Homework 1 (60 pts)
Due 1/16/24, at the start of class

Goals:
• Familiarization with the Framingham Data Set
• Study Design Review
• Examine age-period-cohort effects

Question 1. The Framingham data are described in the pdf document "Framingham Longitudinal Data Documentation" on the Canvas site. Read this documentation thoroughly (really do this – a clear understanding of the structure of the data and the variables included will pay off for a number of future assignments). (30 pts)

Now propose three separate study questions that could be addressed with these data . Propose one study question you would examine with a case-control study, one question you would examine in a cross-sectional study and one question you would examine with a cohort study. They may be relatively obvious questions to address with Framingham data, or more creative questions, simple or complicated.

Frame each question BOTH as an 1. open-ended study question (e.g. Does a vegetarian diet protect against colon cancer, independent of caloric consumption?) and as a 2. hypothesis (e.g. Vegetarian diet decreases risk of incident colon cancer, independent of caloric consumption.) (Note -- this example has no relevance to the Framingham data). This is not necessarily a challenging exercise, but I do want everyone to be careful and specific with your language.

Question 2. This question is meant to give you a more concrete sense of the age-period-cohort problem in data analysis, and also to show you what a Stata data file looks like with a data set that is really small so you can list and view the whole dataset at once. (30 pts)

"lungca.dta" contains data from the Connecticut tumor registry (the first cancer registry in US) and includes the number of incident lung cancer cases by age group among men recorded in the Connecticut tumor registry from 1935 to 1984. Person-years of observation are also included. The variables in the dataset are

age (7 categories from 20-29 to 80-89)

period1 – period5: the number of cases in each of five time periods (period1=1935-1944; period2=1945-1954, period3=1955-1964, period4=1965-1974, period5=1975-1984).

den1-den5: number of person-years of observation in the corresponding 5 denominators (den1=1935-1944; den2=1945-1954...den5=1975-1984).

List the data so you can see what they look like and how they are organized. There are seven lines of data – one for each age group. And there are 11 variables in each line of data (one age group, 5 periods and 5 denominators).

Data Editor (Edit) — lungca.dta

person-years

7R x 5C

	age	period1	period2	period3	period4	period5	den1	den2	den3	den4	den5
1	20-9	1	3	4	6	7	1537781	1380360	1555934	2322128	2769374
2	30-9	10	20	28	31	40	1406807	1615355	1632000	1863489	2343684
3	40-9	70	115	195	289	281	1258708	1493910	1800315	1727315	1800233
4	50-9	247	543	885	1300	1418	1143763	1232189	1514848	1789483	1660060
5	60-9	395	1057	1992	2780	3769	770224	980496	1095932	1336181	1561113
6	70-9	209	790	2001	3017	4354	437017	567892	723242	756609	941025
7	80-9	60	231	673	1453	2270	145147	207148	273417	345493	389562

Age group

	age	Inc1	Inc2	Inc3	Inc4	Inc5
1	20-9	6.50e-07	2.17e-06	2.57e-06	2.58e-06	2.53e-06
2	30-9	7.11e-06	.0000124	.0000172	.0000166	.0000171
3	40-9	.0000556	.000077	.0001083	.0001673	.0001561
4	50-9	.000216	.0004407	.0005842	.0007265	.0008542
5	60-9	.0005128	.001078	.0018176	.0020806	.0024143
6	70-9	.0004782	.0013911	.0027667	.0039875	.0046269
7	80-9	.0004134	.0011151	.0024614	.0042056	.0058271

Calculate incidence rates for each time period within each age group using the corresponding numerator and denominator variables. After you have calculated 5 time-period-specific incidence rates for each age group, you should list the data again. You should now have 16 variables on each line of data. Using the calculated rates, answer the following questions.

a. Describe in words the cross-sectional age effect in each decade from 1935 to 1985 in lung cancer incidence for men in Connecticut.
b. Describe in words the age effect within cohorts.
c. Is there a birth cohort effect?

It is essential practice to construct databases and understand existing datasets. The lungca.dta file is a "wide format" file for *aggregated* data. That means that there are "repeated" observations" for each unit (in this case an age group), all on one line of data. It is aggregated data as the unit of observation is aggregated number of incident cases or person-years, rather than individual. You can imagine a different way to arrange the same data where there would be only one numerator and denominator on each line, and an extra variable indicating which time period they represent. Then each of the current lines of data would turn into five lines, each with FOUR variables: age group, time period, a numerator and a denominator. This would be the "long format file" version of the same data, and you can reshape the wide file to the long file with one command in STATA. Some kinds of data analysis require a wide structure and others require a long structure, but you can shift between them easily.

```
.reshape long period den, i(age) j(decade)
```

Here, "reshape" is the command, "long" specifies the desired new data structure. Then you need to list EVERY VARIABLE to be transferred to the new data structure (just two in this case, period and den). An identification variable (i) and a repeat observation indicator (j) must always be specified. Here, age is the identification variable that needs to be on each record, whether wide or long format, and decade is the variable that tells you the observation number. The choice of the name "decade" was arbitrary – it will be a new variable when the data are reshaped.

Then list the data again. You can switch back with the following command

```
.reshape wide period den, i(age) j(decade)
```

The data should now look as they did originally – with a reordering of the variables.

You may use reshape command to generate dataset, which then could be plotted to help your answer the research questions list above.

On submission files. Microsoft Word documents are preferred, and PDF files are also accepted. Submit two additional files; one file should contain your Stata .do file and the other should contain Stata output .log file.

On formatting Stata output. When you cut and paste Stata output into a Word document, the columns don't line up and the output is difficult or impossible to read. Instead, after pasting the output into your Word document, change the font to a monospaced font such as `Courier New` or Consolata. You may also need to change the font size so that lines don't wrap around; Finally, you

may need to start on a new page in order to keep a table from being split across pages.

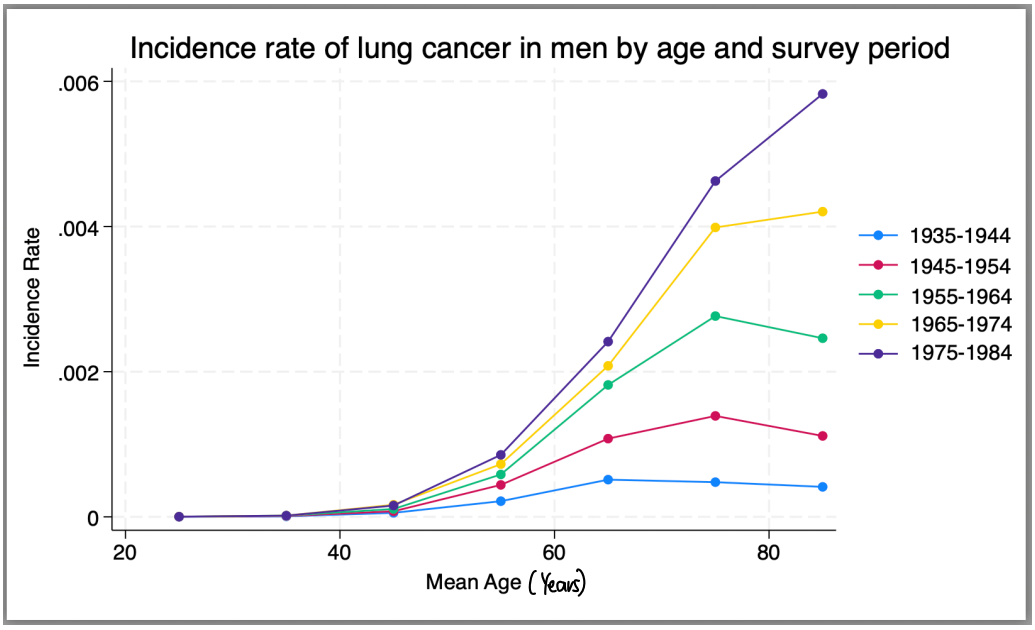


Fig 1.

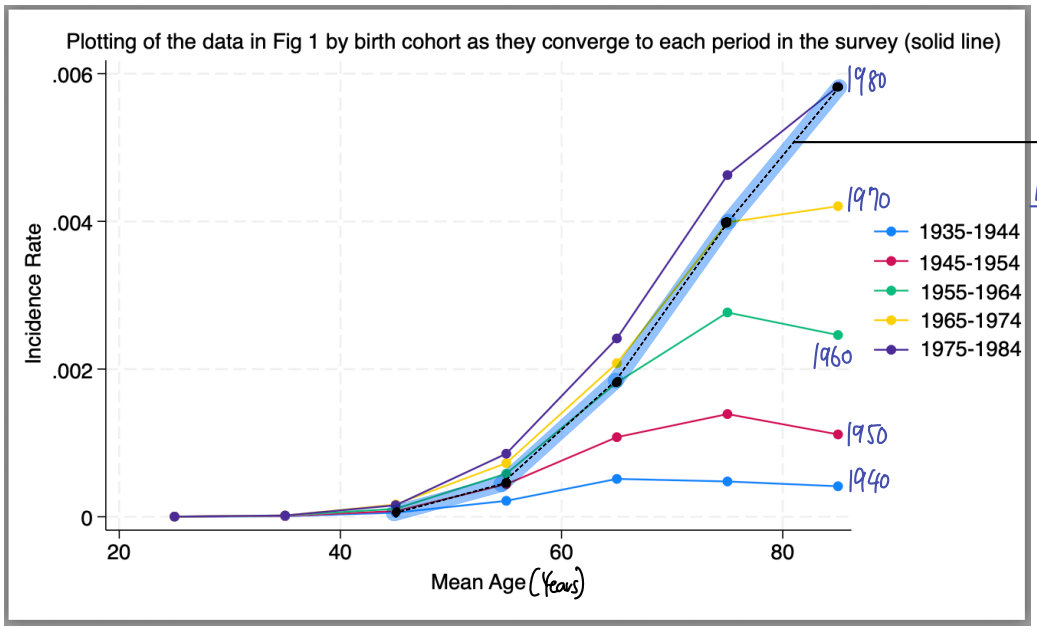


Fig 2.

	age Mean	Inc1 (1940)	Inc2 (1950)	Inc3 (1960)	Inc4 (1970)	Inc5 (1980)	* Birth Year Mean
1	20-9 25	6.50e-07	2.17e-06	2.57e-06	2.58e-06	2.53e-06	1915
2	30-9 35	7.11e-06	.0000124	.0000172	.0000166	.0000171	1905
3	40-9 45	.0000556	.000077	.0001083	.0001673	.0001561	1895
4	50-9 55	.000216	.0004407	.0005842	.0007265	.0008542	1885
5	60-9 65	.0005128	.001078	.0018176	.0020806	.0024143	1875
6	70-9 75	.0004782	.0013911	.0027667	.0039875	.0046269	1865
7	80-9 85	.0004134	.0011151	.0024614	.0042056	.0058271	1855

Epidemiologic Methods (PBHS 31001, Winter 2024)

Alexandra Chang
01/14/2024

Homework #1

Problem 1.

Solution.

Case-control study

- Question: Are individuals with a history of cardiovascular disease (cases) more likely to have *increased cholesterol* than other individuals without cardiovascular disease (controls) in the Framingham population?

- Hypothesis: There is a higher prevalence of *increased cholesterol* among individuals history of cardiovascular disease (cases) compared to those without (controls) in the Framingham population.

Cross-sectional study

- Question: What is the prevalence of *hypertension* among different age groups, and how does it vary by *gender* in the Framingham population?

- Hypothesis: The prevalence of *hypertension* will increase with age, and there will be variances in *gender* distribution of hypertension within each age group in the Framingham population.

Cohort study

- Question: Are individuals with a *cigarette smoking* habit (exposed) more likely to develop cardiovascular disease than other individuals (unexposed) in the Framingham population?

- Hypothesis: Individuals who engage in *cigarette smoking* habit (exposed) will have an increase risk in developing cardiovascular disease compared to those without (unexposed) in the Framingham population.

Problem 2.

Solution. (Refer to the STATA .do and .log file for complete steps)

Age Effect

Cross Sectional Study

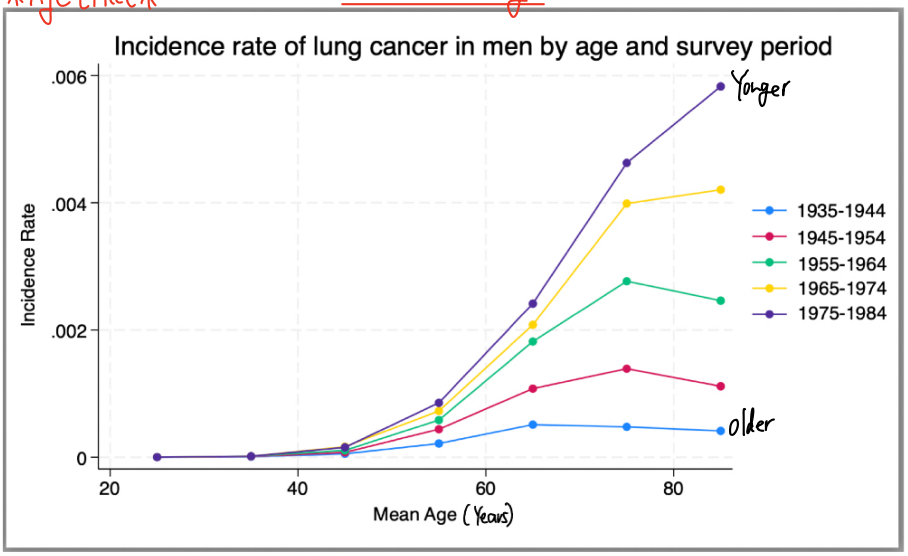
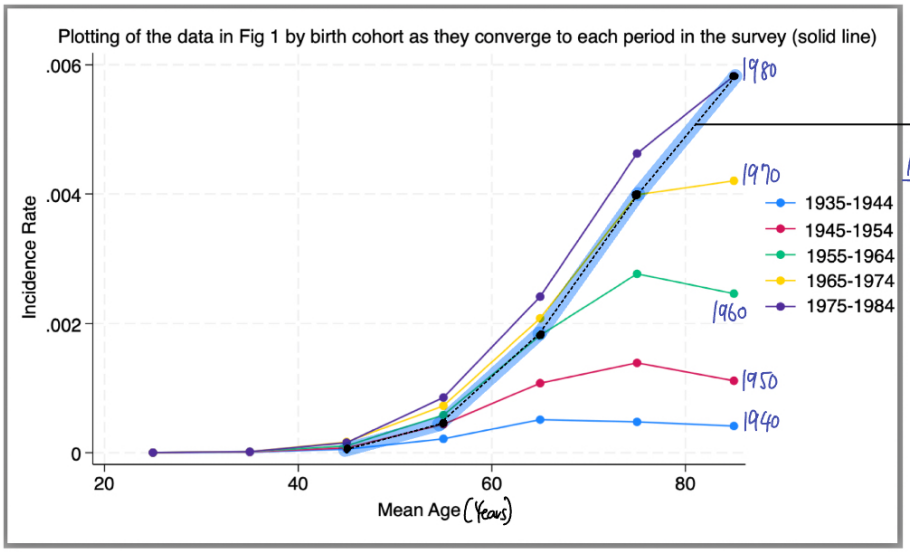


Fig 1.



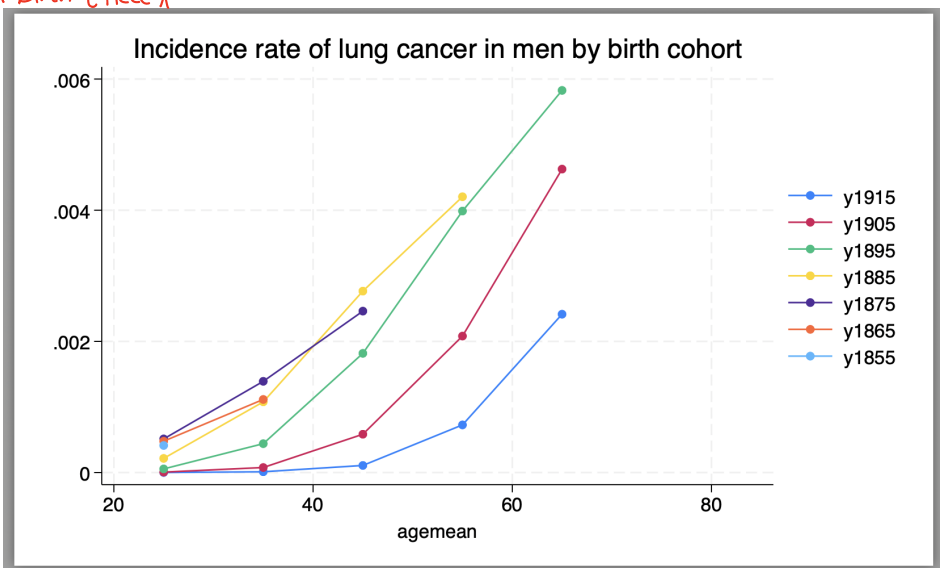
Example:
Group 3: Born in Year mean 1895

Mean
1940
1950
1960
1970
1980

Fig 2.

	age Mean	1940 Inc1	1950 Inc2	1960 Inc3	1970 Inc4	1980 Inc5	* Birth Year Mean
1	20-9 25	6.50e-07	2.17e-06	2.57e-06	2.58e-06	2.53e-06	1915
2	30-9 35	7.11e-06	.0000124	.0000172	.0000166	.0000171	1905
(3)	40-9 45	.0000556	.000077	.0001083	.0001673	.0001561	1895
4	50-9 55	.000216	.0004407	.0005842	.0007265	.0008542	1885
5	60-9 65	.0005128	.001078	.0018176	.0020806	.0024143	1875
6	70-9 75	.0004782	.0013911	.0027667	.0039875	.0046269	1865
7	80-9 85	.0004134	.0011151	.0024614	.0042056	.0058271	1855

Birth Effect



1. In **cross-sectional** studies, age effects are analyzed by *comparing (incidence rates) within different age groups at a particular point in time.*

   (a) The overall trend in lung cancer incidence rates for men in Connecticut from 1935 to 1985 indicates a general increase across all ages across decades. While there is an overall upward trend in lung cancer incidence for all age groups over the decades, there is a distinctive deviation during periods 1, 2, and 3 (blue, red and green), particularly noticeable among individuals aged 80-89, where the rates experience a slight decline in the incidence rates of lung cancer among individuals aged 80-89.

2. In **cohort** studies, age effects are analyzed by *tracking (incidence rates) changes within a specific cohort as they age.*

   (a) The trend of age effect within cohorts for lung cancer incidence rates reveals an overall increase from 1935 to 1985, with variations in the slopes of these trends among different birth cohorts. The cohort born in 1895 (green) shows the most substantial increase, the cohort born in 1915 (dark blue) displays the least steepest slope, while making inferences for the cohort born in 1855 (light blue) remains challenging due to insufficient data for comparison.

3. Yes. In the context of the question about lung cancer incidence rates, a birth cohort effect would imply that individuals born in the same year range exhibit differences in lung cancer incidence rates compared to individuals born in other year ranges.

   (a) Examples: changes in smoking habits, environmental exposures, or advancements in medical care that are specific to certain birth cohorts.

Younger birth cohorts seem to experience longer incidence rates of lung cancer compared to older cohorts at all ages (exceptions present...)

Reference:

Age Effect: Change in the rate according to *age*, irrespective of birth cohort and calendar time.

Cohort Effect: Change in the rate according to *year of birth*, irrespective to age and calendar time.