Cohort Data Analysis (171:242/243)
Section 1: Role of Cohort Studies

Brian J. Smith, Ph.D.

April 4, 2005

Table of Contents

1.1 Study Designs

1.1.1 Cross-Sectional Study

At one point in time data are collected on a sample of the population. Exposure and disease prevalence information are obtained and correlations computed. Such "population correlation" or "ecological" studies are useful in generating interesting hypotheses but are not normally useful in assessing basic causality in an exposure-disease relationship.

1.1.2 Case-Control Study

A sample of individuals with the disease (cases) and a sample of those without (controls) make up the study group. Then their past exposure experience is obtained retrospectively.

1.1.3 Cohort Study

First identify a study group or "cohort" of people about whom you will collect exposure information. Follow them forward in time and note disease occurrence for each individual.

Historical Cohort

By historical records, identify a group with certain exposure characteristics, at some specific point of time in the past, and then follow them forward towards the present, recording their disease experience.

Example: Want to study effects of exposure to levels of a carcinogen which is no longer found in manufacturing and

for which historical data exist and in a group which is such a small fraction of the general population that a case-control study would miss them.

Advantage: Results may be obtained in a short amount of time.

Prospective Cohort

Assemble cohort in the present and follow them prospectively into the future.

Advantage: Collect exactly that information which is needed. The records for a historical cohort study may have been collected for very different reasons and some information may be spotty.

1.2 Historical Role of Cohort Studies

Two landmark papers:

1. Prospective cohort study of British doctors by Doll and Smith (1954), a “preliminary report” on tobacco smoking and lung cancer.
2. Historical cohort study of Case et al. (1954) and Case and Pearson (1954) on bladder cancer in the British chemical industry.

1.2.1 British Doctors Study

Around 1950 results of several case-control studies had been published, including Doll and Hill (1950), demonstrating an association between lung cancer and cigarette smoking. In their 1954 paper Doll and Hill made the case for further prospective studies of the exposure-disease relationship, stating that,

‘In the last five years a number of studies have been made of the smoking habits of patients with and without lung cancer. All these studies agree in showing that there are more heavy smokers and fewer nonsmokers among patients with lung cancer than among patients with other diseases. While, therefore, the various authors have all shown that there is an “association” between lung cancer and the amount of tobacco smoked, they have differed in their interpretation. Some have considered that the only reasonable explanation is that smoking is a factor in the production of the disease; others have not been prepared to deduce causation and have left the association unexplained.’

Thus, a prospective cohort study was begun in 1951 to study lung cancer occurrence in a population whose smoking habits were already known.

	Case-Control	Cohort
Study Start	April 1948	October 1951
Lung Cancers	1,488	411 men 27 women
Total Enrollment	4,342	34,440 mean 6,194 women
Final Results	December 1952	1978 (men) 1980 (women)
References	Doll and Hill (1950, 1952)	Doll and Peto (1976, 1978, 1980)

Comments

- The case-control design was cheaper, quicker, and able to enroll more cases.

- The cohort design acquired more detailed information on health effects of smoking.

1.2.2 Bladder Cancer Study in British Chemical Industry

The purpose was to determine whether the manufacture or use of aniline, benzidine, β-naphthylamine or α-naphthylamine could be shown to produce tumors of the urinary bladder in exposed males. The cohort design was chosen because:

- Only a small percentage of all bladder cancers are due to the chemical industry. A general case-control design would be uninformative.
- Answer needed urgently, current exposure levels were less than past exposure levels. A prospective cohort study wouldn't work.
- Historical cohort study was the only possible approach.

1.3 Strengths and Limitations

1.3.1 Strengths

This section gives the strengths of the cohort study, relative to the case-control design.

1. Cohort study is better at establishing full range of health effects related to a particular exposure. After all, cohort study starts with exposed and unexposed subjects, follows them through time and records all disease experiences. Case-control starts with a

particular disease and a backward look at exposure history.

2. Biases
    a. Recall Bias: The results of a case-control study are questionable if there is a possibility of recall bias. Recall bias should not occur in a properly carried out cohort study.
    b. Precision of Recall: Suppose we have an ordinal exposure variable and there is some (unbiased) random error in the recalled level of exposure. Suppose also that the variability of this error differs between cases and controls (in a case-control study). Then the apparent odds ratio can be quite different from unity even when it shouldn't be. Again, recall bias should be minimized in a cohort study.
    c. Selection Bias:
    In case-control studies, this is possible if a high proportion of those contacted to be population-based controls refuse. If hospital controls are used, which disease categories are eligible?
    In cohort studies, the healthy-worker effect may introduce bias if the employed population is healthier (has lower morbidity rates) than the unemployed population. Also, the chances of a highly-sensitive individual quitting work in a risky industry are probably higher than an insensitive individual.

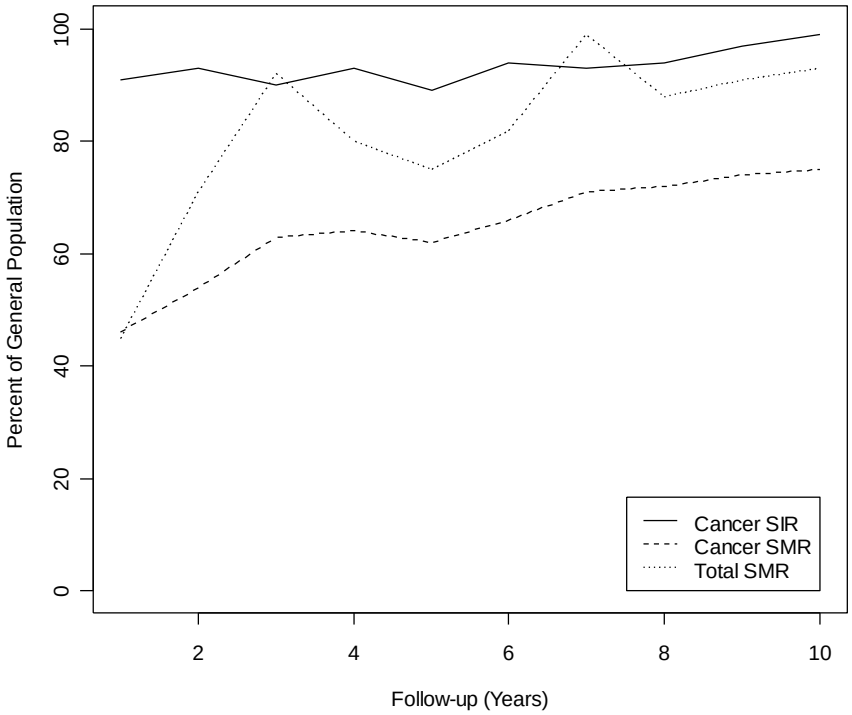


Figure 1. Evolution of the healthy worker effect following entry into a study of Swedish building workers.

3. Efficiency – Cohort studies are more efficient than case-control when the exposure is both rare in the general population and responsible for only a small proportion of the cases. This latter case rules out efficiency of the case-control study; the first case adds to this.
4. Pre-disease exposure information may be impossible to determine retrospectively. One may need blood samples to determine exposure. These are rarely available for case-control studies, but can be handled routinely in prospective cohort studies.
5. Retrospective information may be too inaccurate to be useful; e.g. dietary recall, chemical exposure recall.

6. Cohort studies allow serial measurements of exposure. This will allow not only presence/absence of exposure but time-dependent levels of exposure. This increased accuracy of the exposure over time should improve our inference concerning the exposure-disease relationship.
7. Case-control studies are good for estimating odds ratios. If one also wants to know the actual incidence (morbidity) rates or the absolute risk measurements, a cohort study is necessary.

1.3.2 Limitations

1. Prospective cohort studies require a great commitment over a long period of time. Few people or funding agencies have such patience for any but the most important health issues. Expensive!!! Variations on cohort design are cheaper; e.g. nested case-control and case-cohort.
2. Historical cohort studies can only be done when the cohort of interest exists and complete, accurate information on exposure as well as important confounding variables is available.
3. If the disease is sufficiently rare, even a very large cohort may not develop a sufficient number of cases to make the cohort approach worth while. In this case consideration of the effect or cost per case will favor the case-control over the cohort approach.
4. Cohorts not representative of the general population cannot give estimates (even extrapolated ones) of the population attributable risk. Population-based case-control studies can estimate this.

1.3.3 Summary

Cohort

- Provides well-defined population from which cases arise in an unbiased fashion.
- Complete covariate and exposure experience (duration times, levels, etc.) are available for entire study period.

Case-Control

- Concentrate effort on informative individuals (cases and controls) for whom extensive information is collected.
- Inexpensive

New procedures that combine advantages of these two designs are being developed and implemented.

1.4 Implementation

The two main issues to consider when planning a cohort study are:

1. Is the planned cohort size adequate for detection of real differences?
2. How to implement the study?

Implementation includes consideration of:

1. Inclusion/exclusion criteria – rules for including and excluding individuals should be clear.

2. Dates for subjects
    - Date of enrollment
    - Date of first exposure (often different from date of entry)
    - Date last seen and vital status
3. Follow-up mechanism – the percentage of individuals lost to follow-up is a measure of the quality of the study. The study will be called into question if that percentage is high. The purpose of follow-up is:
    a. Determination of person-years information; who is still under observation and who is lost to follow-up?
        - The follow-up mechanism may vary from country to country.
        - Group-based cohort: labor union, insurance plan, pension plan, professional society, etc.
    b. Identification of cases
        - Death certificates
        - Cancer registry – more accurate, more cases, more information

Table 1. Number of deaths occurring from five through 35 years after onset of work in an amosite asbestos factory, 1941-1945. Cause of death coded in two different ways[a,b].

Underlying cause of death	DC	BE	BE-DC	Expected
All causes	1946	1946	-	1148.0
Cancer, all sites	845	912	+67	259.0
Lung	397	450	+53	81.7
Pleural mesothelioma	23	61	+38	-
Peritoneal mesothelioma	24	109	+85	-
Other mesothelioma	54	0	-54	-
Larynx, buccal and pharynx	21	27	+6	7.5
Esophagus	17	17	0	5.1
Kidney	15	16	+1	8.5
Colon-rectum	54	55	+1	8.5
Stomach	18	21	+3	30.5
Prostate	24	26	+2	12.5
Bladder	7	9	+2	6.7
Pancreas	46	21	-25	16.0
Other	110	83	-27	72.1
Site Unknown	35	17	-18	
Noninfectious pulmonary disease	177	204	+27	68.2
Cardiovascular disease	638	566	-72	660.1
Other and unspecified causes	286	264	-22	160.8

[a] DC, death certificate; BE, best evidence available
[b] From Hammond et al. (1979)

c. Confirmation of case information
   - Use of additional information to “refine” death certificate; e.g. X-rays and asbestos-related disease

d. Coding of disease
   - World Health Organization (WHO) members code death certificates according to current International Classification of Disease (ICD). Disease codes can change from one revision

to another. Be aware of different codings in a cohort spanning different revision periods.

e. Assessment of disease

- Coding an exposure variable as yes/no is insufficient for a dose-response relationship: cannot infer causality or set safety standards. Should quantify level of exposure as much as possible and when exposure occurred, for how long and when it stopped. Such exposure information is needed on an individual level. Mean values for an entire cohort, though not valueless, cannot give dose-response estimates.
- Starting and stopping dates of exposure are often easily obtained.
- Exact level of exposure may be difficult, especially in historical cohorts. One may have to use a categorical measurement of exposure; e.g. low, medium, high. Demonstrating a dose-response relationship on such an ordinal exposure variable is possible.

f. Information on possible confounding factors

- Spurious results arise when confounding factors are not adjusted for in the analysis. We use the term “misclassification” to denote that incorrect information has been collected on a variable.
- For dichotomous variables, misclassification rates of 30% for the confounder can result in very little of the confounding effect being

removed. If the misclassification rate is 10%, then in certain situations nearly half the effect of confounding is still in place.

- Collect as accurate information as possible. If this is not possible, it may be better to try a less expensive approach, like a case-control design and spend the extra money and time on gaining more accurate data.

g. Construction of special comparison groups

- Occasionally, one needs to construct a special group apart from the cohort. For example, cohort consists of smoking and nonsmoking asbestos workers. Need two groups: smoking and nonsmoking people unexposed to asbestos. Unexposed and exposed may be matched in such situations.

h. Power considerations

- Unless your data are merged into a larger study, if your study has too low a power to detect realistic levels of excess risk, your study is most likely not worth doing.

i. Other designs

    i. Synthetic case-control: At each failure time consider the failing person as the case and take a random sample of the rest of the cohort at risk to be the controls. Risk set at time *t* consists of the failing subjects plus the time-matched controls. Use Cox regression to analyze.

ii. Case-cohort: At beginning of study, randomly pick a sub-cohort of the complete cohort. The risk set at time *t* is the intersection of those in the sub-cohort still at risk and those that have failed.
Example: Women's Health Study was to study 15,000 women looking for an association between breast cancer and dietary fat. Dietary forms were to be done and blood drawn on a routine basis. Cost of dietary coding and blood analyses would cost millions. Cheaper if done on a sub-cohort of say 20-25% of the full cohort, plus 5% that develop breast cancer. Case-cohort design is a natural choice for this.

1.5 Interpretation

A discussion of Hill's criteria for assessing whether an association is causal can be found in most introductory Epidemiology text books:

1. Strength of association
2. Biologic credibility
3. Consistency with other investigations
4. Time sequence
5. Dose-response relationship

More and more, what is expected is not just qualitative evidence, but quantification of the degree of risk. Two major aspects of excess risk are the dose response relationship and risk as a function of time.

Dose-Response

Dose response can be assessed when exposure is quantified as a nominal categorical or numerical variable.

Risk over Time

Incidence or mortality rate often are functions of time, since exposure (e.g. excess leukemia rates 5 years after radiation) or of duration of exposure (e.g. lung cancer incidence rates rise with 4^{th} power of smoking duration among continuing smokers). Also of great importance is the change in risk after exposure stops:

- Further evidence of causal relationship
- Show effect of intervention

So it is a good idea that the design of a cohort study accounts for subjects that are formerly exposed.

CAUTION: Need to know why someone stopped smoking; e.g. very poor health and physician told them they had to stop.

Example: Women treated by radiation for cancer of the cervix have four times the risk of lung cancer as expected.

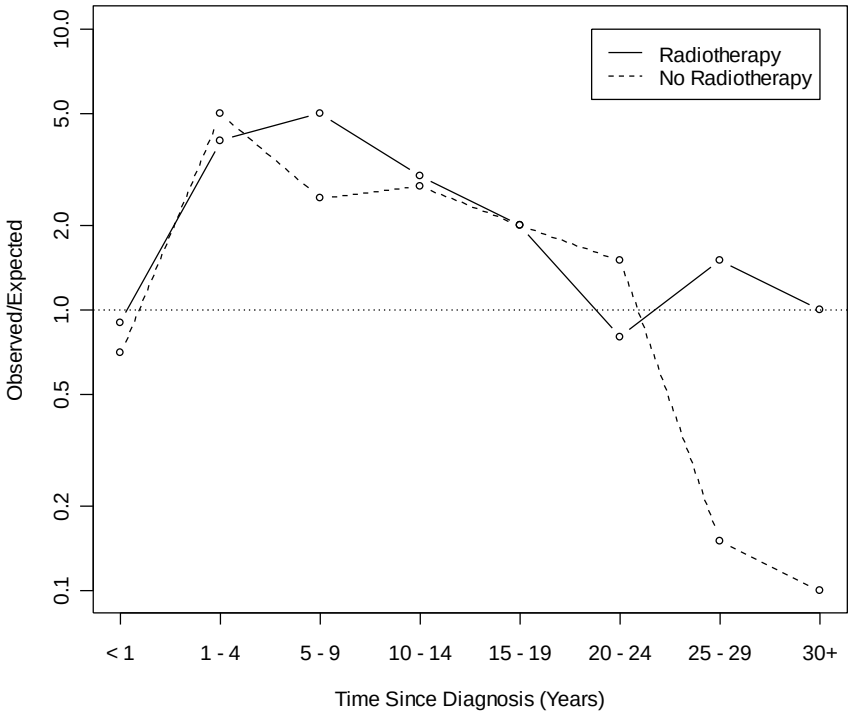


Figure 2. Observed to expected ratios of lung cancer by time since diagnosis of cervical cancer treated with and without radiotherapy.

Upon first inspection it may seem that the excess lung cancer cases are due to the radiotherapy. However, when compared to patients treated without radiotherapy, the same trend is observed. An alternative explanation might be that the excess lung cancers are due to the misclassification of metastases from the original cervical cancer.

1.5.1 Problems with Interpretation

1. Healthy worker effect
    - Can make comparison with external standard population difficulty to interpret. Comparisons between different groups within the cohort should be less affected. Special consideration should also be given to change in employment status (due to ill health?); e.g. retire, change jobs, move to area of lighter work. Mortality is often high a year or two after employment change. One solution is to lag employment status by 2 or 3 years.
    - Analog to healthy worker effect – <u>those who respond to questionnaires</u>. In the British doctors study those who failed to respond had greater mortality rates. In a N.Y. breast cancer screening trial those accepting invitation had half the risk of mortality as those not accepting.
2. Loss to follow-up
    - Incidence rates can be biased downwards if there are people lost to follow-up and we don't know that.
3. Recall bias and misclassification of exposure rates
    - Cohort studies have the advantage of measuring exposure before disease status is ascertained.
4. Lack of information on confounding factors
5. Multiple comparisons
    - Level of test destroyed by number of comparisons. A priori you should have a few hypotheses you want to test. The rest of the many, many things

you can test are not strict statistical tests, but by-products of a hypothesis generating data mining.

6. Identification of forerunners of disease rather than causes
    - An association that looks causal may only reflect an early state of the disease; e.g. cough is the cause of lung cancer or low serum cholesterol levels in people subsequently developing cancer.
7. Conclusions from negative results
    - Can bias or confounding be ruled out?
    - What levels of risk are included within the confidence intervals?
    - How do the levels of exposure in the study compare with the levels in other exposed populations?
    - Had sufficient time elapsed between the start of exposure and the end of follow-up?
    - Is there any reason to suspect that the cohort is at a lower risk than the general population?
    - Are the results consistent with other studies?

1.6 Proportional Mortality Studies

Absolute mortality rates unknown; e.g. don't know annual mortality rate for pancreatic cancer, but do know "proportional mortality rate." For example, 0.1% of all deaths were due to pancreatic cancer. Could also have "proportional incidence rates," possibly from cancer registry.

Study Design: Case-Control, where cases are persons dying from the disease of interest and controls are selected from persons dying of other causes.

Advantage: Quick, cheap look at data. May generate some hypotheses in the initial stage of investigation.

Disadvantage: Excess proportion of one cause of death may mean 1) absolute risk increased for that cause and 2) decrease in rate for some other cause. For example, more hypertensive men survive heart disease and can die of prostate cancer at higher rates. Hypertension is not protective of prostate cancer?!? Serious biases are possible.

Cohort Data Analysis (171:242/243)

Section 2: Rates and Rate Standardization

Brian J. Smith, Ph.D.

April 12, 2005

Table of Contents

2.1 Rates

2.1.1 Crude Rate

Need to estimate disease rate among cohort members during study period; e.g.

$$\text{diseae incidence rate} = \frac{\text{incident cases}}{\text{person-years at risk}}.$$

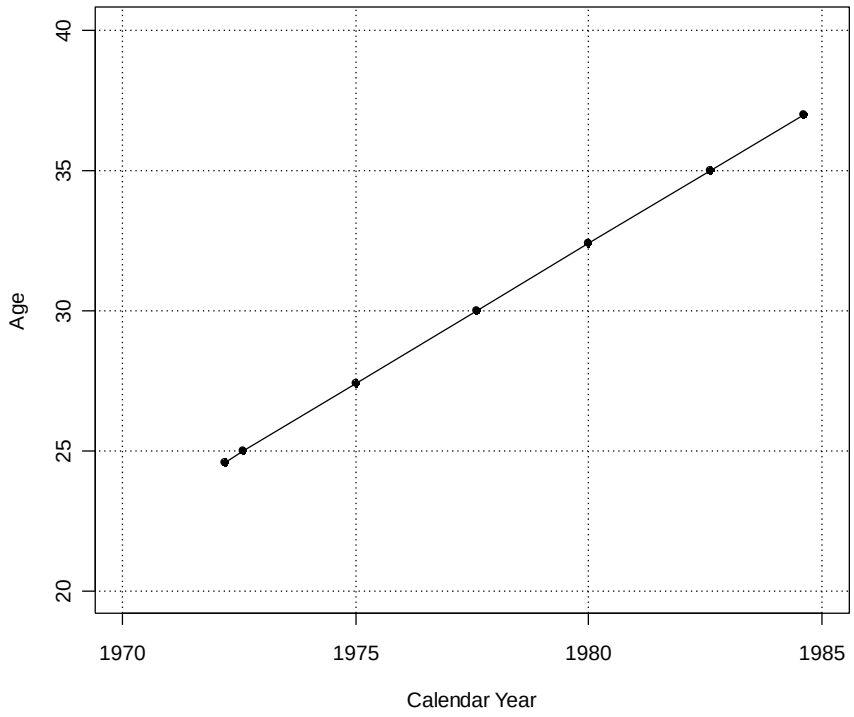
Suppose there are N subjects in the cohort and the l-th subject is at risk for n_l years. Then the number of person years at risk for the entire cohort is $n = \sum_{l=1}^{N} n_l$. If d individuals are diagnosed with the disease during the study period, then the overall or crude incidence rate is

$$\hat{\lambda} = \frac{d}{n} \text{ cases per person-year}.$$

This crude rate ignores any stratification existing within the cohort. It is often of interest to calculate the stratum specific rates. The cohort may be stratified by age intervals and calendar year periods. First, we need to be able to calculate the number of person-years at risk in each stratum.

2.1.2 Calculation of Person-Years

Suppose that subjects are stratified by 5-year age intervals and 5-year calendar periods. Consider a subject who entered the study in 1972.2 at age 24.6 and exited the study in 1984.6 at age 37.0.



The following table demonstrates the calculation of the subject's contribution to the person-years spent in each stratum, where the strata are derived from two factors. Of course there could be more than two factors.

Exact	Approx	Year	Age	Person-Years	
				Exact	Approx
(1972.2, 24.6)	(1972, 24)	-	-	-	-
(1972.6, 25.0)	(1972, 25)	1970-75	20-25	0.4	0.5
(1975.0, 27.4)	(1975, 27)	1970-75	25-30	2.4	2.0
(1977.6, 30.0)	(1977, 30)	1975-80	25-50	2.6	3.0
(1980.0, 32.4)	(1980, 32)	1975-80	30-35	2.4	2.0
(1982.6, 35.0)	(1982, 35)	1980-85	30-35	2.6	3.0
(1984.6, 37.0)	(1984, 37)	1980-85	35-40	2.0	2.5
Totals				12.4	13.0

When using integer dates and ages, assign ½ year to first and last years of age and 1 year to every age in between. Someone entering and exiting the same year gets ¼ year. The exact and approximate methods for computing person-years usually produce similar results.

2.1.3 Stratum-Specific Rates

Suppose there are $j = 1,\ldots,J$ strata, and let d_j and n_j denote the stratum-specific number of incident cases and person-years, respectively. We calculate the number of person-years within each stratum as

$$n_j = \sum_{I=1}^{N} n_{I_j}$$

where N is the total number of subjects in the cohort and n_{I_j} is the amount of time the I-th person spent in stratum j. Then the stratum-specific incidence rate is calculated as

$$\hat{\lambda}_j = \frac{d_j}{n_j}.$$

If d_j represents the number of deaths, then this is interpreted as a mortality rate. $\hat{\lambda}_j$ is an estimate of the true, unknown rate λ_j. Note that the crude rate is

$$\hat{\lambda} = \frac{\sum d_j}{\sum n_j}.$$

Table 1. Respiratory cancer deaths (*d*), person-years at risk (*n*, in thousands), and death rate ($\hat{\lambda}$, per 1000 person-years) in a cohort study of Montana smelter workers.

Age		Calendar Period				Totals
		1938-1949	1950-1959	1960-1969	1970-1977	
40-49	d	5	5	7	4	21
	n	9.217	14.949	16.123	9.073	49.363
	$\hat{\lambda}$	0.542	0.334	0.434	0.441	0.425
50-59	d	11	24	28	17	80
	n	6.421	10.223	13.663	11.504	41.811
	$\hat{\lambda}$	1.713	2.348	2.049	1.478	1.913
60-69	d	14	24	44	35	117
	n	4.006	4.896	7.555	7.937	24.394
	$\hat{\lambda}$	3.495	4.902	5.824	4.410	4.796
70-79	d	4	12	15	27	58
	n	1.507	1.851	2.724	3.341	9.423
	$\hat{\lambda}$	2.654	6.483	5.506	8.081	6.155
Totals	d	34	65	94	83	276
	n	21.151	31.920	40.066	31.855	124.991
	$\hat{\lambda}$	1.608	2.036	2.346	2.606	2.208

2.2 Rate Standardization

The crude rate $\hat{\lambda}$ often depends on the age distribution of the cohort. Crude rates of different cohorts cannot be compared if they have different age distributions; e.g. comparing death rates for ischemic heart disease between a predominantly young smoking cohort and a predominantly older nonsmoking cohort.

Q1: How can we summarize stratum specific rates into a meaningful single rate?

Q2: Can stratum-specific rates be summarized into an appropriate single rate?

Suppose, for now, that the strata are age categories; e.g. 0-4, 5-9, …, 75-79, 80-84, 85+.

Notation

We will use the following notation in our discussion of standardized rates.

Notation	Description
$\hat{\lambda}$	Crude rate in the cohort
$\hat{\lambda}_1,\ldots,\hat{\lambda}_J$	Crude rates in each strata
$d_1,\ldots,d_J$	Number of cases in each strata
$n_1,\ldots,n_J$	Number of person-years in each strata
$p_1,\ldots,p_J$	Proportion of subjects in each strata

A superscript "(s)" will be used to denote quantities that are based on a "standard population."

2.2.1 Direct Standardization

Direct standardization is a method of combining the stratum-specific rates for the age groups so that the age distribution matches some "standard population." Let $p_j^{(s)}$ denote the proportion of people in the standard population that are in stratum *j*. Then the direct standardized rate (DSR) is

$$DSR = \sum_{j=1}^{J} p_j^{(s)} \hat{\lambda}_j .$$

External Standard Population

One can use an external population as a standard population. For example, census counts are often used.

Table 2. Census Bureau 1950 U.S. population (per 1,000,000).

Age	Population		Age	Population
0 – 4	107,258		45 – 49	60,190
5 – 9	85,591		50 – 54	54,893
10 – 14	73,785		55 – 59	48,011
15 – 19	70,450		60 – 64	40,210
20 – 24	76,191		65 – 69	33,199

Age	Population		Age	Population
25 – 29	81,237		70 – 74	22,641
30 – 34	76,425		75 – 79	14,725
35 – 39	74,629		80 – 84	7,025
40 – 44	67,712		85+	3,828

Or one could use some other census year, other country, specific state, gender, shortened age ranges, etc.

Internal Standard Population

If the cohort is large enough, one may calculate direct standardized rates for sub-cohorts using the entire cohort as the standard population. For example, stratification by two sub-cohorts: exposed and unexposed, where the standard population is the complete cohort.

Comparability of Direct Standardized Rates

Suppose we compute the direct standardized rate for two cohorts, using the same standard population,

$$DSR_1 = \sum_{j=1}^{J} p_j^{(s)} \hat{\lambda}_{1j}$$

$$DSR_2 = \sum_{j=1}^{J} p_j^{(s)} \hat{\lambda}_{2j} \quad .$$

Scenario 1

If $\lambda_{1j} = a \cdot c_j$ and $\lambda_{2j} = b \cdot c_j$ for some constants *a* and *b*, then

$$\frac{DSR_1}{DSR_2} = \frac{a\sum p_j^{(s)} c_j}{b\sum p_j^{(s)} c_j} = \frac{a}{b}$$

regardless of the standard population that is used.

Scenario 2

If $\lambda_{1j} = a \cdot c_j$ and $\lambda_{2j} = b \cdot d_j$, then depending on the choice of standard population DSR_1/DSR_2 may by equal to, less than, or greater than a/b.

Example

Consider the following data.

stratum	$\hat{\lambda}_{1j}$	$\hat{\lambda}_{2j}$	$p_j^{s_1}$	$p_j^{s_2}$
1	0.10	0.20	1/3	1/2
2	0.20	0.25	1/3	1/3
3	0.40	0.20	1/3	1/6

The direct standardized rates will differ depending on whether the standard population s_1 or s_2 is used. Using s_1 gives

$$DSR_1 = 0.10/3 + 0.20/3 + 0.40/3 = 0.233$$
$$DSR_2 = 0.20/3 + 0.25/3 + 0.20/3 = 0.217$$

whereas, s_2 gives

$$DSR_1 = 0.10/2 + 0.20/3 + 0.40/6 = 0.183$$
$$DSR_2 = 0.20/2 + 0.25/3 + 0.20/6 = 0.217.$$

Since the stratum-specific rates are not proportional across the two cohorts, the relative magnitude of the two *DSR*s depends on the choice of a standard population.

2.2.2 Standard Errors for the DSR

Standard errors are typically computed under the assumption that the number of incident cases follows a Poisson distribution

$$d_j \sim Poisson\left(n_j \lambda_j\right)$$

where the expected value and variance are

$$E\left(d_j\right) = n_j \lambda_j$$
$$Var\left(d_j\right) = n_j \lambda_j.$$

Under this assumption the estimated variance of the direct standardized rate is

$$Var\left(DSR\right) = Var\left(\sum p_j^{(s)} \hat{\lambda}_j\right) = Var\left(\sum p_j^{(s)} d_j / n_j\right)$$
$$= \sum \left(p_j^{(s)} / n_j\right)^2 Var\left(d_j\right) \cong \sum \left(p_j^{(s)} / n_j\right)^2 n_j \hat{\lambda}_j .$$
$$= \sum \left(p_j^{(s)} / n_j\right)^2 d_j$$

or

$$SE\left(DSR\right) = \sqrt{\sum \left(p_j^{(s)} / n_j\right)^2 d_j}$$

The distribution of the DSR is somewhat skewed. For the purposes of computing confidence intervals it's better to use the log scale

$$SE\left(\ln DSR\right) \cong \frac{SE\left(DSR\right)}{DSR}$$

Consequently, a Wald 95% confidence interval could be constructed as

$$\exp\left\{\ln DSR \pm 1.96 \cdot SE\left(\ln DSR\right)\right\}.$$

Smelter Workers Example

Consider using a standard population with a uniform distribution for the summary age data from Table 1.

	Age			
	40-49	50-59	60-69	70-79
d	21	80	117	58
n	49.363	41.811	24.394	9.423
$\hat{\lambda}$	0.425	1.913	4.796	6.155
$p_j^{(s)}$	0.25	0.25	0.25	0.25

The direct standardized rate is

$$DSR = 0.25 \cdot 0.425 + 0.25 \cdot 1.913 + 0.25 \cdot 4.796 + 0.25 \cdot 6.155 = 3.322$$

with a variance and standard error of

$$Var(DSR) = (0.25/49.363)^2 21 + (0.25/41.811)^2 80 + (0.25/24.394)^2 117 + (0.25/9.423)^2 58 = 0.05651$$

$$SE(DSR) = \sqrt{0.0.05651} = 0.2377$$

Comments

1. The variance estimator given here assumes that the incident rates are independent across strata.
2. When the stratum-specific rates are estimated form cross-sectional data, the $\hat{\lambda}_1,\ldots,\hat{\lambda}_J$ are independent. However, in a cohort study the $d_1,\ldots,d_J$ are dependent. If an individual dies in period *j*, then he

could not die during any previous period. Nevertheless, when the sample size is large, the assumption of independence is reasonable for computing the variance.

3. A potential weakness of the direct method is that the a priori choice of weights is made without regard for the precision with which the stratum-specific rates are estimated.

2.2.3 Indirect Standardization

The indirect standardized rate (ISR) is

$$ISR = \hat{\lambda}^{(s)} \frac{\sum p_j \hat{\lambda}_j}{\sum p_j \hat{\lambda}_j^{(s)}}.$$

Down's Syndrome Example

In Michigan from 1950-64, 731,177 infants were first-borns, of whom 412 had Down's syndrome ($\hat{\lambda} = 56.3$ per 100,000 first-born live births). In the same period 442,811 infants were fifth-born or more to their mothers, of whom 740 were Down's ($\hat{\lambda} = 167.1$ per 100,000 fifth-born or more live births). The two rates cannot be compared directly because maternal age is associated with both birth order and Down's syndrome. Hence, maternal age should be adjusted for in the analysis. We will use indirect adjustment using Down's syndrome crude ($\hat{\lambda} = 89.5$ per 100,000 live births) and age-specific rates for Michigan.

Maternal Age	$\hat{\lambda}_j^{(s)}$	First-born		Fifth-born or more	
		p_j	$p_j\lambda_j^{(s)}$	p_j	$p_j\lambda_j^{(s)}$
<20	42.5	0.315	13.4	0.001	0.0
20-24	42.5	0.451	19.2	0.069	2.9
25-29	52.3	0.157	8.2	0.279	14.9
30-34	87.7	0.054	4.7	0.339	29.7
35-39	264.0	0.019	5.0	0.235	62.0
40+	864.4	0.004	3.5	0.078	67.4
Totals			54.0		176.6

* Rates are per 100,000 individuals

Therefore, the indirect adjusted rates are

First-born	Fifth-born or more
$ISR = 89.5\frac{56.3}{54.0} = 93.3$	$ISR = 89.5\frac{167.1}{176.6} = 84.7$

Comments

It is not necessarily true that the indirect standardized rates will be equal for two cohorts that have equal stratum-specific rates. This is one potential drawback of the indirect method.

2.3 Comparative Measures of Incidence and Mortality

Need to compare rates between a study cohort and a standard population. Comparison needs to be free from effects of confounding factors.

Strategy: Stratify comparison groups so that within each stratum they are homogeneous with respect to the confounding variables. For example, if age is a confounder, stratify the cohort and standard population into age intervals. We then calculate stratum specific rates and summarize these rates. Direct and indirect standardization are two traditional methods for doing this. Later we will discuss some preferred methods based on the Poisson distribution. For now, we will consider:

1. Comparative Mortality Figures
2. Standardized Mortality Ratios

2.3.1 Comparative Mortality Figure

The comparative mortality figure (CMF) or comparative incidence figure (CIF) is linked to direct standardization. CMF is the ratio of DSR to the standard population rate

$$CMF = \frac{\sum p_j^{(s)} \hat{\lambda}_j}{\sum p_j^{(s)} \hat{\lambda}_j^{(s)}}.$$

Note that the ratio of two study cohorts using the same standard population rates is just the ratio of two DSRs. When the standard population is used as the referent group, the *CMF* simplifies to

$$CMF = \frac{\sum n_j^{(s)} / n^{(s)} \times \hat{\lambda}_j}{\sum n_j^{(s)} / n^{(s)} \times d_j^{(s)} / n_j^{(s)}} = \frac{\sum n_j^{(s)} \hat{\lambda}_j}{\sum d_j^{(s)}}.$$

From this expression one can see that the *CMF* can be interpreted as the ratio of the number of deaths expected in the cohort if it had the same age distribution as in the standard population, divided by the number of deaths in the standard population.

Comments

A disadvantage of the *CMF* is that it may not give stable estimates if the stratum-specific rates are based on small numbers of deaths.

2.3.2 Standard Error of the CMF

Assuming that the size of the standard population is large relative to the cohort so that sampling error of the standard rate can be ignored and $d_j \sim Poisson\left(n_j \lambda_j\right)$, the estimated variance and standard error are

$$Var\left(CMF\right) = Var\left(\frac{\sum p_j^{(s)} d_j / n_j}{\sum p_j^{(s)} \hat{\lambda}_j^{(s)}}\right) = \frac{\sum \left(p_j^{(s)} / n_j\right)^2 Var\left(d_j\right)}{\left(\sum p_j^{(s)} \hat{\lambda}_j^{(s)}\right)^2}$$

$$\cong \frac{\sum \left(p_j^{(s)} / n_j\right)^2 d_j}{\left(\sum p_j^{(s)} \hat{\lambda}_j^{(s)}\right)^2}$$

$$SE\left(CMF\right)=\frac{\sqrt{\sum\left(p_j^{(s)}/n_j\right)^2 d_j}}{\sum p_j^{(s)}\hat{\lambda}_j^{(s)}}.$$

A 95% confidence interval could be calculated on the log scales as

$$\exp\left\{\ln CMF \pm 1.96\cdot SE\left(\ln CMF\right)\right\}$$

where $SE\left(\ln CMF\right)\cong SE\left(CMF\right)/CMF$. Likewise, one could test the null hypothesis of $CMF = 1$ with the test statistic

$$X=\frac{\ln CMF}{SE\left(\ln CMF\right)}\sim N\left(0,1\right).$$

However, there is a better way, in terms of age-specific rates which we will see in Section 3.

2.3.3 Standardized Mortality Ratio

The standardized mortality ratio (SMR) or standardized incidence ratio (SIR) is computed as

$$SMR=\frac{D}{E^{(s)}}$$

where D is the observed number of deaths in the cohort and $E^{(s)}$ is the expected number of deaths in the cohort if the standard population stratum-specific rates apply; i.e.

$$D = \sum d_j$$
$$E^{(s)} = \sum n_j \hat{\lambda}_j^{(s)} .$$

Note that the *SMR* is related to indirect standardization since

$$SMR = \frac{\sum n_j \hat{\lambda}_j}{\sum n_j \hat{\lambda}_j^{(s)}} = \frac{ISR}{\hat{\lambda}^{(s)}} .$$

In other words, the standardized ratio can be interpreted as the ratio of the indirect rate divided by the crude rate in the standard population.

Comments

1. Calculation of the *SMR* is made under the assumption that the rate ratios are constant across strata.
2. Advantages of *SMR* over *CMF*:
    a. Suffices to only know *D* (can often calculate the *SMR* for published data).
    b. When analyzing cross-sectional data according to birth cohort rather than calendar period, often the *CMF* cannot be calculated because age

intervals differ for different birth cohorts. *SMR* would work in this situation.

c. *SMR* more stable then *CMF*; not as sensitive to specific rates based on small numbers of deaths.

3. The *SMR* has smaller variance than the *CMF* and is therefore more appropriate for smaller samples.

2.3.4 Standard Error of the SMR

Assuming that the standard population size is large relative to the cohort and $d_j \sim Poisson\left(n_j \lambda_j\right)$, the estimated variance and standard error are

$$Var\left(SMR\right) = \frac{\sum Var\left(d_j\right)}{\left(\sum n_j \hat{\lambda}_j^{(s)}\right)^2} \cong \frac{\sum d_j}{\left(E^{(s)}\right)^2} = \frac{D}{\left(E^{(s)}\right)^2}.$$

$$SE\left(SMR\right) = \sqrt{D} \big/ E^{(s)}$$

Confidence intervals and statistical tests could be performed on the log scale where

$$SE\left(\ln SMR\right) \cong \frac{SE\left(SMR\right)}{SMR} = \frac{\sqrt{D} \big/ E^{(s)}}{D \big/ E^{(s)}} = \frac{1}{\sqrt{D}}.$$

Wald confidence intervals and test statistics can be computed in the usual way. Alternative methods for hypothesis testing are described in the next section.

2.3.5 Hypothesis Testing for the SMR

Conventional Test

Proposed by Monson (1980) this is based on the test statistic

$$X^2 = \frac{\left(\left|D - E^{(s)}\right| - 0.5\right)^2}{E^{(s)}} \sim \chi_1^2$$

where it is assumed that $D \sim Poisson\left(E^{(s)}\right)$.

Exact Method

For small number of deaths, normal approximations to the skewed Poisson distribution are poor, so get exact p-values directly from the Poisson distribution $D \sim Poisson\left(E^{(s)}\right)$.

Byar's Method

The Byar approximation to the exact Poisson test is based on the test statistic

$$X = \sqrt{9\tilde{D}}\left\{1 - \frac{1}{9\tilde{D}} - \left(\frac{E^{(s)}}{\tilde{D}}\right)^{1/3}\right\} \sim N(0,1)$$

where

$$\tilde{D} = \begin{cases} D & \text{if } D > E^{(s)} \\ D+1 & \text{otherwise} \end{cases}.$$

Variance Stabilizing Transformation

The test statistic for this approximate method is

$$X = 2\left\{\sqrt{D} - \sqrt{E^{(s)}}\right\} \sim N(0,1).$$

2.3.6 Confidence Intervals for the SMR

Exact Method

Exact 95% confidence intervals for the SMR are of the form (SMR_L, SMR_U) where

$$SMR_L = \mu_L / E^{(s)}$$

$$SMR_U = \mu_U / E^{(s)}$$

and μ_L and μ_L are obtained from Table 3.

Table 3. Exact multiplies for computing confidence intervals for the SMR.

95% Intervals				99% Intervals		
D	μ_L	μ_L		D	μ_L	μ_L
1	0.025	5.572		1	0.005	7.430
2	0.121	3.612		2	0.052	4.637
3	0.206	2.922		3	0.113	3.659
4	0.272	2.560		4	0.168	3.149
5	0.325	2.334		5	0.216	2.830
10	0.480	1.839		10	0.372	2.140
15	0.560	1.649		15	0.460	1.878
20	0.611	1.544		20	0.518	1.733
25	0.647	1.476		25	0.560	1.640
50	0.742	1.318		50	0.673	1.425

Byar's Method

Byar's approximation to the exact method gives very good results. It is of the form (SMR_L, SMR_U) where

$$SMR_L = \mu_L \big/ E^{(s)}$$
$$SMR_U = \mu_U \big/ E^{(s)}$$

and

$$\mu_L = D\left(1 - \frac{1}{9D} - \frac{z_{1-\alpha/2}}{3\sqrt{D}}\right)^3$$
$$\mu_U = (D+1)\left(1 - \frac{1}{9(D+1)} + \frac{z_{1-\alpha/2}}{3\sqrt{D+1}}\right)^3.$$

Example: If $D = 15$ and $E^{(s)} = 8.33$ then the *SMR* is

$$SMR = \frac{D}{E^{(s)}} = \frac{15}{8.33} = 1.80$$

the Byar approximation to the 95% confidence interval is computed as follows:

$$\mu_L = 15\left(1 - \frac{1}{9(15)} - \frac{1.96}{3\sqrt{15}}\right)^3 = 8.38917$$
$$\mu_U = 16\left(1 - \frac{1}{9(16)} + \frac{1.96}{3\sqrt{16}}\right)^3 = 24.74182$$

and so

$$SMR_L = \mu_L \big/ E^{(s)} = 8.38917/8.33 = 1.007$$
$$SMR_U = \mu_U \big/ E^{(s)} = 24.74182/8.33 = 2.970.$$

The Byar and exact confidence intervals are the same out to two decimal places: $(1.01, 2.97)$. The exact method is the best. The Byar method is the best approximation to the exact.

Comments

1. The exact method for the confidence interval is described by Mulder (AJE, 1983). Formally, the exact 100(1-α)% confidence interval for the *SMR* is

$$\left(\frac{\frac{1}{2}\chi^2_{\alpha/2,2D}}{E^{(s)}}, \frac{\frac{1}{2}\chi^2_{1-\alpha/2,2D+2}}{E^{(s)}} \right).$$

2. Byar's approximate method is based on the so-called Wilson-Hilferty approximation to the chi-square distribution. For this reason you will sometimes see this method referred to as the Wilson-Hilferty approximation.

2.3.7 Comparison of CMF and SMR

A comparison of the *CMF* and *SMR* can be viewed in terms of bias and variance. The *CMF* has greater variance and the *SMR* has greater bias. A disadvantage of the *SMR*, relative to the *CMF*, is that

- Ratio of *SMR*s for two comparison groups may differ substantially from the age-specific rate ratios. The reason for this is analogous to the "summing of 2×2 tables in the presence of confounding."

Q: When can we combine age-groups? When might the pooled *SMR* (odds ratio) differ from the stratum-specific *SMR*s (odds ratios)?

A: If (1) the *SMR*s from each cohort vary across age groups and (2) the age distributions of the two cohorts differ.

Example

In the following table

1. *SMR*s are larger in the 45-64 group
2. There is a larger proportion of older subjects in Cohort 1 than in Cohort 2.

		Age		
		20-44	45-64	Total (20-64)
Cohort 1	D $E^{(s)}$ SMR_1	100 200 50	1600 800 200	1700 1000 170
Cohort 2	D $E^{(s)}$ SMR_2	80 120 67	180 60 300	260 180 144
	$\frac{SMR_1}{SMR_2}$	0.75	0.67	1.18

The stratum specific *SMR*s are smaller in Cohort 1; however, the pooled *SMR* is larger. *CMF* does not have this problem when the stratum-specific rate ratios are proportional. Specifically, if $\hat{\lambda}_{1j} \big/ \hat{\lambda}_{2j} = \theta$ for all j then the

CMF is unbiased for the overall rate ratio λ_1/λ_2, but the *SMR* is biased.

Unbiasedness of CMF

Note that

$$CMF = \frac{\sum p_j^{(s)} \hat{\lambda}_j}{\sum p_j^{(s)} \hat{\lambda}_j^{(s)}} \Rightarrow \frac{CMF_1}{CMF_2} = \frac{\sum p_j^{(s)} \hat{\lambda}_{1j}}{\sum p_j^{(s)} \hat{\lambda}_{2j}}$$

which, under the assumption that $\hat{\lambda}_{1j} \big/ \hat{\lambda}_{2j} = \theta$, is equal to

$$\frac{\sum p_j^{(s)} \theta \hat{\lambda}_{2j}}{\sum p_j^{(s)} \hat{\lambda}_{2j}} = \theta$$

the constant rate ratio.

Biasedness of SMR

Note that

$$\frac{SMR_1}{SMR_2} = \frac{\sum D_{1j} \big/ \sum E_{1j}^{(s)}}{\sum D_{2j} \big/ \sum E_{2j}^{(s)}} = \frac{\sum n_{1j} \hat{\lambda}_{1j} \big/ \sum n_{1j} \hat{\lambda}_j^{(s)}}{\sum n_{2j} \hat{\lambda}_{2j} \big/ \sum n_{2j} \hat{\lambda}_j^{(s)}}$$

which, under the assumption that $\hat{\lambda}_{1j} \big/ \hat{\lambda}_{2j} = \theta$, is equal to

$$\theta \frac{\sum n_{1j} \hat{\lambda}_{2j} \big/ \sum n_{1j} \hat{\lambda}_j^{(s)}}{\sum n_{2j} \hat{\lambda}_{2j} \big/ \sum n_{2j} \hat{\lambda}_j^{(s)}}.$$

This quantity does not equal θ unless either $\hat{\lambda}_{2j} \propto \hat{\lambda}_j^{(s)}$ or $n_{1j} \propto n_{2j}$ for all *j*. Consequently, the *CMF* is unbiased for the common rate ratio $\hat{\lambda}_1 / \hat{\lambda}_2$.

Comments

1. In practice the *CMF* and *SMR* are often close (but not always). Despite the bias of the *SMR*, it is not true that the *CMF* will be closer to the true rate ratio because the *CMF* may have a large variance.
2. In the above discussion we assumed that the rate ratios were constant across strata. If they are not, then we cannot summarize the stratum specific rate ratios; i.e. the *CMF* and *SMR* are not recommended.
3. Interpretations:
    a. The *CMF* is the proportionate increase (or decrease) in the disease rate that would be expected in the standard population if its members had the same exposure as those in the cohort.
    b. The *SMR* is the proportionate increase (or decrease) in the cohort disease rate due to exposures that occurred as a result of cohort membership, relative to the standard population.

Cohort Data Analysis (171:242/243)

Section 3: Comparison of Exposure Groups

Brian J. Smith, Ph.D.

April 20, 2005

Table of Contents

3.1 Introduction

Section 2 focused on the single exposure category problem, comparing mortality rates of a cohort with those of a standard population. In this section we will

1. Describe time-dependent exposure categories and how to allocate person-years at risk to them,
2. Describe the Montana Smelter Workers Study as an example,
3. Explore methods for comparing death rates among several exposure groups:
    a. External *SMR*s
    b. Internal *SMR*s
    c. Relative risk based on Mantel-Haenszel methods
4. Describe methods/warnings for proportional mortality analyses.

Methods discussed for external *SMR*s and internal *SMR*s in this section are largely of historical interest. The most appropriate methods are the so-called Mantel-Haenszel procedures and the Poisson regression analysis for grouped data given in Section 4. The section ends with proportional mortality analysis which is only used when person-years-at-risk data are unavailable and even then is of dubious value.

3.2 Allocation of Person-Years to Time-Dependent Exposure Categories

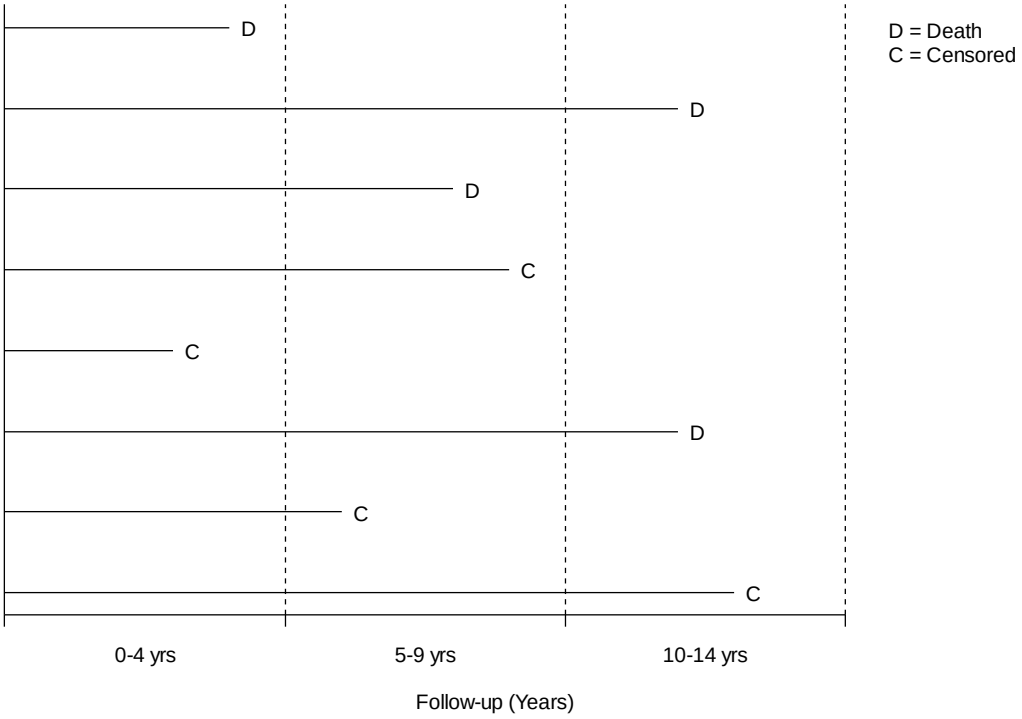
When exposure groups are defined from data available at entry into study, we can just treat each exposure group as a separate cohort and calculate person-years at risk.

However, exposure is often time-dependent; i.e. cumulative exposure changes with time. Each increment in person-years of follow-up is assigned to the same exposure category as would a death had it occurred at that time. Each person may contribute to several categories.

Caution: Do not place someone in an exposure group based solely on total cumulative exposure or duration of employment.

If exposure is continuous over time, then those who live the longest have the highest exposure; the shortest lived have the lowest exposure. This results in the calculated death rates being too low for the high exposure category (number of person-years too larger) and the death rates will be too high in the low exposure category (number of person-years too small). Hence, exposure could be mistakenly found to be beneficial if the time-dependent nature of exposure is ignored.

Person-years Example



	0 – 4 yrs	5 – 9 yrs	10 – 14 yrs
Cases	1	2	1
Person-years			
Correct	37	23	7
Incorrect	7	23	37

3.2.1 Algorithms for Exact Allocation of Person-Years

Can get complicated – for each person must assign his/her person-years to a multidimensional table. For example, time since employment, age, calendar year, time since

cessation of employment $\Rightarrow$ 4-dimensional table and each of these factors is time-dependent.

Clayton's Method (1982)

Appropriate for time-dependent covariates. Need to know exact dates of entry and exit for each cell in the table. The method will be illustrated with an example.

Algorithm

Suppose the three stratification variables are: age, calendar year, and years since first exposure. Consider the strata with age 40-49, calendar period 1950-54, and 5-10 years since first exposure. Let

A = latest date of {date of birth + 40 yrs, 31 December 1949, date of first exposure + 5 yrs}

B = earliest of {date of birth + 50 yrs, 31 December 1954, date of first exposure + 10, date of exit from study}.

If B precedes A, the person contributes no person-years to that strata. Otherwise, the person contributes B – A person-years. Repeat for each person and each stratum in the 3-dimensional table.

3.2.2 Approximate Methods of Allocating Person-Years

A drawback to Clayton's method is that exact dates must be known. Alternatives include

1. Use approximate dates in Clayton's method
2. If integer ages and calendar years are available, use the approximate method from Section 2.

3. Divide each subject's observation period into annual intervals that are allocated in their entirety to a given time-exposure strata.

3.3 Grouped Data from the Montana Copper Smelter Workers Study

Person-years is allocated into a 3-dimensional table defined by age, calendar period, and arsenic exposure. Exposures received during each 10-year period (starting in 1910) are prorated on a linear basis; each individual was classified into the appropriate arsenic exposure duration category at each point in time:

1. Less than one year
2. 1 – 4.9 years
3. 5 – 14.9 years
4. 15+ years

Assignment was based on duration of heavy/moderate exposure at a point two years earlier. This is a crude way of adjusting for bias due to

- Workers who just entered a new cumulative exposure category are necessarily still employed $\Rightarrow$ lower risk of death
- Workers who change employment or retire for health reasons $\Rightarrow$ higher death rates. Also, think about "healthy worker effect". Health status affects hiring, job change, quitting.

Due to a change in the smelting process (average exposure reduced after 1925, the cohort was divided into two groups: 1) 1,482 men employed prior to 1925 and 2) 6,532 men employed on or after 1925. Definition of "high" exposure changes before and after 1925 $\Rightarrow$ very different dose-response curves.

A more appropriate method might have been to classify exposure to pre-1925 and post-1925 in a time-dependent fashion; a person could contribute to both.

For illustrative purposes, the data is summarized in Table 1 by 10-year age and calendar periods.

Table 1. Standard respiratory cancer death rates and standard weights used for comparative analyses of the Montana Smelter Workers data.

Age	Calendar Year				Std. Wgt. (%)
	1938-49	1950-59	1960-69	1970-79	
40-49	0.14817	0.21896	0.28674	0.37391	37.4
50-59	0.47412	0.80277	1.05824	1.25469	30.1
60-69	0.73136	1.55946	2.33029	2.90461	21.5
70-79	0.73207	1.63585	2.85724	4.22945	11.0

3.4 Comparison of Directly Standardized Rates

For the k-th exposure category the direct standardized rate is

$$DSR_k = \sum_{j=1}^{J} p_j^{(s)} \hat{\lambda}_{jk}$$

where $j = 1,\ldots,J$ denote the age-calendar year categories and $p_j^{(s)}$ are weights form a standard population. The comparative mortality figure for the *k*-th exposure category is

$$CMF_k = \frac{DSR_k}{\sum p_j^{(s)} \lambda_j^{(s)}} = \frac{DSR_k}{\lambda^{(s)}}.$$

Comments

1. In practice, comparison of *DSR*s or *CMF*s is limited to studies with a substantial number of deaths in each exposure category; i.e. so that the resulting standardized rates are stable.
2. Hypothesis testing is not commonly done with *CMF*s.

3.5 Comparison of Standardized Mortality Ratios

Note that if data are not extensive and sampling variability is of concern, then *SMR*s are more appropriate than *CMF*s. Define the observed and expected number of deaths (using

rates from an external standard population) for the *k*-th exposure category as follows

$$O_k = \sum_{j=1}^{J} d_{jk}$$
$$E_k^{(s)} = \sum_{j=1}^{J} n_{jk} \lambda_j^{(s)} \quad .$$

Then the *SMR* for the *k*-th exposure category can be written as $SMR_k = O_k \big/ E_k^{(s)}$. We will denote the *SMR* for the entire cohort as $SMR = O_+ \big/ E_+^{(s)}$.

Recall from Section 2 that ratios of *SMR* have serious problems when the ratios of cohort-to-standard-population rates vary widely from one stratum to another. In these cases, the *SMR*s are poor summary measures. For *SMR* analyses to be appropriate, stratum specific rates for each exposure category must be proportional to the external standardized rates.

$$\frac{\lambda_{jk}}{\lambda_j^{(s)}} = \theta_k$$
$$\Rightarrow SMR_k = \theta_k \quad .$$
$$\Rightarrow \frac{SMR_k}{SMR_l} = \frac{\lambda_{jk}}{\lambda_{jl}}$$

If the proportionality assumption holds, then assume that

$$O_k \sim Poisson\left(\theta_k E_k^{(s)}\right)$$

where θ_k is the true, unknown *SMR* for the *k*-th exposure category. Define the relative risk (or rate ratio) as the ratio of the age-specific rates for the *k*-th and first exposure categories ($RR_k = \theta_k / \theta_1$). Note that $RR_1 = 1$. Under the null hypothesis that

$$H_0 : \theta_1 = \ldots = \theta_K \text{ or } H_0 : RR_2 = \ldots = RR_K = 1$$

the "adjusted expected values" of O_k are

$$\tilde{E}_k^{(s)} = O_+ \frac{E_k^{(s)}}{E_+^{(s)}}.$$

3.5.1 Two Dose Levels: Exposed versus Unexposed

Both exact and normal-based approximate methods are available to test the null hypothesis that

$$H_0 : RR_2 = 1.$$

Exact Binomial Distribution

The exact test is based on the test statistic

$$Y \sim Bin\left(O_+, \pi_0\right)$$

where

$$O_+ = \sum d_{ij}$$
$$\pi_0 = E_2^{(s)} / E_+^{(s)}.$$

The two-sided p-value is

$$p = \begin{cases} 2\Pr\left[Y \le O_2\right] & \text{if } O_2 \le \tilde{E}_2^{(s)} \\ 2\Pr\left[Y \ge O_2\right] & \text{otherwise} \end{cases}.$$

Normal Approximation

Usually the number of observed deaths is sufficient to use Normal-theory methods to calculate an approximate p-value. A common choice of test statistic is

$$X^2 = \frac{\left(\left|O_1 - \tilde{E}_1^{(s)}\right| - 0.5\right)^2}{\tilde{E}_1^{(s)}} + \frac{\left(\left|O_2 - \tilde{E}_2^{(s)}\right| - 0.5\right)^2}{\tilde{E}_2^{(s)}} \sim \chi_1^2.$$

This chi-square statistic is inherently two-sided.

3.5.2 Point and Interval Estimation for the Relative Risk

The maximum likelihood estimate of the relative risk is

$$RR_k = \frac{SMR_k}{SMR_1}$$

and there are several methods that can be used to compute confidence intervals.

Exact Confidence Intervals (Pearson and Hartley)

An exact method proposed by Pearson and Hartley in 1962 has the form

$$\left(\frac{\pi_L \tilde{E}_1^{(s)}}{(1-\pi_L)\tilde{E}_k^{(s)}}, \frac{\pi_U \tilde{E}_1^{(s)}}{(1-\pi_U)\tilde{E}_k^{(s)}} \right)$$

where

$$\pi_L = \frac{O_k}{O_k + (O_1 + 1)F_{1-\alpha/2,2O_1+2,2O_k}}$$
$$\pi_U = \frac{(O_k + 1)}{(O_k + 1) + O_1 F_{\alpha/2,2O_1,2O_k+2}}.$$

Approximate Confidence Intervals

Approximate confidence intervals based on the normal approximation to the binomial distribution are of the form (RR_L, RR_U) where RR_L and RR_U are solutions to the quadratic equations:

$$E_1^{(s)}(O_k - 0.5) - RR_L E_k^{(s)}(O_1 - 0.5) = z_{1-\alpha/2}\sqrt{RR_L E_1^{(s)} E_k^{(s)}(O_1 + O_k)}$$
$$E_1^{(s)}(O_k - 0.5) - RR_U E_k^{(s)}(O_1 - 0.5) = -z_{1-\alpha/2}\sqrt{RR_U E_1^{(s)} E_K^{(s)}(O_1 + O_k)}.$$

3.5.3 Testing for Association and Trend in the SMRs

In this section we will consider tests of the null hypothesis

$$H_0 : RR_2 = \ldots = RR_K = 1.$$

Test of General Association

The test for general association is appropriate when the alternative hypothesis is

$$H_A : RR_k \neq RR_l$$

for at least two exposure categories *k* and *l*. The test statistic is

$$X^2 = \sum_{k=1}^{K} \frac{\left(O_k - \tilde{E}_k^{(s)}\right)^2}{\tilde{E}_k^{(s)}} \sim \chi^2_{K-1}.$$

Test for Trend

When it is of interest to test for an ordering of the relative risk, as is the case with alternative hypotheses such as

$$H_0 : 1 < RR_2 < \ldots < RR_K$$

or

$$H_0 : 1 > RR_2 > \ldots > RR_K,$$

a trend statistic of the form

$$X^2 = \frac{\left[\sum_{k=1}^{K} x_k \left(O_k - \tilde{E}_k^{(s)}\right)\right]^2}{\sum_{k=1}^{K} x_k^2 \tilde{E}_k^{(s)} - \left(\sum_{k=1}^{K} x_k \tilde{E}_k^{(s)}\right)^2 \Big/ O_+} \sim \chi_1^2$$

may be used, where $x_1,\ldots,x_k$ are scores associated with the exposure categories. The p-value is inherently two-sided; i.e. a test for an increasing or decreasing trend.

Example: Two Exposure Groups

Suppose $O_1 = 5$, $O_2 = 14$, $E_1^{(s)} = 7.3$, and $E_2^{(s)} = 5.5$. Then the overall standardized mortality ratio is

$$SMR = O_+ \big/ E_+^{(s)} = 19/12.8 = 1.484$$

and the adjusted expected number of deaths are

$$\tilde{E}_1^{(s)} = E_1^{(s)} \frac{O_+}{E_+^{(s)}} = 7.3(1.484) = 10.84$$

$$\tilde{E}_2^{(s)} = E_2^{(s)} \frac{O_+}{E_+^{(s)}} = 5.5(1.484) = 8.16 .$$

The estimated relative risk is

$$RR_2 = \frac{SMR_2}{SMR_1} = \frac{O_2 / E_2^{(s)}}{O_1 / E_1^{(s)}} = \frac{14/5.5}{5/7.3} = 3.7196 .$$

Hypothesis Testing

An approximate test of the hypotheses

$$H_0 : RR_2 = 1$$
$$H_A : RR_2 \neq 1$$

is based on the statistic

$$X^2 = \frac{\left(|5 - 10.84| - 0.5\right)^2}{10.84} + \frac{\left(|14 - 8.16| - 0.5\right)^2}{8.16} = 6.126 \sim \chi_1^2 .$$

The resulting two-sided p-value is

$$p = 2\Pr\left[\chi_1^2 \geq 6.125\right] = 0.0133 .$$

Therefore, at the 5% level of significance, the estimated relative risk is greater than unity.

Confidence Interval

We will use the exact method to compute a 95% confidence interval. Noting that

$$F_{1-\alpha/2,2O_1+2,2O_2} = F_{0.975,12,28} = 2.45$$

$$F_{\alpha/2,2O_1,2O_k+2} = F_{0.025,10,30} = 0.302$$

we compute

$$\pi_L = \frac{O_2}{O_2 + (O_1 + 1)F_{1-\alpha/2,2O_1+2,2O_k}} = \frac{14}{14 + 6(2.45)} = 0.488$$

$$\pi_U = \frac{(O_2 + 1)}{(O_2 + 1) + O_1 F_{\alpha/2,2O_1,2O_k+2}} = \frac{15}{15 + 5(0.302)} 0.909$$

which gives the following confidence interval

$$\left(\frac{\pi_L \tilde{E}_1^{(s)}}{(1-\pi_L)\tilde{E}_2^{(s)}}, \frac{\pi_U \tilde{E}_1^{(s)}}{(1-\pi_U)\tilde{E}_1^{(s)}} \right)$$

$$\left(\frac{0.488(10.84)}{(1-0.488)(8.16)}, \frac{0.909(10.84)}{(1-0.909)(8.16)} \right)$$

$$(1.27, 13.27)$$

Example: Four Exposure Groups

Suppose that we have the following data from four exposure groups:

$$O = \{100, 38, 15, 8\}$$

$$\tilde{E}^{(s)} = \{121.21, 22.63, 11.17, 5.99\}.$$

$$x = \{1, 2, 3, 4\}$$

Test for General Association

The test statistic is

$$X^2 = \frac{(100 - 121.21)^2}{121.21} + \frac{(38 - 22.63)^2}{22.63} + \frac{(15 - 11.17)^2}{11.17} + \frac{(8 - 5.99)^2}{5.99} = 16.1383 \sim \chi_3^2$$

.

At the 5% level of significance, there is a difference between at least two of the relative risks (p = 0.0011).

Test for Trend

The trend statistic is

$$X^2 = \frac{\left[1(100-121.21)+2(38-22.63)+3(15-11.17)+4(6-5.99)\right]^2}{\left\{1(121.21)+4(22.63)+9(11.17)+16(5.99)\right\}-\left\{1(121.21)+2(22.63)+3(11.17)+4(5.99)\right\}^2/161}$$
$$= 8.74 \sim \chi_1^2$$

At the 5% level of significance, there is an increasing trend in the relative risks (p = 0.0031).

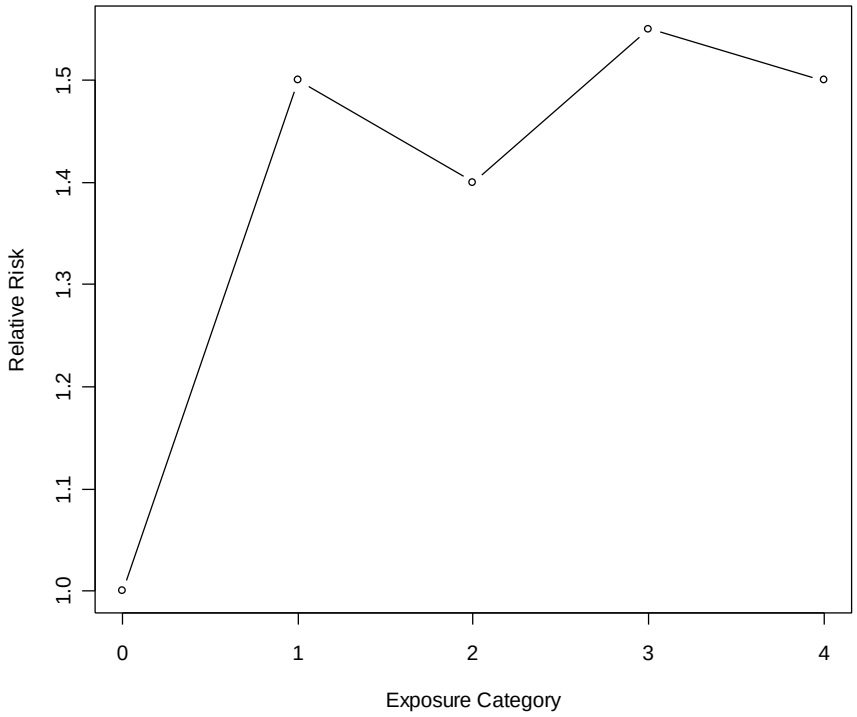
Comments

1. Both of these tests are efficient score tests. This means that they can be computed easily using Poisson regression software.
2. The test of association must be sensitive to all deviations from H_0 whereas the trend test need be sensitive to just two: increasing or decreasing. Hence, the K – 1 degree of freedom association test has far less power to detect a dose-response relationship than the one degree of freedom trend test.

3.5.4 Trend Test for Exposure Effect versus Test for Dose-Response

The trend test may be significant even when the relative risks are not continuously increasing across exposure levels. Fro example, in the figure below, the relative risk is higher for exposure (1-4) categories than for the non-

exposed (0) category; however, the relative risks are not strictly increasing across exposure categories.



Causal Relationship: Trend test may be significant, but causal inference relating exposure to disease is less secure. You may need to worry about how the non-exposed group was chosen (selection bias, confounding). For example, in studies of coffee drinking and bladder cancer, coffee drinkers may be generally more health conscious with respect to diet, exercise, etc. What to try: trend test without the zero exposure category.

Possible Carcinogenic Effect: Need to include zero category. Issue with zero dose group: Should the intercept

for the regression of *SMR* on dose go through 1? Not necessarily.

The trend statistic assumes that the intercept is estimated from the data. This may be appropriate

- To account for healthy worker effect (zero exposure *SMR* < 1).
- When more dying than expected in the zero exposure category.

In these cases a trend test would be biased if the intercept is assumed to be unity.

3.5.5 Selection of the Dose Metameter

The choice of scores $x_1,\ldots,x_K$ to use for the dose categories in the trend test will affect the conclusion. For example, using the observed – expected number of deaths from a hematological cancer study of navy shipyard workers:

Scores	Trend Statistic	p-value
Linear	1.19	0.14
Log	2.25	0.07
Threshold	3.53	0.03

Of course, it is cheating to plot the data, pick the scores, and then do a strict hypothesis test. This is not an easy problem and some serious thought needs to be devoted to it.

If dose level is smoking, how does one quantify an ex-smoker? If exposure is asbestos, it is clear that an ex-

asbestos worker is categorized by cumulative dosage, unlike an ex-smoker. For other exposures, the time-dependent rate of exposure may be more important than the time-dependent cumulative dose.

3.6 Comparison of Internally Standardized Mortality Rates

The appropriateness of using standard rates from an external population is sometimes questionable. Making internal comparisons using only observed data is reasonable in this case.

*The methods described in this section are a rough approximation to the preferred methods discussed in the next section.

Adjusted Expected Values

Previously we computed the expected number of deaths under the assumption that an external standard population was being used to compute the *SMR*s; i.e.

$$E[O_k] = \tilde{E}_k^{(s)} = O_+ \frac{E_k^{(s)}}{E_+^{(s)}}$$

where $E_k^{(s)} = \sum_j n_{jk} \hat{\lambda}_j^{(s)}$.

When internal standardization is used, the expected number of deaths is

$$E[O_k] = E_k$$

where $E_k = \sum_j n_{jk} \hat{\lambda}_j$.

Estimation and Testing using Internal Rates

Same form as those using external rates, just replace $\tilde{E}_k^{(s)}$ by E_k. For example,

$$RR_k = \frac{O_k / E_k}{O_1 / E_1}.$$

Tests of $H_0 : RR_2 = 1$ are based on the statistic

$$X^2 = \frac{\left(|O_1 - E_1| - 0.5\right)^2}{E_1} + \frac{\left(|O_2 - E_2| - 0.5\right)^2}{E_2} \sim \chi_1^2.$$

Tests of $H_0 : RR_2 = \ldots = RR_K$ are based on the general chi-square statistic

$$X^2 = \sum_{k=1}^{K} \frac{\left(O_k - E_k\right)^2}{E_k} \sim \chi_{K-1}^2$$

or the trend statistic

$$X^2 = \frac{\left[\sum_{k=1}^{K} x_k \left(O_k - E_k\right)\right]^2}{\sum_{k=1}^{K} x_k^2 E_k - \left(\sum_{k=1}^{K} x_k E_k\right)^2 \Big/ O_+} \sim \chi_1^2.$$

The latter two tests are conservative.

Comments

1. Internal standardization is a rough approximation to the methods presented in the next section as well as Poisson regression.

2. If age and calendar time (or other stratification variables) confound the exposure-disease relationship, this procedure is conservative.
3. If there are more than two exposure groups, internal standardization does not eliminate the problem of non-comparability of *SMR*s: the pooled “internal” group may be dominated by one or more large exposure groups.

3.7 Preferred Methods of Analysis for Grouped Data

Preferred methods for cohort data are very similar to those for case-control data. Replace "cases" by "deaths", "controls" by "number of person-years" and you have made the link between the two.

In the *j*-th stratum and *k*-th exposure group, let

a_{jk} = number of cases d_{jk} = number of deaths

c_{jk} = number of controls n_{jk} = number of person years

Case-Control Data

The following are properties of case-control analyses:

- The relative measure of risk is

$$OR_{jk} = \frac{a_{jk}/c_{jk}}{a_{j1}/c_{j1}}.$$

- The given 2×K table margins $\{a_{j1},\ldots,a_{jK}\}$ are assumed to have a multivariate hypergeometric distribution.
- Regression is performed with the logistic model.

Cohort Data

The following are properties of cohort analyses:

- The relative measure of risk is

$$RR_{jk} = \frac{d_{jk}/n_{jk}}{d_{j1}/n_{j1}}.$$

- The observed deaths $D_j = \sum_k d_{jk}$ is assumed to have a multinomial distribution.
- Regression is performed with the Poisson model.

3.7.1 Crude Relative Risk

Suppose that we wish to compare two cohorts for which the number of deaths and person-years are O_i and N_i, $i = 1,2$. In the case of no stratification, the relative risk is simply

$$RR = \frac{O_2/N_2}{O_1/N_1}.$$

The approximate chi-square statistic that can be used to test the null hypothesis $H_0 : RR = 1$ is

$$X^2 = \frac{\left(\left|O_2 - E\left[O_2\right]\right| - 0.5\right)^2}{Var\left[O_2\right]}$$

where

$$E\left[O_2\right] = O_+ \frac{N_2}{N_+}$$

$$Var\left[d_2\right] = O_+ \frac{N_1 N_2}{N_+^2}$$

such that $O_+ = O_1 + O_2$ and $N_+ = N_1 + N_2$.

Example

Consider the data:

	Deaths (O_i)	Person-Years (N_i)
Unexposed	5	7300
Exposed	14	5500
Totals	19	12800

The estimated relative risk is

$$RR = \frac{O_2/N_2}{O_1/N_1} = \frac{14/5500}{5/7300} = 3.716$$

and the approximate chi-square statistic is

$$X^2 = \frac{\left(\left|14 - 19\dfrac{5500}{12800}\right| - 0.5\right)^2}{19\dfrac{(7300)(5500)}{(12800)^2}} = 6.115.$$

At the 5% level of significance, the relative risk is different from unity (p = 0.0134).

3.7.2 Mantel-Haenszel Estimator (Two Exposure Groups)

Suppose there are $K = 2$ exposure categories and *J* strata. Denote the *j*-th table of data as

	Deaths	Person-Years
Unexposed	d_{j1}	n_{j1}
Exposed	d_{j2}	n_{j2}
Totals	D_j	N_j

Then the Mantel-Haenszel estimator of the exposure-disease relative risk is

$$RR_{MH} = \frac{\sum_j d_{j2} n_{j1} / N_j}{\sum_j d_{j1} n_{j2} / N_j}$$

with variance and standard error given by

$$Var\left[RR_{MH}\right] = \frac{RR_{MH} \sum_j n_{j1} n_{j2} D_j / N_j^2}{\left(\sum_j \frac{n_{j1} n_{j2} D_j}{N_j \left(n_{j1} + RR_{MH} n_{j2} \right)} \right)^2}.$$

$$SE\left[RR_{MH}\right] = \sqrt{Var\left[RR_{MH}\right]}$$

Since the distribution is so skewed, it is recommended to compute confidence intervals and test statistics on the natural log scale where

$$SE\left[\ln RR_{MH}\right] = \frac{SE\left[RR_{MH}\right]}{RR_{MH}}.$$

Thus, a 95% Wald confidence interval is

$$\exp\left\{\ln RR_{MH} \pm 1.96 SE\left[\ln RR_{MH}\right]\right\}$$

for which the associated test statistic for $H_0 : RR = 1$ is

$$X = \frac{\ln RR_{MH}}{SE\left[\ln RR_{MH}\right]} \sim N(0,1).$$

Approximate Test

Although the Wald statistic can be used to test the null hypothesis $H_0 : RR = 1$, the preferred approximate test statistic is

$$X^2 = \frac{\left\{\left|O_2 - E\left[O_2\right]\right| - 0.5\right\}^2}{Var\left[O_2\right]}$$

$$= \frac{\left\{\left|O_2 - \sum_j D_j n_{j2} / N_j\right| - 0.5\right\}^2}{\sum_j D_j n_{j2} n_{j1} / N_j^2} \sim \chi_1^2$$

British Doctors Example

Consider the data below from a study of coronary heart disease in British male doctors.

Table 2. Deaths from coronary disease among British male doctors.

Age Group (j)	Smoker (i)	Deaths $\left(d_{ji}\right)$	Person-Years $\left(n_{ji}\right)$
35-44	No	2	18,790
	Yes	32	52,407

Age Group (j)	Smoker (i)	Deaths $\left(d_{ji}\right)$	Person-Years $\left(n_{ji}\right)$
45-54	No	12	10,673
	Yes	104	43,248
55-64	No	28	5,710
	Yes	206	28,612
65-74	No	28	2,585
	Yes	186	12,663
75-84	No	31	1,462
	Yes	102	5,317

The Mantel-Haenszel estimate is

$$RR_{MH} = \frac{32 \times 18,790/71,197 + \ldots + 102 \times 1,462/6,779}{2 \times 52,407/71,197 + \ldots + 31 \times 5,317/6,779} .$$
$$= 1.42$$

3.7.3 Tests for Homogeneity of Relative Risks

The fundamental assumption when estimating a common relative risk is that the relative risks across the J strata are the same (homogeneous); i.e. age is not an effect modifier. A test of the homogeneity, that

$$H_0 : RR_{(1)} = \ldots = RR_{(J)}$$

can be carried out based on the chi-square statistic

$$X^2 = \sum \left\{ \frac{\left(d_{j1} - \hat{d}_{j1}\right)^2}{\hat{d}_{j1}} + \frac{\left(d_{j2} - \hat{d}_{j2}\right)^2}{\hat{d}_{j2}} \right\} \sim \chi^2_{J-1}$$

where one typically uses

$$\hat{d}_{j2} = D_j \frac{RR_{MH} n_{j2}}{n_{j1} + RR_{MH} n_{j2}} .$$
$$\hat{d}_{j1} = D_j - \hat{d}_{j2}$$

3.7.4 Tests for Trend

A test for trend across the *J* stratum-specific relative risks can be carried out with the test statistic

$$X^2 = \frac{\left[\sum_j x_j \left(d_{j2} - \hat{d}_{j2}\right)\right]^2}{\sum_j x_j^2 \hat{d}_{j1} \hat{d}_{j2} / D_j - \left(\sum_j x_j \hat{d}_{j1} \hat{d}_{j2} / D_j\right)^2 \Big/ \left\{\sum_j \hat{d}_{j1} \hat{d}_{j2} / D_j\right\}} \sim \chi^2_1$$

where the $x_1,\ldots,x_J$ are stratum-specific scores and, as in the previous section,

$$\hat{d}_{j2} = D_j \frac{RR_{MH} n_{j2}}{n_{j1} + RR_{MH} n_{j2}} .$$
$$\hat{d}_{j1} = D_j - \hat{d}_{j2}$$

British Doctors Example

From Table 2 we have the following

	35-44	45-54	55-64	65-74	75-84
x_j	1	2	3	4	5
$d_{j1}\left(\hat{d}_{j1}\right)$	2 (6.83)	12 (17.13)	28 (28.75)	28 (26.82)	31 (21.52)
$d_{j2}\left(\hat{d}_{j2}\right)$	32 (27.16)	104 (98.87)	206 (205.25)	186 (187.18)	102 (111.48)
D_j	34	116	234	214	133

Note that

$$\sum_j x_j\left(d_{j2}-\hat{d}_{j2}\right)=1(32-27.16)+\ldots+5(102-111.48)$$
$$=-34.92$$
$$\sum_j x_j^2\hat{d}_{j1}\hat{d}_{j2}/D_j=1116.72231$$
$$\sum_j x_j\hat{d}_{j1}\hat{d}_{j2}/D_j=294.23598$$
$$\sum_j \hat{d}_{j1}\hat{d}_{j2}/D_j=86.74386$$

and so

$$X^2=\frac{[-34.920]^2}{1116.72231-\frac{(294.23598)^2}{86.74386}}=10.3\sim\chi_1^2$$

Therefore, at the 5% level of significance, there is an increasing trend in the relative risks across age groups (p = 0.0013).

3.7.5 Mantel-Haenszel Estimator (Multiple Exposure Groups)

There is a generalization of the Mantel-Haenszel estimator to accommodate more two or more exposure categories. In general there are $k = 1,\ldots,K$ exposure categories and we would like to test the hypothesis

$$H_0 : RR_2 = \ldots = RR_K = 1.$$

Under this null hypothesis the $\{d_{j1},\ldots,d_{jK}\}$ have a multinomial distribution with mean and covariance matrix given by

$$E\left[d_{jk}\right] = n_{jk}\frac{D_j}{N_j}$$

$$\left\{\boldsymbol{\Sigma}_j\right\}_{kl} = Var\left[d_{jk}, d_{jl}\right] = \begin{cases} n_{jk}\left(N_j - n_{jk}\right)D_j / N_j^2 & \text{if } k = l \\ -n_{jk}n_{jl}D_j / N_j^2 & \text{if } k \neq l \end{cases}.$$

Define $\mathbf{O}^T = \{O_1,\ldots,O_{K-1}\}$, $\mathbf{E}^T = \{E_1,\ldots,E_{K-1}\}$, and $\boldsymbol{\Sigma}$ such that

$$O_k = \sum_j d_{jk}$$

$$E_k = \sum_j E\left[d_{jk}\right] = \sum_j n_{jk}\frac{D_j}{N_j}.$$

$$\boldsymbol{\Sigma} = \sum_j \boldsymbol{\Sigma}_j$$

The Mantel-Haenszel test for equal risks across exposure categories is based on the statistic

$$X^2 = \left(\mathbf{O} - \mathbf{E}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{O} - \mathbf{E}\right) \sim \chi^2_{K-1}.$$

Tests of trend can be carried out with the statistic

$$X^2 = \frac{\left[\sum_k x_k (O_k - E_k)\right]^2}{\sum_k x_k^2 E_k - \sum_j \left(\sum_k x_k n_{jk} D_j / N_j\right)^2 / D_j} \sim \chi_1^2 .$$

3.7.6 Conservatism of Indirect Standardization

This section provides an example of how statistical confounding can make the internal standardized test conservative.

	Stratum 1		Stratum 2		Totals	
Exposed	Cases	P-Yrs	Cases	P-Yrs	Cases	P-Yrs
No	25	10,000	5	4,000	30	14,000
Yes	5	1,000	25	10,000	30	11,000
RR	2.0		2.0		1.27	

Method of Internal Standardization

Using internal standardization yields the following expected number of deaths:

$$E_1 = \sum_{j=1}^{2} n_{j1} \frac{D_j}{N_j} = 10{,}000 \frac{30}{11{,}000} + 4{,}000 \frac{30}{14{,}000} = 35.844$$

.

$$E_2 = 1{,}000 \frac{30}{11{,}000} + 10{,}000 \frac{30}{14{,}000} = 24.156$$

The resulting relative risk and test statistic are

$$RR = \frac{O_2/E_2}{O_1/E_1} = \frac{30/24.156}{30/35.844} = 1.48$$

$$X^2 = \frac{\{|30 - 35.844| - 0.5\}^2}{35.844} + \frac{\{|30 - 24.156| - 0.5\}^2}{24.156}.$$
$$= 1.98 \sim \chi_1^2$$

for which the two-sided p-value is $p = \Pr\left[\chi_1^2 \geq 1.98\right] = 0.1594$.

Method of Mantel-Haenszel

For this method

$$E[O_2] = \sum_{j=1}^{2} D_j \frac{n_{j2}}{N_j} = 30\frac{1{,}000}{11{,}000} + 30\frac{10{,}000}{14{,}000} = 24.156$$

$$Var[O_2] = \sum_{j=1}^{2} D_j n_{j1} n_{j2} / N_j^2$$
$$= \frac{30(10{,}000)(1{,}000)}{11{,}000^2} + \frac{30(4{,}000)(10{,}000)}{14{,}000^2}.$$
$$= 8.6018$$

The relative risk estimate and test statistic are

$$RR_{MH} = \frac{\sum_j d_{j2} n_{j1}/N_j}{\sum_j d_{j1} n_{j2}/N_j} = 2.0$$

$$X^2 = \frac{\{|30 - 24.156| - 0.5\}^2}{8.6018} = 3.32 \sim \chi_1^2$$

for which the two-sided p-value is $p = \Pr\left[\chi_1^2 \geq 3.32\right] = 0.0684$.

Comments

1. The stratum-specific rate ratios are constant (2.0) and, thus, it is appropriate to estimate an overall relative risk.
2. The internal standardization here uses the ratio of the *SMR*s to estimate the relative risk. The estimate is biased because of the difference in the distribution of person-years for exposed and unexposed across strata.
3. The Mantel-Haenszel estimate is not biased and yields a more powerful statistical test.

3.8 Proportional Mortality and Dose-Response Analysis

Setting: Suppose that one needs to conduct a dose-response analysis using number of deaths only, without person-year information. This may be due to:

1. Person-year data not being available
2. Complete exposure history has been reconstructed for the dead and we want an initial evaluation of relative risk to see if it's worthwhile to go through the lengthy process of acquiring person-year information in the rest of the cohort.

Let d_{jk} denote the number of deaths from cause of interest (stratum j, exposure group k), t_{jk} the number of deaths from all causes, and x_k the dose level associated with exposure group k. Define

$$D_j = \sum_k d_{jk}$$
$$T_j = \sum_k t_{jk}$$

Goal: Determine whether the proportion of death due to cause of interest increases with increasing levels of exposure while adjusting for the stratified variables.

Warning: Competing Risk Problem – If other causes of death are affected by the exposure, then the cause-specific proportion deaths will not allow unbiased estimation of the relative risk.

Assumption: The necessary assumption is that the other causes of death are not related to the exposure – in practice, we may need to exclude those "other" causes that are known to be related to the exposure.

Analysis

Assume that those dying from other causes represent an unbiased sample of the population at risk within each stratum. These will be the controls; those dying from the cause of interest are the cases. Use case-control methods to analyses the data.

Note: The "controls" are assumed to be representative of the population at risk. We are typically uncertain of this. Thus, inference for a proportional mortality study is more tentative than for a case-control study.

3.9 Overview of Estimation and Testing Procedures

Notation

Term	Description
$k = 1,\ldots,K$	Index for the exposure groups
$j = 1,\ldots,J$	Index for the stratification variable
d_{jk}	Number of cases
n_{jk}	Number of person-years
λ_{jk}	Rate in the cohort
$\lambda_j^{(s)}$	Rate in the standard population
$D_j = \sum_k d_{jk}$	Observed cases in stratum *j*
$O_k = \sum_j d_{jk}$	Observed cases at exposure *k*
$E_k^{(s)} = \sum_j n_{jk}\lambda_j^{(s)}$	Expected cases using an external population
$E_+^{(s)} = \sum_k E_k^{(s)}$	Total expected cases
$E_k = \sum_j n_{jk}\hat{\lambda}_j$	Expected cases using the entire cohort as the standard population
$N_j = \sum_k n_{jk}$	Person-years in stratum *j*

Proportionality Assumption

Necessary assumption for comparing *SMR*s from two cohorts; i.e. $\lambda_{jk} / \lambda_{jl} = \theta$ for all $j = 1,\ldots,J$.

Summary of Methods

	External Standardization	Internal Standardization	Mantel-Haenszel
Comparison Across Strata	-	-	Homogeneity (3.7.3) Trend (3.7.4)
Estimation	$RR_k = \frac{O_k / E_k^{(s)}}{O_1 / E_1^{(s)}}$	$RR_k = \frac{O_k / E_k}{O_1 / E_1}$	$RR_k = \frac{\sum d_{jk} n_{j1} / N_j}{\sum d_{j1} n_{jk} / N_j}$
Confidence Interval	Exact (3.5.2) Approximate (3.5.2)	Same as for external; replace $\tilde{E}_k^{(s)}$ by E_k	Exact – Software Approximate (3.7.2)
Test of Association $H_0 : RR_2 = 1$	Exact (3.5.1) Approximate (3.5.1)	Approximate (3.6)*	Approximate (3.7.2)
Test of Association $H_0 : RR_2 = \ldots = RR_K = 1$	Approximate (3.5.3)	Approximate (3.6)*	Approximate (3.7.5)
Trend Test	Approximate (3.5.3)	Approximate (3.6)*	Approximate (3.7.5)

* Mantel-Haenszel methods are preferred

Cohort Data Analysis (171:242/243)
Section 4: Poisson Regression

Brian J. Smith, Ph.D.

April 27, 2005

Table of Contents

4.1 Introduction

In this section we discuss Poisson regression, a method that is appropriate for modeling a discrete response variable that takes on non-negative values (0, 1, 2...).

4.1.1 Montana Smelter Workers Study Revisited

Consider the follow-up data from the Montana study that was grouped by exposure, age, calendar, year, and hiring categories. We will use the variables summarized in the table below to model the risk of death from respiratory cancer in this cohort.

Variables		Description	Values
Outcome	respiratory	Number of deaths from respiratory cancer	numerical
	pyears	Number of person-years	numerical
Predictor	arsenic	Years of arsenic exposure	1 = 0.0-0.9 2 = 1.0-4.9 3 = 5.0-14.9 4 = 15.0+
Confounders	age	Age groups	1 = 40-49 2 = 50-59 3 = 60-69 4 = 70-79
	year	Calendar year	1 = 1938-1949 2 = 1950-1959 3 = 1960-1969 4 = 1970-1977
	period	Hiring periods	1 = before 1925 2 = 1925+

Years of arsenic exposure is the primary risk factor of interest in this study. One could use the Mantel-Haenszel or SMR approaches of Section 3 to estimate the adjusted relative risk of death associated with arsenic exposures. Recall that, in Section 3, stratification was used to adjust for the confounding variables. A limitation of this approach is that it does not allow for the joint estimation of relative risk across both the exposure and stratification variables.

In this section, a multivariate regression approach is presented for the estimation and standardization of rates for cohort data.

4.1.2 The Poisson Distribution

The Poisson distribution can be used to describe a discrete random variable that takes on non-negative values. We will use it to model the observed number of deaths *d* given *n* person-years of follow-up. Let λ denote the true rate of death. Our multivariate regression model will be based on the assumption that *d* is distributed

$$d \sim Poisson(n\lambda).$$

The probability function for this random variable is given by

$$\Pr[d = x] = \frac{e^{-n\lambda}(n\lambda)^x}{x!}$$

For example, suppose that the true death rate for a particular cohort is 1 per 100 person-years $(\lambda = 0.01)$. If a study of individuals in this cohort yielded *n* = 5000 person-years of follow-up, the Poisson distribution function for the number of deaths would be

$$\Pr[x] = \frac{e^{-5000(0.01)}(5000 \cdot 0.01)^x}{x!} = \frac{e^{-50}(50)^x}{x!}$$

and is plotted in Figure 1.

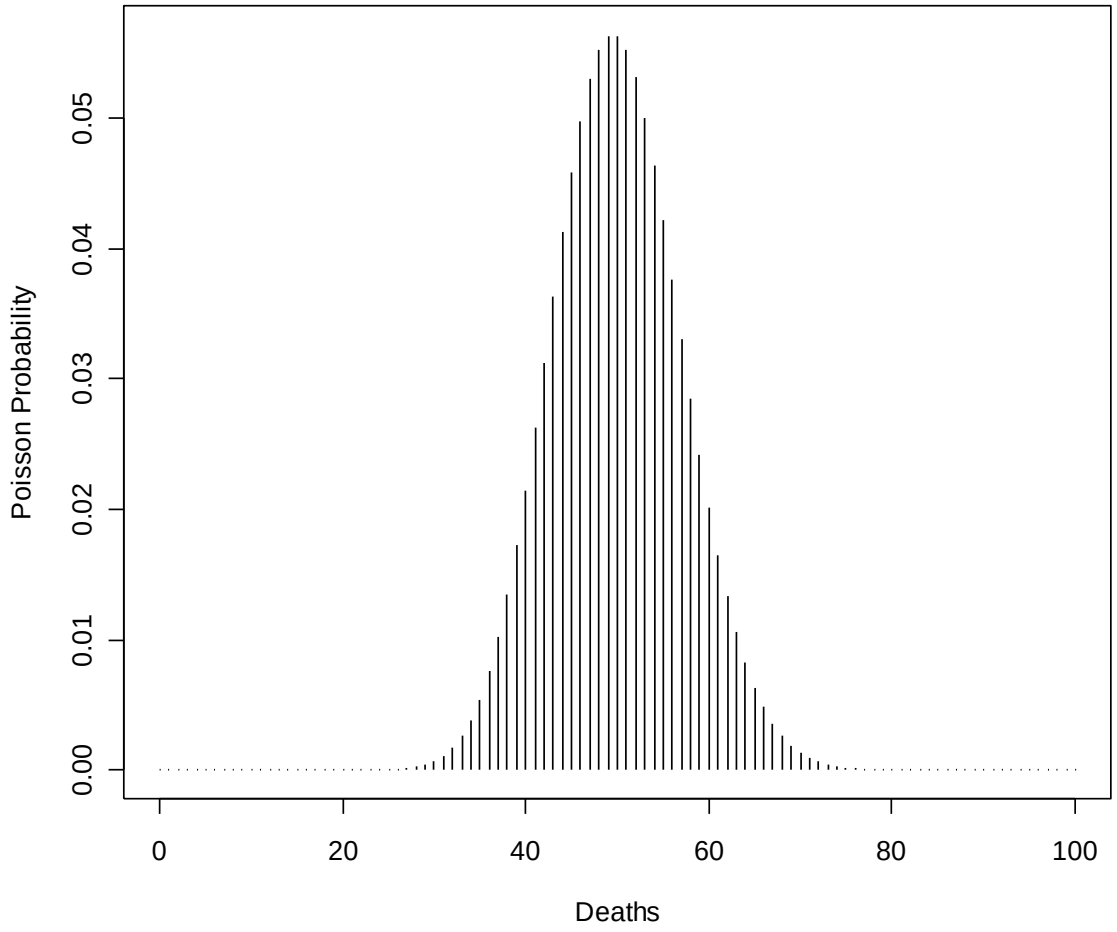


Figure 1. Probabilities for a *Poisson*(50) random variable

Properties:

- A Poisson random variable may take on any non-negative value, including zero.

- The Poisson distribution for the number of deaths has expected value and variance equal to

$$E[d] = Var[d] = n\lambda .$$

- The Poisson parameter $n\lambda$ must be positive.

The Poisson distribution is an approximation to the exact distribution for d/n and will be an adequate approximation provided:

1. The rate of death λ is sufficiently small, and
2. Only a fraction of the cohort members are expected to die during the follow-up period.

4.2 Poisson Regression Model

4.2.1 Model Specification

In Poisson regression, the observed number of deaths is modeled as a multivariate function of the predictor variables $\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)$ according to the following relationship:

$$d \sim Poisson\left(n\lambda\left(\mathbf{x}\right)\right)$$
$$\ln\left[\lambda\left(\mathbf{x}\right)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p .$$

It can be seen that this is a multiplicative model for the rate parameter be re-writing the formula as

$$\lambda\left(\mathbf{x}\right) = \exp\left\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$
$$= e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} \cdots e^{\beta_p x_p}$$

The multiplicative model is commonly used in practice because it ensures that the rate parameter λ will be positive for all possible values of the predictors and coefficients.

4.2.2 Generalized Linear Model

Poisson regression models fall within the generalized linear models framework. Specifically, the mean of the assumed Poisson distribution can be written as a function of the linear predictor,

$$\ln\left[n\lambda(\mathbf{x})\right] = \ln[n] + \ln\left[\lambda(\mathbf{x})\right]$$
$$= \ln[n] + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The $\ln[n]$ term is called an "offset". An offset is a term in the linear predictor with a coefficient that is fixed, rather than estimated. It is needed here because the mean of the Poisson distribution $n\lambda(\mathbf{x})$ must be linked to the linear predictors in the estimation routines.

Smelter Study Example

Define the following indicator variables for the regression analysis:

arsenic2 = I(arsenic = 2) arsenic3 = I(arsenic = 3) arsenic4 = I(arsenic = 4)	year2 = I(year = 2) year3 = I(year = 3) year4 = I(year = 4)
age2 = I(age = 2) age3 = I(age = 3) age4 = I(age = 4)	period2 = I(period = 2)

We will model the reported deaths from pulmonary cancer in this cohort using Poison regression with the following model for the death rate:

$$\lambda(\mathbf{x}) = \exp\left\{\begin{array}{l}\beta_0 + \beta_1 arsenic2 + \beta_2 arsenic3 + \beta_3 arsenic4 \\ \quad + \beta_4 age2 + \beta_5 age3 + \beta_6 age4 \\ \quad + \beta_7 year2 + \beta_8 year3 + \beta_9 year4 \\ \quad + \beta_{10} period2\end{array}\right\}$$

SAS Poisson Regression

```
data smeltermod;
   set smelter;
   arsenic2 = (arsenic = 2);
   arsenic3 = (arsenic = 3);
   arsenic4 = (arsenic = 4);
   age2 = (age = 2);
   age3 = (age = 3);
   age4 = (age = 4);
   year2 = (year = 2);
   year3 = (year = 3);
   year4 = (year = 4);
   period2 = (period = 2);
   lpyears = log(pyears);
run;

proc genmod data=smeltermod;
   model respiratory = arsenic2 arsenic3 arsenic4 age2 age3 age4
                       year2 year3 year4 period2
                       / dist=poisson offset=lpyears;
run;
```

Details

- PROC GENMOD is a SAS procedure for fitting generalized linear models, which includes the linear, logistic, and Poisson models
- Poisson regression is specified with the **dist** option in the model statement. The default is to model the natural log-transformed Poisson parameter as a function of the linear predictor.
- The **offset** option is available to add a linear predictor for which the coefficient is fixed and not estimated. Note that the log-person-years offset must be first defined in the data step.

The GENMOD Procedure

Model Information

Data Set	WORK.SMELTERMOD
Distribution	Poisson
Link Function	Log
Dependent Variable	respiratory
Offset Variable	lpyears

Number of Observations Read	114
Number of Observations Used	114

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	103	105.7389	1.0266
Scaled Deviance	103	105.7389	1.0266
Pearson Chi-Square	103	105.3665	1.0230
Scaled Pearson X2	103	105.3665	1.0230
Log Likelihood		122.6024	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-8.0670	0.2857	-8.6269	-7.5071	797.38	<.0001
arsenic2	1	0.7982	0.1582	0.4880	1.1083	25.44	<.0001
arsenic3	1	0.5734	0.2062	0.1692	0.9776	7.73	0.0054
arsenic4	1	0.9218	0.1810	0.5670	1.2766	25.93	<.0001
age2	1	1.3870	0.2468	0.9034	1.8706	31.60	<.0001
age3	1	2.1926	0.2445	1.7133	2.6719	80.40	<.0001
age4	1	2.3447	0.2702	1.8150	2.8744	75.28	<.0001
year2	1	0.5336	0.2151	0.1120	0.9552	6.15	0.0131
year3	1	0.6870	0.2143	0.2669	1.1071	10.27	0.0013
year4	1	0.6588	0.2305	0.2070	1.1106	8.17	0.0043
period2	1	-0.5116	0.1530	-0.8114	-0.2118	11.19	0.0008
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

4.3 Inference

4.3.1 Relative Risk Estimation

The relative risk (*RR*) is the risk in one group relative to the risk in another. In the multivariate regression setting, it is often of interest to estimate the risk ratio for subjects with covariates $\mathbf{x}'$ relative to those with covariates $\mathbf{x}''$. The general steps for computing relative risks from the results of a Poisson regression are:

1. Write out the ratio of rates using the model specified in the Poisson regression,

$$RR = \frac{\lambda(\mathbf{x}')}{\lambda(\mathbf{x}'')} = \frac{\exp\left\{\beta_0 + \beta_1 x_1' + \beta_2 x_2' + \ldots + \beta_p x_p'\right\}}{\exp\left\{\beta_0 + \beta_1 x_1'' + \beta_2 x_2'' + \ldots + \beta_p x_p''\right\}}.$$

2. Reduce this equation to a form that is the exponential of the estimated regression parameters.

$$RR = \exp\left\{\beta_1\left(x_1' - x_1''\right) + \beta_2\left(x_2' - x_2''\right) + \ldots + \beta_p\left(x_p' - x_p''\right)\right\}$$

3. Insert the regression estimates for the parameters in order to calculate the relative risk.

If the value of a predictor variable is the same in the numerator and denominator rates, then that predictor does not factor into the calculation of the hazard ratio. For instance, if $x_p' = x_p''$ then

$$\beta_p\left(x_p' - x_p''\right) = 0$$

and so the term for the p^{th} predictor drops out of the equation.

4.3.2 Wald Statistics

Estimates of relative risks are often accompanied by confidence intervals and p-values in order to provide measures of statistical significance. Suppose that a Poisson regression model of the form

$$\lambda(\mathbf{x}) = \exp\left\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right\}$$

is fit to a dataset, and interest lies in making inference about the relative risk

$$\begin{aligned} RR &= \exp\left\{\beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_p c_p\right\} \\ &= \exp\left\{\sum_{i=1}^{p} \beta_i c\right\} \end{aligned}$$

where c_i are specified constants. The **Wald 100(1 - α)% confidence interval** for this relative risk is

$$\begin{aligned} CI &= \exp\left\{\sum \beta_i c_i \pm z_{1-\alpha/2} se\left(\sum \beta_i c_i\right)\right\} \\ &= RR \exp\left\{\pm z_{1-\alpha/2} se\left(\sum \beta_i c_i\right)\right\} \end{aligned}$$

Corresponding tests are based on the Wald statistic

$$X^2 = \left(\frac{\sum \beta_i c_i}{se\left(\sum \beta_i c_i \right)} \right)^2 \sim \chi_1^2 .$$

Summary of Regression Results

Variable	Parameter	Estimate	SE
Intercept	β_0	-8.067	0.2857
arsenic2	β_1	0.7982	0.1582
arsenic3	β_2	0.5734	0.2062
arsenic4	β_3	0.9218	0.181
age2	β_4	1.387	0.2468
age3	β_5	2.1926	0.2445
age4	β_6	2.3447	0.2702
year2	β_7	0.5336	0.2151
year3	β_8	0.687	0.2143
year4	β_9	0.6588	0.2305
period2	β_{10}	-0.5116	0.153

- Recall that our model is

$$d \sim Poisson\left(n\lambda\left(\mathbf{x} \right) \right)$$

$$\lambda\left(\mathbf{x} \right) = \exp\left\{ \begin{array}{l} \beta_0 + \beta_1 arsenic2 + \beta_2 arsenic3 + \beta_3 arsenic4 \\ \quad + \beta_4 age2 + \beta_5 age3 + \beta_6 age4 \\ \quad + \beta_7 year2 + \beta_8 year3 + \beta_9 year4 \\ \quad + \beta_{10} period2 \end{array} \right\}.$$

- The coefficients are the estimated effect of the predictors on the log of the respiratory cancer death rate.
- The relative risk of death for the highest, relative to the lowest, arsenic exposures is:

$$RR = \frac{\hat{\lambda}(arsenic4 = 1)}{\hat{\lambda}(arsenic4 = 0)} = \exp\left\{\hat{\beta}_3(1-0)\right\}.$$
$$= \exp\{0.9218\} = 2.51$$

  Thus, the rate of death in the highest exposure category is 2.51 times the rate in the lowest exposure category, after controlling for the effects of age, calendar year, and hiring period.
- The 95% Wald confidence interval is

$$\exp\left\{\beta_3 \pm 1.96\,\mathrm{se}(\beta_3)\right\}$$
$$\exp\left\{0.9218 \pm 1.96(0.181)\right\}.$$
$$(1.76, 3.58)$$

- The Wald test statistics is

$$X^2 = \left(\frac{0.9218}{0.181}\right)^2 = 25.9 \sim \chi_1^2$$

  which gives a p-value of $p = \Pr\left[\chi_1^2 \geq 25.9\right] < 0.0001$.

  Thus, the risk for high, relative to low, arsenic exposure is significant, after controlling for the other covariates in the model.

4.4 Model Fit

4.4.1 Goodness-of-Fit Statistics

Two commonly used measures of model fit are the Pearson Chi-Square statistic

$$X^2 = \sum_{i=1}^{N} \frac{\left(d_i - \hat{d}_i\right)^2}{\hat{d}_i}$$

and the Deviance statistic

$$X^2 = 2\sum_{i=1}^{N} d_i \ln\frac{d_i}{\hat{d}_i} + \left(\hat{d}_i - d_i\right)$$

where $i = 1,\ldots,N$ indexes the observations in the dataset, and $\hat{d}$ is number of deaths predicted from the model. If the model fit is adequate, these statistics have an approximate chi-square distribution with degrees of freedom equal to *N* minus the number of estimated parameters in the model.

- Deviance and Pearson Chi-Square goodness-of-fit statistics are provided in the SAS PROC GENMOD output.
- In the Smelter example, the Deviance and Pearson Chi-square goodness-of-fit results are

GOF	df	Value	Value/df
Pearson Chi-Square	103	105.37	1.02
Deviance	103	105.74	1.03

for which the respective p-values are

$$p = \Pr\left[\chi^2_{103} \geq 105.37\right] = 0.4167$$
$$p = \Pr\left[\chi^2_{103} \geq 105.74\right] = 0.4069 .$$

Therefore, these statistics do not provide evidence of a lack of fit to the data.

- Evidence of a lack-of-fit may indicate that (1) systematic effects are not accounted for or (2) the Poisson assumption is not appropriate.
- The expected value of a chi-square random variable is equal to its degrees of freedom. Thus, the ratio of the goodness-of-fit statistic to its degrees of freedom provides a measure of the observed to expected variability in the residuals. Ratios that are larger than unity are suggestive of a lack of fit; i.e. that there is more residual variability than would be expected under the specified Poisson regression model.

4.4.2 R-Square

The method of Nagelkerke (Biometrika, 1991) can be used to compute an R^2 statistic as

$$R^2 = 1 - \exp\left\{-\frac{2}{N}\left(\ln L\left(\hat{\boldsymbol{\beta}}\right) - \ln L\left(\mathbf{0}\right)\right)\right\}$$

where $\ln L\left(\hat{\boldsymbol{\beta}}\right)$ and $\ln L\left(\mathbf{0}\right)$ denote the log-likelihoods for the Poisson regression models with and without the covariates, respectively. In our Smelter example the resulting R^2 is

$\ln L\left(\hat{\boldsymbol{\beta}}\right)$	$\ln L\left(\mathbf{0}\right)$	N	R^2
122.60	-12.54	114	90.7%

4.4.3 Overdispersion

Poisson regression is particularly susceptible to lack-of-fit problems. Recall that, in Poisson regression, we treat the response variable d as a Poisson random variable with mean equal to $n\lambda\left(\mathbf{x}\right)$. Moreover, the Poisson distribution is such that the mean and variance are equal. Therefore, in our example

$$E\left[d\right] = Var\left[d\right] = n\lambda\left(\mathbf{x}\right)$$
$$= n\exp\left\{\begin{array}{l}\beta_0 + \beta_1 arsenic2 + \beta_2 arsenic3 + \beta_3 arsenic4 \\ \quad + \beta_4 age2 + \beta_5 age3 + \beta_6 age4 \\ \quad + \beta_7 year2 + \beta_8 year4 + \beta_9 year4 \\ \quad + \beta_{10} period2\end{array}\right\}.$$

Perhaps there are omitted covariates or interaction terms that are important predictors of death.

- Omission of important covariates, from the Poisson model could lead us to underestimate the mean *and* the variance.
- Underestimation of the variance leads to underestimation of the standard errors which, in turn, gives test statistics that are more significant than warranted.
- **Overdispersion** refers to the situation where there is more variability in the data than is accounted for by the model.

Pearson Scaled Standard Errors

One way to correct for overdispersion is to multiply the standard error estimates from the Poisson analysis by the square root of the Pearson Chi-Square statistic divided by its degrees of freedom. For instance, the Pearson scale factor of $\sqrt{1.023} = 1.011$ would be used in the Smelter example to correct the standard errors.

Variable	Estimate	SE	Pearson Scaled SE*
Intercept	-8.067	0.2857	$0.2857 \times 1.011 = 0.2889$
arsenic2	0.7982	0.1582	$0.1582 \times 1.011 = 0.1600$
arsenic3	0.5734	0.2062	$0.2062 \times 1.011 = 0.2086$
arsenic4	0.9218	0.181	$0.181 \times 1.011 = 0.1831$
age2	1.387	0.2468	$0.2468 \times 1.011 = 0.2496$
age3	2.1926	0.2445	$0.2445 \times 1.011 = 0.2473$
age4	2.3447	0.2702	$0.2702 \times 1.011 = 0.2733$
year2	0.5336	0.2151	$0.2151 \times 1.011 = 0.2176$
year3	0.687	0.2143	$0.2143 \times 1.011 = 0.2168$

Variable	Estimate	SE	Pearson Scaled SE*
year4	0.6588	0.2305	$0.2305 \times 1.013 = 0.2332$
period2	-0.5116	0.153	$0.153 \times 1.013 = 0.1547$

* The scaled standard errors may be obtained in SAS.

SAS Poisson Regression (Pearson Scaled SE)

```
proc genmod data=smeltermod;
   model respiratory = arsenic2 arsenic3 arsenic4 age2 age3 age4
                       year2 year3 year4 period2
                       / dist=poisson offset=lpyears pscale;
run;
```

Details

- The option **pscale** requests the Pearson scaled standard errors.

```
The GENMOD Procedure

                      Analysis Of Parameter Estimates

                           Standard   Wald 95% Confidence      Chi-
Parameter    DF   Estimate     Error          Limits          Square   Pr > ChiSq

Intercept     1    -8.0670    0.2889    -8.6333    -7.5007    779.47       <.0001
arsenic2      1     0.7982    0.1600     0.4845     1.1118     24.87       <.0001
arsenic3      1     0.5734    0.2086     0.1646     0.9822      7.56       0.0060
arsenic4      1     0.9218    0.1831     0.5629     1.2807     25.35       <.0001
age2          1     1.3870    0.2496     0.8979     1.8761     30.89       <.0001
age3          1     2.1926    0.2473     1.7078     2.6773     78.59       <.0001
age4          1     2.3447    0.2733     1.8090     2.8804     73.59       <.0001
year2         1     0.5336    0.2176     0.1072     0.9601      6.02       0.0142
year3         1     0.6870    0.2168     0.2621     1.1119     10.04       0.0015
year4         1     0.6588    0.2332     0.2018     1.1158      7.98       0.0047
period2       1    -0.5116    0.1547    -0.8148    -0.2084     10.94       0.0009
Scale         0     1.0114    0.0000     1.0114     1.0114

NOTE: The scale parameter was estimated by the square root of Pearson's
      Chi-Square/DOF.
```

4.5 External Standardization

Poisson regression may also be used to estimate and compare standardized mortality ratios (*SMR*s).

4.5.1 Notation

Let us start with a slight redefinition of some notation that was used in past sections.

$j = 1,\ldots,J$	Indexes the strata defined by the variables for which standard rates are available
$k = 1,\ldots,K$	Indexes the strata defined by the variables for which standard rates are not available
$\lambda_j^{(s)}$	Standard population rate in the j^{th} stratum
$O_k = \sum_j d_{jk}$	Total deaths in the k^{th} stratum
$E_k^{(s)} = \sum_j n_{jk}\lambda_j^{(s)}$	Total deaths expected in the k^{th} stratum if the standard rates held true.
$x_{1k},\ldots,x_{pk}$	Regression covariates specific to the k^{th} stratum

Recall that the true *SMR* for the k^{th} stratum is defined as

$$SMR_k = \frac{\sum_j n_{jk}\lambda_{jk}}{\sum_j n_{jk}\lambda_j^{(s)}} = \frac{\sum_j n_{jk}\lambda_{jk}}{E_k^{(s)}}$$

and the corresponding estimate of the *SMR* is

$$\widehat{SMR}_k = \frac{\sum_j d_{jk}}{\sum_j n_{jk}\lambda_j^{(s)}} = \frac{O_k}{E_k^{(s)}}.$$

Smelter Example

Suppose that we are interested in estimating *SMR*s for this cohort study using the U.S. population of white males as the standard population (Table 1).

Table 1. Respiratory death rates* for white U.S. males

Age	Calendar Year			
	1938-49	1950-59	1960-69	1970-77
40-49	0.14817	0.21896	0.28674	0.37391
50-59	0.47412	0.80277	1.05824	1.25469
60-69	0.73136	1.55946	2.33029	2.90461
70-79	0.73207	1.63585	2.85724	4.22945

* Rates are per 1000 person-years

Note that

- Our previous regression analysis of this cohort study included categorical variables for age (4 levels), calendar year (4 levels), hiring period (2 levels), and arsenic exposure (4 levels).
- If we standardize the rates from the cohort by those in the U.S. population, the indexing defined above is as follows: $j = 1,\ldots,16$ for the age-year strata, and

$k = 1,\ldots,8$ for the hiring period-arsenic exposure strata.

- For each of the k stratum we will compute the total O_k and expected $E_k^{(s)}$ number of deaths.

For example, consider the follow-up data in Table 2 for the cohort of smelter workers in the first hiring period and lowest exposure category.

Table 2. Deaths and person-years for the cohort of Montana Smelter Workers in the first hiring period and lowest arsenic exposure category

Age		Calendar Year			
		1938-49	1950-59	1960-69	1970-77
40-49	d	2	0	0	0
	n	3075.27	936.75	0.00	0.00
50-59	d	2	3	3	0
	n	2849.76	2195.59	747.77	0.00
60-69	d	2	7	10	1
	n	2085.43	1675.91	1501.73	440.21
70-79	d	3	6	6	6
	n	833.61	973.32	1027.12	674.44

The observed and expected number of deaths for this hiring period-arsenic exposure stratum are

$$O_1 = 2 + \ldots + 6 = 51$$

$$E_1^{(s)} = 3075.27\left(\frac{0.14817}{1000}\right) + \ldots + 674.44\left(\frac{4.22945}{1000}\right)$$

$$= 21.47$$

Repeating this calculation for all eight strata yields the following results:

Table 3. Observed and expected number of deaths in the Smelter Study

k	period	arsenic	O_k	$E_k^{(s)}$
1	1	1	51	21.47
2	1	2	17	2.95
3	1	3	13	2.76
4	1	4	34	4.44
5	2	1	100	74.12
6	2	2	38	13.84
7	2	3	15	6.83
8	2	4	8	3.66

4.5.2 Model Specification

As before the observed number of deaths is assumed to follow a Poisson distribution

$$d_{jk} \sim Poisson\left(n_{jk}\lambda_{jk}\right)$$

where n_{jk} and λ_{jk} are the number of person years and the true death rate, respectively, in the *j-k* stratum. It can be shown that the sum of Poisson random variables $O_k = \sum_j d_{jk}$ also has a Poison distribution,

$$\begin{aligned} O_k &\sim Poisson\left(\sum_j n_{jk}\lambda_{jk}\right) \\ &\sim Poisson\left(E_k^{(s)} SMR_k\right) \end{aligned}.$$

In particular the outcome variable in the Poisson regression models for the *SMR* will be O_k. The mean of the Poisson distribution will be linked to the linear predictor as follows

$$\ln\left[E_k^{(s)} SMR_k\right] = \ln\left[E_k^{(s)}\right] + \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk}.$$

Note that $\ln\left[E_k^{(s)}\right]$ is included as an offset term in this model. The reason for choosing this offset is that it results in a mean function that can be rewritten as

$$\begin{aligned} \ln\left[E_k^{(s)} SMR_k\right] - \ln\left[E_k^{(s)}\right] &= \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} \\ \ln\left[\frac{E_k^{(s)} SMR_k}{E_k^{(s)}}\right] &= \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} \\ \ln\left[SMR_k\right] &= \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} \end{aligned}$$

or, more specifically,

$$SMR_k = \exp\left\{\beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk}\right\}.$$

Therefore, this formulation of the Poisson regression model allows for estimation and comparison of the *SMR*s across the *k* strata.

Smelter Example

Suppose that we are interested in comparing hiring periods and exposure categories after adjusting the rates in the cohort by the age-year rates for white U.S. males. This can be accomplished with the Poisson model

$$O_k \sim Poisson\left(E_k^{(s)} SMR_k\right)$$

$$SMR_k = \exp\left\{\begin{array}{l}\beta_0 + \beta_1 arsenic2 + \beta_2 arsenic3 + \beta_3 arsenic4 \\ \quad + \beta_4 period2\end{array}\right\}.$$

where the expected deaths $E_k^{(s)}$ are calculated using the age-year rates from the standard population. Table 3 summarizes the data that are needed for this analysis. Note that there are no effects for age or calendar year in the model because they have already been adjusted for in the calculation of expected deaths.

SAS Poisson Regression (SMR Analysis)

```
data smeltersmr;
   input period arsenic O E;
   period2 = (period = 2);
   arsenic2 = (arsenic = 2);
   arsenic3 = (arsenic = 3);
   arsenic4 = (arsenic = 4);
   lnE = log(E);
   cards;
   1 1  51 21.47
   1 2  17  2.95
   1 3  13  2.76
   1 4  34  4.44
   2 1 100 74.12
   2 2  38 13.84
   2 3  15  6.83
   2 4   8  3.66
;

proc genmod data=smeltersmr;
   model O = arsenic2 arsenic3 arsenic4 period2
             / dist=poisson offset=lnE;
run;
```

The GENMOD Procedure

Model Information

Data Set	WORK.SMELTERSMR
Distribution	Poisson
Link Function	Log
Dependent Variable	O
Offset Variable	lnE
Observations Used	8

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	2.8747	0.9582
Scaled Deviance	3	2.8747	0.9582
Pearson Chi-Square	3	2.6964	0.8988
Scaled Pearson X2	3	2.6964	0.8988
Log Likelihood		780.4926	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	0.9570	0.1139	0.7337	1.1803	70.55	<.0001
arsenic2	1	0.7711	0.1577	0.4619	1.0802	23.90	<.0001
arsenic3	1	0.5628	0.2060	0.1590	0.9665	7.46	0.0063
arsenic4	1	0.9491	0.1797	0.5969	1.3014	27.89	<.0001
period2	1	-0.7078	0.1266	-0.9560	-0.4596	31.24	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

4.5.3 Estimation

The parameter estimates from the Poisson regression analysis of the Smelter Study are given in the table below.

Variable	Parameter	Estimate	SE
Intercept	β_0	0.957	0.1139
arsenic2	β_1	0.7711	0.1577
arsenic3	β_2	0.5628	0.206
arsenic4	β_3	0.9491	0.1797
period2	β_4	-0.7078	0.1266

The methods for relative risk estimation and inference are completely analogous to those presented in Section 4.2.

- The relative risk of death for the highest, relative to the lowest, arsenic exposures is:

$$\widehat{RR} = \frac{\widehat{SMR}\left(arsenic4 = 1\right)}{\widehat{SMR}\left(arsenic4 = 0\right)} = \exp\left\{\hat{\beta}_3\left(1-0\right)\right\}.$$

$$= \exp\left\{0.9491\right\} = 2.58$$

  Thus, the rate of death in the highest exposure category is 2.58 times the rate in the lowest exposure category, after controlling for the effects of hiring period and adjusting for the age and calendar year death rates in the U.S. population.

- The 95% Wald confidence interval is

$$\exp\left\{\beta_3 \pm 1.96\,\mathrm{se}\left(\beta_3\right)\right\}$$

$$\exp\left\{0.9491 \pm 1.96\left(0.1797\right)\right\}.$$

$$\left(1.82, 3.67\right)$$

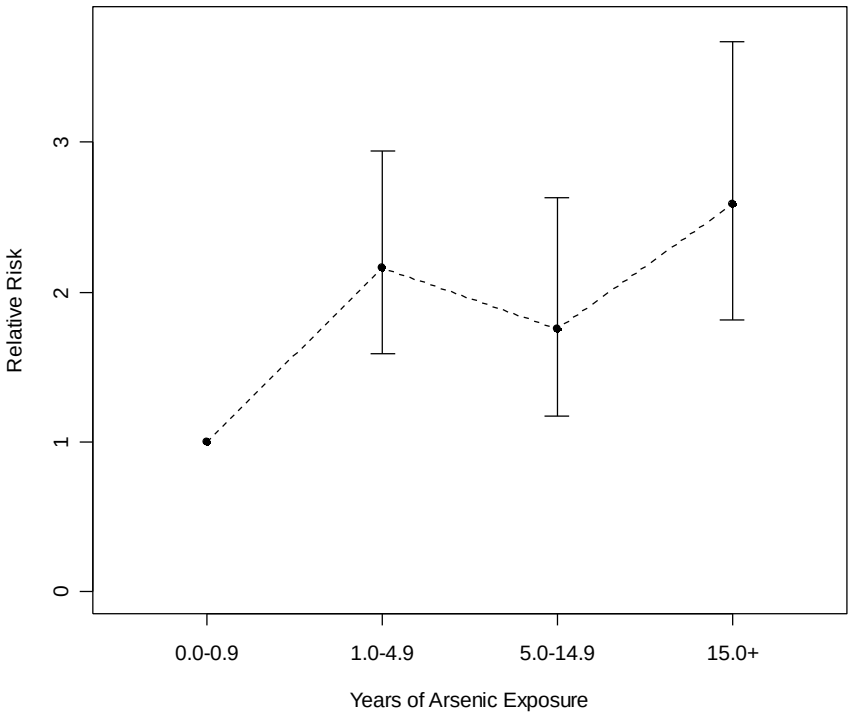
- The Wald test statistics is

$$X^2 = \left(\frac{0.9491}{0.1797}\right)^2 = 27.90 \sim \chi_1^2$$

  which gives a p-value of $p = \Pr\left[\chi_1^2 \geq 27.90\right] < 0.0001$.

  Thus, the risk for high, relative to low, arsenic exposure is significant, after controlling hiring period, age, and calendar year.

- The relative risk estimates for the top three exposure categories, relative to the first, are summarized in the following plot.



4.5.4 Notes

The general steps for performing a Poisson regression analysis of the *SMR*s are

1. Construct appropriate categories for the variables to be included in the analysis.
2. Calculate the number of deaths and person-years within each of the strata defined by the categorical variables.
3. Apply rates from a standard population to compute the expected number of deaths $E_k^{(s)}$, and record the associated observed number of deaths O_k.
4. Fit a Poisson model to the observed number of deaths O_k. Include $\ln\left[E_k^{(s)}\right]$ as an offset term and the other covariates that were not controlled for in the standardization as predictors in the model.

Cohort Data Analysis (171:242/243)
Section 5: Sample Size Estimation

Brian J. Smith, Ph.D.

October 8, 2007

Table of Contents

5.1 Introduction

Suppose that the SMR is to be used to compare the observed number of deaths in the cohort to the number that would be expected if the standard population rates held. Recall that the form of the SMR is

$$SMR = \frac{D}{E^{(s)}} = \frac{D}{n\lambda^{(s)}}.$$

Goal: Perform sample size and power calculations in order to determine the number of subjects needed for a cohort study. Note that

- The calculations involve the number of person-years, which is a function of both the number of subjects and the follow-up periods.
- Exact and approximate methods will be discussed in this section when the hypotheses of interest are

$$H_0 : SMR = 1$$
$$H_A : SMR > 1.$$

- In practice the *SMR* is sometimes referred to as a relative risk.

5.2 Exact Method

5.2.1 Exact Poisson Test

The exact p-value for testing the significance of the *SMR* is

$$p = \Pr\left[Y \geq D\right]$$

where $Y \sim Poisson\left(E^{(s)}\right)$.

Example

Suppose that $E^{(s)} = 9$ and $D = 16$, for which $SMR = D/E^{(s)} = 16/9 = 1.78$. Under the null hypothesis, the observed number of deaths is distributed as $Y \sim Poisson(9)$. Thus, the one-sided p-value is

$$P[Y \geq 16] = 0.0220.$$

5.2.2 Power Calculation

Power estimates for various expected number of deaths and *SMR*s are given in Table 1 and Table 2.

Table 1. Probability of obtaining a result significant at the 5% level (one-sided) for varying values of the expected value *E* assuming no excess risk, and of the true *SMR*

$E^{(s)}$	True *SMR*									
	1.0	1.5	2.0	3.0	4.0	5.0	7.5	10.0	15.0	20.0
1.0	1.90	7	14	35	57	74	94	99	100	100
2.0	1.66	8	21	55	81	93	100			
3.0	3.35	17	39	79	95	99				
4.0	2.14	15	41	84	98	100				
5.0	3.18	22	54	93	100					
6.0	4.26	29	65	97	100					
7.0	2.70	26	64	98	100					
8.0	3.42	32	73	99	100					
9.0	4.15	38	79	100						
10.0	4.87	43	84	100						
11.0	3.22	39	83	100						

12.0	3.74	44	87	100						
13.0	4.27	48	90	100						
14.0	4.79	53	93	100						
15.0	3.27	49	92	100						
20.0	3.43	60	97	100						

Table 2. {continued}

$E^{(s)}$	True *SMR*									
	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
20.0	3.43	9	18	30	45	60	73	83	90	94
25.0	4.98	13	26	42	59	74	85	92	96	98
30.0	4.63	13	27	46	64	79	89	95	98	99
35.0	4.25	13	29	49	69	83	92	97	99	100
40.0	3.87	13	30	52	72	86	94	98	99	100
45.0	4.73	16	36	60	79	91	97	99	100	
50.0	4.24	16	37	61	81	93	98	99	100	
60.0	4.42	18	42	69	88	96	99	100		
70.0	4.48	19	47	75	92	98	100			
80.0	4.46	21	5	80	94	99	100			
90.0	4.39	22	55	83	96	99	100			
100.0	4.28	23	58	86	97	100				

Example

Suppose that we are interested in the power to detect a true *SMR* of 2.0, for a one-sided test performed at the 5% level of significance. If the expected number of deaths is 9, then the estimated power is 79%.

5.2.3 Sample Size

Suppose that we want to have 80% power to detect an *SMR* of 1.50 using a one-sided test at the 5% level of significance.

According to Table 2, $E^{(s)} = 30.0$ expected deaths are needed. Thus, the required number of person-years is

$$E^{(s)} = n\lambda = 30 \Rightarrow n = 30/\lambda^{(s)}.$$

which is a function of the death rate in the standard population. If, say, $\lambda^{(s)} = 0.0001$, then $n = 30/0.0001 = 300{,}000$ person-years of follow-up are needed. This could be obtained in various ways; e.g.

- 10 years follow-up on 30,000 subjects
- 20 years follow-up on 15,000 subjects
- 30 years follow-up on 10,000 subjects

5.3 Approximate Method

In Section 2 several approximate tests were discussed for testing the *SMR*. We will use the one based on the square-root (variance stabilizing) transformation.

5.3.1 Approximate Test

The test statistic is

$$2\left(\sqrt{D} - \sqrt{E^{(s)}}\right) \sim N(0,1).$$

for which the one sided p-value is

$$p = \Pr\left[Z \geq 2\left(\sqrt{D} - \sqrt{E^{(s)}}\right)\right].$$

Alternatively, we could use a critical-value approach to conclude that the $SMR > 1$ if

$$2\left(\sqrt{D} - \sqrt{E^{(s)}}\right) \geq Z_{1-\alpha}$$

or, equivalently,

$$\sqrt{D} \geq \sqrt{E^{(s)}} + Z_{1-\alpha}/2.$$

Note that the probability of rejecting the null hypothesis if there is no difference between the observed and expected number of deaths is

$$\Pr\left[\sqrt{D} \geq \sqrt{E^{(s)}} + Z_{1-\alpha}/2 \middle| SMR = 1\right] = \alpha.$$

5.3.2 Power Calculation

Power $(1-\beta)$ is defined as the probability of rejecting the null hypothesis for a given value (R) of the *SMR*; i.e.

$$\Pr\left[\sqrt{D} \geq \sqrt{E^{(s)}} + Z_{1-\alpha}/2 \middle| SMR = R\right] = 1-\beta.$$

If the $SMR = R$ then $E[D] = RE^{(s)}$. Therefore, the power is computed with standard normal probabilities according to the following formula

$$
\begin{aligned}
1-\beta &= \Pr\left[\sqrt{D}-\sqrt{RE^{(s)}} \geq \sqrt{E^{(s)}}-\sqrt{RE^{(s)}}+Z_{1-\alpha}/2\right] \\
&= \Pr\left[2\left(\sqrt{D}-\sqrt{RE^{(s)}}\right) \geq 2\left(\sqrt{E^{(s)}}-\sqrt{RE^{(s)}}\right)+Z_{1-\alpha}\right] \\
&= \Pr\left[Z \geq 2\left(\sqrt{E^{(s)}}-\sqrt{RE^{(s)}}\right)+Z_{1-\alpha}\right]
\end{aligned}
$$

Note that

$$
2\left(\sqrt{E^{(s)}}-\sqrt{RE^{(s)}}\right)+Z_{1-\alpha} = Z_{\beta}.
$$

Example

Suppose that we are interested in the power to detect a true *SMR* of 2.0, for a one-sided test performed at the 5% level of significance. If the expected number of deaths is 9, then the estimated power is

$$
\begin{aligned}
&\Pr\left[Z \geq 2\left(\sqrt{E^{(s)}}-\sqrt{RE^{(s)}}\right)+Z_{1-\alpha}\right] \\
&= \Pr\left[Z \geq 2\left(\sqrt{9}-\sqrt{2.0(9)}\right)+1.645\right]. \\
&= \Pr\left[Z \geq -0.8403\right] = 80\%
\end{aligned}
$$

5.3.3 Sample Size

For a given significance level (α) and power $(1-\beta)$, we can calculate the required number of expected deaths as

$$2\left(\sqrt{E^{(s)}} - \sqrt{RE^{(s)}}\right) + Z_{1-\alpha} = Z_{\beta}$$

$$\sqrt{E^{(s)}}\left(\sqrt{R} - 1\right) = \frac{Z_{1-\alpha} + Z_{1-\beta}}{2} \quad .$$

$$E^{(s)} = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{4\left(\sqrt{R} - 1\right)^2}$$

Example

Suppose that we want to have 80% power to detect an *SMR* of 1.50 using a one-sided test at the 5% level of significance. The required number of expected deaths is

$$E^{(s)} = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{4\left(\sqrt{R} - 1\right)^2}$$

$$= \frac{\left(1.645 + 0.8403\right)^2}{4\left(\sqrt{1.50} - 1\right)^2} = 30.57 \quad .$$

The number of subjects and associated lengths of follow-up can be computed as before.