Cohort Studies

Outline

- A case study
- Design considerations
- Analytical considerations
- Cohort life tables
- Kaplan-Meier estimation
- Comparison of two sets of survival probabilities

Epidemiology: Study of a lifetime

- **In 1946, scientists started tracking thousands of British children born during one cold March week. On their 65th birthday, the study members find themselves more scientifically valuable than ever before.**
- Published online 1 March 2011 | *Nature* **471**, 20-24 (2011) | doi:10.1038/471020a

National Survey of Health and Development (NSHD)

- March 1946, 16,695 babies in England, Scotland and Wales.
  - Four-page questionnaire
    - Birth weights, father's occupation, the number of rooms and occupants (including domestics) in the baby's home and whether the baby was legitimate or illegitimate.
  - Throughout their school years and young adulthood and on into middle age, researchers weighed, measured, prodded, scanned and quizzed the group's bodies and minds in almost every way imaginable.
  - Participants turn 65 this year.

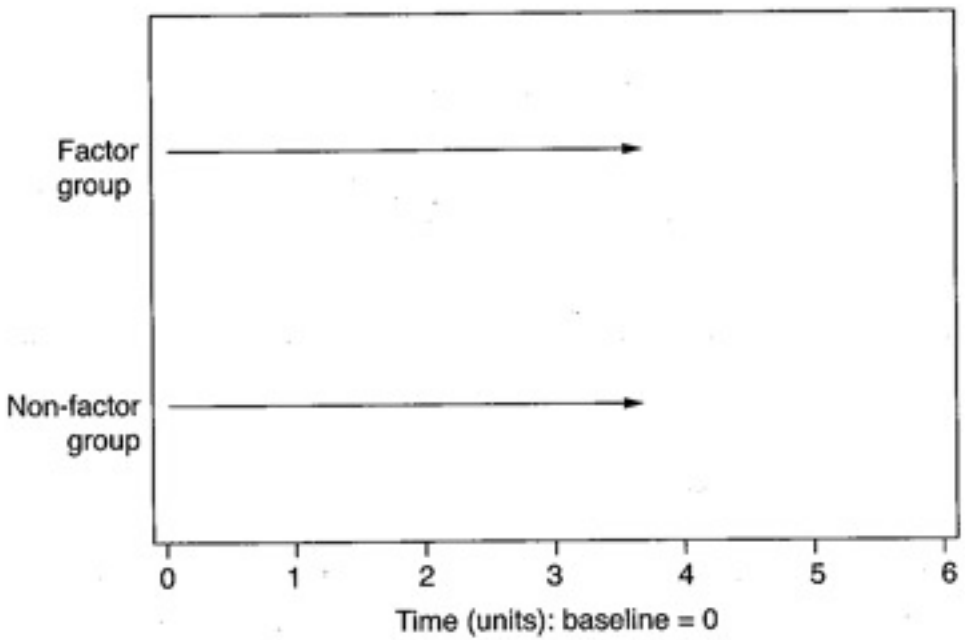
A Generation Under Study

- It's the only study to have chased an entire cohort across its life course
- 8 books, 600 papers
- One of the most important findings: early life matters a lot
- Children who were born into better socioeconomic circumstances were most likely to do well in school and university, escape heart disease, stay slim, fit and mentally sharp and, so far at least, to survive.
- Conclusions important for policy making

- Bright children from the middle classes were more likely to pass the 11+ and do well at school than were equally bright working-class children, although supportive parents and good teachers could better a child's odds.
  - *The Home and the School* (1964) and *All Our Future*(1968) by J.W.B. Douglas

Introduction

- A cohort, or prospective, study is one in which individuals are followed over time to monitor their health outcomes.
- Select two groups of people at the start of the study, the baseline.
  - One group consists of people who possess some special attribute thought to be a possible risk factor for a disease of interest, and the other group does not.
  - Both groups are followed over time and the incidence of disease compared between the groups.
- Study of the hazards of working in the coal industry
  - A group of coal miners and a second group of employees in other heavy industries might be selected.
  - Both groups are monitored for 10 years, after which time the incidence of bronchitis is compared between the groups.



- Advantages:
  - Cohort studies give direct information on the sequence of happenings.
    - Ideal for demonstrating causality
  - Many diseases can be studied simultaneously.
    - Need to record episodes of all the required diseases during the follow-up.

Disadvantages

- Very expensive and time-consuming.
- Not suitable for diseases with a long latency.
  - Smoking and lung cancer
- Not suitable for rare diseases.
  - A large baseline sample or to monitor for a very long time.
- Study effect
  - Someone may act differently simply because of being studied.
- Exposure to the factor of interest may change.
- Withdrawals may occur.
  - Related to the disease and unrelated to the disease.

Studies with a Single Baseline Sample

- Take a single baseline sample and identify the factor and nonfactor groups from the sample data.
  - Advantage: information on the variable used to stratify the risk factor groups is not required beforehand.
  - Disadvantage: distribution of the risk factor cannot be controlled.

Analytical Considerations – Concurrent follow-up

- Fixed cohort studies
  - Everyone starts at the same time and followed up for the same length of time.
  - Risks can be estimated relatively easily.
  - Simple analysis of risks have the disadvantage that they cannot differentiate between short- and long-term effects.
- Variable cohort studies
  - Set of individuals at risk changes during the study for reasons other than loss due to the event of interest.
- A survival analysis
  - Simultaneous analysis of progress for different durations of follow-up
  - The time of events are analyzed rather than the mere fact of the events.
  - Ideal for variable cohort studies.
  - 'survival' here means failure to become diseased, not necessarily relate to lack of death.

Analytical Considerations – Moving Baseline Dates

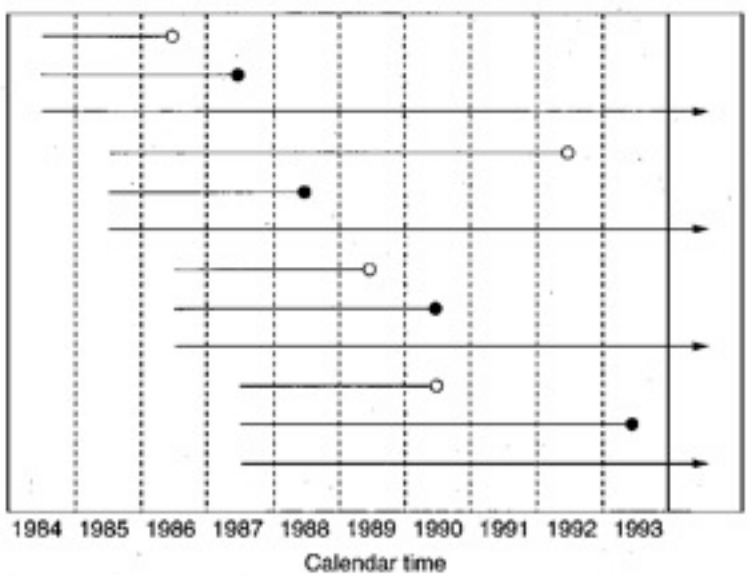
- Recruitment into a study is not simultaneous, but happens over a period of time.
- The issue of different starting time
  - Ignore it
    - Assumes that any effects are homogeneous with respect to calendar time.
    - It is a reasonable assumption provided that the baseline dates do not vary greatly.

Analytical Considerations – Varying Follow-up Durations

- The calendar time at which evaluation of effects is made will generally be the same for each member of the cohort.
  - Length of follow-up may vary
  - In SHHS, the time is from 6.2 to 9.1 years with an average of 7.7 years.
- Treating the study as a fixed cohort
  - Perform a simple risk analysis for all events up to, but not exceeding the minimum elapsed time.
    - Analyze only the first 6 years for all subjects.
  - Ignoring the variation in follow-up durations.
    - Someone will have more chance of an event.
    - Assumes the length of follow-up is not related to the risk factor being studied.

Analytical Considerations – Withdrawals

- Withdrawals are people who are lost to follow-up before experiencing an event.
- Two choices
  - Ignore the withdrawals – overestimate the risk
  - Include them amongst those negative for the event – underestimate the risk
- Decision should be made on whether the withdrawals are related to the event.
- Censored data
  - Anyone who has not yet experienced an event but has a shorter follow-up time than the maximum possible is said to be censored.



Cohort Life Tables

- Cohort life table (life table)
  - A tabular presentation of the progress of a cohort through time.
- To construct a life table
  - Divide the entire follow-up period into consecutive intervals of time.
  - Calculate the following quantities
    - n_t: the number of survivors at time t
    - e_t: the number of events in the study interval that begins at time t.
    - p_t: the estimated probability of surviving the entire study interval that begins at t.
    - q_t: the estimated probability of an event (or failure) during the study interval that begins at t.
    - s_t: the estimated probability of surviving from baseline to the end of the study interval that begins at t.
  - Taking baseline to be time 0, then n_0 will be the total number of subjects in the study.

Cohort Life Tables

- All the *n* and *e* results come from observation
- $q_t = e_t / n_t$
- $p_t = 1 - q_t$
- $s_t = p_0 p_1 p_2 \ldots p_t$.

A cohort of 1000 men at high risk of disease but currently disease free, is recruited. In the first year of study, 5 of the men are newly diagnosed with the disease. In the second year, a further 10; in the third year, 20; in the fourth year, 35; and in the fifth year 50 new cases are identified.

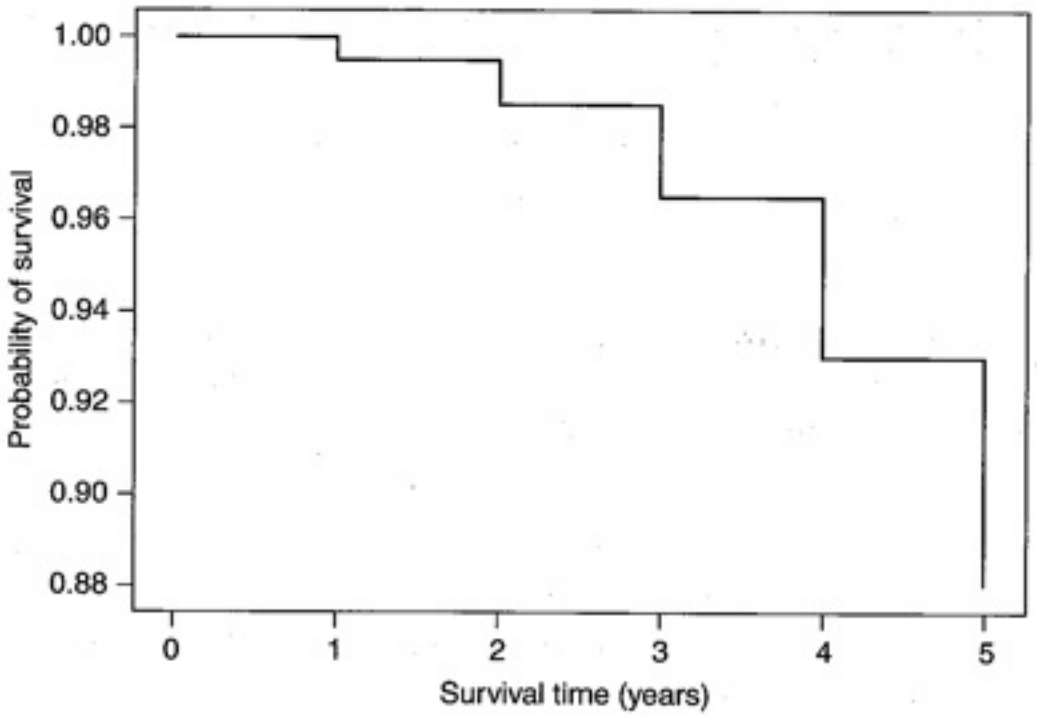
Cohort Life Tables

Table 5.1. Life table for a hypothetical cohort.

Time (t)	No. free of disease (n)	No. of events (e)	Interval risk (q)	Interval survival (p)	Cumulative survival (s)
0	1000	5	0.0050	0.9950	0.9950
1	995	10	0.0100	0.9899	0.9850
2	985	20	0.0203	0.9797	0.9650
3	965	35	0.0363	0.9637	0.9300
4	930	50	0.0538	0.9462	0.8800
5	880				

- The estimated probability of survival for 5 years is 0.88.
  - 880 survivors at time 5 compared 1000 at time 0.

Survival Plot



- A step function
- The chance of survival to a point intermediate to those enumerated in the life table will be estimated to be equal to that chance evaluated at the previous life table time cut-point

Cohort Life Tables - Allowing for Sampling Variation

- We can estimate the standard error of the interval-specific risks by

$$\hat{\text{se}}(p_t) = \sqrt{p_t(1-p_t)/n_t}$$

- Similarly, $\hat{\text{se}}(q_t) = \sqrt{q_t(1-q_t)/n_t}$

$$\hat{\text{se}}(s_t) = s_t \sqrt{\sum_{i=0}^{t} \frac{q_i}{n_i - e_i}}$$

- Standard error of the cumulative survival probability

$$s_t \pm 1.96\, \hat{\text{se}}(s_t)$$

- 95% CI

- Test the null hypothesis that the cumulative survival probability is s, compare

$$\left\{ \frac{s_t - s}{se(s_t)} \right\}$$

Cohort Life Tables – Allowing for Censoring

- Consider all lost subjects as censored.

Table 2. Number of men experiencing a CHD event or being censored by period of observation

Period of follow-up (years)	Censored	Events	Total
0 but less than 1	7	17	24
1 but less than 2	12	22	34
2 but less than 3	24	26	50
3 but less than 4	19	23	42
4 but less than 5	21	37	58
5 but less than 6	15	38	53
6 but less than 7	501	31	532
7 but less than 8	2143	20	2163
8 but less than 9	1375	5	1380
9 but less than 10	66	0	66
Total	4183	219	4402

If the 7 are lost at the beginning estimated failure probability is
17/(4402-7) = 0.003868
If the 7 are lost at the end of first year
17/4402 = 0.003862
A reasonable approximation is
17/(4402-3.5) = 0.003865

$$q_t = e_t / n_t^*, \qquad n_t^* = n_t - c_t / 2$$

This is called actuarial method for analyzing survival data.

Standard error of the cumulative survival probability is

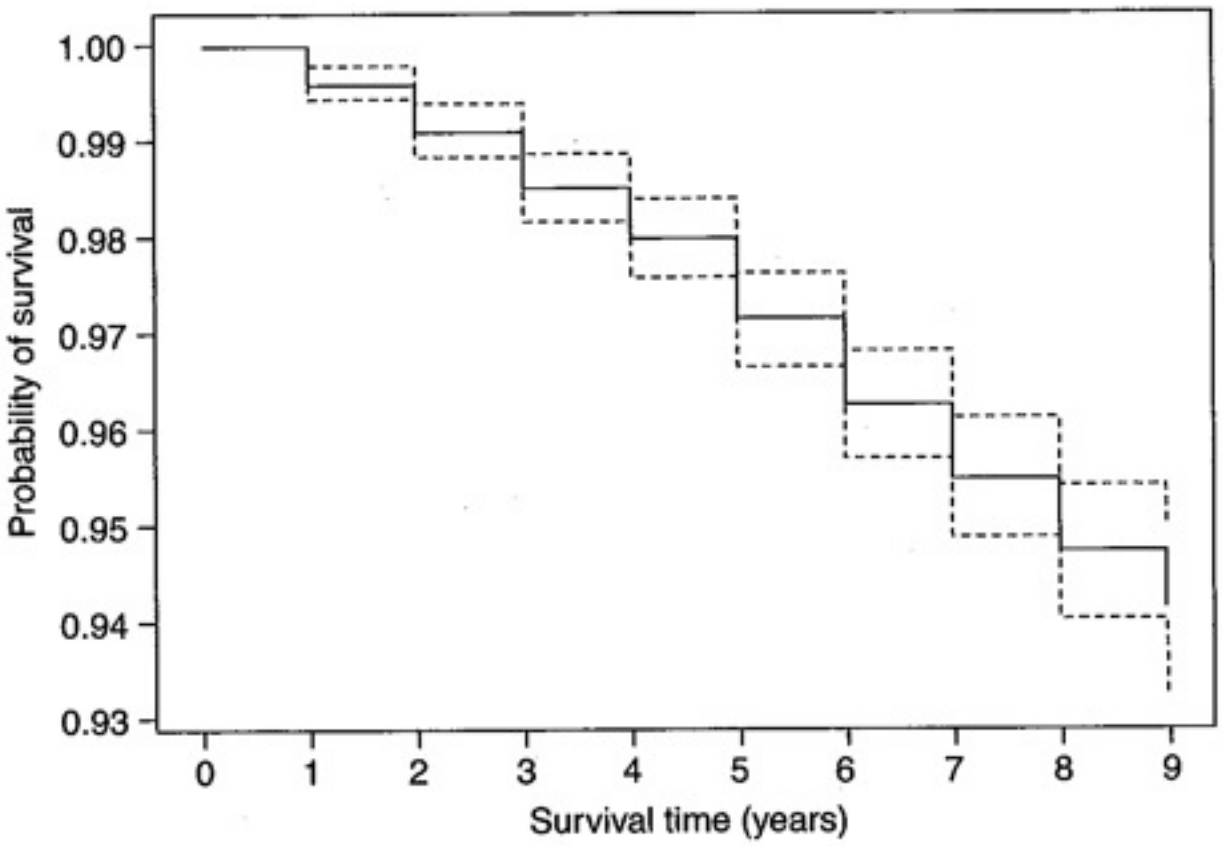
$$\hat{\text{se}}\left(s_t\right) = s_t \sqrt{\sum_{i=0}^{t} \frac{q_t}{n_t^* - e_t}}$$

Cohort Life Tables – Allowing for Censoring

- Life table for coronary events for selected SHHS men.

Time (years)	Number	Censored	Adjusted number	Events	Interval risk	Interval survival	Cumulative survival probability	
							Estimate	Standard error
0	4402	7	4398.5	17	0.003865	0.996135	0.996135	0.0009356
1	4378	12	4372.0	22	0.005032	0.994968	0.991122	0.0014152
2	4344	24	4332.0	26	0.006002	0.993998	0.985174	0.0018253
3	4294	19	4284.5	23	0.005368	0.994632	0.979885	0.0021226
4	4252	21	4241.5	37	0.008723	0.991277	0.971337	0.0025268
5	4194	15	4186.5	38	0.009077	0.990923	0.962521	0.0028804
6	4141	501	3890.5	31	0.007968	0.992032	0.954851	0.0031697
7	3609	2143	2537.5	20	0.007882	0.992118	0.947325	0.0035636
8	1446	1375	758.5	5	0.006592	0.993408	0.941081	0.0045033
9	66	66		0				

Cohort Life Tables – Allowing for Censoring



Actuarial estimates and 95% confidence intervals for cumulative survival probabilities of coronary events.

Cohort Life Tables – Comparison of Two Life Tables

- Life tables can be constructed for each subgroup of the cohort and the survival experiences compared by graphical, or more formal methods.
- The standard error of the difference between two cumulative survival probabilities is estimated as

$$\hat{\text{se}}\left(s_t^{(1)} - s_t^{(2)}\right) = \sqrt{\hat{\text{se}}\left(s_t^{(1)}\right)^2 + \hat{\text{se}}\left(s_t^{(2)}\right)^2}$$

- An approximate 95% CI for the true difference is

$$s_t^{(1)} - s_t^{(2)} \pm 1.96\hat{\text{se}}\left(s_t^{(1)} - s_t^{(2)}\right)$$

- An approximate test of the null hypothesis that the two survival probabilities are equal is given by comparing

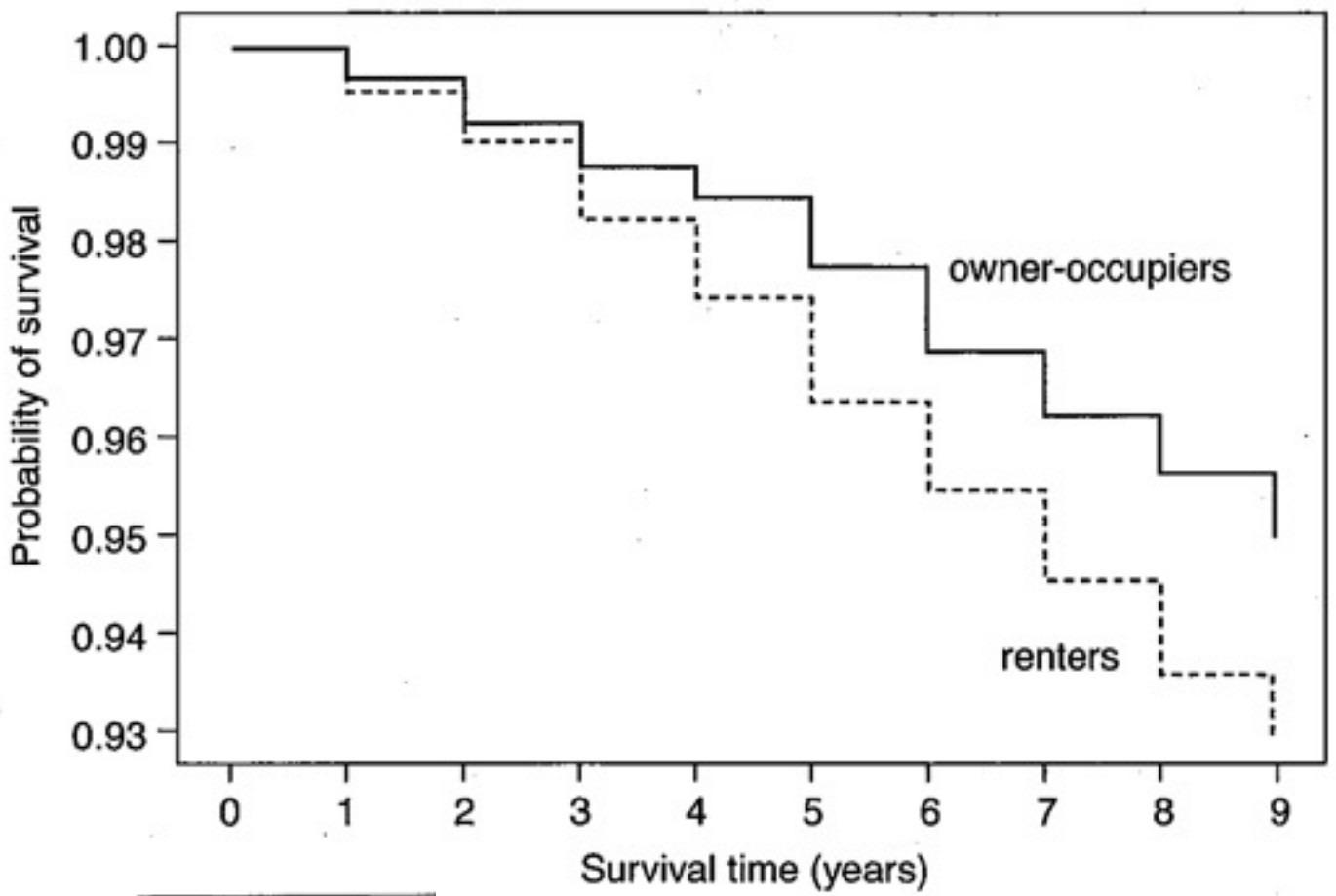
$$\left\{\frac{s_t^{(1)} - s_t^{(2)}}{\hat{\text{se}}\left(s_t^{(1)} - s_t^{(2)}\right)}\right\}^2$$

to chi-square with 1 d.f.

Cohort Life Tables – Comparison of Two Life Tables

- Ex. The SHHS data of were disaggregated by housing tenure status and separate life tables were constructed for owner-occupiers and renters.

	Owner–occupiers					Renters				
				Cumulative survival probability					Cumulative survival probability	
Time (years)	Number	Censored	Events	Estimate	Standard error	Number	Censored	Events	Estimate	Standard error
0	2482	2	8	0.996776	0.0011382	1920	5	9	0.995306	0.0015609
1	2472	5	12	0.991932	0.0017968	1906	7	10	0.990075	0.0022657
2	2455	10	11	0.987478	0.0022349	1889	14	15	0.982184	0.0030282
3	2434	9	8	0.984227	0.0025058	1860	10	15	0.974241	0.0036323
4	2417	12	17	0.977287	0.0030006	1835	9	20	0.963597	0.0043024
5	2388	4	21	0.968686	0.0035126	1806	11	17	0.954499	0.0047943
6	2363	247	15	0.962197	0.0038679	1778	254	16	0.945249	0.0052762
7	2101	1286	9	0.956258	0.0043212	1508	857	11	0.935617	0.0059684
8	806	755	3	0.949563	0.0057661	640	620	2	0.929946	0.0071534
9	48	48	0			18	18	0		



$$\hat{\text{se}}\left(s_t^{(1)} - s_t^{(2)}\right) = \sqrt{\hat{\text{se}}\left(s_t^{(1)}\right)^2 + \hat{\text{se}}\left(s_t^{(2)}\right)^2}$$

$$\left\{\frac{s_t^{(1)} - s_t^{(2)}}{\hat{\text{se}}\left(s_t^{(1)} - s_t^{(2)}\right)}\right\}^2$$

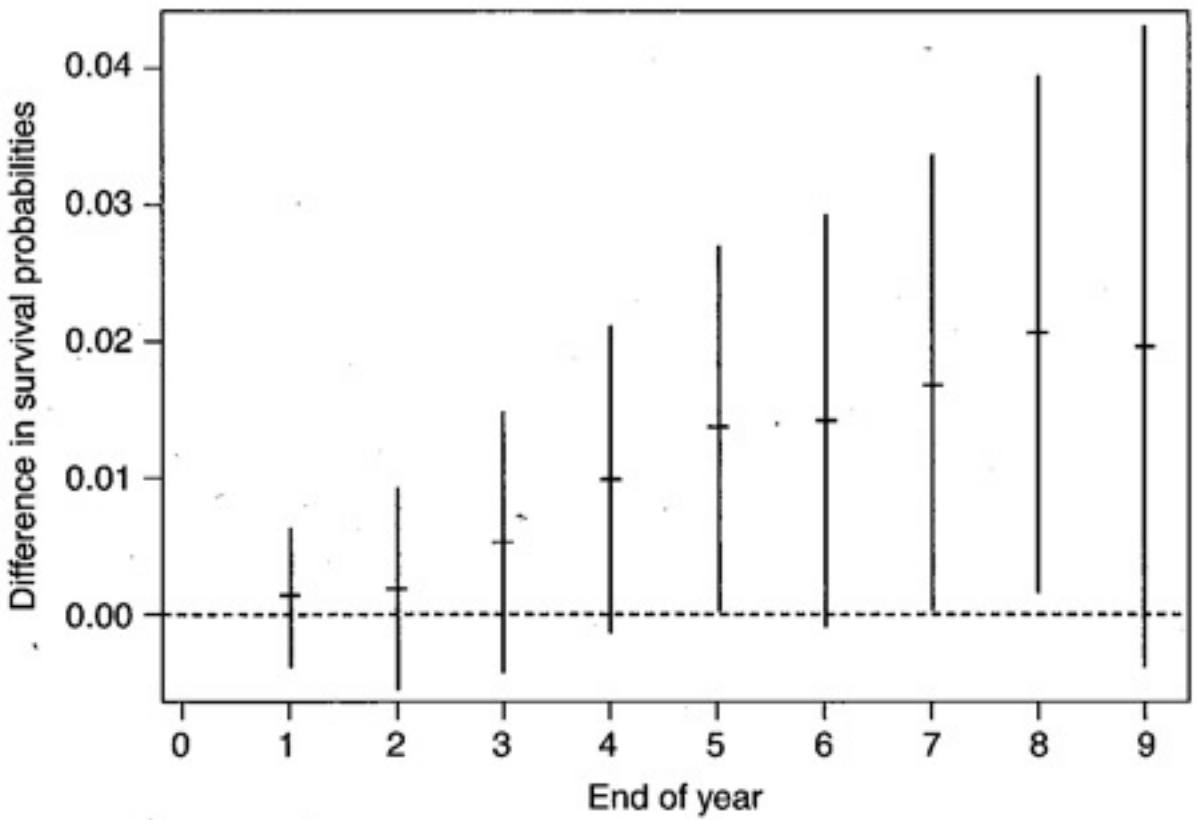
$$\frac{(0.949563 - 0.929946)^2}{0.0057661^2 + 0.0071534^2} = 4.56$$

$0.05 > p > 0.025$

If 5-year survival is of interest, an approximate 95% CI for the difference in the probability of survival for 5 years for owner-occupiers compared with renters is

$$0.977287 - 0.963597 \pm 1.96\sqrt{0.0030006^2 + 0.0043024^2}$$

The figure below shows the difference in estimated survival probabilities up to each whole number of years of survival, where 99% confidence intervals are shown.



Cohort Life Tables – Limitations

- Life table approach produces a step function, leading to overestimation of survival probabilities at points intermediate to the cut-points that define the life table intervals.
  - It is better to choose small intervals whenever several specific probabilities are likely to be of interest.
- The actuarial method assumes that the censoring occurs uniformly within the interval.
- The approximation may not be valid.
  - The average time to censoring is not ½, but a_t, which is the average proportion of the interval that is survived before censoring occurs.

$$n_t^* = n_t - (1 - a_t)c_t$$

- The risk at different period of a life span may be different.
  - Again, small intervals are recommended.

Kaplan-Meier Estimation

- Kaplan-Meier (KM) or product-limit approach addresses the limitations of the standard life table.
- In KM, the observed event times for the cohort studied define the values of t at which s_t is evaluated.
  - Leads to a life table with smallest possible intervals.
  - Requires more computation
- The choice of approximation used to deal with censoring is unlikely to be important when the intervals in the equivalent life table are small.
- The tabular display is less useful.

Kaplan-Meier Estimation cont.

Ex. During the first year of follow-up in the SHHS, the completed survival times (in days) for the male subset are

1, 46, 91*, 101, 101, 103, 119, 133*, 137, 145*, 156, 186*,

208, 215, 235, 242, 251, 294, 299, 300, 309*, 312, 336*, 357*.

* denote a censored observation.

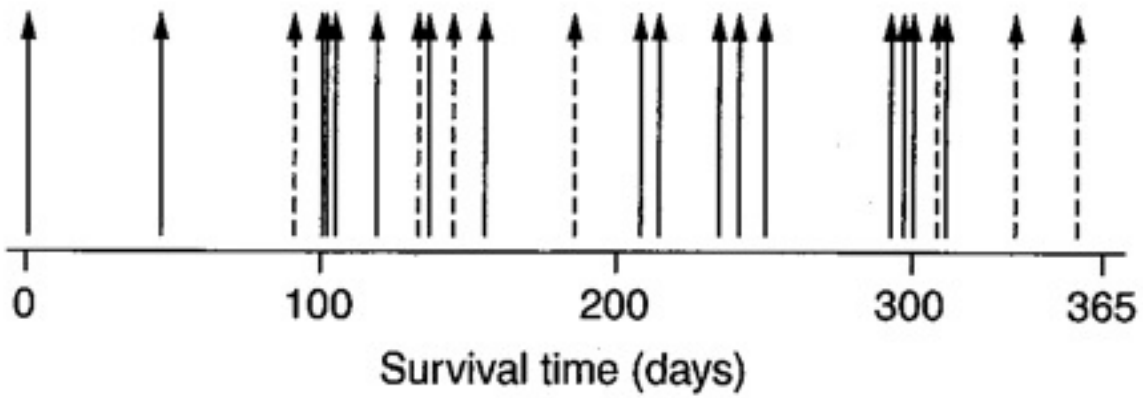


Table 5. Data and Kaplan-Meier estimation of the survival function year 1 of follow-up for the selected subset of SHHS men.

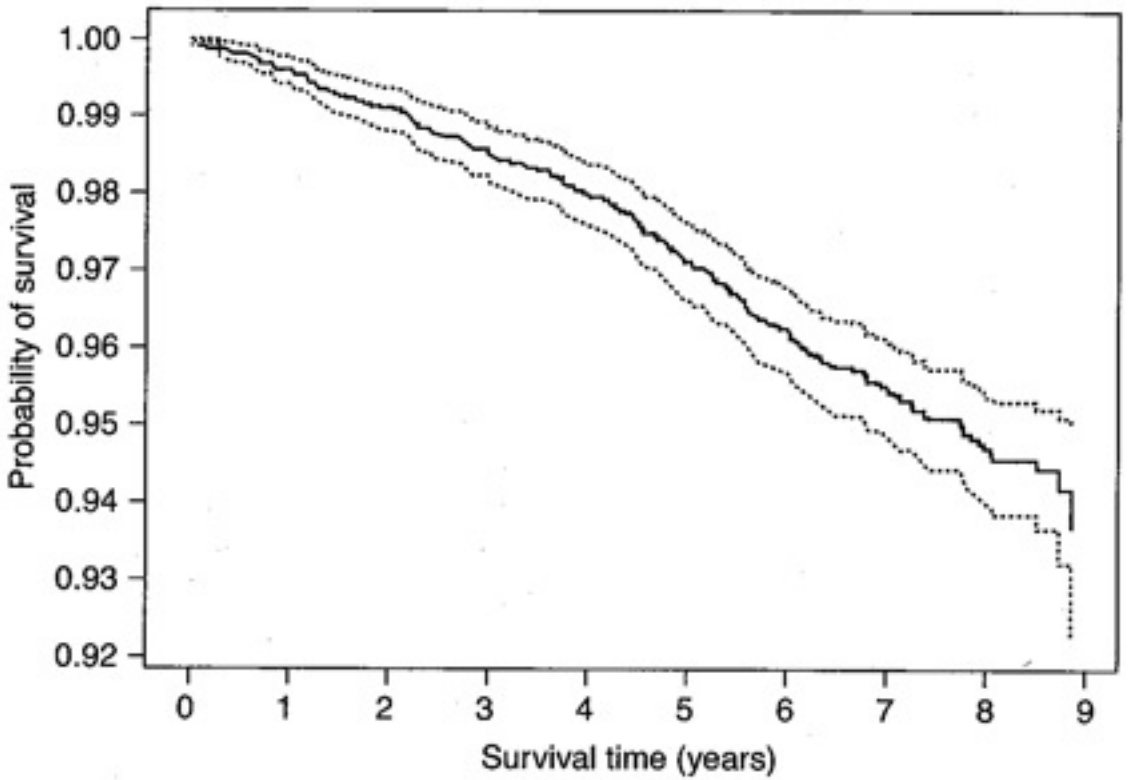
Time (days)	Number	Events	Survival probability	
			Estimate	Standard error
0	4402		1	
1	4402	1	0.999773	0.0002271
46	4401	1	0.999546	0.0003212
101	4399	2	0.999091	0.0004542
103	4397	1	0.998864	0.0005077
119	4396	1	0.998637	0.0005561
137	4394	1	0.998410	0.0006007
156	4392	1	0.998182	0.0006421
208	4390	1	0.997954	0.0006810
215	4389	1	0.997727	0.0007178
235	4388	1	0.997500	0.0007528
242	4387	1	0.997273	0.0007862
251	4386	1	0.997045	0.0008183
294	4385	1	0.996818	0.0008491
299	4384	1	0.996591	0.0008788
300	4383	1	0.996363	0.0009075
312	4381	1	0.996136	0.0009354
365	4378			

Kaplan-Meier Estimation – An Empirical Comparison

- Difference due to the distinct methods for dealing with withdrawals will exist.

Table 6. Comparison of actuarial and Kaplan-Meier (KM) results for the survivor function.

Time (years)	Estimate		Standard error (×10 000)	
	Actuarial	KM	Actuarial	KM
1	0.996135	0.996136	9.356	9.354
2	0.991122	0.991125	14.152	14.148
3	0.985174	0.985177	18.253	18.249
4	0.979885	0.979885	21.226	21.226
5	0.971337	0.971339	25.268	25.267
6	0.962521	0.962525	28.804	28.800
7	0.954851	0.954982	31.697	31.607
8	0.947325	0.947144	35.636	36.308
9	0.941081	0.936608	45.033	70.003



Kaplan-Meier estimates and 95% confidence intervals for cumulative survival probabilities for coronary events.

Comparison of Two Sets of Survival Probabilities – Mantel-Haenszel Methods

	Owner–occupiers		Renters	
Interval (years)	Number[a]	Events	Number[a]	Events
0 but less than 1	2482	8	1920	9
1 but less than 2	2472	12	1906	10
2 but less than 3	2455	11	1889	15
3 but less than 4	2434	8	1860	15
4 but less than 5	2417	17	1835	20
5 but less than 6	2388	21	1806	17
6 but less than 7	2363	15	1778	16
7 but less than 8	2101	9	1508	11
8 but less than 9	806	3	640	2

- Consider the first interval in the standard life table.

Survival experience by housing tenure status during the first year of follow-up

Housing tenure	Event	No event	Total
Renters	9	1911	1920
Owner–occupiers	8	2474	2482
Total	17	4385	4402

Comparison of Two Sets of Survival Probabilities – Mantel-Haenszel Methods

- We can construct similar tables for each interval and seek a summary measure of the chance of event across all the intervals.

9	1911	1920	10	1896	1906	15	1874	1889
8	2474	2482	12	2460	2472	11	2444	2455
17	4385	4402	22	4356	4378	26	4318	4344
$E(a_1) = 7.415$			$E(a_2) = 9.578$			$E(a_3) = 11.306$		
$V(a_1) = 4.166$			$V(a_2) = 5.382$			$V(a_3) = 6.353$		
15	1845	1860	20	1815	1835	17	1789	1806
8	2426	2434	17	2400	2417	21	2367	2388
23	4271	4294	37	4215	4252	38	4156	4194
$E(a_4) = 9.963$			$E(a_5) = 15.968$			$E(a_6) = 16.364$		
$V(a_4) = 5.618$			$V(a_5) = 9.000$			$V(a_6) = 9.235$		
16	1762	1778	11	1497	1508	2	638	640
15	2348	2363	9	2092	2101	3	803	806
31	4110	4141	20	3589	3609	5	1441	1446
$E(a_7) = 13.310$			$E(a_8) = 8.357$			$E(a_9) = 2.213$		
$V(a_7) = 7.540$			$V(a_8) = 4.839$			$V(a_9) = 1.230$		

- The continuity-corrected Cochran-Mantel-Haenszel test statistic for the null hypothesis of no overall difference in survival experience between the housing tenure groups is

$$\frac{\left(\left|\sum a_i - \sum \mathrm{E}(a_i)\right| - \frac{1}{2}\right)^2}{\sum \mathrm{V}(a_i)} = \frac{\left(\left|115 - 94.474\right| - 0.5\right)^2}{53.363} = 7.52$$

- p-value = 0.006 for Chi-square with 1 df.
- There is evidence of a difference in the overall chance of survival.