Notes for first lectures Statistics 215 DA Freedman

Text. Freedman, *Statistical Models: Theory and Practice*, Cambridge (2005).

Bring the book to class.

Read chapters 1-2-3. Work exercises! W O R K E X E R C I S E S ! ! !

Office hour: 339 Evans 10–11 Thursday

Lab, 330 Evans, F 12–2. TA: Johann Gagnon-Bartsch

Read. Talk. Work exercises.

No Midterm. No Homework. Pop quizzes? Labs. Project. Final.

Final exam: Group 9, Saturday 12/15/07, 5–8 pm

The final exam will be given at the scheduled time.

The final exam will not be given at any other time.

What does B+ mean?

Boot camp on random variables (chap. 2) and matrices (chap. 3):

Wednesday 9/5/06, 3–5 pm, 332 Evans

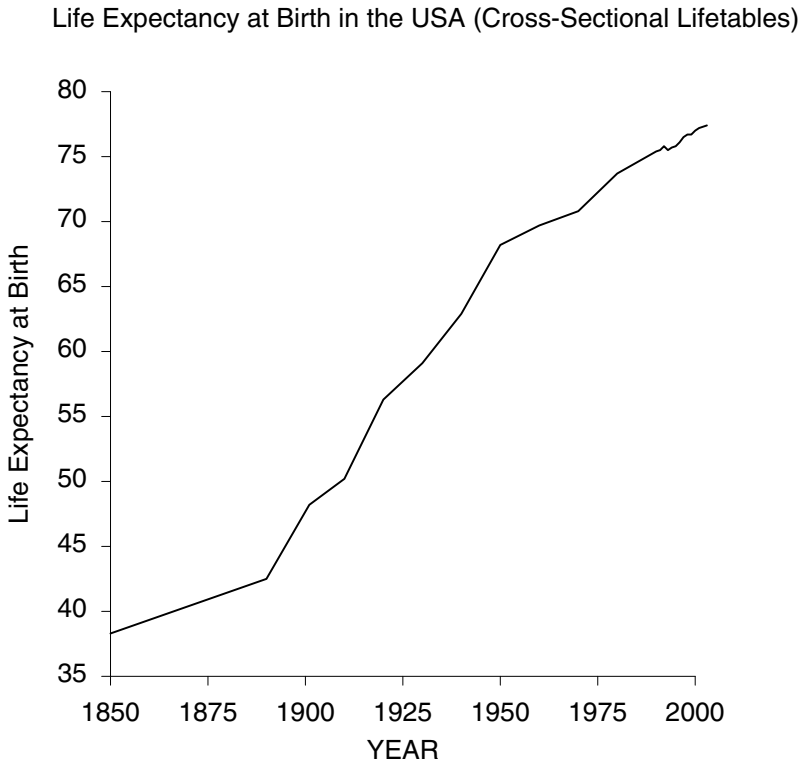
http://www.stat.berkeley.edu/users/census/rv.pdf

See my web page for corrections, other handouts, schedule.

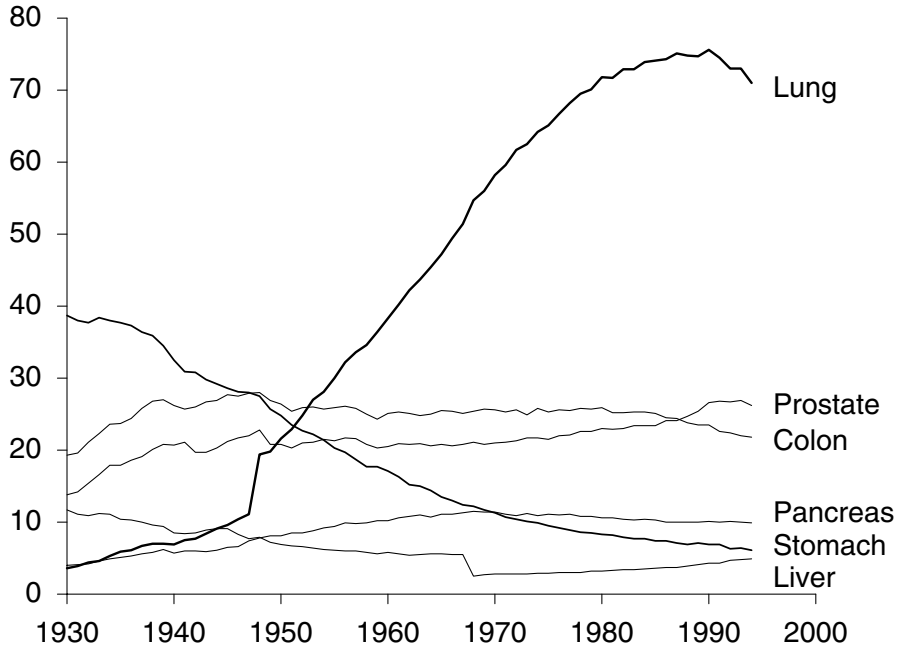
www.stat.berkeley.edu/users/census/index.html

If your e-mail has a short, crisp question with a short, crisp answer, I might answer it. Otherwise, try office hours or after lecture—or even during lecture.

The Demographic Transition. Life expectancy at birth was for centuries around 40 years, but death rates start to fall in Europe and North America around 1800, i.e., life expectancy started to go up.



Age-Standardized Cancer Death Rates for Males, 1930–96. Per 100,000. Direct method. Reference population is the US population of 1970. Data from the American Cancer Society.



Hospital-based case-control study. Smoking and lung cancer. Doll and Hill (1952).

	Cases	Controls
Smoker	1350	1296
Nonsmoker	7	61

The "odds ratio" is

$$\frac{1350/7}{1296/61} \approx 9,$$

i.e., lung cancer is about 9 times as common among the smokers. (This study "matched" the controls to the cases, which we're ignoring here.)

STUDIES VS ANECDOTES: THE CONTROL GROUP

Cohort studies. Follow people over time, see who smokes and who dies.

Men smoke more and get cancer more... Compare male smokers to male non-smokers. Controlling for gender.

People age 65–74 smoke more and get cancer more... Compare male smokers age 65–74 to male non-smokers age 65–74. Controlling for age and gender.

Indirect standardization: ratio of observed to expected deaths. Expected numbers are computed by applying age-specific rates for non-smokers to person-years at risk among current smokers of cigarettes. Veterans cohort.

Cause of death	Observed	Expected	Ratio
All causes	36,143	20,857	1.73
Respiratory	2,107	488	4.31
Cancer	7,608	3,590	2.12
Lung & bronchus	2,609	231	11.28
Larynx	94	8	11.49
Liver	176	75	2.33
Bladder	326	151	2.16
Brain	160	152	1.05
Cardiovascular	21,413	13,572	1.58
Stroke	2,728	2,075	1.32
Other			
Ulcer	365	92	3.97
Cirrhosis	404	150	2.69
No death certificate	849	390	2.17

RT Ravenholt, Population and Development Review, 1990. E Rogot & JL Murray, Public Health Reports, 1980, Vol 95, pp213–222. American Veterans Cohort, 16 years of followup. Study began in 1954/57, with 300,000 subjects. Mainly white males, veterans of World War I, with US Gov't life insurance.

EXPERIMENTS VS OBSERVATIONAL STUDIES. SOME SUBJECTS GET TREATMENT (EXPOSURE). SOME DON'T (THE CONTROLS). WHO DECIDES? HOW??

CONFOUNDING

Direct standardization. The standardized death rate from cancer of type j in year t is

$$\sum_i n_i d_{ijt} \Big/ \sum_i n_i,$$

where n_i is the number of men in age group i in the reference population, and d_{ijt} is the death rate from cancer of type j among men in age group i in the population corresponding to year t. Actual death rates for year t are applied to a population whose age distribution does not change with t . . . .

Indirect standardization. The expected number of deaths among the smokers from cause j is

$$\text{Exp} = \sum_i n_i d_{ij},$$

where (i) n_i is the number of person years at risk in age group i among the smokers, and (ii) d_{ij} is the death rate from cause j among non-smokers in age group i, i.e., the number of deaths in age group i among non-smokers divided by number of non-smoker person years at risk. The ratio Obs/Exp compares age-specific death rates of smokers to non-smokers.

Person years at risk.

(i) Tom is born in 1900. A non-smoker, he is recruited to the Veterans study in 1954. He dies in a hunting accident in 1980. He contributes one person-year of observation for age 54, 55, . . . , 79. From the point of view of our table, he is "censored" at age 80. If we had a line for "accidents," he would contribute an event there at age 80.

(ii) Dick is born in 1890. A smoker, he is recruited to the Veterans study in 1957. He dies of a heart attack in 1980. He contributes one person-year of observation for age 67, 68, . . . , 89. He contributes an event to the "Cardiovascular" line at age 90. From the point of view of the other lines, he is censored at age 90.

(iii) Indirect standardization is just like direct, except, the "standard population" consists of the exposed, and we count person-years of exposure rather than persons. Indeed, let n_i be person-years at risk among the smokers, and $n_i{}'$ be person-years at risk among non-smokers, in age group i. Likewise, let e_i and $e_i{}'$ be the event counts. Then Obs/Exp is

$$\frac{\sum_i e_i}{\sum_i n_i (e_i{}'/n_i{}')} = \frac{\sum_i n_i (e_i/n_i)}{\sum_i n_i (e_i{}'/n_i{}')}$$

(iv) The Veterans study determined exposure at baseline. Not unusual, but, not good.

Issues.

If we start modeling this (and they do, e.g., proportional hazards), two key assumptions to watch:

(a) Risk depends on age not "cohort" (i.e., birth year) or "period" (calendar time).
(b) Independence of competing risks.

The HIP trial: intention-to-treat analysis of experimental data. Breast cancer is one of the most common malignancies among women in the U.S. If it is detected early enough—before the cancer spreads—chances of successful treatment are much better. Do screening programs speed up detection by enough to matter? The first large-scale trial was run by the Health Insurance Plan of Greater New York, starting in 1963. The subjects (all members of the plan) were 62,000 women age 40 to 64. These women were divided at random into two equal groups. In the treatment group, women were encouraged to come in for annual screening, including examination by a doctor and X-rays. About 20,200 women in the treatment group did come in for the screening; but 10,800 refused. The control group was offered usual health care. All the women were followed for many years. Results for the first 5 years are shown in the table below. ("HIP" is the usual abbreviation for the Health Insurance Plan.)

Deaths in the first five years of the HIP screening trial, by cause. Rates per 1,000 women.

		Cause of Death			
		Breast cancer		All other	
		Number	Rate	Number	Rate
Treatment group					
Examined	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control group	31,000	63	2.0	879	28

Epidemiologists who worked on the study found that (i) screening had little impact on diseases other than breast cancer; (ii) poorer women were less likely to accept screening than richer ones; and (iii) most diseases fall more heavily on the poor than the rich.

(a) Does screening save lives? Which numbers in the table prove your point?
(b) Why is the death rate from all other causes in the whole treatment group ("examined" and "refused" combined) about the same as the rate in the control group?
(c) Why is the death rate from all other causes higher for the "refused" group than the "examined" group.
(d) Breast cancer (like polio, but unlike most other diseases) affects the rich more than the poor. Which numbers in the table confirm this association between breast cancer and income?
(e) The death rate (from all causes) among women who accepted screening is about half the death rate among women who refused. Did screening cut the death rate in half? If not, what explains the difference in death rates?