Measuring Disease and Exposure

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of Science, whatever the matter may be."

Lord Kelvin (quoted in Kenneth Rothman, *Modern Perspectives in Epidemiology*, 1 ed. Boston, Little Brown, 1986, pg 23)

Key aspects of epidemiology

- Epidemiology deals with **populations**
- Epidemiology involves **measurement**
- Epidemiologic studies involve **comparison**
- Epidemiology is fundamentally **multidisciplinary**

Numeracy: applying numbers to phenomena

Conceptual models underlie measures

How we apply numbers and what type of measures we construct:

A. purpose of the measure

B. nature of the data available to us.

C. conceptualization of the phenomenon

- Cannot observe/record all aspects of reality; must identify essential ones.
- Preserve important features and not overburden us with superfluous data.
- Conceptual models provide the basis for selection, operational definitions, classifications, measures, and analysis.

Counts and measures

Denominators

Terminology for ratios

Ratio: the quotient of two numbers (e.g., "593 persons/square mile"). This term is the most general and includes any expression with a numerator and a denominator.

Proportion: the fractional component of a quantity ("10% of NC residents are over age 65"); the numerator is "contained in" the denominator. A proportion must fall between 0 and 1 (inclusive).

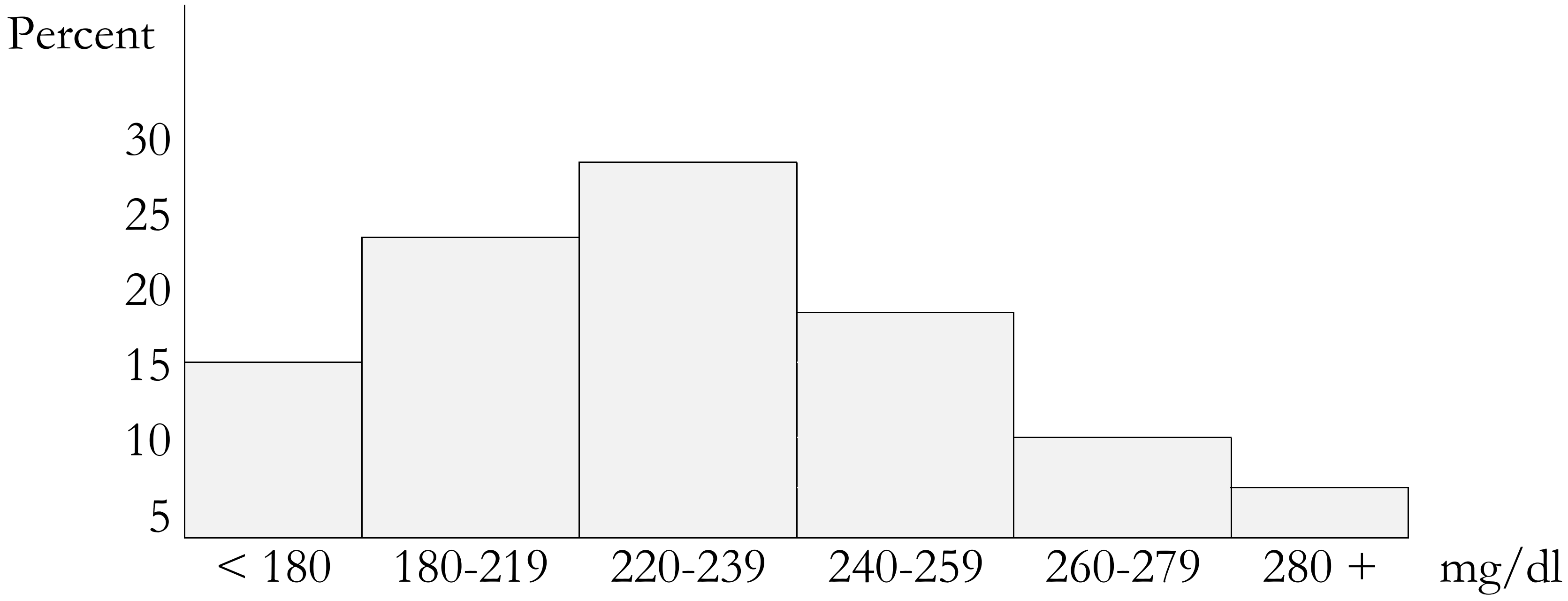
Rate: the ratio of a change in one quantity to a change in another quantity, usually time. (An event is regarded as a "change"). The rate is "relative" if the change in the numerator is divided by its base, e.g., "90 lung cancer cases per 100,000 population per year". In contrast, "120,000 new lung cancer cases per year" is an absolute rate.

“Capturing the phenomenon”

Survivorship

Distributions – the fuller picture

Serum cholesterol levels - Distribution



Common summary statistics for description and comparison

Mean – "Average" value of the variable

Median – Middle of the distribution of the variable – half of the values lie below and half lie above

Quartiles – Demarcate the 1st, 2nd, and 3rd quarter of the distribution of the variable

Percentiles – Demarcate a percentage of the distribution, e.g., 20th percentile (second decile) is the value below which the lowest 20% of the observations fall.

Standard deviation – Distance of a "typical" observation from the mean [not the same as "standard error"]

Interquartile range – Distance between the 1st and 3rd quartiles.

Skewedness – Degree of asymmetry about the mean value. Positively or right-skewed means distribution extends to the right: the mean lies to the right of the median, due to outlying values.

Kurtosis – Degree of peakedness relative to the length and size of its tails. Highly peaked is "leptokurtic"; flat is "platykurtic".

Summary statistics may not tell the whole story

Community health promotion

Mean alcohol consumption reduced by 1 drink/day:

- 5 drink/day reduction for each person in the highest consumption 20 percent of the population?
- 1.25 drink/day reduction for all people but those in the highest consumption 20%?

Summary statistics may not tell the whole story

Black-white differences in birth weight

Birth weight distribution has Gaussian ("normal") shape, minimum 500 g., maximum 5,000 g., mean 3,000 g.

Statistically smooth and reasonably symmetrical, but <u>biological implications</u> vary greatly:

- Birth weight 1,000-2,000 g. mortality rate = 33%
- Birth weight < 1,000 g. mortality rate = 75%.

Moral: need to know the shape of the distribution and implications for health of different values.

Heterogeneity and distributions of unknown factors

Any summary is a weighted average

Populations differ in characteristics which affect health.

Any overall number (e.g., mean, proportion) conceals subgroups that differ from the overall.

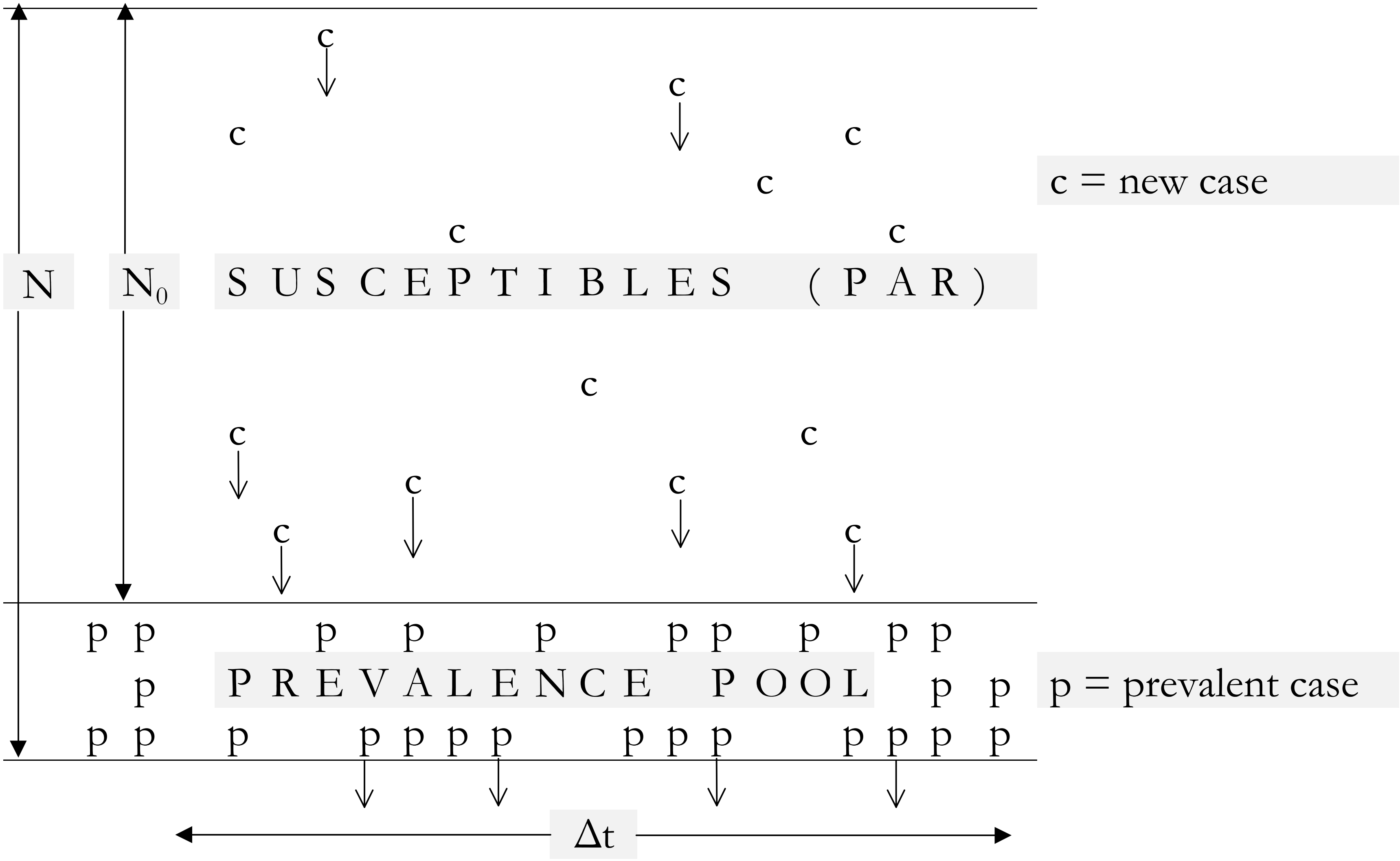
Important to know composition of group – basic demographic characteristics, important exposures, genetic make-up (?)

- Workforce experiences 90 lung cancer deaths per l00,000 per year

Most (overall) epidemiologic measures are weighted averages of across component subgroups (specific).

Can "adjust" or "standardize" to improve comparability.

(Rusty on weighted averages? See the appendix.)



N

N_0

SUSCEPTIBLES (PAR)

c = new case

PREVALENCE POOL

p = prevalent case

Δt

c's are incident (new) cases

Δt indicates the time interval

p's are prevalent (existing) cases

↓ exits from unaffected population or prevalence pool

prevalence odds = incidence × duration

$$\begin{aligned} \text{population size} = N &= \text{disease-free persons} + \text{prevalent cases} \\ &= N_0 + \text{prevalence pool} \end{aligned}$$

In stable population with constant incidence and prevalence:

$$\begin{aligned} \text{incident cases} &= \text{terminations} \\ \text{incidence} \times N_0 \times \Delta t &= \text{termination rate} \times \text{prevalence} \times N \times \Delta t \end{aligned}$$

$$\text{Prevalence} \times \frac{N}{N_0} = \frac{\text{Incidence}}{\text{Termination rate}}$$

$$\text{Termination rate} = \frac{\text{Terminations}}{\text{No. of cases} \times \Delta t} = \frac{1}{\Delta t}$$

Relationships among incidence, mortality, and prevalence

Virulence of the disease - Is it rapidly fatal?

Health care - When do cases come to medical attention?

Can cases be cured?

Does earlier detection alter prognosis?

Behavior - Do people recognize and act promptly on symptoms?

Do patients comply with treatment?

Competing causes of death - Are people with the disease likely to die of other causes?

Migration - Are people with the disease likely to leave the area?

Are people with the disease like to migrate to the area?

Prevalence versus incidence

	Prevalence	Incidence
Cases	Entities	Events
Source population	At risk to be a case	At risk to become a case
Time	Static (point)	Dynamic (interval)
Uses	Planning	Etiologic research

Considerations relevant for both prevalence and incidence

(Cases

1. **Case definition** – What is a case?

   Examples: arthritis, cholelithiasis, cardiovascular disease, diabetes, psychiatric disorder, epidemiologic treatment of syphilis or gonorrhea, prostate cancer

2. **Case development** – When is a case?

   Issues: induction, latency, progression, reversibility

   Examples: atherosclerosis, cancer, cholelithiasis, diabetes, hypertension, AIDS

3. **Case detection** – When is a case a "case"?

   Issues: Detectability is a function of technology and feasibility. What can be detected is not the same as what is detected.

   Examples: Atherosclerosis, breast cancer, cholelithiasis, osteoporosis, asymptomatic infections, prostate cancer

Considerations relevant for both prevalence and incidence

Source population [Population at risk (PAR)]

A. What is the relevant population — who is really "at risk"?
E.g., age (most diseases), sex (breast cancer), STD's and sexual activity, uterine cancer and hysterectomy, gallbladder cancer and cholecystectomy, genotypes?

B. What about previous manifestations?

Of the same disease? (influenza, tumors, injuries)

Of a related disease (stroke after CHD, cancer at a different site)

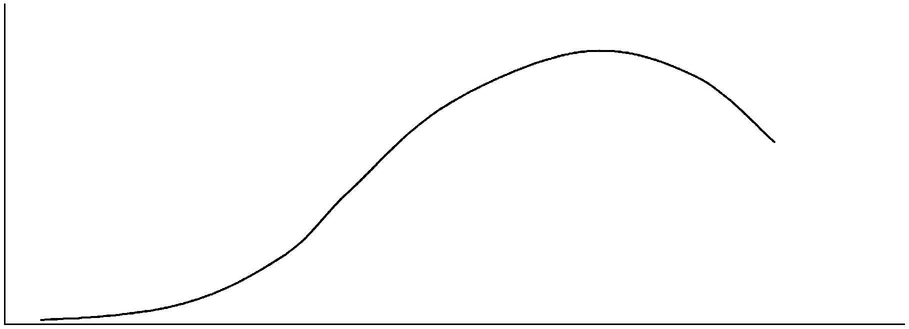
C. What about death from other causes? (competing risks)
E.g., deaths for diabetes reduce the rate of death from coronary artery disease, heart disease deaths reduce the rate of death from lung cancer to the extent that smokers are at excess risk for both

Passage of time [incidence only] – what period of observation?

A. Natural history of the disease - period of risk versus period of observation
E.g., atom bomb survivors and solid tumors, motor vehicle injury, congenital malformations

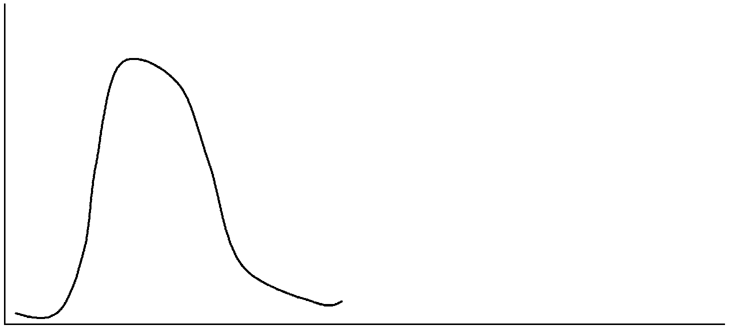
B. Different periods of observation for different subjects (does 1 person observed for 2 years = 2 people observed 1 year?)

C. Changes in incidence during the period (e.g., seasonal variation, secular change)



40 years

Cancer in atomic bomb survivors



3 years

Congenital malformations

Types of populations for follow-up

Cohort – defined at a point in time and monitored ("unit trust").

- Entrances into the cohort are not permitted; exits are problematic.
- Becomes smaller and older over time.
- May have been defined in the past (e.g., based on employment records) (**retrospective or historical cohort**)

Dynamic population – defined over a period ("mutual fund" analogy).

- Entrances and exits are expected
- Size and age distribution can change in either direction.

Types of incidence measures:

Cumulative incidence (CI) – proportion of a population who experience an event or develop a condition during a stated period.

$$\boxed{\mathrm{CI} \quad = \quad \frac{\text{New cases during stated period}}{\text{Number of persons at risk}}}$$

Incidence density (ID) – relative rate at which new cases develop in a population.

$$\boxed{\mathrm{ID} \quad = \quad \frac{\text{New cases during stated period}}{\text{Population-time}}}$$

Cumulative incidence (CI)

$$CI = \frac{\text{New cases during stated period}}{\text{Number of persons at risk}}$$

A. Population free of the outcome at baseline (the cohort);

B. All members are at risk of the event;

C. All first events are detected.

Example:

1,000 newly-trained police officers enter patrol duties.

33 suffer handgun injury during first six months.

967 officers carry out patrol duties with no handgun injuries.

6-month CI of handgun injury is 33/1,000 = 0.033.

Estimates six-month risk of handgun injury to new patrol officers.

Some things to note about CI

A. Explicit or implicit period of time (e.g., "6-month CI");

B. Proportion, so each person can be counted as a case only once, even has multiple events;

C. Proportion – direct estimate of risk (probability of event).

Sample calculation:

200 people free of disease X observed over 3 years

10 cases of X develop

3-year CI = 10 cases / 200 people = 10/200 = .05

3-year risk of X of average person, conditional on not dying from another cause, is 0.05.

Risk and odds

Risk – probability that event will occur in a stated or implied time interval – conditional on remaining "at risk" and "in view".

Odds – probability for / probability against: p/(1–p)

Example:

Risk = 0.05 ↔ odds = .05/.95 = 0.0526 (odds always > risk).

- Can be any non-negative number.
- Ln(odds) ["logit"] can be any real number
- Logit = 0 ↔ odds = 1.0 ("fifty-fifty") ↔ probability = 0.5.

[Rusty on logarithms? See the appendix on logarithms and exponents.]

Cumulative incidence when loss to follow-up

What if 20 of 200 died before developing X (i.e., not at risk for 3 years)?

Four principal alternatives:

A. Ignore the deaths: 3-year CI = 10/200 = .05

B. Ignore the people (analyze only those followed all 3 years):

3-year CI = 10/(200-20) = .056

C. Include only half:

3-year CI = 10/(200-20/2) = .053

D. Product-limit (if know when deaths occur): estimate risk by segment; (b) take inverse (1–risk); (c) multiply inverses; (d) convert back to 3-year risk.

All methods involve assumptions.

Incidence density (ID)

$$\boxed{\text{ID} = \frac{\text{New cases during stated period}}{\text{Population-time}}}$$

Sample calculation (10 cases among 200 people in 3 years):

ID = 10 cases / (200 people × 3 years) = 10 / 600 person-years

= .167 cases per person-year (py) = 0.167 / py = 167 / 1000py

Note:

A. ID is a <u>relative rate</u>, not a proportion.

B. <u>Units</u> must be stated or value is ambiguous (e.g., 15 cases/100,000 person-years = 15 cases/1,200,000 person-months).

C. In principle, instantaneous; in practice, compute average ID.

Calculating ID

Cases are same as for CI but can count multiple events per person.

Computation of population-time:

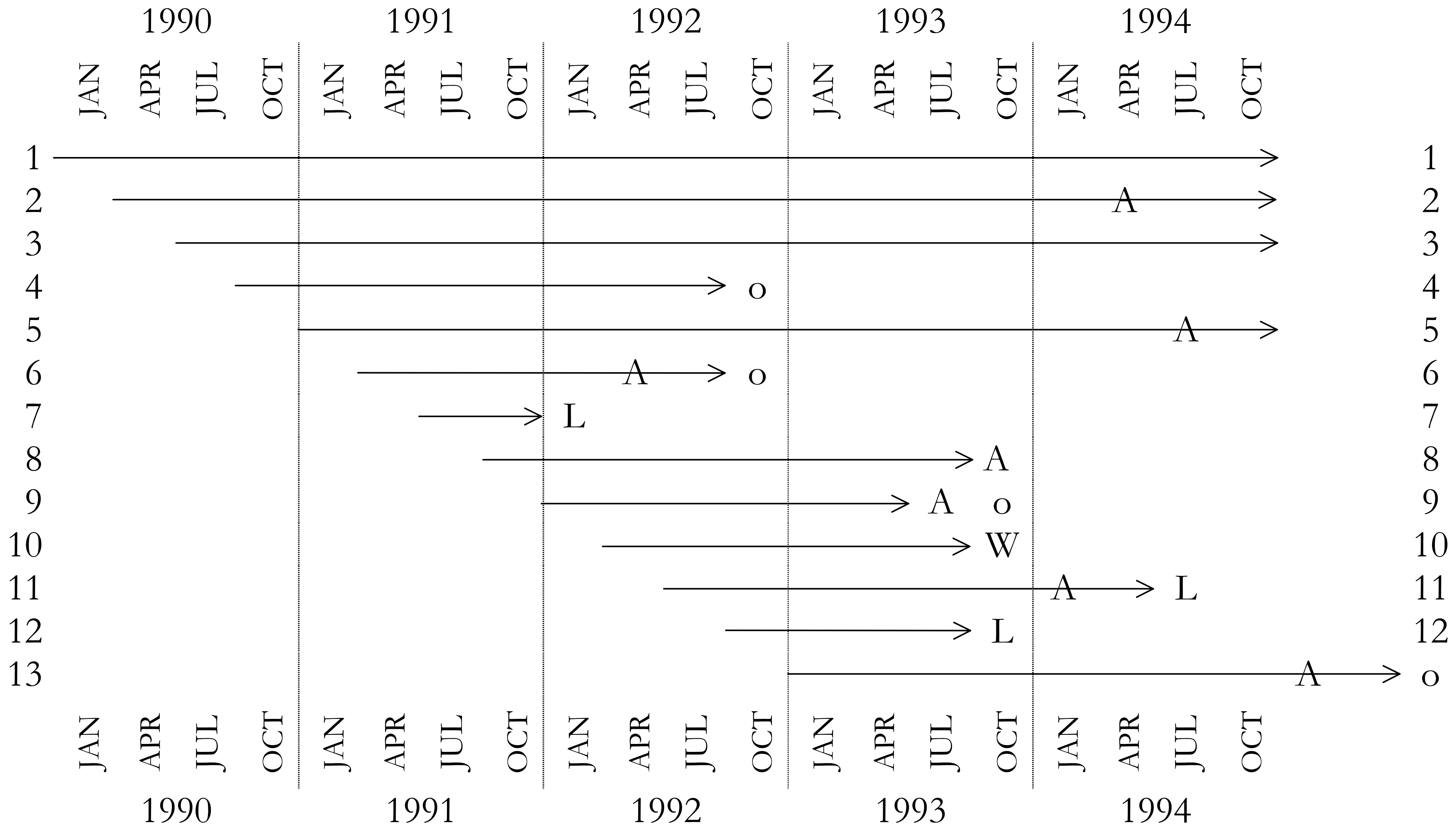
A. population-time = Σ (disease-free time for each person)

B. population-time = average population × length of period

a. Cohort: $\dfrac{N_0 + (N_0 - a - D - W)}{2} \times$ time interval

b. Dynamic population: $N_0 \times$ time interval

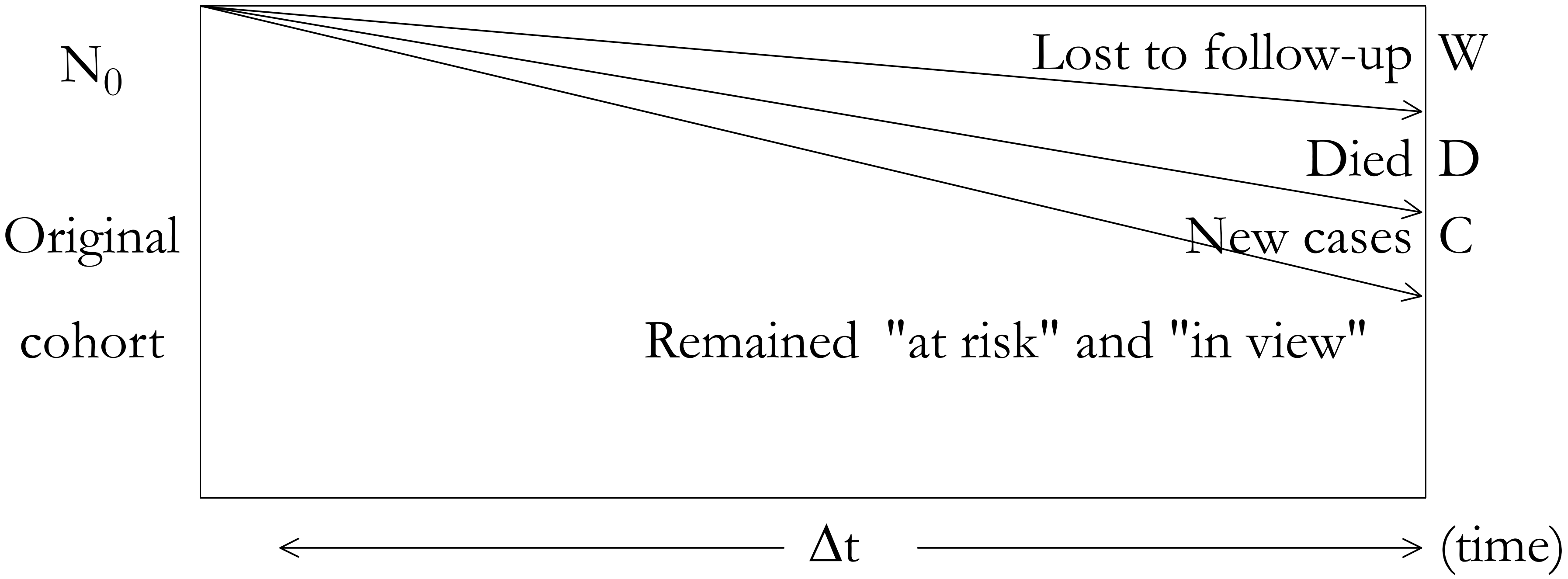
Person-time in a cohort, individual follow-up times are known



Key: A = admitted to nursing home care L = lost to follow-up W = withdrew o = died

Hypothetical experience of 13 advanced Alzheimer's patients cared for at home January 1990 - December 1993 and followed until December 31, 1994 for admittance to a nursing home, in order by study entrance date (after Kleinbaum, Kupper, and Morgenstern, 1982).

Cohort, individual follow-up times not known



N_0

Original

cohort

Lost to follow-up W

Died D

New cases C

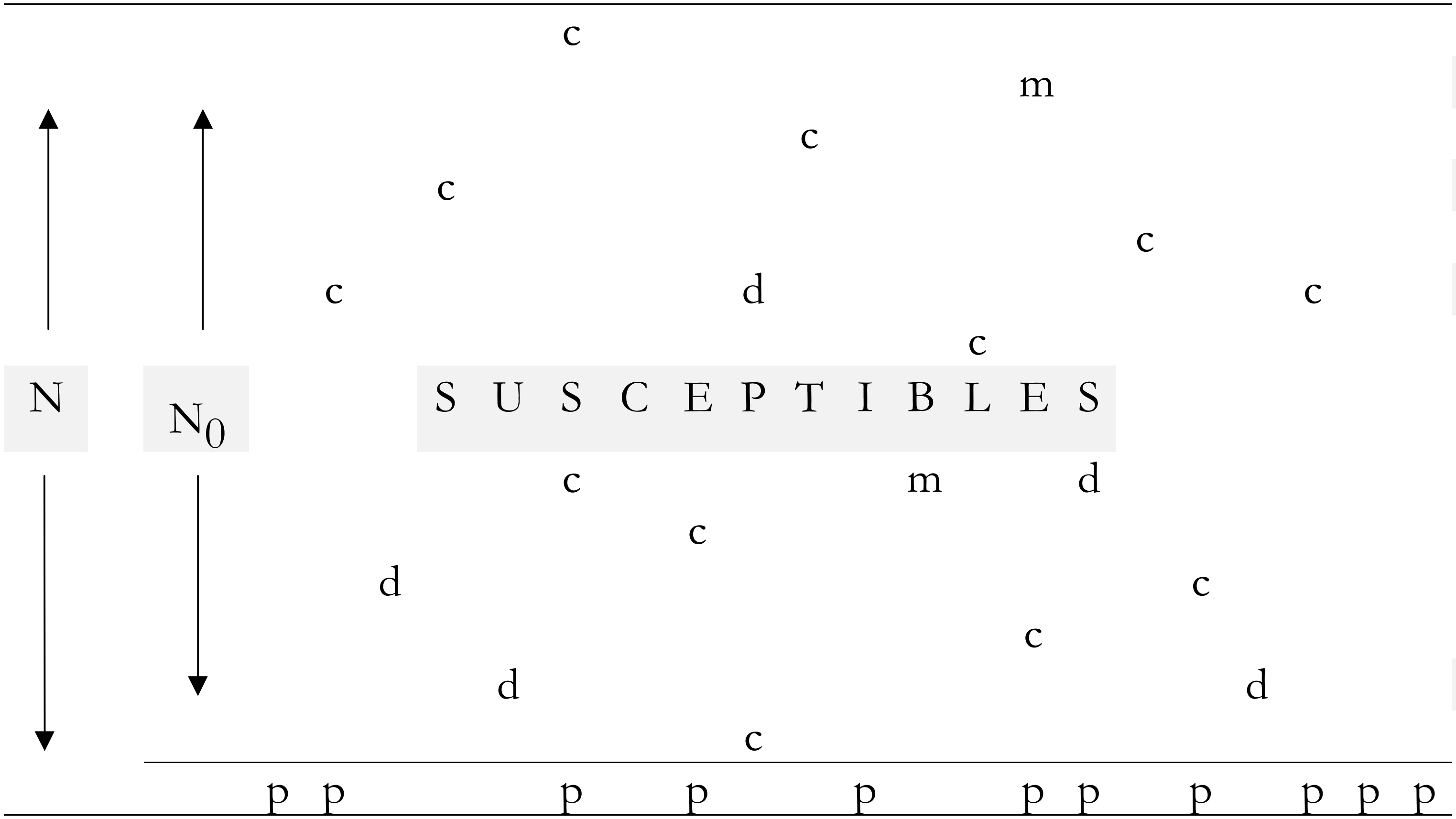
Remained "at risk" and "in view"

Δt (time)

$$ID = \frac{C}{(N_0 - W/2 - D/2 - C/2)\,\Delta t}$$

(Δt = time interval)

(Calculation of person-time in a stable, dynamic population



c = new cases

d = deaths

m = migrants

p = prevalent cases

$$ID = \frac{\text{cases}}{N_0(\Delta t)} \quad \text{or} \quad ID = \frac{\text{cases}}{N(\Delta t)}$$

(CI vs. ID - a real-life example

". . . substantially different occurrence rates of breast cancer: about **6.7 per thousand** (601/89,538) in the nurses cohort and about **18.2 per thousand** (131/7,188) in the NHANES cohort." (Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 1988;242:1259 quoted in Savitz DA et al., p.79, emphasis added)

Implication:

A. Different rates suggest errors in ascertainment of breast cancer

B. With under/overascertainment, there may be biased ascertainment

C. Bias may produce more complete or overdiagnosis among drinkers

CI vs. ID - a real-life example

". . . substantially different occurrence rates of breast cancer: about **6.7 per thousand** (601/89,538) in the nurses cohort and about **18.2 per thousand** (131/7,188) in the NHANES cohort." (Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 1988;242:1259 quoted in Savitz DA et al., p.79, emphasis added)

However:

Nurses Study: 601 cases/89,538 women over 4 years

4-year CI = 6.7 per 1,000, ID = 1.68 per 1,000 women-years

NHANES: 121 cases/7,188 women over 10 years (10 cases should have been excluded by Feinstein)

10-year CI = 16.8 per 1,000, ID = 1.68 per 1,000 women-years

Moral: Hazardous to compare CI's for different lengths of follow-up.

Two complementary measures of incidence: CI and ID

Cumulative incidence (CI)

A. increases with period of observation (i.e., it is "cumulative")

B. has problems with: multiple events, follow-up times that vary

C. does not require knowing exact time of onset of the disease

D. directly estimates risk

Incidence density (ID)

A. suggests ability to extrapolate over time - "duration free";

B. accommodates: multiple events, follow-up times that vary

C. does not require a cohort to estimate or interpret

D. may be more appropriate for etiologic inference

Choosing between CI and ID:

A. Objective

Estimate rate or risk

B. Natural history

Does the period of interest fit within the period of observation (restricted versus extended risk period)?

E.g., lifetime risk (CI) of death would be useless for comparing relative longevity of men and women.

C. Availability of data, e.g.

Fixed cohort, dynamic cohort, dynamic population

Different follow-up times

Knowing when events occur may favor one method or the other.