

Practicals - Dealing with confounding

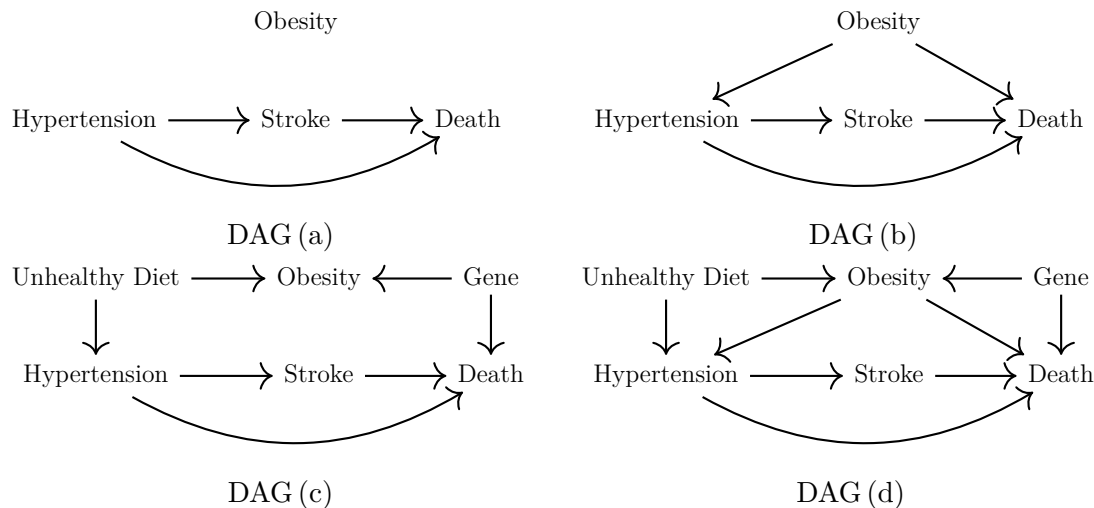
(I) DAGs and stratification

Epidemiological methods in medical research 2022

2 February 2023

Exercise 1: warming-up with DAGs

We would like to plan a study investigating the causal effect of hypertension on the risk of death. We plan to recruit patients just after their first diagnosis of hypertension (baseline), measure their weight at baseline, whether they experienced a stroke after baseline, and when they died. To simplify, we assume that the weight is constant over time, the diagnostic of hypertension is accurate, and that we could find an appropriate control group. As we are writing the "statistical analysis" section of the protocol, we wonder on which variable we should adjust on.




1. Assume that the "true" DAG is DAG (a) in the figure above. Should we adjust on stroke or on obesity? Why?
2. Assume that the "true" DAG is DAG (b) in the figure above. What is the difference with the previous DAG? Should we adjust on stroke or on obesity? Why?

We run a logistic regression to assess the effect of hypertension on death, adjusting for obesity. A collaborator points out that it could also be interesting to report about the effect of obesity on the risk of death and suggest to use the estimated obesity effect in that model. Is it a good idea?

3. Look now at the DAGs (c) and (d).


- can you guess whether it is possible to estimate a causal effect? If yes which variable should we adjust on?
- check your intuition using DAGitty (<http://www.dagitty.net/dags.html>). You will find on the course webpage a short tutorial `dagitty-tutorial.pdf`.

Exercise 2: analysis of the UC Berkeley data

The aim of this exercise is to reproduce the analysis of the admission data from UC Berkeley 1973 shown during the lecture. The dataset is available on the course webpage in .txt: `UCBAdmissions.txt`.  users can also install the `datasets` package and load the dataset using:

```
data(UCBAdmissions, package = "datasets")
ftable(UCBAdmissions)
```

		Dept	A	B	C	D	E	F
Admit	Gender							
Admitted	Male		512	353	120	138	53	22
	Female		89	17	202	131	94	24
Rejected	Male		313	207	205	279	138	351
	Female		19	8	391	244	299	317

The dataset contains the number of admitted and rejected applicants per department and by gender. A "long" format of this dataset (i.e. each line contains the number of applications for a given department, gender, and admission status) is also available on the course webpage in .txt: `UCBAdmissions_long.txt`. In , it can be obtained doing:

```
dfAll <- as.data.frame(UCBAdmissions)
dfAll$N <- dfAll$Freq
dfAll$D <- (dfAll$Admit=="Admitted")*(dfAll$N)
df <- aggregate(cbind(N,D) ~ Gender + Dept, data = dfAll, FUN = "sum")
head(df)
```




	Gender	Dept	N	D
1	Male	A	825	512
2	Female	A	108	89
3	Male	B	560	353
4	Female	B	25	17
5	Male	C	325	120
6	Female	C	593	202

Some of the questions will refer to the slides shown during lecture 5. They can be found on the course webpage, module "Day 3: Bias and confounding", file "L5-confounding.pdf".

1. Compare the percentage of admission between males and females over all departments (i.e. reproduce the first table of slide 40).

We now compare the percentage of admission between males and females, stratifying on department. We will use three different approaches:

2. "Common effect"

- a) perform a Cochran-Mantel-Haenszel test (slide 52, function `mantelhaen.test` in  and `proc freq` with the option `cmh` in SAS). Can you compute the common odd-ratio yourself?
- b) what assumption(s) are making with this approach? Are they fulfilled? You can have a look to the Breslow-Day test (`DescTools::BreslowDayTest` in ).
- c) compare the results with a logistic model with an additive effect of gender and department on the log odd scale. In  you can use the function `glm(..., family = binomial(link="logit"))` and in SAS the `proc logistic`.
- d) what would you conclude? The formula slide 30 of lecture 2 to deduce the risk ratio from the odd ratio may be useful.




3. Full stratification using a likelihood ratio test (slide 45):

- a) fit a logistic model under the null hypothesis of no gender effect.
- b) fit a logistic model under the alternative hypothesis of a gender effect specific to each strata.
- c) perform a log-likelihood ratio test (LRT).
- d) what would you conclude?
- [Extra time] Have a look to the slides 51-52 and try to compute the log-likelihood and the p-value of the LRT yourself.

```
qchisq(0.95, df = 6) ## quantile of the chi-squared
```

```
[1] 12.5915941.
```

4. Full stratification using strata-specific tests:

- a) compare the percentages of admission between males and females in each department (table slide 41).
- b) to get a p-value you can use the `Epi::twoby2`, `Publish::table2x2`, or `exact2x2::uncondExact2x2` in  (I used the latter one to get the uncorrected p-value).
- c) adjust for multiple comparisons, e.g. using `p.adjust(..., method = "bonferroni")` in .
- d) What would you conclude?
- [Extra time] Look at the predicted probabilities of the logistic models from 3a) and 3b) and compare them to the ones calculated in question 1 and 4a). In  you can use the function `predict(..., type = "response")`.