Statistical methods in public health

Tommi Härkänen

National Institute for Health and Welfare (THL)
Department of Health (TERO)

September 1, 2015

Contents

Course outline

Background

Example data sets

Binary outcome

Conditional probability and Bayes rule

Comparing two groups

Outline

The lectures and (computer) excersises will be given in C128 (2pm to 6pm) from September 1 to October 13. Course materials are available via wiki and Moodle pages.

- Different study designs
- Binary model
- Survival analysis
- Regression analysis
- Graphical models
- Case control studies

Practical work is used for the evaluation.

- The assignment of the practical work will be available via Moodle (October 13).

Epidemiology (short definition)

British Medical Journal (BMJ): "Epidemiology is the study of **how often diseases occur** in different groups of people and **why**."

What is it used for? "Epidemiological information is used to plan and evaluate **strategies to prevent** illness and as a guide **to the management** of patients in whom disease has already developed."

Study designs

Cross-sectional All data are observed at one point in time

Cohort Group of individuals are measured at *baseline*, and after that they are followed up a certain period of time until the *outcome* occurs

Repeated measurements As in cohort study, but individuals are measured one or more times after the baseline

Retrospective designs E.g. in case-control studies the outcome is measured first, and after the the covariates (risk factors etc.) are measured retrospectively

Randomized studies Most reliable to assess causal effects, but cannot be applied in many research areas due to economic or ethical reasons.

Statistical tools

Descriptive statistics Mean, prevalence, ...

Description of the outcome Incidence, ...

Associations Bivariate measures, regression models, ...

Data sets

Health and well-being for residents (ATH) (Cross-sectional study.) Data representing the Finnish population on, for instance, residents' lifestyles and experiences (not found in registers) were collected [1].

Framingham Heart Study (Repeated measurements cohort study.) The objective was to identify the common factors or characteristics that contribute to cardiovascular disease (CVD).
5,209 men and women between the ages of 30 and 62 were recruited from the town of Framingham, Massachusetts in 1948[2]. The open data set[3] is used here.

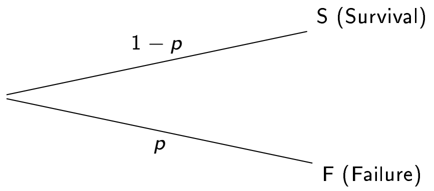
[1] https://www.thl.fi/fi/web/thlfi-en/research-and-expertwork/population-studies/ath-health-a

[2] https://www.framinghamheartstudy.org/about-fhs/history.php

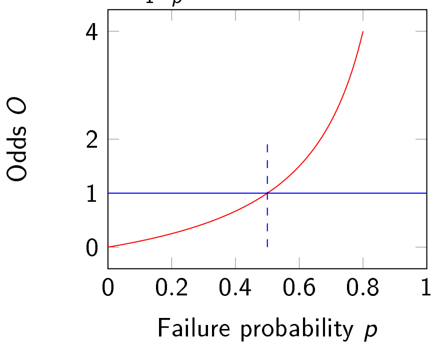
[3] http://www.stat.ncsu.edu/courses/sibs/datasets/framingham/, not suitable for research

Binary model

Probability of failure is p.



Odds: $O := \frac{p}{1-p}$:



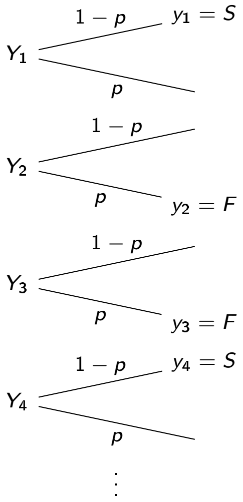
Likelihood

Assume that we have observed n for individuals a binary outcome $y = (y_1, y_2, \dots, y_n)$, $y_i \in \{S, F\}$. For example, $n = 10$:

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
S	F	F	S	S	S	S	F	S	F

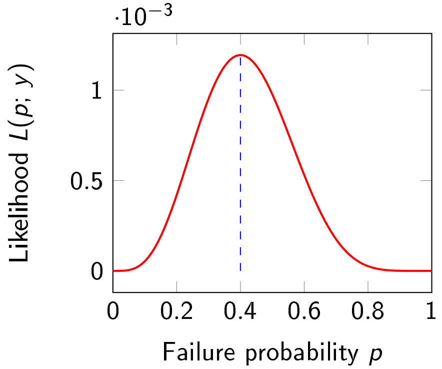
Failure probability $p = \mathbb{P}\{Y_i = F\}$ is unknown, but same for all individuals (identically distributed).
We also assume *independence* of observations.
The *likelihood* is a function of the possible parameter values (given the observations):

$$L(p;\, y) := (1-p) \times p \times p \times (1-p) \times (1-p)\times (1-p) \times (1-p) \times p \times (1-p) \times p = p^4 \times (1-p)^6 = p^{\sum_i \mathbb{1}\{y_i = "F"\}} \times (1-p)^{\sum_i \mathbb{1}\{y_i = "S"\}}$$



Maximum Likelihood

Binary data, 10 observations, 4 failures (cases):



Value $p = 0.4$ is where the likelihood attains it's maximum value and it is called the *maximum likelihood estimate* (MLE) $\hat{p}$.
The higher the likelihood value, the better *supported* by the data.

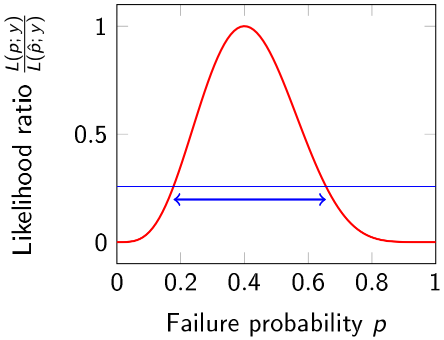
Supported values of the parameter

In the previous example MLE was $\hat{p} = 0.4$. We might be interested if the true (unknown) value of $p = 0.5$. Does the data support this?

Consider *likelihood ratio* (LR) $L(p;\, y)/L(\hat{p};\, y)$.

Supported values are called those p for which LR exceed a certain threshold value, e.g. 0.258.

How to choose the threshold?



Likelihood vs. frequentist inference

A key question in *likelihood inference* (above) is how to choose the threshold. There is no concensus.

Frequentist inference

Uncertainty is measured in terms of *probability*.

- Assume large number of *repeated "experiments"* in similar conditions (same sample size, same model, same parameter values, etc.)
- What is the *coverage probability* that the supported ranges of the experiments contain the true (unknown) parameter value?

Log likelihood

The normal model with continuous outcome $Y_i \in \mathbb{R}$:

$$\ell(\mu, \sigma^2; y) := \log L(\mu, \sigma^2; y) =$$

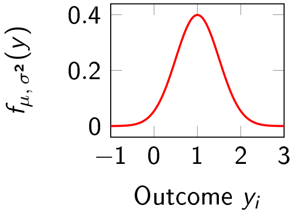
$$\log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \frac{(y_i-\mu)^2}{\sigma^2} \right\} =$$

$$\sum_{i=1}^{n} -\frac{1}{2} \log 2\pi - \frac{1}{2} \frac{(y_i-\mu)^2}{\sigma^2}$$

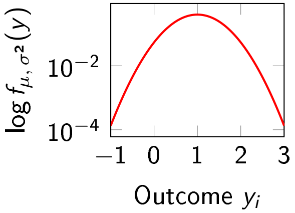
The logarithm of the likelihood is simpler to handle for several reasons:

- Finding the ML estimate using the gradient and finding the root is easy
- Avoids numerical instability (especially in case of large data sets)

Probability density, $\mu = 1$, $\sigma^2 = 0.25$:



Log density with $\mu = 1$, $\sigma^2 = 0.25$:



Coverage probability

Assuming a single Gaussian random variable Y (and σ^2 known) with observed value $y = \hat{\mu}$.
The log LR is

$$\log \text{LR} := \ell(\mu, \sigma^2; y) - \ell(\hat{\mu}, \sigma^2; y) = -\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} - -\frac{1}{2}\frac{(y-y)^2}{\sigma^2} = -\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}$$

As $Y \sim N(\mu, \sigma^2)$, the standardized Y follows the chi-squared distribution with one degrees of freedom: $((Y-\mu)/\sigma)^2 \sim \chi^2(1)$.

Supported range based on LR 0.258 is based on −1.35 for log LR.
The probability that the log LR values are above −1.35 is

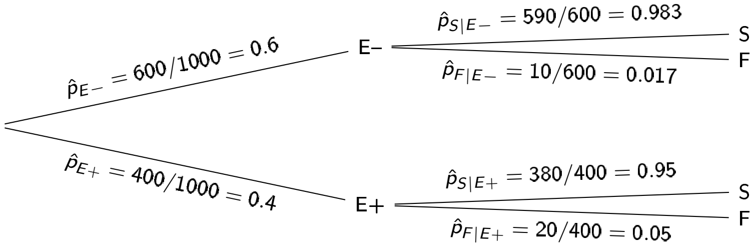
$$F_{\chi^2(1)}(-2 \times -1.35) = 0.9$$

This result **holds approximately for other log likelihoods** e.g. the binary model. One needs to find the ML estimate and standard deviation.

Conditional probability

Assume that some individuals have been *exposed* (E+) to a factor influencing the failure probability, and other have not (E–).
Observed (artificial) data:

	F	S	Sum
E+	20	380	400
E–	10	590	600
Sum	30	970	1000


$\hat{p}_{E-} = 600/1000 = 0.6$
E–
$\hat{p}_{S|E-} = 590/600 = 0.983$
S
F
$\hat{p}_{F|E-} = 10/600 = 0.017$
$\hat{p}_{E+} = 400/1000 = 0.4$
E+
$\hat{p}_{S|E+} = 380/400 = 0.95$
S
F
$\hat{p}_{F|E+} = 20/400 = 0.05$

Bayes rule

Assume that conditional probabilities $\mathbb{P}\{A \mid B\}$ are known, but we want "reversed" conditional probabilities $\mathbb{P}\{B \mid A\}$. Recall that

$$\mathbb{P}\{A \mid B\} = \frac{\mathbb{P}\{A, B\}}{\mathbb{P}\{B\}}.$$

We need also $\mathbb{P}\{B\}$ in order to calculate the joint distribution $\mathbb{P}\{A, B\} = \mathbb{P}\{A \mid B\}\mathbb{P}\{B\}$.

After that $\mathbb{P}\{A\}$ can be calculated, and

$$\mathbb{P}\{B \mid A\} = \frac{\mathbb{P}\{A, B\}}{\mathbb{P}\{A\}}.$$

Risk ratio (RR)

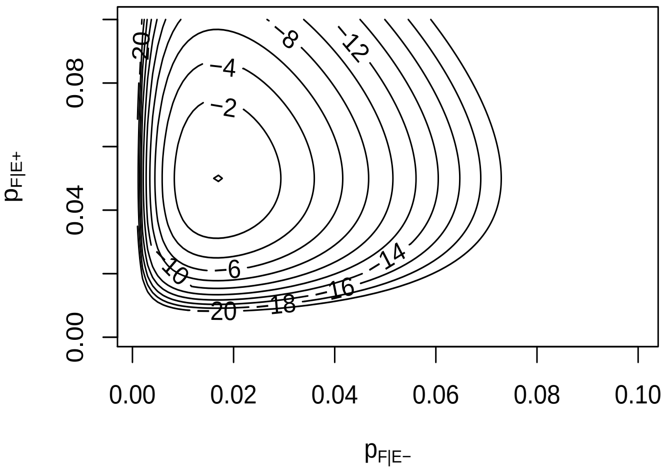
Commonly used statistic to illustrate the difference between the exposure groups is the risk ratio (RR):

$$RR := \frac{p_{F|E+}}{p_{F|E-}}$$

In the previous example the RR estimate is
$\widehat{\text{RR}} = \hat{p}_{F|E+}/\hat{p}_{F|E-} = 0.05/0.017 = 3$

Contour plot of two parameters

Log likelihood ratio



Reparameterization

We want to estimate the difference of the two groups using RR
$\theta := p_{F|E+}/p_{F|E-}$.
As $p_{F|E+} = \theta p_{F|E-}$, we still have two parameters:

Parameter of interest θ

Nuisance parameter $p_{F|E-}$

Recall the log likelihood

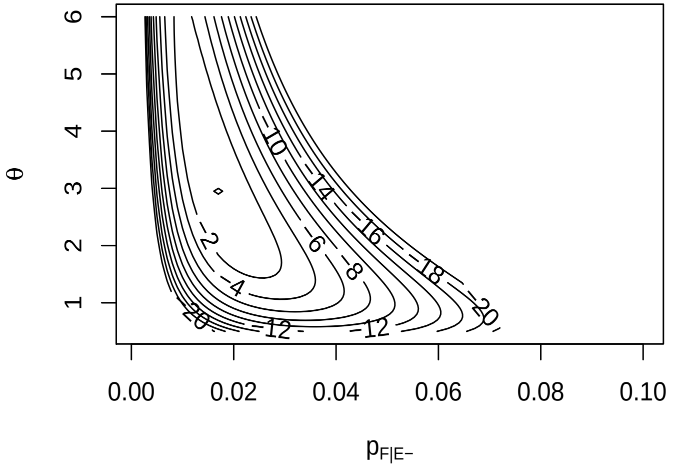
$$\ell(p_{F|E+}, p_{F|E-}; \text{data}) = n_{F|E+} \log p_{F|E+} + n_{F|E+} \log(1 - p_{F|E+}) + \\ n_{F|E-} \log p_{F|E-} + n_{F|E-} \log(1 - p_{F|E-}). \quad (1)$$

How to get rid of a nuisance parameter?

- Estimate it.
- Profile likelihood.
- (Conditional likelihood.)

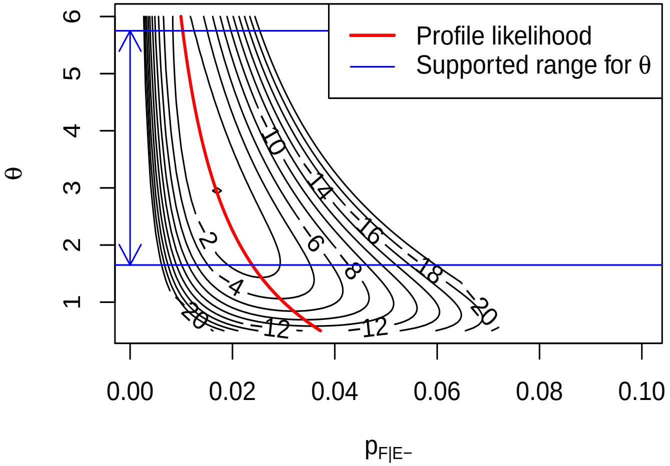
Log LR surface for the risk of unexposed and RR

$p_{F|E-}$ and $\theta := p_{F|E+}/p_{F|E-}$



Profile likelihood

Find $\hat{p}_{F|E-}$ for each θ


Profile likelihood
Supported range for θ
θ
1
2
3
4
5
6
0.00
0.02
0.04
0.06
0.08
0.10
$p_{F|E-}$
-2
-4
-6
-8
-10
-12
-12
-14
-16
-18
-20
-20