be assumed that I, the number of incident cases throughout the total time under risk, ΔT, follows a Poisson distribution with parameter $I_r \cdot \Delta T$.

Thus, the expected value and the variance of $\hat{I}_r$ are I_r and $I_r / \Delta T$, respectively. This leads to the approximate confidence interval under large sample conditions:

$$\mathrm{CI}(I_r; 1-\alpha) = \hat{I}_r \pm z_{1-\alpha/2} \cdot \sqrt{\hat{I}_r / \Delta T}.$$

Comparison of incidence rates

Wald test for two incidence rates

The Wald test can be used to test the following hypothesis of equal incidence rates in two populations given the data from two independent samples:

$$\mathrm{H}_0 : I_{r_0} = I_{r_1} \text{ vs. } \mathrm{H}_1 : I_{r_0} \neq I_{r_1}.$$

Both incidence rates can be estimated as the ratios of the incident cases, I_0 and I_1, divided by the total times under study, ΔT_0 and ΔT_1. Given both the total observation time, $\Delta T = \Delta T_0 + \Delta T_1$, and the total number of incident cases, $I = I_0 + I_1$, under H_0, I_1 follows a binomial distribution with parameters $n = I$ and $p = \Delta T_1 / \Delta T$. Hence,

$$E_1 = \mathrm{E}(I_1) \overset{\mathrm{H}_0}{=} I \cdot \Delta T_1 / \Delta T.$$

Comparison of incidence rates (cont.)

Wald test for two incidence rates (cont.)

The corresponding variance of I_1 is

$$V_1 = \mathrm{V}(I_1) \overset{\mathrm{H}_0}{=} I \cdot \Delta T_1 \Delta T_0 / (\Delta T)^2.$$

It can then be shown that, under large sample conditions,

$$\chi^2 = \frac{(I_1 - E_1)^2}{V_1} \sim \chi^2_1,$$

or, equivalently,

$$\chi = \frac{I_1 - E_1}{\sqrt{V_1}} \sim \mathcal{N}(0, 1).$$

⇝ R-functions `rateratio` and rateratio.wald (`epitools`)

Relation between prevalence and incidence

Relation between prevalence and incidence rate

In a steady-state population, i.e., a population with constant incidence and disease duration distribution, the following equation holds:

$$P = \frac{I_r \cdot \mathrm{E}(D_d)}{1 + I_r \cdot \mathrm{E}(D_d)} \quad \Longleftrightarrow \quad \frac{P}{1-P} = I_r \cdot \mathrm{E}(D_d),$$

where D_d stands for duration of disease. That implies that the longer the average duration of disease and/or the higher the incidence rate, the higher the prevalence.

For further reading, see:

Freeman, J. and G. Hutchison (1980). Prevalence, incidence and duration. *American Journal of Epidemiology 112(5)*, 707–723.

References

- Aragon, T. (2012). epitools: Epidemiology Tools. R package version 0.5-7. `http://CRAN.R-project.org/package=epitools`
- Clopper, C. and S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*, 404–413.
- Brecht Devleesschauwer, Paul Torgerson, Johannes Charlier, Bruno Levecke, Nicolas Praet, Sophie Roelandt, Suzanne Smit, Pierre Dorny, Dirk Berkvens and Niko Speybroeck (2014). prevalence: Tools for prevalence assessment studies. R package version 0.4.0. `http://cran.r-project.org/package=prevalence`
- Friedman G., W. Kannel, T. Dawber, and P. McNamara (1966). Comparison of prevalence, case history and incidence data in assessing the potency of risk factors in coronary heart disease. *American Journal of Epidemiology 83*, 366–378.
- Jewell N. (2004). *Statistics for Epidemiology*. Chapman & Hall/ CRC.

References

- Joint United Nations Programme on HIV/AIDS (UNAIDS) (2010). Global report. UNAIDS report on the global AIDS epidemic. 2010.
- Kreienbrock, L. and S. Schach (1995). *Epidemiologische Methoden*. Gustav Fischer Verlag Stuttgart·Jena.
- Martín-Santos, R., E. Gelabert, S. Subirà, A. Gutiérrez-Zotes, K. Langohr, M. Jover, M. Torrens, R. Guillamat, F. Mayoral, F. Cañellas, J.L. Iborra, M. Gratacòs, J. Costas, I. Gornemann, R. Navinés, M. Guitart, M. Roca, R. De Frutos, F. Vilella, M. Valdés, L. García-Estévez, J. Sanjuán (2012). Research Letter: Is neuroticism a risk factor for postpartum depression? *Psychological Medicine*, 42(7), 1559–1565.
- Muga, R., A. Sanvisens, F. Bolao, J. Tor, J. Santesmases, R. Pujol, C. Tural, K. Langohr, C. Rey-Joly, A. Muñoz (2006). Significant reductions of HIV prevalence but not of hepatitis C virus infections in injection drug users from metropolitan Barcelona: 1987–2001. *Drug and Alcohol Dependence* 82 Suppl. 1, S29–S33.

References (cont.)

- Muga, R., I. Ferrero, K. Langohr, P. García Olalla, J. del Romero, M. Quintana, I. Alastrue, J. Belda, J. Tor, S. Pérez-Hoyos, J. del Amo and the Spanish Multicenter Study Group of Seroconverters (GEMES) (2007). Changes in the incidence of tuberculosis in a cohort of HIV-seroconverters before and after the introduction of HAART. *AIDS* 21, 2521–2527.
- Porta, M. (2008). *A Dictionary of Epidemiology.* 5th ed. Oxford University Press.
- Rothman K. and S. Greenland (1998). *Modern epidemiology.* Lippincott Williams & Wilkins.
- World Health Organization and UNAIDS (2015). Global Aids Response Progress Reporting 2015.