Epidemiologic Data Analysis using R
Part 2: Basic analysis of proportions

Janne Pitkäniemi
(Esa Läärä)

Finnish Cancer Registry, Finland, <janne.pitkaniemi@cancer.fi>
(University of Oulu, Finland, <esa.laara@oulu.fi>)

University of Tampere
Faculty of Social Sciences
Feb 26- Apr 9 2018

Contents

1. Binary outcomes and proportions
2. Comparative parameters of risks and their estimation
3. Binomial regression models and comparative parameters
4. Adjustment for confounding and evaluation of modification by binomial regression

Main R functions covered:

- `twoby2()` (Epi package)
- `glm()`
- `ci.lin()` (Epi package)

Outcomes in epidemiologic research

Epidemiologic studies address the occurrence of diseases and other health related phenomena:

(a) cross-sectional: **prevalence** of diseases,

(b) longitudinal: disease **incidence**, and mortality

Often we want to compare the prevalence or incidence of disease between two groups defined by a binary *risk factor* X

$$X = 1\text{: “exposed”}, \quad X = 0\text{: “unexposed”}$$

Types of outcome variables

- *Binary* (0/1) variables at individual level
  - disease *status* at a *time point*
  - *change* of status, *event* or *transition* (*e.g.* from healthy to diseased)
- *Proportions* at group level
  - prevalence
  - incidence proportion or cumulative incidence,
- *Rates* of events
  - incidence or mortality rate (per 1000 y)
  - car accidents (per million km)
- *Time* to event
  - survival time (often censored)

Incidence and prevalence proportions

- **Incidence proportion** (R) of a binary (0/1) outcome (disease, death etc.) over a fixed risk period is defined

$$R = \frac{D}{N} = \frac{\text{number of new cases during period}}{\text{size of population-at-risk at start}}$$

  Also called **cumulative incidence** (or even "risk").

  NB. This formula requires complete follow-up, i.e. no *censorings*, and absence of *competing risks*.

- **Prevalence (proportion)** P of disease at time point t

$$P = \frac{\text{no. of existing cases at } t}{\text{total population size at } t}.$$

Two-group comparison

- Binary risk factor X: exposed vs. unexposed.
- Summarizy results from cohort study with fixed risk period and no losses:

Exposure	Cases	Non-cases	Group size
yes	D_1	C_1	N_1
no	D_0	C_0	N_0
total	D_+	C_+	N_+

- Incidence proportions in the two exposure groups

$$R_1 = \frac{D_1}{N_1}, \qquad R_0 = \frac{D_0}{N_0}.$$

- These are crude *estimates* of the true *risks* π_1, and π_0 of outcome in the two exposure categories.

Example: Observational clinical study

Treatment failure in two types of operation for renal calculi (Charig *et al.* 1986. *BMJ* 292: 879-882)

- OS = open surgery (invasive)
- PN = percutaneous nephrolithotomy

Treatment group (j)	Failure (D_j)	Success (C_j)	Patients (N_j)	Failure-% (R_j)
OS ($j = 1$)	77	273	350	*22.0*
PN ($j = 0$)	60	290	350	*17.1*

Crude incidence proportions of treatment failure:

$$R_1 = 77/350 = 22.0\%, \qquad R_0 = 60/350 = 17.1\%$$

Risks and their comparative parameters

The **risk** or probability of binary outcome (*e.g.* new case of disease) in the "exposed" π_1 and in the "unexposed" π_0 as to binary risk factor X (values 1 and 0) are typically compared by

- **risk difference** $\theta = \pi_1 - \pi_0$
- **risk ratio** $\phi = \pi_1/\pi_0$
- **odds ratio** (risk odds ratio)

$$\psi = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

The odds ratio is close to the risk ratio when the risks are "small" (less than 0.1 – the "rare-disease assumption").

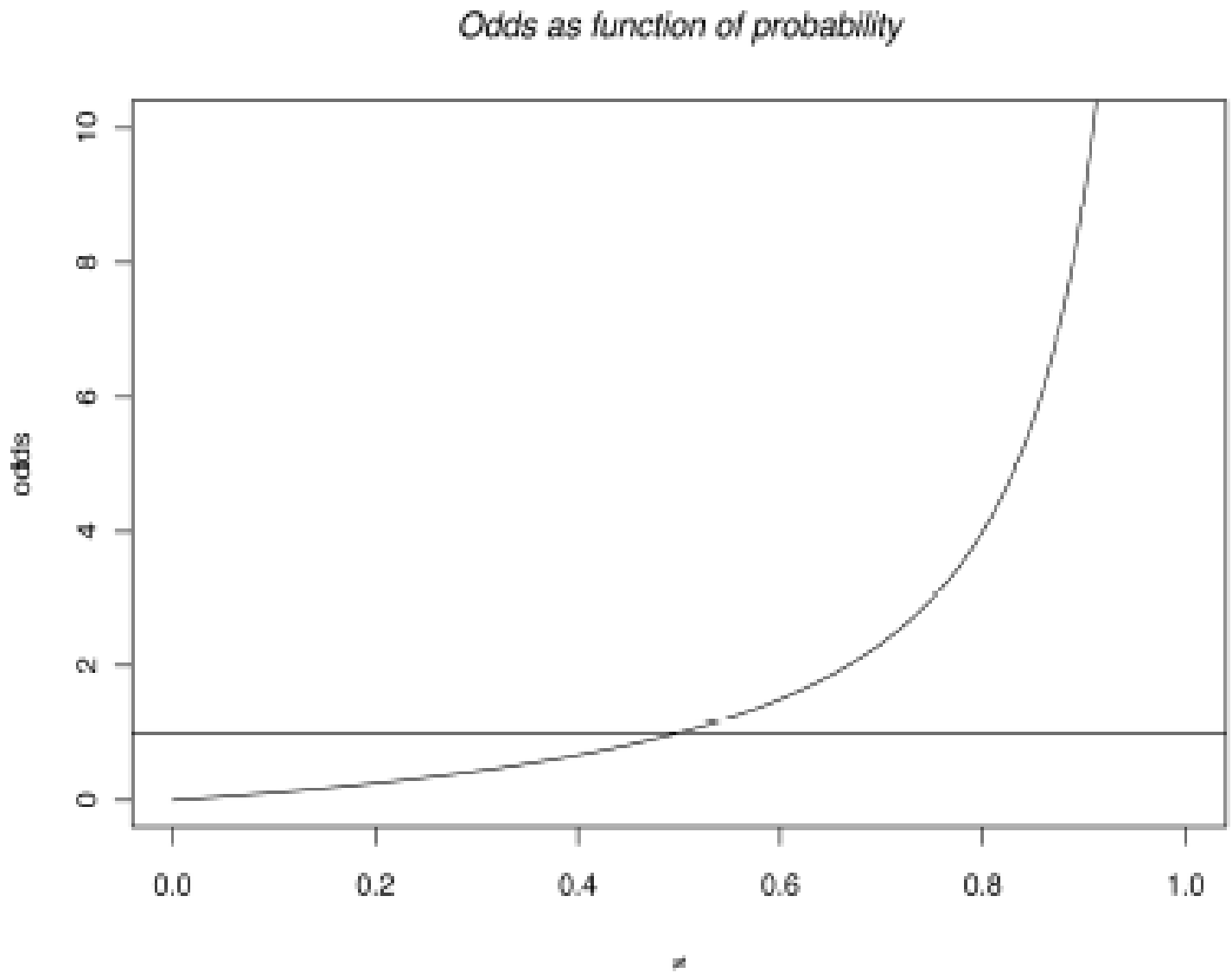
Odds and Odds Ratio (OR)

The **odds (Ω)** is the probability of binary outcome $P(Y = 1) = \pi$ divided by the the probability of binary outcome $P(Y = 0) = 1 - \pi$.

$$\Omega = \frac{\pi}{1 - \pi}$$

- Odds of 2.5 means that the probability of Y=1 (success) is two and half times higher than the probability of Y=0 (failure)
- Odds 0.5 means that success probability of success is 50the probability of failure
- Odds of 1 implies that probability of both outcomes 0.5 (equal)

Probability and odds



Risks and comparative parameters estimated

The risks π_1 and π_0 are estimated by empirical incidence proportions $R_1 = D_1/N_1$, and $R_0 = D_0/N_0$.

Crude estimates of comparative parameters

- **incidence proportion difference** $\quad \mathrm{RD} = R_1 - R_0$
- **incidence proportion ratio** $\quad \mathrm{RR} = R_1/R_0$
- **incidence odds ratio**

$$\mathrm{OR} = \frac{R_1/(1-R_1)}{R_0/(1-R_0)}$$

NB. To remove *confounding*, the estimated must be adjusted for relevant *confounders*.

Example: OS vs. PN (cont'd)

Crude estimates of true risk difference θ, risk ratio ϕ, and odds ratio ψ between OS and PN:

$$\begin{aligned}
\mathrm{RD} &= \frac{77}{350} - \frac{60}{350} = 0.22 - 0.171 = +\mathbf{0.049}\ (+4.9\%) \\
\mathrm{RR} &= \frac{77/350}{60/350} = \frac{77/60}{350/350} = \frac{0.22}{0.171} = \mathbf{1.283} \\
\mathrm{OR} &= \frac{77/273}{60/290} = \frac{0.22/(1-0.22)}{0.171/(1-0.171)} = \mathbf{1.363}
\end{aligned}$$

PN appears more successful than OS.

Is this (a) true, (b) due to bias, or (c) due to chance?

Tools for assessing precision in estimation

(a) Difference parameters:

- *standard error* of estimate SE,
- *error margin* at level c: $\mathrm{EM} = z_c \times \mathrm{SE}$,
- *confidence interval* at level c: $\mathrm{CI} = [\mathrm{est} - \mathrm{EM}, \mathrm{est} + \mathrm{EM}]$,

(b) Ratio parameters:

- standard error of log-estimate SEL,
- error factor at level c: $\mathrm{EF} = \exp(z_c \times \mathrm{SEL})$,
- confidence interval at level c: $\mathrm{CI} = [\mathrm{est}/\mathrm{EF}, \mathrm{est} \times \mathrm{EF}]$,

where z_c is an appropriate standard Gaussian quantile.

NB. These simple approximate CI formulas are based on the Wald statistics. More accurate but complicated procedures are recommended when computationally available.

Precision in risk difference estimation

Standard error of RD, error margin at level 95% ($c = 0.95$, $z_c = 1.96$), and 95% confidence interval for θ:

$$\begin{aligned} \mathrm{SE} &= \sqrt{\frac{R_1(1-R_1)}{N_1} + \frac{R_0(1-R_0)}{N_0}} \\ &= \sqrt{\frac{R_1^2(1-R_1)}{D_1} + \frac{R_0^2(1-R_0)}{D_0}} \\ \mathrm{EM} &= 1.96 \times \mathrm{SE} \\ \mathrm{CI} &= [\mathrm{RD} - \mathrm{EM}, \mathrm{RD} + \mathrm{EM}] \end{aligned}$$

NB. Standard error depends inversely on numbers of outcome cases D_1 and D_0 in the two exposure groups.

Example: OS vs. PN (cont'd)

$$\begin{aligned} \mathrm{SE} &= \sqrt{\frac{0.22 \times (1 - 0.22)}{350} + \frac{0.171 \times (1 - 0.171)}{350}} \\ &= \mathbf{0.030} \quad (3 \text{ \%-points}), \\ \mathrm{EM} &= 1.96 \times 0.030 = \mathbf{0.059}, \end{aligned}$$

95 % CI for true risk difference θ:

$$[0.049 - 0.059,\ 0.049 + 0.059] = [-\mathbf{0.010}, +\mathbf{0.108}]$$

i.e. from -1.0 to +10.8 percent points.

PN appears more successful than OS, although the evidence is quite weak.

Precision in risk ratio estimation

Standard error of log(RR), 95% error factor (EF) of RR, and 95% CI for true risk ratio ϕ:

$$\begin{aligned} \mathrm{SEL} &= \sqrt{\frac{1}{D_1} + \frac{1}{D_0} - \frac{1}{N_1} - \frac{1}{N_0}} \approx \sqrt{\frac{1}{D_1} + \frac{1}{D_0}} \\ \mathrm{EF} &= \exp\{1.96 \times \mathrm{SEL}\} \\ \mathrm{CI} &= [\mathrm{RR}/\mathrm{EF}, \mathrm{RR} \times \mathrm{EF}]. \end{aligned}$$

NB. Precision depends essentially on the numbers of exposed and unexposed cases, especially in a "rare disease" setting, where the numbers of cases are small in relation to group sizes.

Example: OS vs. PN (cont'd)

Standard error of log(RR), 95% error factor (EF) of RR, and 95% CI for true risk ratio ϕ:

$$\begin{aligned} \mathrm{SEL} &= \sqrt{\frac{1}{73} + \frac{1}{60} - \frac{1}{350} - \frac{1}{350}} \\ &= \mathbf{0.1547} \end{aligned}$$

$$\begin{aligned} \mathrm{EF} &= \exp\{1.96 \times 0.1547\} \\ &= \mathbf{1.3543} \end{aligned}$$

$$\begin{aligned} \mathrm{CI} &= [1.2833/1.3543, 1.2833 \times 1.3543] \\ &= [\mathbf{0.9476}, \mathbf{1.7380}]. \end{aligned}$$

Precision in odds ratio estimation

Standard error of log(OR), 95% error factor (EF) of OR, and 95% CI for true odds ratio ψ:

$$\begin{aligned} \mathrm{SEL} &= \sqrt{\frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0}} \\ \mathrm{EF} &= \exp\{1.96 \times \mathrm{SEL}\} \\ \mathrm{CI} &= [\mathrm{OR}/\mathrm{EF}, \mathrm{OR} \times \mathrm{EF}]. \end{aligned}$$

NB. Precision depends again on the numbers of exposed and unexposed cases, especially in a "rare disease" setting, where the numbers of cases are small in relation to group sizes.

Example: OS vs. PN (cont'd)

Standard error of log(OR), 95% error factor (EF) of OR, and 95% CI for true odds ratio ψ:

$$\begin{aligned} \mathrm{SEL} &= \sqrt{\frac{1}{77}+\frac{1}{60}+\frac{1}{273}+\frac{1}{290}} \\ &= \mathbf{0.1917} \end{aligned}$$

$$\begin{aligned} \mathrm{EF} &= \exp\{1.96 \times 0.1547\} \\ &= \mathbf{1.4562} \end{aligned}$$

$$\begin{aligned} \mathrm{CI} &= [1.3632/1.4562, 1.3632 \times 1.4562] \\ &= [\mathbf{0.9362}, \mathbf{1.9851}]. \end{aligned}$$

Estimating comparative parameters in R

- A multitude of R functions in several packages are readily available for point estimation and CI calculation using either "exact" or/and various approximative methods.
- We shall here demonstrate the use of function `twoby2()` in the `Epi package`. It applies the simple Wald approximations as described above, but for
  - risk difference: the Newcombe method is used, and
  - odds ratio: the "exact" conditional method is also available.

  Hence, similar results are expected as obtained above.

Use of function `twoby2()`

- Loading the `Epi package`:
  ```
  > library(Epi)
  ```
- Reading the counts of the 2 x 2-table into a matrix:
  ```
  > counts <- c(77, 273, 60, 290)
  > tab <- matrix( counts, nrow=2, byrow=T)
  ```
- Viewing the contents of the matrix/table:
  ```
  > tab
       [,1] [,2]
  [1,]   77  273
  [2,]   60  290
  ```
- Calling the function with `tab` as its argument:
  ```
  > twoby2(tab)
  ```

Output from twoby2()

```
Outcome   : Col 1
Comparing : Row 1 vs. Row 2

      Col 1 Col 2    P(Col 1) 95% conf. interval
Row 1    77   273      0.2200    0.1797   0.2664
Row 2    60   290      0.1714    0.1355   0.2146

                                    95% conf. interval
            Relative Risk: 1.2833    0.9476   1.7380
          Sample Odds Ratio: 1.3632    0.9362   1.9851
Conditional MLE Odds Ratio: 1.3626    0.9206   2.0237
    Probability difference: 0.0486   -0.0188   0.1155

              Exact P-value: 0.1272
          Asymptotic P-value: 0.1061
------------------------------------------------------
```

Analyses based on binary regression model

Crude estimates and CIs for the comparative parameters can also be obtained by fitting appropriate **binary regression models** for the numbers D_j or proportions R_j.

Special cases of **generalized linear models** (GLM) with

(i) **random part**: D_j is assumed to obey the binomial distribution or **family** with index N_j and probability π_j,

(ii) **systematic part**: **linear predictor** $\eta_j = \alpha + \beta X_j$, in which $X_j = 0$ for unexposed and $X_j = 1$ for exposed,

(iii) **link function**: $g(.)$ that connects the probability π_j and the systematic part η_j by:

$$g(\pi_j) = \eta_j = \alpha + \beta X_j$$

Link functions and comparative parameters

General model: $g(\pi_j) = \alpha + \beta X_j$ for the risks by binary X

- **identity** link: $g(\pi_j) = \pi_j = \alpha + \beta X_j$:
  $\Rightarrow \beta = \pi_1 - \pi_0 = \theta$,
  $=$ risk difference btw $X_j = 1$ and $X_j = 0$
- **logarithmic** link: $g(\pi_j) = log(\pi_j)$
  $\Leftrightarrow \pi_j = \exp(\alpha + \beta X_j) = e^{\alpha} e^{\beta X_j}$
  $\Rightarrow \beta = \log(\pi_1) - \log(\pi_0) = \log(\phi)$,
  $\Rightarrow e^{\beta} = \phi =$ risk ratio btw exposed and unexposed,
- **logit** link: $g(\pi_j) = \log[\pi_j/(1 - \pi_j)]$

$$\Leftrightarrow \quad \pi_j = \frac{1}{1 + \exp\{-(\alpha + \beta X_j)\}} = \text{expit}(\alpha + \beta X_j)$$

Logit model for odds ratio

Substituting logit function for $g(.)$ and values of X_j we get

$$\begin{aligned} \log\left(\frac{\pi_0}{1-\pi_0}\right) &= \alpha = \text{baseline logit} \\ \log\left(\frac{\pi_1}{1-\pi_1}\right) &= \alpha + \beta. \end{aligned}$$

This implies

$$\begin{aligned} \pi_0 &= \frac{1}{1+\exp(-\alpha)}, \quad \pi_1 = \frac{1}{1+\exp\{-(\alpha+\beta)\}}, \\ \beta &= \log\left\{\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right\} \\ e^{\beta} &= \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \psi = \underline{\text{odds ratio}} \text{ btw exp'd and unexp'd.} \end{aligned}$$

Fitting binary regression models in R

Function `glm()`

- estimation method: **maximum likelihood** (ML),
- computation algorithm: IWLS.

Key arguments of `glm()`:

- model formula: *"response"* $\sim$ *"expression of regressors"*
- `weights` = group sizes N_j when proportions R_j are given as the response (outcome) variable,
- `family = binomial(link = 'log')`, if risk ratio,
  `family = binomial(link = 'logit')`, if odds ratio,
  `family = binomial(link = 'identity')`, if risk diff'ce
  is the parameter of interest.

Example: Treatment of renal calculi (cont'd)

Grouped data set comprises

- two observations (one for each treatment group),
- three variables:

  `treat` = treatment, with values 1 = OS, 0 = PN,
  `fail` = number of failures D_j,
  `npat` = number of patients N_j,

- variable vectors defined:

```
> treat <- c(0, 1)
> fail <- c(60, 77)
> npat <- c(350, 350)
> prop <- fail/npat
```

Estimation of risk ratio

- Defining the *model object*:

```
> RRmodel <- glm( prop ~ treat,
+   family=binomial(link='log'), weights=npat)
```

- Estimation results extracted by function `ci.lin()` in `Epi` (two columns of the whole output omitted for clarity)

```
> round( ci.lin(RRmodel, Exp=T)[, -(3:4)], 4)
             Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)  -1.7636 0.1175    0.1714 0.1362 0.2158
treat         0.2495 0.1547    1.2833 0.9476 1.7380
```

- The estimate of β is $\widehat{\beta} = 0.2495 = \log(1.2833)$, and that of risk ratio ϕ is $\mathrm{RR} = \exp(0.2495) = 1.2833$.
- Estimate of α is $\widehat{\alpha} = -1.7636 = \log(0.1714) = \log(R_0)$.
- Previous results recovered also for SEL and CI.

Estimation of odds ratio

```
> ORmodel <- glm( prop ~ treat,
        fam = binomial(link='logit'), weights=npat)

> round( ci.lin(ORmodel, Exp=T)[, -(3:4)], 4)
             Estimate StdErr exp(Est.)   2.5%  97.5%
(Intercept)  -1.5755 0.1418    0.2069 0.1567 0.2732
treat         0.3099 0.1917    1.3632 0.9362 1.9851
```

- The estimate of β is $\widehat{\beta} = 0.3099 = \log(1.3632)$, and the estimated ψ is $\text{OR} = \exp(0.3099) = 1.3632$.
- The estimate of α is $\widehat{\alpha} - 1.5755 = \log(0.2069)$, in which $0.2069 = 0.1714/(1 - 0.1714)$ is the estimated baseline odds $R_0/(1 - R_0)$.

Estimation of risk difference

```
> RDmodel <- glm( prop ~ treat,
    fam=binomial(link='identity'), w=npat)

> round( ci.lin(RDmodel), 3)
            Estimate StdErr     z     P   2.5% 97.5%
(Intercept)    0.171   0.02 8.510 0.000  0.132 0.211
treat          0.049   0.03 1.623 0.105 -0.010 0.107
```

- Again, same results obtained as with *e.g.* `twoby2()`, although CIs are here based on Wald statistic.
- **NB**. Fitting binomial model with this link easily fails with more complicated models, especially involving continuous variables.

Confounding and effect modification

Consider another factor Z which is

- also a risk factor of the outcome,
- possibly associated with exposure X in study population,
- not a causal consequence of X.

$\Rightarrow$ Adjustment for possible *confounding* and evaluation of *effect modification* needed.

Example: OS vs. PN (cont'd)

Failure of treatment depends also on initial condition of patient, like extent and severity of disease.

Results stratified by initial diameter size of the stone:

Size	Trt	Fails	Npats	Fail-%	**RD(%)**
$<$ 2 cm	OS	6	87	*6.9*	
	PN	35	270	*13.0*	$-$**6.1**
$\geq$ 2 cm	OS	71	263	*27.0*	
	PN	25	80	*31.3*	$-$**4.3**

OS seems more successful in both subgroups, even though overall PN appeared better.

IS THERE A PARADOX?

Confounding

Solution: Treatment groups are not *comparable* w.r.t. initial size. Size of the stone is a **confounder** of the association between operation type and failure risk,because it is

1. an *independent determinant* of outcome (failure), based on external knowledge,
2. *statistically associated* with operation type in the study population,
3. *not causally affected* by operation type.

This is an instance of "confounding by indication":

- patient status affects choice of treatment

$\rightarrow$ *bias* in comparing treatments.

This bias would be best to avoid in planning:
$\rightarrow$ *randomized allocation* of treatments!

Stratified analysis

Stratification of cohort data with proportions

- at each level k of factor Z results are summarized:

Exposure to risk factor	Number of cases	Number of non-cases	Group size
yes	D_{1k}	C_{1k}	N_{1k}
no	D_{0k}	C_{0k}	N_{0k}
Total	D_{+k}	C_{+k}	N_{+k}

Stratum-specific incidence proportions by exposure group:

$$R_{1k} = \frac{D_{1k}}{N_{1k}}, \qquad R_{0k} = \frac{D_{0k}}{N_{0k}}$$

Adjusted estimation of risk difference

- Let π_{jk} be true risk in exposure group j ($j = 0, 1$) as to X and stratum k ($k = 0, \ldots, K$) of Z. Let also

$$\theta_k = \pi_{1k} - \pi_{0k}$$

  be the risk difference in stratum k.

- Many approaches to combine stratum specific results into one summary estimator that adjusts for confounding.

  These are all somehow *weighted averages* of stratum-specific estimators.

- Different weighting principles:
  - Maximum likelihood (ML),
  - Mantel-Haenszel (MH) weights,
  - Standardization either by external standard population or by "indirect" standardization.

Model-based adjustment of risk difference

Define generalized linear model for binary outcome with

- one binary exposure variable X and
- one binary stratifying factor or covariate Z (easily generalized to polytomous factors).

Random part: Number of cases D_{jk} in exposure group j ($j = 0, 1$) of X and level k ($k = 0, 1$) of Z is assumed to be binomially distributed

$$D_{jk} \sim \text{Binomial}(N_{jk}; \pi_{jk}),$$

Model-based adjustment of risk difference (cont'd)

Systematic part:

$$\pi_{jk} = \alpha + \beta X_j + \gamma Z_k,$$

where X_j is 0/1-indicator as before, and

$Z_k = 1$ for level $k = 1$ of Z, otherwise $Z_k = 0$,
$\alpha = \pi_{00} =$ baseline ("corner cell") risk,
$\gamma = \pi_{01} - \pi_{00} = \pi_{11} - \pi_{10}$,
$\beta = \pi_{10} - \pi_{00} = \theta_0 = \pi_{11} - \pi_{01} = \theta_1$,

How do we read this?

Model-based adjustment of risk difference (cont'd)

Implications of model definition

- the model assumes **homogeneity** of true risk difference θ associated with factor X (exposed vs. unexposed) across levels of Z: $\theta_1 = \theta_0 = \beta$,
- inclusion of Z in the model leads to adjustment of it when estimating the "true" effect θ of X,
- $\gamma =$ risk difference between levels 1 and 0 of Z; this is same in both exposure groups ($j = 0, 1$)
  $\Rightarrow$ homogeneity of the effect of Z is assumed, too.

Example. Treatment of renal calculi (cont'd)

- Define new variable

  `size` = initial stone size (0 for "small", 1 for "large")

- Extended data matrix comprises four observational units (rows) and four variables (columns):

size	trt	fails	npats
0	1	6	87
0	0	35	270
1	1	71	263
1	0	25	80

- These may be read in as before, *e.g.*

```
size <- c( 0, 0, 1, 1) ; trt <- c( 1, 0, 1, 0)
fails <- c( 6, 35, 71, 25)
npats <- c( 87, 270, 263, 80)
props <- fails/npats
```

Fitting model for adjusted risk difference

As before, but model formula supplemented by `+ size`

```
> RDmod2 <- glm( props ~ trt + size,
+       fam = binomial(link='identity'), w = npats)

> round( ci.lin(RDmod2), 3)
             Estimate StdErr      z     P   2.5% 97.5%
(Intercept)     0.128  0.019  6.596 0.000  0.090 0.166
trt            -0.056  0.030 -1.888 0.059 -0.114 0.002
size            0.195  0.032  6.106 0.000  0.132 0.258
```

Reading the results:

$\widehat{\alpha}$ $= \mathbf{0.128} = \widehat{\pi}_{00}$, fitted baseline risk,

$\widehat{\gamma}$ $= \mathbf{0.195}$, RD between large and small stones,

$\widehat{\beta}$ $= -\mathbf{0.0561}$ $[-\mathbf{0.114}, \mathbf{0.002}]$, estimated common treatment effect $\widehat{\theta}$ for OS *vs.* PN.

$=$ Weighted average of $\widehat{\theta}_0 = -0.061$ and $\widehat{\theta}_1 = -0.043$.

Effect modification

Homogeneity assumption – true differences were put equal:

$$\theta_k = \pi_{1k} - \pi_{0k} = \theta$$

across all levels k of covariate Z. *Is this realistic?*

Example. Is the true risk difference for treatment failure between OS and PN similar for small and big stones?

Empirical differences of failure proportions were

$$\begin{aligned} \text{small stones} : \widehat{\theta}_0 &= 0.069 - 0.130 = \mathbf{-0.061} \\ \text{large stones} : \widehat{\theta}_1 &= 0.270 - 0.313 = \mathbf{-0.043} \end{aligned}$$

Is the contrast $-0.043 - (-0.061) = 0.018$ between these differences due to chance only, or is there essential **effect modification** present?

Modelling modification of risk difference

The random part is the same, but the systematic part is

$$\pi_{jk} = \alpha + \beta X_j + \gamma Z_k + \tau U_{jk}.$$

$U_{jk} = X_j \times Z_k$, product of values of X and Z,
$\alpha = \pi_{00} =$ baseline ("corner cell") risk,
$\beta = \pi_{10} - \pi_{00} = \theta_0, \qquad \gamma = \pi_{01} - \pi_{00},$
$\tau = \theta_1 - \theta_0 = (\pi_{11} - \pi_{01}) - (\pi_{10} - \pi_{00}),$
interaction parameter

τ describes, how much greater is the risk difference between levels 1 and 0 of risk factor X among those at level 1 of factor Z than in those at level 0.

Fitting model with modification

- Generation of an interaction of **product term**:

```
> trtsize = size*treat
```

- Expanded and rearranged data matrix:

fails	npats	size	trt	trtsize
35	270	0	0	0
6	87	0	1	0
25	80	1	0	0
71	263	1	1	1

- Fitting the model including the product term:

```
> RDmod3 <- glm( props ~ trt + size + trtsize,
+      fam = binòmial(link='identity', w = npats)
```

Fitting model with modification (cont'd)

Results and interpretation:

```
> round( ci.lin(RDmod3)[ , -(3:4)], 4)
             Estimate StdErr    2.5%  97.5%
(Intercept)    0.1296 0.0204  0.0896 0.1697
trt           -0.0607 0.0340 -0.1273 0.0060
size           0.1829 0.0557  0.0737 0.2921
trtsize        0.0181 0.0678 -0.1147 0.1509
```

- $\widehat{\beta} = -0.061 = \widehat{\theta}_0 =$ RD for OS *vs.* PN in small stones,
- $\widehat{\gamma} = -0.183 =$ RD btw large and small stones for OS.
- estimate [95 % CI] of the interaction parameter:

$$\widehat{\tau} = \mathbf{0.0181}[-\mathbf{0.115}, \mathbf{0.151}]$$

$\Rightarrow$ no evidence for essential modification of risk difference.

Interpretation

This model is *saturated*: It has as many coefficients as there are observations. Hence

- residual df $= 0$,
- fitted cell probabilities $=$ observed proportions:

$$\begin{aligned}
\widehat{\pi}_{00} &= 0.130, \\
\widehat{\pi}_{10} &= 0.130 - 0.061 = 0.069, \\
\widehat{\pi}_{01} &= 0.130 + 0.183 = 0.313, \\
\widehat{\pi}_{11} &= 0.130 - 0.061 + 0.183 + 0.018 = 0.027
\end{aligned}$$

- residual X^2 and deviance are both 0.

Final comments

- When risk ratio ϕ or odds ratio ψ is the parameter of interest, adjustment for confounding and evaluation of modification can be done by fitting an analogous binomial GLM with relevant link function.
- Modelling can easily be extended to cover one or more polytomous and/or continuous covariates. Flexible functional forms may be specified to describe the effects of the latter type of variables.
- Binomial models are not limited to grouped data but may be fitted on individual data with binary outcomes, too.
- With more complicated models, especially involving continuous variables, the identity link (sometimes log link, too) violates the basic range restriction:
  outcome probabilities π must remain within 0 and 1.