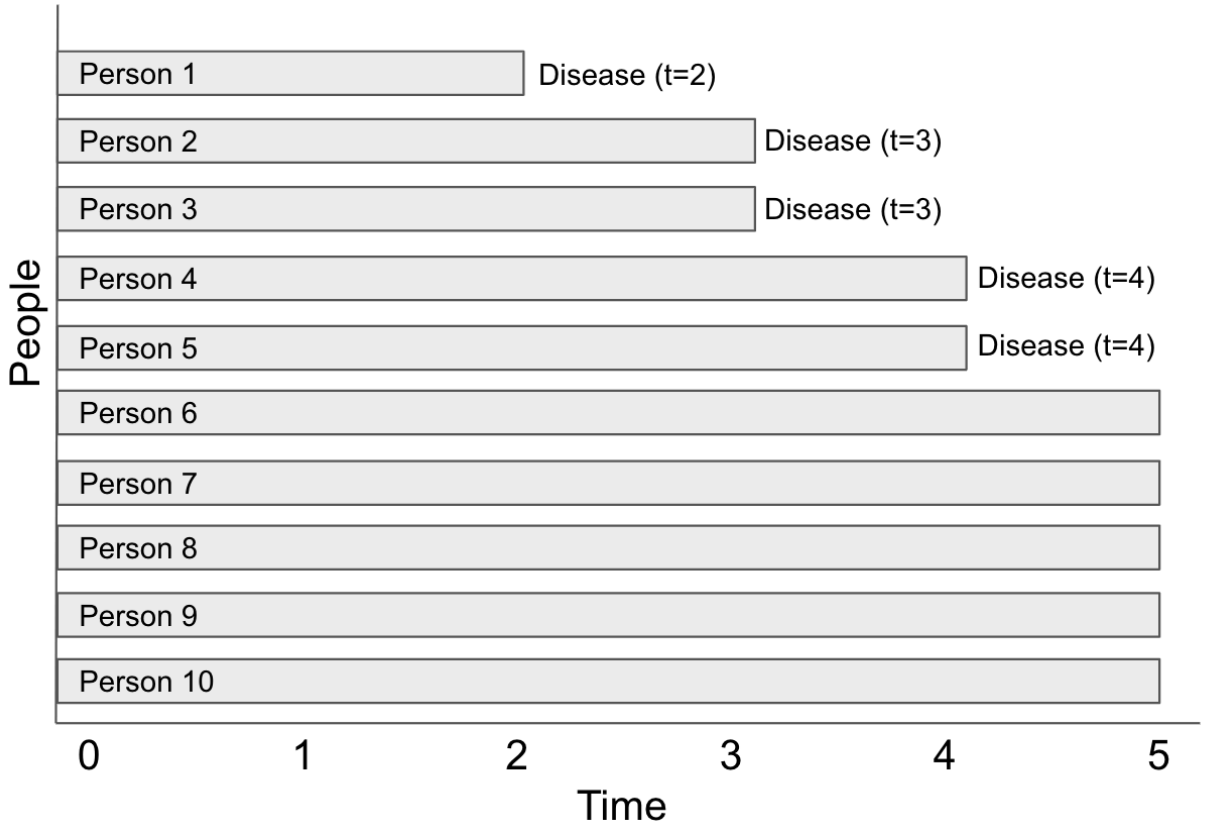
PH 250B Week 3, Tab 2 Practice Problems
Topic: Measures of Disease

The first five problems in this set are practice calculating cumulative incidence using the different techniques that we've learned and these five problems build on each other. We'd recommend doing them in order. In the answer key, we give a summary that compares and contrasts the different calculation methods in each of these scenarios. Our advice is that after you do each set of calculations you compare your results from the different methods and consider how and why they are different before looking at the answer summary.
The subsequent problems provide additional practice.

<u>Cumulative Incidence Problems</u>

Problem 1: Closed population, follow-up time for non-cases is equal, interval risk is not rare



Simple cumulative method

CI $= I / N_0$

CI = 5 / 10 = 0.50

Actuarial method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	0	0	0 / 10 = 0	1
2	1,2	10	1	0	1 / 10 = 0.1	0.9
3	2,3	9	2	0	2 / 9 = 0.222	0.778
4	3,4	7	2	0	2 / 7 = 0.286	0.714
5	4,5	5	0	0	0 / 5 = 0	1

$CI_{t0, tj}$ = 1 - (1 x 0.9 x 0.778 x 0.714 x 1) = 0.500

Kaplan Meier method

Time j	pop. at risk N'_j	incident disease I_j	interval risk R_j	interval survival S_j
2	10	1	1 / 10 = 0.1	0.9
3	9	2	2 / 9 = 0.222	0.778
4	7	2	2 / 7 = 0.286	0.714

$CI_{t0, tj}$ = 1 - (0.9 x 0.778 x 0.714) = 0.500

Density method

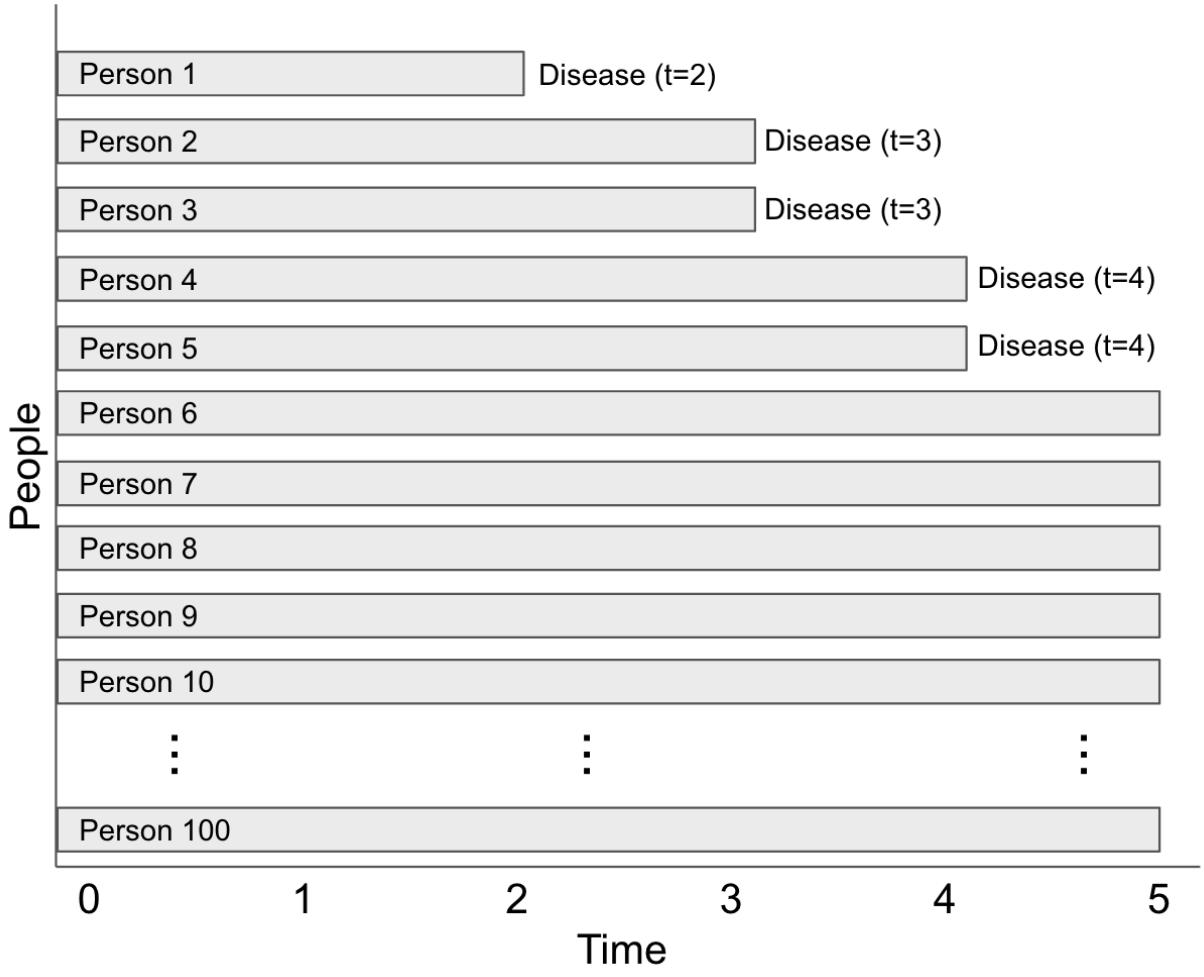
Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	10	0	1	0 / (10 x 1) = 0	0 x 1 = 0
2	1,2	10	1	1	1 / (10 x 1) = 0.1	0.1 x 1 = 0.1

3	2,3	9	2	1	2 / (9 x 1) = 0.222	0.222 x 1 = 0.222
4	3,4	7	2	1	2 / (7 x 1) = 0.286	0.286 x 1 = 0.286
5	4,5	5	0	1	0 / (5 x 1) = 0	0 x 1 = 0

$CI_{t0,\ tj}$ = 1 - e^(-(0 + 0.1 + 0.222 + 0.286 + 0)) =0.456

Summary: Each of the methods **except the density method** provides the same answer in the scenario in which there is equal follow-up time for all non-cases and no withdrawals. The density method differs because the interval risk was not rare (i.e., it was >10%).

Problem 2: Closed population, follow-up time for non-cases is equal, interval risk is rare



Simple cumulative method

CI = I / N_0

CI = 5 / 100 = 0.05

Actuarial method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	100	0	0	0 / 100 = 0	1
2	1,2	100	1	0	1 / 100 = 0.0100	0.9900

3	2,3	99	2	0	2 / 99 = 0.0202	0.9798
4	3,4	97	2	0	2 / 97 = 0.0206	0.9794
5	4,5	95	0	0	0 / 95 = 0	1

$CI_{t0, tj}$ = 1 - (1 x 0.9900 x 0.9798 x 0.9794 x 1) = 0.0500

Kaplan Meier method

Time j	pop. at risk N'_j	incident disease I_j	interval risk R_j	interval survival S_j
2	100	1	1 / 100 = 0.0100	0.9900
3	99	2	2 / 99 = 0.0202	0.9798
4	97	2	2 / 97 = 0.0206	0.9794

$CI_{t0, tj}$ = 1 - (0.9900 x 0.9798 x 0.9794) = 0.050

Density method

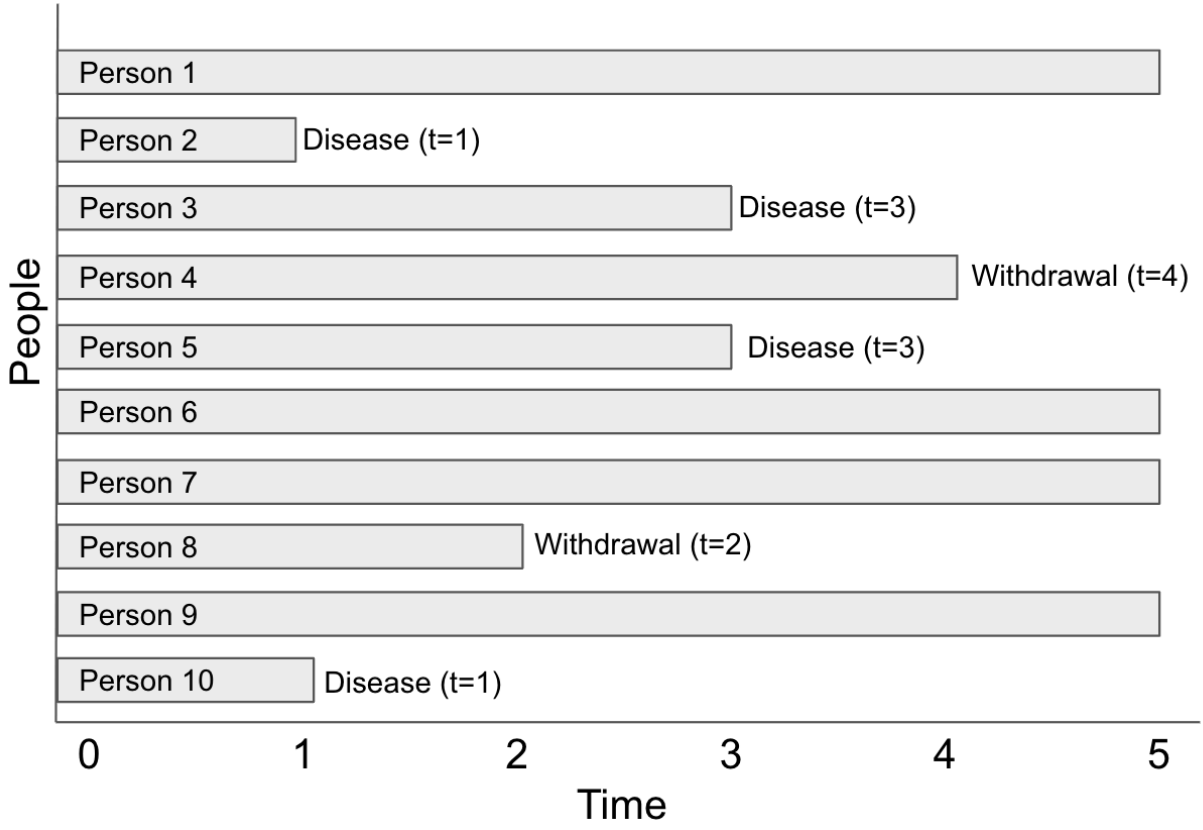
Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Interval duration Δt	Interval incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	0,1	100	0	1	0 / (100 x 1) = 0	0 x 1 = 0
2	1,2	100	1	1	1 / (100 x 1) = 0.01	0.01 x 1 = 0.01
3	2,3	99	2	1	2 / (99 x 1) = 0.0202	0.0202 x 1 = 0.0202
4	3,4	97	2	1	2 / (97 x 1) = 0.0206	0.0206 x 1 = 0.0206
5	4,5	95	0	1	0 / (95 x 1) = 0	0 x 1 = 0

$CI_{t0, tj}$ = 1 - e^(-(0 + 0.01 + 0.0202 + 0.0206 + 0)) =0.0495

Summary: Each of the methods **except the density method** provides the same answer when rounding to the three decimal points in the scenario in which there is equal follow-up time for all non-cases and no withdrawals and the interval risk is rare. This example shows that even when the interval risk is rare, the density method is still only an approximation of the other methods. For this reason when information other than the interval specific incidence densities is available, it is preferable not to use the incidence density methods.

Problem 3: Open population, follow-up time for non-cases is not equal, no interval has both withdrawal and disease, interval risk is not rare

(note: this was updated on 9/10/19)



Simple cumulative method

Not appropriate because there are withdrawals

Actuarial method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	2	0	2 / 10 = 0.200	0.800
2	1,2	8	0	1	0 / (8 - 1/2) = 0	1
3	2,3	7	2	0	2 / 7 = 0.286	0.714
4	3,4	5	0	1	0 / (5 - 1/2) = 0	1

5	4,5	4	0	0	0 / 4 = 0	1

$CI_{t0, tj}$ = 1 - (0.800 x 1 x 0.714 x 1 x 1) = 0.429

Kaplan Meier method

Time j	pop. at risk N'_j	incident disease I_j	interval risk R_j	interval survival S_j
1	10	2	2/10 = 0.200	0.800
3	7	2	2/7 = 0.286	0.714

$CI_{t0, tj}$ = 1 - (0.800 x 0.714) = 0.429

Density method

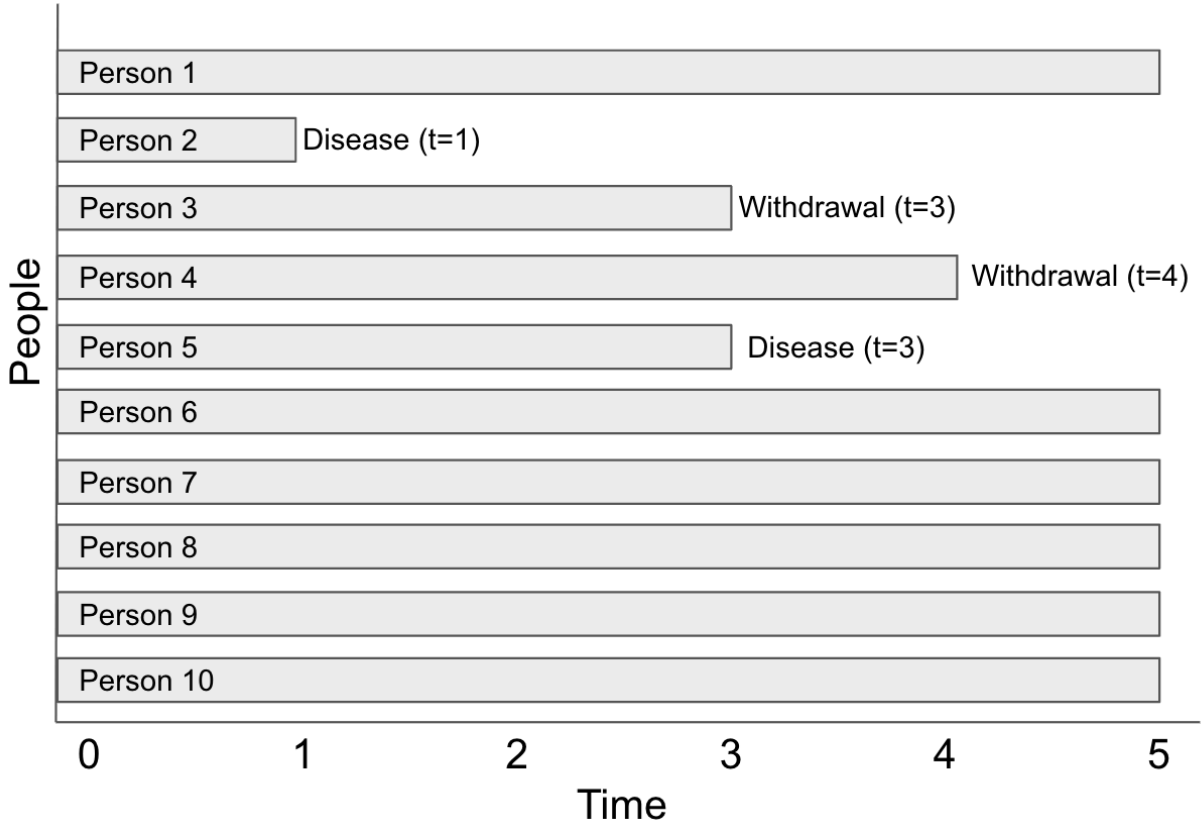
Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Interval duration $\triangle t$	Interval incidence density ID_j	Interval risk $ID_j \triangle t_j$
1	0,1	10	2	1	2/(10 x 1) =0.2	0.2 x 1 = 0.2
2	1,2	8	0	1	0 / (8 x 1) = 0	0 x 1 = 0
3	2,3	7	2	1	2 / (7 x 1) = 2/7=0.286	0.286 x 1=0.286
4	3,4	5	0	1	0 / (5 x 1) = 0	0 x 1 = 0
5	4,5	4	0	1	0 / (4 x 1) = 0	0 x 1 = 0

$CI_{t0, tj}$ = 1 - e^(-(0.2 + 0 + 0.286 + 0 + 0)) =0.385

Summary: The actuarial and Kaplan Meier methods provide the same answer in the scenario in which the population is open, follow-up time for non-cases is not equal, and no interval has both withdrawal and disease. The estimate from the density method differs because the interval risk is not rare (i.e., it is >10%).

Problem 4: Open population, follow-up time for non-cases is not equal, one interval has both withdrawal and disease, interval risk is not rare

(note: this was updated on 9/10/19)



Simple cumulative method

Not appropriate because there are withdrawals

Actuarial method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	0	1 / 10 = 0.1	0.900
2	1,2	9	0	0	0 / 9 = 0	1
3	2,3	9	1	1	1 / (9 - 1/2) = 0.118	0.882
4	3,4	7	0	1	0 / (7 - 1/2) = 0	1

5	4,5	6	0	0	0 / 6 = 0	1

$CI_{t0, tj}$ = 1 - (0.900 x 1 x 0.882 x 1 x 1) = 0.2062

Kaplan Meier method

Time j	pop. at risk N'_j	incident disease I_j	interval risk R_j	interval survival S_j
1	10	1	1/10 = 0.1	0.900
3	9	1	1/9 = 0.11	0.889

$CI_{t0, tj}$ = 1 - (0.9 x 0.889) = 0.200

Density method (note: assume withdrawals were followed through the end of the interval in which they withdrew)

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Interval duration $\triangle t$	Interval incidence density ID_j	Interval risk $ID_j \triangle t_j$
1	0,1	10	1	1	1 / (10 x 1) = 0.1	0.1 x 1 = 0.1
2	1,2	9	0	1	0 / (9 x 1) = 0	0 x 1 = 0
3	2,3	9	1	1	1 / (9 x 1) = 0.111	0.111 x 1 = 0.111
4	3,4	7	0	1	0 / (7 x 1) = 0	0 x 1 = 0
5	4,5	6	0	1	0 / (6 x 1) = 0	0 x 1 = 0

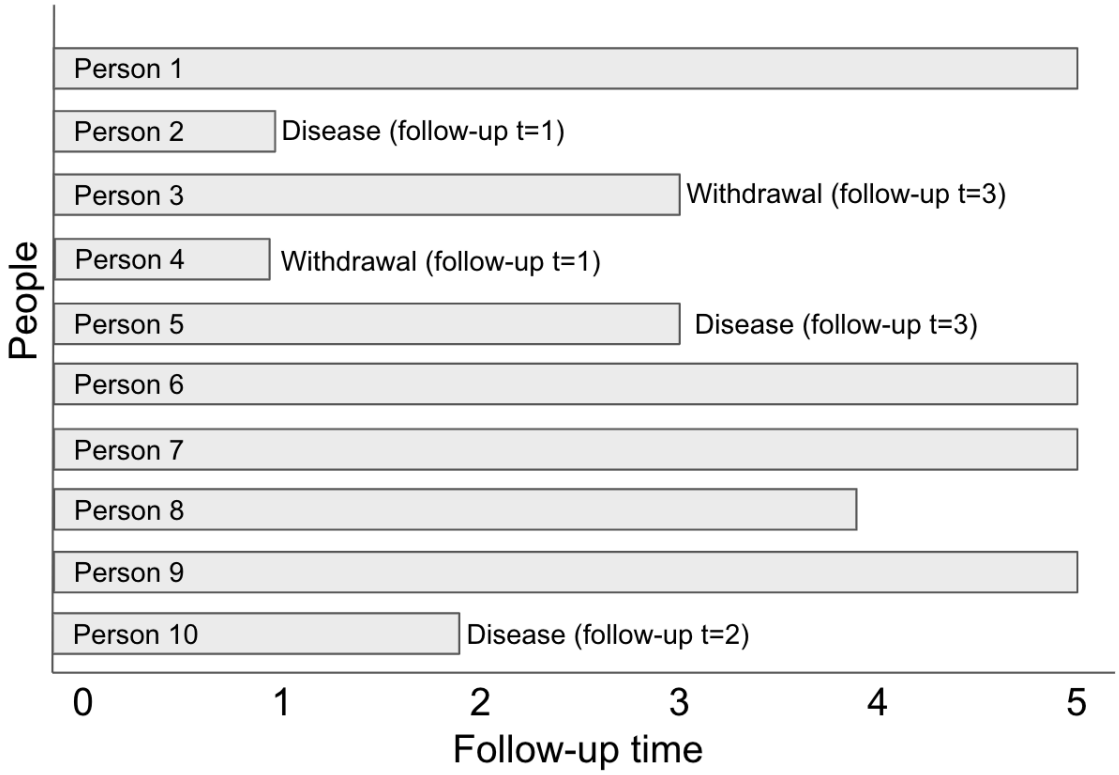
$CI_{t0, tj}$ = 1 - e^(-(0.1 + 0 + 0.111 + 0 + 0)) = 0.190

Summary: Each method provides a different answer in the scenario in which the population is open, follow-up time for non-cases is not equal, and at least one interval has both withdrawal and disease. The actuarial and Kaplan Meier methods differ because of the interval with both a withdrawal and an incident case (interval 2,3). This is because the actuarial method assumes that the withdrawal occurred halfway through the interval, but the Kaplan Meier method does

not. The estimate from the density method differs because the interval risk is not rare (i.e., it is >10%).

Problem 5: Open population, follow-up time for non-cases is not equal, at least one interval has both withdrawal and disease, and interval risk is not rare

In this example, “Time” on the x-axis is in months.



Calculate the incidence density

Person	Person-months of followup
1	5
2	1
3	3
4	1

5	3
6	5
7	5
8	4
9	5
10	2
Total	34

Incidence density: 3 / 34 = 0.088 per person-months

Cumulative incidence using the simple cumulative method

Not appropriate because there are withdrawals

Cumulative incidence using the Actuarial method

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Withdrawals W_j	interval risk (conditional risk) R_j	Interval survival S_j
1	0,1	10	1	1	1 / (10-1/2) = 0.105	0.895
2	1,2	8	1	0	1 / 8 = 0.125	0.875
3	2,3	7	1	1	1 / (7 - 1/2) = 0.154	0.846
4	3,4	5	0	0	0 / 5 = 0.000	1.000
5	4,5	4	0	0	0 / 4 = 0.000	1.000

$CI_{t0, tj}$ = 1 - (0.895 x 0.875 x 0.846 x 1.000 x 1.000) = 0.337

Cumulative incidence using the Kaplan Meier method

Time j	pop. at risk N'_j	incident disease I_j	interval risk R_j	interval survival S_j
1	10	1	1/10 = 0.100	0.900

2	8	1	1/8 = 0.125	0.875
3	7	1	1/7 = 0.143	0.857

$CI_{t0, tj}$ = 1 - (0.900 x 0.875 x 0.857) = 0.325

Cumulative incidence using the density method (note: assume withdrawals were followed through the end of the interval in which they withdrew)

Interval j	time period t_{j-1}, t_j	pop. at risk N'_{0j}	incident disease I_j	Interval duration $\triangle t$	Interval incidence density ID_j	Interval risk $ID_j \triangle t_j$
1	0,1	10	1	1	1/10 = 0.100	0.1 x 1 = 0.100
2	1,2	8	1	1	1/8 = 0.125	0.125 x 1 = 0.125
3	2,3	7	1	1	1/7 = 0.143	0.143 x 1 = 0.143
4	3,4	5	0	1	0/5 = 0.000	0.000 x 1 = 0.000
5	4,5	4	0	1	0/4 = 0.000	0.000 x 1 = 0.000

$CI_{t0, tj}$ = 1 - e^(-(0.100 + 0.125 + 0.143 + 0.000 + 0.000)) = 0.308

Summary: Each method provides a different answer in the scenario in which the population is open population, follow-up time for non-cases is not equal, at least one interval has both withdrawal and disease, and interval risk is not rare. The actuarial and Kaplan Meier methods differ because of the intervals with both a withdrawal and an incident case (intervals 0,1 and 2,3). This is because the actuarial method assumes that the withdrawal occurred halfway through the interval, but the Kaplan Meier method does not. The estimate from the density method differs because the interval risk is not rare (i.e., it is >10%).

Additional Practice Problems

Problem 6. (Fall 2017 250B Practice Exam Problem Set)

You have been hired by the Chinese CDC to investigate rates of chronic lung disease (CLD) among coal mine workers in China. You are given the following information about CLD cases, drawn from studies collecting data among miners all over the country.

Age group	No. of CLD cases	Person-time accrued (years)	Incidence density
16-25	1,500	750,000	0.00200
26-30	2,500	862,000	0.00290
31-35	3,000	811,000	0.00370
36-40	3,000	1,250,000	0.00240
Total	10,000	3,673,000	0.00272

a. Calculate and interpret the relevant overall (crude) measure of disease. What is this measure?

   Incidence density.
   The rate of CLD among Chinese coal miners is 27.2 cases per 10,000 person-years.

b. Calculate the relevant measure of disease for each age group, and transcribe your results to the space provided in the table above.

   See table

Problem 7. (Fall 2017 250B Practice Exam problem set)

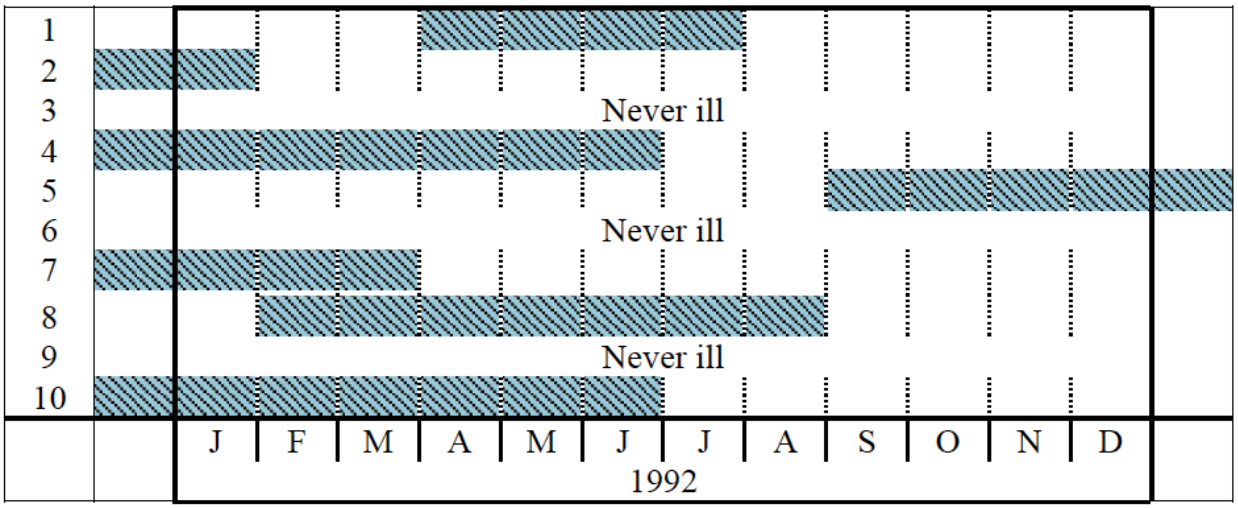
Incidence density is: (circle all that apply)

a) A proportion
b) The fundamental measure of disease occurrence
c) The same as the hazard
d) Best interpreted at the individual level
e) None of the above

Problem 8. (Fall 2017 250B problem set)

Time to disease in populations

The figure below summarizes the disease status of 10 individuals followed over a period of one year. The blue shaded bars represent time measured continiously from the beginning of disease until death.



a. What is the point prevalence on May 1, 1992?
4 cases / 8 people alive in the population = 0.5
Please note that the denominator is the population of interest. This includes those who have developed the disease (otherwise it would not be a proportion) and excludes those who have died.

b. What is the cumulative incidence for March 1, 1992 to December 31, 1992? Interpret this measure.

2 incident cases in time period / 5 at risk at beginning of time period = 0.4

Only 5 at risk because person 2 had died and persons 4, 7, 8, and 10 already had the disease.
Interpretation: The probability of getting the disease between Mar 1 and Dec 31 among the at-risk population on Mar 1 was 0.4

c. What is the incidence density from March 1, 1992 to December 31, 1992? Interpret this measure.
Calculate the person-time at risk for the 5 people at risk at the beginning of the specified time period.
Person 1 was at risk in March = 1 month
Person 3 was at risk for the whole time period = 10 months
Person 5 was at risk from March through August = 6 months
Person 6 was at risk for the whole time period = 10 months
Person 9 was at risk for the whole time period = 10 months
Total person-time at risk = 37 months
Number of incident cases / person-time at risk = 2 cases/ 37 person-months = 0.054 cases per person-month
Interpretation: The rate of disease in the population at risk from Mar 1 1992 to Dec 31 1992 was 0.054 cases per person-month.

d. Which of the measures from parts b and c estimate risk and which estimate rate? What is the difference? Which measure can be applied to individuals?
CI estimates risk, which is the average risk of getting the disease over the time period, in the population. ID estimates rate or the average rate of disease occurrence over the time period in the population.
“The concept of risk for disease is widely used and reasonably well understood by many people. It is measured on the same scale and interpreted in the same way as a probability. Epidemiologists sometimes speak about risk applying to an individual in which case they are describing the probability that a person will develop a given disease. It is usually pointless, however, to measure risk for a single person because for most diseases the person simply either does or does not contract the disease. For a larger group of people we can describe the proportion who developed the disease” (Source: Epidemiology An Introduction, 2nd Edition. Kenneth J, Rothman. 2012).

Problem 9. (Fall 2017 250B problem set)
The following hypothetical data represent the 5-year follow-up experience of 1,000 newly diagnosed cases of ovarian cancer.
NOTE: if you do not get the exact same answer, it is likely due to rounding differences. Make sure that you did the calculations correctly, and don’t worry if the answer is slightly different.

a. To use the actuarial method to calculate the cumulative incidence of mortality due to ovarian cancer in each interval, what assumption(s) do you need to make?

1) All deaths and withdrawals occur uniformly throughout the 1-year intervals
2) Independence between censoring and survival
3) Lack of secular trends in incidence

J	Interval (t_{j-1}, t_j)	# study participants at risk at start of period N'_{0j}	# new cases I_j	# withdrawals W_j
1	(0, 1)	1000	118	58
2	(1, 2)	824	232	40
3	(2, 3)	552	80	32
4	(3, 4)	440	67	18
5	(4, 5)	355	29	48

b. Use the actuarial method to calculate the cumulative incidence of ovarian cancer mortality in each interval. Interpret the cumulative incidence for interval five.
Calculate the risk in each interval:

$$R_{(tj-1,tj)} = \frac{I_j}{N'_{0j} - (\frac{W_j}{2})}$$

Interpretation: Individuals diagnosed with ovarian cancer at time 0 who survive until time 4 have a 0.088 probability of dying between time 4 and 5. This measure is conditional on survival to time 4.

J	N'_{0j}	I_j	W_j	Interval Risk R_j
1	1000	118	58	118/(1000-(58/2))=0.122
2	824	232	40	232/(824-(40/2))=0.289
3	552	80	32	80/(552-(32/2))=0.149

4	440	67	18	67/(440-(18/2))=0.155
5	355	29	48	29/(355-(48/2))=0.088

c. Calculate the 5-year cumulative incidence. Interpret this measure.

J	N'_{0j}	I_j	W_j	Interval risk R_j	Interval survival S_j	Cumulative incidence $R_{(t0,tj)}$
1	1000	118	58	0.122	1-0.122=0.878	1-0.878=0.122
2	824	232	40	0.289	1-0.289=0.711	1-(0.878*0.711)=0.376
3	552	80	32	0.149	1-0.149=0.851	1-(0.878*0.711*0.851)=0.469
4	440	67	18	0.155	1-0.155=0.845	0.551
5	355	29	48	0.088	1-0.088=0.912	0.591

First calculate survival in each interval: S(6C,7,6C) = 1 − R(6C,7,6C)
Next calculate cumulative incidence based on the survival in each interval:
R(t0,tj) = 1 − Π 1 − R (tj-1, tj) = 1 − Π(S tj-1, tj)
The 5-year cumulative incidence is 0.591.
Interpretation: The probability/risk of dying of ovarian cancer over the 5-year period after diagnosis was 0.591

d. Calculate the 5-year cumulative survival. Interpret this measure.

J	N'_{0j}	I_j	W_j	Interval risk R_j	Interval survival S_j	Cumulative incidence $R_{(t0,tj)}$	Cumulative survival $S_{(t0,tj)}$

1	1000	118	58	0.122	0.878	0.122	1-0.122=0.878
2	824	232	40	0.289	0.711	0.376	1-0.376=0.624
3	552	80	32	0.149	0.851	0.469	1-0.469=0.531
4	440	67	18	0.155	0.845	0.551	1-0.511=0.449
5	355	29	48	0.088	0.912	0.591	1-0.591=0.409

Calculate the cumulative survival based on the cumulative incidence at each interval.
$S(t0,tj) = 1 - R(t0,tj)$
The 5-year cumulative survival is 0.409.
Interpretation: The probability of surviving for five years after diagnosis was 0.409.

e. Using the density method (using incidence density), estimate the interval risk (CI), cumulative 5-year risk, and cumulative 5-year survival for new cases of ovarian cancer. What assumptions must you make to calculate person-time (PT) using this data? What assumptions must you make to use ID (rate) to calculate CI (risk)?

J	N'_{0j}	I_j	Δt	Incidence density ID_j	Interval risk $ID_j \Delta t_j$
1	1000	118	1	118/1000=0.118	0.118 x 1 = 0.118
2	824	232	1	0.282	0.282
3	552	80	1	0.145	0.145
4	440	67	1	0.152	0.152
5	355	29	1	0.082	0.082

Calculations of person-time **assume** withdrawals occur at midpoint, on average

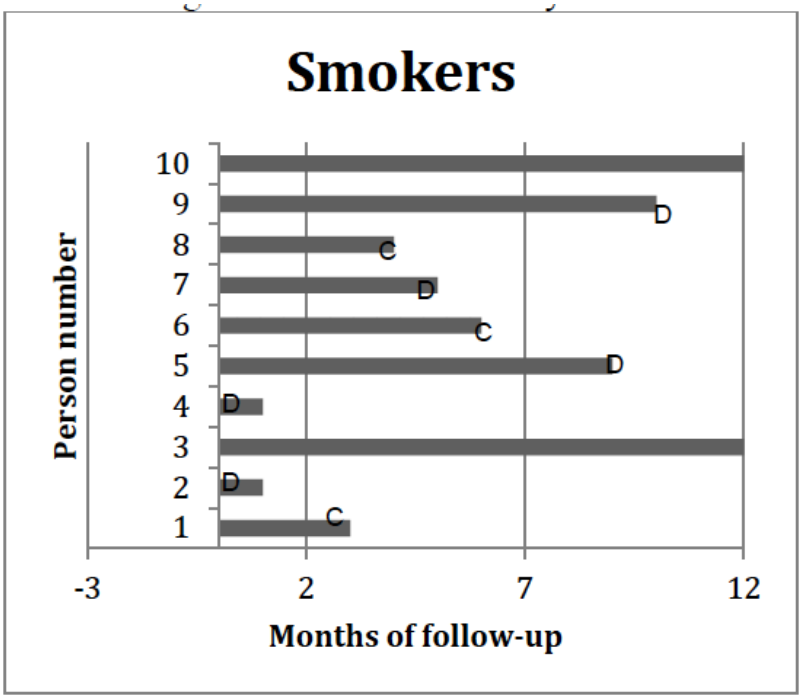
Using ID (rate) to calculate CI (risk) **assumes** a closed cohort with no competing risks and that the ID is constant within the interval (this last assumption is reasonable for short intervals)

$CI_{t0, tj}$ = 1 - e^(-(0.118 + 0.282 + 0.145 + 0.152 + 0.082)) = 0.541

f. Imagine you are conducting a study to measure the incidence of mortality following acute myocardial infarction. You want to use the actuarial method to calculate cumulative incidence. The probability of mortality is greatest in the period when the symptoms first show and decreases over time. Given this information, how would you construct the time intervals for data collection so that you could use the actuarial method to calculate cumulative incidence?

The actuarial method assumes that events and withdrawals occur at a uniform rate in each interval. Intervals should be shorter in the period following the onset of symptoms and longer later in the study. (Szklo Pg 52)

Problem 10. (Fall 2017 250B Practice Exam problem set)
Smoking and all-cause mortality



a. In this study of smoking and mortality, data were collected continuously for 12 months or until event (mortality). D indicates death, and C indicates last known contact with censored individuals. Note that 2 individuals were alive and uncensored at the end of the study. Use the Kaplan-Meier method to estimate the cumulative incidence by filling in the table below (use as many rows as necessary).
Cumulative incidence among smokers

$Time_J$	N_J	I_J	Interval risk (R_J)	Interval survival S_j	Cumulative incidence	Cumulative survival
1	10	2	2/10=0.2	0.8	0.2	0.8

5	10-2-2=6	1	⅙=0.167	0.8333	1-(0.8*0.833)=0.333	0.667
9	6-1-1=4	1	¼=0.25	0.75	1-(0.8*0.8333*0.75)=0.500	0.5
10	4-1=3	1	⅓=0.333	0/667	1-(0.8*0.833*0.75*0.667)=0.667	0.333

Cumulative incidence = 0.667 over 12 months (no need to state units besides time)

b. The cumulative incidence among 10 non-smokers during the same period was 0.260. Calculate an absolute measure of association between smoking and cumulative incidence of death. Name this measure. Interpret this measure.

Absolute measure of association: Risk Difference = Re-Ru=0.667-0.260=0.407
Interpretation: The risk of death among smokers was 40.7 percentage points higher over 12 months compared to non-smokers.
Could accept other absolute measures as well.

c. The following issues arose during this study. Which of the following violates an assumption needed in using the Kaplan-Meier method to estimate cumulative incidence? Multiple answers may be selected.

i. Smokers were more likely to be lost to follow-up than non-smokers.
Ii. People who were lost to follow-up did not withdraw from the study at the midpoint of intervals of observation.
Iii. the rate of smoking (cigarettes smokers per person-day) in the study population increased starting in the 7th month of follow up.

Issue (i) might appear to be referring to the assumption of non-independence (i.e. association between) censoring and the outcome (in this case, mortality), but because it concerns the exposure (smoking) and not mortality, it is not a violation of that assumption. Were people lost to follow-up more likely to die than those who weren't lost to follow-up, the independence of censoring and survival assumption would be violated, and none of the methods, including Kaplan-Meier (KM), could validly estimate the cumulative incidence.

Issue (ii) is related to the assumption of "uniformity of events and losses within an interval" – it's getting at the idea that withdrawals (losses) need to happen evenly throughout an interval. This assumption is not required for the KM method. (Of note, the uniformity assumption isn't actually articulated correctly here: it doesn't require withdrawals to occur in the middle of the interval – it just requires them to be evenly spaced throughout the interval, such that on average withdrawal occurs in the middle of the interval, which allows us to divide withdrawals by 2 in the actuarial method, for example).

Issue (iii) describes a secular trend in the exposure. A key assumption of all cumulative incidence calculation methods is that of no secular trends over the time period for which cumulative incidence is being calculated. Therefore, this issue prevents us from validly estimating cumulative incidence. To work around this, we could estimate cumulative incidence before and after the 7^{th} month so no secular trend is occurring during the period during which we're calculating cumulative incidence.