Help! Statistics!
Mediation Analysis

Sacha la Bastide-van Gemert
Medical Statistics and Decision Making
Epidemiology, UMCG

Help! Statistics! Lunch time lectures

What? Frequently used statistical methods and questions in a manageable timeframe for all researchers at the UMCG.
No knowledge of advanced statistics is required.

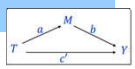
When? Lectures take place every 2nd Tuesday of the month, 12.00-13.00 hrs.

Who? Unit for Medical Statistics and Decision Making

When?	Where?	What?	Who?
Mar 14, 2017	Room 16	Mediation analysis	S. la Bastide
Apr 11, 2017	Rode Zaal	Basics of survival analysis	D. Postmus
May 9, 2017	Rode Zaal	Multiple linear regression; some do's and don'ts	H. Burgerhof
June 13, 2017	Room 16	Multiple testing	C. zu Eulenburg

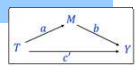
Slides can be downloaded from:
http://www.rug.nl/research/epidemiology/download-area

Mediation analysis: overview



- Introduction: example (*smoking during pregnancy, birth-weight*)
- Traditional approaches and their limitations (B&K, Sobel's test)
- Better alternative: bootstrapping test
- Underlying assumption: uncorrelated errorterms
  - *Intermezzo: causal graphs*
- Causal mediation analysis
  - *Intermezzo: the counterfactual framework*
- R package ``mediation'': application and sensitivity analyses
- Concluding remarks & literature references

Mediation: example



M = DNA methylation

a b

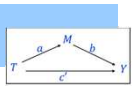
T= smoking of the mother during pregnancy

c'

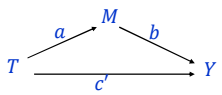
Y= birth-weight of the child

X= {possible confounding covariates like age of the mother, SES, gestational age, …}

Mediation analysis
Introduction and terminology



- Mediation analysis: exploring the underlying mechanism of a relationship by which one variable (exposure/treatment T) influences another variable (outcome Y) through a third variable (mediator M):

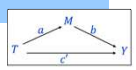


… hence a causal notion!

- Mediator M is a variable on the causal pathway from T to Y
- Total effect, direct effect and indirect effect
- Full and partial mediation

Linear systems (under assumptions):
total effect *c* of T on Y consists of direct effect *c'* and indirect effect *a*b*

Baron and Kenny's ``Causal Steps approach''



B&K (1986) popularized the "causal steps approach" to distinguish mediation:

Step 1: H_0: $c = 0$ has to be rejected, i.e. *c* (``total effect of T on Y'') must be significant

$$Y = \beta_{10} + cT + \varepsilon_1$$

Step 2: H_0: $a = 0$ has to be rejected, i.e. *a* must be significant

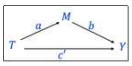
$$M = \beta_{20} + aT + \varepsilon_2$$

Step 3: H_0: $b = 0$ has to be rejected, i.e. *b* is significant AND *c'* should be smaller in absolute value than the total effect *c*

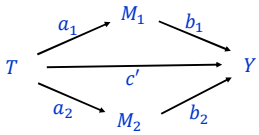
$$Y = \beta_{30} + c'T + bM + \varepsilon_3$$

Then: the non-zeroness of an intermediate *a*b* effect is logically claimed to be existing

B&K's approach has been criticized a lot (but is still used!):

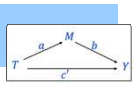


- step 1 ("significant total effect") is not necessary: the pathways could cancel each other out, and *c'* becomes noticeable only when the mediator is controlled for
- can easily `miss' mediating effects when not all paths are included in the formal model:



- based not on the quantification of the intervening effect but on separate tests of the paths T→M and M→Y
- lowest in power among methods for testing intervening variable effects

Sobel's test



Sobel's test: tests whether the mediation effect is significantly different from 0

$$z = \frac{a*b}{\sqrt{b^2 s_a{}^2 + a^2 s_b{}^2}}$$ with s_a and s_b se's of the effects

, assuming a normal distribution of z

Better than B&K: Sobel's test is more accurate than B&K steps, actually tests the thing we are looking for!

But: very low statistical power due to normality assumption (too strong!) and inadequate estimation of the se of $a * b$.

Hayes (and many others): use bootstrap test of the indirect effect instead!

Intermezzo: the principle of bootstrapping

Boots… sample: sample drawn from the original sample with replacement, …sing the s… same sample size (n)

k boo… samp…

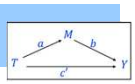
k boo… replic…

We'll save this topic for another
Help! Statistics!- lecture ...

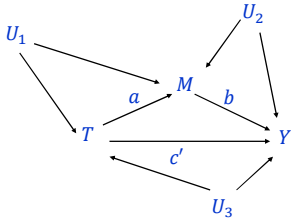
For now:
note that it is a distribution-free estimation method, so no additional assumptions need to be made on the mediation effect

Result: bootstrap confidence interval (percentile based)

Underlying assumptions

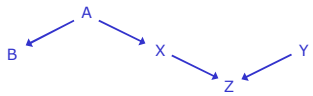


Assumption for mediation analysis: uncorrelated error terms ε_1, ε_2, ε_3 for T, M and Y, i.e.: no unmeasured variables U_i that confound the effects



What would go wrong?

Intermezzo: causal graphs (1)



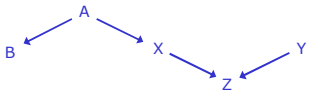
Causal graphs: a graphical representation of causal relationships between variables
Parents, children, ancestors and *descendants*

Z is a *collider:* a particular node on a path such that both the preceding and subsequent nodes on the path have directed edges going into that node
here: X → Z ← Y

In general, a path on a causal diagram does not need to follow the directions of the arrows: Z – X – A , B – A – X

Any path which contains a collider, is called a *blocked path*
A – X – Z – Y (Z is a collider on this path) ; otherwise *unblocked* B – A – X – Z

Intermezzo: causal graphs (2)



Two variables will only be statistically associated in the population as a whole if:

EITHER one is a cause of the other
X causes Z; A causes B

OR they share a common cause or ancestor
B and X are caused by A, as are B and Z

Intermezzo: causal graphs (3)

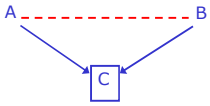
Conditioning on a variable is graphically represented by placing a box around that variable

Conditional on its parents, a variable C will be independent of all variables which are not descendants of C

Intermezzo: causal graphs (4)

Conditioning on children influences (introduces or alters) associations between parents/ancestors of that variable.

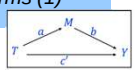
A: the battery is low
B: the gas tank is empty
C: the car does not start



Summarized, conditioning can:

- remove marginal dependencies
- introduce new (conditional) dependencies
- alter the magnitude of already existing dependencies

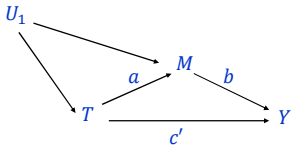
The assumption of uncorrelated error terms (1)



Assumption: uncorrelated error terms for T, M and Y,
i.e.: no unmeasured variables U_i that confound the various effects

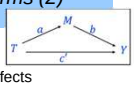
Assume the existence of a variable U_1.

Then instead of just the effect a, simultaneously the effect through the unblocked path $T - U_1 - M$ would be estimated!



Example:
T = smoking during pregnancy
M = DNA methylation
Y = birth-weight

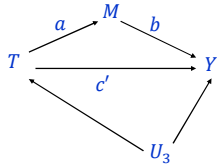
The assumption of uncorrelated error terms (2)



Assumption: uncorrelated error terms for T, M and Y,
i.e.: no unmeasured variables U_i that confound the various effects

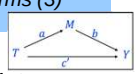
Assume the existence of a variable U_3.

Then instead of just the direct effect c' , simultaneously the effect through the unblocked path $T - U_3 - Y$ would be estimated!



Example:
T = smoking during pregnancy
M = DNA methylation
Y = birth-weight

The assumption of uncorrelated error terms (3)

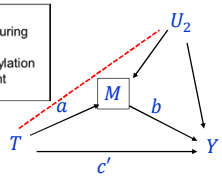


Assumption: uncorrelated error terms for T, M and Y,
i.e.: no unmeasured variables U_i that confound the various effects

Assume the existence of a variable U_2.

Then, conditioning on M in the mediation analysis would introduce bias:

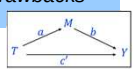
Example:
T = smoking during pregnancy
M = DNA methylation
Y = birth-weight



In addition to estimating c', we would also be estimating some spurious effect via the ``new'' unblocked path T- U_2 - Y!

Moreover: we would not be able to correctly estimate *b*

The approach so far: assumptions and drawbacks



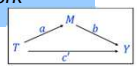
- Assumption: uncorrelated error terms, no unmeasured confounding (partly ignored by B&K!)
- Only valid for relatively simple and linear systems
- No exposure T- mediator M interaction possible

i.e. the estimation of the direct effect must not depend on the value of M (if so, we would need a population summary of the effects at different levels of the mediator)

Hence: we want more options!

Good news: a more general approach to Causal Mediation Analysis provides just that, within the counterfactual framework

Intermezzo: the counterfactual framework



Counterfactual/potential outcomes: the big ``what if?''

Example: baby from mother that actually smoked (T=t=1): *``What would this baby's birth-weight have been if its mother did not smoke?''*

Unobservable! (parallel universe…)

$M(t)$: potential value of mediator under treatment status $T = t$
$Y(t, m)$: potential value of outcome for $T = t$ and mediator value $M = m$

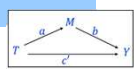
Actual observed variables M_i and Y_i for a subject *i* can be rewritten as:

$$M_i = M_i(T_i) \quad \text{and} \quad Y_i = Y_i(T_i, M_i(T_i))$$

The concept of counterfactuals provides a better definition of the causal effects involved…

Here: T=0,1, but generalizable to arbitrary reference points, T=t, T=t'

The Average Causal Mediation Effect (Natural Indirect Effect)



Definition: causal mediation effect under treatment status t for subject i:

$$Y_i(t, M_i(1)) - Y_i(t, M_i(0)), \qquad t = 0,1$$

Example: baby from mother that actually smoked (T=t=1):

actual observed birth-weight, with observed mediator value $M_i(1)$

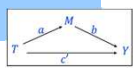
birth-weight potentially obtained if the mediator took the value as if the mother did not smoke

Average Causal Mediation (or Indirect) Effect (ACME):

$$E\{Y(t, M(1)) - Y(t, M(0))\}, \quad \text{for} \quad t = 0,1$$

ACME is the expected change in Y when one lets M change *as if* T did, while holding T constant → the effect of the T on Y through M

The Average Causal Direct Effect (Natural Direct Effect)



Definition: direct effect under treatment status t:

$$Y_i(1, M_i(t)) - Y_i(0, M_i(t)), \qquad t = 0,1$$

Example: baby from mother that actually smoked (T=t=1):

actual observed birthweight

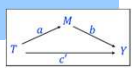
birthweight potentially obtained if its mother did not smoke, with unchanged mediator value $M_i(1)$

Average Direct Effect (ADE):

$$E\{Y(1, M(t)) - Y(0, M(t))\}, \quad \text{for} \quad t = 0,1$$

ADE is the expected change in Y when one lets T change, but M is held constant → represents all effects of T on Y, other than through *M*

Total Causal Effect



Total Causal Effect (TCE):

$$E\{Y(1, M(1)) - Y(0, M(0))\},$$

i.e. expected increase in the outcome Y as the treatment changes from T=0 to T=1, while the mediator M is allowed to track that change

Now, for the good news:

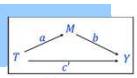
… conveniently skipping mathematical mediation formulas, underlying theorems and their – rather intimidating – mathematical proofs…

ACME, ADE (and TCE) can be estimated!

... by averaging over levels of M and measured covariates X (estimated by f.e. bootstrapping)

and at a relatively small cost: meeting (weaker version of) sequential ignorability assumption (≈``uncorrelated errorterms'')

Causal mediation analysis



Causal mediation analysis in a counterfactual framework hence provides:

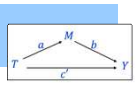
- a general, nonparametrical (!) measure of mediation, including formal definitions of direct and indirect effects (ADE, ACME) which not only …
- enhances understanding, but also allows…
- more & better estimation methods, improving validity, interpretation and
- application in a much wider range of models than the linear one (different types of variables, nonlinear effects (interaction, moderation…))

➢ *the R package ``mediation'' provides just that!*

NB: linear structural equation models (including B&K's approach) can be interpreted as an ACME estimator (adding parametric assumptions), so that:

ACME = $a * b$
ADE = c'
TCE = $c = c' + a * b$

R package ``mediation''



Two statistical models are needed:

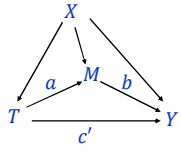
1) conditional distribution of M: $f(M|T, X)$

f.e.: *model.m ← lm(M ~T+X, data = …)*

2) conditional distribution of Y: $f(Y|T, M, X)$

f.e.: *model.y ← lm(Y ~T+M+X, data = …)*

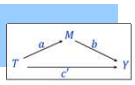
models specified by the researcher



Then *mediate(model.m,model.y)* uses these models to estimate ACME, ADE and TCE, with CI's based on bootstrapping (or other simulation methods)

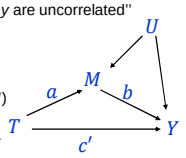
>>> non-parametrically, works for a large number of types of models/variables, works with interaction terms (f.e. T*M), with nonlinear effects of M on Y, …

R package ``mediation'' sensitivity analyses



Sequential ignorability assumption:
``the error terms ε_M and ε_Y from *model.m* and *model.y* are uncorrelated''

Sensitivity analysis: let's say they are not!
(``There is an unobserved confounder U,
responsible for part of the variances of both M and Y'')

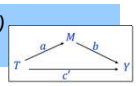


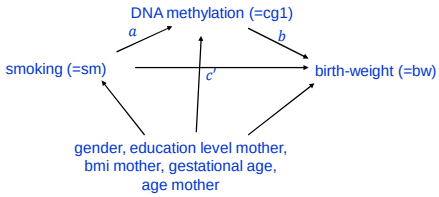
$\rho = corr(\varepsilon_M, \varepsilon_Y)$, ρ is sensitivity parameter

How does ACME change when ρ changes?
For which values of ρ does the ACME's CI contain zero?

> quantify how large U must be in order for your original conclusion to be reversed

R mediation: output example (1) Defining the models



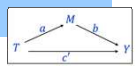



```
> model.m <- lm(cg1 ~ sm + gender + edum + bmi + gestage + agem, ...)

> model.y  <- lm(bw ~ cg1+ sm + gender + edum + bmi + gestage + agem, ...)

> out1.1 <- mediate(model.m, model.y, sims = 1000, boot = TRUE, treat = "sm", mediator = "cg1")
```

R mediation: output example (2) ACME, ADE and total effect




```
> summary(out1.1)
Causal Mediation Analysis
Nonparametric Bootstrap Confidence Intervals with the Percentile Method
```

$a * b$

	Estimate	95% CI Lower	95% CI Upper	p-value
ACME	-120.253	-185.437	-59.576	0.00
ADE	-143.085	-265.763	-13.228	0.03
Total Effect	-263.338	-377.250	-144.633	0.00
Prop. Mediated	0.457	0.221	0.918	0.00

```
> model.m
Call: lm(formula = cg1 ~ sm + gender + edum + bmi + gestage + agem, data = dd1)
Coefficients:
```

a

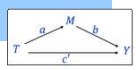
(Intercept)	sm	gender	edum	bmi	gestage	agem
0.6898096	-0.1019811	0.0355022	0.0009650	0.0008642	-0.0017235	0.0004985

```
> model.y
Call: lm(formula =bw ~ cg1 + sm + gender + edum + bmi + gestage + agem,data =dd1)
Coefficients:
```

b

(Intercept)	cg1	sm	gender	edum	bmi	gestage	agem
-5169.639	1179.167	-143.085	164.076	-27.690	10.210	182.721	8.379

R mediation: output example (3) Sensitivity analysis

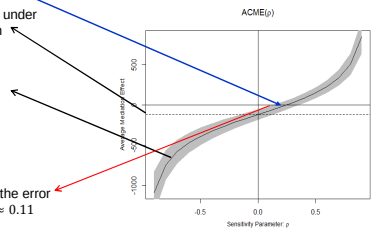


Mediation Sensitivity Analysis for Average Causal Mediation Effect
Sensitivity Region

Rho	ACME	95% CI Lower	95% CI Upper
[1,] 0.2	-24.4837	-82.4734	33.5060
[2,] 0.3	27.2946	-30.7656	85.3549

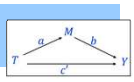
Estimated ACME (= -120) under the seq. ignor. assumption (ρ=0)

Estimated ACME under various values of ρ



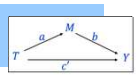
So: ACME < 0 as long as the error correlation ρ is less than ≈ 0.11

Concluding remarks



- Traditional mediation analysis approaches (B&K, Sobel) should be avoided, newer methods (f.e. based on bootstrapping of the mediation effect) provide better estimation through more reliable CI's
- Within the counterfactual framework: causal mediation analysis provides applications beyond simple linear models: nonlinear effects, moderation and interaction effects, various types of outcomes variables/models, mixed effects models, ...
- R package "mediation'': offers two estimation approaches (bootstrapping or approximate asymptotic distribution-based) and additional sensitivity analyses for testing robustness of violation of the assumptions

Literature on (causal) mediation:



- T.J. VanderWeele, J.M. Robins, `Four types of effect modification. A classification based on directed acyclic graphs', *Epidemiology* 18 (2007), 561-568
- A.F. Hayes, `Beyond Baron and Kenny: statistical mediation analysis in the new millennium', *Communication Monographs* 76 (2009) 408-420
- J. Pearl, `The mediation formula: a guide to the assessment of causal pathways in nonlinear models', in: C. Berzuini, D. Dawid, L. Bernardinelli (eds) *Causality: statistical perspectives and applications* (2012)
- J. Pearl, `Interpretation and identification in causal mediation', *Psychological Methods* 19 (2014) 459-481
- K. Imai, et al, `Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* 25 (2010) 51-71
- L. Küpers et al, `DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring', *Int.J. Epidemiol.* 44 (2015) 1224-1237

... S. Vansteelandth: R package ``medflex''

Next Help! Statistics! Lunchtime Lecture

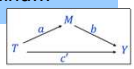
Basics of survival analysis

Douwe Postmus

April 11, 2017

Rode Zaal

Causal mediation comics are hard to find…



Mediator Trick #273: "The Mom Phone Call"
Used to break deadlocked mediations.

c.16CharlesFincher05.12 LawComixHome.com