```
course = "Transdisciplinary Data Science"


lesson_iteration = 6
lesson_title = "Data simulation: use cases and applications"


auth = "Ian Hussey"
dept = "Psychology of Digitalisation; Institute of Psychology"
```

```
course = "Psychology of Digitalisation"


lesson_iteration = 6
lesson_title = "Simulated data and computational simulation studies"


auth = "Ian Hussey"
dept = "Psychology of Digitalisation"
```

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\Gamma(n) = (n-1)!$$

$$p = 2 \cdot P(T_{df} > |t|)$$

Why I’m here

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\Gamma(n) = (n-1)!$$

$$p = 2 \cdot P(T_{df} > |t|)$$

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Why I’m here

p-hacking

and how to avoid it

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Why I’m here

Background

of the lecture

- Bad at math
  - I don’t have the matrix algebra to solve problems
  - but I can brute force them with relatively simple code

- ‘How do I learn this new method, while not conditioning my results on my data?

- ‘Violating assumptions of statistical tests is bad’
  - How bad?
  - Under what circumstances?

- What questions about human (researcher) behavior can we answer based on modelling the necessary implications of known facts?

Background

of the lecture

You

- Faculty / Institute / Subject?

Not a tutorial

Be aware that you can use simulated data & simulation studies to:

- Reduce p-hacking
  - Separate learning an analysis method from using that method

- Understand the methods themselves better
  - By knowing the ground truth

- Model and understand phenomena

Avoid p-hacking

using simulated data

Most common masters project:

1. Collect data (takes ages)
2. Once data collection is complete
    1. Panik
    2. Figure out how to process data
    3. Figure out how to analyse

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Avoid p-hacking

using simulated data

Best project

1. Start data collection
    1. Complete the study a few times yourself
    2. Understand & check data it produces
    3. Simulate more data
    4. Use that to write processing and analysis code
2. Data collection finishes
    1. Run data processing and analysis script once, report.

Realism

in simulated data

How realistic is simulated data?

As realistic as you need.

As lazy as it can be.

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Realism

in simulated data

How realistic is simulated data?

As realistic as you need.

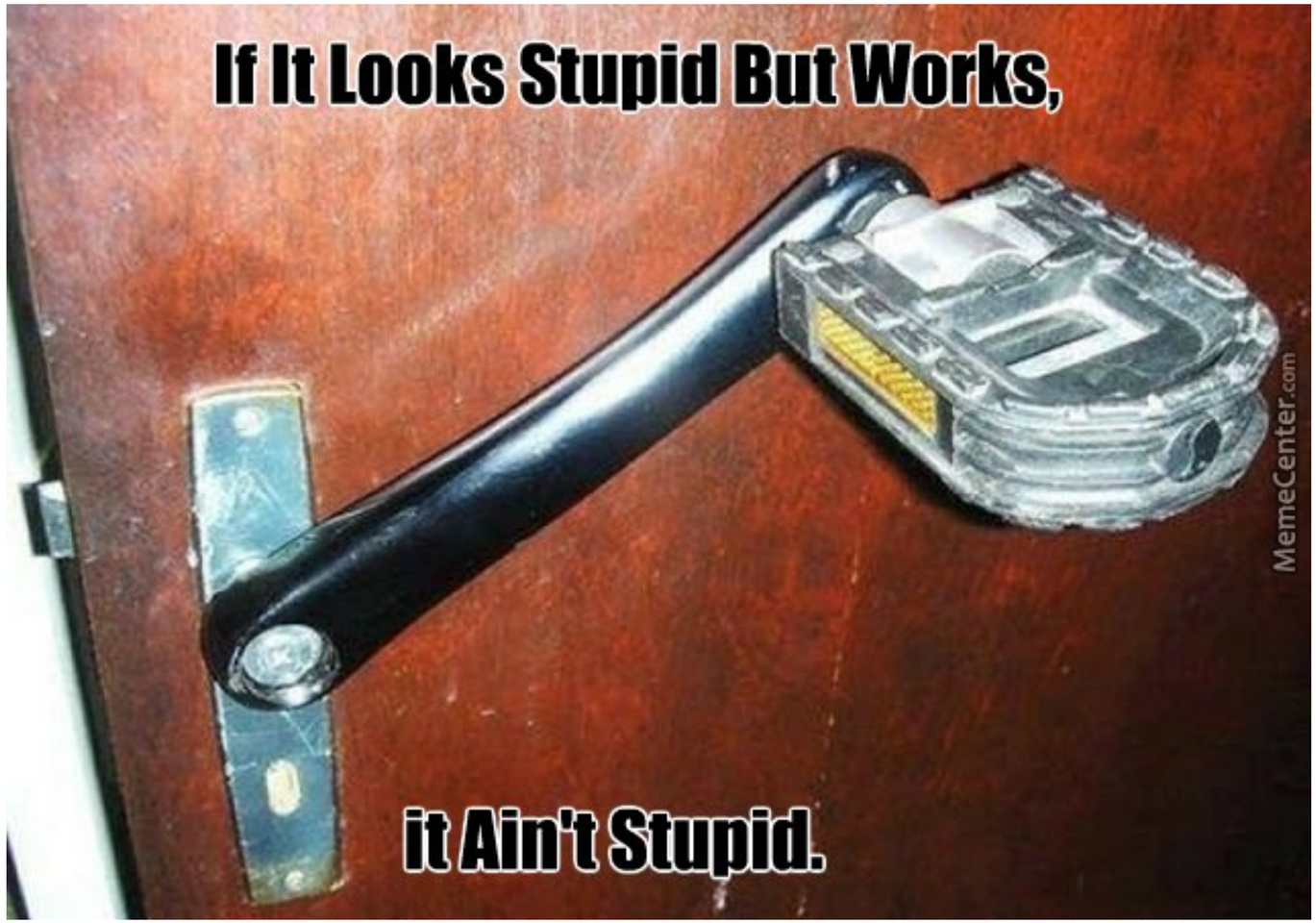
As lazy as it can be.

- Complete your own study a few times
- Simply type random data into excel
- Get fancy and do it in R
  - Distribution
  - Associations
  - Causal structure

Realism

in simulated data

How realistic is simulated data?



aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"


WARNING
LABEL SIMULATED DATA
VERY CLEARLY
SO IT IS NOT CONFUSED
WITH REAL DATA

Compare & contrast

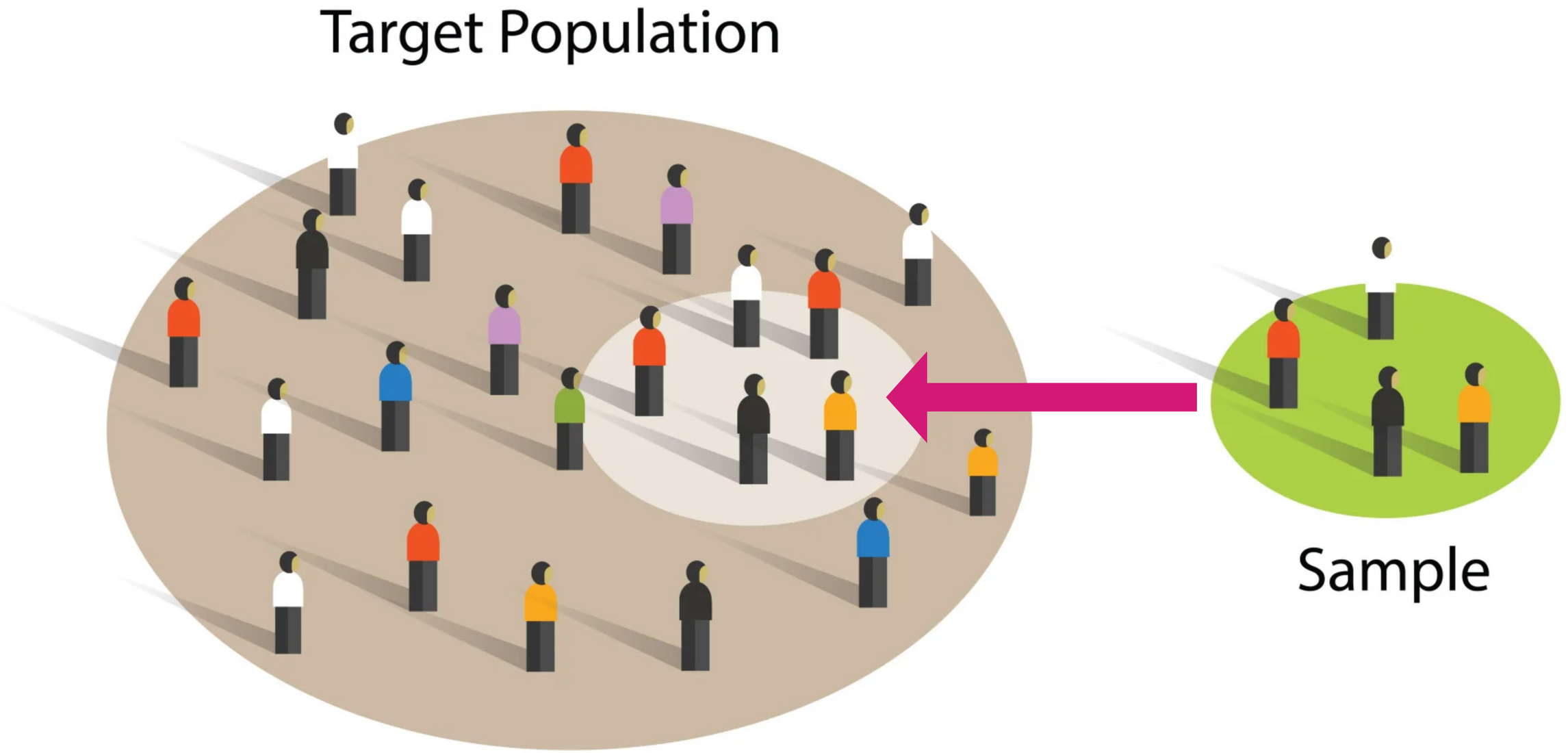
Real world vs Monte Carlo simulation studies

Logic of simulation studies

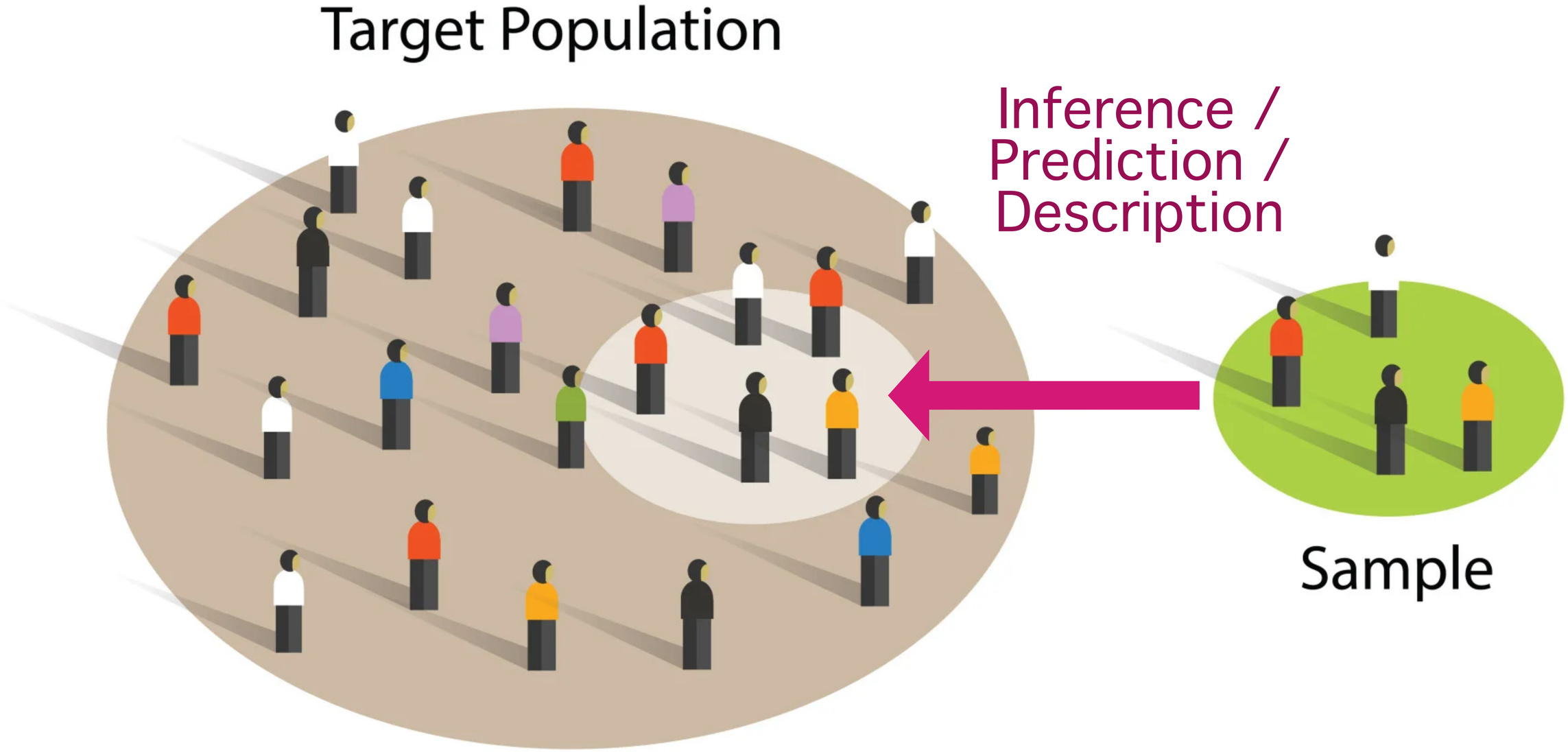
- Real-world research
  - Descriptions, inferences, predictions about the world in general using limited data
  - You never know when you're right!
  - Probabilistic statements only

- Simulation studies
  - Create ground truth, use it to understand other things
  - Understand methods themselves
  - Create principled models of the world derived from data

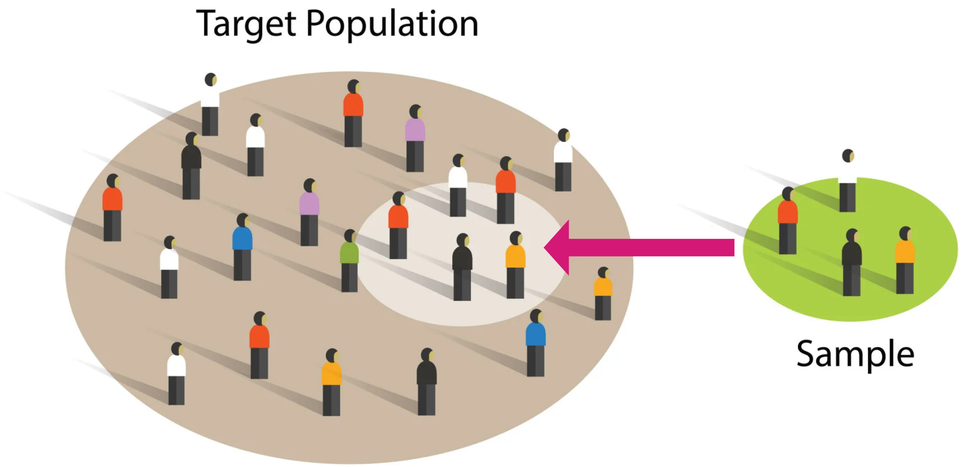
Empirical research



Empirical research

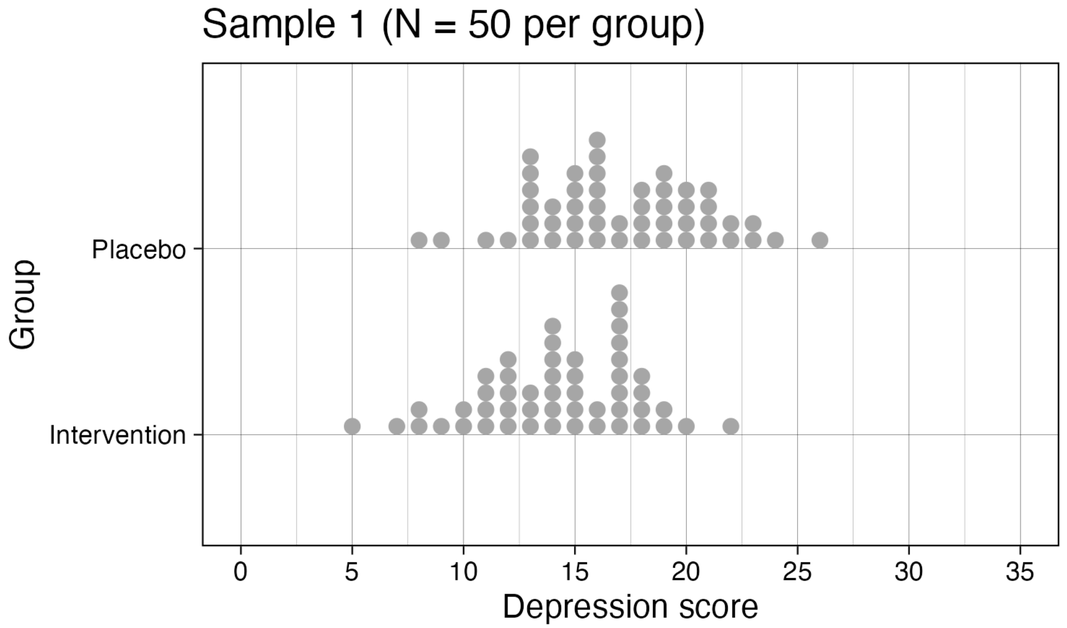


Empirical research

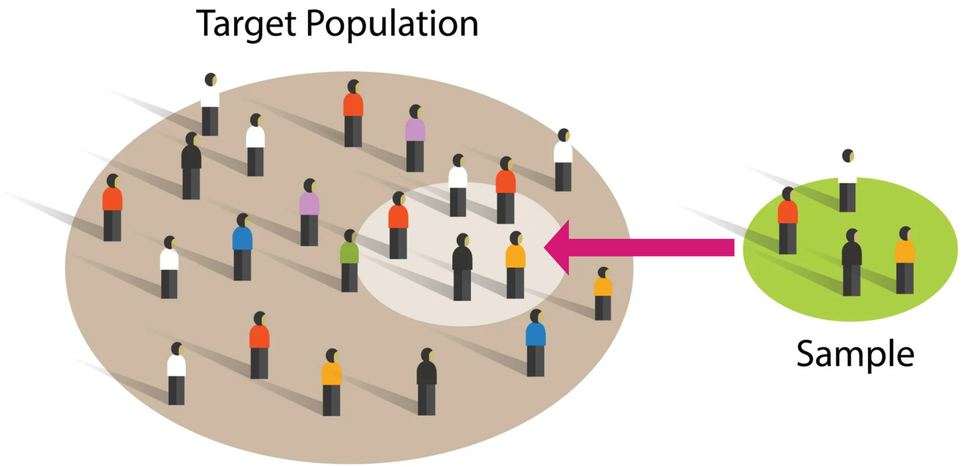


Population Distribution

Intervention works?
Medium effect size?

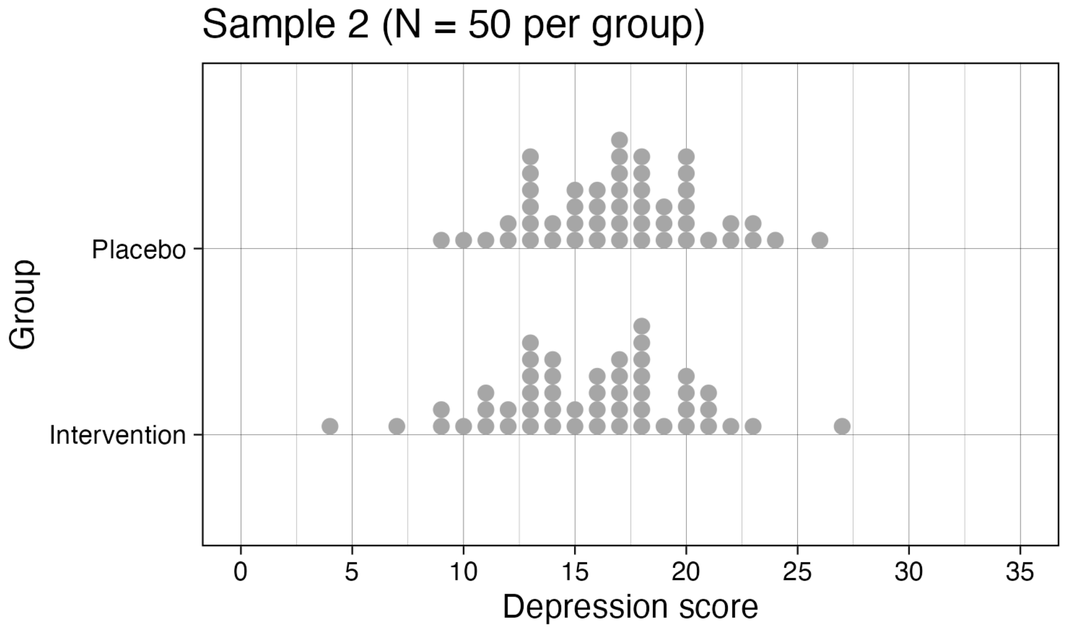


Empirical research

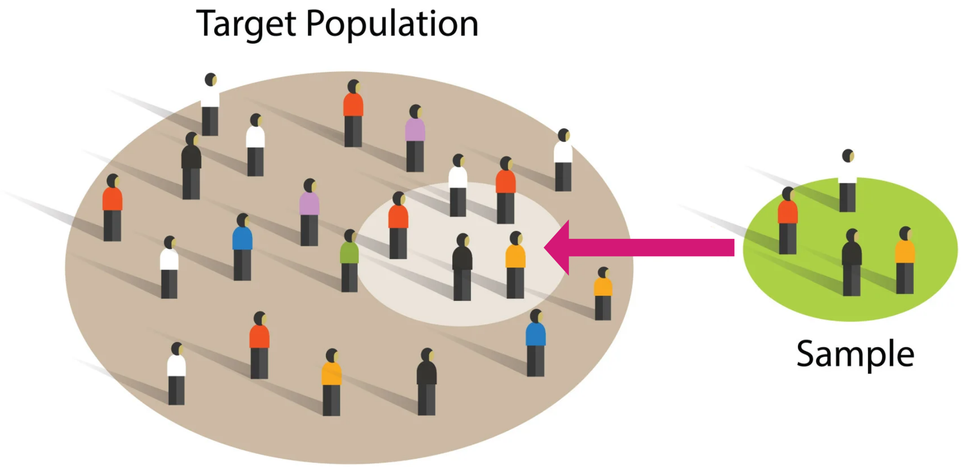


Population Distribution

Intervention works?
Small effect size?

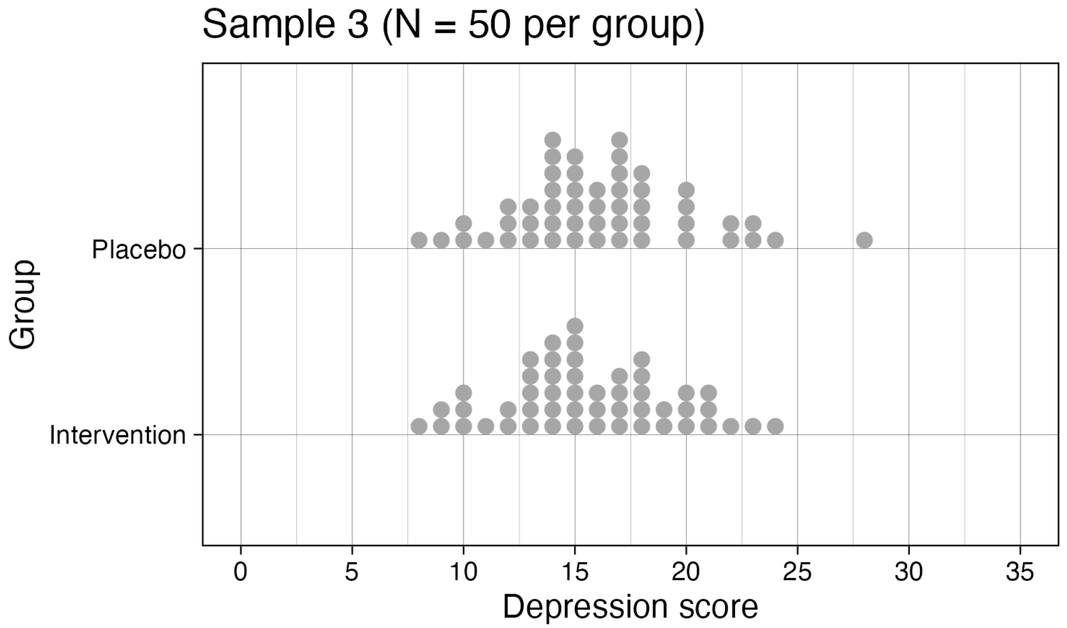


Empirical research

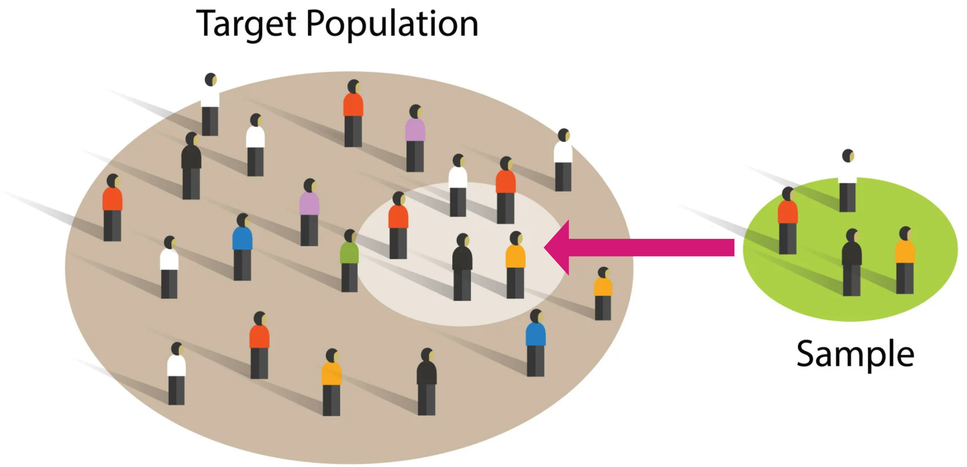


Population Distribution

Intervention doesn't work?
No effect?

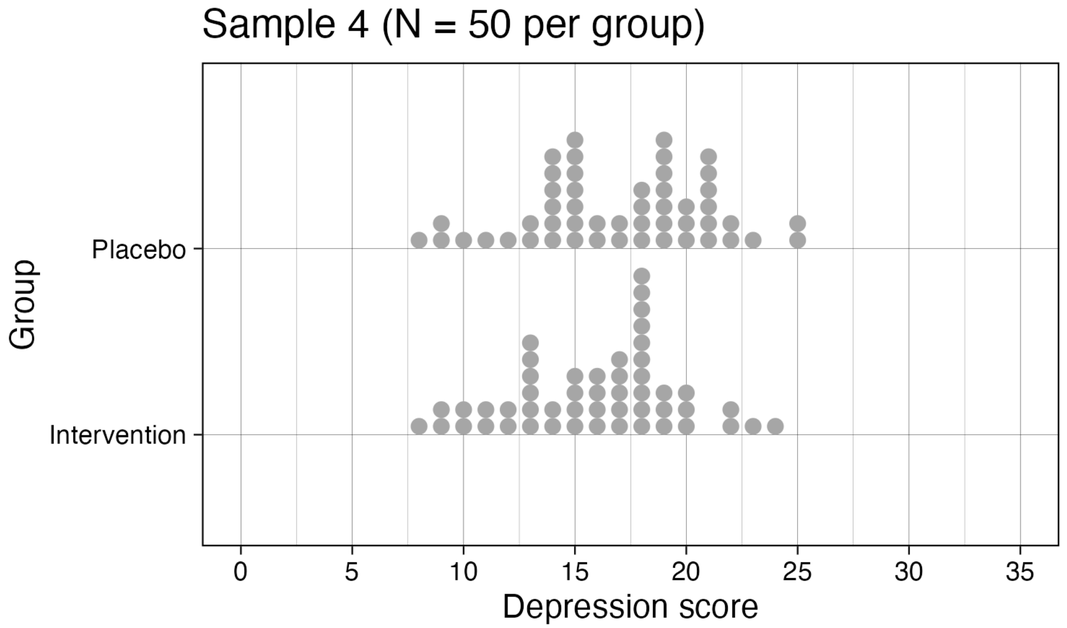


Empirical research



Population Distribution

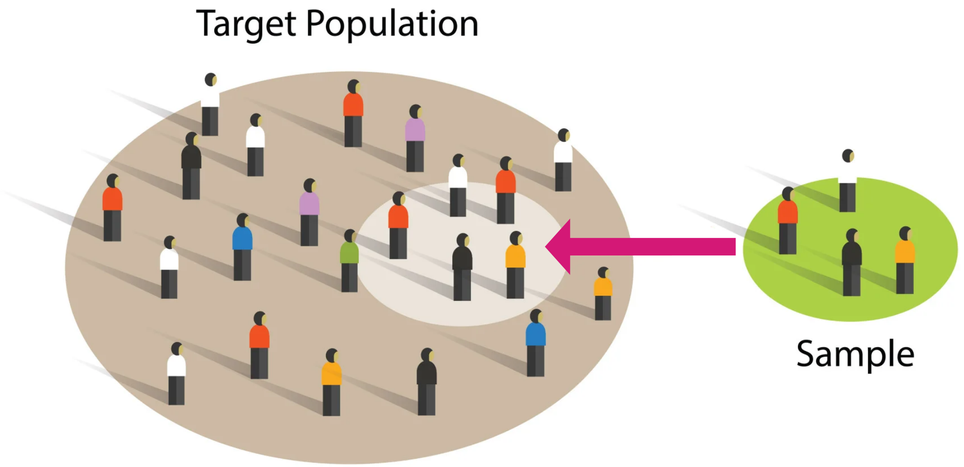
Intervention doesn’t work?
No effect?



aut = "Ian Hussey";

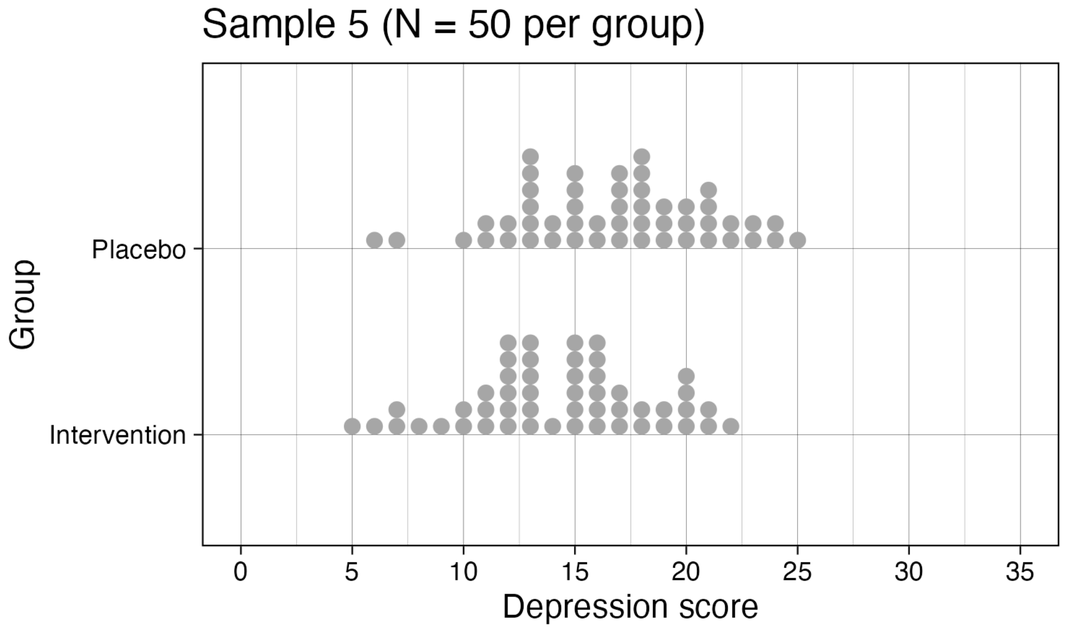
dept = "Psychc

Empirical research

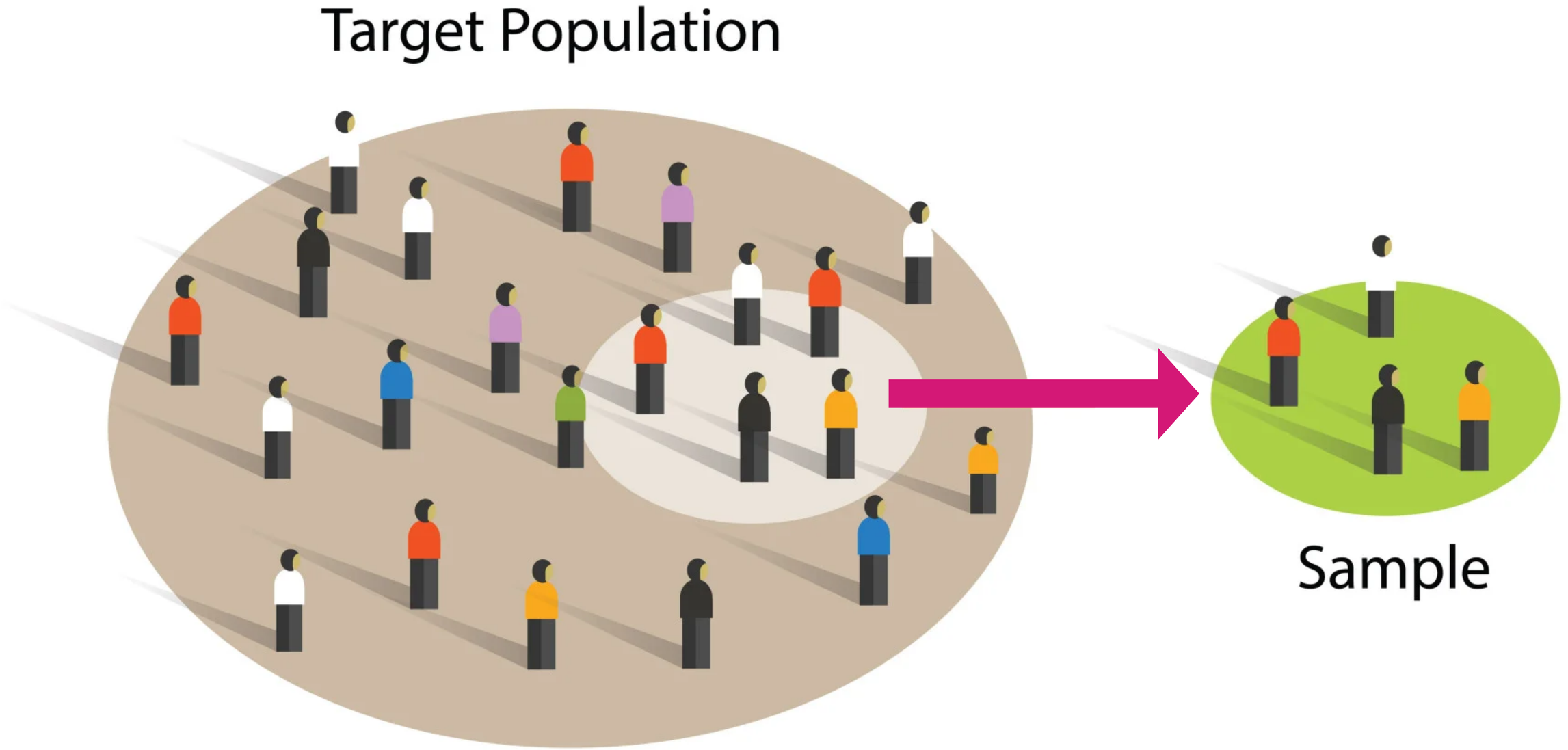


Population Distribution

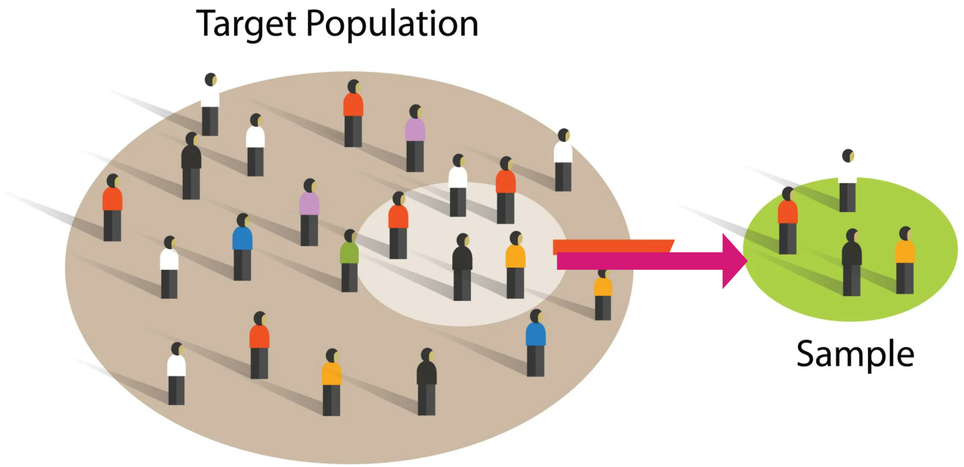
Intervention works?
Small effect size?

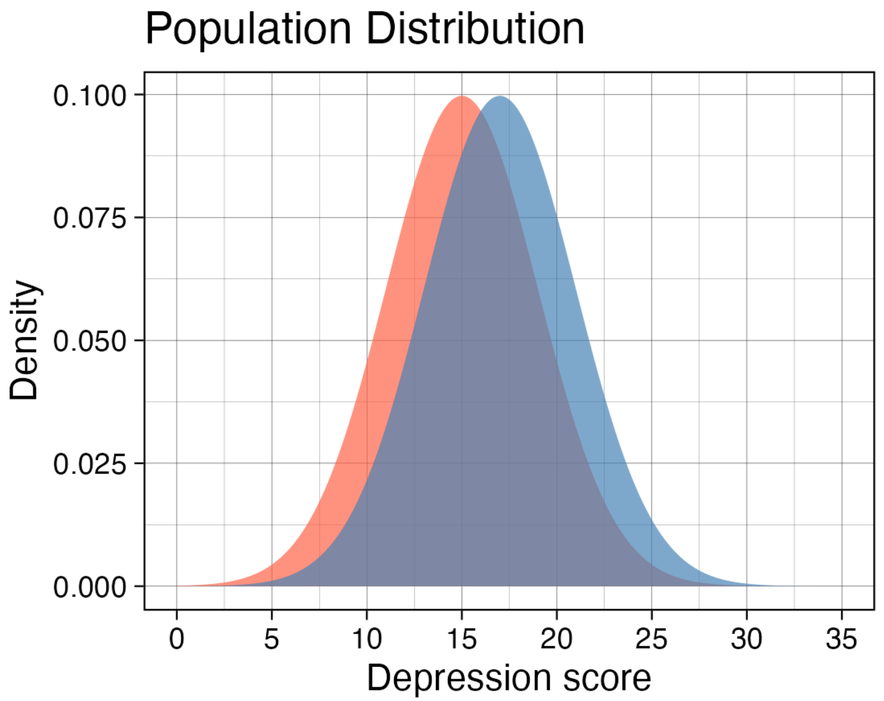


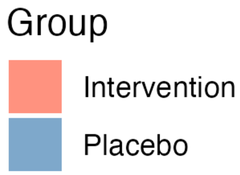
Simulation studies



Simulation studies



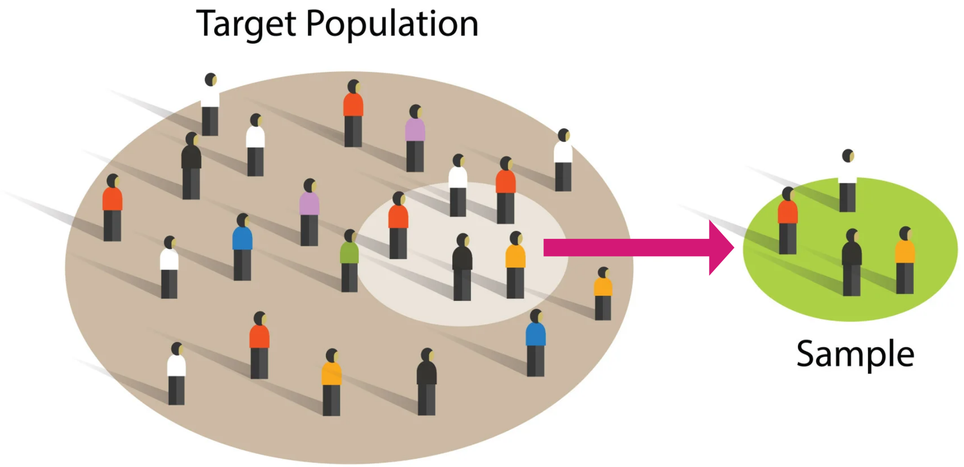


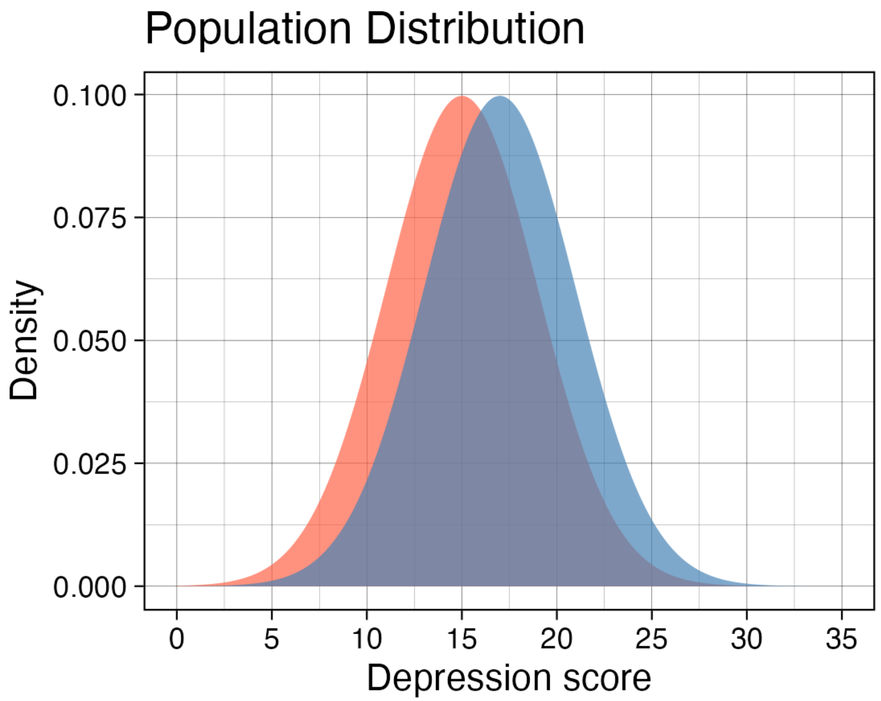


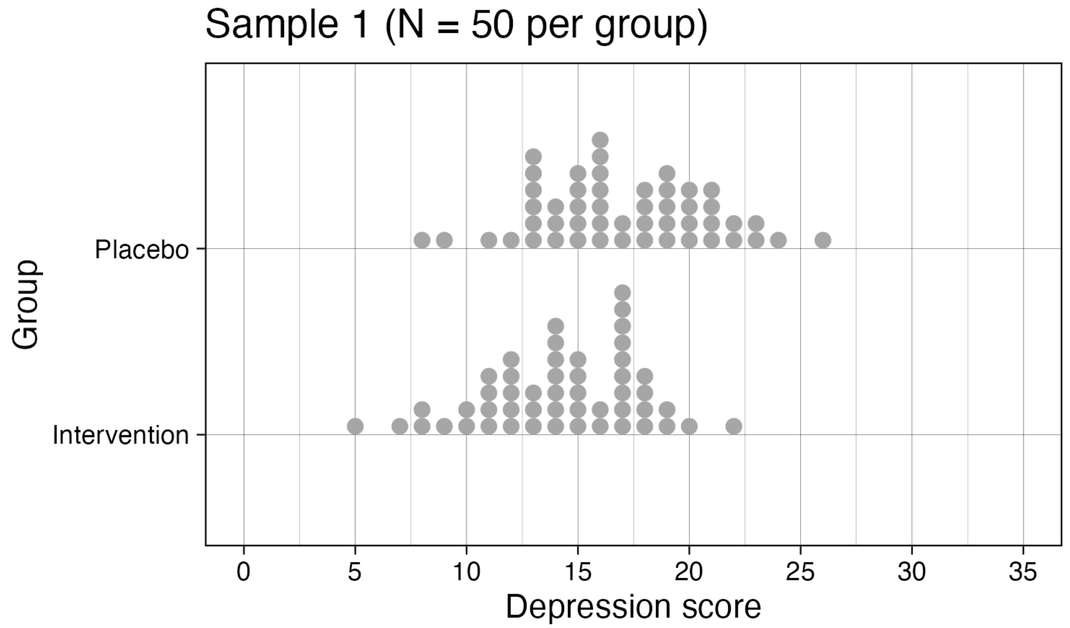
Create the ground truth

Intervention works
Small effect size (SMD = 0.4)

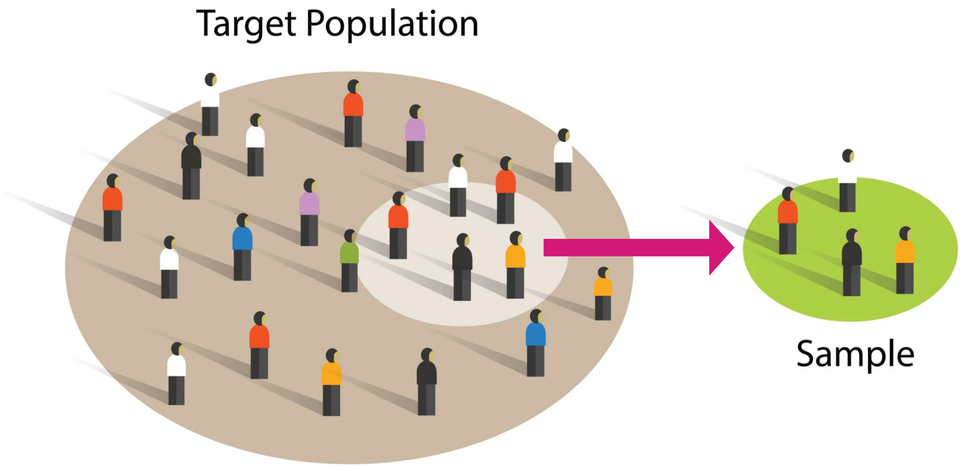
Simulation studies

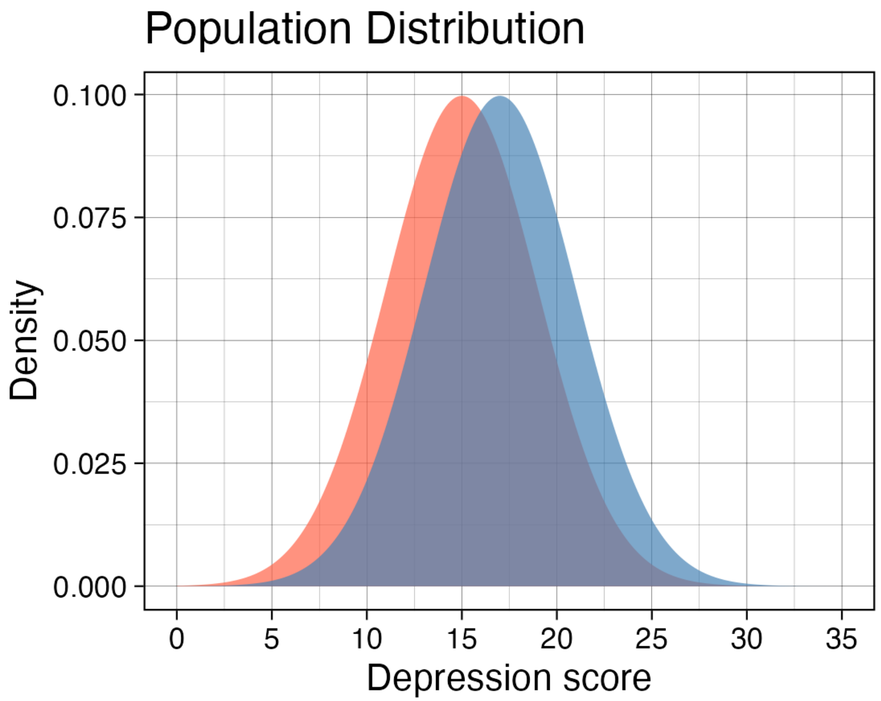


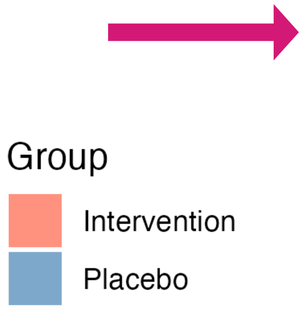


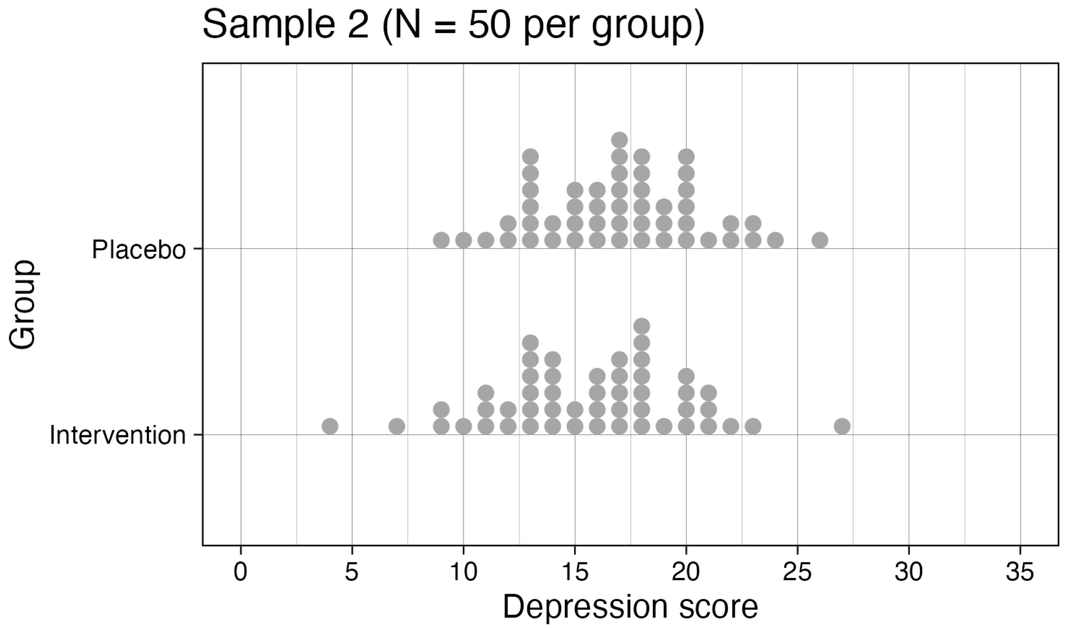


Simulation studies

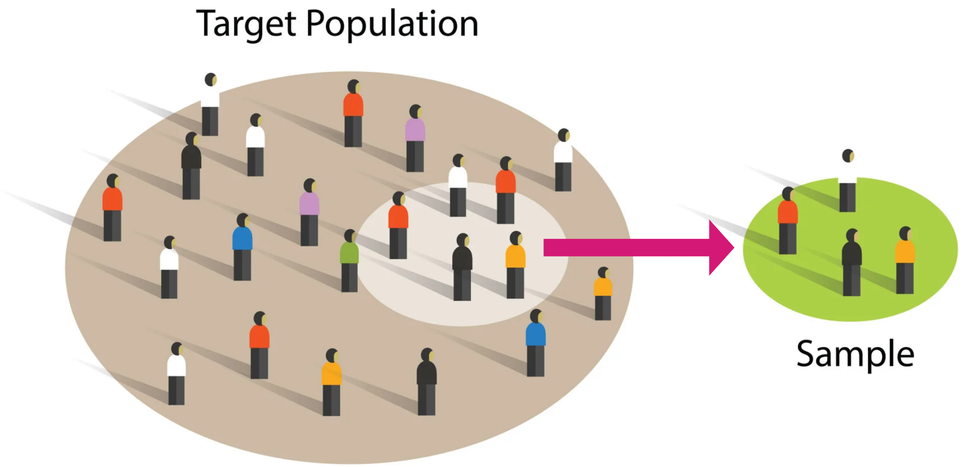


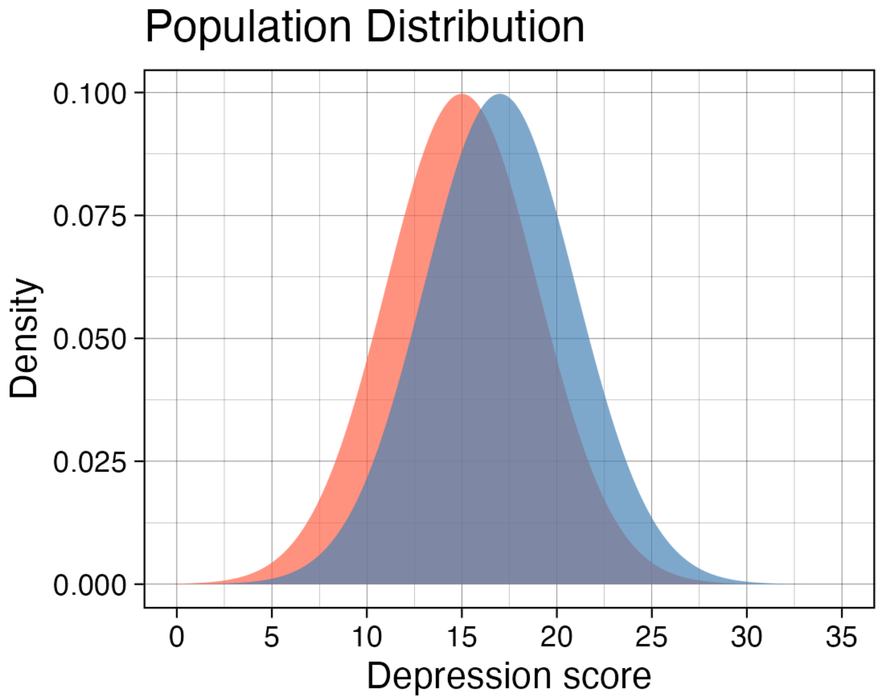


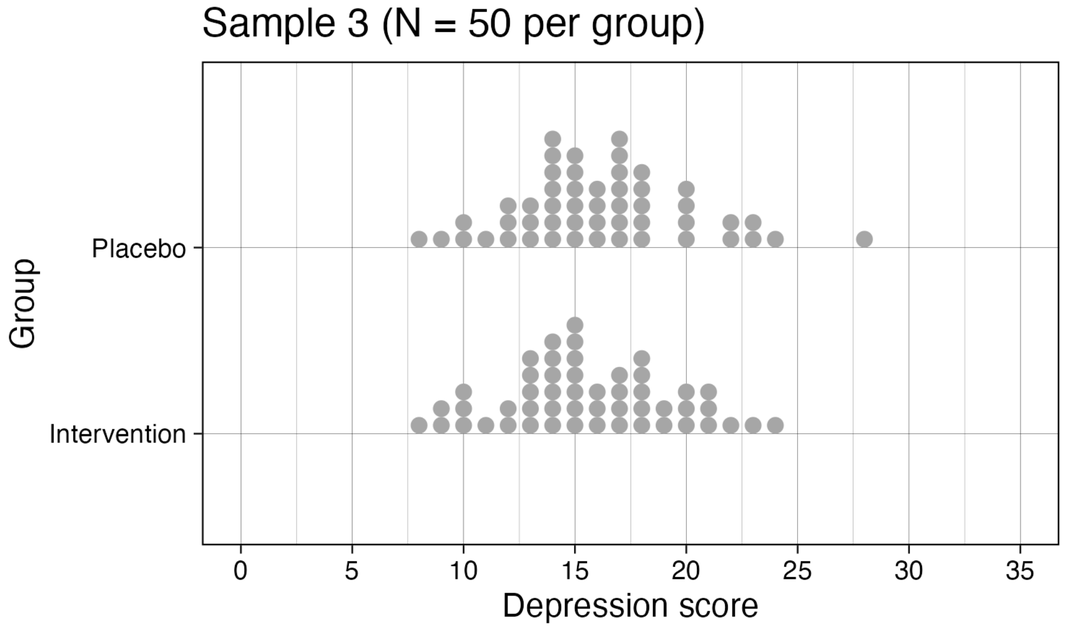




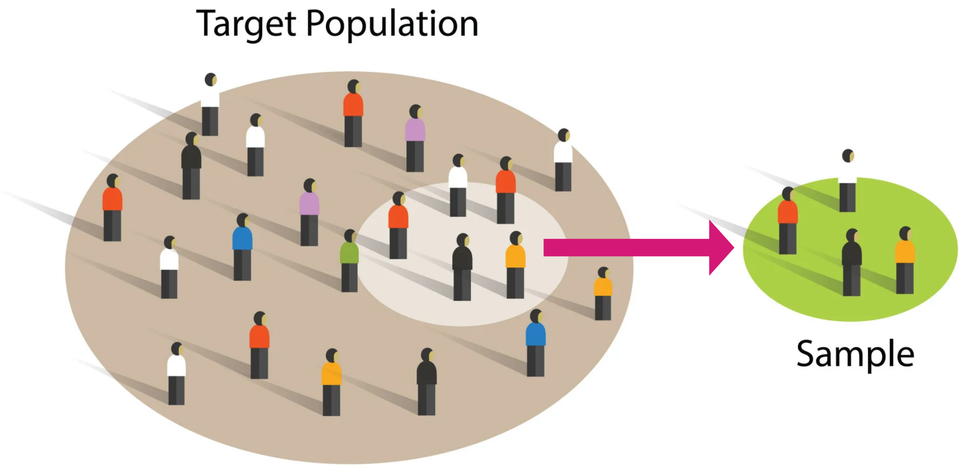
Simulation studies

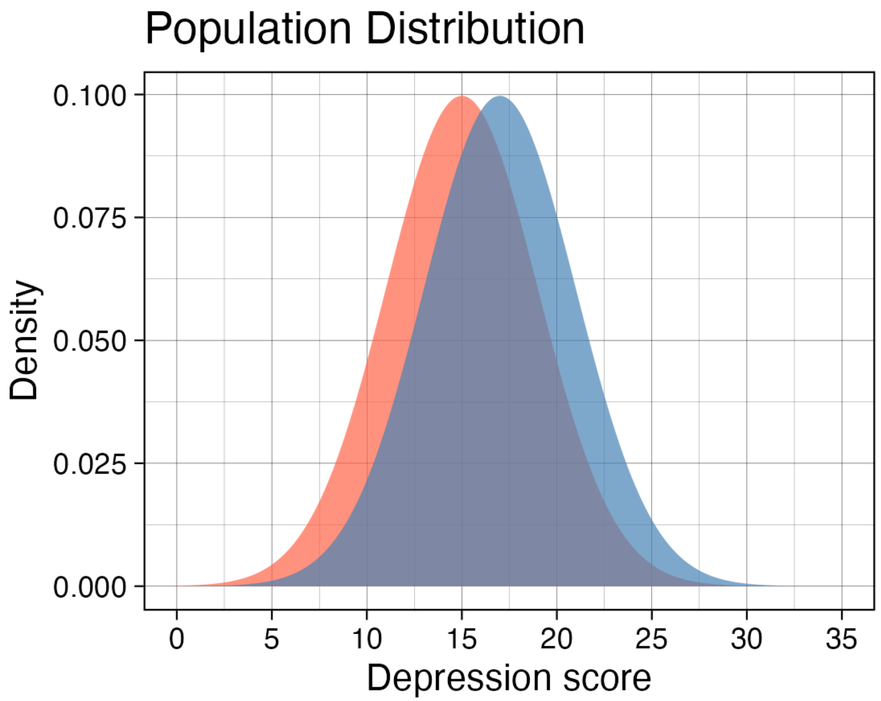


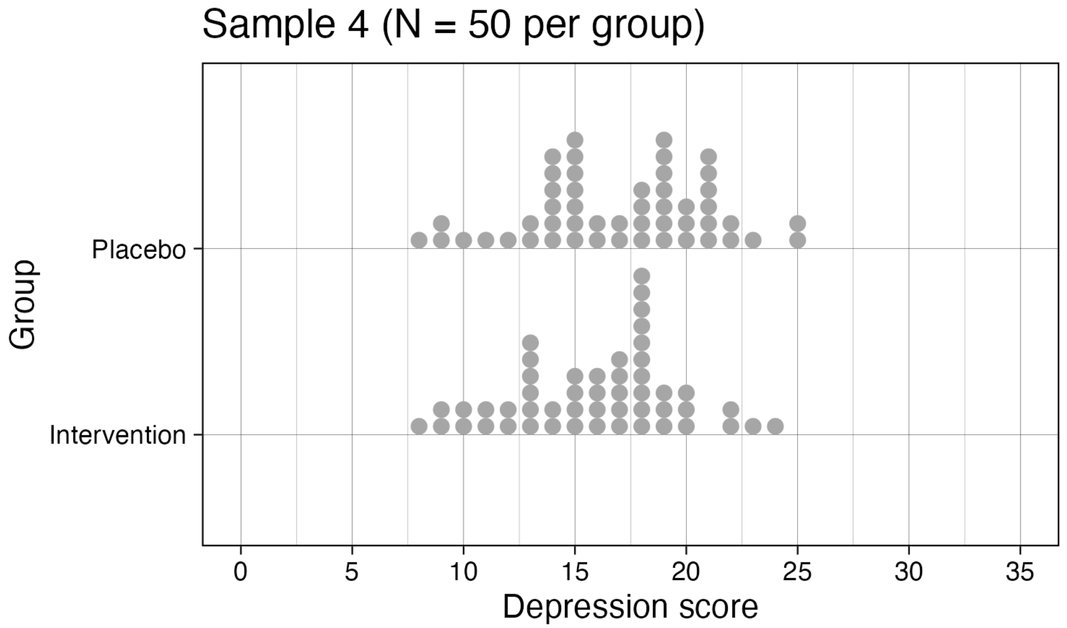




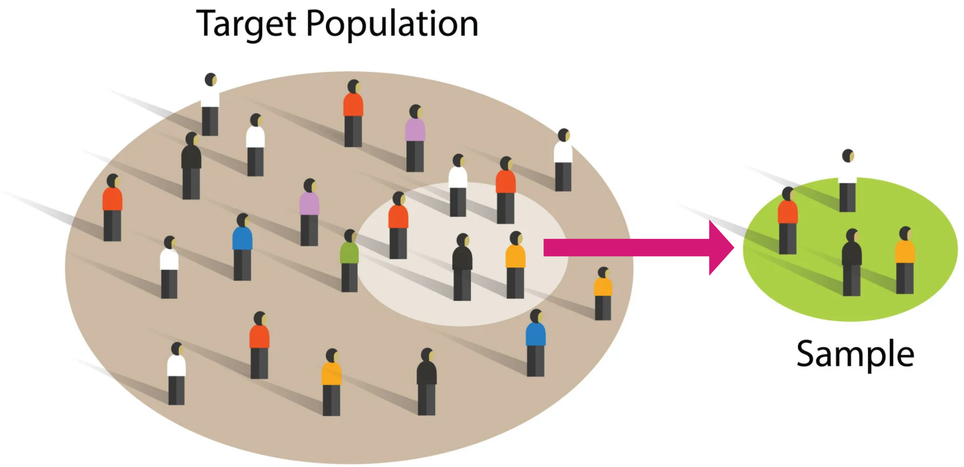
Simulation studies

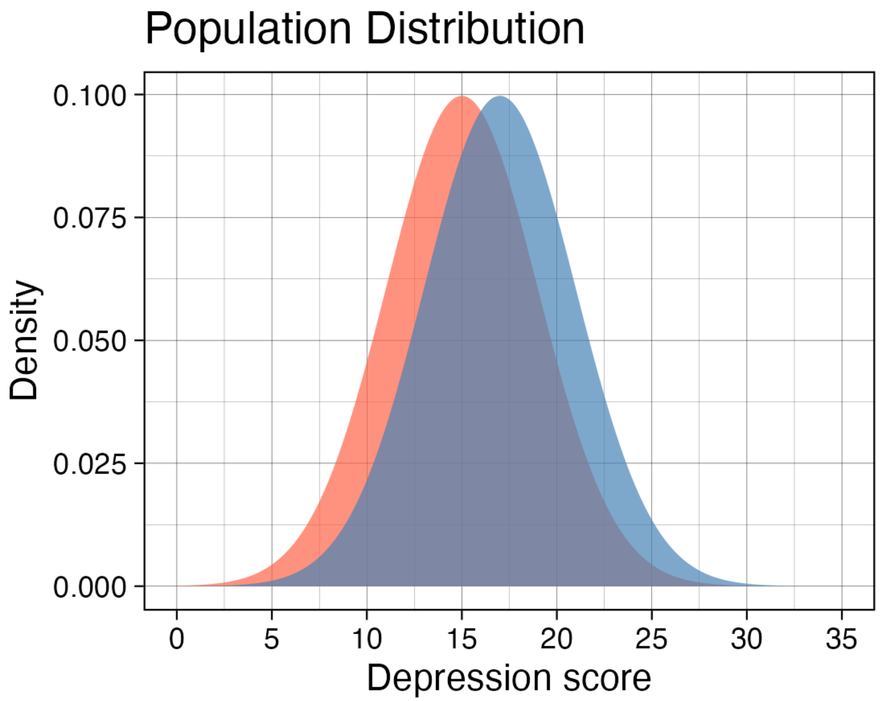


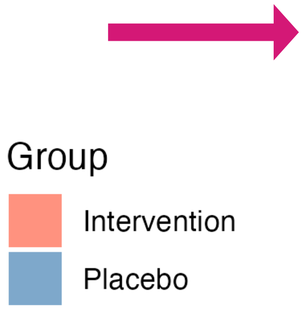


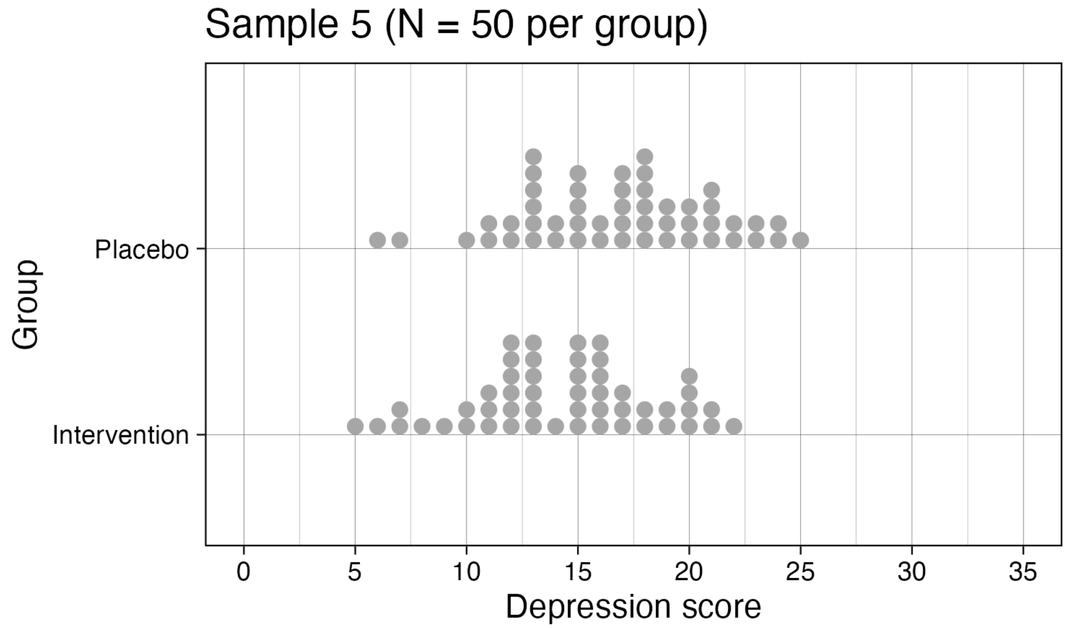


Simulation studies









Core Components

in Monte Carlo simulation studies

Experimental research

1. Experimentally manipulate variable(s)
2. Generate data set (data collection)
3. Analyse data with a statistical method
4. Repeat 1 & 2 several times (‘studies’)
5. Summarize results across studies (‘meta-analysis’)

Monte Carlo simulation

Core Components

in Monte Carlo simulation studies

1. Experimentally manipulate variable(s)
2. Generate pseudo-random data set with known properties
3. Analyse data with a statistical method
4. Repeat 1 & 2 thousands of times (‘iterations’)
5. Summarize results across iterations

Pseudo-Random Number Generators

in Monte Carlo simulation studies

How is simulated data generated?

Technically: Using pseudo-random number generators
- Asymptotically random but reproducible

Practically: Using fairly simple R/Python/etc code

Pseudo-Random Number Generators

in Monte Carlo simulation studies

How is simulated data generated?

Technically: Using pseudo-random number generators
- Asymptotically random but reproducible

Practically: Using fairly simple R/Python/etc code

```{r}
library(janitor)

runif(n = 3, min = 1, max = 10) ▷ round_half_up(digits = 0)

runif(n = 3, min = 1, max = 10) ▷ round_half_up(digits = 0)

runif(n = 3, min = 1, max = 10) ▷ round_half_up(digits = 0)
```

```
[1] 8 7 6
[1] 8 2 7
[1] 7 5 7
```

Pseudo-Random Number Generators

in Monte Carlo simulation studies

How is simulated data generated?

Technically: Using pseudo-random number generators
- Asymptotically random but reproducible

Practically: Using fairly simple R/Python/etc code

```{r}

library(janitor)

set.seed(42)
runif(n = 3, min = 1, max = 10) ▷ round_half_up(digits = 0)

set.seed(42)
runif(n = 3, min = 1, max = 10) ▷ round_half_up(digits = 0)

set.seed(42)
runif(n = 3, min = 1, max = 10) ▷ round_half_up(digits = 0)

```

```
[1] 9 9 4
[1] 9 9 4
[1] 9 9 4
```

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Pseudo-Random Number Generators

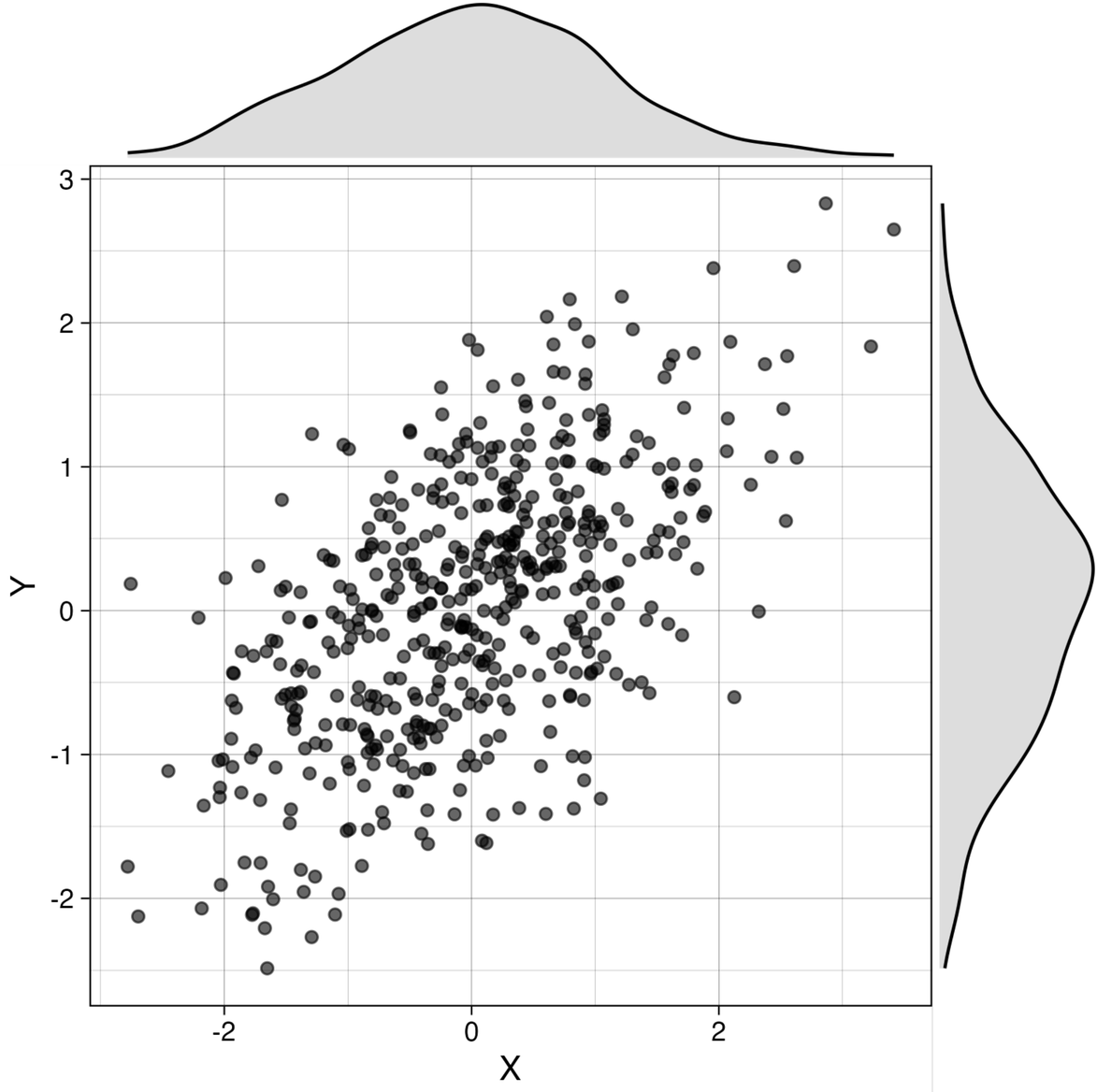
in Monte Carlo simulation studies

How is simulated data generated?

Multivariate normal:

```r
library(faux)

dat <- rnorm_multi(n = 500,
                   vars = 2,
                   varnames = c("X", "Y"),
                   mu = 0,
                   sd = 1,
                   r = 0.6)
```



aut = "Ian Hussey";

dept = "Psycho

Deeper understanding

through Monte Carlo simulation studies

p-values

The probability of observing data at least as extreme given the null hypothesis is true

	Population effect exists (alternative hypothesis)	Population effect does not exist (null hypothesis)
Significant p value ($p < .05$)	True positive	False postiive
Non-significant p value ($p >= .05$)	False negative	True negative

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the null hypothesis?

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the null hypothesis?

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$$\Gamma(n) = (n-1)!$$

$$p = 2 \cdot P(T_{df} > |t|)$$

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the null hypothesis?

```
replicate(n = 100000, # repeat the following 10k times:
          expr = t.test( # run a t-test
            # two groups of 50 simulated participants
            # with a true difference in means of 0
            x = rnorm(n = 50, m = 0, sd = 1),
            y = rnorm(n = 50, m = 0, sd = 1)
          )$p.value) ▷ # extract the p value
  hist() # plot a histogram of the p values
```

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the null hypothesis?

```
replicate(n = 100000, # repeat the following 10K times:
          expr = t.test( # run a t-test
            # two groups of 50 simulated participants
            # with a true difference in means of 0
            x = rnorm(n = 50, m = 0, sd = 1),
            y = rnorm(n = 50, m = 0, sd = 1)
          )$p.value) ▷ # extract the p value
  hist() # plot a histogram of the p values
```

Analyse

Do it many times

Generate data

Summarise across iterations

aut = "Ian Hussey";

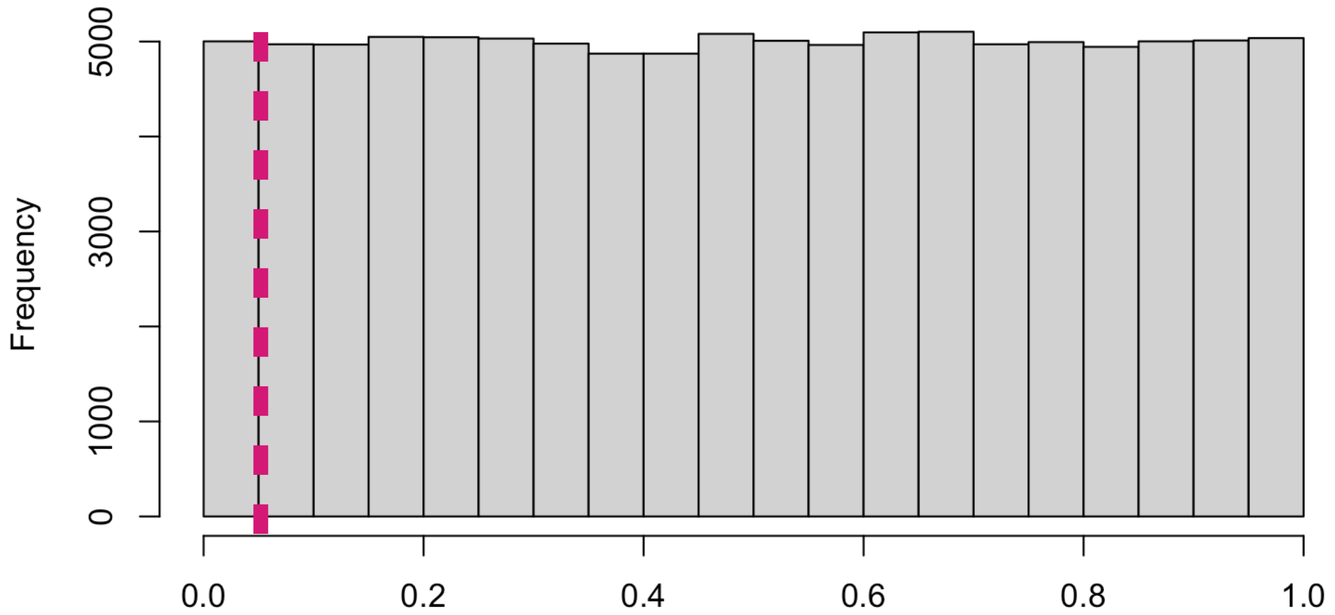
dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the null hypothesis?

```
replicate(n = 100000, # repeat the following 10k times:
          expr = t.test( # run a t-test
            # two groups of 50 simulated participants
            # with a true difference in means of 0
            x = rnorm(n = 50, m = 0, sd = 1),
            y = rnorm(n = 50, m = 0, sd = 1)
          )$p.value) ▷ # extract the p value
  hist() # plot a histogram of the p values
```



aut = "Ian Hussey";

ogy"

Deeper understanding

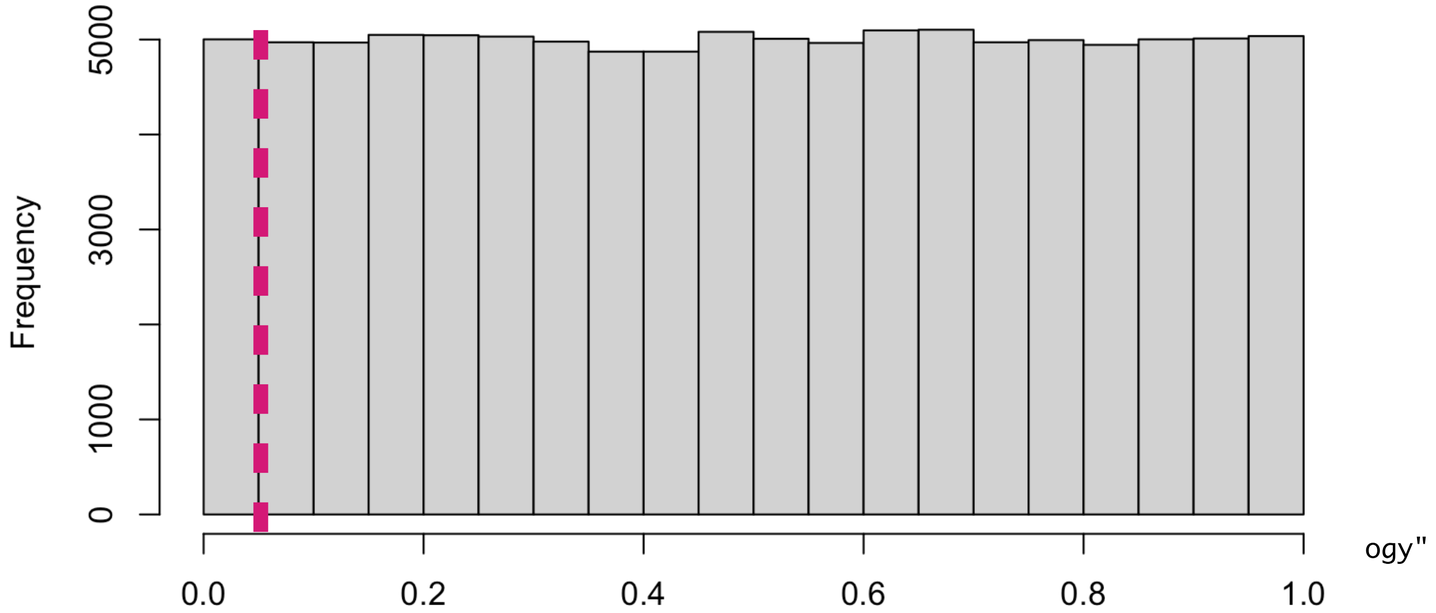
through Monte Carlo simulation studies

What is the distribution of p-values
under the null hypothesis?

5% of p values are below .05

Because all values of p are equally likely
when the null hypothesis is true

False positive rate



aut = "Ian Hussey";

ogy"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the alternative hypothesis?

```
replicate(n = 100000, # repeat the following 10k times:
          expr = t.test( # run a t-test
            # two groups of 50 simulated participants
            # with a true difference in means of 0
            x = rnorm(n = 50, m = 0, sd = 1),
            y = rnorm(n = 50, m = 0.5, sd = 1)
          )$p.value) ▷ # extract the p value
  hist() # plot a histogram of the p values
```

aut = "Ian Hussey";

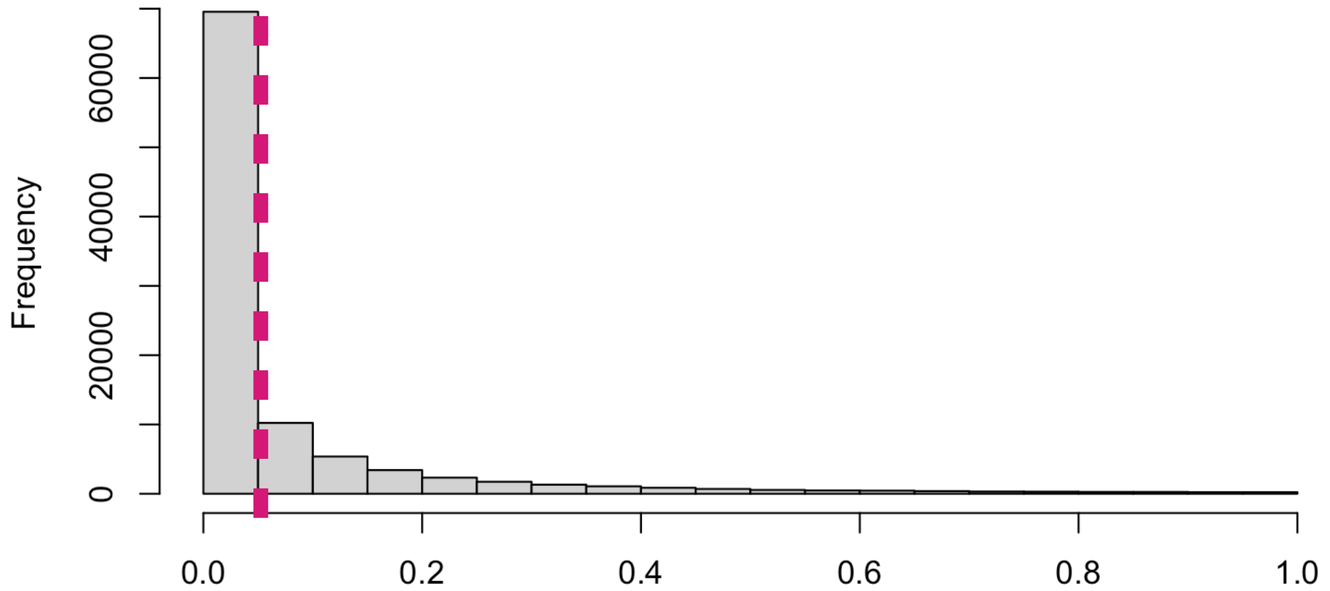
dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the alternative hypothesis?

```
replicate(n = 100000, # repeat the following 10k times:
          expr = t.test( # run a t-test
            # two groups of 50 simulated participants
            # with a true difference in means of 0
            x = rnorm(n = 50, m = 0, sd = 1),
            y = rnorm(n = 50, m = 0.5, sd = 1)
          )$p.value) ▷ # extract the p value
  hist() # plot a histogram of the p values
```



aut = "Ian Hussey";

ogy"

Deeper understanding

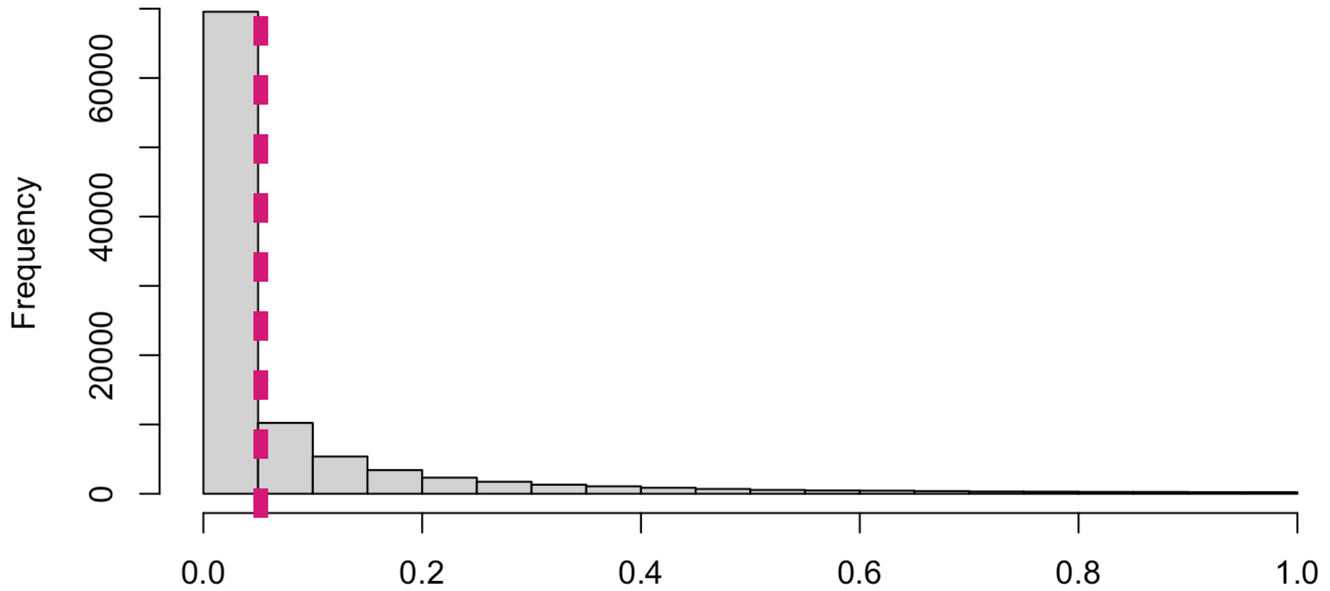
through Monte Carlo simulation studies

What is the distribution of p-values
under the alternative hypothesis?

['statistical power']% of p values are below .05

Because all values of p are equally likely
when the alternative hypothesis is true

True positive rate



aut = "Ian Hussey";

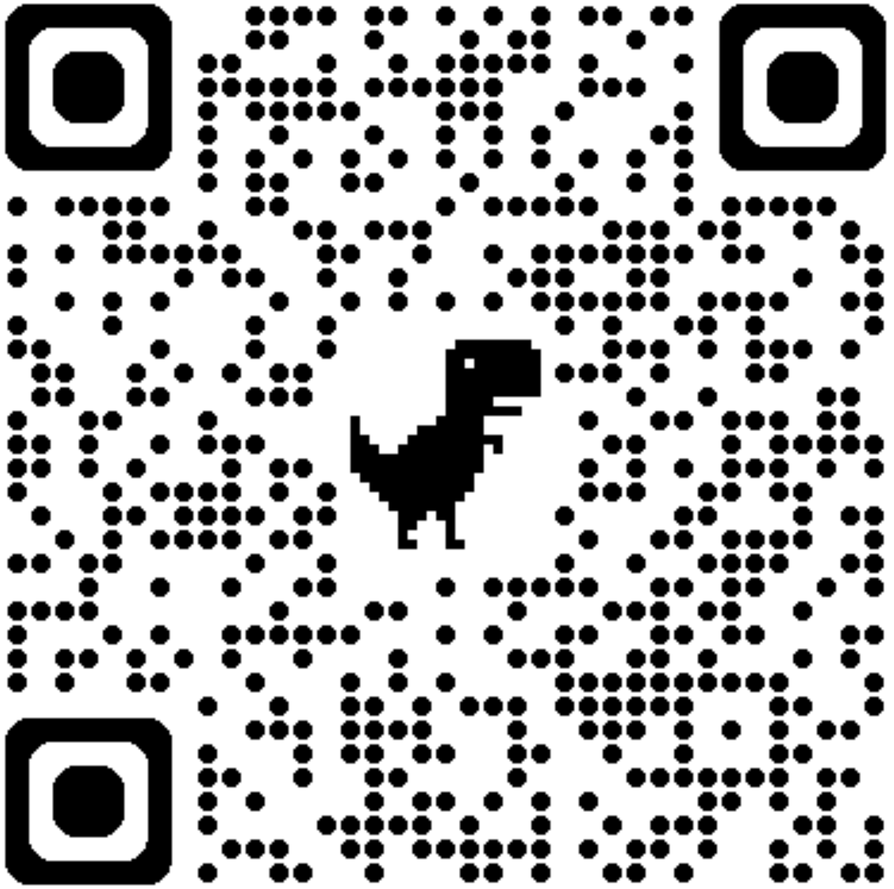
ogy"

Deeper understanding

through Monte Carlo simulation studies

What is the distribution of p-values under the alternative hypothesis?

https://rpsychologist.com/d3/pdist/



Deeper understanding

through Monte Carlo simulation studies

p-values

The probability of observing data at least as extreme given the null hypothesis is true

	Population effect exists (alternative hypothesis)	Population effect does not exist (null hypothesis)
Significant p value ($p < .05$)	True positive [statistical power]	False postiive
Non-significant p value ($p >= .05$)	False negative [should equal alpha]	True negative

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Examples

of Monte Carlo simulation studies

Example 1:

Violation of statistical assumptions is ‘bad’

In what way?

How bad?

Violate assumptions on purpose in simulated data

e.g., normality, homoscedacity

Observe impact on test’s properties

Examples

of Monte Carlo simulation studies

Example 1:

Textbooks tell us to:

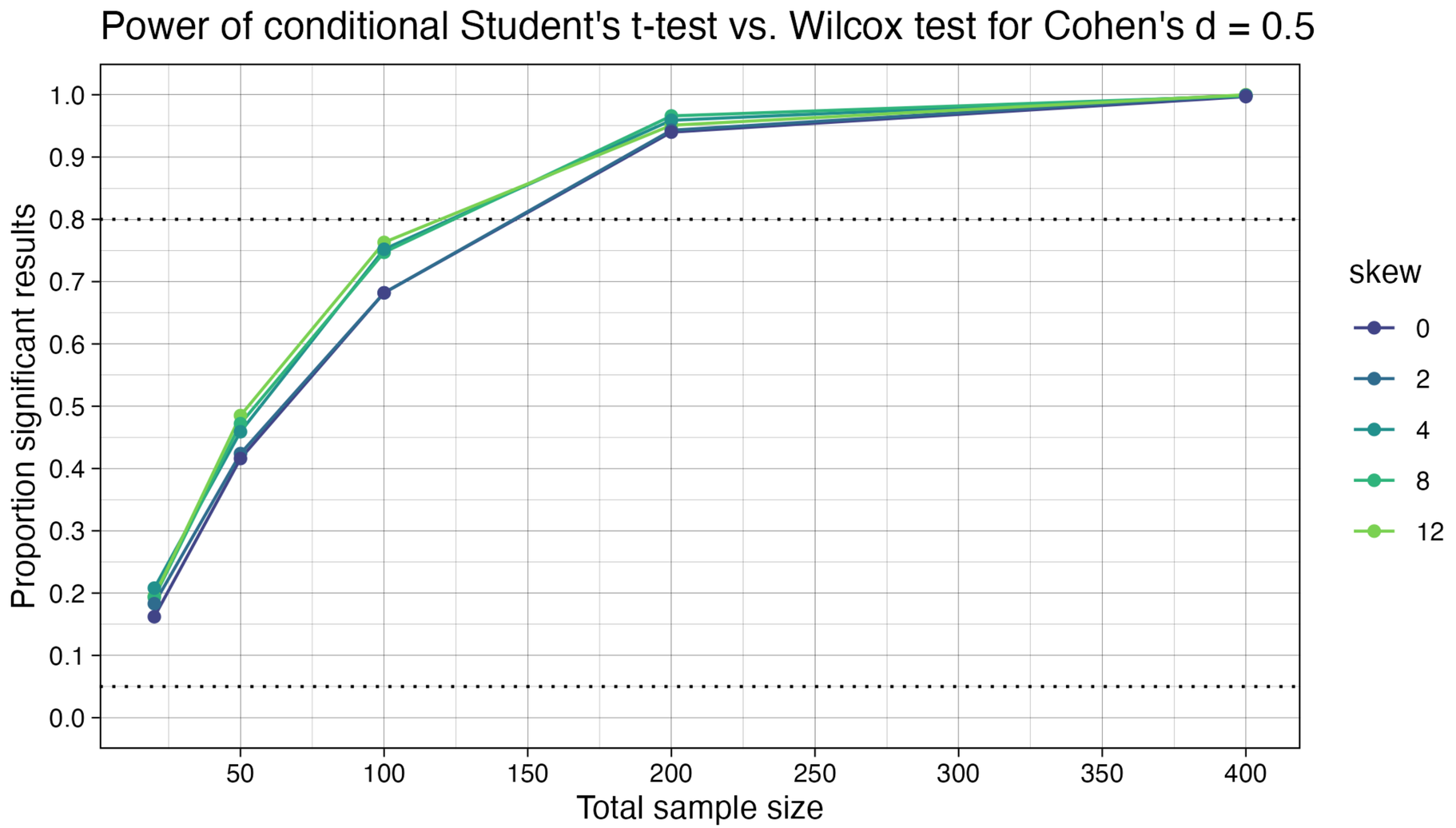
1. Test statistical assumption
2. Run a parametric test if assumptions met
3. Run a non-parametric test if assumptions violated

Why not just run the non-parametric test by default?

Examples

of Monte Carlo simulation studies

Example 1:



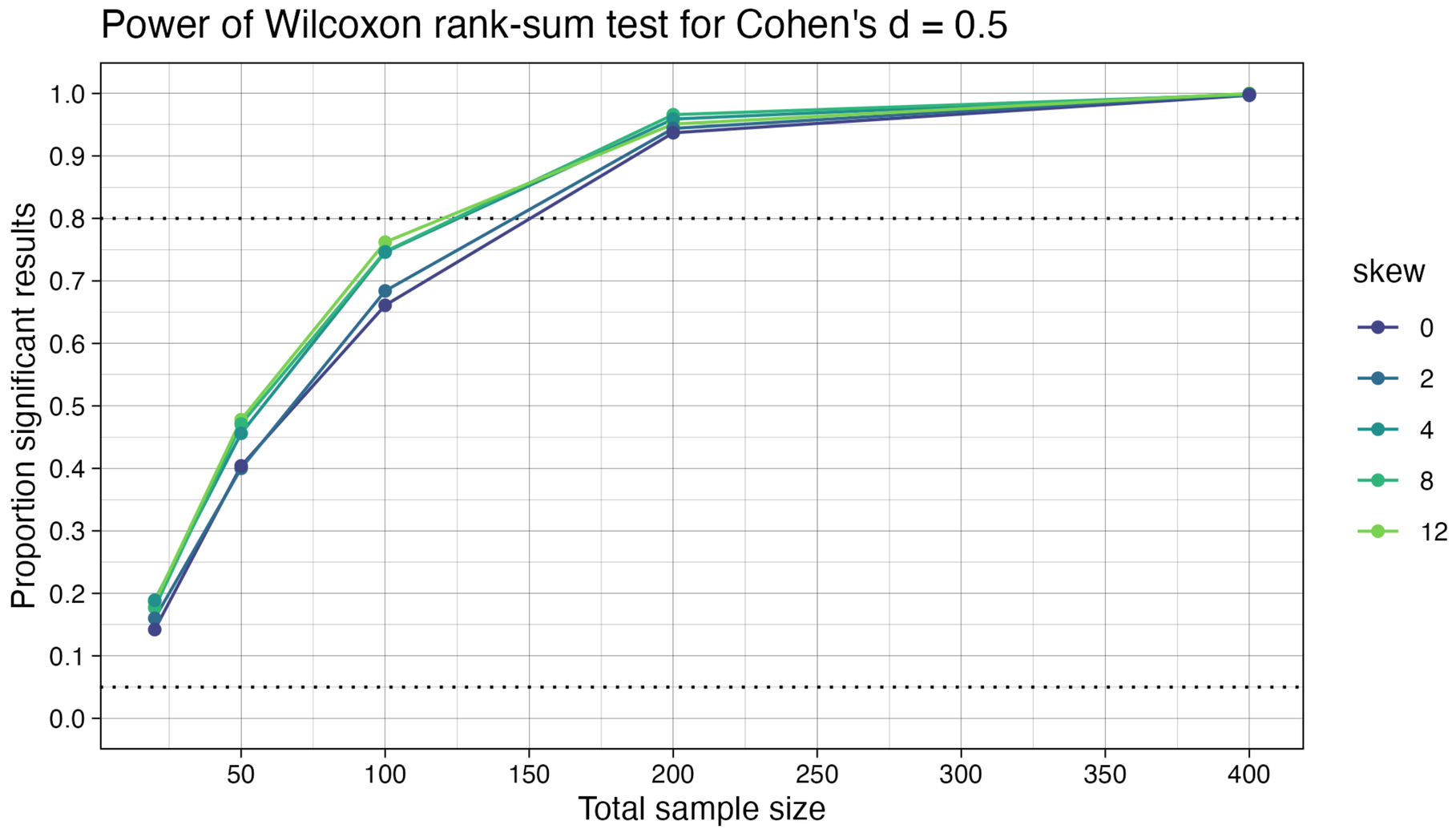
aut = "Ian Hussey";

Examples

of Monte Carlo simulation studies

aut = "Ian Hussey";

Example 1:



Example 2:

p-hacking is ‘bad’

In what way?

How bad?

Model common p hacking practices using simulated data

Observe impact on test’s properties

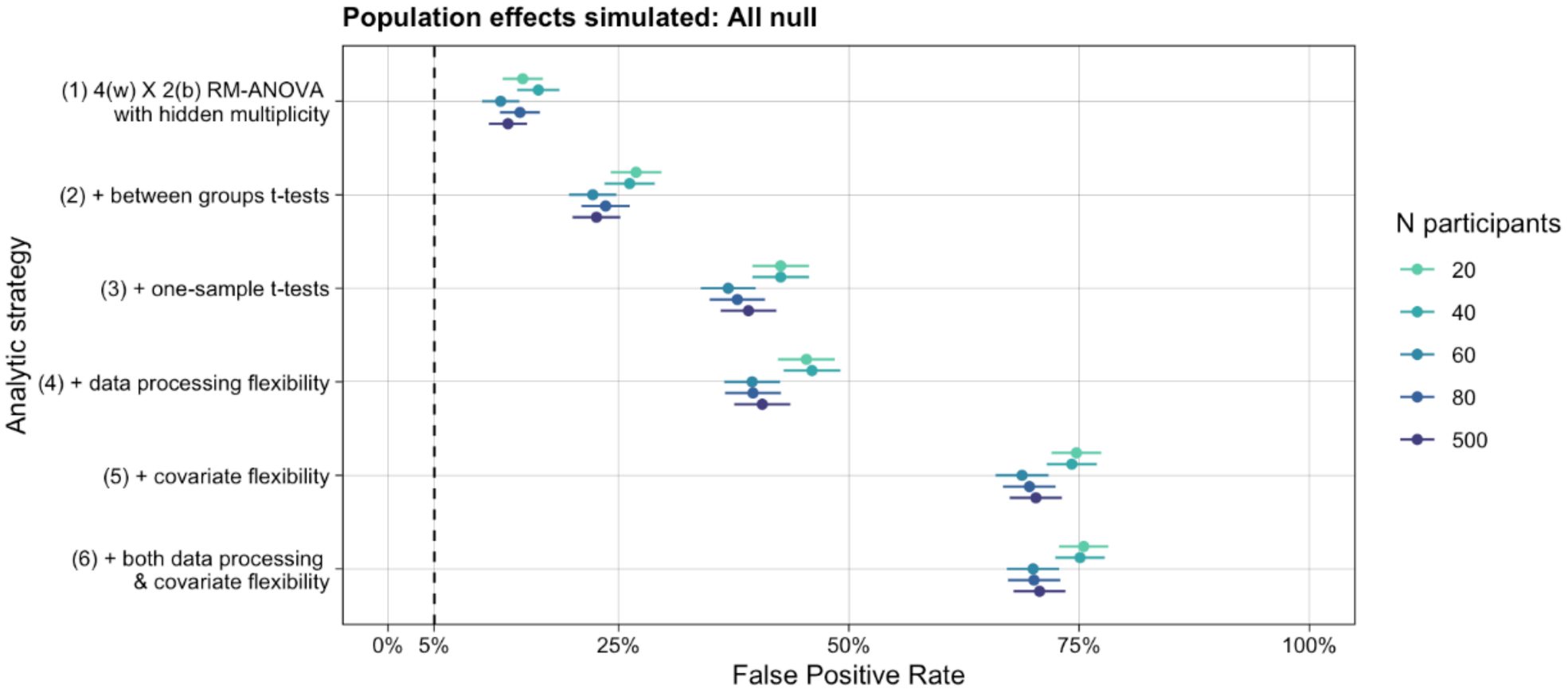
Eg false positive rate

Examples

of Monte Carlo simulation studies

aut = "Ian Hussey";

Example 2:



Example 3:

Systematic noise in your data is ‘bad’

In what way?

How bad?

Model common sources of noise
using simulated data
(e.g., careless responding on questionnaires)

Observe impact on test’s properties
Eg correlations

Examples

of Monte Carlo simulation studies

Example 3:

Systematic noise in your data is ‘bad’

[shiny app]

Examples

of Monte Carlo simulation studies

Example 4:

Expose and understand statistical paradoxes and artefacts

+

Causal modelling

When is a more effective treatment less effective?

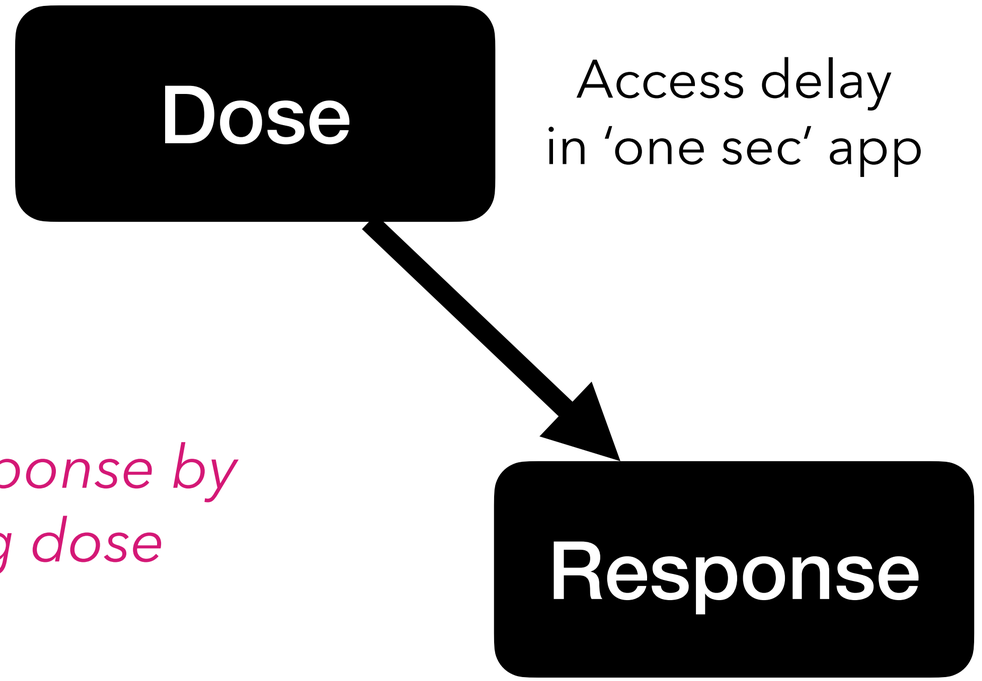
aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Examples

of Monte Carlo simulation studies

When is a more effective treatment less effective?



aut = "Ian Hussey";

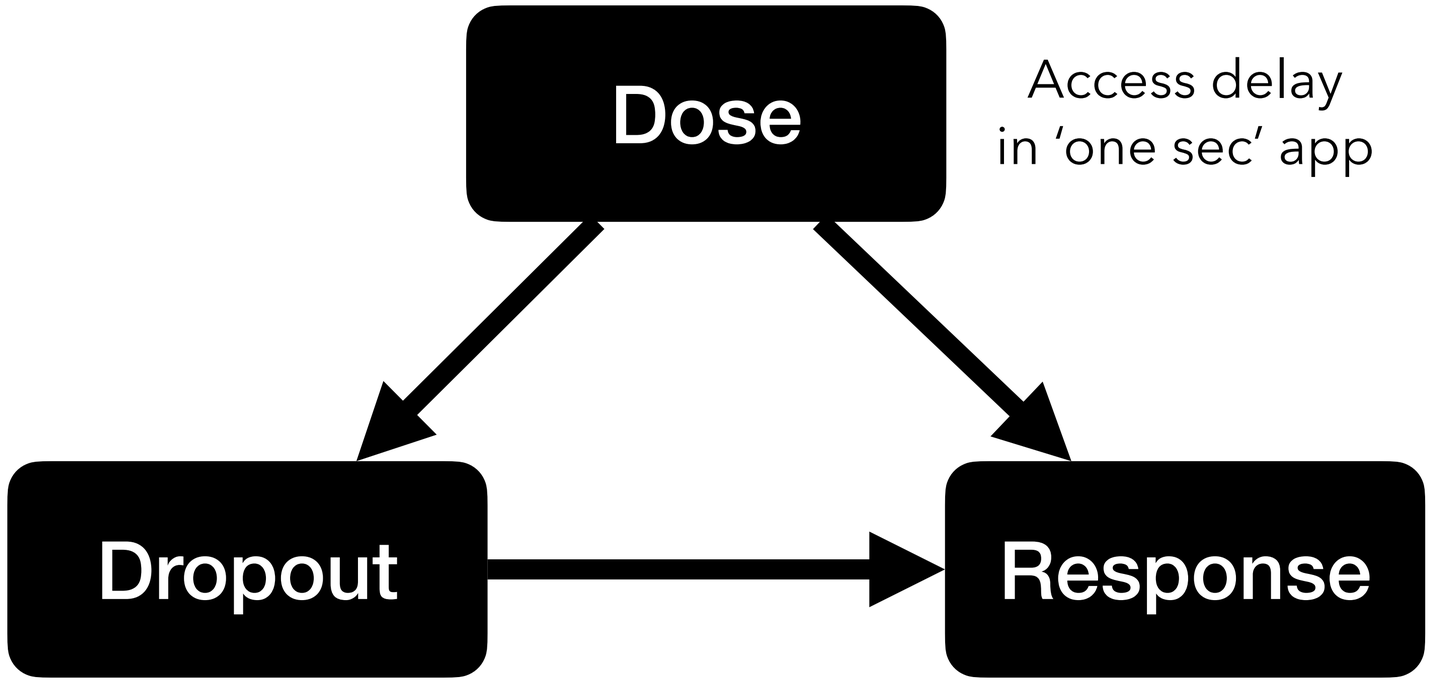
dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Examples

of Monte Carlo simulation studies

When is a more effective treatment less effective?

When there are confounding variables



Access delay
in ‘one sec’ app

“Gopfertammi, this is annoying..”
[deletes app]

Reduced social media usage

aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Examples

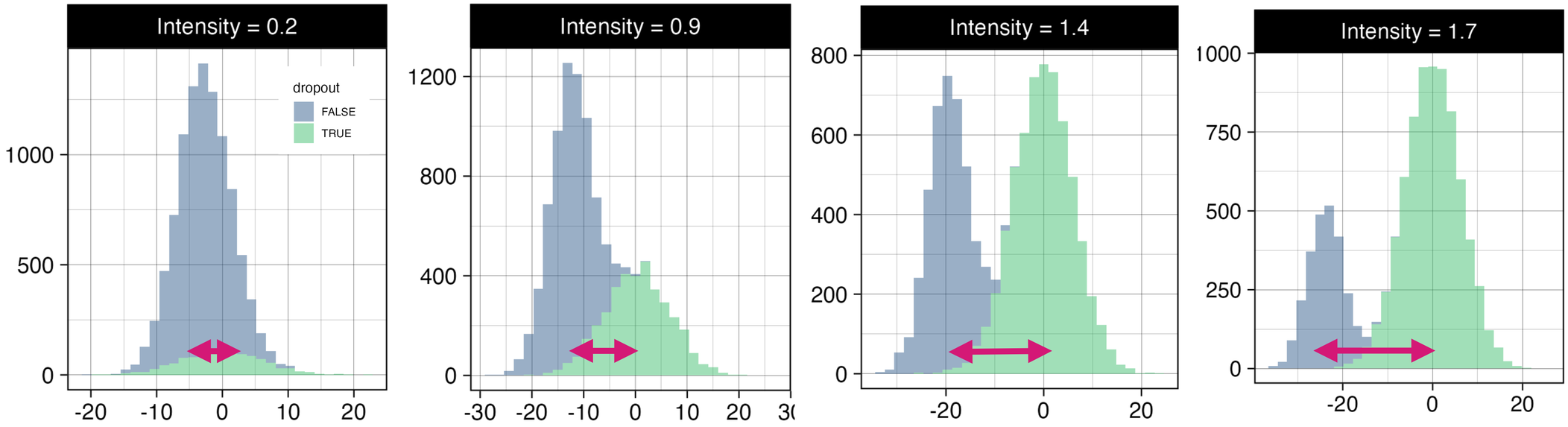
of Monte Carlo simulation studies

When is a more effective treatment less effective?

When there are confounding variables

Should we try and help most people a little bit,

Or a small proportion of people a lot?



aut = "Ian Hussey";

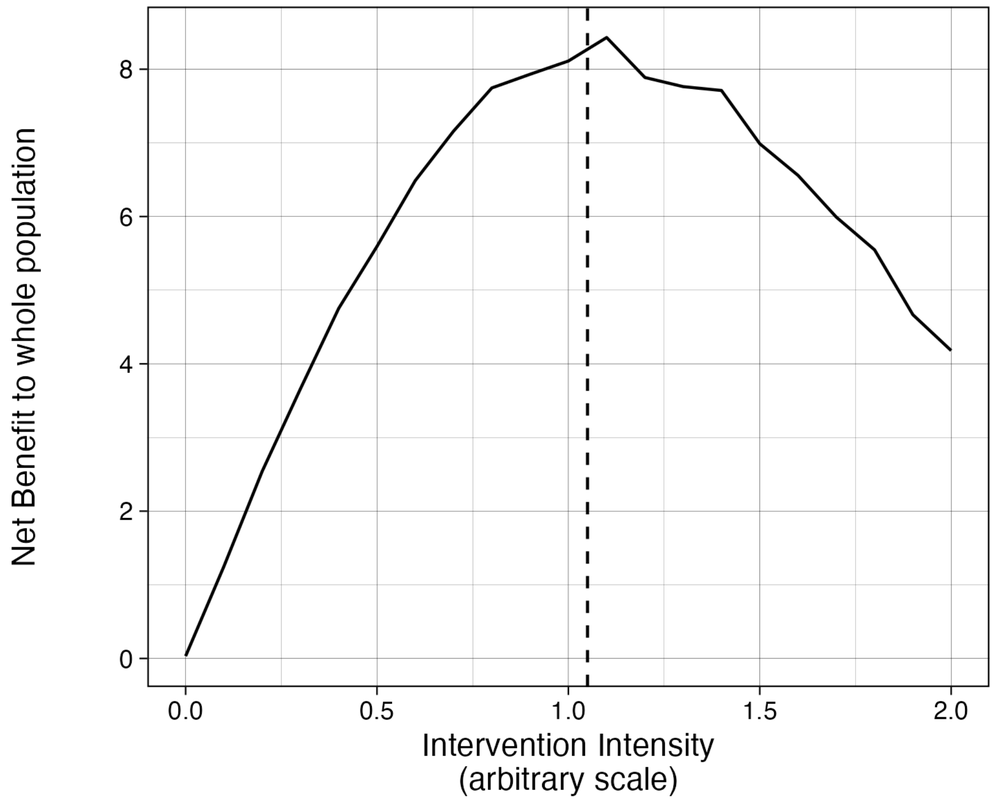
dept = "Psychology of Digitalisation || Digitalisation of Psychology"

Examples

of Monte Carlo simulation studies

When is a more effective treatment less effective?

When there are confounding variables



aut = "Ian Hussey";

dept = "Psychology of Digitalisation || Digitalisation of Psychology"


WARNING
LABEL SIMULATED DATA
VERY CLEARLY
SO IT IS NOT CONFUSED
WITH REAL DATA

Books & Articles

- Hussey (2024) Improving your statistical inferences using Monte Carlo simulation studies https://github.com/ianhussey/simulation-course
- Miratrix & Pustejovsky (2025) Designing Monte Carlo Simulations in R https://jepusto.github.io/Designing-Simulations-in-R/
- Siepe et al. (2024) Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting
- Smaldino (2023) Modeling Social Behavior: Mathematical and Agent-Based Models of Social Dynamics and Cultural Evolution

R packages

Data simulation

- faux
  - Factorial designs
  - Multilevel models
  - Simple correlational data
- lavaan::simulateData()
  - Cross sectional data
  - Causal models

Monte Carlo simulation studies

- purrr
- SimDesign

Vielen dank für Ihre Aufmerksamkeit