Basic statistics

Categorical data

Thomas Alexander Gerds

Categorical data: examples

- Binary (yes/no)
  - exposure
    - male/female
    - exposed/non-exposed
    - positive test/negative test

- More classes
  - nominal
    - green/red/blue/yellow
    - hospitals, study centers, doctors
    - race
  - ordinal
    - lightgreen/green/darkgreen
    - disease stage
    - age ≤ 40/40–50/50–60/≥ 60

Categorical outcome

Binary outcome

\[ Y = \begin{cases} 
1 & \text{event / positive / disease} \\
0 & \text{no event / negative / non-disease} 
\end{cases} \]

Parameters

- **Prevalence:** proportion of the population with event at fixed time point.
  
  *How many have the disease right now?*

- **Incidence/hazard rate:** probability of event relative to time unit:
  
  *How many per year newly acquire the disease?*

- **Risk:** personalized probability that event occurs in given time period:
  
  *How likely will a given subject acquire the disease?*
Statistical inference

Estimating risks and prevalence

\[ \hat{p} = \text{Relative frequency} = \frac{\text{Number of events}}{\text{Number of subjects}} = \frac{x}{n} \]

Confidence limits (normal approximation)

\[ \left[ \hat{p} - 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} ; \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right] \]

Confidence limits (exact)

\[
\text{binom.test}(x,n)
\]

Example: Randomized clinical trial on Dalteparin

Response to dalteparin therapy of 85 diabetic patients with peripheral arterial occlusive disease and chronic foot ulcers in a placebo controlled randomized trial\(^1\).

Treatment groups:

Dalteparin (N=43), Placebo (N=42)

Outcome:

<table>
<thead>
<tr>
<th>Category</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>intact skin</td>
<td>healed</td>
</tr>
<tr>
<td>ulcer area decreased ≥ 50%</td>
<td>improved</td>
</tr>
<tr>
<td>decreased or increased ulcer area &lt; 50%</td>
<td>unchanged</td>
</tr>
<tr>
<td>increased ulcer area ≥ 50%</td>
<td>impaired</td>
</tr>
<tr>
<td>amputation above/below ankle</td>
<td>amputation</td>
</tr>
</tbody>
</table>

\(^1\) Kalani et al. *Diabetes Care* 26: 2575-2580, 2003

Frequency table

<table>
<thead>
<tr>
<th></th>
<th>Dalteparin</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healed</td>
<td>14 (33%)</td>
<td>9 (21%)</td>
</tr>
<tr>
<td>Improved</td>
<td>15 (35%)</td>
<td>11 (26%)</td>
</tr>
<tr>
<td>Unchanged</td>
<td>7 (16%)</td>
<td>9 (21%)</td>
</tr>
<tr>
<td>Impaired</td>
<td>5 (12%)</td>
<td>5 (12%)</td>
</tr>
<tr>
<td>Amputation</td>
<td>2 (5%)</td>
<td>8 (19%)</td>
</tr>
<tr>
<td>total (100%)</td>
<td>43</td>
<td>42</td>
</tr>
</tbody>
</table>

Warm-up exercises

1. Estimate the risks of the event (amputation or increased ulcer area) separately in both treatment groups and supply the results with exact 95% confidence intervals. (there is a R-tutorial called Exact-binomial-confidence-limits).

2. Write down the interpretation of the confidence intervals. Do the confidence intervals indicate a significant difference?

3. Include the middle category (decreased or increased ulcer area) in the event and repeat the whole statistical stunt (1. and 2.).
Barplot

Here we pool the outcome categories as follows:

<table>
<thead>
<tr>
<th>Category</th>
<th>Dichotomized outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>intact skin</td>
<td>better</td>
</tr>
<tr>
<td>ulcer area decreased $\geq 50%$</td>
<td></td>
</tr>
<tr>
<td>decreased or increased ulcer area $&lt; 50%$</td>
<td></td>
</tr>
<tr>
<td>increased ulcer area $\geq 50%$</td>
<td>worse</td>
</tr>
<tr>
<td>amputation above/below ankle</td>
<td></td>
</tr>
</tbody>
</table>

Group comparison

Dalteparin group

Risk of worse outcome $= \frac{14}{43} = \hat{p}_1$

Placebo group

Risk of worse outcome $= \frac{22}{42} = \hat{p}_2$

Association measures

Relative risk: $\frac{\hat{p}_1}{\hat{p}_2}$
Odds ratio: $\frac{\frac{\hat{p}_1}{1-\hat{p}_1}}{\frac{\hat{p}_2}{1-\hat{p}_2}}$
Risk difference: $\hat{p}_1 - \hat{p}_2$
2x2 contingency table

<table>
<thead>
<tr>
<th>Response</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>no</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>total</td>
<td>a+c</td>
<td>b+d</td>
<td>N</td>
</tr>
</tbody>
</table>

Risk estimates

$$\hat{p}_1 = \frac{a}{a+b} \quad \hat{p}_2 = \frac{c}{c+d}$$

Relative risk

$$\hat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

Standard error of log(\(\hat{RR}\)) and confidence interval

$$\hat{\sigma} = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$$

$$CI_{95\%} = [\hat{RR} \exp(-1.96 \hat{\sigma}); \hat{RR} \exp(1.96 \hat{\sigma})]$$

Relative risk: placebo versus dalteparin

$$\hat{RR} = \frac{22/42}{14/43} = 1.609$$

Standard error of log(\(\hat{RR}\)) and confidence interval

$$\hat{\sigma} = \sqrt{\frac{1}{22} - \frac{1}{42} + \frac{1}{14} - \frac{1}{43}} = 0.264$$

$$CI_{95\%} = [0.959; 2.7] \text{ (does include 1)}$$
Odds

Needed for
▶ case-control studies
▶ logistisk regression

Difficult to interprete, but if risks are small, then risks \(\approx\) odds.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0013</td>
<td>0.0013</td>
</tr>
<tr>
<td>0.025</td>
<td>0.0256</td>
</tr>
<tr>
<td>0.17</td>
<td>0.2048</td>
</tr>
<tr>
<td>0.39</td>
<td>0.6393</td>
</tr>
</tbody>
</table>

Can compute odds from risks and back
▶ \(\text{odds} = \frac{p}{1-p}\)
▶ \(p = \frac{\text{odds}}{1+\text{odds}}\)

Odds ratio: placebo versus dalteparin

\[
\hat{OR} = \frac{22 \cdot 29}{14 \cdot 20} = 2.279
\]

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Outcome</th>
<th>worse</th>
<th>better</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>placebo</td>
<td></td>
<td>22</td>
<td>20</td>
<td>42</td>
</tr>
<tr>
<td>dalteparin</td>
<td></td>
<td>14</td>
<td>29</td>
<td>43</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>36</td>
<td>49</td>
<td>85</td>
</tr>
</tbody>
</table>

Standard error of \(\log(\hat{OR})\) and confidence interval

\[
\hat{\sigma} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = 0.449
\]

\(Cl_{95\%} = [0.946; 5.491] \; \text{(does include 1)}\)
Reporting results

The relative risk of group 1 (Dalteparin) versus group 2 (Placebo) is estimated as

\[
\frac{14}{43} \div \frac{22}{42} = 0.622
\]

Equivalent statements:

- The risk in group 2 is reduced by a factor 0.622 compared to group 1.
- The risk in group 2 is 37.8% lower than in group 1.
- The risk in group 1 is 1.609 times higher than in group 2.
- The risk in group 1 is 60.9% higher than in group 2.

Testing independence in a randomized clinical trial

Null hypothesis: the treatment has no effect.

\[
\text{Prob}(\text{worse given dalteparin}) = \text{Prob}(\text{worse given placebo}) \\
\iff p_1 = p_2 \\
\iff p_1 = 1 \\
\iff \frac{p_1}{1 - p_1} = 1 \\
\iff \frac{p_1}{1 - p_2} = \frac{1}{1 - p_2}
\]

Tests of independence between the treatment group and the outcome groups:

- \(\chi^2\) test (normal approximation)
- Fisher’s exact test

The \(\chi^2\) test statistic

\[
\chi^2 = \frac{(\text{observed counts} - \text{expected counts})^2}{\text{expected counts}}
\]

Observed counts

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Responses</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>no</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

The expected counts are calculated under the null hypothesis.

Rule of thumb: a valid analysis requires that all expected counts are \(\geq 5\).
Test results

Null hypothesis: dalteparin treatment has no effect for chronic foot ulcers.

<table>
<thead>
<tr>
<th>Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher’s exact test</td>
<td>0.0808</td>
</tr>
<tr>
<td>Pearson's $\chi^2$ test</td>
<td>0.0644</td>
</tr>
<tr>
<td>Pearson’s $\chi^2$ test with Yates’ continuity correction</td>
<td>0.1032</td>
</tr>
</tbody>
</table>

```
tab <- matrix(c(22,20,14,29),ncol=2,byrow=TRUE)
fisher.test(tab)
chisq.test(tab,correct=FALSE)
chisq.test(tab,correct=TRUE)
```

Exercise

Compute the expected counts under the null hypothesis of the following two tables:

<table>
<thead>
<tr>
<th></th>
<th>event</th>
<th>no event</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>non-exposed</td>
<td>6</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>event</th>
<th>no event</th>
</tr>
</thead>
<tbody>
<tr>
<td>exposed</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>non-exposed</td>
<td>18</td>
<td>10</td>
</tr>
</tbody>
</table>

Hint: see help(chisq.test)

Larger contingency tables

If the table is not 2x2 but, e.g., 3x4 or 5x2, the $\chi^2$ test and Fisher’s exact test are testing an ANOVA null hypothesis.

For example,

```
tab <- matrix(c(14,15,7,5,2,9,11,9,5,8),ncol=2)
fisher.test(tab)
```

Laboratory experiment: diffuse large B cell lymphoma (DLBCL)

Group 1: "LP type"
33 cases of Nodular lymphocyte predominant Hodgkin lymphoma

Group 2: conventional DLBCL
41 de novo DLBCL samples from Finnish men

The aim of the present study was to characterize morphology and immunophenotype of NLPHL and compare it with conventional DLBCL.
Comparison of tumor types

<table>
<thead>
<tr>
<th></th>
<th>EMA negative</th>
<th>EMA positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>conventional DLBCL</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td>LP type</td>
<td>9</td>
<td>24</td>
</tr>
</tbody>
</table>

Based on these data, is the probability to see EMA positive higher in conventional DLBCL compared to "LP type"?

Exercises

- Make a barplot which shows the numbers of the contingency table
  → R-tutorial: Bar plots

- Calculate and compare the probabilities of EMA positive. Write down conclusion statements which include 95% confidence interval(s) and p-value(s).
  → R-tutorial: 2x2 tables

Confounding

A confounding variable is an extraneous variable in a research model that should have been experimentally controlled, but was not.

Failing to take a confounding variable into account can lead to a false conclusion that the dependent variables are in a causal relationship with the independent variable.
Simpson’s paradox

Suppose that in addition to the outcome and the exposure group a categorical confounder variable (e.g., gender) is measured for each individual.

- **Subgroup analysis**
  Analyze 2x2 contingency tables separately in each strata defined by the confounder variable.

- **Cochran-Mantel-Haenszel test**
  Computes a weighted average of the subgroup analyses.

- **Logistic regression**
  Similar results. Applicable also with continuous confounders.

Observational study design

In a **cohort study**, an outcome or disease-free study population is first identified by an exposure (e.g., onset of diabetes) or other inclusion criteria and followed in time until the disease or outcome of interest occurs.

**Case-control** studies identify subjects by outcome status at the outset of the investigation. Once outcome status is identified and subjects are categorized as cases. For each case a given number of controls (e.g., 4) are selected. A candidate control is a subject without the outcome but from the same source population.

Cohort study

For example consider the multiple risk factor intervention trial (MRFIT) including more than 300000 men. Table from Kannel et al. (2003)²

<table>
<thead>
<tr>
<th>Blood pressure</th>
<th>yes</th>
<th>no</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 140</td>
<td>6122</td>
<td>259479</td>
<td>265601</td>
</tr>
<tr>
<td>≥ 140</td>
<td>4927</td>
<td>77450</td>
<td>82377</td>
</tr>
<tr>
<td>total</td>
<td>11049</td>
<td>336929</td>
<td>347978</td>
</tr>
</tbody>
</table>

Relative risk = 2.595 \ (C_{95\%} : [2.501; 2.692])

²Is the relation of systolic blood pressure to risk of cardiovascular disease continuous and graded, or are there critical values?, Hypertension. 2003;42:453.
Case-control study

The famous example:

<table>
<thead>
<tr>
<th>Smoker</th>
<th>Lung cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>no</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>total</td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

Since the number of controls (b+d) is defined by the study design it is not possible to estimate the prevalence of lung cancer in the population. The statistic $\hat{RR}$ depends also on the number of controls per case and should not be used for measuring association in case-control studies.

The statistic $\hat{OR}$ works.

Diagnostic study

A diagnosis is an estimate of the patient’s current status

A prediction is an estimate of the patient’s future status

The estimates can be based on patient’s genotype, phenotype and exposure history.

Medical test

A medical diagnostic test is a decision rule

\[ X = \begin{cases} 
1 & \text{positive / disease} \\
0 & \text{negative / non-disease} 
\end{cases} \]

The test can be based on a biomarker. A biomarker is any biological measurement made on a patient which is related to the disease status, extent, or activity.

Example: screening for prostate cancer

The first commercial PSA (Prostate Specific Antigen) test:

\[ X = \begin{cases} 
1 & \text{positive if PSA > 4.0 ng/mL} \\
0 & \text{negative if PSA \leq 4.0 ng/mL} 
\end{cases} \]
Example: screening for prostate cancer

The first commercial PSA (Prostate Specific Antigen) test:

\[
X = \begin{cases}
1 & \text{positive if PSA} > 4.0\,\text{ng/mL} \\
0 & \text{negative if PSA} \leq 4.0\,\text{ng/mL}
\end{cases}
\]

▶ The reference range of serum PSA is 0.0–4.0 ng/mL (based on a study of 472 healthy men where 99% had a total PSA level below 4 ng/mL).

▶ There are some that feel that this level should be lowered to 2.5 ng/ml in order to detect more cases of prostate cancer.

Sensitivity and Specificity

\(Y\): Outcome (disease status) E.g. prostate cancer

\(X\): Test result (biomarker) E.g. PSA test

\[
\begin{array}{c|c|c}
Y = 1 & Y = 0 \\
\hline
X = 0 & \text{False negative} & \text{True negative} \\
X = 1 & \text{True positive} & \text{False positive}
\end{array}
\]

▶ True positive rate (sensitivity): \(P(X = 1 \mid Y = 1)\)
▶ True negative rate (specificity): \(P(X = 0 \mid Y = 0)\)

Predictive Values

The patient is not directly interested in sensitivity and specificity but in the predictive values of the test.

The positive predictive value:

\[
P(Y = 1 \mid X = 1)
\]

The probability that a patient with a positive test is diseased

The negative predictive value:

\[
P(Y = 0 \mid X = 0)
\]

The probability that a patient with a negative test is not diseased
Paired data

Example: rater agreement

<table>
<thead>
<tr>
<th>Diagnostic tool II</th>
<th>positive</th>
<th>negative</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>negative</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>total</td>
<td>a+c</td>
<td>b+d</td>
<td>N</td>
</tr>
</tbody>
</table>

Keywords: Cohen’s Kappa coefficient, McNemar’s test

Similar: two doctors, before and after measurement on each patient, matched case control design, ...

Take home messages

2x2 contingency tables are used to assess associations between exposure and risk, medical tests, and rater agreement.

When risk is very low or very high then odds ratios and relative risks are close.

A case-control study cannot estimate the disease prevalence and hence also not parameters that depend on the disease prevalence such as the relative risk.