

Lecture 5: Dealing with confounding

DAGs and stratification

Key concepts

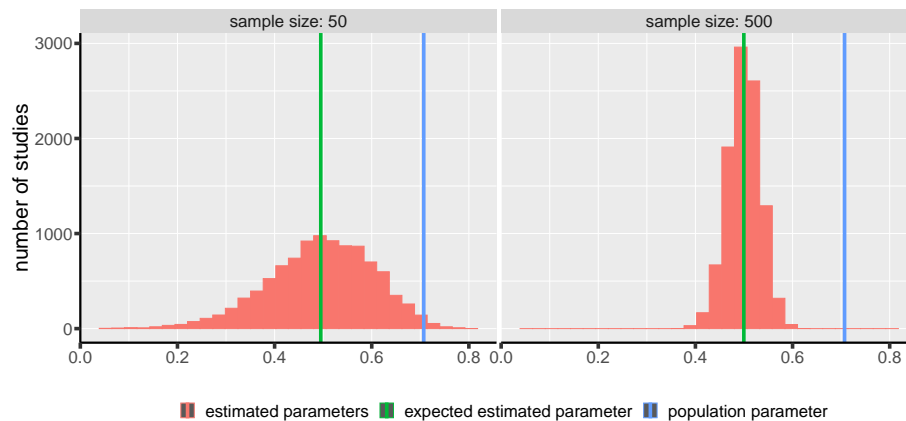
Brice Ozenne

1 Error decomposition

Every study involves a parameter of interest (β , also called population parameter) which we would like to estimate based on some data and a statistical method. For a given statistical method, the estimated parameter $\hat{\beta}$ will vary as a function of the collected data and we will denote $\mathbb{E}[\hat{\beta}]$ the expected estimated parameter.

Example:

- β is the true (but unknown) efficacy of a vaccine
- $\hat{\beta}$ is the estimated efficacy of the vaccine based on a single study
- $\mathbb{E}[\hat{\beta}]$ is (approximately) the average estimated efficacy of the vaccine over multiple studies.



The error made can be decomposed in two terms: $\hat{\beta} - \beta = \underbrace{\hat{\beta} - \mathbb{E}[\hat{\beta}]}_{\text{sampling error}} + \underbrace{\mathbb{E}[\hat{\beta}] - \beta}_{\text{bias}}$

Bias: systematic difference between the estimated quantity and the population parameter. It is stable across replication studies.

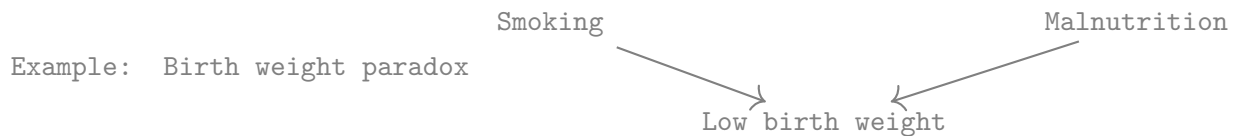
Sampling error: fluctuation in the estimated quantity due to the finite number of samples. It will differ from study to study.

The sampling error can be reduced by increasing the sample size or pooling the results from several replication studies. This is why we mostly worry about the bias:

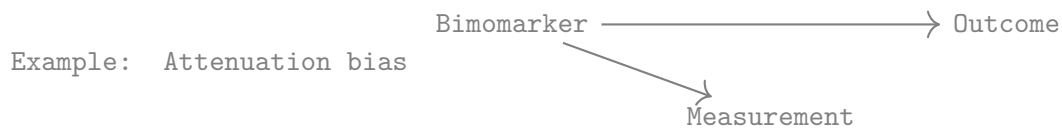
- **Confounding bias:** fictitious association between the exposure and the outcome resulting from a third variable.



- **Selection bias:** bias arising when the observed data is a non-random sample of the population. Typically the sampling probability is a function of the outcome, e.g. patients who fully recovered or got very ill leave the study.



- **Information bias:** bias due to imperfect measurement of the exposure or the outcome (measurement error).



- **Estimation bias:** bias due to the use of an uncorrect statistical method or approximations to facilitate the computations. In the latter case, "good" approximations lead to a consistent estimator, i.e. the bias vanishes as the sample size increases.

⚠ In presence of confounding, the "naive" estimates can still carry some information (e.g. for prediction) and be reported. However, they are not useful for assessing the causal effect the exposure.

Example: the admission rate to UC Berkeley for the fall 1973 were higher for men (41%) than for women (35%). These numbers are not helpful for deciding whether there is gender discrimination as, with the data at hand, the sex-admission association is confounded by the department in which the applicant applied.

2 Graphical representation of a study

2.1 Directed acyclic graphs (DAGs)

Graph: graphical representation composed of composed of vertices (variables) and edges (connection between variables).

Example: $X \longrightarrow Y \longrightarrow Z$ and $A \text{ --- } B \text{ --- } C$ are two graphs (denoted \mathcal{G}_1 and \mathcal{G}_2).

(Undirected) path: sequence of edges linking two vertices.

Example: \mathcal{G}_2 contains 6 paths, e.g. X to Z , Z to X , Y to Z

Oriented edge: edge that has a starting vertex and a ending vertex.

Example: $X \longrightarrow Y$ is an oriented edge but $A \text{ --- } B$ is not.

Directed path: sequence of oriented edges linking two vertices. The ending vertex of each edge must be the same as the starting vertex of the next edge.

Example: \mathcal{G}_1 contains 3 directed paths (X to Y , X to Z , and X to Z).

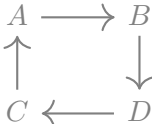
Ancestors of \bullet : vertices where at least one directed path ending at \bullet is starting.

Example: X has no ancestor but is an ancestor of Y and Z .

Descendants of \bullet : vertices where at least one directed path starting at \bullet is ending.

Example: Z has no descendant but is a descendant of X and Y .

Cycle: directed path where the starting vertex of the first edge is the ending vertex of the last edge.

Example: the graph \mathcal{G}_3  contains a cycle.

DAG: a DAG is a graph containing no cycle and where each edge is oriented.

Example: \mathcal{G}_1 is a DAG but not \mathcal{G}_2 or \mathcal{G}_3 .

Causal effect:¹ directed path from X to Y . If the directed path contain no other vertex, it correspond to a **direct** causal effect, otherwise to an **indirect** causal effect.

Example: In \mathcal{G}_1 , X has a (direct) causal effect on Y
and an (indirect) causal effect on Z .

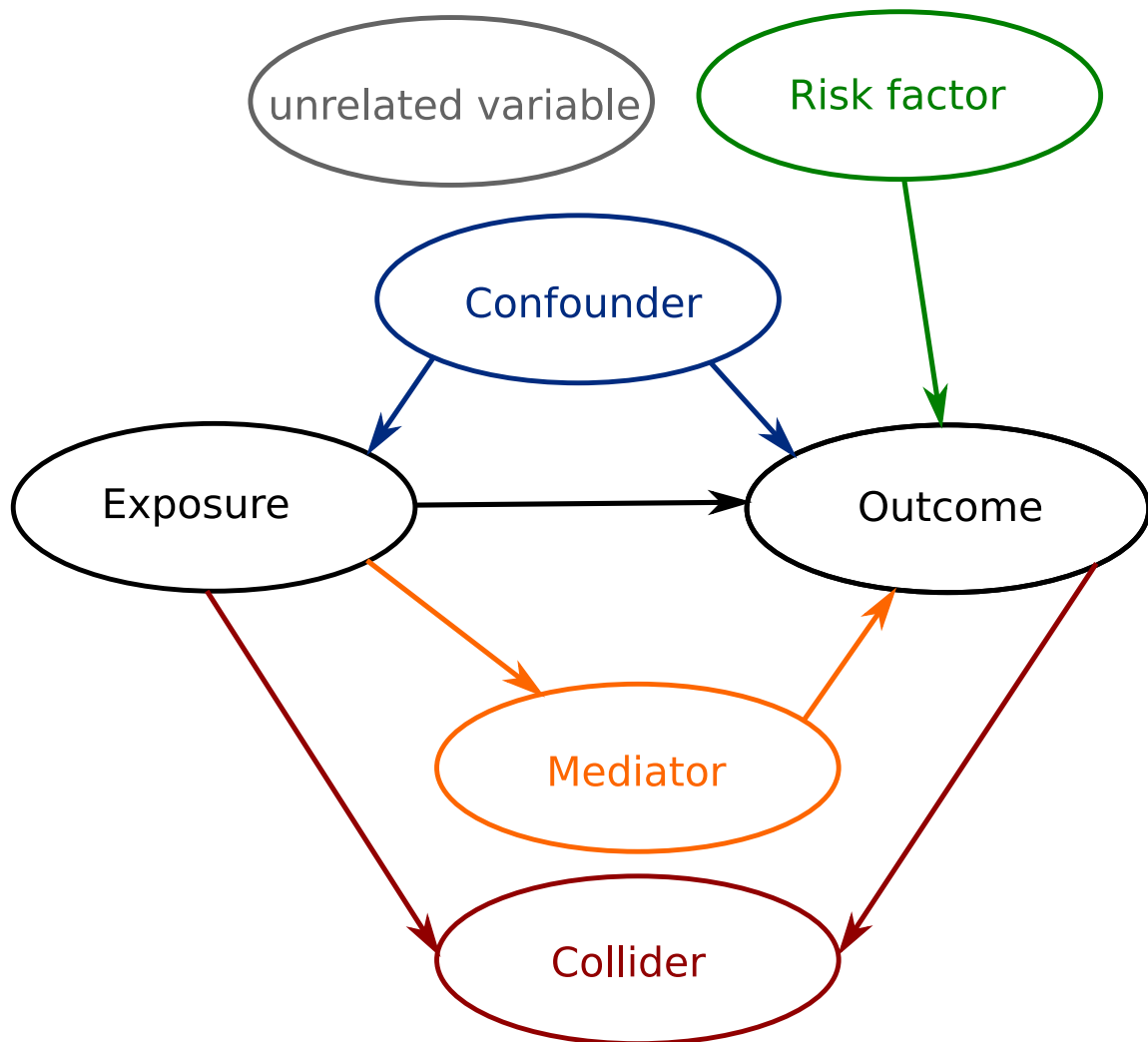
In \mathcal{G}_2 $B \leftarrow A \rightarrow C$, B has no causal effect on C .

¹ X has a causal effect on Y if the distribution of Y changes when only X changes (and its descendants but only due the change of X).

2.2 Nomenclature of the variables in a simple DAG

Given an exposure and outcome variable, we can now define:

- **Risk factor:** ancestor of only the outcome
- **Confounder:** ancestor of both the exposure and the outcome
- **Collider:** descendant of both the exposure² and the outcome.
- **Mediator:** variable on a directed path relating the exposure to the outcome.
- **Unrelated variable:** none of the previous
 - not related to the outcome and nor to the exposure
 - only an ancestor or a descendant of the exposure
 - only a descendant of the outcome



²there must be a least one directed path relating the collider to the exposure that does not contain the outcome

3 What to control for?

We would like to control for a set of variable Z in order to:

- leave all directed paths between E and Y unperturbed
- block all spurious paths between E and Y
- create no new spurious paths between E and Y

3.1 Simple DAG

- **Risk factor:** yes - less uncertainty on the estimate.
- **Confounder:** yes - otherwise biased estimate Simpson paradox
- **Collider:** no - otherwise biased estimate Selection bias/Bergson paradox
- **Mediator:** depends of the parameter of interest: Mediation analysis
 - adjustment: direct causal effect
 - no adjustment: total causal effect
- **Unrelated variable:** no - (slightly) more uncertainty on the estimate

⚠ In non linear models, adding a variable will likely affect the interpretation of a parameter in the model (conditional effect). Standardisation can be used to recover the parameter of interest (marginal effect).

3.2 Complex DAG

There is no confounding between X and Y when controlling for Z if:

- no variable in Z is a descendent of X
- every path between X and Y is blocked - except the direct paths from X to Y

d-separation: the path between X and Y is blocked by Z if it either

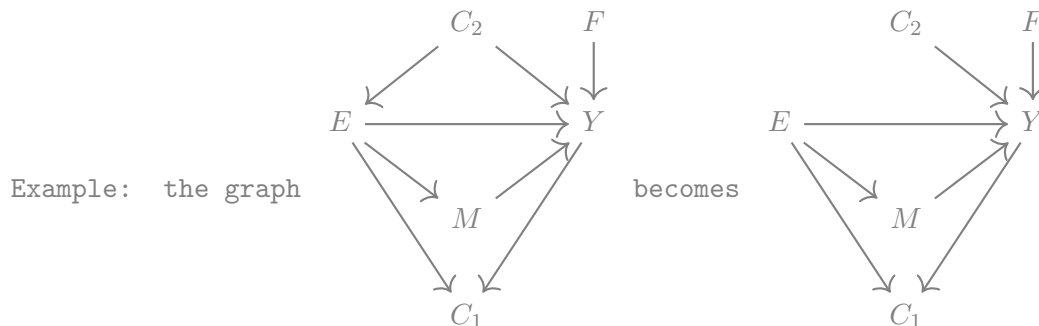
- contain a collider that is not in Z (or its descendents)
- contain a non-collider that is in Z .

Note: using the d-seperation criteria requires that the DAG contains all parents (first order ancestors) of any pair of variables.

4 How to control for a set of variable?

4.1 Using a specific study design

- **randomization:** the exposure is randomly allocated among participants. This "removes" all arrows directed to the exposure variable.
 - ✓ control for known and unknown confounders (in large enough samples).
 - ✓ can (usually) be analyzed with simple statistical methods.
 - ✗ can be difficult to implement for ethical, logistical, or economical reason.
 - ✗ not well suited to study rare events (would require a very large sample).



when the exposure is randomized. An unadjusted analysis will lead to an unbiased estimate. Care should be taken when adjusting for covariates:

- adjusting for C_2 and F will lower the sampling error (more efficient estimator).
- adjusting for M or C_1 may bias the analysis.

- **restriction:** only participants with a specific value of the (known) confounders are included in the study (e.g. 50-year old females). Within the sample the (known) confounders have no effect on the exposure nor on the outcome.

- ⚠ only controls for known confounders. Residual confounding due to unknown confounders is possible.
- ⚠ almost impossible to implement exactly as it would lead to too restrictive inclusion criteria. Instead, a range of values for the confounder is used (e.g. age between 40 and 50). A narrow range means a better control of the confounding but also more stringent inclusion criteria.
- ✓ can (usually) be analyzed with simple statistical methods.

Note: another commonly used approach is matching and will be presented later in the course.

4.2 Adjusting for confounding in the statistical analysis

We consider the case of a binary outcome Y , a binary exposure E , and a categorical confounder C with K categories. We are interested in a parameter θ which can be the prevalence, or the incidence rate, or the τ -years risk. We would like to test the null hypothesis:

- \mathcal{H}_0 : the parameter is the same for exposed and non-exposed in each strata

versus the alternative hypothesis:


- \mathcal{H}_1 : the parameter is different for exposed and non-exposed in at least one strata

The following table present three statistical models:

strata	no effect		"full" stratification		multiplicative "common effect" stratification	
	non-exposed	exposed	non-exposed	exposed	non-exposed	exposed
1	θ_1	θ_1	$\theta_{0,1}$	$\theta_{1,1}$	θ_1	$\beta\theta_1$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
K	θ_K	θ_K	$\theta_{0,K}$	$\theta_{1,K}$	θ_K	$\beta\theta_K$

We can assess the association between exposure and outcome by:

"Full" stratification: we estimate the (denoted $\theta_{e,k}$) for each exposure e and confounder value c . When the group are independent, we can use the estimators introduced in lecture 3. We can then:

- *strata-specific tests*: compare the estimates between exposure groups in each strata: $\hat{\theta}_{0,1}$ vs. $\hat{\theta}_{1,1}$, \dots , $\theta_{0,K}$ vs. $\theta_{1,K}$, using for instance risk difference and associated confidence intervals (lecture 3). Evidence for any difference is evidence against \mathcal{H}_0
 remember to adjust for multiple comparisons when computing p-values or confidence intervals
- *likelihood ratio test*: obtain a more powerful test by comparing the likelihood of the no effect model and the "full" stratification model. A large difference in likelihood is evidence for an association exposure-outcome³.

"Common effect" stratification: we assume a similar exposure effect in each strata (denoted β) and a different parameter value in each strata (denoted $\theta_1, \dots, \theta_K$). We can then assess the exposure effect by testing whether there is evidence to reject $\beta = 1$.

³in at least one strata but the test will not indicate which one(s).

Example: for each strata k , we can create the following 2×3 table:

Group \ Outcome	Outcome		
	Survival	Failure	Time at risk
non-exposed	a_k	b_k	$\tilde{T}_{\bar{e},k}$
exposed	c_k	d_k	$\tilde{T}_{e,k}$

An alternative to the multiplicative model is to use the Cochran-Mantel-Haenszel (CMH) test. It models a common odd-ratio over several strata:

$$\widehat{OR}^{MH} = \frac{\sum_{k=1}^K \frac{a_k d_k}{n_k}}{\sum_{k=1}^K \frac{b_k c_k}{n_k}}$$

where $n_k = a_k + b_k + c_k + d_k$ is the number of observations in the k -th strata. Testing whether $OR=1$ tests the presence of an association between outcome and exposure. Under the null hypothesis, $\widehat{OR}^{MH} \sim \chi_K^2$.

In chapter 15, [Rothman et al. \(2008\)](#) provides formula for the CMH applied to the risk difference or risk ratio

$$\widehat{RD}^{MH} = \frac{\sum_{k=1}^K \frac{d_k(a_k+b_k) - b_k(c_k+d_k)}{n_k}}{\sum_{k=1}^K \frac{(a_k+b_k)(c_k+d_k)}{n_k}} \quad \widehat{RR}^{MH} = \frac{\sum_{k=1}^K \frac{d_k(a_k+b_k)}{n_k}}{\sum_{k=1}^K \frac{b_k(c_k+d_k)}{n_k}}$$

incidence rate difference or incidence rate ratio:

$$\widehat{IRD}^{MH} = \frac{\sum_{k=1}^K \frac{d_k \tilde{T}_{\bar{e},k} - b_k \tilde{T}_{e,k}}{\tilde{T}_{\bar{e},k} + \tilde{T}_{e,k}}}{\sum_{k=1}^K \frac{\tilde{T}_{e,k} \tilde{T}_{\bar{e},k}}{\tilde{T}_{\bar{e},k} + \tilde{T}_{e,k}}} \quad \widehat{IRR}^{MH} = \frac{\sum_{k=1}^K \frac{d_k \tilde{T}_{\bar{e},k}}{\tilde{T}_{\bar{e},k} + \tilde{T}_{e,k}}}{\sum_{k=1}^K \frac{b_k \tilde{T}_{e,k}}{\tilde{T}_{\bar{e},k} + \tilde{T}_{e,k}}}$$

Corresponding formula for the variance of the estimator can be found in [Rothman et al. \(2008\)](#).

The CMH estimate is more variable than the corresponding maximum likelihood estimate. However it can be computed even with sparse data (i.e. a_k , b_k , c_k , d_k , $\tilde{T}_{\bar{e},k}$, or $\tilde{T}_{e,k}$ is 0).

Note 1: holding the confounder at a fixed value within strata "removes" the arrows from the confounder to the exposure and outcome. Finer stratification leads to better control for confounding but also sparser our data (within strata).

Note 2: other statistical approaches to handle confounding (inverse probability weighting, G-formula) will be presented later in the course.

5 In R

For illustration, we will use the following dataset:

```
data(Whickham, package = "mosaicData")
Whickham$ageC <- cut(Whickham$age, c(0,35,50,65,100))
summary(Whickham)
```

```
outcome  smoker      age      ageC
Alive:945   No :732   Min.    :18.00   (0,35]   :430
Dead :369   Yes:582   1st Qu.:32.00   (35,50]  :321
              Median :46.00   (50,65]  :334
              Mean   :46.92   (65,100] :229
              3rd Qu.:61.00
              Max.   :84.00
```

5.1 Tables

- `table` to obtain perform a 2 by 2 table between the outcome (Y) and the exposure (E) for each level of the confounder (C):

```
t23 <- table(Y = Whickham$outcome,
             E = Whickham$smoker,
             C = Whickham$ageC)
t23
```

, , C = (0,35]

	E	
Y	No	Yes
Alive	236	182
Dead	7	5

, , C = (35,50]

	E	
Y	No	Yes
Alive	126	155
Dead	13	27

, , C = (50,65]

	E	
Y	No	Yes
Alive	115	100
Dead	55	64

, , C = (65,100]

	E	
Y	No	Yes
Alive	25	6
Dead	155	43

- `ftable` for a concise display

```
ftable(t23)
```

		C (0,35]	(35,50]	(50,65]	(65,100]
Y	E				
Alive	No	236	126	115	25
	Yes	182	155	100	6
Dead	No	7	13	55	155
	Yes	5	27	64	43

- `stats::aperm` to change the order of the entries in a table:

```
t23.bis <- aperm(t23,c(2,1,3))
ftable(t23.bis)
## same as
## ftable(E = Whickham$smoker, Y = Whickham$outcome, C = Whickham$ageC)
```

		C (0,35]	(35,50]	(50,65]	(65,100]
E	Y				
No	Alive	236	126	115	25
	Dead	7	13	55	155
Yes	Alive	182	155	100	6
	Dead	5	27	64	43

5.2 Estimating odd ratios from aggregated data

- by hand, one at a time

```
a <- t23.bis["No","Alive","(0,35)"]
b <- t23.bis["No","Dead","(0,35)"]
c <- t23.bis["Yes","Alive","(0,35)"]
d <- t23.bis["Yes","Dead","(0,35)"]
(a*d)/(b*c)
```

```
[1] 0.9262166
```

- by hand, all at once

```
vec.a <- t23.bis["No","Alive",]
vec.b <- t23.bis["No","Dead",]
vec.c <- t23.bis["Yes","Alive",]
vec.d <- t23.bis["Yes","Dead",]
(vec.a*vec.d)/(vec.b*vec.c)
```

```
(0,35]   (35,50]   (50,65]   (65,100]
0.9262166 1.6883375 1.3381818 1.1559140
```

- by hand, using the `apply` function to perform arbitrary computation within each strata. The argument `MARGIN` indicate on which dimension of the table we want to repeatedly apply the calculations defined by the argument `FUN` (`x` is an abstract representation of a given 2 by 2 table):

```
apply(t23.bis, MARGIN = 3,
      FUN = function(x){
        r1 <- x["No","Dead"]/(x["No","Alive"]+x["No","Dead"])
        r2 <- x["Yes","Dead"]/(x["Yes","Alive"]+x["Yes","Dead"])

        out <- c("risk(No)" = r1,
                  "risk(Yes)" = r2,
                  "risk difference" = r2-r1,
                  "risk ratio" = r2/r1,
                  "odds ratio(risk)" = (r2/(1-r2))/(r1/(1-r1))
                  )

        return(out)
      })
```

```

C
      (0,35]   (35,50]   (50,65]   (65,100]
risk(No)      0.028806584 0.09352518 0.32352941 0.86111111
risk(Yes)      0.026737968 0.14835165 0.39024390 0.87755102
risk difference -0.002068616 0.05482647 0.06671449 0.01643991
risk ratio      0.928189458 1.58622147 1.20620843 1.01909151
odds ratio(risk) 0.926216641 1.68833747 1.33818182 1.15591398
```

- using a dedicated package like `vcd::loddsratio` (the argument `correct` would add 0.5 to each cell to handle empty cells):

```
library(vcd)
loddsratio(t23, log = FALSE, correct = FALSE)
```

odds ratios for Y and E by C

```
(0,35] (35,50] (50,65] (65,100]
0.9262166 1.6883375 1.3381818 1.1559140
```

- `stats::mantelhaen.test` to perform a Cochran-Mantel-Haenszel test and estimate a common odds ratio ⁴

```
mantelhaen.test(t23)
```

Mantel-Haenszel chi-squared test with continuity correction

```
data: t23
Mantel-Haenszel X-squared = 2.7287, df = 1, p-value = 0.09856
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.9623196 1.8830784
sample estimates:
common odds ratio
 1.346151
```

- `DescTools::BreslowDayTest` or `vcd::woolf_test` to check the assumption of homogeneity of the odds ratios across strata:
(the usefulness of this test is questionable as it does not indicate how important the heterogeneity is)

```
library(DescTools)
BreslowDayTest(t23)
```

Breslow-Day test on Homogeneity of Odds Ratios

```
data: t23
X-squared = 0.9007, df = 3, p-value = 0.8253
```

⁴The `metafor::rma.mh` implements the Cochran-Mantel-Haenszel test for risk difference, risk ratio, incidence rate ratio, incidence rate difference (⚠ I haven't tested it)

5.3 Estimating odd ratios from individual data

- using `glm` to fit a logistic regression (with the argument `family=binomial(link="logit")`)

```
## full stratification
e.strata <- glm(outcome ~ ageC + ageC:smoker,
               data = Whickham, family = binomial(link="logit"))
## names of the coefficients containing the character smoker
name.coefsmoker <- grep("smoker",names(coef(e.strata)), value = TRUE)
exp(coef(e.strata)[name.coefsmoker])
```

```
ageC(0,35]:smokerYes  ageC(35,50]:smokerYes  ageC(50,65]:smokerYes
               0.9262166                1.6883375                1.3381818
ageC(65,100]:smokerYes
               1.1559140
```

```
## common effect stratification
e.common <- glm(outcome ~ ageC + smoker,
               data = Whickham, family = binomial(link="logit"))
exp(coef(e.common)["smokerYes"])
```

```
smokerYes
1.346148
```

- `anova` to perform a likelihood ratio test between two logistic regressions (with the argument `test = "LRT"`)

```
e.H0 <- glm(outcome ~ ageC,
            data = Whickham, family = binomial(link="logit"))

LRT.strata <- anova(e.strata, e.H0, test = "LRT")
LRT.common <- anova(e.common, e.H0, test = "LRT")
c("full stratification" = LRT.strata["Pr(>Chi)"][2,1],
  "common effect stratification" = LRT.common["Pr(>Chi)"][2,1])
```

```
full stratification  common effect stratification
               0.41220871                0.08083229
```

- display fitted values, i.e. fitted probability for each age and smoking status:

```
grid <- unique(Whickham[,c("ageC", "smoker")])
grid$fit.H0 <- 100*predict(e.H0, newdata=grid, type="response")
grid$fit.common <- 100*predict(e.common, newdata=grid, type="response")
grid$fit.strata <- 100*predict(e.strata, newdata=grid, type="response")
print(grid, digits = 2)
```

	ageC	smoker	fit.H0	fit.common	fit.strata
1	(0,35]	Yes	2.8	3.3	2.7
3	(65,100]	Yes	86.5	89.0	87.8
4	(65,100]	No	86.5	85.8	86.1
5	(50,65]	No	35.6	32.3	32.4
6	(35,50]	Yes	12.5	13.8	14.8
9	(0,35]	No	2.8	2.4	2.9
26	(35,50]	No	12.5	10.7	9.4
35	(50,65]	Yes	35.6	39.1	39.0

5.4 Visualizing heterogeneity

First extract the log odd ratios:

```
df.strata <- data.frame(name = gsub(":"smokerYes", "", name.coefsmoker),
                        estimate = coef(e.strata)[name.coefsmoker],
                        lower = confint(e.strata)[name.coefsmoker,1],
                        upper = confint(e.strata)[name.coefsmoker,2])
df.strata
```

Waiting for profiling to be done...

Waiting for profiling to be done...

	name	estimate	lower	upper
ageC(0,35]:smokerYes	ageC(0,35]	-0.07664712	-1.3082302	1.0812790
ageC(35,50]:smokerYes	ageC(35,50]	0.52374430	-0.1610087	1.2543342
ageC(50,65]:smokerYes	ageC(50,65]	0.29131184	-0.1570201	0.7423584
ageC(65,100]:smokerYes	ageC(65,100]	0.14489135	-0.7522483	1.1850019

and the common odd ratio:

```
df.common <- data.frame(name = "common",
                        estimate = coef(e.common)["smokerYes"],
                        lower = confint(e.common)["smokerYes",1],
                        upper = confint(e.common)["smokerYes",2])
df.common
```

Waiting for profiling to be done...

Waiting for profiling to be done...

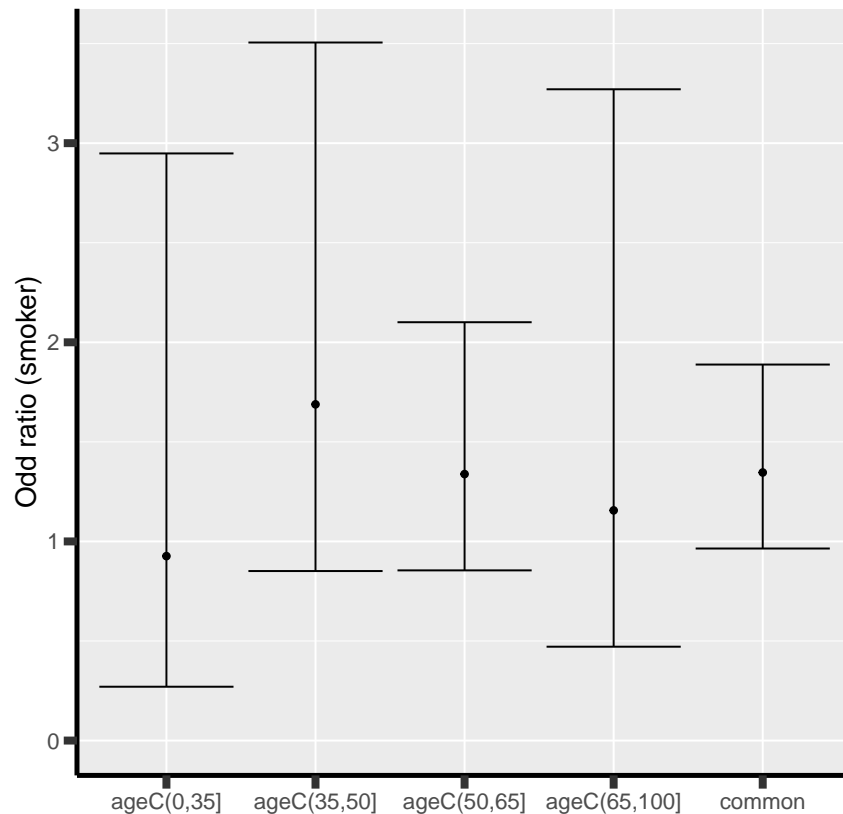
	name	estimate	lower	upper
smokerYes	common	0.2972473	-0.03625327	0.6356513

Combine both datasets and take the exponential to obtain odd ratios:

```
df.all <- rbind(df.strata, df.common)
df.all$estimate <- exp(df.all$estimate)
df.all$lower <- exp(df.all$lower)
df.all$upper <- exp(df.all$upper)
```

Display the results, e.g. using ggplot:

```
library(ggplot2)
gg <- ggplot(df.all, aes(x = name, y = estimate, ymin = lower, ymax =
  upper))
gg <- gg + geom_point() + geom_errorbar() + ylab("Odd ratio (smoker)") +
  xlab("")
gg <- gg + coord_cartesian(ylim = c(0,3.5))
gg
```



6 Reference

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern epidemiology*. Lippincott Williams & Wilkins.