PSY 5939: Longitudinal Data Analysis

Contents

1 Mediation

1.1 Theory

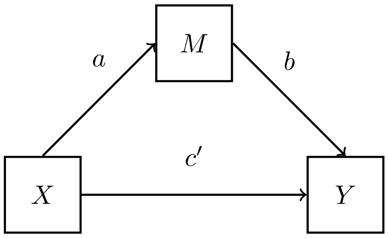
1.1.1 Statistical mediation

The effect of one variable (X) on a second variable (Y) is transmitted via a third variable (M)

M **mediates** the effect of X on Y

How does X **cause changes** in Y? By causing changes in M

1.1.2 Mediation model



1.1.3 Mediation theory

Many sources of theory for mediation models

- Stimulus → Organism → Response (SOR)
- Attitudes → Intentions → Behavior (Azjen)
- Intervention → Proximal behavior → Distal behavior

1.1.4 Why mediation?

Design and evaluation of multi-component interventions

- Can identify which specific components are changed by the intervention and in turn affect the outcome

Better understanding of causal ordering of the variables in time

- Three time points are better than two

Mediation is a model of **process** and processes unfold **over time**, so mediation is inherently a **longitudinal** model

- (Even though people us it cross-sectionally – more later)

1.2 Study design

1.2.1 Prevention and intervention

Many prevention / intervention studies are designed around the idea of **modifiable mediators**

The intervention changes a mediator, which in turn changes some outcome of interest

In this framework, mediators tend to be things like attitudes, intentions, norms, skills, etc., which can (potentially) be **changed**

Things like genetic susceptibility, height, gender are not really "modifiable"

1.2.2 Choosing mediators

Mediators are often chosen because they have a known relationship with the outcome of interest

Improved parent-child interaction lead to less child depression

- We design a depression intervention to *target parent-child interaction*

Blood lipid level has been shown to predict later heart disease

- We design a heart health diet to *decrease blood lipids*

Organizational skills are related to improved employee efficiency

- We design an employee program to *improve worker organizational skills*

1.3 Causal ordering

1.3.1 Mediation in time

Several different ways that X, M, and Y can exist in time

- Cross-sectional: $X_1 \rightarrow M_1 \rightarrow Y_1$
- Semi-longitudinal:

$X_1 \rightarrow M_1 \rightarrow Y_2$

$X_1 \rightarrow M_2 \rightarrow Y_2$

$X_1 \rightarrow M_1 \rightarrow Y_3$

$X_1 \rightarrow M_3 \rightarrow Y_3$

- Longitudinal*: $X_1 \rightarrow M_2 \rightarrow Y_3$

1.3.2 Cross-sectional estimates of longitudinal effects

Maxwell & Cole (2007) and Maxwell, Cole, & Mitchell (2011)

Cross-sectional mediation *almost always* produces **biased** estimates of longitudinal mediation effects

Cross-sectional mediation effects may be **higher** OR **lower** than the longitudinal effects

The reason for this should be fairly obvious

- How two variables are related at one time doesn't have much to do with how they are related across time

1.4 Third variable effects

1.4.1 Mediation vs moderation (interaction)

Both mediation and moderation involve a *third variable*

- There is a relationship between X and Y, and then there's this third variable (Z or M)

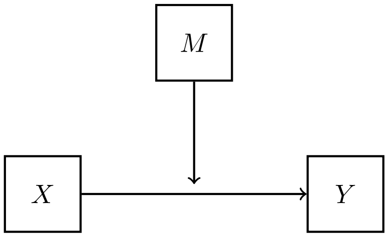
Moderation: Z **changes** the relationship between X and Y
Effects for one group, no effects for the other group
Positive effect for one group, negative effect for other

- *Who* does it work for?

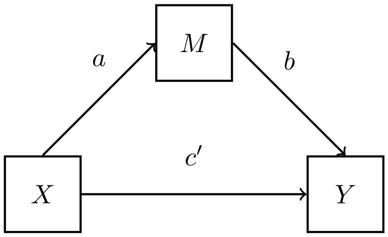
Mediation: Z **causes** the relationship between X and Y
Z is the mechanism by which X is related to Y

- *How* does it work?

1.4.2 Moderation / interaction



1.4.3 Mediation

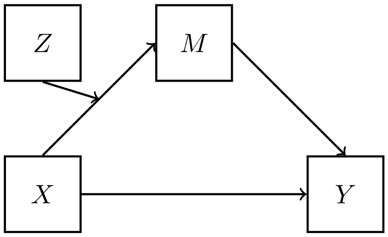


1.4.4 Moderated mediation (and mediated moderation)

You can have both moderation and mediation in the same model

Moderated mediation (below) is common

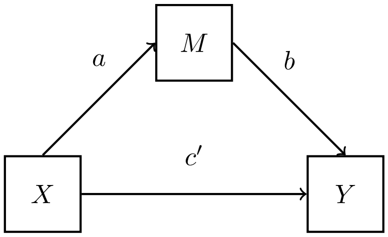
More recent work refers to this as “causal process” model (Hayes) because moderated mediation is hard to say and people get confused



2 Indirect effects

2.1 Mediation model

2.1.1 Mediation model



2.1.2 Mediation equations

a path:

$$\hat{M} = i_{MX} + aX$$

b and c’ paths:

$$\hat{Y} = i_{YXM} + bM + c'X$$

c path:

$$\hat{Y} = i_{YX} + cX$$

2.1.3 Simulated dataset

a path:

```
a_model <- tidy(lm(m ~ x, med_data))
a_model
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   0.0662     0.311     0.213  0.832
## 2 x             0.584      0.273     2.13   0.0353
```

b and c’ paths:

```
bcp_model <- tidy(lm(y ~ x + m, med_data))
bcp_model
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -0.174     0.478    -0.364  0.717
## 2 x              0.130     0.431     0.302  0.763
## 3 m              0.310     0.156     1.99   0.0489
```

c path:

```
c_model <- tidy(lm(y ~ x, med_data))
c_model
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -0.154     0.485    -0.317   0.752
## 2 x              0.311     0.427     0.728   0.468
```

2.1.4 Quantifying mediated effects

All of these equations describe the complex (sort of) relationships among X, M, and Y

But what we really want is a **single number** that tells us about the mediated effect

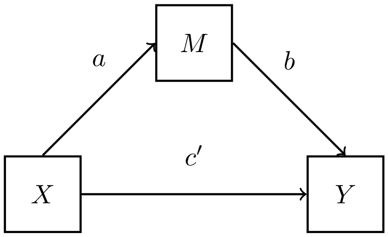
- Preferably with a **statistical test**

There are two different (but *generally* equivalent) ways to quantify the mediated effect

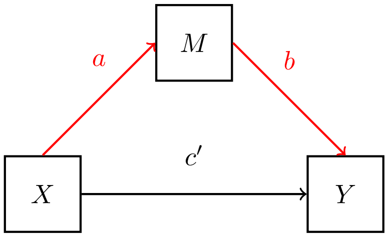
- Product method
- Difference method

2.2 Mediated effect as product

2.2.1 Mediated effect as product



2.2.2 Mediated effect as product



2.2.3 Mediated effect as product

The mediated effect is the effect of X on Y via M

In SEM, such a path is described as the **product** of the regression coefficients that go into it

The a coefficient reflects the X → M path

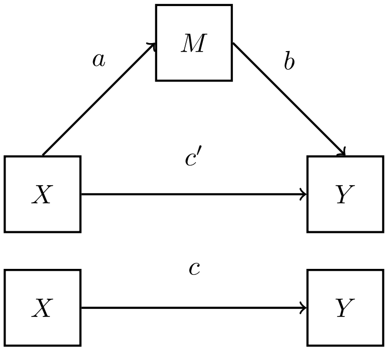
The b coefficient reflects the M → Y path

The mediated effect is $a \times b$

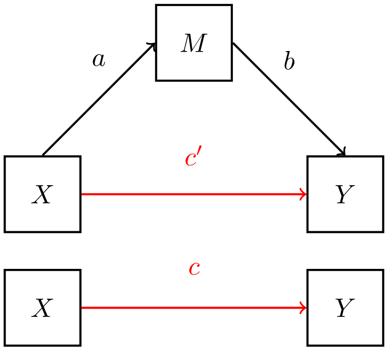
- Simulated data: $a \times b = 0.584 \times 0.31 = 0.181$

2.3 Mediated effect as difference

2.3.1 Mediated effect as difference



2.3.2 Mediated effect as difference



2.3.3 Mediated effect as difference

The mediated effect is the effect of X on Y via M

The c coefficient reflects the **total effect** of X on Y

The c' coefficient reflects the **direct effect** of X on Y

The **difference** between total effect and the direct effect must be the **indirect** portion that is transmitted through M

The mediated effect is $c - c'$

- Simulated data: $c - c' = 0.311 - 0.13 = 0.181$

2.4 Mediated (indirect) effect

2.4.1 Mediated (indirect) effect

In the basic mediation model, when M and Y are both **continuous** and **normally distributed** and **linear regression** is used,

$a \times b = c - c'$

(So it doesn't matter which way you calculate the mediated effect)

- Simulated data: $a \times b = c - c' = 0.181$

2.4.2 Equivalence

However, the two estimates of the mediated effect will typically **not** be equivalent in other situations, such as

- M and/or Y are not continuous and non-OLS regression is used (such as logistic regression for binary outcome)
- There is clustering within the data (such as multiple children from the same family) and multilevel/mixed models are used
- When moderation of the a and/or b paths occurs

Studies show that $a \times b$ is closer to the population mediated effect in these situations

- We will focus on tests of the a × b estimate

2.4.3 Categorical M or Y

Alternative methods when M and/or Y are non-normal

Geldhof, G. J., Anthony, K. P., Selig, J. P., & Mendez-Luck, C. A. (2018). Accommodating binary and count variables in mediation: A case for conditional indirect effects. International Journal of Behavioral Development, 42(2), 300-308.

The slopes (a and b) are the **derivative** of their respective equations

- Use the derivative of the equation even when the equation is nonlinear

Use the **derivative** of the equations

- Special derivative that depends on the model that was used
- Count M and Poisson regression: $ae^{i_{MX}+aX}$ instead of a
- Binary M and logistic regression: $\frac{ae^{i_{MX}+aX}}{(1+e^{i_{MX}+aX})^2}$ instead of a

3 Statistical tests

3.1 "Classic" methods

3.1.1 Causal steps ("Baron & Kenny")

Judd & Kenny (1981), Baron & Kenny (1986)

A series of *logical* steps to establish mediation

If there is "failure" at any point, there is not mediation

1. Test for significance of the c path (X → Y, ignoring M)
2. Test for significance of the a path (X → M)
3. Test for significance of the b path (M → Y, controlling for X)
4. B & K later add: Test for non-significance of c' path (X → Y , controlling for M)
5. Even later add: Test that c' is at least smaller than c

Later authors suggest relaxing step 1, continue to step 2 even if step 1 fails

3.1.2 Simulated data: Causal steps ("Baron & Kenny")

Simulated data:

- Step 1: c path is significant, c = 0.311, p = 0.468 :'(

With later criteria: continue even if step 1 fails

- Step 2: a path is significant, a = 0.584, p = 0.035 ✓

- Step 3: b path is significant, b = 0.31, p = 0.049 ✓
- Step 4: c' path is NS, c' = -0.174, p = 0.717 ✓

3.1.3 Sobel test

Sobel (1982, 1986)

Sobel calculated the standard error to conduct a Wald test (z-test) of the mediated effect ($a \times b$):

$$z_{a \times b} = \frac{a \times b}{se_{a \times b}}$$

The standard error of the product is calculated using the a and b coefficients and their respective standard errors:

$$se_{a \times b} = \sqrt{a^2 \sigma_b^2 + b^2 \sigma_a^2}$$

where σ_b^2 is the **squared** standard error of b and σ_a^2 is the **squared** standard error of a

- Simulated data: $z_{a \times b} = 0.181/0.124 = 1.457$

3.1.4 Why not to use these "classic" methods

Causal steps is great in terms of the **logic** of mediation, BUT

- No actual statistical test of the whole framework
- Steps 1, 4, and 5 make the process conservative overall
- Low power and high type I error

Sobel test is a Wald test, which assumes a *normal distribution*

- Low power and high type I error
- Even if both a and b are normal, $a \times b$ is typically **not** normal

The true sampling distribution of $a \times b$ is complicated (not normal) and depends on a, b, n, corr(a, b)

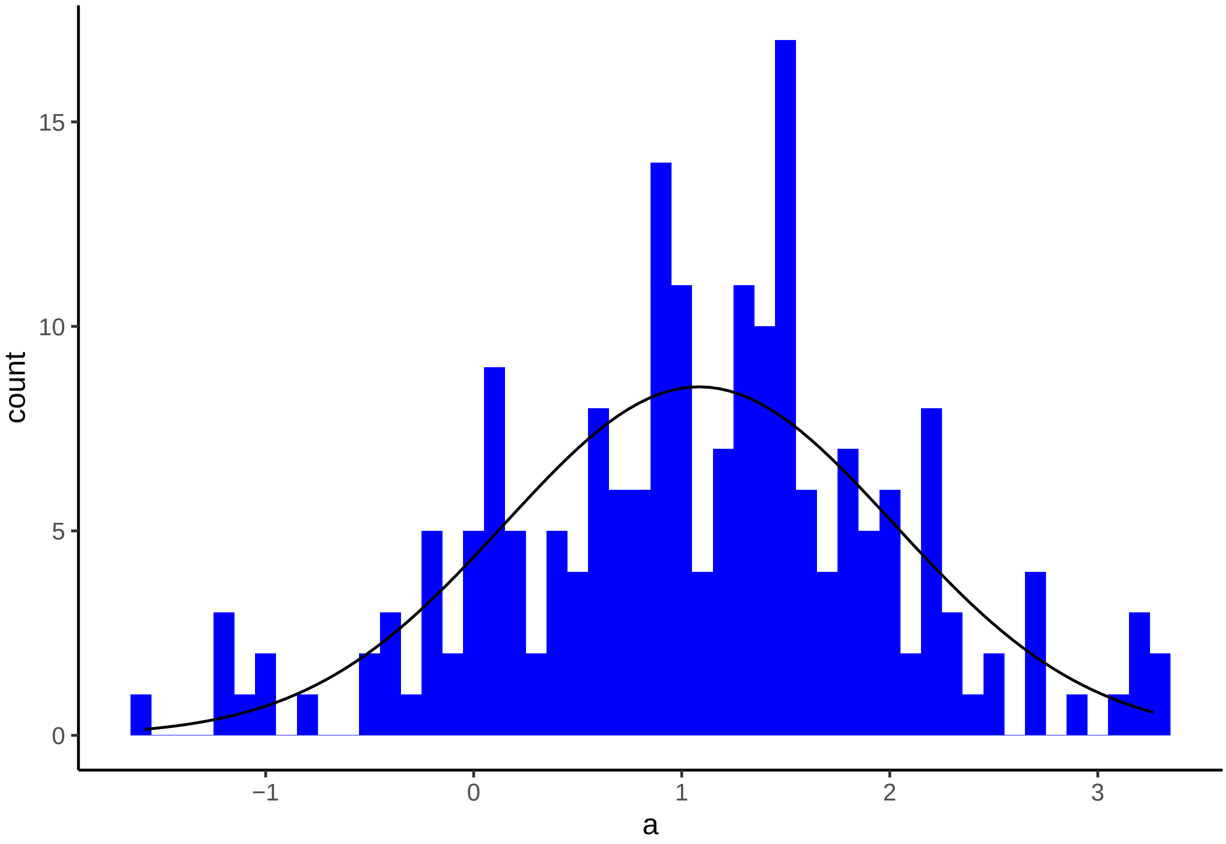
3.1.5 Modern methods for mediation

There are four major methods that you should be using to evaluate the statistical significance of the mediated effect

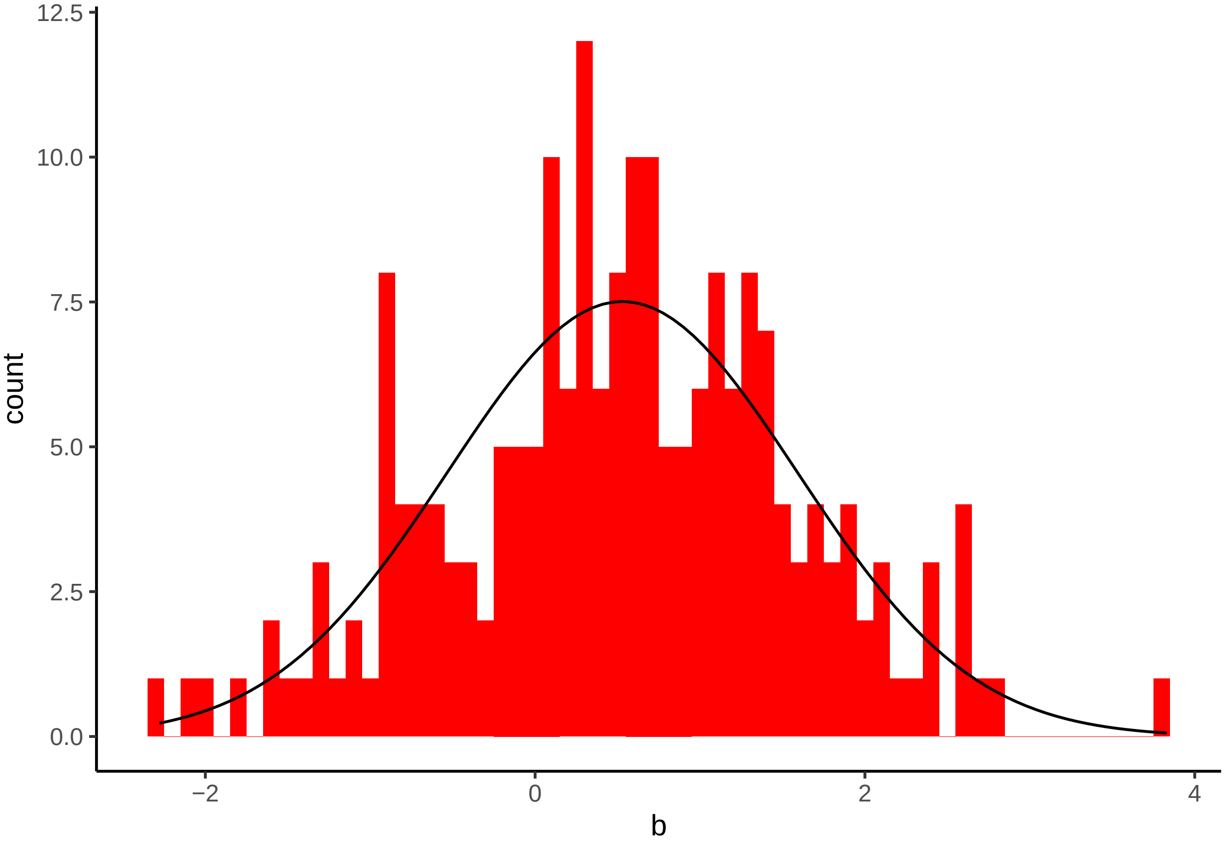
- Distribution of the product
- Joint significance
- Bootstrapping
- Monte Carlo simulation

3.2 Distribution of the product

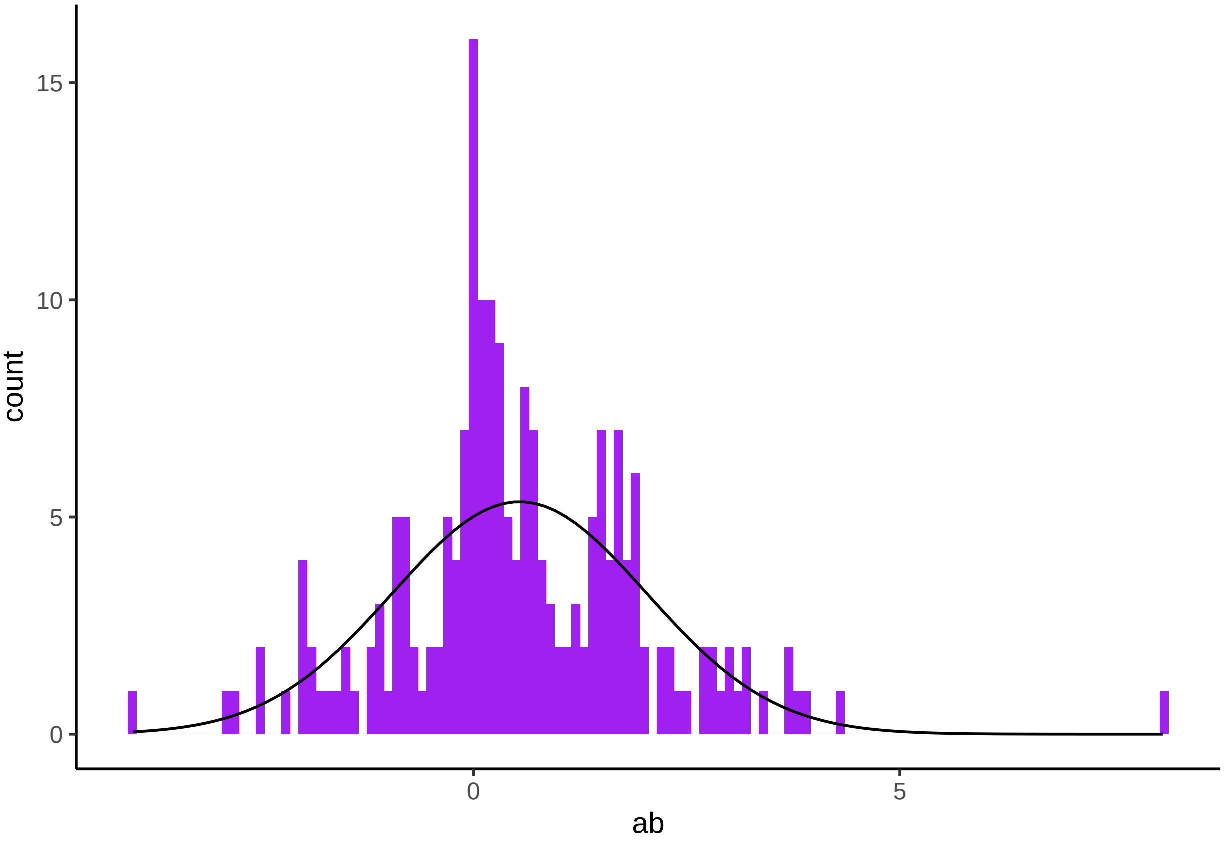
3.2.1 a is normally distributed ($N(1, 1)$)



3.2.2 b is normally distributed ($N(0.5, 1)$)



3.2.3 a × b is NOT normally distributed



3.2.4 Distribution of the product of 2 normal variables

Meeker, Cornwell, & Aroian (1981) presented tables of the distribution of the product of two standard normal variables

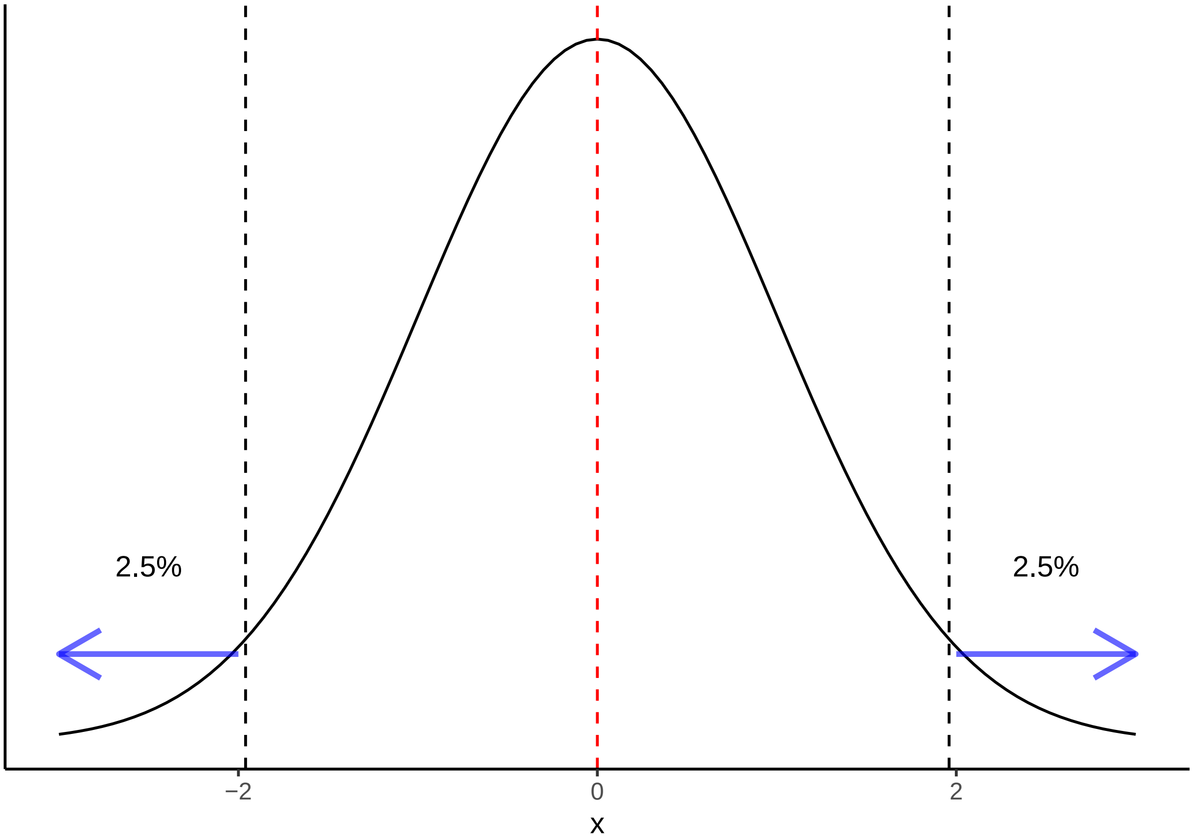
MacKinnon et al. (2007) implemented and expanded these tables with their PRODCLIN program: Distribution of the **PROD**uct **C**onfidence **L**imits for **IN**direct effects

Programs for SAS, SPSS, R, stand-alone
Web-based: https://amplab.shinyapps.io/MEDCI/

- Input the values for a and b, their respective standard errors, the correlation between a and b, and the Type I error rate

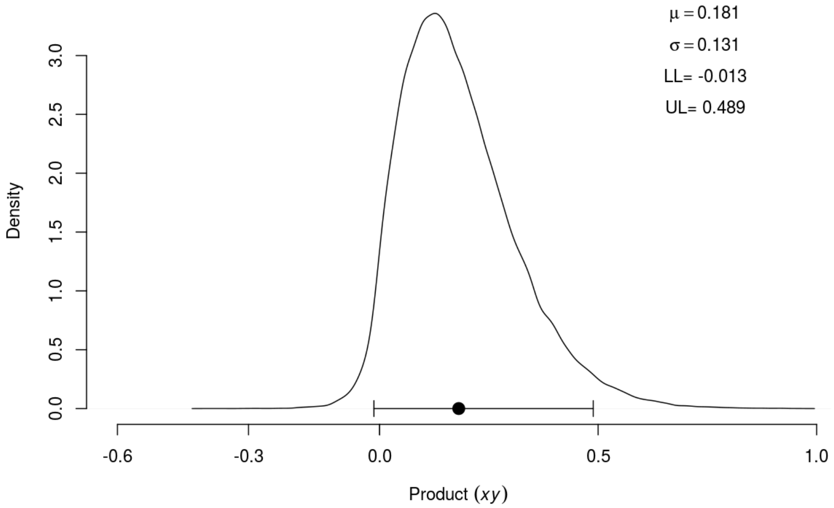
- Get back the **asymmetric** confidence limits on the mediated effect

3.2.5 Symmetric confidence limits



3.2.6 Asymmetric confidence limits

For asymmetric distributions, the confidence limits are **asymmetric** around the parameter estimate



3.2.7 a × b estimate and asymmetric CI

Look back at the last slide, showing the distribution of a × b

The estimate (mean) for that distribution was 0.181 ($a \times b$)

If we were creating a 95% confidence interval on that estimate, we would look for a *lower confidence limit* (2.5% of the distribution is lower) and an *upper confidence limit* (2.5% of the distribution is higher)

The lower 95% confidence limit is -0.013

The upper 95% confidence limit is 0.489

3.3 Joint significance

3.3.1 Joint significance

Test for the significance of both the a and b paths

Using standard t-tests of the regression coefficients

- a path needs to be significant
- b path needs to be significant

Advantages: easy, can be used for complicated models, good statistical power and type I error rate

Disadvantages: no actual estimate of the mediated effect (but you can use a × b), no confidence interval on the mediated effect

MacKinnon et al. (2002) describes the statistical properties

Simulated data:

- a path is significant, a = 0.584, p = 0.035 ✓
- b path is significant, b = 0.31, p = 0.049 ✓

3.4 Bootstrap

3.4.1 Bootstrapping

Most statistical tests assume a **theoretical** distribution

- e.g., z-test assumes normal distribution, regression coefficients are usually tested against a t-distribution

In some situations, you can't depend on theoretical distributions

- theoretical distribution is unknown
- theoretical distribution is very complicated
- sample size is too small for theory to hold

We can instead use an **empirical** distribution

An empirical distribution uses the data to create a distribution

Empirical distributions are created using statistical simulations

3.4.2 Bootstrapping

Resampling method to create an empirical distribution

Efron & Tibshirani (1994) developed the bootstrap as a *nonparametric* method to calculate variance of an estimate (i.e., standard error)

Have a sample of size n

1. Draw a sample (with replacement) of size n from the sample

2. Perform statistical analysis (e.g., regression) on the sample

3. Record the statistic you get from that sample

Repeat steps 1 through 3 many times (1000 or more)

3.4.3 What do we do with those 1000 resamples?

Create confidence intervals using the bootstrap distribution

Order 1000 resamples

For a 95% CI, 2.5% of observations in each tail

- 25th value is the lower limit
  - (2.5% <= this value)
- 976th value is the upper limit
  - (2.5% >= this value)

(Unless you write your own program, this all happens behind the scenes)

3.4.4 Simulated data: Bootstrap

```
library(lavaan)
```

```
## This is lavaan 0.6-10
## lavaan is FREE software! Please report any bugs.
```

```
model <- ' # direct effect
             y ~ cp*x
           # mediator
             m ~ a*x
             y ~ b*m
           # indirect effect (a*b)
             ab := a*b
           # total effect
             total := cp + (a*b)
         '
fit <- sem(model, data = med_data, se = "bootstrap")
summary(fit, ci = TRUE)
```

```
## lavaan 0.6-10 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         5
##
##   Number of observations                           100
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
```

```
## Parameter Estimates:
##
##   Standard errors                            Bootstrap
##   Number of requested bootstrap draws             1000
##   Number of successful bootstrap draws            1000
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##   y ~
##     x         (cp)    0.130    0.434    0.300    0.765   -0.742    0.900
##   m ~
##     x          (a)    0.584    0.233    2.508    0.012    0.104    1.023
##   y ~
##     m          (b)    0.310    0.168    1.848    0.065   -0.021    0.632
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##    .y                21.150    2.909    7.271    0.000   14.970   26.453
##    .m                 9.012    1.192    7.558    0.000    6.499   11.237
##
## Defined Parameters:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##     ab                0.181    0.127    1.422    0.155   -0.019    0.488
##     total             0.311    0.462    0.674    0.500   -0.618    1.089
```

3.4.5 Bootstrap programs

Lockwood & MacKinnon (1998)

Preacher: http://www.quantpsy.org/supp.htm

Hayes' PROCESS: http://www.processmacro.org/

Various for R: `confint()`, `boot()`, `mediate()` in the **psych** package, option in **lavaan**

Mplus, AMOS, other SEM programs can do bootstrap tests too

3.4.6 Bootstrap for mediation

The distribution of a × b is asymmetric and not normal

Use bootstrap to create an empirical distribution of the mediated effect (a × b)

Use empirical distribution to construct confidence intervals / test significance of a × b

3.5 Monte Carlo simulation

3.5.1 Monte Carlo simulation

Simulate MANY observations that follow a specified distribution

For mediation, we know that the sampling distributions of a and b are each normally distributed

- The problem is the product of them, ab, it's NOT normal

But we could simulate a and b (which we know are normal) and then create ab from those values

(This is actually exactly what I did in the slides to show you that ab is not normally distributed)

3.5.2 Monte Carlo simulation

The standard error for a regression coefficient is the **standard deviation** of its sampling distribution

- Simulate 1000+ cases of a ~ N(mean(a), s.e.(a))
- Simulate 1000+ cases of b ~ N(mean(b), s.e.(b))
- Create the product, ab
- As with bootstrap, order the observations
- As with bootstrap, select the values at the 2.5%ile and 97.5%ile (for a 95% CI)

3.5.3 Monte Carlo for mediation

The distribution of a × b is asymmetric and not normal

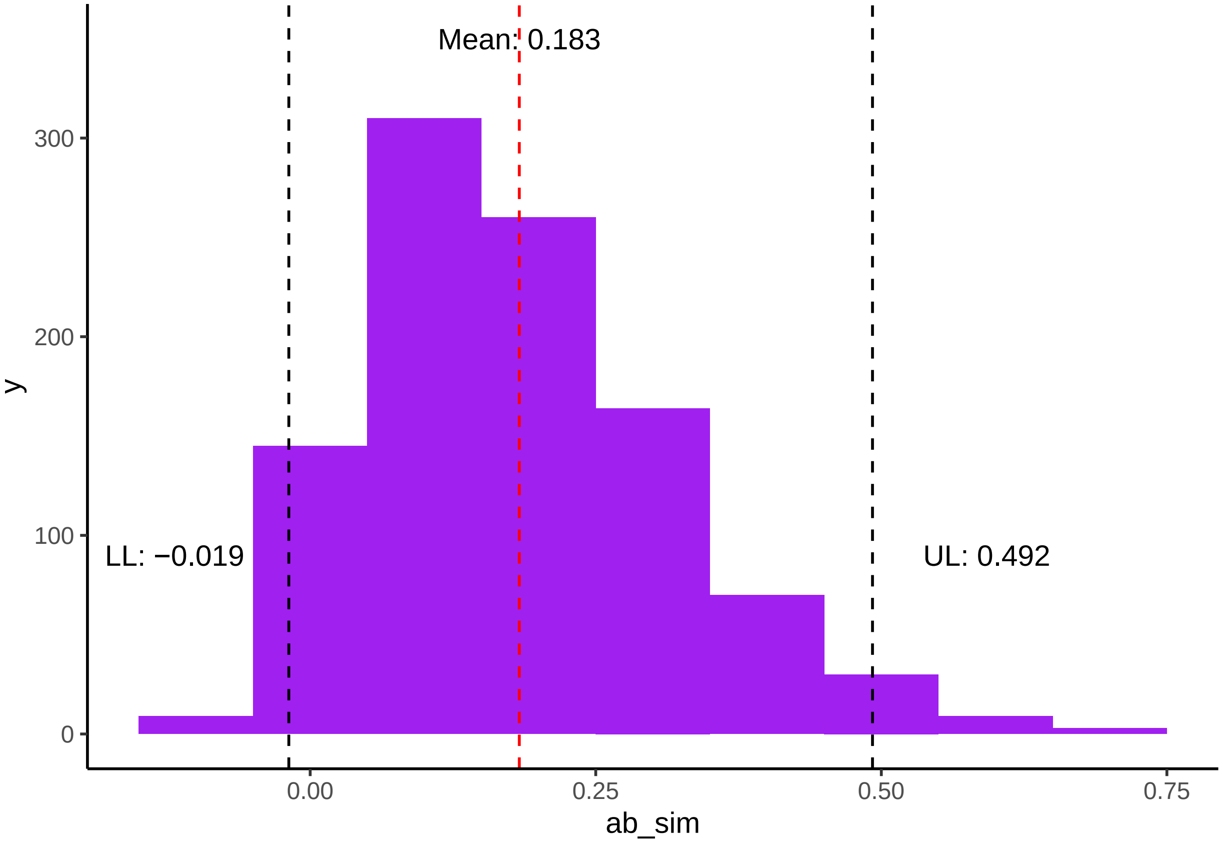
Use Monte Carlo to create an **empirical** distribution of the mediated effect (a × b)

Use empirical distribution to construct confidence intervals / test significance of a × b

- Preacher & Selig (2012) describe use for indirect effects
- A simple tool I made here: https://stefany.shinyapps.io/SimpleMediation/

3.5.4 Asymmetric confidence limits

For asymmetric distributions, the confidence limits are **asymmetric** around the parameter estimate



3.6 Summary of tests

3.6.1 Summary of modern mediation tests

MacKinnon et al. (2002), MacKinnon et al. (2004)

Joint significance: best balance of type I error and statistical power across conditions (sample size, effect size)

Product of coefficients: pretty good but difficult to actually use until PRODCLIN

Bootstrap: better confidence intervals than most other methods, requires programming skill or use of additional program, very flexible for more complex designs

Monte Carlo: better confidence intervals than most other methods, requires programming skill or use of additional program, very flexible for more complex designs

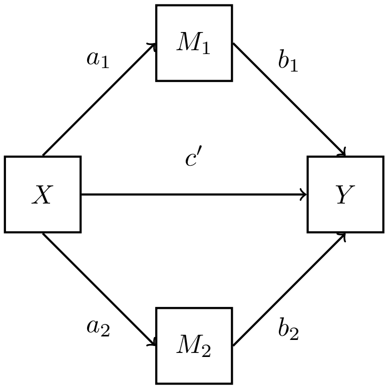
3.6.2 Summary of results in simulated data

Method	Indirect effect	Test statistic	Confidence interval	Significant?
Baron and Kenny	0.181			Yes
Sobel test	0.181	z = 1.457	[0.057, 0.305]	No
Dist of product	0.181		[-0.013, 0.489]	No
Joint significance	0.181			Yes
Bootstrap	0.181	z = 1.422	[-0.019, 0.488]	No
Monte Carlo	0.181		[-0.019, 0.492]	No

3.6.3 More complex designs

More than just a single observed variable as X, M, and / or Y

Multiple mediators: simultaneous (below) or sequential


M_1
a_1
b_1
c'
X
Y
a_2
M_2
b_2