Lecture notes on epidemiology

The aims of epidemiology:

Epidemiology is the study of the distribution and determinants of health related states or events in specified populations and the application of this study to control health problems (Mould).

Typically one tries to find associations between risk factors (exposures) and diseases by studying the occurrence of diseases in populations. The final aim is to conclude casual relations.

Examples of epidemiological findings (from Rothman):

- Smoking and lung cancer
- Smoking and other health related events
- Passive smoking
- Low-level ionising radiation and leukaemia
- Saccharin and bladder cancer
- Swine flu and Guillain-Barré syndrome
- The effect of diethylstilbestrol on offspring
- Tampons and toxic-shock syndrome
- Coffee drinking and pancreatic cancer

Epidemiological studies are not generally amenable to being investigated by randomised trials and observational studies are therefore the more practical to study factors and exposure which can not be controlled by the investigators (Mould).
Hill’s criteria for distinguishing between causal and non-causal associations:

1. Strength (the association should be sufficiently strong)

2. Consistency (it should be seen in different populations under different circumstances)

3. Specificity (a cause should be specific to its effect, e.g. smoking leads to a specific form of lung cancers)

4. Temporality (cause precedes effect)

5. Biological gradient (greater dose should give larger effect)

6. Plausibility (seen in animal models)

7. Coherence (not in conflict to known biology)

8. Experimental evidence (if possible a randomised experiment should be carried through)

9. Analogy (the cause make sense)

**Type of studies:**

Studies may be classified as prospective or retrospective. In a prospective study measures of exposures and covariates are made before illness occur. In a retrospective study these measurements are made after the case have already occurred.

- Cross-sectional study

  A study that includes all persons, in the population, at the time of ascertainment, or a representative sample of such persons. Disease and exposure are observed simultaneously.
• Cohort study

Two groups (or more) groups of people that are free from the disease and that differ according to the extent of their exposure to a potential cause of the diseases are studied during a time interval.

• Case-Control study

A case control study includes people with a disease and a suitable control group of people unaffected by the disease. The occurrence of the possible cause (exposure) is compared between cases and controls. A relevant situation is to think of a case-control study as a cohort study where exposure is investigated for all cases and for a sample of the non-cases.

**Measures of disease frequency:**

**Prevalence** is the number of cases of a disease in a defined population. The prevalence can be measured by a prevalence rate, i.e., the proportion of the population that has the disease at a specific time. This can be measured by in a cross-sectional study.

**Incidence** is the number of new cases of disease in a population. The incidence can only be measured by considering a population during a time interval. It can be measured by the incidence rate. The incidence rate is defined as

\[
\text{(The number of new cases in the population during the time considered)/(the total number of years that people are under risk to develop the disease)}.
\]

**Cumulative incidence** is the number of people in a cohort that gets the disease during the time of study.

The relation between prevalence and incidence is somewhat complicated. It depends on the duration of the disease state. An approximate relation is

\[
P = I D
\]

Where D is the mean duration which diseased people carry the disease.
If D is increased (e.g. the mortality is lowered by new medicines or if the disease is diagnosed at an earlier state) the prevalence rate increases even if the incidence rate is the same.

Epidemiological are most often concerned with the study of incidences. The main aim is to compare incidences for exposed and non-exposed people.

**Measures of effect and association:**

Incidence rates in exposed and non-exposed people can be compared in several way. Obvious alternatives are differences and ratios. Let \( I_1 \) be the incidence rate for exposed people and \( I_0 \) be the incidence rate of non-exposed people. The the two alternatives are

\[
I_1 - I_0
\]

and

\[
I_1/I_0.
\]

If \( p_1 \) is the probability that an exposed person gets ill during a time interval and \( p_0 \) is the corresponding probability that an unexposed person gets ill then we can measure the difference either as a risk difference

\[
p_1 - p_0,
\]

Or a risk ratio

\[
p_1/p_0.
\]

Still another alternative is the odds ratio

\[
\frac{p_1/(1-p_1)}{p_0/(1-p_0)}.
\]

These measures gives different results in all cases other then \( p_1 = p_0 \). In that case the difference is 0 and the risk ratio and the odds ratio are both equal to 1.

It should be observed that these comparisons say little of the absolute values of the risk.

Other measures (which relates to a certain population) is attributable risk or etiological fraction.
**Statistical analysis:**

**One strata:**

In the simplest situation the outcome of a cohort study can be summarized in a 2x2-table:

<table>
<thead>
<tr>
<th></th>
<th>exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ill</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>Not ill</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td></td>
<td>$n_{-1}$</td>
<td>$n_{-2}$</td>
</tr>
</tbody>
</table>

An interesting hypothesis is that the risks of getting ill are the same for exposed and non-exposed people. This corresponds to a $\chi^2$-test of homogeneity (or Fisher’s exact test). If this hypothesis is rejected it is of interest to estimate the effect of the exposure. One effect measure is the odds ratio. The odds-ratio is estimated by the cross-product ratio:

$$\frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Most common is however to estimate the logarithm of the odds ratio. The ML-estimate is

$$\ln(n_{11}) - \ln(n_{12}) - \ln(n_{21}) + \ln(n_{22})$$

If the observations in the cells are large it is possible to prove that this estimate is asymptotically normal distributed with the true value of the logarithm of the odds ratio as mean and an asymptotic variance that can be estimated by

$$\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

This can be used to calculate an approximate confidence interval for the logarithm of the odds ratio by the following formula

$$\ln(n_{11}) - \ln(n_{12}) - \ln(n_{21}) + \ln(n_{22}) \pm z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

From this an approximate confidence region for the log odds ratio can be derived as
The reason for calculate estimates and confidence intervals for log odds instead of the odds directly is that the normal approximation often works better for the log of the cross-product ratio than for the cross-product ratio itself.

In a case-control study the observations can also be summarized in a 2x2 table, but with a different interpretation:

<table>
<thead>
<tr>
<th></th>
<th>exposed</th>
<th>Not exposed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>n_{11}</td>
<td>n_{12}</td>
<td>n_{1.}</td>
</tr>
<tr>
<td>controls</td>
<td>n_{21}</td>
<td>n_{22}</td>
<td>n_{2.}</td>
</tr>
<tr>
<td></td>
<td>n_{.1}</td>
<td>n_{.2}</td>
<td>n_{..}</td>
</tr>
</tbody>
</table>

The same analysis can be applied. Observe that in a case-control study you can estimate the odds-ratio but not the risk ratio. However if the risk is small these measure are almost identical.

**Several strata:**

Sometimes the analysis has to be stratified into different groups of people. In that case we have to summarize the observations into several 2x2-tables (one for each strata). The reason for this is that there may be some other factor that relates to the exposure and to the disease that we want to control for. Making a separate analysis for strata in which such factors are constant does this. Typical stratification variables are sex, age, smoking, drinking and other life style factors. Such factor may either influence the effect as being either effect modifiers or confounders.

Confounding is a systematic bias in the estimate of the effect that follows from that the cases and controls differs with regard to a risk factor not considered in the study. The analysis have to consider this possibility as is seen by the following example

**Stratum 1**

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Controls</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>130</td>
</tr>
</tbody>
</table>
This gives the estimated odds ratio 4.0

Stratum 2

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Not exposed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>8</td>
<td>92</td>
<td>100</td>
</tr>
<tr>
<td>Controls</td>
<td>2</td>
<td>98</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>190</td>
<td>200</td>
</tr>
</tbody>
</table>

This gives the estimated odds ratio 4.3

Summing these two tables gives

Strata 1 + 2

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Not exposed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>58</td>
<td>142</td>
<td>200</td>
</tr>
<tr>
<td>Controls</td>
<td>22</td>
<td>178</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>320</td>
<td>400</td>
</tr>
</tbody>
</table>

Which gives the estimated odds ratio 3.3.

That it not possible to do a fair result by just summarising (collapsing) over strata is referred to as non-collapsibility. It is easy to construct other examples that give even more deviating estimates.

If the effect is different in different strata this effect modification has to be described and analysed. Formulating a model for how the effect is modified by various covariates may do this. If the effect is the same in all strata there is need for a statistical method to do testing and estimating simultaneously for several 2x2-tables:

The null hypothesis of no effect in any of the strata can be tested by a so-called Mantel-Haenszel test.
Stratum 1:

<table>
<thead>
<tr>
<th></th>
<th>exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>$a_1$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>controls</td>
<td>$c_1$</td>
<td>$d_1$</td>
</tr>
</tbody>
</table>

...

Stratum i

<table>
<thead>
<tr>
<th></th>
<th>exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>$a_i$</td>
<td>$b_i$</td>
</tr>
<tr>
<td>controls</td>
<td>$c_i$</td>
<td>$d_i$</td>
</tr>
</tbody>
</table>

...

Stratum n

<table>
<thead>
<tr>
<th></th>
<th>exposed</th>
<th>Not exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>$a_n$</td>
<td>$b_n$</td>
</tr>
<tr>
<td>controls</td>
<td>$c_n$</td>
<td>$d_n$</td>
</tr>
</tbody>
</table>

The observed number of exposed cases is

$$O = \sum_i a_i$$

The expected number of exposed cases is (according to hypergeometric distribution in each strata)

$$E = \sum_i \frac{(a_i + b_i)(a_i + c_i)}{a_i + b_i + c_i + d_i}$$

The variance of $O$ is (according to the hypergeometric distribution)

$$V = \sum_i \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{t_i^2(t_i - 1)}$$
Where

\[ t_i = a_i + b_i + c_i + d_i \]

The null hypothesis of no effect in all strata is rejected if

\[ \frac{(O - E)^2}{V} \]

is large compared to a \( \chi^2(1) \)-percentile.

**Matched case-control studies:**

In order to minimise the possibility of introducing confounders epidemiological studies are sometimes matched. This means that controls are selected specific to each case (or a group of cases) in a way that makes them as close to the cases as possible as regards covariates that are not under study. An typical examples of this is given in Rice section 13.5. This describes a 1-1 matching. It is also possible to have more controls, so called 1-m matching. Data from matched studies should be analysed by methods for simultaneous analysis of several 2x2-tables. Two remarks:

- In matched studies the precision gained by increasing the number of controls to each cases decreases rather fast.
- There may be a disadvantage of matched studies since it prevents any study of the effect of the variables that has been used to define the matching.

**Analysis of matched case-control studies**

The methods for estimating matched case-control studies are the same as for analysing several 2x2-tables where same effect is assumed in all strata. In a matched study each case define a stratum. An estimate of the common effect is obtained through the Mantel-Haenszel estimator:

\[
\frac{\sum a_i d_i / t_i}{\sum b_i c_i / t_i}
\]
Logistic regression

In many situations covariates can be attached to each individual under study. This means that we do not only observe the Bernoulli variable $Y$ which can take the values 0 and 1 according to if the individual is diseases or not but we also have other measurements ($x_1, \ldots, x_k$) which may influence the risk of getting ill. In such cases there is need for a model that describes the influence of the covariates (independent variables) on the distribution of $Y$. Such a model which is similar to normal linear regression assumes that

$$
Pr(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}
$$

This model is called a logistic regression model. Observe that the log odds is a linear combination of the covariates, i.e.,

$$\ln \left( \frac{Pr(Y = 1)}{Pr(Y = 0)} \right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

One of the covariates may be exposure, either defined as a 0/1-variable (exposed or not) or as a dose (that measure the amount of exposure). In such case the corresponding regression coefficient measures the effect of the exposure. In fact the estimate of the regression coefficient is the logarithm of the odds ratio effect of exposure that differs with one unit of the scale in which the exposure is measured in.

Conditional logistic regression

In matched case control studies we have a case with corresponding covariates and one or more controls with possible different covariates. We can then use a logistic regression model to calculate a conditional probability that says how probable it is that among a number of people (the case and the controls) with a certain set of covariate values it is that just the case have the disease under study.

Let consider a 1-1 matching situation and assume that the case has covariates $x$ and the control covariate $z$. (For simplicity we assume a one-dimensional covariate.) Then we can calculate the conditional probability that the person with covariates $x$ have the disease given that exactly one of the two persons have the disease. Elementary (but not simple) calculations give the result
\[
\frac{\exp(\beta_1 x)}{\exp(\beta_1 x) + \exp(\beta_1 z)}.
\]

The statistical analysis of the matched data can be analysed using these probabilities to build up a likelihood. ML estimates can be found and likelihood ratio test derived.

**Possible sources of Bias in epidemiological studies:**

- Confounding
- Selection bias (e.g. Healthy worker effect)
- Recall Bias (cases may report exposures more accurately than controls)
- Publication Bias (significant results tends to be published easier than studies with significance is not found)