Epidemiological and Statistical notes

www.pubhealth.ku.dk/~bxc/STAR/Epi.2005.02

STAR research course in
Diabetes Epidemiology and Biostatistics

Aurangabad, India
February 2005

Bendix Carstensen, Senior Statistician
Steno Diabetes Center
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk
http://www.biostat.ku.dk/~bxc

Charlotte Glümer, Senior Epidemiologist
Steno Diabetes Center
chgl@steno.dk

Preface

This set of notes is written as part of the teaching in diabetes Epidemiology and Biostatistics in the STAR-research courses in India.

It is not intended as a text to be followed closely during the courses, but rather as a reminder and reference for the simplest concepts mentioned in the course, i.e. as a handbook of "how to" in the most common instances encountered in (diabetes) epidemiology.

It contains some simple examples and exercises that hopefully will serve as illustration and encouragement for the reader to get familiar with the material.

It only contains some very basic concepts and methods that are essential to understanding of epidemiological concepts and analyses. It is not a text that can stand alone. It is a text for looking up the formulae and methods. Therefore, do not expect to find solutions in here to all problems that you encounter.

Bendix Carstensen & Charlotte Glümer
Steno Diabetes Center
January 2005.

Contents

Chapter 1

Epidemiological concepts

1.1 Epidemiological studies

Different study types are used for different purposes. In principle a complete enumeration and meticulous follow-up of an entire population could answer any relevant epidemiological question. But most epidemiological questions could be answered much cheaper. Hence other study types are used widely.

However, in order to understand the different study types it may be of help to imagine that we had access to detailed follow-up of the entire population, as illustrated in figure 1.1.

Each study type in epidemiology corresponds to a *partial* observation of this process, both with respect to which persons are included, and with respect to which variables are recorded and with respect to the time they are recorded.

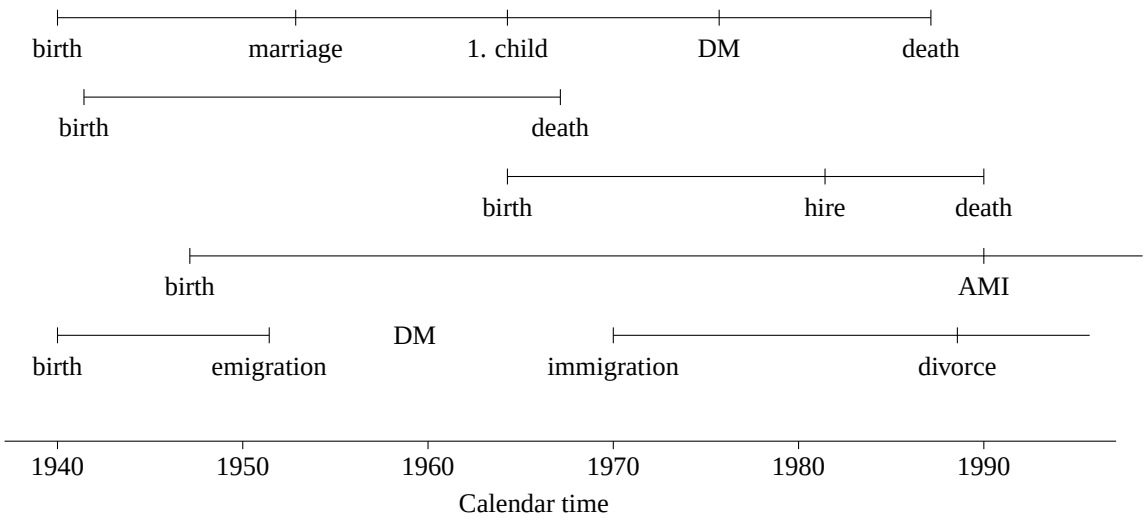

Figure 1.1: *Complete follow-up of a population. Dates of any event of interest is recorded, so the current status (age, no. children, no. previous marriages, etc.) is known at any point in time.*

1.2 Descriptive measures

1.2.1 Measures of disease occurrence

Incidence rate: The number of new cases during a specified time in a specified population. Measured in cases per person-year, i.e. the unit is time^{-1}.
Example: The incidence rate of DM among Danish children (5–9 years) is 8.5 cases per 100,000 person-years.

Prevalence of disease: The fraction or number of cases in a population at specified time *point*. Measured as a fraction (or number).
Example: The prevalence of T2D among 60 year old Danish women in 1999 was 12% [2].

Duration of disease: The *distribution* of time from diagnosis of disease to death or cure.
Example: The median duration of disease among T2D cases was 7.4 years.

In addition to these three measures we also frequently use the

(Cumulative) risk of disease: The probability of acquiring a disease during a specified interval of time.
Example: The 5-year risk of DM among persons with IGT is 7%.

Odds of disease: The probability of acquiring disease during a specified interval of time, divided by the probability of avoiding it.
Example: The 5-year odds of DM among persons with IGT is $0.07/0.93 = 0.075$.

For diseases that are rare, the risk over a given period is approximately equal to the rate multiplied by the period. This is the reason that rate and risk are often confused in terminology.

These measures will give an exhaustive description of the time-course of the disease in a homogeneous population. To the extent that the population is not homogeneous we must subdivide the population by age, sex and other factors believed to influence the course of disease.

1.2.2 Comparative measures

In addition to the descriptive measures that apply to a homogeneous population there is also a need for comparative measures that relates disease occurrence in two populations or two groups.

Rate Ratio: The ratio of two rates.
Example: The rate ratio of DM among boys versus girls is 1.2.

Relative Risk: Ratio of two risks, i.e. of two disease probabilities. Because of the close connection between rates and risks, a rate ratio is often also termed “relative risk”.

Disease Odds-ratio: Ratio of two disease odds. This measure is mostly used in case-control studies, mainly because it is the only estimable quantity from such studies (see section 1.5).

Attributable Risk: The fraction of cases *among exposed individuals* that can be attributed to an exposure.

Population Attributable Risk: The fraction of cases *among all individuals* that can be attributed to an exposure.

Number needed to treat: The number of persons one must treat in order to obtain one cured person. Equals the inverse of the risk difference between treatment and placebo.

1.3 Cross-sectional studies

When a sample of a population is surveyed at a fixed point in time we call it a *cross-sectional study*.

The only thing we record in such a study is the current status of persons, i.e. for example their fasting plasma glucose level, their smoking habits for the last 10 years (as they are recalled at the time of study).

One particular feature (drawback) of a cross-sectional study is that only persons alive at a given point in time are eligible. Therefore, the only direct measure of disease occurrence available is the *prevalence* of disease. This is measured as the affected fraction of the population at a given point in time.

Data will be current status information on a sample of the population. For example, age, sex, blood pressure, fasting plasma glucose.

That will enable us to compute for example:

- Prevalence of IFG (as the fraction with fasting plasma glucose > 7.1mmol/l).
- Sex-specific prevalences of IFG.
- Male-female ratio of IFG prevalence.
- Distribution of systolic blood pressure for males and females.

1.4 Cohort or follow-up studies

These are studies where a population (the *cohort*) is followed over a time period and disease occurrence (*events*) is recorded. Thus for each person we will know when disease occurred and what periods were disease free.

From a cohort study of death or a chronic disease we must for each person know:

- date of entry into the study
- date of exit from the study
- status at exit (diseased / non-diseased)

Aggregating the number of events and the risk time (date of exit $-$ date of entry) for the entire cohort gives:

- Number of events (D)
- Amount of follow-up time (person-years) (Y)

so the average *rate* of disease (or death) can be computed as D/Y.

In short: Cohort studies are studies that allows the estimation of rates.

Normally, explanatory variables will be recorded not only at the entry into the cohort but also as time passes. Variables that will require updating during follow-up are for example parity (no. children born), smoking habits and marital status.

Such variables will allow subdivision of events and person-years by for example parity, so that rates can be compared between different parity groups. Note that for variables such as parity that varies *within* each person, each person may contribute risk time in more than one group.

1.5 Case-control studies

These are studies where a sample of cases, is compared with a sample of non-cases (controls) from the population.

The overall rationale is that if a condition is much more prevalent among *cases* than among *controls* then this condition is likely to be a risk factor for occurrence of the disease.

In a case-control study information is obtained for less money than in a cohort study where cases may be scarce. The drawback is that it is difficult to obtain the same data quality as in a cohort study, and that the data are inherently cross-sectional.

For example, if obesity is found to be more prevalent among cases of T2D, than among controls sampled from the population we may be inclined to infer that obesity is a risk factor for T2D. However if obesity is a consequence of T2D we would see exactly the same kind of data.

This refers to the situation where cases were sampled as *prevalent* cases of T2D. If we instead had sampled cases at the time of T2D onset (which would be a logistically very complicated design indeed), we would not have the possibility of mixing determinants of T2D with consequences.

1.5.1 Sampling plans

Inference from case-control studies is only valid if cases and controls are drawn from the same population, i.e. if any control would eligible as a case had she been a case.

Thus the way we sample cases and controls influences the range and validity of the conclusions that we can draw.

Prevalent cases

If cases are taken as prevalent cases in a given population and controls as a sample of the population at the same time, we are only able to draw inferences that we could also have drawn from a cross-sectional study.

Incident cases

If cases are included as they are diagnosed, and controls are taken at the same time as the cases, we have a design which is known as *incidence density sampling*, which allows inference as from cohort studies, i.e. estimation of rate-ratios (but not rates). This is under under the assumption that all information is recorded in the same quality as in a cohort study.

We may get biased or inaccurate information on variables that subjects are asked to recall retrospectivley at data collection, whereas other variables such as genotype can be recorded virtually without error at any time.

1.5.2 Analysis

For the sake of simplicity, let us consider estimation of the effect of an exposure with prevalence p. Suppose the disease risk *in the population* is π_1 among exposed and π_0 among non-exposed, and assume that a case is sampled with probability s_1 (a number close to 1) and a control with probability s_0 (a number close to 0).

The disease odds *in the study* is for *exposed* individuals:

$$\omega_1 = \frac{\text{P}\left\{\text{Case, exposed, in the study}\right\}}{\text{P}\left\{\text{Control, exposed, in the study}\right\}} = \frac{p \times \pi_1 \times s_1}{p \times (1-\pi_1) \times s_0} = \frac{\pi_1}{(1-\pi_1)} \times \frac{s_1}{s_0}$$

and for *unexposed* individuals:

$$\omega_1 = \frac{\text{P}\left\{\text{Case, exposed, in the study}\right\}}{\text{P}\left\{\text{Control, exposed, in the study}\right\}} = \frac{p \times \pi_0 \times s_1}{p \times (1-\pi_0) \times s_0} = \frac{\pi_0}{(1-\pi_0)} \times \frac{s_1}{s_0}$$

and hence we have the *odds-ratio* of disease between exposed and unexposed *in the study*:

$$\text{OR}_{\text{study}} = \frac{\pi_1}{(1-\pi_1)} \times \frac{s_1}{s_0} \Big/ \frac{\pi_0}{(1-\pi_0)} \times \frac{s_1}{s_0} = \frac{\pi_1}{(1-\pi_1)} \Big/ \frac{\pi_0}{(1-\pi_0)} = \text{OR}_{\text{population}}$$

Thus we see that the odds-ratio in the population can be estimated from the case-control sample. The core assumption is that the sampling fractions s_0 and s_1 only depend on case-control status and not on exposure status.

If a study is based on prevalent cases, the OR estimates the prevalence-odds-ratio; if it is based on incidence density sampling the OR estimates the cumulative incidence odds-ratio for a very short period, which is the same as the rate-ratio.

The extension of these formulae from a binary exposure to several possibly continuous exposures will lead to logistic regression models. This topic is outside the scope of these notes.

1.6 Case-cohort studies

In a case-control study the cases and controls are selected at the same time, either as a cross-sectional sample (prevalent cases) or by incidence density sampling (incident cases), so the sampling of the controls is tied to the occurrence of the cases.

In a case-cohort study a sub-cohort is selected at the start of the study period. Information is then collected on the entire sub-cohort and on all cases that occur during the study period (both inside and outside the sub-cohort).

The advantage of this study design is the ability to study several different outcomes *with the same set of controls* (the sub-cohort). In the incidence density sampled case-control study the selection of controls is tied to the occurrence of the cases.

A primary example of this design is in studies based in the EPIC data where a large group of people in Europe ($\approx$ 500,000) were included and asked about dietary habits. To analyse the effect on occurrence of different cancers the case-cohort design is ideal, because only questionnaires for a subcohort need be processed, and not a separate control-sample for each type of cancer.

The drawback of the design is a somewhat more complicated statistical analysis, which requires software for survival analysis with special features. The analysis of theses studies is outside the scope of these notes.

1.7 Randomized studies

In order to establish a causal link between an exposure and a disease outcome it is necessary to manipulate the exposure and subsequently follow persons w.r.t. disease occurrence and compare the events in an exposed and a non-exposed group. This is the basic idea behind randomised studies.

Although the conceptual difference is small, one often distinguishes between intervention studies and (randomized) clinical trials.

The terminological difference comes from the fact that intervention studies are aimed at lifestyle changes by means of intervention programs, whereas clinical trials are aimed at disease cure by means of drugs or medical procedures.

1.7.1 Intervention studies

Intervention studies are studies where a (random) subset of a target population is offered an intervention (exercise, smoking cessation,...) and the population then is followed up for relevant (disease or lifestyle) outcomes. The intervention group is then compared with the non-intervention group.

If intervention has been allocated by randomization, then the only two possible explanations of differences between the groups are 1) chance and 2) effect of intervention. The effect of chance can be quantified, and hence ruled out with any desired degree of certainty (provided of course that the study is sufficiently large).

1.7.2 Randomized clinical trials

The rationale behind randomized studies is the same as behind intervention studies: In order to establish a causal effect of a drug or medical procedure one needs to manipulate the allocation of it by randomzing patients to different regimens.

Then by the virtue of the randomization only chance and drug effect will remain as explanation for any difference between drugs.

1.7.3 Analysis of randomized studies

The statistical analysis of randomized studies depends on the outcome.

If the outcome is disease occurrence or recovery (binary, yes/no outcome), the methods is as for a follow-up study, analysis of incidence rates (survival analysis) or proportions.

If the outcome is a measurement as e.g. blood-pressure or change in blood pressure, the methods will be analysis by linear models (anova, t-test, regression). These methods are outside the scope of these notes.

1.8 Bias

Bias means distortion of results in a systematic manner. In epidemiology it refers specifically to distortion arising from errors in design and data-collection. Some biases may be correctable if discovered, others may not.

Bias can arise in several ways; Sackett [6] has compiled a long list of possible biases that may arise in epidemiological studies.

The main sources of bias are:

Selection bias: If subjects included in a study do not represent the population they are supposed to.
Example: If obese diabetics are more likely to be recruited than non-obese diabetics, the effect of obesity will be exaggerated. If the magnitude of this over-representation is unknown, there will be no way to correct it.

Information bias: If data collected on individuals are systematically wrong.
Example: If obese persons under-report their weight, the effect of weight on a health outcome may be over-estimated.

Recall bias: If data from individuals are collected retrospectively, cases and non-cases may recall exposures differently, influenced by the disease.
Example: T2D cases may be more likely to recall changes in dietary habits than non-cases. This would distort the estimate of dietary effect on T2D.

1.9 Confounding

If a variable is associated both to the exposure of interest and to the disease outcome, then the effect of the exposure may be estimated wrongly.

If we want to assess the effect of obesity on the prevalence of T2D, we should be aware that age is associated both with obesity and T2D. If we fail to control for age we will attribute some of the age effect to obesity, we say that the effect of obesity is *confounded* by age.

The analysis should estimate the effect of obesity at a given age, i.e.separately in each age and then combine the estimates. In practice this would be done by using a regression model, "controlling" for age. The technical details of this are outside the scope of these notes.

The formal definition of a (potential) *confounder* is a variable which is:

1. associated with exposure of interest.
2. associated with disease.
3. an *independent* risk factor.

Note the last requirement which is often overlooked:

Suppose we instead of implicitly assuming that genotype were an independent risk factor for T2D, considered genotype as a *cause* of obesity, i.e. assumed that the only effect of genotype on T2D was through obesity:

$$\text{Genotype} \longrightarrow \text{Obesity} \longrightarrow \text{T2D}$$

In this case it would not be meaningful to control for obesity in the analysis. This situation cannot be distinguished in the data from the situation where there is an independent effect of obesity.

Confounding is not a feature of *data*, it is a feature of the epidemiological conceptualization of data.

Therefore, there is no such thing as a statistical test for confounding.

1.10 Diagnostic tests and screening

1.10.1 Diagnostic tests

A diagnostic test classifies persons as "diseased" or "non-diseased". Such a test is not infallible, so some truly diseased persons may be classified as non-diseased, and some non-diseased may be classified as diseased.

The performance of a diagnostic test can be evaluated by testing a number of persons with known disease status. Results are usually summarized in a 2×2 table, and the sensitivity and specificity of the *test* relative to the *true disease status* are computed as shown.

	Disease status			
Test	$+$	$-$	$+$	$-$
$+$	True positive	False positive	a	b
$-$	False negative	True negative	c	d
		Sensitivity:	sens $= a/(a+c)$	
		Specificity:	spec $= d/(b+d)$	
		Predictive value of positive test:	PV$+ = a/(a+b)$	
		Predictive value of negative test:	PV$- = d/(c+d)$	

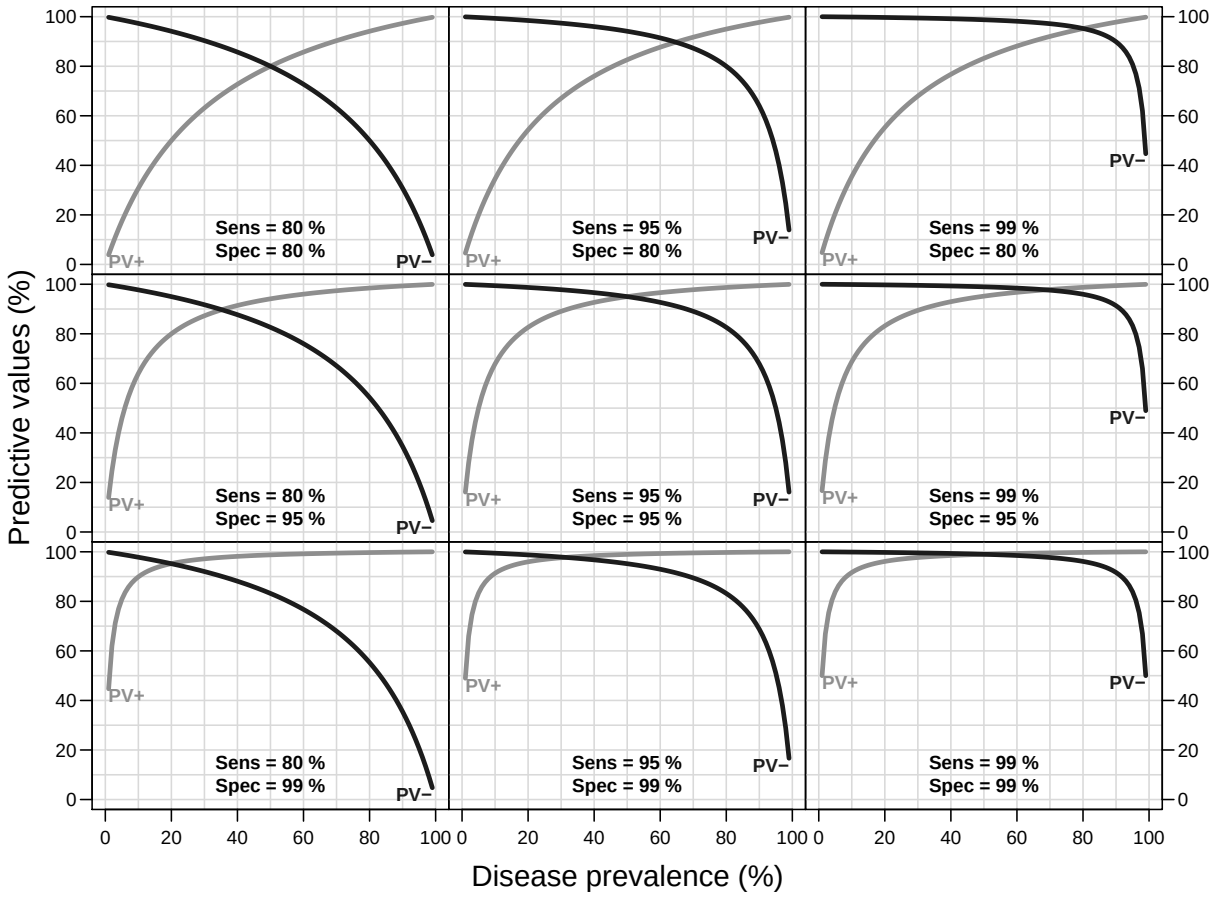

Figure 1.2: *Predictive values as functions of disease prevalence and sensitivity and specificity. Note that* PV+ *increases with disease prevalence and specificity, but almost not with sensitivity (same curves in each row). Conversely* PV− *decreases with increasing disease prevalence and increases with increasing sensitivity, but is almost unaffected by increasing specificity (same curve in each column).*

A predictive value is the probability of disease status given test — predictive value of a positive test (PV+) is the probability of being diseased given a postive test and predictive value of a negative test (PV−) is the probability of being non-diseased given a negative test. In practical applications it is the predictive values that are of interest, since the true disease status of persons will not be known.

The predictive values are not only determined by the sensitivity and specificity, but also depend on the prevalence of the disease, specifically:

$$\begin{aligned} \text{PV}+ &= \frac{\text{sens} \times p}{\text{sens} \times p + (1 - \text{spec}) \times (1 - p)} \\ \text{PV}- &= \frac{\text{spec} \times (1 - p)}{\text{spec} \times (1 - p) + (1 - \text{sens}) \times p} \end{aligned}$$

Thus for common diseases (prevalence p close to 1) it is easy to obtain a large PV+, and for rare diseases (p close to 0) it is easy to get large PV−.

In figure 1.2 is illustrated how the predictive values depend on the disease prevalence and sensitivity and specificity.

1.10.2 The likelihood ratio

Instead of positive predictive value $\mathrm{P}\{\ \mathrm{D}+ \mid \mathrm{T}+\ \}$, we may consider the positive predictive **odds**:

$$\frac{\mathrm{P}\{\ \mathrm{D}+ \mid \mathrm{T}+\ \}}{\mathrm{P}\{\ \mathrm{D}- \mid \mathrm{T}+\ \}}$$

This can also be written:

$$\begin{aligned}\frac{\mathrm{P}\{\ \mathrm{D}+ \mid \mathrm{T}+\ \}}{\mathrm{P}\{\ \mathrm{D}- \mid \mathrm{T}+\ \}} &= \frac{\mathrm{P}\{\ \mathrm{T}+ \mid \mathrm{D}+\ \}\,\mathrm{P}\{\ \mathrm{D}+\ \}/\mathrm{P}\{\ \mathrm{T}+\ \}}{\mathrm{P}\{\ \mathrm{T}+ \mid \mathrm{D}-\ \}\,\mathrm{P}\{\ \mathrm{D}-\ \}/\mathrm{P}\{\ \mathrm{T}+\ \}} \\ &= \frac{\mathrm{P}\{\ \mathrm{D}+\ \}}{\mathrm{P}\{\ \mathrm{D}-\ \}} \times \frac{\mathrm{P}\{\ \mathrm{T}+ \mid \mathrm{D}+\ \}}{\mathrm{P}\{\ \mathrm{T}+ \mid \mathrm{D}-\ \}}\end{aligned}$$

This equation can be expressed as:

$$\text{Posterior odds} = \text{Prior odds} \times \text{Likelihood ratio}$$

where:

Posterior odds: The odds of disease given data (i.e. the positive test)

Prior odds: The odds of disease **prior** to the test.

Likelihood ratio: How much is the odds of disease increased in the light of the data (the positive test). This is a measure of how much we improve the prediciton of disease by doing a test, instead of just guessing based on the prevalence odds ($\mathrm{P}\{\ \mathrm{D}+\ \}/\mathrm{P}\{\ \mathrm{D}-\ \}$).

The likelihood ratioa for a positive and negative test are functions of sensitivity and specificity:

$$\frac{\mathrm{P}\{\ \mathrm{T}+ \mid \mathrm{D}+\ \}}{\mathrm{P}\{\ \mathrm{T}+ \mid \mathrm{D}-\ \}} = \frac{\text{sens}}{1-\text{spec}} \qquad \frac{\mathrm{P}\{\ \mathrm{T}- \mid \mathrm{D}+\ \}}{\mathrm{P}\{\ \mathrm{T}- \mid \mathrm{D}-\ \}} = \frac{1-\text{sens}}{\text{spec}}$$

A likelihood ratio measures how much disease odds is changed in the light of data (i.e. by doing a test).

It is *independent* of the prevalence of disease.

1.10.3 Screening tests based on a continuous measurement

The dependence of PV+ on the prevalence of DM is part of the rationale for targeted screening, as the prevalence of the disease is higher in the high-risk population than in the general population , then the predictive value of a positive test is also higher.

Screening for diabetes typically involves determination of a fasting plasma glucose (or random blood glucose), and so a diagnostic test is defined as positive if the measuremnt exceeds a certain threshold, a cut-off, e.g. 7.1 mmol/l.

Ideally, a screening test should be characterized by high sensitivity and specificity. However, increasing the sensitivity by lowering the cut-off for positivity of a test, results in a decrease in the specificity. There is no such monotone relationship between the cut-off and the predictive values.

So for each threshold chosen as cut-off for the diagnostic test we get a set of (sensitivity, specificity, PV+, PV−)

1.10.4 ROC curves

When evaluating a screening test on the basis of pairs of sensitivity and specificity related to a specific cut-point will only give a small glimpse of a test's performance, and its real diagnostic ability may not be revealed.

Therefore we could calculate the sensitivity and specificity for *any* cutpoint. Plotting the sensitivity versus $1-$specificity would give a picture of the overall discrimination power of the variable. The ROC curve will start in (0,0) and end in (1,1), see the example in figure 1.3.

A Receiver Operating Characteristic (ROC) plot provides a view of the entire spectrum of sensitivities and specificities because all possible sensitivity/specificity pairs for a particular test are displayed.

The *area under the curve* (AUC) is a measure of the diagnostic accuracy. The AUC represents the probability that a randomly selected individual in the diseased group has a higher value of the test than a randomly selected individual in the non-diseased group. An AUC of 1 indicates a perfect separation of the test values in the two groups, whereas an AUC of 0.5 indicates no apparent difference in distribution of the measurements between the two groups.

In clinical settings a decision cut-off (either diagnostic or for screening) must be chosen. The optimal cut-off in the literature of screening tests is often defined as the point where the 45° tangent touches the ROC curve. This maximises the sum of sensitivity and specificity, but with this cut-off it is assumed that the sensitivity and specificity are of equal clinical significance, which is not always the case.

In the case of stepwise screening for diabetes where the first step includes a risk score, which is relatively cheap and harmless for the population, the cut-off value may be low and result in a high sensitivity but a relatively low specificity. This is not necessarily the "optimal" cut-off point. On the other hand, the next step may introduce a blood measurement and thus the specificity may be of greater importance than the sensitivity. The optimal cut-off is a function of the cost of false positive and false negative test results where the cost is not only including the financial cost due to misclassification, but also includes of the potential harm/personal cost of a test.

Example

In a Danish screening study of T2D, HbA_{1c} was considered as posibble tool for diagnostic testing. The ROC-curve based on this measurement alone is shown in figure 1.3.

1.10.5 Questionnaires / Risk scores

A risk scores based on e.g. a simple questionnaire or other relatively cheaply obtainable information can be a useful screening tool in targeted screening. It is used to identify a group of people at high risk of diabetes, and diagnostic testing is then restricted to this high-risk population, reducing the burden on participants and personnel and hence the cost of screening.

1.11 Measurement error and misclassification

Associations between variables are in general diluted if the variables are measured with error. But only if the measurement error is independent of both the measured and other variables, i.e. the same for small and large values of these. We call this independent measurement error.

A way of seeing this is to imagine the extreme sitaution where measurement error dominates the value of the variable. Then the variable is nothing but random error which of course is unrelated to any other variable.

Misclassification is in principle the same phenomenon as measurement error, only for categorical variables. A categorical variable is measured with error if individuals are occasionally placed in the

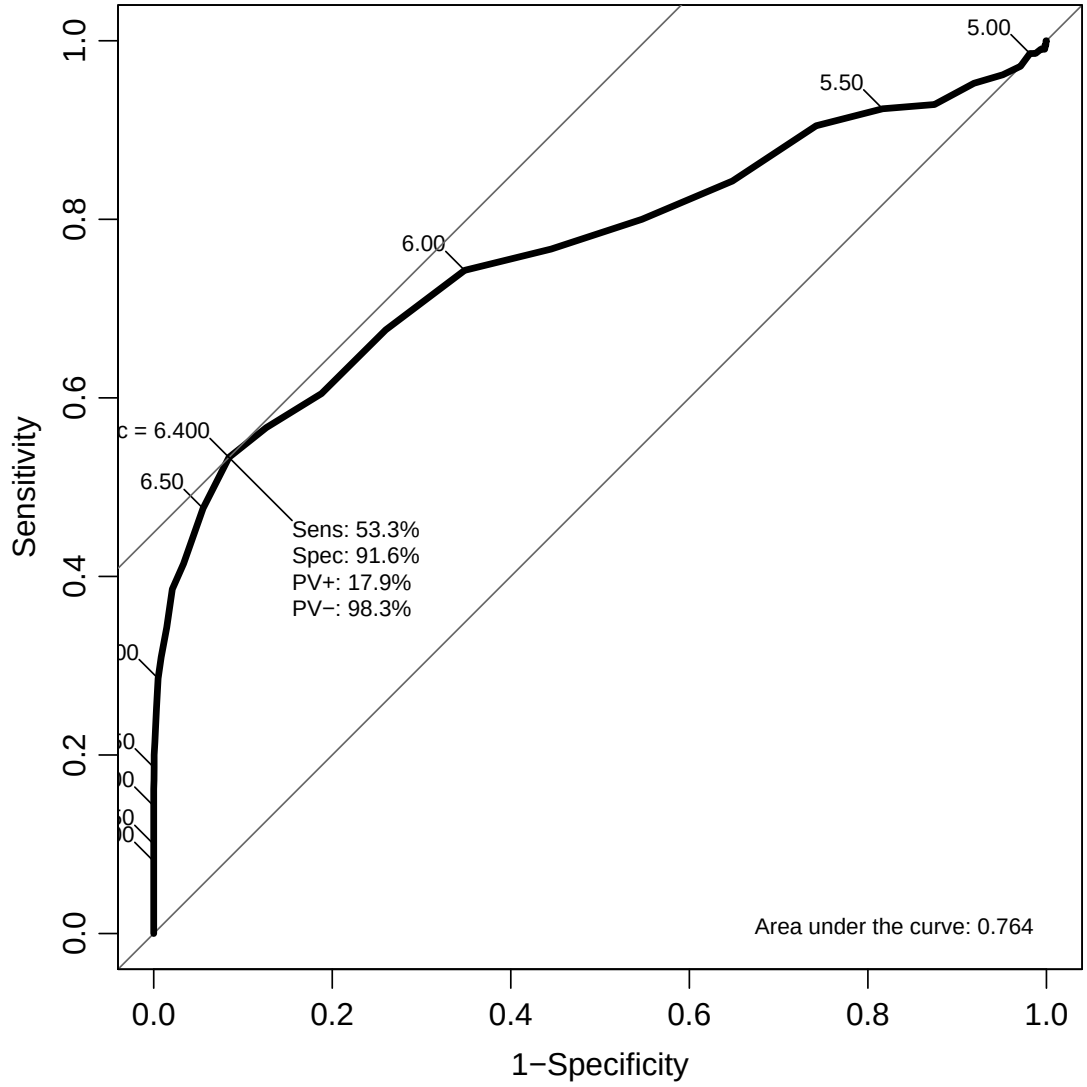

Figure 1.3: *ROC-curve for DM-prevalence (SDM) based on measurement of HbA_{1c}. The "optimal" cutpoint, 6.4% is indicated on the curve. Neither the sensitivity, specificity nor the predictive values are impressively high.*

wrong category, misclassified. If the misclassification is independent of the category and other varaibles we talk about "non-differential misclassification".

The effect of independent measurement error and non-differential misclassification is to dilute associations between variables. Thus we will have more conservative estimates as eroors increase.

If measurement error or misclassification depends on other variables results may be biased. The direction and magnitude of bias depends on how the measurement errors depends on variables.

1.12 Data

Epidemiological data are usually collected by individual, and are mostly stored with one record per person, and all information on each person in that. Here is an example from a study from SDC:

id	help1	age	sex	ddur	insdur	bmi	suglocat	aware2	hba1c	insreg	inskilo
1	0	84	1	14	11	22.30815	NA	NA	11.2	1	0.2333333
2	1	65	2	29	25	25.30864	1	2	8.5	3	0.5121951
3	0	61	1	7	4	20.96211	2	1	7.1	2	0.5270655
5	0	68	2	13	9	23.27608	1	1	7.8	2	0.5547850
7	0	48	1	12	5	27.05515	2	1	9.7	4	0.8307692
8	0	69	2	7	7	31.30918	NA	NA	7.9	2	0.3629032
10	0	78	2	18	3	25.66115	2	1	8.5	2	0.2352941
14	0	71	2	7	5	23.96566	1	1	8.6	2	0.6770099
15	1	70	2	20	3	30.82747	1	2	7.4	2	0.2850877
16	2	63	1	17	7	25.45807	1	1	9.0	4	0.9295775
18	1	64	2	20	9	41.00407	2	1	9.1	2	0.8673895

In this dataset each record is identified by the `id` of the person.

If we have follow-up data on persons, we can either have separate variables as date of first visit, blood pressure at first visit, date of second visit, blood pressure at second visit, and so on. This is normally very clumsy, and the preferred way of organizing data is with one record per person and visit. In such a dataset each record would be identified by patient id *and* date of visit.

1.12.1 Types of variables

Normally we distinguish between continuous and categorical variables.

Continuous variables are variables that (in principle) can take any of an infinity of values (blood pressure, height, age). The actual value is of course limited by the precision with which we can measure.

Categorical variables can only take a few different values (sex: M / F, nationality: DK / GB / IND / CHN / ..., glucose tolerance: NGT / IFG / IGT / DM). Some categorical variables are derived by grouping one or more continuous variables, for example glucose tolerance which is derived from the two continuous variables FPG and 2hPG.

1.12.2 Terminology

In epidemiological/statistical terms, we talk about a *data set*, each line is termed a *record* and each column a *variable*, and the unique identifier of the record is called the *(record)-id*.

In database terms a data set is called a *table*, records are called *records*, variables *fields*, and the record identifier the *key*. Thus in the last example mentioned above the key would be (person-id,visit-date).

Chapter 2

Statistical concepts

2.1 Statistical models

All of the simple calculations shown in these notes are special cases of more complex statistical models, either a logistic regression model (for proportions, and odds-ratios) or a Poission regression model (for rates).

2.2 Statistical concepts

When conducting a study in order to investigate a certain hypothesis, the conclusion must be based on the collected data. It is essential translate the clinical/epidemiological questions into quantitative terms in order to clarify what the quantity of interest is, for example:

- The rate-ratio of DM occurrence between persons who have sedentary vs. physically active work.
- The mean blood-pressure in newly dignosed T2D patients.
- The number needed to treat with drug X in order to prevent one stroke.

The target of a statistical analysis is to provide an *estimate* of the quantity of interest, and of the *uncertainty* associated with this estimate.

2.2.1 Estimate

The *estimate* is a function of the data, such as the ratio of rates or the average of the observed blood pressures. But it may very well also be a result from a more complicated computation when a more elaborate statistical model is used for the analysis.

An estiamte is a quantity which reflects the biological processes that generated data. As such it is not likely to change by design of the study beyond what is dictated by sampling error.

2.2.2 Confidence intervals

Uncertainty of an estimate is quantified in the form of a *confidence interval*, usually a 95% confidence interval. The interpretation of a confidence interval is:

"With 95% probability the interval will contain the true value of the hazard ratio"

or slightly more informal:

"The interval are those values of the hazard ratio that are supported by the data".

This implies that values outside the confidence interval are *not* supported by the data. So if the confidence interval for the rate ratio (hazard ratio) is (1.2;3.5) then we can exclude that the rate ratio is smaller than 1.2 and larger than 3. In particular we can exclude that the two groups have the same T2D rates (hazard ratio=1). But we cannot exclude a hazard ratio of 1.5. So if a hazard ratio of 1.5 is considered epidemiologically or clinically irrelevant, then the study cannot exclude a difference between the groups that is irrelavent, and so in this sense is inconclusive.

As opposed to the estimate, the width of a confidence interval is stronly influenced by the design of the study, notably by the size of the study.

2.2.3 Tests and p-values

If we are only interested in whether a given quantity is different from a null value (1 for a ratio and 0 for a difference), we may use a p-value. The p-value is a probabilistic measure of the distance between the data and a defined *null hypothesis*. In the case of comparing two rates a typical null hypothesis would be that "rate ratio=1". A null hypothesis is typically a hypothesis of no effect. Rejecting the null hypothesis is then the same as claiming effect.

The p-value gives the probability of observations that disagree with the null hypothesis at least as much as the observed data; "observations that are further away from the null than the observed data". If this is small it must mean that the observed data are far away from the null. Hence the null must be wrong. The usual convention is to reject the null hypothesis if the p-value is less than 0.05 (the *significance level*).

This exercise is called to *test a hypothesis*.

However, a small p-value can arise either from a large and imprecise estimate or from a small and precise estimate. If the estimate is large and imprecise, the study is inconclusive, the confidence interval may contain effect values so small to be of little importance and effect values so larger that they are of great importance. But if the estimate is small and precise the confidence interval may only contain values that are of limited importance, and thus be conclusive of absence of relevant effects.

Thus a p-value does not tell whether a study is conclusive or not. This is the reason that the use of p-values alone is discouraged, and that confidence intervals for the effects of interest are required by many medical journals.

Technical form of a test

Most tests are test of whether an estimated quantity is 0 or some other fixed value (the null value). For example if the prevalence observed equals 6% or whether the rate-ratio between two groups is 1.

In most cases we take the estimate, work out the distance from the null value. This distance is then related to the standard error of the estimate. In practise the deviation is divided by the standard error, and squared to yield the *test statistic*. This is compared to the χ^2-distribution with 1 degree of freedom.

Modern computer programs have facilities to work out the p-value, i.e. the probability that a random variate with a χ^2-distribution exceeds the value of the test statistic. A value of 3.84 of the test-statstic corresponds to a p-value of 5%.

2.3 One proportion

When a population or a sample is classified in one of two categories (well / diseased) we may ask what *proportion* of the population is diseased.

2.3.1 Formulae

In the observation of a proportion (prevalence) we have:

x**:** Number of affected persons.

n**:** Total number of persons.

The prevalence or proportion affected is computed as:

$$p = \frac{x}{n}$$

For small values of x or n use the formula based on log-odds[1]. First compute the error factor:

$$\text{ef} = \exp\bigl(1.96/\sqrt{np(1-p)}\bigr)$$

then the confidence interval is:

$$\frac{p}{p + (1-p) \times \text{ef}} \quad ; \quad \frac{p}{p + (1-p)/\text{ef}}$$

This does not work when $x = 0$ (then $p = 0$) or $x = n$ (then $1 - p = 0$).

$x = 0$ **or** $x = n$

These are cases where either none or all are affected.

In this case the exact 95% limits can be computed easily:

$$x = 0 : \qquad 0 \quad ; \quad 1 - 0.025^{1/n}$$

$$x = n : \qquad 0.025^{1/n} \quad ; \quad 1$$

2.3.2 Example

A population survey in China found that 72 out of 1424 persons surveyed had IFG (i.e. a fasting plasma glucose value ≥ 7.1).

The prevalence estimate is thus:

$$p = \frac{72}{1424} = 5.06\%$$

The error factor becomes

$$\text{ef} = \exp\bigl(1.96/\sqrt{np(1-p)}\bigr) = \exp\bigl(1.96/\sqrt{1424 \times 0.0506 \times 0.9494}\bigr) = 1.268$$

Then the 95% c.i. becomes:

$$\begin{aligned}\left(\frac{p}{p + (1-p) \times \text{ef}}; \frac{p}{p + (1-p)/\text{ef}}\right) &= \left(\frac{0.0506}{0.0506 + 0.9494 \times 1.268}; \frac{0.0506}{0.0506 + 0.9494/1.268}\right) \\ &= (0.0403, 0.0632) = (4.0\%; 6.3\%)\end{aligned}$$

[1] This is a formula based on computing a confidence interval for $\log(p/(1-p))$ and back transforming this to the prevalence scale. In text-books is most often used the formula $p \pm \sqrt{p(1-p)/n}$ which is highly inaccurate for small samples and for proportions close to 0 and 1.

2.3.3 Exercise

In the chinese population survey there were 554 males and 870 females. Among the diabetic individuals there were 25 males and 47 females.

Compute the prevalence of diabetes among men and women separately, and give 95% confidence intervals for each of these figures.

2.4 Two proportions

When we observe a prevalence in two groups we want a comparative measure, some numerical expression of how the two groups differ.

The observations which form the basis for this are:

x_1, x_0: Number of affected persons in group 1 and 0 respectively.

n_1, n_0: Total number of persons in group 1 and 0 respectively.

There are a number of options for description of the differences between the two groups:

1. Difference of the proportions.
2. Ratio of the proportions — relative risk.
3. Odds-ratio.

The different measures address different situations, and the decisiaon on what to use is a subject matter decision, *not* a statistical one.

Testing wheter two proportions are equal can be based on either of the statistics, and may give slightly different results, but the hypothesis is the same wheter it is formulated as “relative risk=1”, as “odds-ratio=1” or “difference of proportions=0”.

2.4.1 Difference of two proportions

Differences of proportions are relevant if the parameter of interest is the fraction of the population that will benefit from an intervention. The difference in response proportions in treated and untreated groups will be the fraction of persons that benefit from a treatment.

Formulae

If we have two proportions and want to compare by the difference, we use a formula for the s.e. of each proportion and get:

$$\text{s.e.}[p_1 - p_0] = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

95% c.i. for difference of proportions is then:

$$p_1 - p_0 \pm 1.96 \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

Example

If the Chinese data are subdivided by sex we find 25 male cases out of 554 ($p = 0.045$), and 47 female cases out of 870 ($p = 0.054$).

Thus we have a difference in proportions of $0.0451 - 0.0540 = -0.0089$ and a 95% c.i. for this

$$\begin{aligned} -0.0089 \pm 1.96 \times \sqrt{\frac{0.0451 \times 0.9549}{554} + \frac{0.0540 \times 0.9460}{870}} &= -0.0089 \pm 1.96 \times 0.0117 \\ &= (-0.0317; 0.0140) \\ &= (-3.2\%; 1.4\%) \end{aligned}$$

Thus the male-female difference in prevalences is not significantly different from 0. In fact the data would support claims from a 3% lower prevalence among men to a 1% higher.

2.4.2 Number needed to treat

If the two groups we compare is a treated and an untreated group and we record the fraction in each that recovers from a disease, we can calculate the *number needed to treat* to bring about one cured patient.

The reasoning is as follows: If the extra fraction of recoveries in the treated group is say 10% then we obviously need to treat 10 persons to bring about one extra recovery. Note that we must relate it to the *extra* fraction of recoveries, i.e. the *difference* in recovery probability between treated and untreated.

If the recovery fraction is p_1 in the treated group and p_0 in the untreated, then:

$$\text{fraction benefitting from treatment} = p_1 - p_0$$

and hence, in order to bring about one extra recovery we need to treat:

$$n_{\text{ntt}} = \frac{1}{p_1 - p_0}$$

If we have a confidence interval for the probability difference, $p_1 - p_0$, we just need to take the inverse of the ends of the confidence interval for the number needed to treat.

Example: Treatment of foot ulcers with Dalterapin

Kalani *et al.* [5] reported on 85 diabetes-patients with foot ulcers that were treated with either Dalterapin or Placebo. Their findings can be summarized in the following table:

	Treatment	
Outcome	Dalterapin	Placebo
Improved	29	20
Worse or unchanged	14	22
Total	43	42

$$p_{\text{Dal}} = \frac{29}{43} = 0.674 \qquad p_{\text{Pl}} = \frac{20}{42} = 0.476$$

Difference in improvement probabilities is the fraction of the patients that benefit from treatment: $p_{\text{Dal}} - p_{\text{Pl}}$

$$\begin{aligned} p_{\text{Dal}} - p_{\text{Pl}} &= 0.198 \approx 20\% \\ \text{s.e.}[p_{\text{Dal}} - p_{\text{Pl}}] &= \sqrt{\frac{p_{\text{Dal}}(1-p_{\text{Dal}})}{n_{\text{Dal}}} + \frac{p_{\text{Pl}}(1-p_{\text{Pl}})}{n_{\text{Pl}}}} \\ &= 0.105 \\ 95\%\text{ c.i.} \quad : \quad & 0.198 \pm 1.96 \times 0.105 = (-0.008; 0.404) \approx (0\%, 40\%) \end{aligned}$$

The number needed to treat in order to bring about improvement in one patient is:

$$n_{\text{ntt}} = 1/0.198 = 5.04$$

and a confidence interval for this is:

$$1/(0.0; 0.4) = (2.5, \infty)$$

So it looks as if the treatment is effective but the uncertainty of the statement is rather large. Since a difference of 0 is in the confidence interval, the effect is not significant, and hence the confidence interval for the n_{ntt} goes all the way to infinity.

Exercise

If the relevant outcome of the Dalteparin study is considered amputation or not the resulats are:

	Treatment	
Amputation	Dalterapin	Placebo
Yes	2	8
No	41	34
Total	43	42

Compute the difference in ampuation frequencies between the two therapies and a confidence interval for this.

Also compute the number needed to treat in order to avoid one amputation. Produce a 95% confidence interval for this as well.

2.4.3 Ratio of two proportions

Formulae

If we have two proportions and want to compare them by the ratio, the *relative risk*, we first work out:

$$p_1 = \frac{x_1}{n_1}, \qquad p_0 = \frac{x_0}{n_0}, \qquad \Rightarrow \qquad \text{RR} = p_1/p_0$$

To form a confidence interval we form the *error factor*:

$$\text{ef} = \exp\left(1.96 \times \sqrt{\frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_0} - \frac{1}{n_0}}\right)$$

and the 95% c.i. for the RR is then:

$$\text{RR} \overset{\times}{\div} \text{ef}$$

The term inside the expression for the error factor is the standard error of the $\log(\text{RR})$, so a test for $\text{RR} = 1$ must be transformed to a test of $\log(\text{RR}) = 0$, so the test statistoc becomes

$$\chi^2(1) = \left(\frac{\log(\text{RR}) - 0}{\sqrt{\frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_0} - \frac{1}{n_0}}} \right)^2 = \frac{\log(\text{RR})^2}{\frac{1}{x_1} - \frac{1}{n_1} + \frac{1}{x_0} - \frac{1}{n_0}}$$

Examples

If the Chinese data are subdivided by sex we find 25 male cases out of 554 ($p_\text{M} = 0.0451$), and 47 female cases out of 870 ($p_\text{F} = 0.0540$).

Thus we have $\text{RR} = p_\text{M}/p_\text{F}$, i.e. a M/F ratio of $0.0451/0.0540 = 0.790$ and a 95% c.i. for this:

$$0.790 \overset{\times}{\div} \exp\left(1.96 \times \sqrt{\frac{1}{25} - \frac{1}{554} + \frac{1}{47} - \frac{1}{870}}\right) = 0.790 \overset{\times}{\div} 1.60 = (0.492; 1.269)$$

Thus the male-female ratio of prevalences is not significantly different from 1. In fact the data would support most claims from a 50% smaller prevalence among men (M/F-ratio 0.5) to a 25% higher (M/F-ratio 1.25).

2.4.4 Odds-ratio from two proportions

Instead of using prevalence (proprotion affected) we may use the *odds* as the basis. The odds is defined as the fraction of affected relative to the fraction of non-affected:

$$\text{odds} = \omega = \frac{p}{1-p}$$

The ratio of odds from two groups is called an *odds-ratio*. Odds and odds-ratios have a particular role in the analysis of *case-control studies*.

Formulae

If we have two proportions and want to compare the odds, we use the *odds-ratio*: First work out the *odds* in each of the groups:

$$\omega_1 = \frac{p_1}{1-p_1} = \frac{x_1}{n_1 - x_1} \qquad \omega_0 = \frac{p_0}{1-p_0} = \frac{x_0}{n_0 - x_0}$$

The *odds-ratio*, OR, is then:

$$\text{OR} = \frac{\omega_1}{\omega_0} = \frac{x_1}{n_1 - x_1} \Big/ \frac{x_0}{n_1 - x_0}$$

The *error factor* for an OR is:

$$\text{ef} = \exp\left(1.96 \times \sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_0} + \frac{1}{n_0 - x_0}}\right)$$

and so the 95% c.i. for the OR becomes:

$$\text{OR} \overset{\times}{\div} \text{ef}$$

The term inside the expression for the error factor is the standard error of the $\log(\text{OR})$, so a test for $\text{OR} = 1$ must be transformed to a test of $\log(\text{OR}) = 0$, so the test statistic becomes

$$\chi^2(1) = \left(\frac{\log(\text{OR}) - 0}{\sqrt{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_0} + \frac{1}{n_0 - x_0}}} \right)^2 = \frac{\log(\text{OR})^2}{\frac{1}{x_1} + \frac{1}{n_1 - x_1} + \frac{1}{x_0} + \frac{1}{n_0 - x_0}}$$

Example

In a case-control study of risk factors for community acquired pneumococcal bacteriaemia Thomsen *et al.* [8] performed a population based case-control study and found the following:

	Pneumococcal bacteriaemia	
	Cases	Controls
Diabetic	53	298
Non-diabetic	545	5682

The odds-ratio of bacteriaemia between DM and non-DM subjects is:

$$\text{OR} = \frac{53/298}{545/5682} = 1.854$$

and the error-factor is:

$$\text{ef} = \exp\left(1.96 \times \sqrt{\frac{1}{53} + \frac{1}{298} + \frac{1}{545} + \frac{1}{5682}} \right) = 1.357$$

and hence a 95% c.i. for the OR is:

$$1.854 \overset{\times}{\div} 1.357 = (1.367, 2.516)$$

Thus the presence of diabetes approximately doubles the risk of pneumococcal bacteriaemia, but we cannot be sure that the increase is only about 40%, or that the increase is 2.5 fold.

A test for the null hypothesis OR=1 is performed by taking the square of the log-OR and dividing by the variance:

$$\chi^2(1) = \frac{\log(1.854)^2}{\frac{1}{53} + \frac{1}{298} + \frac{1}{545} + \frac{1}{5682}} = 15.732$$

which gives a p-value of 0.000072.

Exercise

Chaturvedi *et al.* [1] performed a case-control study of risk factors for amputation among different ethnic groups of diabetics and found the following overall result:

	Amputation	
Ethnicity	Cases	Controls
South Asians	8	47
Europeans	164	329

Compute the odds-ratio of amputation between the two ethnic groups, with a 95% confidence interval.

Compute the χ^2 test statistic for the hypothesis of no association between etnicity and amputation.

The authors found that the effect of ethnicity was even larger after adjusting for age, but smaller (and non-significant) after additionally adjusting for peripheral vascular disease, neuropathy and smoking. Discuss which of the two ways of controlling / adjusting for other variables you would prefer.

2.5 Rates

Estimation of a rate requires observation of two quantities:

D : Number of affected persons.

Y : Amount of risk time (number of person-years).

Formulae

The rate is estimated as:

$$\lambda = \frac{D}{Y}$$

The error factor for the rate is:

$$\text{ef} = \exp\bigl(1.96/\sqrt{D}\bigr)$$

and so a 95% confidence interval:

$$\lambda \overset{\times}{\div} \text{ef} = \lambda \overset{\times}{\div} \exp\bigl(1.96/\sqrt{D}\bigr)$$

2.5.1 Example

In the paper by Hu *et al.* [4] a cohort of Finnish persons were followed up for the occurrence of T2D.

A total of 373 cases were recorded in the course of 88045 person-years. The rate of T2D occurrence is thus:

$$\begin{aligned} \lambda = 373/88,045 \text{ cases/person-year} &= 0.004235 \text{ cases/person-year} \\ &= 4.235 \text{ cases/1000 person-year} \end{aligned}$$

The 95% confidence interval of the rate is:

$$4.235 \overset{\times}{\div} \exp(1.96/\sqrt{373}) = 4.235 \overset{\times}{\div} 1.107 = (3.828; 4.689) \text{ cases/1000 person-years}$$

2.5.2 Exercises

In the paper [4] there is a subdivision of the male part of the cohort by the amount of occupational physical acitivity:

	Cases	Person-years	Rate per 1000 p-y	95% c.i.
Light	97	29,216		
Moderate	32	18,874		
Active	71	32,955		

Compute the rates and the confidence intervals in each of the three groups.

2.6 Two rates

2.6.1 Formulae

The rate-ratio between two rates λ_1 and λ_0 is estimated by the ratio between the empirical rates:

$$\text{RR} = \frac{D_1/Y_1}{D_0/Y_0}$$

The error factor for the rate ratio is:

$$\text{ef} = \exp\left(1.96 \times \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}\right)$$

and thus a 95% confidence interval is:

$$\text{RR} \overset{\times}{\div} \exp\left(1.96 \times \sqrt{\frac{1}{D_1} + \frac{1}{D_0}}\right)$$

The term $\sqrt{\frac{1}{D_1} + \frac{1}{D_0}}$ inside the expression for the error factor is the standard error of the log (natural log) of the rate ratio. Hence a χ^2 test with 1 d.f. for equality of two rates is:

$$\chi^2(1) = (\log(\text{RR})/\text{s.e.}[\log(\text{RR})])^2 = \frac{\log(\text{RR})^2}{1/D_1 + 1/D_0}$$

2.6.2 Example

In the paper by Hu *et al.* [4], there were 32 cases of T2D during 18,874 person-years among in the group with moderate physical exercise at work and 71 cases during 32,955 person years in the group with active physical exercise. We find for the comparison of T2D incidence rates between persons with moderate and persons with active occupational work a RR of:

$$\text{RR} = \frac{32/18,874}{71/32,955} = 0.787$$

and:

$$\text{s.e.}(\log(\text{RR})) = \sqrt{\frac{1}{32} + \frac{1}{71}} = 0.213 \quad \Rightarrow \quad \text{erf} = \exp(1.96 \times 0.213) = 1.52$$

so the 95% confidence interval for this RR is:

$$0.787 \overset{\times}{\div} 1.52 = (0.52, 1.19)$$

The confidence interval contains 1, so there is no significant difference between the rates in the two groups. But it can be excluded that those with moderate physical exercise has more than a 20% elevated rate of T2D.

A test for the null hypothesis that $\text{RR} = 1$ or $\log(\text{RR}) = 0$ is

$$\chi^2(1) = \left(\frac{\log(\text{RR}) - \log(1)}{1/D_0 + 1/D_1}\right)^2 = \frac{-0.2395^2}{1/32 + 1/71} = 1.26$$

which corresponds to a p-value of 0.26.

Thus there is no difference in the T2D rates between people with moderate and active occupational physical exercise.

2.6.3 Exercise

Compute the estimates and 95% c.i.s for the relative risk comparing persons with light occupational work on one side and the two other groups (moderate and active) on the other side (using light occupational work as the reference):

	Cases	Person-years	Rate	Rate ratio	95% c.i.
Light	97	29,216		1.00 (ref.)	—
Moderate	32	18,874			
Active	71	32,955			

2.7 Power and precision

Power and precision are two different aspects of determination of study size. Power focuses on the p-value alone and thus mixes study size and anticipated effect, whereas precision disregards the p-value.

2.7.1 Power

Computation of power is to answer a question of the type:

- How large a study do we need to be able to detect a relative risk of 2?
- *Given* that the *true* effect is RR $= 2$, how do we make sure that the probability of rejecting RR $= 1$ is 80%.

It has nothing to do with the size of the actual estimate that comes out of the study.
A power calculation requires four quantities to be decided on:

1. α, the significance level used. This is usually taken to be 5%.
2. δ, the distance to the null hypothesis (i.e. the alternative).
3. n, the sample size.
4. γ, the power: P $\{$Rejecting the null hypotesis $\mid \delta,\ n,\ \alpha\}$ (in the example above, 80%). This is usually taken to be 80 or 90%.

For a given test situation, if three of these are known, the fourth can always be computed, but remember:

- The basis for power calculations is p-values, and hence hypothesis testing and null hypoteses.
- Power calculation is strictly a pre-study tool.

  Post-hoc power calculations are meaningless. When data are available, the information is in the data, **not** in computations on what otherwise might have happened under different assumptions.
- Power computations may be useful as guidelines to whether a given sample size is adequate.

In realistic situations, power calculations are mathematically extremely complicated.

Instead one may use "brute force": Simulate a dataset under the assumed conditions on the computer and compute test (i.e p-value). Repeat the procedure, say 1000 times. The fraction of times the test is significant is the power.

2.7.2 Precision

This concerns the question: How precisely do we want to determine the quantity under study, i.e. prevalence, rate ratio or other measure.

If we concentrate on the *precision* of the quantity under study (prevalence, rate ratio, number needed to treat,...), the question is largely reduced to one of the width of the confidence interval.

Thus the quantities needed to compute the precision are:

1. The confidence level to be used, $1 - \alpha$, normally 95%.
2. θ, the effect anticipated.
3. n, the sample size.
4. δ, half the width of the confidence interval, normally taken as the difference (or ratio) of the upper limit and the estimate, i.e. the confidence interval is assumed to be $\theta \pm \delta$ or $\theta \overset{\times}{\div} \delta$, depending on the measure.

There are two advantages of this over power calculations:

1. Given the target of inference (prevalence, rate-ratio,...), explicit standards for the desired precision of results must be defined.
2. The resulting answers are normally not very dependent on the assumed effect.

2.7.3 Prevalence studies

In a prevalence study which is mainly descriptive there is not much point in computing power as there is no natural null hypothesis to test.

For prevalence studies of one population we have the expression for error factor we use in the calculation of a confidence interval, and hence for evaluation of the precision.

Thus for any combination of study size (n) and prevalence (p) we can compute the width of the confidence interval for the prevalence, either in absolute terms (upper limit $-$ estimate) or relative terms (upper limit/estimate).

These are illustrated in figure 2.2.

2.7.4 Comparing two proportions

In the case of a simple comparison of two groups we are comparing two proportions. The precision (and power to detect a difference) will depend on the prevalence of the outcome in the unexposed (placebo, reference) group, the sizes of the two groups and the effect of exposure (treatment, intervention).

There are three different ways of showing precision and power for comparison of two proportions, as shown in figure 2.2.

The mian message of this plot is that regradless of whether one is using RR, OR or probability difference, there is a steep increase in power with an increasing effect, but an almost constant precision. This is because we are looking at a situation with a fixed number of subjects in each groups.

To choose a precision which is the same as the assumed effect, means aiming at a study which gives an estimate with a lower bound equal to the null value (OR=1, say). This will correspond to a power of 50%.

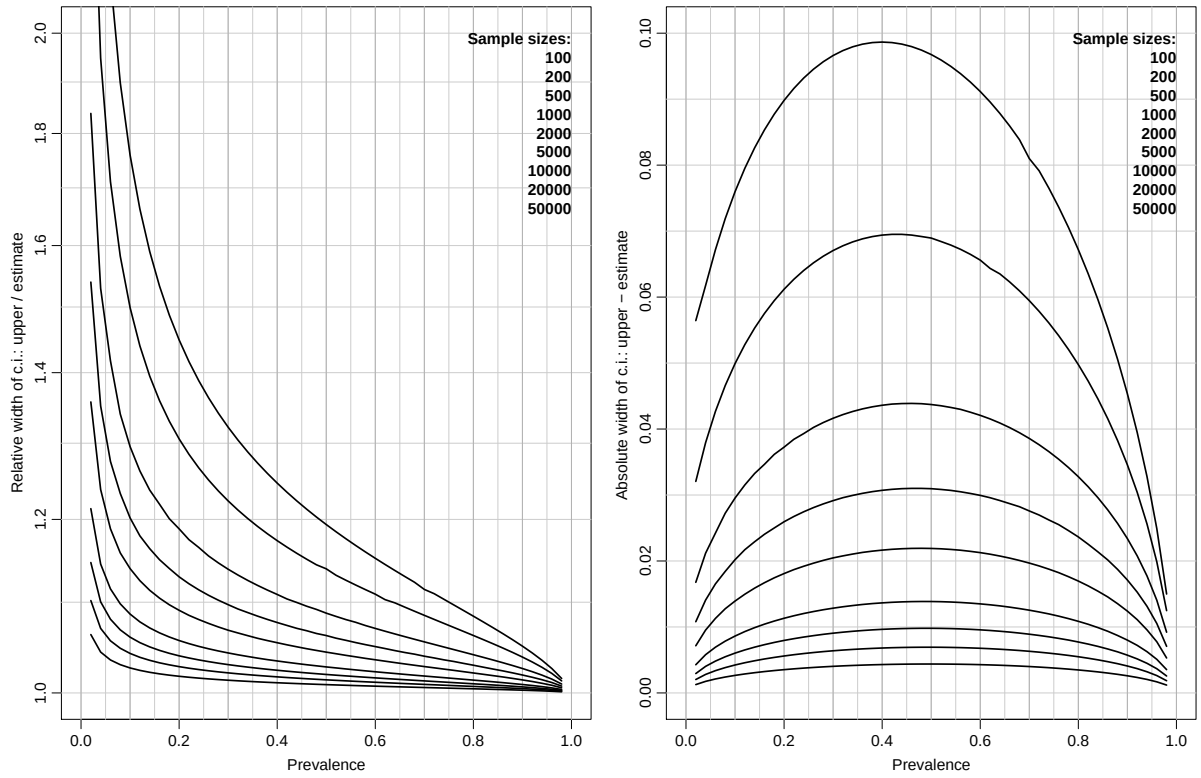

Figure 2.1: *Precision of prevalence studies: For example for a study with 1000 persons (4th curve from top), a prevalence of 30% will give a relative precision of 1.1 (left panel), i.e. the ratio of upper bound and the estimate is 1.1, corresponding to an absolute precison of 0.03, i.e. the difference between upper bound of the c.i. and the estimate is 0.03 (right panel, 4th curve from top).*

2.7.5 Case-control studies

A special case of comparing two proportions is the case-control study. The parameters determining the precision and power is the numer of cases and controls, and the prevalence of exposure in the controls (i.e. in the population).

In figure 2.3 is shown precision and power for case-control studes with an equal number of cases and controls, and in figure 2.4 for case-control studies with three times as many controls as cases.

2.7.6 Cohort studies

In cohort studies the main interest is in comparison of groups. The drawings in figure 2.5 gives power and precision for cohort studies of two groups where the follow-up time (person-years) is the same in both groups. The parameter used to describe the size of the study is the number of expected cases in the unexposed group (i.e. the cumulative incidence rate over the follow-up period for the reference group).

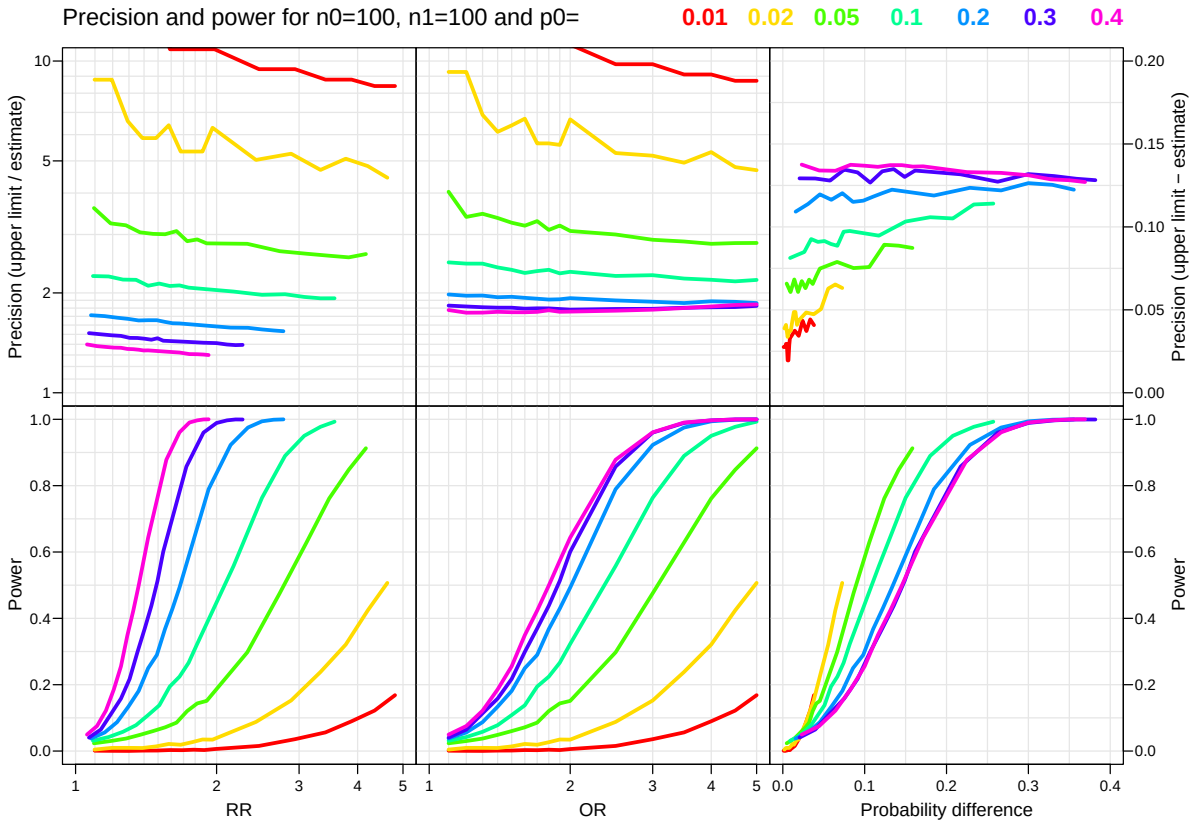

Figure 2.2: *Precision (upper panels) and power (lower panels) for comparing two proportions based on 100 persons in each group.*

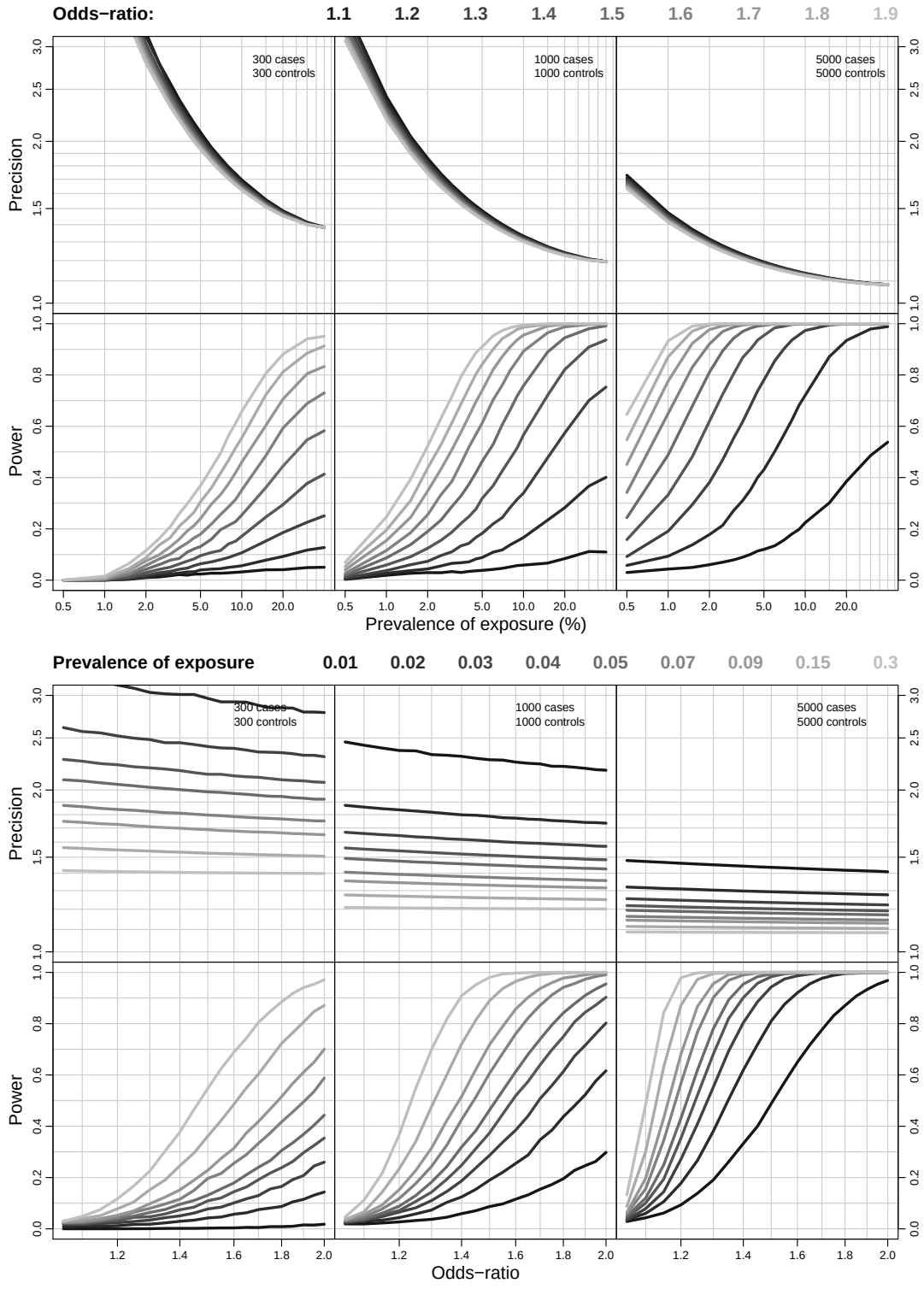

Figure 2.3: *Case-control studies with equal number of cases and controls. Precision in estimation of the OR and power to detect an OR different from 1, plotted against prevalence of exposure (top frame) against odds-ratio (bottom frame)*

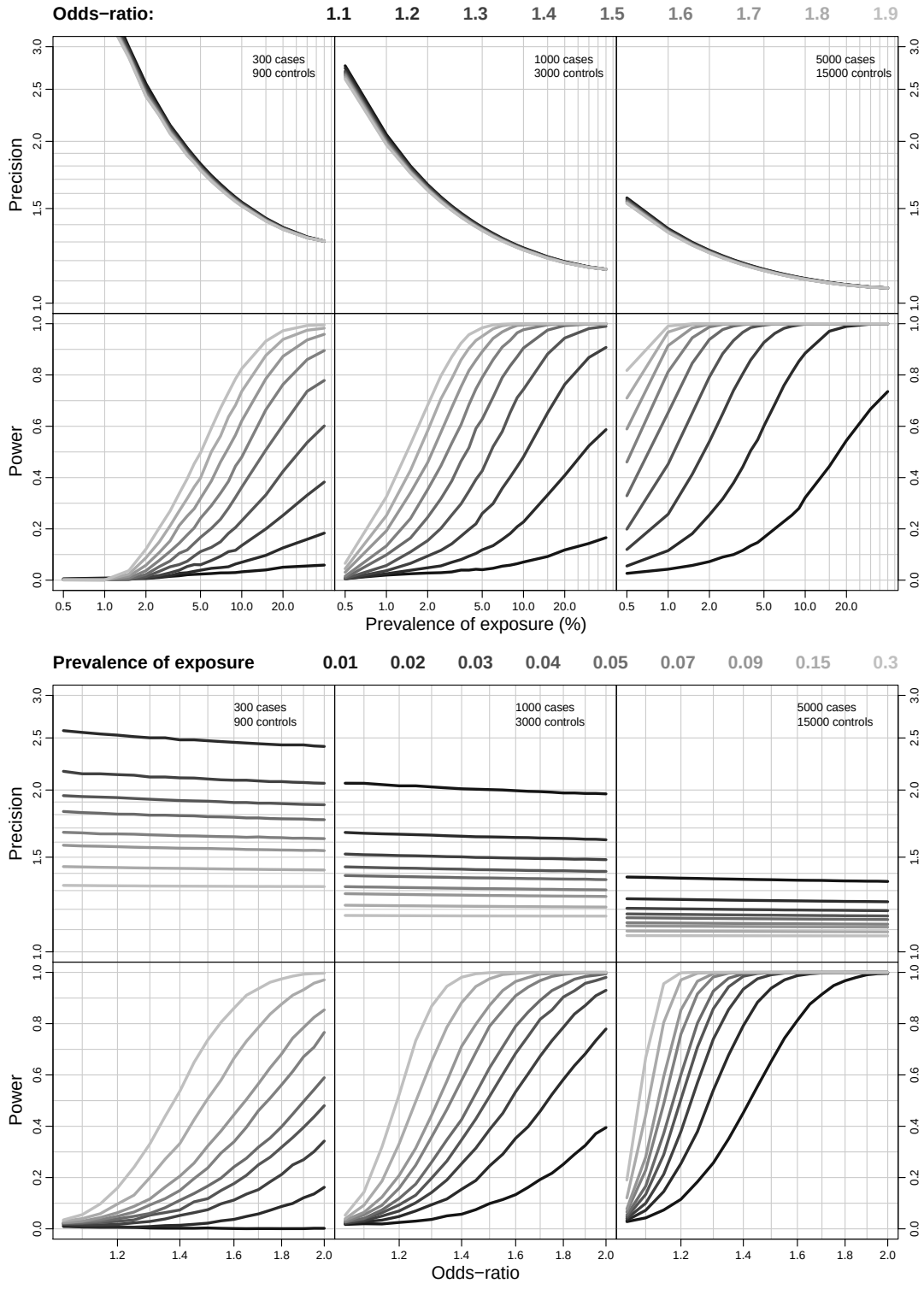

Figure 2.4: *Case-control studies with unequal number of cases and controls. Precision in estimation of the OR and power to detect an OR different from 1, plotted against prevalence of exposure (top frame) against odds-ratio (bottom frame)*

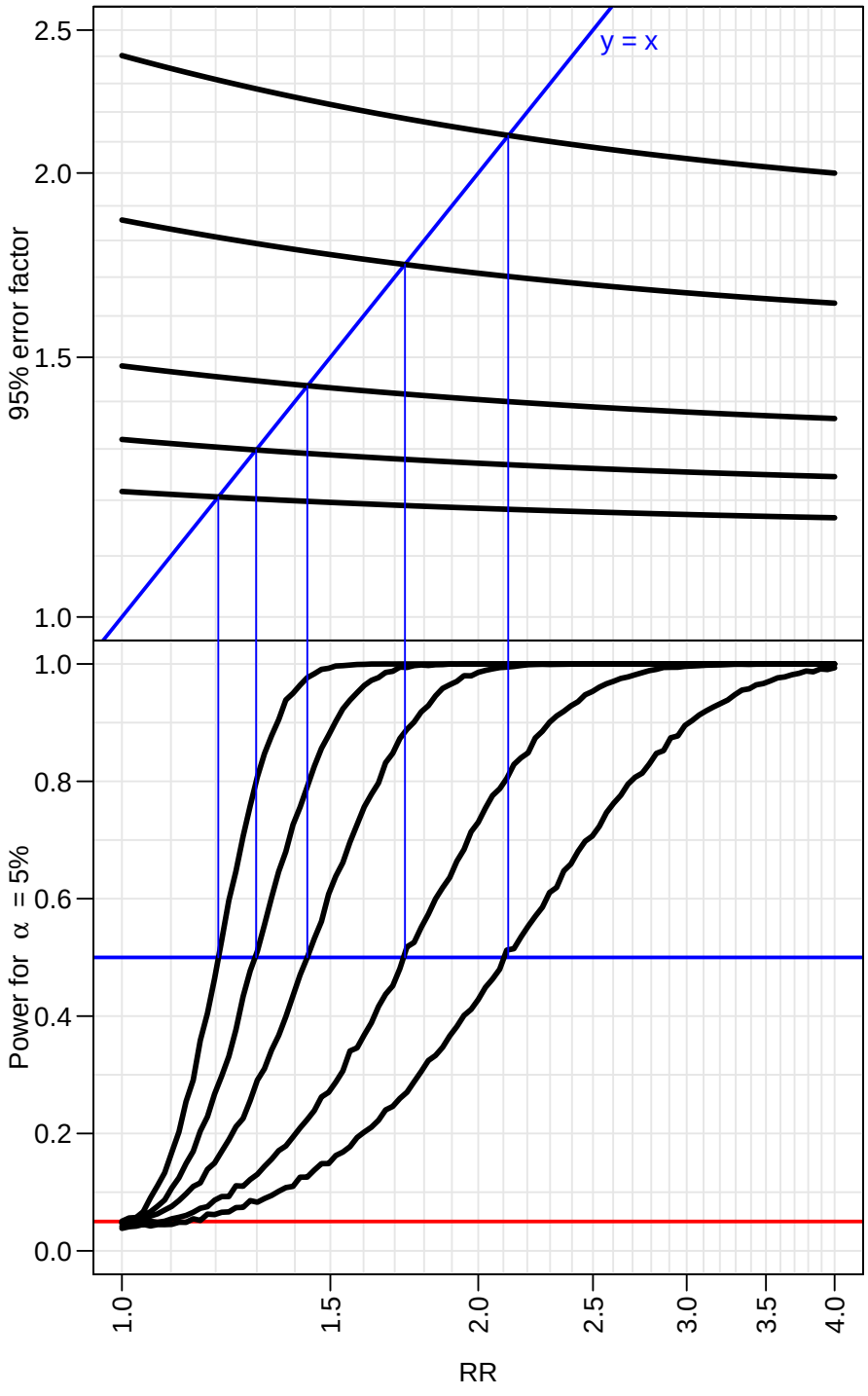

Figure 2.5: *Precision (top) and power (bottom) to detect a RR different from 1, assuming that the amount of follow-up in both groups is the same in both groups and that the expected number of cases in the unexposed group is 10, 20, 50, 100 and 200 respectively.*
If a study is designed to have a precision equal to the effect, i.e. so that the lower bound of the c.i. is 1 for agiven effect, then the power to detect this effect is 50%.

Chapter 3

Mathematical concepts

3.1 Logarithms and exponentials

3.1.1 Powers

$$
\begin{aligned}
10^2 &= 10 \times 10 \\
10^3 &= 10 \times 10 \times 10 \\
10^2 \times 10^3 &= 10^5 \\
10^3/10^2 &= 10^1 \\
(10^3)^2 &= 10^6
\end{aligned}
$$

$$
\begin{aligned}
10^2/10^2 &= 10^0 = 1 \\
10^2/10^3 &= 10^{-1} = 1/10 \\
10^{1/2} \times 10^{1/2} &= 10^1 \\
10^{1/2} &= \sqrt{10}
\end{aligned}
$$

3.1.2 Multiplication and division with logarithms

Base e ($\mathrm{e} = 2.718281828459045$):

$$
\begin{aligned}
\log(xy) &= \log(x) + \log(y) \\
\log(x/y) &= \log(x) - \log(y) \\
\log(x^a) &= a\log(x) \\
\log(1/x) &= -\log(x) \\
\log(1) &= 0 \\
\log(\mathrm{e}) &= 1
\end{aligned}
$$

Base 10:

$$
\begin{aligned}
\log_{10}(xy) &= \log_{10}(x) + \log_{10}(y) \\
\log_{10}(x/y) &= \log_{10}(x) - \log_{10}(y) \\
\log_{10}(x^a) &= a\log_{10}(x) \\
\log_{10}(1/x) &= -\log_{10}(x) \\
\log_{10}(1) &= 0 \\
\log_{10}(10) &= 1
\end{aligned}
$$

3.1.3 Base for exponential functions

Base e: $\mathrm{e}^x = \exp(x)$

$$
\begin{aligned}
\mathrm{e}^x \times \mathrm{e}^y &= \mathrm{e}^{x+y} \\
\mathrm{e}^x/\mathrm{e}^y &= \mathrm{e}^{x-y} \\
(\mathrm{e}^x)^y &= \mathrm{e}^{x\times y} \\
1/\mathrm{e}^x &= \mathrm{e}^{-x} \\
\mathrm{e}^1 &= \mathrm{e} \\
\mathrm{e}^0 &= 1
\end{aligned}
$$

Base 10:

$$
\begin{aligned}
10^x \times 10^y &= 10^{x+y} \\
10^x/10^y &= 10^{x-y} \\
(10^x)^y &= 10^{x\times y} \\
1/10^x &= 10^{-x} \\
10^1 &= 10 \\
10^0 &= 1
\end{aligned}
$$

Bibliography

[1] N Chaturvedi, CA Abbott, A Whalley, P Widdows, SY Leggetter, and JM Boulton. Risk of diabetes-related amputation in south asians vs. europeans in the UK. *Diabetic Medicine*, 19:99–104, 2002.

[2] C Glümer, T Jørgensen, and K Borch-Johnsen. Prevalences of diabetes and impaired glucose regulation in a Danish population. *Diabetes Care*, 26(8):2335–2340, August 2003.

[3] YH Hamid, SA Urhammer, DP Jensen, C Glümer, K Borch-Johnsen, T Jørgensen, T Hansen, and O Pedersen. Variation in the Interleukin-6 receptor gene associates with type 2 diabetes in Danish whites. *Diabetes*, 53:3342–3345, 2004.

[4] G Hu, Q Qiao, K Silventoinen, JG Eriksson, P Jousilahti, J Lindstöm, TT Valle, A Nissinen, and J Tuomilehto. Occupational, commuting and leisure-time physical activity in relation to risk for type 2. *Diabetologia*, 46:322–329, 2003.

[5] M Kalani, J Apelqvist, M Blombäck, K Brismar, B Eliasson, B Fagrell JW Eriksson, A Hamsten, O Torffvit, and G Jörneskog. Effect of dalteparin on healing of chronic foot ulcers in diabetic patients with peripheral arterial occlusive disease. *Diabetes Care*, 26:2575–2580, 2003.

[6] DL Sackett. Bias in analytic research. *Journal of Chonic Diseases*, 32(1–2):51–63, 1979.

[7] DL Streiner and GR Norman. *PDQ Epidemiology*. B.C.Decker, Hamilton, London, 2nd ed. edition, 1998.

[8] RW Thomsen, HH Hundborg, H-H Lervang, SP Johnsen, HC Schønheyder, and HT Sørensen. Risk of community-acquired pneumococcal bacteremia in patients with diabetes. *Diabetes Care*, 27:1143–1147, 2004.

Chapter 4

Solution to exercises

4.1 One proportion

In the chinese population survey there were 554 males and 870 females. Among the diabetic individuals there were 25 males and 47 females.

Hence the prevalences in each of the two genders are:

$$\text{Males:} \quad p_{\text{M}} = \frac{25}{554} = 0.0451 \qquad \text{Females:} \quad p_{\text{F}} = \frac{47}{870} = 0.0540$$

To compute a 95% confidence interval for each of these prevalences we use the formulas:

$$\text{ef} = \exp\bigl(1.96/\sqrt{np(1-p)}\bigr) \qquad \text{c.i.:} \quad \frac{p}{p + (1-p) \overset{\times}{\div} \text{ef}}$$

which gives the following for males:

$$\begin{aligned} \text{ef} &= \exp(1.96/\sqrt{554 \times 0.0451 \times 0.9549}) = 1.494 \\ \text{c.i.:} &\qquad \frac{0.0451}{0.0451 + 0.9549 \overset{\times}{\div} 1.494} = (0.0307; 0.0659) \end{aligned}$$

and for females:

$$\begin{aligned} \text{ef} &= \exp(1.96/\sqrt{870 \times 0.0540 \times 0.9460}) = 1.342 \\ \text{c.i.:} &\qquad \frac{0.0540}{0.0540 + 0.9460 \overset{\times}{\div} 1.342} = (0.0408; 0.0711) \end{aligned}$$

4.2 Two proportions

4.2.1 Difference in proportions: Avoiding amputation

The table of results from the study by Kalani *et al.* [5] is, when reduced to a question of amputation or not:

	Treatment	
Amputation	Dalterapin	Placebo
Yes	2	8
No	41	34
	43	42

$$p_{\text{Dal}} = \frac{2}{43} = 0.0465 \qquad p_{\text{Pl}} = \frac{8}{42} = 0.1905$$

Difference in amputation probabilities is the fraction of the patients that benefit from treatment: $p_{\text{Pl}} - p_{\text{Dal}}$. Note the order of subtraction — we are discussing the probaility of the *adverse* outcome (amputation).

$$\begin{aligned} p_{\text{Pl}} - p_{\text{Dal}} &= 0.1440 \\ \text{s.e.}[p_{\text{Pl}} - p_{\text{Dal}}] &= \sqrt{\frac{p_{\text{Dal}}(1-p_{\text{Dal}})}{n_{\text{Dal}}} + \frac{p_{\text{Pl}}(1-p_{\text{Pl}})}{n_{\text{Pl}}}} \\ &= 0.0686 \\ 95\%\text{ c.i.} \quad : \quad & 0.1440 \pm 1.96 \times 0.0686 = (0.0096; 0.2784) \approx (1\%, 28\%) \end{aligned}$$

The number needed to treat in order to avoid one amputation is

$$n_{\text{ntt}} = 1/0.1440 = 6.94$$

and a confidence interval for this is:

$$1/(0.01; 0.28) = (3.5, 100)$$

So it looks as if the treatment is effective but the uncertainty of the effect size w.r.t. avoiding amputation is rather large.

4.2.2 Odds-ratio from case-control study.

In a case-control study of risk factors for amputation among different ethnic groups of diabtics, Chaturvedi *et al.* [1] performed a case-control study and found the following overall result:

	Amputation	
Ethnicity	Cases	Controls
South Asians	8	47
Europeans	164	329

The odds-ratio of amputation between the two ethnic groups is:

$$\text{OR} = \frac{8/47}{164/329} = 0.3415$$

and the error-factor is:

$$\text{ef} = \exp\left(1.96 \times \sqrt{\frac{1}{8} + \frac{1}{47} + \frac{1}{164} + \frac{1}{329}}\right) = 2.166$$

and hence a 95% c.i. is

$$0.3415 \overset{\times}{\div} 2.166 = (0.158, 0.739)$$

A test of the null hypothesis OR=1 is performed by taking the square of the log-OR and dividing by the variance:

$$\chi^2(1) = \frac{\log(0.3415)^2}{\frac{1}{8} + \frac{1}{47} + \frac{1}{164} + \frac{1}{329}} = 7.428$$

which gives a p-value of 0.0064.

Adjusting for confounding

In the present case it seems entirely reasonable to adjust for the potential confounding by age.

Adjusting for the variables peripheral vascular disease and neuropathy may not seem so reasonabale, because they may be viewed either as precursors of the outcome, amputation or as other manifestations of the same underlying propensity to vascular disease that leads to amputation and which apparently differe between ethnic groups. Thus this may be viewed as over-adjustment.

This argument is clearly not tenable in the case of smoking, so one might argue that the relevant adjustment would be only for age and smoking, which in the article is reported to give an OR of 0.38 (0.15–0.95), p=0.04.

4.3 Rates

In the paper [4] there is a subdivision of the male part of the cohort by the amount of occupational physical acitivity as in the table below.

The rate in the "Light" groups is $97/29,216 = 3.320$ cases per 1000 person-years. The 95% error-factor for this rate is:

$$\text{ef} = \exp(1.96/sqrt97) = 1.220$$

so the confidence interval becomes $3.320 \overset{\times}{\div} 1.220 = c(2.721, 4.051)$

Similar calculations for the other two groups gives the results in the table below:

	Cases	Person -years	ef	Rate per 1000 p-y	95% c.i.
Light	97	29,216	1.22	3.32	(2.72; 4.05)
Moderate	32	18,874	1.41	1.70	(1.20; 2.40)
Active	71	32,955	1.26	2.15	(1.71; 2.72)

4.3.1 Comparing rates

The estimates for the rate ratio relative to "light occupational work" is computed by computing first the rates as shown in colum 3 in the table.

The rate-ratio for "Moderate" is $1.695/3.320 = 0.511$, and the error-factor for this is:

$$\text{ef} = \exp\big(1.96 * \sqrt{\frac{1}{97} + \frac{1}{32}}\big) = 1.491$$

so the 95% c.i. is $0.511 \overset{\times}{\div} 1.491 = (0.342, 0.761)$ Similarly, we find for comparison of "Active" with "Light" RR $= 2.154/3.320 = 0.649$ and the error factor:

$$\text{ef} = \exp\big(1.96 * \sqrt{\frac{1}{97} + \frac{1}{71}}\big) = 1.358$$

so the 95% c.i. is $0.649 \overset{\times}{\div} 1.358 = (0.478, 0.881)$.

The results are summarized in the table below:

	Cases	Person -years	Rate	Rate ratio	95% c.i.
Light	97	29,216	3.320	1.00	—
Moderate	32	18,874	1.695	0.51	(0.34; 0.76)
Active	71	32,955	2.154	0.65	(0.48; 0.88)