NMST532 Design and Analysis of Medical Studies

Course notes

Michal Kulich

Last modified on May 26, 2021.

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University

These course notes contain an overview of the contents of the course "NMST532 Design and Analysis of Medical Studies", which is a part of the curriculum of the Master's program "Probability, Mathematical Statistics and Econometrics".

This material undergoes continuing development. The author will appreciate notifications by the reader of potential errors, typos or misprints.

České překlady odborných termínů uvedené v tomto textu jsou pouhé návrhy a pokusy autora o jejich přijatelné převedení do češtiny. Většina odborných termínů v oblasti, kterou se zabývá tento text, nemá všeobecně přijaté a používané české ekvivalenty. Jejich překlady kolísají a často jsou suplovány nepřeloženými anglickými výrazy.

Michal Kulich
kulich@karlin.mff.cuni.cz

In Karlín on May 26, 2021

Contents

Foreword

This course offers a brief introduction to most commonly used statistical methods in medical research.

The first part of the course focuses on statistical methods in epidemiology. It starts with a summary of basic epidemiological terminology and an overview of descriptive methods for estimating disease incidence. The most important epidemiological study designs, case-control and cohort studies, are introduced and discussed. Classical methods for the analysis of these designs are briefly summarized. However, the main focus is on approaches based on regression models, in particular logistic regression and loglinear models.

In the second part, we give a short summary of diagnostic methods. This part provides basic terminology for statistical properties of diagnostic methods and explains a few descriptive approaches to their estimation.

The final part is devoted to randomized clinical trials. A short overview of the drug development process is provided. Then focus shifts to randomized Phase III trials. It is explained how these trials are planned and what is the role of the statistician at the planning stage (choice of the primary outcome, selection of study population, enrollment considerations, randomization methods, statistical analysis plan and sample size calculation). A few variants of the basic study design are discussed. This part is concluded by a universal approach to power and sample size calculation and brief summary of group sequential testing methods.

In this course, it is assumed that the student is familiar with a relatively wide range of statistical theory: apart from introductory statistics we rely on sufficient familiarity with linear regression, generalized linear models, and parametric and non-parametric survival analysis methods. These methods are not explained in this course. It is shown how they are used to analyze common medical data problems. When needed, the standard methods are extended to handle more general problems (censoring with left truncation, analysis of aggregated data, analysis of non-standard sampling designs). Thus this course is not self-containing – it relies on relatively wide and thorough training in data analysis methods and practice.

1. Descriptive Epidemiology

1.1. Subject of epidemiology

Epidemiology is an interdisciplinary branch of medicine that describes and investigates the occurrence of a disease in a human population and studies potential risk factors that cause the development of the disease or protective factors that prevent it. The disease of interest need not be infectious, as the name "epidemiology" might suggest, but it may be any kind of disease such as cancer, cardiovascular disease, type I diabetes or dental carries.

Causes of diseases and means to protect oneself against disease have been always a topic of human interest. For thousands of years, people were using various naive methods for protection against disease, without any understanding how and why diseases develop. It is not surprizing that these methods were rarely effective and frequently harmful.

The first written records about successful use of rational logical thinking about disease occurrence leading to correct conclusions are only a few hundred years old. In 1670, an Alsatian physician Louis Thuillier correctly determined that the cause of the terrible condition called *St. Anthony's fire*, now known as *ergotism*, was *Claviceps purpurea*[*] fungus that infects rye and other cereals. *St. Anthony's fire* was common in Alsatia at Thuillier's time and he made a correct conclusion about its cause by taking records of its occurrence in different conditions over many years and carefully considering all factors that might explain where and when the disease was most frequent.

Another commonly cited example is John Snow's investigation of cholera outbreak in London in 1854. He made a map of the neighborhood where cholera cases appeared and marked on it each individual death. He figured out that the victims were clustered around a water pump on Broad Street and concluded that they got sick of the water drawn from that pump. He demanded that the pump be closed; when it happened, the outbreak receded. Later, it was found out that the Broad Street pump was built next to an old cesspit from which infected feces contaminated the drinking water.

Both Thuillier and Snow did not understand the real mechanism causing the disease (poisonous alkaloids of *Claviceps purpurea* in Thuillier's case, bacterium *Vibrio cholerae* in Snow's case). However, both were able to use observations of occurrence of disease to get very close to the real cause and to draw correct conclusions. Today, there are still many diseases for which the real causes are not known or are too complex to understand. The only way to discover which environmental/genetic/lifestyle factors affect the disease is to conduct empirical

[*] Česky *paličkovice nachová neboli námel*

studies where observations of occurrence of disease are analyzed by statistical methods. Thus, statistics is the core part of epidemiology.

1.2. Disease prevalence and incidence

Let us think about a population of individuals and a disease that they might get. Some of those individuals might be immune to the disease temporarily or permanently. The individuals that are not immune are called *susceptible*. Any of the susceptible individuals might get the disease at some moment and become *a case*[*]. They could stay in the diseased category for some time. In the end some would die of the disease, and some would recover. Those who recover may enter the susceptible population again or become immune for a period of time or permanently.

Let us consider this framework to define *prevalence*[†] of the disease. Prevalence is the proportion of the whole population that has the disease, or more precisely, the probability of having the disease. The proportion is then an estimate of the probability. We can talk about *point prevalence*, where we look at the disease at a particular fixed time and count all cases who had the disease at that time, or *prevalence over period*, where we count all subjects who had the disease at any time during the period. We can condition the prevalence on being susceptible (leaving the unsusceptible out of the population) or we can calculate prevalence regardless of susceptibility.

Statistically, prevalence analyses involve just a simple binary outcome (1 = disease/case and 0 = no disease/control). Classical statistical methods for binary outcomes are sufficient to conduct analyses of prevalence if data on disease status are available.

A more interesting measure of disease occurrence is *incidence*[‡]. Incidence is the rate of occurrence of *new* cases of the disease in the population of interest. Thus, for incidence we only take into account the first time the subject gets the disease (over lifetime or within a defined period of interest). It is not important what happens after the subject gets the disease for the first time. Incidence is a better measure to investigate causes of the disease than prevalence. One of the disadvantages of prevalence is that it is affected by duration of the disease. Hence changes in duration (caused, e.g., by better treatments) affect prevalence even if the true causes of the disease remain the same and act in the same way. Differences and changes in prevalence are generally hard to interpret. Incidence does not share this drawback. For these reasons we focus primarily on statistical methods for incidence analysis.

Mortality[§] is a special case of incidence when disease occurrence is replaced by the occurrence of death. One can consider either overall mortality (death from any cause) or cause-specific mortality (death from a particular disease). Mortality analyses follow the same framework as incidence analyses.

[*] Česky *případ* [†] Česky *prevalence* [‡] Česky *incidence* [§] Česky *úmrtnost*

1.3. Incidence as hazard for left-truncated survival data

For investigating incidence, we can simplify our framework. Assign to each subject three latent random variables (E, T, C) that fully describe the subject's history of getting (or not getting) the disease. Their meaning is:

- E is the entry time, i.e., the time when the subject becomes susceptible and enters the observation,
- T is the time when the subjects gets the disease for the first time
- C is the exit time, i.e., the time when the subject leaves observation because of death or other reasons.

All of these variables are measured on the same time scale. Time 0 usually means birth (then all the variables are ages at particular events) or it can be some other well defined moment, specific calendar time etc. The three latent variables determine what is observed for the subject.

- If $E \leq T \leq C$, the subject is a case with observed occurrence of the disease at the time T.
- If $E \leq C < T$, the subject is a control with no observed occurrence of the disease.
- If $C < E$ or $T < E$, the subject is not observed at all.

The continuous non-negative random variable T is the time of the first occurrence of the disease and the incidence is in fact the hazard function of this variable. So we can define incidence more formally as

$$\lambda(t) = \lim_{h \searrow 0} \frac{1}{h} \mathrm{P}\big[t \leq T < t + h \,\big|\, T \geq t\big].$$

for any $t \geq 0$. Incidence can be investigated by the methods for censored failure time data (survival analysis) with T being the failure time and C being the censoring time. However, standard methods for censored data assume that each subject is observed from time 0.

Thus, we need to extend the survival analysis methods to allow random entry times E as well. The fact that the subject is unobserved if $\min(T, C) < E$ is called *left truncation* and we have a combination of right-censored and left-truncated data.

Introduce an event indicator

$$\delta = \begin{cases} 1 & \text{if } E \leq T \leq C, \\ 0 & \text{if } E \leq C < T. \end{cases}$$

Thus, a subject is a case if $\delta = 1$ and a control if $\delta = 0$. Extend the follow-up indicator (at-risk process) to take entry times into account as well:

$$Y(t) = \mathbb{1}(E \leq t, t \leq \min(C, T)).$$

The follow-up indicator starts at 0 at the time 0, jumps to 1 at the entry time and stays at 1 until disease or censoring occurs. Then it drops back to zero.

Finally, the process counting the number of observed occurrences of the disease is

$$N(t) = \mathbb{1}(T \leq t, E \leq T \leq C).$$

We need to extend the independent censoring condition to handle random entry times. We do it by assuming that the usual form of the compensator applies to the current situation. In particular,

$$M(t) = N(t) - \int_0^t Y(s)\lambda(s)\,ds \quad \text{is an } \mathcal{F}_t\text{-martingale,} \tag{1.1}$$

where $\mathcal{F}_t = \sigma\{N(s), Y(s), 0 \leq s \leq t\}$. A sufficient condition to guarantee the validity of (1.1) is that the pair (E, C) is independent of T.

Under this condition, the counting process theory used in right-censored data also holds for left-truncated data. The definitions and properties of nonparametric estimators and tests (Nelson-Aalen, Kaplan-Meier, logrank statistics, Cox model) all work, the only difference being the modified definition of the at-risk process $Y(t)$.

Suppose the population consists of n independent subjects that are observed for occurrence of disease. Each subject has associated a triplet of latent variables (E_i, T_i, C_i), $i = 1, \ldots, n$. The data observed on the i-th subject can be summarized by a pair of processes $N_i(t) = \mathbb{1}(T_i \leq t, E_i \leq T_i \leq C_i)$ and $Y_i(t) = \mathbb{1}(E_i \leq t, t \leq \min(C_i, T_i))$. Let $U_i(t) = \int_0^t Y_i(s)\,ds$ be the duration of follow-up of the i-th subject over the interval $(0, t)$.

Let us introduce notation we will use in the following sections. Let

$$\overline{N}(t) = \sum_{i=1}^{n} N_i(t)$$

be the number of cases observed by the time t,

$$\overline{Y}(t) = \sum_{i=1}^{n} Y_i(t)$$

the number of subjects that are followed at the time t, and

$$\overline{U}(t) = \sum_{i=1}^{n} U_i(t)$$

the total follow-up time for the whole population over the interval $(0, t)$.

1.4. Empirical incidence estimates

If individual data are available, the cumulative incidence $\Lambda(t) = \int_o^t \lambda(s)\,ds$ could be estimated nonparametrically by the Nelson-Aalen estimator. However, we would like to have an estimator of the incidence rate itself. One could apply kernel smoothing methods to obtain a

smooth estimate of incidence from the Nelson-Aalen estimator but we will not take that approach. Instead, we will derive a simple histogram-like empirical estimate of incidence based on a piecewise exponential model.

MLE with exponential distribution

Let us start with exponentially distributed time to disease occurrence. Assume $T_1, \ldots, T_n$ are independent, with distribution $\text{Exp}(\lambda)$ and constant incidence λ. Suppose (E_i, C_i) is independent of T_i. Define $T_i^* = T_i - E_i, C_i^* = C_i - E_i$. In all observed subjects, T_i^* and C_i^* are non-negative. Consider the data $U_i = \min(T_i^*, C_i^*)$ and $\delta_i = \mathbb{1}(T_i^* \leq C_i^*) = \mathbb{1}(E_i \leq T_i \leq C_i)$. Then $T_i^* \sim \text{Exp}(\lambda)$ because $\text{P}\left[T_i > t + s \middle| T_i > s\right] = \text{P}\left[T_i > t\right]$ and E_i (playing the role of s) is independent of T_i. Hence we can estimate λ from (T_i^*, δ_i) by maximum likelihood methods for exponential distribution with random censoring and get

$$\widehat{\lambda} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n U_i}.$$

So, constant incidence can be estimated by the number of observed cases divided by the total follow-up time. With the notation introduced at the end of the previous section,

$$\widehat{\lambda} = \overline{N}(\infty)/\overline{U}(\infty). \tag{1.2}$$

MLE of piecewise constant incidence

Now divide the time axis into M disjoint intervals defined by points $0 \equiv t_0 < t_1 < \cdots < t_{M-1} < t_M < \infty$. The length of the k-th interval is $d_k = t_k - t_{k-1}$. Suppose the incidence $\lambda(t)$ is piecewise constant on those M intervals, that is, $\lambda(t) = \lambda_k$ for $t \in \langle t_{k-1}, t_k)$. This corresponds to a piecewise exponential distribution of T_i.

Estimation of λ_k proceeds as follows: start with $k = 1$, the first interval. Censor all observations at t_1, that is, redefine C_i to $\min(C_i, t_1)$. Then we are dealing with a censored sample from $\text{Exp}(\lambda_1)$ and the MLE is given by equation (1.2) stopped at t_1, that is,

$$\widehat{\lambda}_1 = \overline{N}(t_1)/\overline{U}(t_1).$$

Now consider the second interval. Redefine all entry times to start at t_1 (set them to $\max(E_i, t_1)$) and redefine all censoring times to censor at t_2 (take $\min(C_i, t_2)$). Now we are dealing with an exponential sample with incidence λ_2 and can use equation (1.2) again to estimate λ_2 except that the number of cases and total follow-up time are calculated over the interval $\langle t_1, t_2)$. The same procedure can be repeated for each successive interval.

Define $\overline{N}(t_{k-1}, t_k) = \overline{N}(t_k) - \overline{N}(t_{k-1}) \equiv n_k$ the number of cases observed on the k-th interval and define $\overline{U}(t_{k-1}, t_k) = \sum_{i=1}^n \int_{t_{k-1}}^{t_k} Y_i(s)\, ds \equiv u_k$ the total follow-up time on the k-th interval. It follows that the MLE of λ_k is

$$\widehat{\lambda}_k = \overline{N}(t_{k-1}, t_k)/\overline{U}(t_{k-1}, t_k) = n_k/u_k$$

for $k = 1, \ldots, M$.

We will call these piecewise constant estimates of incidence *empirical incidence estimates*. They can be plotted to get a histogram-like picture of estimated incidence. With enough data and suitable partitioning, the piecewise constant incidence estimate can reasonably approximate a relatively complicated incidence function. It is important to select the partitioning carefully, so that each interval has enough observed cases.

Empirical incidence estimates with aggregated data

Epidemiological data sets rarely include information on individual subjects. More frequently, the data are aggregated over subgroups. Such aggregated data include the time intervals, the numbers of cases observed in the intervals and the numbers of subjects observed at the start of the intervals.

Empirical incidence estimates can be calculated from aggregated data by replacing the exact total follow-up time u_k by its estimate obtained from the duration of the interval d_k and the number of subjects $y_k = \overline{Y}(t_{k-1})$ who are at risk at the start of the interval. Suppose that there are no entries into the population and no censoring during the interval (except at the cutpoints). The we can estimate u_k simply by $d_k y_k$ to get

$$\widehat{\lambda}_k = \frac{n_k}{d_k y_k}$$

or, taking into account the number of cases observed within the interval,

$$\widehat{\lambda}_k = \frac{n_k}{d_k(y_k - n_k/2)},$$

giving each case only half of the follow-up of a control. If the disease is rare, n_k is relatively small compared to y_k and the adjustment by $n_k/2$ is quite negligible. With common diseases, however, the adjustment substantially improves the bias in the incidence estimate.

The estimated incidence depends on the units of time used for d_k. Follow-up is usually measured in person-years, p.y. (1 person-year corresponds to one person followed for one year). Incidence of rare diseases is frequently expressed per 10^5 person years (one hundred thousand) and understood as the number of cases expected in one year in a population of 10^5 subjects.

1.5. Age-specific incidence, age-standardized incidence, cumulative incidence

Age-specific incidence

Age-specific incidence is simply piecewise-constant incidence estimate of the previous section calculated for pre-specified age groups (the time scale is age, age 0 is the birth). The age groups

are usually taken as five or ten years wide. Incidence of most diseases strongly depends on age so it rarely makes sense to report the overall incidence in the population without taking into account its age composition.

Similarly, sex-specific incidence is calculated separately by gender, and calendar-time-specific incidence is calculated separately by predefined intervals of calendar time.

Age-standardized incidence

For a simple comparison of incidences in two sub-populations (e.g., men vs. women), it would be helpful to characterize incidence in the subgroup by a single number instead of an age-dependent way. However, the overall incidence in the subgroup is affected by its age composition. Instead, the epidemiologists combine age-specific incidences into a single number with weights that are the same for all subgroups and represent the age composition of some selected "standard" population. This is called *age-standardized incidence*. Such standardized incidences can be compared even between subgroups with different age compositions.

Weights $w_1, \ldots, w_M$ are chosen for age intervals $\langle t_{k-1}, t_k)$ so that $w_k \geq 0$ and $\sum_{k=1}^{M} w_k = 1$. They describe the relative representation of the age groups in a standard reference population (which should be identified). The reference population can be, for example:

- the general population of the country the study population is taken from;
- some idealized standard population (published, e.g., by the World Health Organization);
- the combined population of the current study (all subgroups together).

The age-standardized incidence is a weighted average of age-specific incidences,

$$\widehat{\lambda}_S = \sum_{k=1}^{M} w_k \widehat{\lambda}_k.$$

Cumulative incidence

The cumulative incidence is integrated incidence rate over a time interval, $\Lambda(t) = \int_0^t \lambda(s)\,ds$. If individual follow-up data is available, cumulative incidence can be estimated by the Nelson-Aalen estimator. Under the piecewise exponential model with constant incidence λ_k on intervals $\langle t_{k-1}, t_k)$, $k = 1, \ldots, M$, the cumulative incidence can be expressed as $\Lambda(t_k) = \sum_{j=1}^{k} d_j \lambda_j$, where $d_j = t_j - t_{j-1}$ is the duration of the j-th interval. An estimator for the cumulative incidence is obtained as a linear combination of empirical incidence estimates,

$$\widehat{\Lambda}(t_k) = \sum_{j=1}^{k} d_j \widehat{\lambda}_j.$$

In the context of rare diseases, cumulative incidence has an interesting interpretation. We know it is related to the survival function S by $S(t) = \mathrm{e}^{-\Lambda(t)}$. When the disease is rare, $\Lambda(t)$ is

small and we can apply a one term Taylor approximation $e^{-x} \approx 1 - x$ to write $S(t) \approx 1 - \Lambda(t)$, or $F(t) \approx \Lambda(t)$, where F is the distribution function of age of disease occurrence. Thus, for rare diseases, the cumulative incidence at the age t approximates the probability of getting the disease by the age t.

Confidence intervals for age-standardized and cumulative incidence

Recall that the empirical incidence estimates are $\widehat{\lambda}_k = n_k/u_k$, where n_k is the number of cases observed on the k-th interval and u_k is the total follow-up time on the k-th interval (person-years of follow-up), either exact (if individual data are available), or approximated as $d_k y_k$, where y_k is the number of subjects at risk at the start of the interval (for aggregated data).

The number of cases n_k is a realization of a counting process with constant intensity λ_k. Conditionally on u_k, it can be shown that the counting process is a Poisson process and hence $n_k \sim \mathrm{Po}(\lambda_k u_k)$. Also, $n_1, \ldots, n_M$ are independent. For $u_k \to \infty$ (increasing total follow-up time), Poisson distribution can be approximated by a normal distribution. More precisely,

$$\frac{n_k - \lambda_k u_k}{\sqrt{\lambda_k u_k}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1) \quad \text{as } u_k \to \infty. \tag{1.3}$$

After a simple manipulation, we get

$$\frac{\widehat{\lambda}_k - \lambda_k}{\sqrt{\lambda_k/u_k}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1) \quad \text{as } u_k \to \infty. \tag{1.4}$$

Take $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_M)^\mathsf{T}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^\mathsf{T}$. Since $\widehat{\lambda}_k$, $k = 1, \ldots, M$, are asymptotically normal and they are independent of each other, we get joint asymptotic normality

$$\Sigma^{-1/2}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{\mathrm{D}} \mathrm{N}_M(\mathbf{0}, \mathbb{I}_M),$$

where

$$\Sigma = \mathrm{diag}\left(\frac{\lambda_k}{u_k}\right).$$

For any non-zero vector of constants $\boldsymbol{c}$, it holds

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\lambda}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\lambda}}{\sqrt{\boldsymbol{c}^\mathsf{T}\Sigma\boldsymbol{c}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1). \tag{1.5}$$

Finally, we replace Σ by its estimator $\widehat{\Sigma} = \mathrm{diag}\,(\widehat{\lambda}_k/u_k) = \mathrm{diag}\,(n_k/u_k^2)$ and calculate

$$\boldsymbol{c}^\mathsf{T}\widehat{\Sigma}\boldsymbol{c} = \sum_{k=1}^{M}\left(\frac{c_k}{u_k}\right)^2 n_k.$$

An asymptotic confidence interval for a linear combination of incidence rates $\boldsymbol{c}^\mathsf{T}\boldsymbol{\lambda}$ is

$$\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\lambda}} \mp u_{1-\alpha/2}\sqrt{\sum_{k=1}^{M}\left(\frac{c_k}{u_k}\right)^2 n_k}.$$

This gives us asymptotic confidence intervals for both age-standardized incidence and cumulative incidence. For age-standardized incidence, we set $c_k = w_k$ and for cumulative incidence, $c_k = d_k$. For cumulative incidence calculated on $(0, t_M)$ with aggregated data, we can plug in $d_k y_k$ for u_k and express the variance as $\boldsymbol{c}^\mathsf{T}\widehat{\Sigma}\boldsymbol{c} = \sum_{k=1}^{M} n_k/y_k^2$. Hence a confidence interval for cumulative incidence with coverage probability converging to $1-\alpha$ is

$$\widehat{\Lambda}(t_M) \mp u_{1-\alpha/2}\sqrt{\sum_{k=1}^{M}\frac{n_k}{y_k^2}}.$$

1.6. Exposures and exposure-disease associations

Excess risk and relative risk

Now consider a potential risk factor Z, which could affect the risk of developing the disease. In epidemiology, such a variable is called *an exposure*. The exposure can be some environmental factor (radiation, air pollution), a characteristic of the lifestyle of the subject (diet, smoking, alcohol consumption), an innate factor (genetics, ethnicity), a cultural factor, or just anything. We would like to learn whether the exposure affects the incidence of the disease and evaluate its effect.

Classical epidemiology works with discrete (or discretized) variables. Let us assume at this moment that the exposure Z is discrete with values $0, 1, \ldots, J$. Value 0 means an "unexposed" subject, who was never subjected to the risk factor at all (e.g., a non-smoker). Subsequent values of Z indicate increasing levels of the exposure. The simplest special case for $J = 1$ leads to a binary exposure that allows only two values: 0 = unexposed (non-smoker) and 1 = exposed (smoker).

Denote by λ_{jk} the incidence of the disease at the j-th level of the exposure over the k-th age interval, $j = 0, \ldots, J$, $k = 1, \ldots, M$. There are two ways to express the effect of the exposure on the incidence. We can take differences relative to the unexposed, i.e.,

$$ER_{jk} = \lambda_{jk} - \lambda_{0k}.$$

This is called *the excess risk*. The excess risk can be interpreted as the expected number of additional new cases of the disease occurring in a population with exposure level j (compared to the unexposed population of the same size). If both incidences are expressed in the usual way, per 10^5 person-years of follow-up, ER_{jk} is the number of cases attributable to the exposure that occur in a population of 100, 000 people over one year.

Another measure of exposure effects is provided by *the relative risk*

$$RR_{jk} = \frac{\lambda_{jk}}{\lambda_{0k}}.$$

The relative risk RR_{jk} shows the proportional increase in incidence for the exposure level j (relative to the unexposed subjects) at the age category k.

In many practical examples the relative risk does not depend on the age category, i.e., $RR_{jk} = r_j$ for all $k = 1, \ldots, K$. Realizing analogy with the Cox model, we say that the *proportional hazards condition* is satisfied. Then the effect of exposure level j can be expressed by a single number r_j.

Empirical estimates of relative risk and excess risk can be obtained from empirical estimates of λ_{jk} calculated for each exposure-by-age subgroup by the methods explained in Section 1.4. However, as we show below, relative risk can be estimated even from data that only include disease status, without any recorded follow-up information. This is the most important advantage of relative risk over excess risk.

However, relative risk should not be interpreted without consideration of the wider context, especially when interested in the practical importance of various risk factors. For example, a rare exposure Z_1 with a seemingly high relative risk of 5 or 10 for a rare disease represents a much smaller practical problem for the overall human health than a common exposure Z_2 with modest relative risk of 1.2 acting on a common disease. Efforts to eliminate Z_1 will bring much smaller benefits than efforts to eliminate Z_2. For such considerations, the excess risk would be more relevant measure than the relative risk.

Estimating relative risk when no follow-up data are available

The relative risk is defined as ratio of incidences. Estimation of incidence requires information of individual follow-up or at least number of subjects who are at risk at different age intervals. However, under some circumstances, it is possible to estimate relative risk from binary data that only includes information on disease status of the subjects, but not on follow-up. In this section, we show why this is so and under what circumstances such simplified estimation provides reasonable results.

It is shown in Section 1.5 on p. 12 that for rare diseases, the cumulative incidence approximates the probability of not having the disease at a given age. The cumulative incidence for the exposure group j at the end of the age interval k is

$$\Lambda_j(t_k) = \sum_{l=1}^{k} \lambda_{jl} d_l \approx F_j(t_k) = \mathrm{P}\big[T \le t_k \,\big|\, Z = j\big],$$

where d_l is the duration of the l-th age interval and F_j is the conditional distribution function of age at occurrence of disease given exposure group j. To write it in this way, we need to assume that the exposure is time-invariant (each subject stays in the same exposure group for the whole life).

Now make the proportional hazards assumption, i.e., let there be no interaction between exposure and age. Take λ_{0k} the incidence of the unexposed group in the k-th age interval and express $\lambda_{jk} = r_j \lambda_{0k}$. Then

$$\Lambda_j(t_k) = \sum_{l=1}^{k} r_j \lambda_{0l} d_l = r_j \Lambda_0(t_k).$$

Using the approximate equality of cumulative incidence and distribution function, we get

$$r_j = \frac{\Lambda_j(t_k)}{\Lambda_0(t_k)} \approx \frac{\mathrm{P}\left[T \leq t_k \,\middle|\, Z = j\right]}{\mathrm{P}\left[T \leq t_k \,\middle|\, Z = 0\right]} \quad \text{for all } k = 1, \ldots, M.$$

Taking $k = M$, we have $r_j = \mathrm{P}\left[T \leq t_M \,\middle|\, Z = j\right] / \mathrm{P}\left[T \leq t_M \,\middle|\, Z = 0\right]$, the ratio of overall probabilities of being a case (by the upper age limit t_M) in the exposure group j relative to the unexposed. We can estimate the probability of being a case empirically by relative frequencies. Let $\overline{N}_{j\cdot}$ be the total number of observed cases in the exposure group $j \in \{0, \ldots, J\}$ and let n_j be the number of subjects in the exposure group j. Then $\overline{N}_{j\cdot}/n_j$, $j = 1, \ldots, J$, estimates $\mathrm{P}\left[T \leq t_M \,\middle|\, Z = j\right]$ and $\overline{N}_{0\cdot}/n_0$ estimates $\mathrm{P}\left[T \leq t_M \,\middle|\, Z = 0\right]$. The relative risk r_j can be estimated by

$$\widehat{r}_j = \frac{\overline{N}_{j\cdot}/n_j}{\overline{N}_{0\cdot}/n_0}.$$

To calculate this estimator of relative risk, we only need to know the numbers of cases and controls in each exposure category, nothing more. This approach works well if the following conditions are satisfied:

- the disease is rare;
- the exposure is time invariant;
- the proportional hazards assumption holds.

Supplementary reading

- Esteve et al. (1994, Chap. 1, pp. 1–34; Chap. 2, pp. 49–62)
- Breslow and Day (1980, Chap. II, pp. 42–49)
- Breslow and Day (1987, Chap. 2, pp. 48–57)

2. Analyzing Exposure-Disease Associations

2.1. Epidemiological study design: Cohort studies and case-control studies

In epidemiology, we are frequently interested in estimating the association between an exposure and a disease and making statistical inference about it. We want to answer questions such as "Is the exposure related to the disease or not?" or "How strong is the association and what is the uncertainty about its size?" So, we need methods for hypothesis testing and construction of confidence intervals. Which methods are most appropriate for a particular problem depends on the nature of the problem, structure of the data available for analysis, and data collection methods (study design). Classical statistical methods usually require the *iid paradigm*, availability of independent and identically distributed observations–random vectors that contain data observed on each participant. However, in epidemiology the data are rarely collected in this way. Therefore we first consider the important issue of study design.

Let us start with the simplest possible setup of a single binary exposure X. For each of N participating subjects, we observe a pair (X_i, Y_i), where $X_i \in \{0, 1\}$ is the exposure (0 = unexposed and 1 = exposed) and $Y_i \in \{0, 1\}$ is the disease status (0 = control, disease-free subject and 1 = case, subject having the disease). No follow-up information is available. The data can be summarized in the form of a 2×2 contingency table as displayed in Table 2.1.

Suppose the subjects in the table came from a population with classification probabilities of the four cells denoted by $\boldsymbol{\pi} = (\pi_1^E, \pi_1^U, \pi_0^E, \pi_0^U)$. The exposure-disease association can be expressed, e.g., by the relative risk discussed in Section 1.6. Under the conditions described

Table 2.1.: Observed frequencies in a case-control study with a single binary exposure.

	$X = 1$ (Exposed)	$X = 0$ (Unexposed)	Total
Cases ($Y = 1$)	n_1^E	n_1^U	n_1
Controls ($Y = 0$)	n_0^E	n_0^U	n_0
Total	n^E	n^U	N

there, the relative risk can be approximated even in the absence of follow-up data. We have

$$RR = \frac{\pi_{1|E}}{\pi_{1|U}}, \tag{2.1}$$

where $\pi_{1|E} = \mathrm{P}\big[Y = 1 \big| X = 1\big] = \pi_1^E/(\pi_1^E+\pi_0^E)$ is the probability of disease in the exposed group and $\pi_{1|U} = \mathrm{P}\big[Y = 1 \big| X = 0\big] = \pi_1^U/(\pi_1^U + \pi_0^U)$ is the probability of disease in the unexposed group.

Now consider three different ways how the data forming the contingency table could have been obtained.

Cohort study

Let the N participating subjects be independently drawn from the population of interest. Then the data $(n_1^E, n_1^U, n_0^E, n_0^U)$ have a joint multinomial distribution. All components of $\boldsymbol{\pi}$ and any of their functions can be consistently estimated with such data. In particular, the empirical estimate of the relative risk is simply $(n_1^E n^U)/(n^E n_1^U)$. It is also the MLE under the multinomial model.

This is a design that satisfies the *iid paradigm*. The participating subjects are independently drawn and form a representative sample from a single underlying population. In epidemiology, this design is called *the cohort design*[*]. All the classical statistical methods for the analysis of two-way tables can be used with the cohort design.

However, data collected in this way are likely to yield very small counts in the contingency table if the disease is rare and/or if the exposure is rare. In epidemiology, most diseases are quite rare. This is, by the way, an assumption needed for a reasonable approximation of the relative risk by (2.1). Cohort studies are quite inefficient because they need very large sample sizes N to yield sufficient information about the disease.

Case-control study

The problem with small numbers of cases of rare diseases can be alleviated in the following way. Fix the desired number of cases n_1 and the number of controls n_0 beforehand. Then obtain a random sample of size n_1 from subjects who have the disease and another random sample of size n_0 from subjects who do not have the disease, and observe their exposures. This design is called *the case-control design*[†].

The case-control data forms two independent samples from two different subpopulations and it does not follow the *iid paradigm*. We have $n_1^E \sim \mathrm{Bi}(n_1, \pi_{E|1})$ and $n_0^E \sim \mathrm{Bi}(n_0, \pi_{E|0})$ independent, where $\pi_{E|1} = \mathrm{P}\big[X = 1 \big| Y = 1\big] = \pi_1^E/(\pi_1^E + \pi_1^U)$ is the probability of being exposed for the cases and $\pi_{E|0} = \mathrm{P}\big[X = 1 \big| Y = 0\big] = \pi_0^E/(\pi_0^E + \pi_0^U)$ is the probability of being exposed for the controls. Because the model for this design depends only on $\pi_{E|1}$ and $\pi_{E|0}$, it

[*] Česky *kohortová studie* [†] Česky *studie případů a kontrol*

allows consistent estimation of exposure distributions given the outcome but not the marginal or conditional distributions of the outcome. For example, with the case control design, we cannot estimate the prevalence of the disease $\pi_1 = \mathrm{P}[Y = 1]$. The consistent ML estimator of π_1 for the cohort design is n_1/N. However, in the case-control study, n_1 is set by the investigator and n_1/N is not a consistent estimator of π_1. Similarly, we cannot estimate $\pi_{1|E}$, the probability of disease in the exposed group, or $\pi_{1|U}$, the probability of disease in the unexposed group, or the relative risk $RR = \frac{\pi_{1|E}}{\pi_{1|U}}$ from case-control data.

The case-control study is an example of *outcome-dependent sampling*, a data collection mechanism for which inclusion into the study depends on the outcome (here, disease status). Such data have to be analyzed by specialized methods.

Exposure-dependent sampling

It is also possible to create the contingency table by setting the numbers of exposed subjects n^E and the number of unexposed subjects n^U and taking independent random samples from these two groups. This design may be called *exposure-dependent sampling*.

The data (n_1^E, n_1^U) can be represented by two independent samples from two binomial distributions, this time column-wise: $n_1^E \sim \mathrm{Bi}(n^E, \pi_{1|E})$ and $n_1^U \sim \mathrm{Bi}(n^U, \pi_{1|U})$. Here, we cannot estimate the exposure probability $\pi_E = \mathrm{P}[X = 1]$ or the conditional probabilities $\pi_{E|1}$ and $\pi_{E|0}$ (but these are of little interest anyway), but we can estimate the relative risk $RR = \frac{\pi_{1|E}}{\pi_{1|U}}$.

This sampling design could be useful with rare exposures affecting a common disease but it is not used as frequently in epidemiological studies as the case-control design.

Odds ratio and its invariance

The question is whether we can estimate anything useful about exposure-disease association from the case-control study. The answer is, of course, *"Yes, we can!"* The parameter we will focus on is the *odds ratio*[*]. In the cohort design, the odds ratio is defined as

$$\theta \equiv OR = \frac{\pi_1^E \pi_0^U}{\pi_0^E \pi_1^U}.$$

The odds ratio is 1 if and only if there is no association between the exposure X and the disease Y, that is, if and only if the relative risk is 1. However, the odds ratio can be rewritten in terms of parameters that are identifiable under each of the three designs, in particular

$$\theta \equiv OR = \frac{\pi_1^E \pi_0^U}{\pi_0^E \pi_1^U} = \frac{\pi_{1|E}\pi_{0|U}}{\pi_{0|E}\pi_{1|U}} = \frac{\pi_{E|1}\pi_{U|0}}{\pi_{E|0}\pi_{U|1}}.$$

[*] Česky *poměr šancí*

Hence, the odds ratio is invariant to the study design and can be always estimated consistently. Furthermore, if the disease is rare,

$$\theta \equiv OR = \frac{\pi_{1|E}\pi_{0|U}}{\pi_{0|E}\pi_{1|U}} = \frac{\pi_{1|E}(1-\pi_{1|U})}{\pi_{1|U}(1-\pi_{1|E})} \approx \frac{\pi_{1|E}}{\pi_{1|U}} = RR$$

so the odds ratio well approximates the relative risk of rare diseases.

The empirical estimate of odds ratio is the same for all three study designs:

$$\widehat{\theta} = \frac{n_1^E n_0^U}{n_0^E n_1^U}. \tag{2.2}$$

It is also the maximum likelihood estimator, with consistency and asymptotic normality guaranteed by the general ML theory.

2.2. Odds ratio estimation and testing: classical methods

In this section, we review classical methods for estimation and testing of odds ratios in the simplest case of a single binary exposure, which can be transformed into a 2×2 contingency table.

As explained in the previous section, it is possible to use odds ratios for describing exposure-disease associations instead of relative risks. The odds ratio can be estimated consistently in the same way with each of the study designs and approximates the relative risk closely if the disease is rare. In cohort studies, however, follow-up information is usually available, so we prefer analyses that estimate relative risks using survival analysis methods. Therefore the methods covered in this section are primarily applicable to case-control studies.

Large sample methods

Large sample methods for estimating and testing the odds ratio in a 2×2 table are taught in basic statistics courses. The MLE $\widehat{\theta}$ is given by (2.2). Even though this estimator is asymptotically normal, it is better to transform it by the log-transformation and use the convergence

$$\frac{\log\widehat{\theta} - \log\theta}{\sqrt{\widehat{V}_\theta}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1), \tag{2.3}$$

where

$$\widehat{V}_\theta = \frac{1}{n_1^E} + \frac{1}{n_1^U} + \frac{1}{n_0^E} + \frac{1}{n_0^U} \tag{2.4}$$

and $n\widehat{V}_\theta$ is a consistent estimator of the asymptotic variance of $\sqrt{n}(\log\widehat{\theta} - \log\theta)$. This can be proven by the delta method and Slutsky Theorem.

We can use (2.3) to construct asymptotic tests of $H_0 : \theta = \theta_0$ and asymptotic confidence intervals for θ.

The most common asymptotic test of independence in a two-way table is the Pearson χ^2 test. The general form of the test statistic is

$$\chi^2 = \frac{\left(n_1^E - \frac{n^E n_1}{N}\right)^2}{\frac{n^E n_1}{N}} + \frac{\left(n_1^U - \frac{n^U n_1}{N}\right)^2}{\frac{n^U n_1}{N}} + \frac{\left(n_0^E - \frac{n^E n_0}{N}\right)^2}{\frac{n^E n_0}{N}} + \frac{\left(n_0^U - \frac{n^U n_0}{N}\right)^2}{\frac{n^U n_0}{N}}.$$

Under the null hypothesis $H_0 : \theta = 1$, this statistic converges in distribution to χ_1^2. For a 2×2 table, the statistic can be rewritten as (see Anděl 2002, Theorem 13.5)

$$\chi^2 = \frac{N}{n^E n^U n_0 n_1}(n_1^E n_0^U - n_0^E n_1^U)^2 = \frac{N(n_0^E n_1^U)^2}{n^E n^U n_0 n_1}(\widehat{\theta} - 1)^2.$$

The Pearson χ^2 test is actually a Rao-type test based on the asymptotic normality of the untransformed odds ratio $\sqrt{N}(\widehat{\theta} - \theta)$ with $\theta = 1$. It can be used to test $H_0 : \theta = 1$ but it cannot be easily generalized to test $H_0 : \theta = \theta_0$ or to construct confidence intervals for θ.

Small sample methods

The large sample methods require large enough counts in the whole table. Methods based on asymptotic normality of log-transformed odds ratio (2.3) completely break down when at least one of the cell counts is zero. The Pearson χ^2 test is unreliable when any of the cell counts is too small (e.g., < 5). The MLE (2.2) takes an infinite value when $n_0^E = 0$ or $n_1^U = 0$. Fortunately, exact small sample methods are available for 2×2 tables.

The small sample methods are based on the conditional distribution of the number of exposed cases n_1^E given all the marginals n_1, n_0, n^E, n^U. It can be shown that, when the true odds ratio is θ,

$$p_k(\theta) \equiv \mathrm{P}\left[n_1^E = k \,\middle|\, n_1, n_0, n^E, n^U; \theta\right] = \frac{\binom{n_1}{k}\binom{n_0}{n^E - k}\theta^k}{\sum_{i \in \mathcal{I}} \binom{n_1}{i}\binom{n_0}{n^E - i}\theta^i} = \frac{\binom{n^E}{k}\binom{n^U}{n_1 - k}\theta^k}{\sum_{i \in \mathcal{I}} \binom{n^E}{i}\binom{n^U}{n_1 - i}\theta^i}, \tag{2.5}$$

for $k \in \mathcal{I}$, where the set of permitted values $\mathcal{I}$ includes all natural numbers i such that $\max(0, n^E - n_0) \leq i \leq \min(n^E, n_1)$. This is called a *non-central hypergeometric distribution* . In the special case of $\theta = 1$, we get the standard hypergeometric distribution. Because the small sample methods condition on all the marginal counts, they are invariant with respect to the sampling design and can be used with cohort studies, case-control studies, and exposure-dependent sampling in the same way.

The conditional maximum likelihood estimator $\widetilde{\theta}$ of odds ratio is obtained by maximizing $p_{n_1^E}(\theta)$ over $\theta > 0$ (with the observed value of n_1^E). It is the value of θ that equates the observed value of n_1^E with its expectation under the distribution (2.5), i.e., $\widetilde{\theta}$ solves the equation

$$n_1^E = \frac{\sum_{k \in \mathcal{I}} k \binom{n_1}{k} \binom{n_0}{n^E - k} \widetilde{\theta}^k}{\sum_{i \in \mathcal{I}} \binom{n_1}{i} \binom{n_0}{n^E - i} \widetilde{\theta}^i}.$$

This leads to the calculation of the roots of a polynomial of the degree $\min(n^E, n_1) - \max(0, n^E - n_0)$. One has to be careful to evaluate $p_{n_1^E}(\theta)$ at all the existing roots in order to find the global maximum.

We can use the distribution (2.5) to obtain exact tests and construct exact confidence intervals. Consider the hypothesis $H_0 : \theta = \theta_0$. An exact two-sided test will reject H_0 if

$$\sum_{k \le n_1^E, k \in \mathcal{I}} p_k(\theta_0) \le \frac{\alpha}{2} \quad \text{or} \quad \sum_{k \ge n_1^E, k \in \mathcal{I}} p_k(\theta_0) \le \frac{\alpha}{2}.$$

In the special case of $\theta_0 = 1$, this test is called *the Fisher exact test*[*], see Anděl (2002, Sec. 13.5). It is important to realize that, even though the test is called "exact", its true level is $\le \alpha$ because the discrete reference distribution does not allow reaching an arbitrary level exactly.

An "exact" confidence interval for the odds ratio based on the distribution (2.5) has a lower limit θ_L and an upper limit θ_U that satisfy the equations

$$\sum_{k \ge n_1^E, k \in \mathcal{I}} p_k(\theta_L) = \frac{\alpha}{2}$$

and

$$\sum_{k \le n_1^E, k \in \mathcal{I}} p_k(\theta_U) = \frac{\alpha}{2}.$$

2.3. Odds ratio estimation and testing: regression methods

We have only considered a single binary exposure so far. This is not sufficient for most practical applications, for example multi-level exposures, continuous exposures, or multiple exposures affecting the same disease. With cohort study or exposure-dependent sampling, such problems can be analyzed by logistic regression with disease status as the binary outcome. Logistic regression can be used to deal with fairy general situations and is the method of choice for estimation and testing of exposure-disease associations. With case-control sampling, the situation is less clear because disease status is not the outcome in this design and the *iid paradigm*

[*] Česky *Fisherův faktoriálový test*

Figure 2.1.: Graphical representation of (a) confounding effect of variable C and (b) mediating effect of variable C. The exposure is denoted by X, the disease by Y.

is violated. Nevertheless, we will show in the next chapter, section 3.3, that logistic regression can be used even with case-control sampling to estimate the parameters of interest properly.

Before we get there, however, we need to introduce and discuss the fundamental issue of confounding.

2.4. Confounding in epidemiological studies

The ultimate goal of epidemiological studies is to ascertain whether the exposure causes the disease or not. However, this cannot be done by considering only the exposure X and the disease Y out of the context of all other factors that may be related to them.

Imagine that there exists another variable C, which is associated both with the exposure and with the disease. If such a variable is not properly considered in the analysis, it can distort the true causal relationship between the exposure and the outcome. Figure 2.1 displays two possible ways how C could be related to the exposure X and the disease Y. The variables are pictured as nodes of a graph and causal relationships between them are marked by oriented edges with arrows.

In the left panel (a) of the figure, the variable C affects both the exposure X and the disease Y. Such a variable is called *a confounder*[*]. In the presence of a confounder, the true causal association between the exposure X and the disease Y cannot be estimated correctly. In order to estimate the true effect of X on Y, it is necessary to measure the confounder C and to remove its confounding effect by methods discussed below. Confounders can distort the true exposure effect both ways: they can create spurious effects that do not in fact exist, or they can mask a true exposure effect so that it cannot be detected.

In the right panel (b), the variable C is affected by the exposure X and, in turn, it affects the disease Y. Such a variable is called *a mediator*[†]. A mediating variable must not be treated as a confounder in the analysis; if its effect is removed, the analysis would estimate only the part of the true causal effect of X on Y, which is not mediated through C. In epidemiology, this error is called *over-matching*. Over-matching may also occur when X and C are two alternative measures expressing the same true cause of the disease. In that case it is also wrong to consider

[*] Česky *matoucí veličina* nebo *zkreslující veličina* [†] Česky *zprostředkující veličina*

C a confounder and to remove its effect.

It is also possible to encounter a variable that is a consequence of the disease. Such a variable is neither a confounder nor a mediator and, of course, it should not play any role in the analysis.

Thus, it is very important to distinguish confounders from mediators. Confounders must be taken into account in the analysis but mediators must not. It is impossible to tell the difference between them from the data alone. We can check which variables are associated both with the exposure and with the outcome but we usually cannot determine the direction of the association between X and C from the data. This requires an external expert knowledge that carefully considers the meaning of the measured variables and their role in the process leading to the disease.

Confounding is always a concern in observational (non-randomized) studies. The key principle of epidemiological studies is that all potential confounders should be considered when the experiment is planned, should be appropriately measured and accounted for. There are three main strategies to deal with confounding in case-control studies at the stage of sampling (subject selection) and/or analysis: stratification, matching, and adjustment.

Stratification [*] means that we classify the subjects into K groups (called *strata*) according to the value of the confounder. Thus, within each stratum, the subjects have fairly similar confounder values. Stratification can be performed at the sampling stage, by taking separate samples of cases and controls (of a pre-specified size) from each stratum (*stratified sampling*[†]). Or it can be performed at the analysis stage, by conducting separate analyses within each stratum and then combining stratum-specific effect estimates into one overall estimate (*stratified analysis*[‡]). Statistical methods for the analysis of stratified case-control studies will be discussed in the next chapter.

Matching [§] is actually a very fine stratification, with strata that are so small that each of them includes just a single case. It allows choosing controls for each case individually so that the controls are as similar to the case in terms of confounding variables as possible. Methods for the analysis of matched studies are explained in Chapter 4.

Adjustment [¶] means including the confounder as a covariate in the binary response regression model used for estimation of the exposure effect on the disease. Because the regression model conditions on the values of the covariates, it estimates the exposure effect conditionally on having equal values of the confounder.

Cohort studies are usually prospective, subjects are enrolled before their confounders and outcomes are known. Therefore, in cohort studies, adjustment is the only available approach. In case-control studies, stratification and matching are frequently done, combined with adjustment at the analysis stage.

[*] Česky *stratifikace* [†] Česky *stratifikovaný výběr* [‡] Česky *stratifikovaná analýza* [§] Česky *matching*
[¶] Česky *adjustment*

2.5. Practical issues

2.5.1. Causality

Assessing causality from observational studies is always difficult and not quite reliable. There are many sources of potential biases that may create a non-existing effect or mask a true effect of the exposure on the disease. Confounding is always the top concern but there are many other mechanisms that introduce bias. The only reliable way for estimation of causal effects is randomization. However, in epidemiology, randomization is only possible when protective measures designed to prevent the disease are investigated. Randomizing human subjects to harmful risk factors is not an option. This situation is not specific to epidemiological research, it applies to data analysis problems in general.

So, can we in fact make any conclusions about causal effects from epidemiological studies? The answer is a cautious "yes", under certain conditions. Breslow and Day (1980, pp. 86–90) discuss criteria that need to be satisfied in order to consider an association suggested by an observational study causal. Among those criteria are:

- **Dose-response relationship**. If the exposure is a true cause of the disease, the risk of the disease should increase with higher levels of the exposure and longer exposure duration. If such trends are not observed or are not monotone, the association is likely not to be causal.
- **Strength of the association**. If the estimated relative risk (odds ratio) is quite large (say, 5 or more), it is less likely that it could be brought about by confounding. On the other hand, a weak (though significant) relative risk such as 1.3 can easily be an artefact of confounding or other biases.
- **Credible temporal relationships**. If it can be shown that the exposure precedes the disease by a reasonable amount of time and/or that the increased risk occurs in time periods when the exposure is likely to have the largest effect, the arguments for causality are strengthened. If, on the other hand, the increased risk is observed in time periods when the exposure could not have had a strong impact, the association cannot be causal.
- **Plausible biological explanation**. There must be a plausible explanation of biological processes leading from exposure to disease, supported by findings of basic science.
- **Confirmation by multiple studies**. A positive finding has a higher credibility if it can be replicated by other studies conducted by a different methodology on different populations. A single positive study cannot be considered a definite proof of causality.

With respect to temporal relationships between the exposure and the disease, longitudinal studies that follow the evolution of exposure, disease, and confounders over a period of time provides much stronger evidence than a cross-sectional study, where all the data are collected at a single moment.

To summarize, interpretation of observational studies must be done carefully and thoughtfully. Their findings need not correspond to the true causal relationships. This is why we prefer

to describe the results as "associations" between X and Y rather than giving false impressions of causal meaning by calling them "effects" of X on Y.

With epidemiological studies, the problem of causality is serious enough. However, most of these studies are at least well planned, conducted according to a specific study design, with data carefully checked and processed. The situation is much worse with the analysis of data sets collected in ad-hoc ways where the study population is not well defined, participation rates are not under control, and variables and outcomes are recorded in a haphazard way. This criticism is especially relevant for so called "big data" problems, in which the size of the data set cannot make for the lack of control over sampling procedures and data quality.

2.5.2. Sources of bias

Confounding belongs to the most serious sources of bias in case-control studies but it is not the only source of bias. In order to come up with reliable results, all sources of bias should be minimized.

Selection of controls

In case-control studies, separate samples from cases and controls are obtained. It is important to make sure that the two groups come from the same underlying population. Ideally, cases would be taken from disease registry capturing occurences of the disease in a certain population (geographical area). Controls should be sampled from the same population to which the disease registry applies using, e.g., lists of registered voters, register of inhabitants, etc. However, this is rarely done due to practical constraints, such studies are expensive and time-consuming.

Commonly, cases are identified from medical records in a hospital where their disease is treated. The selection of appropriate controls for such population may be questionable. Sometimes controls are taken from other patients who are treated or hospitalized in the same hospital with a different condition (so called *hospital controls*), however, this may not be a random sample from the same population as the cases. The most serious concern is that the controls might have a condition caused by a factor closely related to the exposure, which would lead to a biased odds ratio estimate.

Case ascertainment, diagnosis accuracy

It is not surprizing that errors in case ascertainment (classification of subjects as cases/controls) may have very detrimental effects on the case-control study. It is a bit more surprizing that these errors are not as uncommon as it might seem. When cases are taken from routinely maintained disease registries or according to diagnoses marked in hospital records, the possibility of an incorrect diagnosis is real. Also the controls may not be as healthy as they should, some of them may be undiagnosed cases of the disease of interest.

Errors in diagnoses can be minimized by careful verification of the status of each of the study participants. This can be done, e.g, by review of medical records by an independent experienced physician or by conducting additional laboratory procedures to verify the diagnosis.

Misclassification of exposures

Exposures are frequently subject to measurement error. Sometimes past exposures can be determined relatively precisely using recorded information (employment records, medical records, etc.). More commonly, exposures must be estimated from imperfect sources or reported by the participants themselves or their relatives. Some exposures (intake of dietary fat) cannot be ascertained by any reliable method. Also, exposure levels may vary over time and it may be difficult to summarize them in a meaningful way. Measurement errors in exposures typically lead to shrinking estimated associations towards one.

Prospective vs. retrospective studies. Recall bias.

Another source of bias may arise by not keeping data collection procedures the same for both cases and controls. Cohort studies can be done in a prospective way (enrolling subjects and following them over time). This is a great advantage because exposure assessment can be standardized and verified. In such studies, all participants have data collected in the same way, measurements can be verified, errors identified and fixed, missing items resolved. Also, prospective studies allow better control of temporal relationships (does the suspected risk factor precede the development of the disease or not?).

Case-control studies (and some cohort studies) are done retrospectively: cases and controls are identified at a certain moment and their past exposures are ascertained. This often means that the way the exposures are measured cannot be controlled. Sometimes, it is necessary to ask the subjects to report their exposures over their lifetimes (or to ask their families if the subject is severely ill or deceised). This raises a concern that exposures could be reported differently by the controls (healthy subjects) than by the cases who are severely ill and have pondered about possible causes of their disease for some time (not speaking about families of subjects who died of the disease). This differential error in exposure assessment is called *recall bias*; it may substantially deteriorate the results of the study.

2.5.3. Sample size in case-control studies

Case-control studies are especially powerful for rare diseases. With rare diseases, cases are scarce so we include in the study all the cases that can be identified. Then the question is how many controls we should get. The number of controls should be at least equal to the number of cases. If we take more controls, we improve precision in the odds ratio estimate and increase the power to detect an exposure effect. However, if we take more than four controls per case, the additional gain in power is negligible.

If the disease is less rare we specify the total number of subjects to be enrolled and take samples of cases and controls of equal size.

Supplementary reading

- Breslow and Day (1980, Chap. III, pp. 84–115, Chap. IV, pp. 122-136)

3. Stratified Case-Control Studies

3.1. Stratified sampling vs. stratified analysis

Consider one or several confounders that we want to take into account in the analysis of a case-control study and define K strata based on the values of the confounder(s). Each subject from the population of interest is classified into one of the strata and the values of the confounder(s) within each stratum are similar. Thus, when we compare two subjects from the same stratum, they would not differ in the values of the confounder(s) too much.

For example, if the confounder is age, we can take subjects younger than 20 years as the first stratum and create additional strata as 10-year age groups 20 – 30 years, 30 – 40 years etc. The last stratum could be subjects 80 years old or older. We have created a total of $K = 8$ strata.

Stratified case-control sampling is performed by choosing a fixed number of cases and controls in each stratum. In small strata, we usually take all available cases and up to 4 times as many controls. In large strata, we can take a subsample of the cases and a suitable number of controls. Denote the number of cases in the k-th stratum by n_{1k} and the number of controls by n_{0k}. The total number of subjects in the stratum is n_k.

With a single binary exposure, which will be the setting investigated first, the data can be expressed as K 2×2 contingency tables. The notation is summarized in Table 3.1.

Stratified samples must be analyzed by *stratified analysis*. This is performed by estimating the parameter of interest (here, odds ratio) in each stratum and then combining the stratum-specific estimates into a single estimate/single test statistic. Because subjects in the stratum do not vary too much in terms of the confounders (used for setting up the strata), the con-

Table 3.1.: Observed frequencies in the k-th stratum of a stratified case-control study with a single binary exposure.

Stratum: k	$X = 1$ (Exposed)	$X = 0$ (Unexposed)	Total
Cases ($Y = 1$)	n_{1k}^E	n_{1k}^U	n_{1k}
Controls ($Y = 0$)	n_{0k}^E	n_{0k}^U	n_{0k}
Total	n_k^E	n_k^U	n_k

founding effect is (nearly) eliminated in each stratum-specific analysis. The overall estimate or test created by combining the stratum-specific analyses is therefore much less affected by confounding than a non-stratified analysis.

Stratified analysis can be performed even on data that was originally collected by ordinary case-control sampling. After the confounders are assessed, the subjects are divided into strata and the data are formed as in Table 3.1. The only difference is that the row totals are random instead of fixed. However, this does not affect the analysis. Forming strata on data that was not selected by stratified sampling is called *post-stratification*. As opposed to stratified sampling, in post-stratified data the numbers of cases and controls in each stratum are not under control and may turn quite unbalanced.

Stratification is suitable for controlling a small number of important confounders, typically age and sex. With too many confounders, we would either obtain a very large number of strata or stratification would be so rough that confounding would not be removed. For stratified sampling, it is necessary that stratification variables are already available when the cases and controls are selected for the study; one cannot stratify on variables that are difficult or expensive to measure. Most analysis methods for stratified case-control studies break down when the data include a large number of small strata – there are some exceptions, though, which are noted in the subsequent sections.

3.2. Classical methods for stratified case-control studies

As noted above, stratified analysis proceeds by performing the analysis within each stratum and combining the results across strata. If this strategy is to be effective for the odds ratio, we need to assume that all strata have the same odds ratio θ between the exposure and the disease. In other words, we assume that there is no interaction between the exposure and the confounder, or that the exposure affects the disease in the same way at all levels of the confounder.

With the notation of Table 3.1, the odds ratio can be estimated from the k-th stratum by the empirical estimator

$$\widehat{\theta}_k = \frac{n^E_{1k} n^U_{0k}}{n^E_{0k} n^U_{1k}}, \tag{3.1}$$

see (2.2). If there is no interaction, all these estimators estimate the same parameter θ. When $\theta \approx 1$, the variance of this estimator can be approximated by

$$\hat{\mathrm{var}}\, \widehat{\theta}_k \approx \left(\frac{n^U_{1k} n^E_{0k}}{n_k} \right)^{-1}. \tag{3.2}$$

(the variance estimator can be justified by the asymptotic normality of empirical relative frequencies and the delta method).

3.2.1. Cochran-Mantel-Haenszel test

Now consider the null hypothesis $H_0 : \theta = 1$ against the alternative $H_1 : \theta \neq 1$. Fix the marginals of each of the 2×2 contingency tables from Table 3.1. Under H_0 and fixed marginals, the number of exposed cases n_{1k}^E has a hypergeometric distribution with expectation

$$\mathsf{E}\, n_{1k}^E \equiv m_k = n_{1k} \frac{n_k^E}{n_k}$$

and exact variance

$$\operatorname{var} n_{1k}^E \equiv V_k = \frac{n_{1k} n_{0k} n_k^E n_k^U}{n_k^2 (n_k - 1)}.$$

The test statistic is composed in the same way as the logrank test for comparing two censored samples (even the expectation and variance are exactly the same as in that case, only with a different notation and meaning of the counts). Take

$$\chi^2_{\mathsf{MH}} = \frac{\left[\sum_{k=1}^K (n_{1k}^E - m_k)\right]^2}{\sum_{k=1}^K V_k}. \tag{3.3}$$

Under the null hypothesis, the test statistic has an asymptotic χ_1^2 distribution. The asymptotic distribution can be justified by similar arguments as in the case of the logrank test, although here the situation is simpler because the tables for the individual strata are independent of each other. The asymptotic distribution is valid under both kinds of asymptotics: a small number of strata with large number of subjects in each, or a large number of small strata. This is a great advantage of this test.

The null hypothesis is rejected at the asymptotic level of α if and only if

$$\chi^2_{\mathsf{MH}} \geq \chi_1^2(1 - \alpha),$$

where $\chi_1^2(1 - \alpha)$ is the $1 - \alpha$ quantile of the χ_1^2 distribution.

This test is called *the Cochran-Mantel-Haenszel test*, or, in short *the Mantel-Haenszel test*. It was first suggested by Cochran (1954), with a somewhat different V_k, and later made popular by Mantel and Haenszel (1959).

3.2.2. Woolf estimator

The next task is to estimate the common odds ratio parameter θ. All the stratum-specific estimators $\widehat{\theta}_k$ estimate this parameter so we need to combine them in a suitable way into a single estimator. Woolf (1955) proposed to work on the log odds scale and to use the asymptotic distribution (2.3) for $n_k \to \infty \; \forall k$. *The Woolf estimator* is defined by

$$\log \widehat{\theta}_{\mathsf{W}} = \frac{\sum_{k=1}^K w_k \log \widehat{\theta}_k}{\sum_{k=1}^K w_k},$$

where w_k are weights. The optimal weights for a linear combination of consistent estimators are inverses of variances of the individual estimators; in this case we get

$$w_k = \left(\frac{1}{n_{1k}^E} + \frac{1}{n_{1k}^U} + \frac{1}{n_{0k}^E} + \frac{1}{n_{0k}^U} \right)^{-1},$$

see (2.4). It is not difficult to see that the asymptotic variance of $\log \widehat{\theta}_W$ with these weights can be estimated by $\left(\sum_{k=1}^K w_k \right)^{-1}$. It follows from the asymptotic normality of $\log \widehat{\theta}_k$ that

$$\frac{\log \widehat{\theta}_W - \log \theta}{\sqrt{\left(\sum_{k=1}^K w_k \right)^{-1}}} \xrightarrow{D} N(0, 1)$$

as $n_k \to \infty$ for all $k = 1, \ldots, K$. This allows to construct asymptotic confidence intervals for θ.

The Woolf estimator works well when the number of strata is small and the number of subjects is large in each stratum. However, stratum-specific estimates of $\log \theta$ cannot even be calculated when any of the cell counts in the stratum-specific contingency table is 0. Even if all the counts are positive, but small, the practical performance of the Woolf estimator is very poor.

3.2.3. Mantel-Haenszel estimator

A much better estimator of the common odds ratio was proposed by Mantel and Haenszel (1959). They used empirical odds ratio estimators directly, without taking logs, together with asymptotic variances (3.2).

The Mantel-Haenszel estimator is obtained as the weighted average

$$\widehat{\theta}_{MH} = \frac{\sum_{k=1}^K \omega_k \widehat{\theta}_k}{\sum_{k=1}^K \omega_k},$$

with weights taken as inverses of (3.2), that is $\omega_k = n_{1k}^U n_{0k}^E / n_k$. When we plug in expressions for $\widehat{\theta}_k$ and ω_k, we get

$$\widehat{\theta}_{MH} = \frac{\sum_{k=1}^K \omega_k \widehat{\theta}_k}{\sum_{k=1}^K \omega_k} = \frac{\sum_{k=1}^K \frac{n_{1k}^U n_{0k}^E}{n_k} \frac{n_{1k}^E n_{0k}^U}{n_{0k}^E n_{1k}^U}}{\sum_{k=1}^K \frac{n_{1k}^U n_{0k}^E}{n_k}} = \frac{\sum_{k=1}^K \frac{n_{1k}^E n_{0k}^U}{n_k}}{\sum_{k=1}^K \frac{n_{1k}^U n_{0k}^E}{n_k}}. \tag{3.4}$$

It can be shown that the Mantel-Haenszel estimator is 1 if and only if the Mantel-Haenszel test statistic is exactly zero. Indeed,

$$\chi^2_{MH} = 0 \quad \Leftrightarrow \quad \sum_{k=1}^K n_{1k}^E \frac{n_k}{n_k} = \sum_{k=1}^K n_{1k} \frac{n_k^E}{n_k}.$$

Write the k-th term on the left-hand side as

$$\frac{n_{1k}^E(n_{1k}^E + n_{1k}^U + n_{0k}^E + n_{0k}^U)}{n_k} = \frac{(n_{1k}^E)^2 + n_{1k}^E n_{1k}^U + n_{1k}^E n_{0k}^E + n_{1k}^E n_{0k}^U}{n_k}$$

and the k-th term on the right-hand side as

$$\frac{(n_{1k}^E + n_{1k}^U)(n_{1k}^E + n_{0k}^E)}{n_k} = \frac{(n_{1k}^E)^2 + n_{1k}^E n_{1k}^U + n_{1k}^E n_{0k}^E + n_{1k}^U n_{0k}^E}{n_k}.$$

Clearly, the two sides (after summation) are equal if and only if

$$\sum_{k=1}^K \frac{n_{1k}^E n_{0k}^U}{n_k} = \sum_{k=1}^K \frac{n_{1k}^U n_{0k}^E}{n_k},$$

that is, $\widehat{\theta}_{\mathrm{MH}} = 1$.

The Mantel-Haenszel estimator has good properties even if some of the cell counts are very small or even zero. Is is actually consistent under both kinds of asymptotics: for K constant and $n_k \to \infty$ for all k, as well as for $K \to \infty$ and n_k small.

It took over 25 years to prove that $\log \widehat{\theta}_{\mathrm{MH}}$ is asymptotically normal and to develop a variance estimator. Eventually, it was done by Robins et al. (1986). Their asymptotic variance formula (for $\log \widehat{\theta}_{\mathrm{MH}}$) is rather complicated and we show it here just to share its aesthetic beauty:

$$\frac{\sum_{k=1}^K \frac{n_{1k}^E + n_{0k}^U}{n_k}\frac{n_{1k}^E n_{0k}^U}{n_k}}{2\left(\sum_{k=1}^K \frac{n_{1k}^E n_{0k}^U}{n_k}\right)^2} + \frac{\sum_{k=1}^K \left(\frac{n_{1k}^E + n_{0k}^U}{n_k}\frac{n_{0k}^E n_{1k}^U}{n_k} + \frac{n_{0k}^E + n_{1k}^U}{n_k}\frac{n_{1k}^E n_{0k}^U}{n_k}\right)}{2\left(\sum_{k=1}^K \frac{n_{1k}^E n_{0k}^U}{n_k}\right)\left(\sum_{k=1}^K \frac{n_{0k}^E n_{1k}^U}{n_k}\right)} + \frac{\sum_{k=1}^K \frac{n_{0k}^E + n_{1k}^U}{n_k}\frac{n_{0k}^E n_{1k}^U}{n_k}}{2\left(\sum_{k=1}^K \frac{n_{0k}^E n_{1k}^U}{n_k}\right)^2}.$$

This can be used to construct confidence intervals for $\log \theta$ (and thence θ).

3.3. Logistic regression for stratified case-control studies

The classical methods discussed in the previous section are limited to a single binary exposure and do not allow further adjustment for confounders not used for stratification. Logistic regression represents a much more general and flexible tool for the analysis of stratified (as well as non-stratified) case-control studies. It allows simultaneous estimation and testing of effects of multiple exposures of an arbitrary nature (binary, ordinal, continuous). It also allows adjustment for additional confounders and interactions between stratum and exposure, or exposure and confounder. However, theoretical properties of logistic regression have been developed in the context of iid data, so it is necessary to justify its use with stratified case-control sampling.

Let us introduce a framework for investigating this problem. Let C be a discretized confounder used for stratification. The values of C are $k = 1, \ldots, K$ and $q_k = \mathrm{P}[C = k]$ (in the general population), so that $\sum_{k=1}^{K} q_k = 1$. Consider a vector of covariates $\boldsymbol{X}$ that includes exposures and confounders not included in C, transformed in a suitable way. The disease status is captured by the variable Y ($Y = 1$ means disease case, $Y = 0$ means disease-free control).

Consider the joint distribution of $(Y, \boldsymbol{X}, C)$ in the general population and write its density as

$$f(y, \boldsymbol{x}, k) = f(y|\boldsymbol{x}, k) \cdot f(\boldsymbol{x}|k) \cdot f(k) = f_k(y|\boldsymbol{x}) \cdot f_k(\boldsymbol{x}) \cdot q_k,$$

where $f_k(y|\boldsymbol{x})$ is the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ in the k-th stratum and $f_k(\boldsymbol{x})$ is the density of $\boldsymbol{X}$ in the k-th stratum. Assume that $f_k(y|\boldsymbol{x})$ can be expressed through a stratum-specific logistic regression model, in particular

$$\mathrm{P}\big[Y = 1 \,\big|\boldsymbol{X}, C = k\big] = \frac{\mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{X}}}{1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{X}}}.$$

The parameters are stratum-specific and have the following interpretation: take β_{kj}, the j-th component of $\boldsymbol{\beta}_k$, then $\mathrm{e}^{\beta_{kj}}$ is the odds ratio for the disease due to a unit increase in the j-th covariate in the k-th stratum, and e^{α_k} is the odds of the disease for a subject with $\boldsymbol{X} = \boldsymbol{0}$ in the k-th stratum.

This setup provides a model for exposure-disease association in the general population. Notice that the implied population model that ignores strata is

$$\mathrm{P}\big[Y = 1 \,\big|\boldsymbol{X}\big] = \sum_{k=1}^{K} q_k \mathrm{P}\big[Y = 1 \,\big|\boldsymbol{X}, C = k\big] = \sum_{k=1}^{K} q_k \frac{\mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{X}}}{1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{X}}},$$

and it cannot be transcribed as a logistic regression model unless $K = 1$ or all coefficients are independent of strata: $\boldsymbol{\beta}_k = \boldsymbol{\beta}$ and $\alpha_k = \alpha$ for all $k = 1, \ldots, K$. In the latter case, C is not a confounder at all (because it is not related to the disease) and stratification is useless.

Stratified case-control sampling means that we obtain samples from the conditional distribution of $\boldsymbol{X}$ given Y and C. Introduce a sampling indicator ξ so that $\xi = 1$ means that the subject (from the general population) is included in the stratified case-control sample and $\xi = 0$ means that the subject is not included. Write the complete data as $(Y, \boldsymbol{X}, C, \xi)$. The observations consist of $(Y, \boldsymbol{X}, C)$ for the subjects with $\xi = 1$. From such data, we can estimate elements of the conditional distribution of $(Y, \boldsymbol{X}, C)$ given $\xi = 1$.

Stratified case-control sampling is obtained by setting the sampling probabilities as follows

$$\mathrm{P}\big[\xi = 1 \,\big|Y = 1, \boldsymbol{X}, C = k\big] = \pi_{1k} \quad \text{and} \quad \mathrm{P}\big[\xi = 1 \,\big|Y = 0, \boldsymbol{X}, C = k\big] = \pi_{0k}.$$

These probabilities depend on the outcome (disease) and stratum (confounder) but not on anything included in $\boldsymbol{X}$. With rare diseases we would set $\pi_{1k} = 1$ (sample all cases from each stratum) but this is not necessary. In all real applications, $\pi_{1k} \gg \pi_{0k}$.

Now we calculate the probability that the subject is a case ($Y = 1$) given $\boldsymbol{X}$ and C among subjects included in the stratified case-control sample, i.e., within the data we actually have available for analysis. In the process, we use the assumptions formulated above.

$$\begin{aligned}
\mathrm{P}\left[Y = 1 \middle| \boldsymbol{X}, C = k, \xi = 1\right] &= \frac{\mathrm{P}\left[Y = 1, \xi = 1 \middle| \boldsymbol{X}, C = k\right]}{\mathrm{P}\left[\xi = 1 \middle| \boldsymbol{X}, C = k\right]} \\
&= \frac{\mathrm{P}\left[\xi = 1 \middle| Y = 1, \boldsymbol{X}, C = k\right]\mathrm{P}\left[Y = 1 \middle| \boldsymbol{X}, C = k\right]}{\mathrm{P}\left[\xi = 1 \middle| Y = 1, \boldsymbol{X}, C = k\right]\mathrm{P}\left[Y = 1 \middle| \boldsymbol{X}, C = k\right] + \mathrm{P}\left[\xi = 1 \middle| Y = 0, \boldsymbol{X}, C = k\right]\mathrm{P}\left[Y = 0 \middle| \boldsymbol{X}, C = k\right]} \\
&= \frac{\pi_{1k}\dfrac{\mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}}{1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}}}{\pi_{1k}\dfrac{\mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}}{1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}} + \pi_{0k}\dfrac{1}{1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}}} = \frac{\dfrac{\pi_{1k}}{\pi_{0k}}\mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}}{\dfrac{\pi_{1k}}{\pi_{0k}}\mathrm{e}^{\alpha_k + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}} + 1} = \frac{\mathrm{e}^{\alpha_k^* + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}}{1 + \mathrm{e}^{\alpha_k^* + \boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{X}}},
\end{aligned}$$

where $\alpha_k^* = \alpha_k + \log\frac{\pi_{1k}}{\pi_{0k}}$ does not depend on $\boldsymbol{X}$ (because π_{0k} and π_{1k} do not depend on $\boldsymbol{X}$ when $C = k$ is fixed).

This means that, in data collected by stratified case-control sampling, the disease status Y satisfies the same logistic regression model as if we were working with an iid sample from the general population, except for the main effects of the strata α_k (strata-specific intercepts), which are shifted by a quantity depending on the sampling probabilities. Because $\pi_{1k} \gg \pi_{0k}$, there is an upward bias in α_k. The parameters $\boldsymbol{\beta}_k$ which may include main effects of the exposures, main effects of potential additional confounders, and their interactions with stratum, will be estimated consistently by the logistic regression model.

It is not surprising that strata-specific odds of disease cannot be estimated consistently — the case-control sampling design lacks the information about them. However, if we knew the sampling probabilities for the cases and controls, we would be able to evaluate and subtract the bias in strata effects and get consistent estimates of α_k. Because non-stratified case-control design is a special case (with $K = 1$), we have also verified the validity of logistic regression for ordinary case-control studies, as foreseen in Section 2.3.

The arguments we laid out here only justify consistency of estimates of $\boldsymbol{\beta}_k$ obtained from logistic regression. Additional development is needed to show that the estimated parameters are asymptotically normal, that their asymptotic variance can be estimated by the observed information from logistic regression fits, and that likelihood ratio (deviance) tests are valid. All of this has been shown by Prentice and Pyke (1979) for ordinary case-control studies and by Scott and Wild (1991) for stratified case-control studies.

To summarize, stratified case-control data can be analyzed by logistic regression with disease status as the response, as long as main effects of the stratum are included in the linear predictor. The model correctly estimates the effects of exposures, additional confounders, mutual interactions among them, and interactions of these variables with stratum. Confidence intervals and tests for these parameters provided by standard logistic regression software are

all asymptotically valid. The asymptotics for logistic regression, however, does not allow the number of parameters α_k^* to grow to infinity. The asymptotic results require that K is bounded and the numbers of subjects per stratum n_k all converge to ∞. Thus, logistic regression cannot be used for a large number of small strata.

Last, we note that there is a connection between the Mantel-Haenszel test and logistic regression. Take a binary exposure $X \in \{0, 1\}$ and fit the logistic model

$$\mathrm{P}\left[Y = 1 \,\middle|\, X, C = k, \xi = 1\right] = \frac{\mathrm{e}^{\alpha_k^* + \beta X}}{1 + \mathrm{e}^{\alpha_k^* + \beta X}}$$

with β independent of stratum (to satisfy the assumption of no interaction between exposure and stratum). Then the score statistic for testing the null hypothesis $H_0 : \beta = 0$ (no exposure effect) is equivalent to Cochran's formulation of the Cochran-Mantel-Haenszel test statistic.

Supplementary reading

- Breslow and Day (1980, Chap. IV, pp. 136–146, Chap. VI, pp. 192–242)

4. Matched Case-Control Studies

4.1. Principles of matching

Matching is just a very fine stratification: the strata are made so small that there is a single case in each, with a small number m of controls. The controls are selected to match the particular case in terms of important confounders as close as possible. The number of controls per case is pre-determined by the study plan. An important and common special case of $m = 1$ is called a matched-pair design, with a single control selected for each case.

Typical matching variables in epidemiological studies are age and sex – the controls have the same sex as the case and are just as old. One can include other variables in the matching procedure. The advantage of matching is that we can select controls who look in some aspect "similar" to the case without having to define precisely what exactly "similarity" means. This can be achieved, e.g., by picking controls from the same social or geographical environment the case belongs to. Thus, by matching we can achieve adjustment for confounders that are difficult to quantify or difficult to measure (lifestyle, genetics, culture).

When investigating risk factors for development of Type 1 diabetes in children, we take cases who attend elementary school and select matched controls from the case's classmates of the same sex who do not have diabetes. The matched controls will have similar age and cultural and socio-economic background as the cases.

When investigating the effect of some occupational exposure on certain type of cancer, we select controls from the siblings or close relatives of the case who do not have that type of cancer. The matched controls will have similar lifestyle, culture, and genetic predispositions for the disease as the case.

One should resist the temptation to match on too many potential confounders at once. Strict matching on too many variables can make finding suitable controls almost impossible. Loose matching would sacrifice the main advantages of the matched design. Also, too much effort to match controls closely to the cases could result in inadvertently matching on the exposures (*over-matching*). If the matched controls have similar exposures as the cases, the data contain very little information on the exposure effects and the effects cannot be reliably estimated.

The matched case-control design can be very powerful if conducted thoughtfully and responsibly. It is undoubtedly one of the most efficient and bias-resistant study designs for epidemiological research. However, it is also somewhat dangerous – obtaining matched controls may not be as easy as expected and over-matching on exposures can make the collected data

Table 4.1.: Observed frequencies in a pair-matched case-control study with a single binary exposure.

Pair #: k	$X = 1$ (Exposed)	$X = 0$ (Unexposed)	Total
Cases ($Y = 1$)	n_{1k}^E	n_{1k}^U	$n_{1k} = 1$
Controls ($Y = 0$)	n_{0k}^E	n_{0k}^U	$n_{0k} = 1$
Total	n_k^E	n_k^U	$n_k = 2$

Table 4.2.: Four possible outcomes of one matched pair with a single binary exposure.

(a)

	E	U	Σ
Case	1	0	1
Cont.	1	0	1
Σ	2	0	2

n_{11} tables

(b)

	E	U	Σ
Case	1	0	1
Cont.	0	1	1
Σ	1	1	2

n_{10} tables

(c)

	E	U	Σ
Case	0	1	1
Cont.	1	0	1
Σ	1	1	2

n_{01} tables

(d)

	E	U	Σ
Case	0	1	1
Cont.	0	1	1
Σ	0	2	2

n_{00} tables

worthless.

4.2. Classical methods for matched case-control studies

In this section, we again consider just the simplest possible situation: a pair-matched case-control design with a single binary exposure. We keep the same notation as in Chapter 2 (see Table 4.1), k indexes the pairs/strata. The number of pairs K is equal to the number of available cases because each stratum includes exactly one case.

In a pair-matched design, the contingency table displayed in Table 4.1 has only four possible outcomes, see Table 4.2. The first outcome (a) and the last outcome (d) in Table 4.2 are concordant outcomes: both the case and the control have the same exposures (either both exposed or both unexposed). These outcomes do not contain any information about the association between the exposure and the disease. All that information is included in the other, discordant outcomes: in outcome (b), the case is exposed and the control is unexposed, in outcome (c) it is vice versa. The whole dataset can be summarized without any loss of information as the numbers of outcomes of each of the four kinds that appeared in the K matched pairs. Let n_{11} be the number of pairs with outcome (a), n_{10} the number of pairs with outcome (b), n_{01}

the number of pairs with outcome (c), and n_{00} the number of pairs with outcome (d). The sum $n_{11} + n_{10} + n_{01} + n_{00}$ gives the total number of pairs K.

Estimation and testing is based on the numbers of discordant tables n_{10} and n_{01}. Recall the notation from Section 2.1 but make it pair-dependent: $\pi_{E|1}^{(k)} = \mathrm{P}\left[X = 1 \middle| Y = 1, \text{pair } k\right]$ for the probability that the case in the k-th pair is exposed, and $\pi_{E|0}^{(k)} = \mathrm{P}\left[X = 1 \middle| Y = 0, \text{pair } k\right]$ for the probability that the control in the k-th pair is exposed. Let the exposure status of the case and the control (conditionally on the pair) be independent.

The odds ratio for disease among the exposed relative to the unexposed in the k-th pair can be expressed as

$$\theta = \frac{\pi_{E|1}^{(k)}(1 - \pi_{E|0}^{(k)})}{(1 - \pi_{E|1}^{(k)})\pi_{E|0}^{(k)}}.$$

We assume that the odds ratio is the same for all pairs (no interaction between the exposure and the pair), so it is expressed as a parameter θ that does not depend on k.

The probability that the case is exposed and the control is unexposed in the k-th stratum is (by conditional independence) $\pi_{E|1}^{(k)}(1 - \pi_{E|0}^{(k)})$. This is the probability that the pair will generate outcome (b). The probability that the control is exposed and the case is unexposed in the k-th stratum is $(1 - \pi_{E|1}^{(k)})\pi_{E|0}^{(k)}$. This is the probability that the pair will generate outcome (c).

Summarizing these steps, if we get a discordant table, where either the case of the control (but not both) is exposed, we can express the conditional probability that the case is exposed (the outcome is (b)) as

$$\frac{\pi_{E|1}^{(k)}(1 - \pi_{E|0}^{(k)})}{\pi_{E|1}^{(k)}(1 - \pi_{E|0}^{(k)}) + (1 - \pi_{E|1}^{(k)})\pi_{E|0}^{(k)}} = \frac{\theta}{\theta + 1} \equiv \pi.$$

Because θ is the same in all pairs, the probability π that the case is exposed in a discordant table is also the same in all pairs.

The pairs are independent, so the distribution of n_{10} given the total number $n_{10} + n_{01}$ of discordant tables is $\mathrm{Bi}(n_{10} + n_{01}, \pi)$ – this is the number of successes observed in $n_{01} + n_{01}$ independent binary experiments with success probability π. This finding is the key to making statistical inference about the odds ratio.

In the conditional binomial model, the MLE of π is $\widehat{\pi} = \frac{n_{10}}{n_{10}+n_{01}}$. Inverting $\pi = \frac{\theta}{\theta+1}$, we get $\theta = \frac{\pi}{1-\pi}$ and the conditional MLE of θ is

$$\widehat{\theta} = \frac{\widehat{\pi}}{1 - \widehat{\pi}} = \frac{n_{10}}{n_{01}}. \tag{4.1}$$

The odds ratio in a pair-matched case-control study can be estimated by the ratio of the number of pairs where the case is exposed and the control is unexposed to the number of pairs with the opposite exposure status.

We could also calculate the Mantel-Haenszel estimator using equation (3.4), with strata sizes $n_k = 2$ for all k. It turns out that the Mantel-Haenszel estimator is the same as the conditional MLE (4.1).

Let us turn attention to the hypothesis of no exposure effect, $H_0 : \theta = 1$. This is equivalent to $H_0 : \pi = \frac{1}{2}$. Using the conditional MLE $\widehat{\pi} = \frac{n_{10}}{n_{10}+n_{01}}$ of the probability π and the standard central limit theorem, we get

$$\sqrt{n_{10} + n_{01}}\, \frac{\widehat{\pi} - 1/2}{\sqrt{1/4}} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1)$$

under H_0. The left-hand side can be rewritten as

$$\frac{2n_{10} - (n_{10} + n_{01})}{\sqrt{n_{10} + n_{01}}} = \frac{n_{10} - n_{01}}{\sqrt{n_{10} + n_{01}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1)$$

and hence

$$\frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \xrightarrow{\mathrm{D}} \chi^2_1$$

under H_0. We reject $H_0 : \theta = 1$ if and only if

$$\chi^2_{\mathrm{MN}} \equiv \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \geq \chi^2_1(1 - \alpha).$$

This is the test known in classical statistics as the McNemar test (Anděl 2002, Chapter 13.6). It was originally developed as a paired test of equality of probabilities.

Let us evaluate the Mantel-Haenszel test statistic (3.3). We have $n_{1k} = n_{0k} = 1$, $n_k = 2$,

$$m_k = \mathrm{E}\, n^E_{1k} = n_{1k} \frac{n^E_k}{n_k} = \frac{n^E_k}{2},$$

and

$$V_k = \mathrm{var}\, n^E_{1k} = \frac{n_{1k} n_{0k} n^E_k n^U_k}{n^2_k (n_k - 1)} = \frac{n^E_k n^U_k}{4}.$$

In concordant tables, (n^E_k, n^U_k) are either $(0, 2)$ or $(2, 0)$, $n^E_{1k} = m_k$ and $V_k = 0$. The contributions of concordant tables to both the numerator and the denominator of the Mantel-Haenszel statistic are zero. Discordant tables have $n^E_k = n^U_k = 1$, $m_k = 1/2$ and $V_k = 1/4$. Thus, the Mantel-Haenszel test statistic is (the summation goes over discordant tables)

$$\chi^2_{\mathrm{MH}} = \frac{\left[\sum (n^E_{1k} - 1/2)\right]^2}{\sum 1/4} = \frac{\left(n_{10} - \frac{n_{10} + n_{01}}{2}\right)^2}{\frac{n_{10} + n_{01}}{4}} = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}}$$

and we end up with the McNemar test again.

Confidence intervals for θ can be obtained by a transformation of confidence intervals for π calculated by any of the standard methods (Wald, Wilson, Clopper-Pearson, etc.).

The classical methods explained in this section can be applied only to the simplest pair-matched designs. To a certain extent, they can be generalized to handle $1 : m$ matching with $m > 2$, variable numbers of controls per case, or multi-level exposures. However, these generalizations are complicated and do not share the simplicity and beauty we have enjoyed here. A much more flexible alternative for analyzing matched case-control studies in general can be obtained by a modification of logistic regression explained in the following section.

4.3. Conditional logistic regression for matched case-control studies

Consider a matched case-control study with K cases, and m_k matched controls for the k-th case, $k = 1, \ldots, K$. We allow an arbitrary variable number of controls – matched pairs are a special case with $m_k = 1$ for each k. Denote the total number of sampled controls by $M = \sum m_k$; the total number of study subjects is $K + M$.

Form a covariate vector $\boldsymbol{X}$ that includes all exposures, and confounders that were not used for matching (after a suitable transformation). Suppose that the probability of disease follows the following logistic regression model:

$$\mathrm{P}\big[Y = 1 \,\big|\boldsymbol{X}, C = k\big] = \frac{\mathrm{e}^{\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{X}}}{1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}^\top \boldsymbol{X}}}, \quad k = 1, \ldots, K. \tag{4.2}$$

This model includes a separate intercept for each stratum (which includes a case together with his matched controls). There is no stratum-by-exposure interaction (because it cannot be consistently estimated from the matched data anyway). The number of parameters in this model tends to ∞ as the number of cases K increases and the MLE's are inconsistent – not only $\widehat{\alpha}_k$ but even $\widehat{\boldsymbol{\beta}}$, see Breslow and Day (1980, Sec. 7.1) for an example.

However, consistent estimates of $\boldsymbol{\beta}$ can be obtained by an appropriate conditioning. We know that there is a single case in each stratum. Take a single stratum and take a subject with an observed covariate vector $\boldsymbol{x}_0$. Denote observed covariate vectors of the m_k remaining subjects in the stratum by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m_k}$. Now calculate the conditional probability that the subject with covariates $\boldsymbol{x}_0$ is the sole case in the stratum given the covariate vectors of all the other subjects and assuming the validity of the logistic regression model (4.2). To facilitate the notation, denote by D the event that the considered subject is a case and by $\overline{D}$ the event that the subject is a control. Then the desired conditional probability is in fact the probability that the covariates of the diseased subject are $\boldsymbol{x}_0$ rather than any the other covariate values observed in the stratum,

i.e.,

$$\frac{\mathrm{P}\left[X = \boldsymbol{x}_0 \,\middle|\, D\right] \prod_{i=1}^{m_k} \mathrm{P}\left[X = \boldsymbol{x}_i \,\middle|\, \overline{D}\right]}{\mathrm{P}\left[X = \boldsymbol{x}_0 \,\middle|\, D\right] \prod_{i=1}^{m_k} \mathrm{P}\left[X = \boldsymbol{x}_i \,\middle|\, \overline{D}\right] + \sum_{l=1}^{m_k} \left(\mathrm{P}\left[X = \boldsymbol{x}_0 \,\middle|\, \overline{D}\right] \mathrm{P}\left[X = \boldsymbol{x}_l \,\middle|\, D\right] \prod_{i \neq l} \mathrm{P}\left[X = \boldsymbol{x}_i \,\middle|\, \overline{D}\right] \right)}. \tag{4.3}$$

The numerator expresses the probability of the only constellation that leads to the desired result, the denominator sums the probabilities of all possible constellations, respecting that the case-control sampling scheme generates covariate values conditionally on disease status. For any $i = 0, \ldots, m_k$ we have

$$\mathrm{P}\left[X = \boldsymbol{x}_i \,\middle|\, D\right] = \frac{\mathrm{P}\left[D \,\middle|\, X = \boldsymbol{x}_i\right] \mathrm{P}\left[X = \boldsymbol{x}_i\right]}{\mathrm{P}(D)}$$

and

$$\mathrm{P}\left[X = \boldsymbol{x}_i \,\middle|\, \overline{D}\right] = \frac{\mathrm{P}\left[\overline{D} \,\middle|\, X = \boldsymbol{x}_i\right] \mathrm{P}\left[X = \boldsymbol{x}_i\right]}{1 - \mathrm{P}(D)}$$

where all the probabilities on the right-hand sides are stratum-specific. When we plug these expressions into (4.3), the numerator and each summand in the denominator all include

$$\mathrm{P}(D)[1 - \mathrm{P}(D)]^{m_k} \prod_{i=0}^{m_k} \mathrm{P}\left[X = \boldsymbol{x}_i\right]$$

so these terms cancel each other throughout. Thus, the probability (4.3) can be expressed in terms of conditional disease probabilities as follows:

$$\frac{\mathrm{P}\left[D \,\middle|\, X = \boldsymbol{x}_0\right] \prod_{i=1}^{m_k} \mathrm{P}\left[\overline{D} \,\middle|\, X = \boldsymbol{x}_i\right]}{\mathrm{P}\left[D \,\middle|\, X = \boldsymbol{x}_0\right] \prod_{i=1}^{m_k} \mathrm{P}\left[\overline{D} \,\middle|\, X = \boldsymbol{x}_i\right] + \sum_{l=1}^{m_k} \left(\mathrm{P}\left[\overline{D} \,\middle|\, X = \boldsymbol{x}_0\right] \mathrm{P}\left[D \,\middle|\, X = \boldsymbol{x}_l\right] \prod_{i \neq l} \mathrm{P}\left[\overline{D} \,\middle|\, X = \boldsymbol{x}_i\right] \right)}. \tag{4.4}$$

Plug in the logistic regression model (4.2) for each of the conditional disease probabilities. In the numerator of (4.4), this gives

$$\mathrm{P}\left[D \,\middle|\, X = \boldsymbol{x}_0\right] \prod_{i=1}^{m_k} \mathrm{P}\left[\overline{D} \,\middle|\, X = \boldsymbol{x}_i\right] = \frac{\mathrm{e}^{\alpha_k + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{x}_0}}{\prod_{i=0}^{m_k} \left(1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{x}_i}\right)}.$$

The expression $\left[\prod_{i=0}^{m_k} \left(1 + \mathrm{e}^{\alpha_k + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{x}_i}\right)\right]^{-1}$ appears not only in the numerator of (4.4) but also in each term in the denominator, so it cancels out again.

After all this development, the conditional probability that the subject with covariates $\boldsymbol{x}_0$ is the case given the covariate vectors of all the other subjects in the stratum can be written as

$$\frac{\mathrm{e}^{\alpha_k+\boldsymbol{\beta}^\top \boldsymbol{x}_0}}{\sum_{i=0}^{m_k} \mathrm{e}^{\alpha_k+\boldsymbol{\beta}^\top \boldsymbol{x}_i}} = \frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{x}_0}}{\sum_{i=0}^{m_k} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}},$$

where, at the last step, we were able to cancel the stratum-specific intercept α_k and ended up with an expression that only depends on $\boldsymbol{\beta}$ and is the same for all strata. This is taken as the likelihood contribution from a single stratum. Because the strata are independent from each other, the conditional likelihood is the product of such terms. Let $\boldsymbol{X}_{ki}$ be the covariate vector for the ith subject in the k-th stratum ($i = 0$ is the case). The conditional likelihood is

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{k0}}}{\sum_{i=0}^{m_k} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ki}}} = \prod_{j=1}^{K+M} \left(\frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_j}}{\sum_{l=1}^{K+M} Y_{lj} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_l}} \right)^{\delta_j},$$

where $Y_{lj} = 1$ if subjects j and l belong to the same stratum, and $Y_{lj} = 0$ otherwise, and $\delta_j = 1$ if the subject j is a case, and $\delta_j = 0$ if the subject j is a control. The last expression makes clear that the likelihood for the conditional logistic model has the same form as the partial likelihood of the Cox model, with δ_j playing the role of the failure indicator and Y_{lj} playing the role of the at-risk indicator.

Therefore all the developments for the conditional logistic model (score function, likelihood equations, observed and expected information, asymptotic properties, consistency and asymptotic normality) can be deduced from the results known for the Cox model. Asymptotic theory for the conditional logistic regression model is actually simpler because the likelihood terms are independent of each other (unlike in the Cox model). To perform the calculation on data, only slight modifications of software designed for the Cox model is needed. In R, this is done by the function `clogit` in library `survival`.

Now consider the conditional likelihood for the special case of matched pairs. We get

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{k0}}}{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{k0}} + \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{k1}}} = \prod_{k=1}^{K} \frac{\mathrm{e}^{\boldsymbol{\beta}^\top (\boldsymbol{X}_{k0}-\boldsymbol{X}_{k1})}}{1 + \mathrm{e}^{\boldsymbol{\beta}^\top (\boldsymbol{X}_{k0}-\boldsymbol{X}_{k1})}},$$

which is the likelihood of a logistic regression model with responses $Y_1 = \cdots = Y_K = 1$, no intercept term, and covariates taken as differences between the covariates of the case and the matched control. When we set up the data set and the model according to this prescription, we can estimate the parameters of conditional logistic regression using standard logistic regression software. All results generated by the logistic regression fit (standard errors, test statistics, confidence intervals) are valid.

The last form of the likelihood also makes obvious that if a covariate attains the same value for the case as for the control, the pair does not contribute to the estimation of that covariate's parameter.

Supplementary reading

- Breslow and Day (1980, Chap. V, pp. 162–169, Chap. VII, pp. 248–268)

5. Cohort Studies

5.1. Cohort study design

Cohort studies use random samples from the target population (or, potentially, exposure-dependent samples). With rare diseases, very large sample sizes are needed to come up with a sufficient number of cases. Cohort studies are usually conducted in a prospective way. At the start of a prospective cohort study, a large cohort of participants is enrolled. All of them are followed for a sufficient time, incident cases are captured and exposures are recorded. Prospective cohort studies take a long time to complete and are quite expensive.

A great advantage of prospective cohort studies is the possibility to reduce some biases that plague case-control studies. Case ascertainment can be done in a standardized and reliable way, reducing the potential for misdiagnosis or misclassification. Exposure and confounder records can be obtained with better precision and reliability. Biases caused by an inappropriate selection of controls are avoided.

Cohort studies can also be done retrospectively, by sampling from databases or existing records where the necessary information had been captured. Such studies are sometimes combined with a case-control subsample to reduce the cost of covariate measurement. Suppose that a retrospective study included a cohort of 10,000 subjects, of which 250 acquired the disease of interest. Suppose further that basic demographic and follow-up data as well as some imperfect information on exposures and confounders is available for all 10,000 participants. Obtaining precise exposure information for all 10,000 people would be extremely time-consuming and expensive. So, instead, this "ideal" information is only collected on the 250 cases and a random sample of a comparable size (say, also 250) from the controls. Then, the data available for the large cohort is combined with the detailed assessment collected on the case-control subsample. Such designs are called *case-cohort design* or *nested case-control study*. Statistical methods for the analysis of such designs have been developed but they are out of the scope of this course.

5.2. Models for ungrouped cohort data

Prospective cohort studies satisfy the i.i.d. assumption of standard statistical methods and usually include individual-level data on follow-up, exposures, and occurrence of the disease of interest. Thus, they can be analyzed by standard survival analysis methods.

The follow-up data include the entry time E_i (usually we work with age as the time scale, so this would be interpreted as age at the entry into the study), the exit time X_i and disease

outcome δ_i for participants $i = 1, \ldots, n$. The exposures and confounders are collected into the covariate vector $\boldsymbol{Z}_i(t)$, which may have time-varying components. The covariate vector is a random process observed in the interval (E_i, X_i).

The follow-up data can be transformed into the counting process $N_i(t) = \mathbb{1}(X_i \leq t, \delta_i = 1)$ and the at-risk process $Y_i(t) = \mathbb{1}(E_i \leq t \leq X_i)$. The theory of counting process martingales is available to justify the properties of various statistical methods.

The Cox model

The most commonly used model for the analysis of ungrouped cohort data is the *Cox model*, of course. The incidence rate at the time (age) t is expressed as

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t)\mathrm{e}^{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}(t)},$$

where $\lambda_0(t)$ is the incidence rate for a subject with zero exposures and confounders (an arbitrary unknown hazard function) and e^{β_j} is the relative risk associated with a unit increase in the j-th covariate (while keeping other exposures and confounders unchanged).

The time-varying components of $\boldsymbol{Z}(t)$ are supposed to affect the incidence through their current value. It is the responsibility of the analyst to transform the observed histories of exposures and confounders so that the resulting model makes sense for the disease of interest. There are various possibilities how to summarize exposure histories: one can use cumulative exposures over the whole lifetime, cumulative exposures over some period of time (10 years back), lagged cumulative exposures, where exposures acquired for some time prior to the current age are ignored, some moving averages, smoothed trajectories, etc. The choice depends on how the exposure is believed to affect the disease: immediate vs. cumulative effect, expected duration of increased risk after exposure, expected duration of latent time between exposure and disease diagnosis, etc. The Cox model allows a great flexibility in transformations of time-varying covariates, and with sufficient data, one can determine the correct form by testing significance of alternative transformations.

The basic form of the Cox model requires the proportional hazards assumption, i.e., the relative risk of the exposures and confounders must not depend on age. However, this assumption can be relaxed and tested by the inclusion of interactions between exposures/confounders and age. For example, one can factorize age into several age groups and allow different relative risks in each age group.

General relative risk models

The Cox model can be extended by considering other functions to express the influence of the linear predictor on the hazard function. Let

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t)r(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}(t)),$$

where r is a known sufficiently smooth (twice continuously differentiable) strictly increasing function such that $r(0) = 1$. This model is called *the general relative risk model*. The function r may be called *the relative risk function*, it plays a role of an inverse link function in the GLM terminology. The Cox model is a special case with $r(x) = \mathrm{e}^x$. Another useful relative risk function is $r(x) = 1 + x$. The resulting model has the form

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t)(1 + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}(t)).$$

This is called *the additive relative risk (ARR) model*. With a single exposure z, the relative risk with respect to an unexposed subject is linear in z: $RR(z) = 1 + \beta z$. Thus, the relative risk of an ARR model increases much more slowly than with a Cox model, where $RR(z) = e^{\beta z}$. The additive relative risk model is used, e.g., for modeling radiation effects on cancer incidence – it is believed to express the true risk better than the Cox model in this case.

Consider what happens when you fit the Cox model with a shifted log-transformed exposure $\log(a + z)$*, where* a *is some positive constant. What is the functional form of* $RR(z)$ *then?*

The additive relative risk model has a problem when $\beta Z < -1$*. However, if we are convinced that the exposure* Z *cannot have a protective effect (*$\beta \geq 0$*) then the problem cannot occur.*

The analysis of the general relative risk model is based on partial likelihood. The calculations follow the same route as those for the Cox model, only some of the expressions are more complicated. The partial likelihood is

$$L_P(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{r(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}_i(X_i))}{\sum_{j=1}^{n} Y_j(X_i) r(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}_j(X_i))} \right)^{\delta_i}.$$

Denote

$$\begin{aligned}
u(x) &= \log r(x), \\
u'(x) &= \frac{d \log r(x)}{dx}, \\
S^{(0)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{i=1}^{n} Y_i(t) r(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}_i(t)), \\
S^{(1)}(\boldsymbol{\beta}, t) &= \frac{\partial S^{(0)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}}, \\
S^{(2)}(\boldsymbol{\beta}, t) &= \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \boldsymbol{Z}_i(t)^{\otimes 2} [u'(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}_i(t))]^2 r(\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}_i(t)), \\
\overline{\boldsymbol{Z}}(\boldsymbol{\beta}, t) &= \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}.
\end{aligned}$$

The score statistic is

$$\boldsymbol{U}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \left[Z_i(t) u'(\boldsymbol{\beta}^\mathsf{T} Z_i(t)) - \overline{Z}(\boldsymbol{\beta}, t) \right] dN_i(t),$$

the estimator solves the system of equations $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}, \tau) = \mathbf{0}$, and the estimated Fisher information matrix has the form

$$\widehat{I}(\widehat{\boldsymbol{\beta}}, t) = \frac{1}{n} \int_0^t \left[\frac{S^{(2)}(\widehat{\boldsymbol{\beta}}, t)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, t)} - \overline{Z}(\widehat{\boldsymbol{\beta}}, t)^{\otimes 2} \right] d\overline{N}(t).$$

Details and proofs of consistency and asymptotic normality are provided in Prentice and Self (1983). Unfortunately, the observed information matrix is complicated and not necessarily positive definite at all $\boldsymbol{\beta}$. Therefore, the likelihood need not be concave and we need to pay attention to finding the global maximum when solving the likelihood equations.

It is also possible to introduce general relative risk models that combine various relative risk functions used for subsets of covariates. Suppose the covariate vector Z is decomposed into two parts $Z^\mathsf{T} = (Z_1^\mathsf{T}, Z_2^\mathsf{T})$ and the regression parameter $\boldsymbol{\beta}$ is divided correspondingly into $\boldsymbol{\beta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T})$. Then define

$$r(\boldsymbol{\beta}^\mathsf{T} Z) = r_1(\boldsymbol{\beta}_1^\mathsf{T} Z_1) r_2(\boldsymbol{\beta}_2^\mathsf{T} Z_2),$$

where $r_1(x) = \mathrm{e}^x$ and $r_2(x) = 1 + x$. Then, Z_1 affects the RR exponentially, Z_2 affects the RR additively, and their respective effects are multiplied.

Additive hazards model

Lin and Ying (1994) studied the additive model

$$\lambda(t \mid Z) = \lambda_0(t) + \boldsymbol{\beta}^\mathsf{T} Z(t).$$

This is called *the additive hazards model*. Here, $\lambda_0(t)$ is again the incidence rate for a subject with zero exposures and confounders (an arbitrary unknown hazard function) and covariates act additively on this function. The parameter β_j can be interpreted as the expected number of cases that will occur per unit of time if the corresponding covariate is increased by 1. Thus, the model is useful when the interest is in estimating *excess risks* rather than relative risks.

Lin and Ying (1994) proposed an estimator for this model based on the method of moments and proved its consistency and asymptotic normality. The estimator is not efficient but it is very simple (it can be even expressed by an explicit formula and is easy to calculate).

5.3. Models for grouped cohort data

Cohort data are sometimes available in a grouped format, which is unsuitable for analysis by the models mentioned in the previous section. Those models all require individual-level data on follow-up, exposures and occurrence of failures.

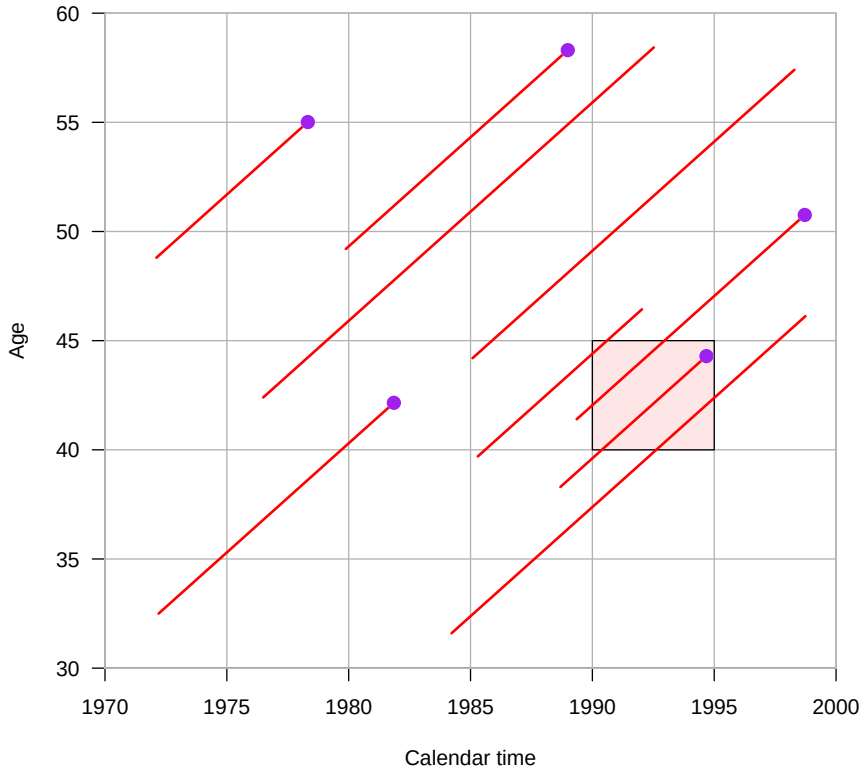

Figure 5.1.: Lexis diagram recording the development of age during follow-up. Purple dots denote observed events.

The setup of grouped cohort data can be explained on a Lexis diagram, which provides a graphical representation of the follow-up of subjects over time. Figure 5.1 shows an example of a Lexis diagram showing development of age during time. (Other examples of a Lexis diagram are in [BD1], Fig. 2.2 on p. 48 and in [BD2], Fig. 2.1 on p.50.) In Figure 5.1, calendar time is on the horizontal axis, the vertical axis represents age. Individual subjects are plotted as red line segments starting at the time of entry and ending at the time of exit. The cases are indicated by purple dots at exit times. Because age increases at the same rate as calendar time, all the line segments are running parallel at an angle of 45° with the horizontal axis.

Now divide the calendar time axis into intervals, e.g. every 5 years, and do the same with the age axis. This divides the diagram into rectangular cells (squares if the age and time intervals have the same length). Approximate the incidence in the (i, j)-th cell by a constant λ_{ij}. Consider a particular cell, for example year 1990–1995 and age 40–45, as highlighted in Figure 5.1. The total follow up time u_{ij} in this cell is equal to total length of projections of

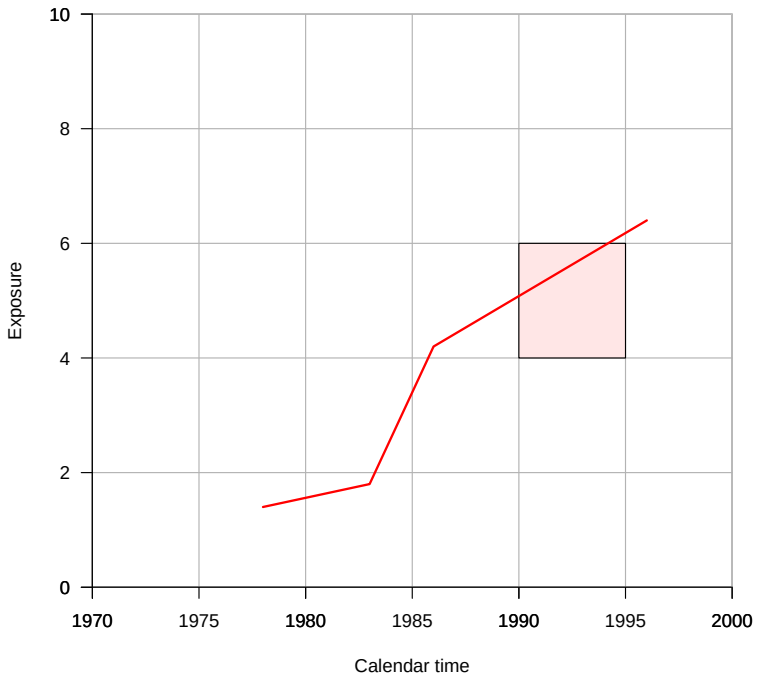

Figure 5.2.: Lexis diagram recording the development of a time-varying exposure during the follow-up of a single subject.

the red segments (restricted within the cell) onto the horizontal axis; this is the total time spent by all subjects withing the (i, j)-th cell. Then count the number of cases (purple dots) observed within the cell and denote it by n_{ij}. If the incidence is constant within the cell, then $n_{ij} \sim \text{Po}(u_{ij}\lambda_{ij})$ and the counts from different cells are independent of each other.

The cell-specific incidence λ_{ij} can be estimated empirically using the methods from Sec. 1.4, i.e. $\widehat{\lambda}_{ij} = n_{ij}/u_{ij}$. But we want to separate calendar time effects from age effects. We can do that by fitting a Poisson loglinear model with the observed cell counts n_{ij} as the response and offset $\log u_{ij}$. The main effects of that model estimate the log relative risks of the i-th (j-th) level of calendar time (age) relative to the first level (the baseline). This approach is related to fitting loglinear models to contingency tables as explained in the course on Advanced Regression Models and also to using loglinear models to estimate covariate effects on intensities of Poisson processes observed for variable duration of time. See the course notes for Advanced Regression, Sec. 3.2 and 3.3.

We can create the Lexis diagram for any exposure, not just age. If the exposure is constant in time (gender, race, highest education achieved before the entry), the path of a subject through the diagram is horizontal. An example of a Lexis diagram for a time-varying exposure (e.g,

cumulative alcohol consumption) is shown in Figure 5.2. Here, the subject had different trends in the exposure during different intervals. Again, we can calculate the total follow-up time in each cell by summing projections of the red segments onto the horizontal axis, count the number of cases in each cell and fit Poisson loglinear models as explained above.

In practice, a two-way classification is not sufficient because we need more than two factors to consider simultaneously. But the idea of Lexis diagram and accumulation of observed cases and follow-up time over the individual cells easily extends to multi-factor settings as well. Then we in fact perform analysis of a multi-dimensional contingency table via loglinear models with offsets.

Example:

A grouped cohort data may come, e.g., in the following format

Cases n	Total f-up u	Age group	Age midpoint	Cal. year	Cal. yr midpoint	Exposure group	Expos. midpoint
6	2132	35–40	37.5	1995–2000	1997.5	0–10	5.0
15	3759	55–60	57.5	2005–2010	2007.5	35–50	42.5
⋮	⋮	⋮					

The midpoints we assign to each interval are useful for fitting functional forms of relative risks rather that treating each interval as a separate category. For example, we may consider the following loglinear model for this data set:

$$\log m_{ijk} = \log u_{ijk} + \alpha + \beta_i + \gamma(c_j - 1995) + \delta \log(e_k - 4),$$

where m_{ijk} is the expected number of cases in the i-th level of age, j-th level of calendar time and k-th level of exposure, u_{ijk} is the total follow-up time in that cell, c_j is the midpoint of the j-th calendar time interval, and e_k is the midpoint of the k-th exposure interval. The parameters to be estimated are α, β_i for $i > 1$, γ and δ.

Since $m_{ijk} = \lambda_{ijk} u_{ijk}$, we have

$$\log \lambda_{ijk} = \alpha + \beta_i + \gamma(c_j - 1995) + \delta \log(e_k - 4).$$

The interpretation of the parameters is as follows:

- e^{α} is the incidence for the youngest age group ($i = 1$) in the year 1995 at exposure 5.
- e^{β_i} is the relative risk of age group i with respect to the youngest age group.
- $e^{10\gamma}$ is the proportional increase in the risk over 10 year time period.
- The relative risk of exposure level e_k relative to exposure 5 is

$$\frac{\lambda_{ijk}}{\lambda_{ij1}} = e^{\delta \log(e_k - 4)} = (e_k - 4)^{\delta},$$

this is a power function of the exposure midpoint.

This model does not include interactions between age and exposure, or any other pairs of variables. Such interactions could be added into the model and tested.

The loglinear model requires that there are enough data in each cell. The fitted cell counts $\hat{m}_{ijk}$ should be at least 5 for the asymptotics to work. Thus, fitting a model with finely categorized multiple exposures, confounders and other factors in a multi-way table requires a huge amount of data. One should be really careful not to fit these models to data that is too sparse.

The Poisson model can be used with other link functions to fit additive risks or excess relative risks. For example, with identity link function, we can fit the additive model

$$\lambda_{ijk} = \boldsymbol{\beta}^\top \boldsymbol{X}_{ijk},$$

where $\boldsymbol{X}_{ijk}$ is a suitably selected vector of covariates for the cell (i, j, k). For such models, the total follow-up time u_{ijk} is used as a "prior weight" in the GLM fitter rather than as an offset and the response is provided in the form of empirical incidence estimates n_{ijk}/u_{ijk}. The prior weight in GLM terminology is a known constant specific to each observation that divides the common dispersion parameter.

To summarize, the use of Poisson loglinear models for aggregated cohort data provides an interesting alternative to standard survival analysis methods. Of course, if the data comes in an individual format, we prefer the Cox model. An artificial aggregation would lose information. However, if the data arrive in an aggregated form, Poisson loglinear models are the only choice.

5.4. Discrete Cox model

Sometimes it is only the information on the timing of events that gets aggregated, while exposures and confounders are available for each individual separately. This may happen for two reasons:

- only calendar year (month etc.) of diagnosis is known, with no exact date;
- diagnosis is done by repeated diagnostic testing according to some schedule; we only know the date of the last negative test and the date of the first positive test (for the cases) but not the exact date the disease was acquired.

The second case is typical for infectious diseases. We never know the exact date when the infection was acquired.

With such data, we only know an interval when the event of interest occurred but not the time. This aggregates the original continuous distribution of T into a discrete distribution. The

standard Cox model is unsuitable for such data because it assumes a continuous distribution of the event time T.

There are modifications of the partial likelihood that allow treatment of ties in observed failure times (Efron, Breslow methods). They are good for occasional ties but not for purely discrete failure times. There is also so called "exact" method that would work for dicrete failure times but is computationally intensive.

The discrete Cox model is developed as follows. Consider a continuous random variable $T \geq 0$, the exact event time. Partition the positive half-line, the support of T, into r disjoint intervals $A_i = \langle a_{i-1}, a_i)$, $i = 1, \ldots, r$, where $0 = a_0 < a_1 < \cdots < a_{r-1} < a_r = \infty$. The intervals need not have the same length. Take discrete values $t_1 < \ldots < t_r$ and define a discretized version of T as

$$T^* = t_j \quad \text{when} \quad T \in A_j, \text{ i.e., } a_{j-1} \leq T < a_j.$$

This setup describes both the rounding issue and repeated diagnostic testing problem (however, the intervals must be the same for all subjects).

Suppose the original continuous event time T satisfies the Cox model

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}(t)}.$$

The covariates $\boldsymbol{Z}(t)$ may be time varying but we will assume that they are constant within each interval A_j, so that they can change values only at the times a_j. Denote $\boldsymbol{Z}(t) = \boldsymbol{z}_j$ for $t \in A_j$.

Now we will calculate the distribution of T^*. The survival function of T can be written as

$$S_T(t) = \exp\Big\{-\int_0^t \lambda_0(s) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}(s)} ds\Big\}.$$

Denote $p(t_k) \equiv \mathrm{P}\left[T^* = t_k\right]$ and write

$$\begin{aligned} p(t_k) = \mathrm{P}\left[T^* = t_k\right] = \mathrm{P}\left[a_{k-1} \leq T < a_k\right] = S_T(a_{k-1}) - S_T(a_k) = \\ = \exp\Big\{-\int_0^{a_{k-1}} \lambda_0(s) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}(s)} ds\Big\} \Big[1 - \exp\Big\{-\int_{a_{k-1}}^{a_k} \lambda_0(s) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_k} ds\Big\}\Big] \end{aligned}$$

Next, decompose the integral from 0 to a_{k-1} into the individual intervals.

$$\exp\Big\{-\int_0^{a_{k-1}} \lambda_0(s) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}(s)} ds\Big\} = \prod_{l=1}^{k-1} \exp\Big\{-\int_{a_{l-1}}^{a_l} \lambda_0(s) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_l} ds\Big\}.$$

Because $\boldsymbol{Z}$ is constant on each interval, the integrals only integrate the baseline hazard. Denote

$$\alpha_k \equiv \exp\Big\{-\int_{a_{k-1}}^{a_k} \lambda_0(s) ds\Big\}.$$

Plug it into the expression for $p(t_k)$ to get

$$p(t_k) = \mathrm{P}\left[T^* = t_k\right] = \left(\prod_{l=1}^{k-1} \alpha_l^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_l\}}\right)\left(1 - \alpha_k^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_k\}}\right).$$

Now express the survival function of the discretized T^*.

$$S(t_k) = \mathrm{P}\left[T^* > t_k\right] = \mathrm{P}\left[T > a_k\right] = \exp\left\{-\int_0^{a_k} \lambda_0(s) \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}(s)} ds\right\} = \prod_{l=1}^{k} \alpha_l^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_l\}}. \tag{5.1}$$

For a discrete failure time, the hazard function can be expressed as $\lambda(t_k) = p(t_k)/S(t_{k-1})$. In our case,

$$\lambda(t_k) = \frac{p(t_k)}{S(t_{k-1})} = 1 - \alpha_k^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_k\}}. \tag{5.2}$$

Consider the discretized data $(X_i^*, \delta_i, \boldsymbol{Z}_i(t))$, independent vectors for $i = 1, \ldots, n$. We have $X_i^* = \min(T_i^*, C_i^*)$, $\delta_i = \mathbb{1}(T_i^* \le C_i^*)$. The censoring variable C_i^* is conditionally independent of T_i^* given $\boldsymbol{Z}_i$ and its support is the same as the support of T_i^*, the discrete times $t_1 < \cdots < t_r$. Denote

$$p_i(t_k) = \mathrm{P}\left[T_i^* = t_k\right], \quad S_i(t_k) = \mathrm{P}\left[T_i > t_k\right], \quad \lambda_i(t_k) = \frac{p_i(t_k)}{S_i(t_{k-1})}.$$

Let $\boldsymbol{z}_{ik}$ be the value of $\boldsymbol{Z}_i(t)$ over the interval A_k. Denote by m_i the index of the time when X_i^* occurred: $t_{m_i} = X_i^*$.

The likelihood for the observed data is (see the notes for the course Censored Data Analysis, Theorem 2.2):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i(t_{m_i})^{\delta_i} S_i(t_{m_i})^{1-\delta_i} = \prod_{i=1}^{n} \lambda_i(t_{m_i})^{\delta_i} S_i(t_{m_i - 1})^{\delta_i} S_i(t_{m_i})^{1-\delta_i}.$$

Plug in the expressions for $S_i(t_k)$ from (5.1) and for $\lambda_i(t_k)$ from (5.2) to get

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{\left(1 - \alpha_{m_i}^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_{i m_i}\}}\right)^{\delta_i} \left(\prod_{l=1}^{m_i - 1} \alpha_l^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_{il}\}}\right)\left(\alpha_{m_i}^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_{i m_i}\}}\right)^{1-\delta_i}\right\}.$$

This likelihood can be transcribed as follows. For all $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ define

$$Y_{ij} = \begin{cases} 1 & \text{if } \delta_i = 1 \text{ and } j = m_i \\ 0 & \text{otherwise,} \end{cases}$$

and

$$p_{ij} = 1 - \alpha_j^{\exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{z}_{ij}\}}.$$

Then

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} p_{ij}^{Y_{ij}} (1 - p_{ij})^{1 - Y_{ij}},$$

which is the likelihood of independent alternative variables with success probability p_{ij}!

Now express α_j as $\exp\{-e^{\gamma_j}\}$. Then we can write

$$p_{ij} = 1 - \alpha_j^{\exp\{\boldsymbol{\beta}^\mathsf{T} z_{ij}\}} = 1 - \exp\{-e^{\gamma_j + \boldsymbol{\beta}^\mathsf{T} z_{ij}}\}.$$

Invert this to obtain the linear predictor.

$$\eta_{ij} = \gamma_j + \boldsymbol{\beta}^\mathsf{T} z_{ij} = \log(-\log(1 - p_{ij})) = g(p_{ij}),$$

where g is the complementary log-log link. The linear predictor includes the factorized discrete time in addition to the linear predictor from the original Cox model.

Thus, we can fit the discrete Cox model by binary data regression with complementary log-log link. These two models have the same likelihood and therefore all the expressions derived from the likelihood have the same form and share the same asymptotic behavior.

The data for fitting the binary model must be arranged in the following way in order to generate the correct likelihood. For the i-th subject, we create m_i rows (m_i is the value of the discrete censored Y_i^*).

Y_{ij}	t_{ij}	Z_{ij}
0	t_1	z_{i1}
0	t_2	z_{i2}
$\vdots$		$\vdots$
0	t_{m_i-1}	z_{i,m_i-1}
δ_i	t_{m_i}	z_{im_i}

After arranging the data in this way, we would call, e.g., in R:

```
glm(y~factor(time)+z1+z2+..., family=binomial, link=cloglog)
```

and use the usual functions to evaluate the results of the fit (`summary`, `anova`, `drop1`, etc.).

This adaptation of the Cox model was first proposed and investigated by Prentice and Gloeckler (1978). However, the connection to binary regression with the complementary log-log link escaped the attention of these authors. This idea can be found, e.g., in Allison (1982), who unfortunately did not provide detailed justification of the validity of this approach.

The fact that the complementary log-log link can be used for the analysis of censored data is occasionally mentioned in the literature on binary regression and GLM. However, details or clarifications are nowhere to be found.

Bookmark this chapter or learn it by heart. This is something you will not be able to find using Google or to discover by searching the Web of Science.

Supplementary reading

- Breslow and Day (1987, Chap. 3, pp. 82–91, Chap. 4, pp. 120–171, Chap. 5, pp. 178–197)

6. Diagnostic Tests

6.1. Diagnostic markers

In medicine, diseases are frequently diagnosed by measuring a level of a biomarker X, which is a continuous variable associated with the presence of the disease D. Suppose that patients who have the disease have on average higher levels of the biomarker and we want to use the biomarker to distinguish them from disease-free individuals. We select a threshold c for the biomarker and define a diagnostic test $T(c) = \mathbb{1}(X \geq c)$. Thus,

- when the marker X exceeds the threshold c then $T = 1$, the test is positive and the subject is diagnosed;
- when the marker X is under the threshold c then $T = 0$, the test is negative and the subject is considered free of the disease.

This is a classification problem. With a univariate observation X, there is no other reasonable classification rule than the threshold test $T(c)$. We assume that the conditional densities $f(x \mid D)$ (density of X among subjects with the disease) and $f(x \mid H)$ (density of X among

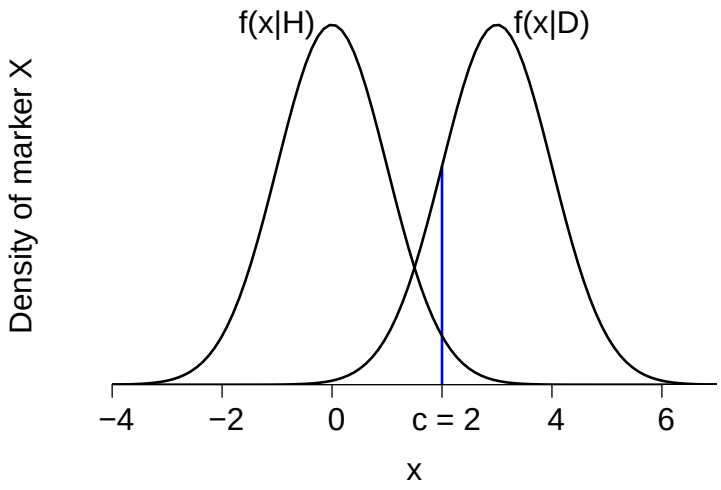

Figure 6.1.: Conditional densities of the biomarker in disease-free individuals (left curve) and those who have the disease (right curve). The threshold for diagnosis is set at the value 2.

healthy subjects) are different from each other, ideally, shifted in the location. An illustrative example is provided in Figure 6.1.

Considering the true disease status D or H and the test result $T(c)$, four different outcomes are possible. Their probabilities form the following table.

	Disease	Healthy	All
$T(c) = 1$ (positive test)	π_{D1}	π_{H1}	$\pi_{\cdot 1}$
$T(c) = 0$ (negative test)	π_{D0}	π_{H0}	$\pi_{\cdot 0}$
All	$\pi_{D\cdot}$	$\pi_{H\cdot}$	1

The probability $\pi_{D\cdot}$ is the true prevalence of the disease. All probabilities except $\pi_{D\cdot}$ and $\pi_{H\cdot}$ depend on the choice of the threshold c (though it is not expressed in the notation).

6.2. Sensitivity and specificity

Consider a subject with the disease and consider the probability that the test is positive. This is called *the sensitivity of the test*. We have

$$\text{Sensitivity:} \qquad \mathrm{P}\big[T = 1 \,\big|D\big] = \frac{\pi_{D1}}{\pi_{D\cdot}} \stackrel{\text{df}}{=} \eta \equiv \eta(c).$$

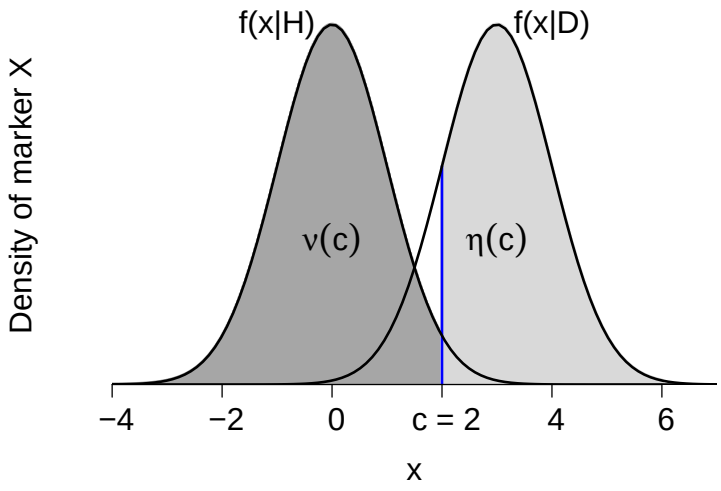

Figure 6.2.: Graphical illustration of sensitivity $\eta(c)$ (lighter shade) and specificity $\nu(c)$ (darker shade).

Now take a subject who is healthy and consider the probability that the test is negative. This is called *the specificity of the test*. We have

$$\text{Specificity:} \qquad \mathrm{P}\big[T = 0 \,\big|H\big] = \frac{\pi_{H0}}{\pi_{H\cdot}} \overset{\text{df}}{=} \nu \equiv \nu(c).$$

See also Figure 6.2. Sensitivity and specificity are probabilities of the two desired outcomes, so we would like the test to have a high sensitivity and a high specificity. This is not possible, of course, unless the conditional densities of the marker have a disjoint support allowing a perfect classification (and hence error-free diagnosis). In general, sensitivity and specificity depend on the threshold c. When we increase the threshold, we increase specificity but decrease sensitivity (and vice versa).

Sensitivity η is the rate of true positive tests. The probability $1 - \eta$ of the error that is made by obtaining a negative test result on a diseased subject is called *the false negative rate*. Specificity ν is the rate of true negative tests. The probability $1 - \nu$ of the error that is made by obtaining a positive test result on a healthy subject is called *the false positive rate*.

Sensitivity and specificity are the properties of the test, they are not affected by the population to which the test is applied (as long as the conditional distributions of the marker stay the same). Other important aspects of the practical performance of the test depend on the population through the prevalence $\pi_{D\cdot}$ of the disease. In particular, the probability that the subject who tested positive indeed has the disease is called *positive predictive value*. It is related to sensitivity, specificity and prevalence as follows.

$$\text{Positive predictive value:} \qquad \mathrm{P}\big[D \,\big|T = 1\big] = \frac{\pi_{D1}}{\pi_{\cdot 1}} = \frac{\eta\pi_{D\cdot}}{\eta\pi_{D\cdot} + (1-\nu)(1-\pi_{D\cdot})}.$$

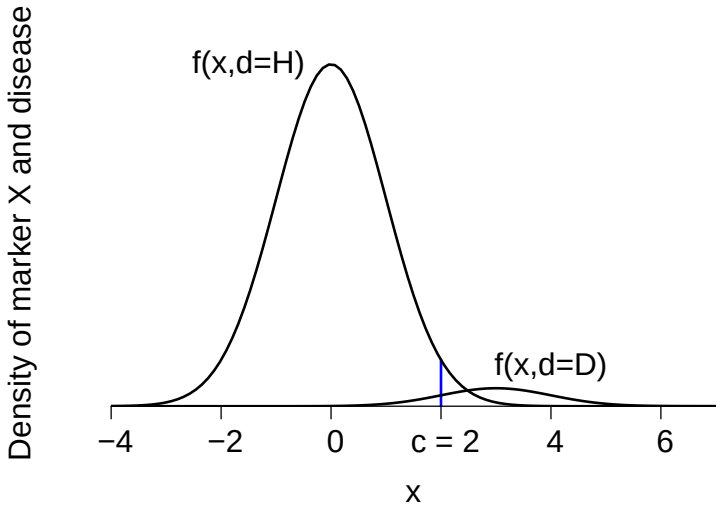

Figure 6.3.: Joint distribution of marker and disease, with disease prevalence 0.05.

Similarly, the probability that the subject who tested negative is indeed disease-free is called *negative predictive value.*

$$\text{Negative predictive value:} \qquad \mathrm{P}\big[H \,\big|\, T = 0\big] = \frac{\pi_{H0}}{\pi_{\cdot 0}} = \frac{v(1-\pi_{D\cdot})}{v(1-\pi_{D\cdot}) + (1-\eta)\pi_{D\cdot}}.$$

Because most diseases are rare (prevalence is small), negative predicitive value is usually quite close to 1. However, positive predictive value can be quite poor unless the specificity v is very close to one. This can be illustrated by plotting the joint distribution of the marker and disease (see Figure 6.3). This figure plots the marker distributions among the healthy and the diseased in the scale corresponding to the prevalence. It is clear that a large fraction of the subjects who test positive will be misdiagnosed healthy idividuals.

Simple example:
Suppose the population prevalence of Covid19 antibodies is 0.05. Antibodies are detected by a rapid test with sensitivity 0.99 and specificity 0.9. What fraction of the positive test results will truly have antibodies against Covid19?
We have $\pi_{D\cdot} = 0.05$, $\eta = 0.99$, $v = 0.9$. *Hence, we have positive predictive value*

$$\mathrm{P}\big[D \,\big|\, T = 1\big] = \frac{0.99 \cdot 0.05}{0.99 \cdot 0.05 + 0.1 \cdot 0.95} = \frac{0.0495}{0.1445} = 0.34.$$

Only about a third of the subjects detected by the test will have Covid19 antibodies. Two thirds will be false positives.

Diagnostic tests are sometimes compared using so called *diagnostic accuracy*, which is the probability that the test provides a correct classification. Diagnostic accuracy can be expressed as

$$\pi_{D1} + \pi_{H0} = \eta\pi_{D\cdot} + v\pi_{H\cdot}.$$

This is not a suitable criterion for comparing tests because it depends on the prevalence of the disease.

Estimation of these parameters is straightforward because they are all derived from a 2×2 contingency table and have a probability interpretation. It is enough to use empirical proportions to estimate these probabilities. Confidence intervals are also straightforward. However, one needs to know the true status of the subjects to estimate any of these parameters. If only the test results are available (or test results of two alternative tests), none of these characteristics can be evaluated.

There is no generally applicable rule for choosing the threshold c for diagnostic testing. It is all about the tradeoff between sensitivity and specificity. This consideration needs to take into account the consequences of obtaining a false positive or a false negative test result and weigh them against each other. It all depends on the disease to be diagnosed, available treatment options, purpose of the test. Imprecise tests are sometimes followed by more precise but more expensive or more aggresive diagnostic procedures.

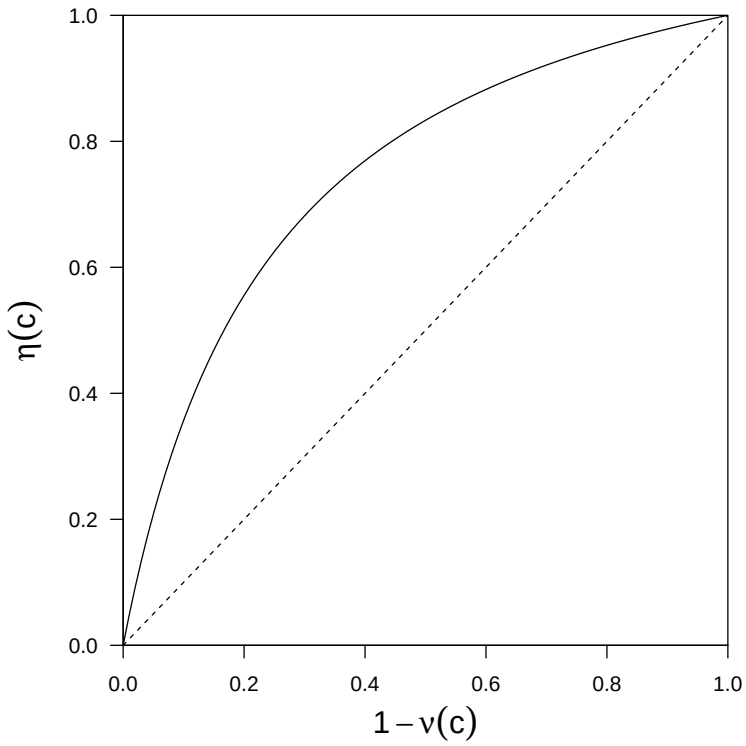

Figure 6.4.: Example ROC curve.

6.3. ROC curves

It is of advantage to describe the properties of a diagnostic test across many different threshold levels at once. This is done by the so called *ROC curve* (ROC is an acronym for *Receiver Operating Characteristic*). A ROC curve is a plot of sensitivity against the false positive rate (1 minus specificity) created for all values of $c \in (-\infty, \infty)$. The whole curve is contained in the unit square. For $c = -\infty$, the test always returns a positive result. Then, sensitivity is one, $\eta(-\infty) = 1$, specificity is zero, $\nu(-\infty) = 0$, and the ROC curve starts at the upper right corner of the graph. For $c = \infty$, the test never returns a positive result. Then, sensitivity is zero, $\eta(\infty) = 0$, specificity is one, $\nu(\infty) = 1$, and the ROC curve ends at the lower left corner of the graph (see Figure 6.4).

If the value of the marker X is independent of the disease status, the ROC curve runs along the diagonal of the unit square, sensitivity is equal to false positive rate because the result of the test does not depend on the disease status. Such test is worthless, of course. At the other extreme, the ROC of the perfect test shifts from $[1, 1]$ to $[0, 1]$ (specificity increases while sensitivity is still one) and then jumps to $[0, 0]$ (specificity is one and sensitivity decreases to zero). The upper left corner is only reachable with the ideal test at a c that provides a perfect

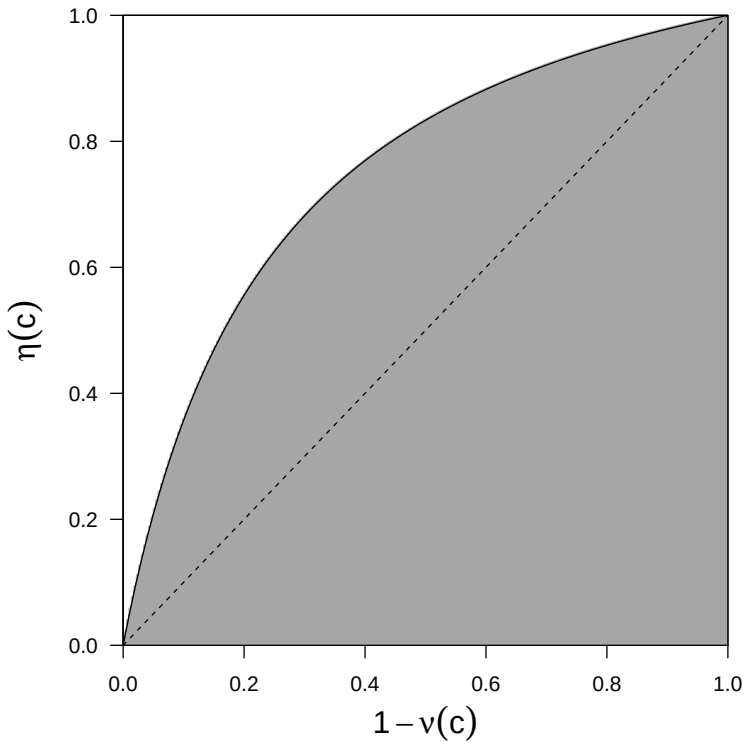

Figure 6.5.: Illustration of Area Under the Curve (AUC).

classification.

ROC curves are used to choose c that just provides the desired specificity (or sensitivity) and to compare the performance of two tests. They can be estimated from the data by taking empirical estimates of sensitivity and specificity over a range of values of c. Because observed data points are discrete, estimated ROC curves are not smooth but look like step functions.

Sometimes we want to express the performance of a test by a single number. A reasonable and common way to do that is to calculate the area under the ROC curve (AUC), see Figure 6.5. The ideal test has AUC = 1, the worthless test has AUC = 0.5. The closer the AUC to 1 the better, of course. The AUC can be used to compare even tests that have crossing ROC curves.

6.4. Diagnostic tests based on multiple markers

We can generalize our approach to diagnostic testing to diagnoses made from multiple markers as well. Involving multiple markers in the testing result can, of course, substantially reduce the errors.

The most straightforward generalization is based on logistic regression. Suppose we ob-

serve multiple markers arranged into a covariate vector $\boldsymbol{Z}$ together with the true disease status D on a sample of n independent subjects. We build a logistic model

$$\log \frac{\mathrm{P}\big[D\,\big|\boldsymbol{Z}\big]}{\mathrm{P}\big[H\,\big|\boldsymbol{Z}\big]} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{\beta}$$

by the usual methods and obtain the estimated coefficients $\widehat{\boldsymbol{\beta}}$. We consider the classification rule

$$T(c) = 1 \quad \text{if} \quad \widehat{\pi}(\boldsymbol{Z}) > c$$

for some $c \in (0, 1)$, where $\widehat{\pi}(\boldsymbol{Z})$ is the fitted value (estimated disease probability) from the logistic model. However, the estimated disease probabilities are increasing functions of the linear predictor $\boldsymbol{Z}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}$. Thus, we can take the linear rule

$$T(c) = 1 \quad \text{if} \quad \boldsymbol{Z}^{\mathsf{T}}\widehat{\boldsymbol{\beta}} > c,$$

where $c \in (-\infty, \infty)$. Then we can investigate sensitivity and specificity as functions of c, plot ROC curves, calculate AUC, all as before.

Alternatively, we can use any other classification or prediction procedure known from other areas of statistics: Bayesian classification, cluster analysis, principal components, classification trees, neural networks etc.

Bibliography

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories, *Sociological Methodology* **13**: 61–98.

Anděl, J. (2002). *Základy matematické statistiky*, Matfyzpress, Praha.

Breslow, N. and Day, N. (1980). *Statistical Methods in Cancer Research. Volume I: The Analysis of Case-Control Studies*, IARC Scientific Publication No. 32, International Agency for Research on Cancer, Lyon.

Breslow, N. and Day, N. (1987). *Statistical Methods in Cancer Research. Volume II: The Design and Analysis of Cohort Studies*, IARC Scientific Publication No. 82, International Agency for Research on Cancer, Lyon.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**(4): 417–451.

Esteve, J., Benhamou, E. and Raymond, L. (1994). *Statistical Methods in Cancer Research. Volume IV: Descriptive Epidemiology*, IARC Scientific Publication No. 128, International Agency for Research on Cancer, Lyon.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**(1): 61–71.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**(4): 719–748.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics* **34**: 57–67.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* **66**(3): 403–411.

Prentice, R. and Self, S. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form, *Annals of Statistics* **11**(3): 804–813.

Robins, J., Breslow, N. and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models, *Biometrics* **42**(2): 311–323.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies, *Biometrics* **47**(2): 497–510.

Woolf, B. (1955). On estimating the relation between blood group and disease, *Annals of Human Genetics* **19**(4): 251–253.

List of Tables

List of Figures

Index

A. Appendix

A.1. A universal approach to sample size and power calculation

The outcome Y will be observed in two independent iid samples for two treatments A and B:

$$Y_{A1}, \ldots, Y_{An_A} \quad \text{for treatment A} \quad \text{and} \quad Y_{B1}, \ldots, Y_{Bn_B} \quad \text{for treatment B.}$$

Let the treatment effect be expressed as the difference in the expectations of Y between treatments A and B: denote $\mu_A = \mathsf{E}\, Y_{Ai}$, $\mu_B = \mathsf{E}\, Y_{Bi}$, and $\theta = \mu_A - \mu_B$. We consider the two-sided test of the null hypothesis

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0.$$

Consider asymptotically normal estimators $\widehat{\mu}_A$, $\widehat{\mu}_B$ of the treatment-specific expectations that satisfy

$$\sqrt{n_A}(\widehat{\mu}_A - \mu_A) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \sigma_A^2),$$
$$\sqrt{n_B}(\widehat{\mu}_B - \mu_B) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \sigma_B^2)$$

when the null hypothesis $\mu_A = \mu_B$ is true.

The test statistic for the two-sample test is defined as*

$$Z = \frac{\widehat{\mu}_A - \widehat{\mu}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}.$$

Under H_0, $Z \xrightarrow{\mathsf{D}} \mathsf{N}(0, 1)$. The null hypothesis is rejected (at the asymptotic two-sided level α) if $|Z| > u_{1-\alpha/2}$.

Now let's calculate the power of this test under a fixed alternative $\mu_A - \mu_B = \theta > 0$. The variances of $\widehat{\mu}_A$ and $\widehat{\mu}_B$ might be changed under the alternative. To calculate the power, we will assume that

$$\sqrt{n_A}(\widehat{\mu}_A - \mu_A) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \tau_A^2),$$
$$\sqrt{n_B}(\widehat{\mu}_B - \mu_B) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \tau_B^2)$$

* The real statistic would include estimated variances. However, at the planning stage, these are not available and the variances must be replaced by sensible guesses.

when $\mu_A = \mu_B + \theta$. Thus, under this alternative,

$$\frac{\widehat{\mu}_A - \widehat{\mu}_B - \theta}{\sqrt{\frac{\tau_A^2}{n_A} + \frac{\tau_B^2}{n_B}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

We rewrite the test statistic Z as follows:

$$Z = \frac{\widehat{\mu}_A - \widehat{\mu}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \sqrt{\frac{\frac{\tau_A^2}{n_A} + \frac{\tau_B^2}{n_B}}{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \left(\underbrace{\frac{\widehat{\mu}_A - \widehat{\mu}_B - \theta}{\sqrt{\frac{\tau_A^2}{n_A} + \frac{\tau_B^2}{n_B}}}}_{\equiv U \xrightarrow{\mathrm{D}} \mathrm{N}(0,1)} + \frac{\theta}{\sqrt{\frac{\tau_A^2}{n_A} + \frac{\tau_B^2}{n_B}}} \right).$$

Now introduce $r = n_B/n_A$ and express Z as

$$Z = \underbrace{\sqrt{\frac{\tau_A^2 + \frac{\tau_B^2}{r}}{\sigma_A^2 + \frac{\sigma_B^2}{r}}}}_{\equiv \delta_\sigma} \left(U + \underbrace{\frac{\theta}{\sqrt{\tau_A^2 + \frac{\tau_B^2}{r}}}}_{\equiv \delta_\theta} \sqrt{n_A} \right) = \delta_\sigma (U + \delta_\theta \sqrt{n_A}).$$

Let us approximate the power

$$\begin{aligned} \mathrm{P}\left[\delta_\sigma \left| U + \delta_\theta \sqrt{n_A} \right| > u_{1-\alpha/2}\right] &= \underbrace{\mathrm{P}\left[\delta_\sigma (U + \delta_\theta \sqrt{n_A}) < -u_{1-\alpha/2}\right]}_{\approx 0} + \mathrm{P}\left[\delta_\sigma (U + \delta_\theta \sqrt{n_A}) > u_{1-\alpha/2}\right] \\ &\approx \mathrm{P}\left[U > \frac{1}{\delta_\sigma} u_{1-\alpha/2} - \delta_\theta \sqrt{n_A}\right] = 1 - \Phi\left(\frac{1}{\delta_\sigma} u_{1-\alpha/2} - \delta_\theta \sqrt{n_A}\right). \end{aligned}$$

So, the power for a given n_A, n_B and θ is done. We just need some idea about variability in both samples under the hypothesis and the alternative to calculate δ_σ and δ_θ.

How many observations do we need to achieve a power of at least $1 - \beta$? Solve the inequality

$$\begin{aligned} 1 - \Phi\left(\frac{1}{\delta_\sigma} u_{1-\alpha/2} - \delta_\theta \sqrt{n_A}\right) &> 1 - \beta \\ \Phi\left(\frac{1}{\delta_\sigma} u_{1-\alpha/2} - \delta_\theta \sqrt{n_A}\right) &< \beta \\ \frac{1}{\delta_\sigma} u_{1-\alpha/2} - \delta_\theta \sqrt{n_A} &< u_\beta \\ \delta_\theta \sqrt{n_A} - \frac{1}{\delta_\sigma} u_{1-\alpha/2} &> u_{1-\beta} \end{aligned}$$

$$\sqrt{n_A} > \frac{\frac{1}{\delta_\sigma} u_{1-\alpha/2} + u_{1-\beta}}{\delta_\theta}$$

$$n_A > \frac{\left(\frac{1}{\delta_\sigma} u_{1-\alpha/2} + u_{1-\beta}\right)^2}{\delta_\theta^2}.$$

This is the number of observations needed in arm A. In arm B, we need r times as many. The total sample size for the whole study is

$$(1+r)\frac{\left(\frac{1}{\delta_\sigma} u_{1-\alpha/2} + u_{1-\beta}\right)^2}{\delta_\theta^2},$$

where

$$\delta_\theta = \frac{\theta}{\sqrt{\tau_A^2 + \frac{\tau_B^2}{r}}}$$

and

$$\delta_\sigma = \sqrt{\frac{\tau_A^2 + \frac{\tau_B^2}{r}}{\sigma_A^2 + \frac{\sigma_B^2}{r}}}.$$

How to select the parameters we are free to choose? In planned experiments, we usually take $r = 1$. The level α always takes the standard 0.05 value. The power $1 - \beta$ should be set to at least 0.8, ideally to 0.95 to make probabilities of type I and type II errors the same. The alternative θ should express the clinically relevant effect, that is, the effect size that would make a substantial enough difference in clinical practice. This is something to discuss with the clinicians that initiated the study. Of course, small effect sizes require astronomical sample sizes.

This approach can be used to calculate sample size for **continuous outcomes**. For these outcomes, variances σ_A^2 and σ_B^2 under the null hypothesis are set to the same value (obtained from previous studies or by an "educated" guess) and the variances τ_A^2 and τ_B^2 are the same as under the null hypothesis. If there is no way even to guess the single variance parameter, a sensitivity analysis can be performed (evaluating the power/sample size for a range of sensible variance values).

For **binary outcomes** the means are success probabilities and the variances can be determined from them. We set π_0 the probability under the null hypothesis (the same for both arms) and the alternative π_1 in arm A. We can directly apply the sample size calculation to the test statistic based on the difference in the estimated probabilities, with

$$\theta = \pi_1 - \pi_0, \quad \sigma_A^2 = \sigma_B^2 = \tau_B^2 = \pi_0(1-\pi_0), \quad \tau_A^2 = \pi_1(1-\pi_1).$$

Or, we can consider a test based on the log odds ratio, with $\mu_A = \log \frac{\pi_1}{1-\pi_1}$, $\mu_B = \log \frac{\pi_0}{1-\pi_0}$

$$\theta = \log \frac{\pi_1}{1-\pi_1} - \log \frac{\pi_0}{1-\pi_0},$$
$$\sigma_A^2 = \sigma_B^2 = \tau_B^2 = \frac{1}{\pi_0(1-\pi_0)}$$
$$\tau_A^2 = \frac{1}{\pi_1(1-\pi_1)}.$$

For **time to event outcomes**, the asymptotics is based on the approximation (1.3) of Poisson distribution by normal distribution combined with the Δ-method. Let λ_A be the event rate in arm A (assumed to be constant for the purpose of power calculation) and λ_B be the event rate in arm B. Let u_A and u_B be the expected total duration of follow-up in both arms, depending on the number of subjects n_A, n_B and the follow-up durations y_A and y_B. Then the number of events observed in arm A is $N_A \sim \mathrm{Po}(\lambda_A u_A)$ and in the arm B it is $N_B \sim \mathrm{Po}(\lambda_B u_B)$. From (1.3) and (1.4) we have

$$\frac{\frac{N_A}{u_A} - \lambda_A}{\sqrt{\frac{\lambda_A}{u_A}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1) \quad \text{and} \quad \frac{\frac{N_B}{u_B} - \lambda_B}{\sqrt{\frac{\lambda_B}{u_B}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

Denote $\widehat{\lambda}_A = \frac{N_A}{u_A}$ and $\widehat{\lambda}_B = \frac{N_B}{u_B}$. By the Δ-method,

$$\frac{\log \widehat{\lambda}_A - \log \lambda_A}{\sqrt{\frac{1}{\lambda_A u_A}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1) \quad \text{and} \quad \frac{\log \widehat{\lambda}_B - \log \lambda_B}{\sqrt{\frac{1}{\lambda_B u_B}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

Consider the null hypothesis $H_0 : \lambda_A = \lambda_B$ and the alternative $H_1 : \lambda_B = \mathrm{e}^{\theta}\lambda_A$ for $\theta > 0$. Then, under H_1,

$$\frac{\log \widehat{\lambda}_A - \log \widehat{\lambda}_B - \theta}{\sqrt{\frac{1}{\lambda_A u_A} + \frac{1}{\lambda_B u_B}}} \xrightarrow{\mathrm{D}} \mathrm{N}(0,1).$$

We take $\lambda_B = \mathrm{e}^{\theta}\lambda_A$, $u_A = n_A y_A$, $u_B = r n_A y_B$ (better approximations of the total follow-up time are possible if the outcome is not rare) and use the previously developed sample size formula with

$$\tau_A^2 = \frac{1}{\lambda_A y_A}, \quad \tau_B^2 = \frac{1}{\mathrm{e}^{\theta}\lambda_A y_B}$$

and (under the null hypothesis, with $\theta = 0$)

$$\sigma_A^2 = \frac{1}{\lambda_A y_A}, \quad \sigma_B^2 = \frac{1}{\lambda_A y_B}.$$

Of course, usually $y_A = y_B$ but this calculation can handle even unequal follow-up duration.

If the primary analysis is **more complex** than a two-sample test (e.g., a linear regression model with some baseline measurements taken as additional predictors), we perform the sample size calculation on the two-sample test anyway, arguing that the regression analysis will only improve the power that will be available for the two-sample test. The problem with complex methods is that the sample size calculation would require the knowledge of additional parameters (such as regression parameters, residual standard error) that are not available before the data are collected.

Sometimes the calculated sample size is adjusted upwards for the expected attrition rate (subjects that will be enrolled but will not complete the study and will not be included in the final analysis).

As far as power calculation is concerned, it is important to be aware that it should be done only **before** the study is conducted. When the data is already collected, doing a power calculation does not make sense because it is already known whether or not the hypothesis has been rejected (the power is the probability of rejecting the null hypothesis under the alternative). Despite that, some experts who do not understand probability theory or hypothesis testing principles request a *post-hoc* power calculation as a measure of uncertainty of the study result. However, a better and a totally sufficient measure of this uncertainty is the confidence interval for the treatment effect.