ADS2 Week11- Power and Sample Size

2019-11-25

Wanlu Liu

wanluliu@intl.zju.edu.cn

Lab website: labw.org

Learning Objectives

- Understand intuition behind power calculations
- Know how to perform power/sample size analysis with formula or in R
- Reveal the relationship among significance level, power, effect size and sample size
- Demonstrate different stages in clinical trails and stopping rules

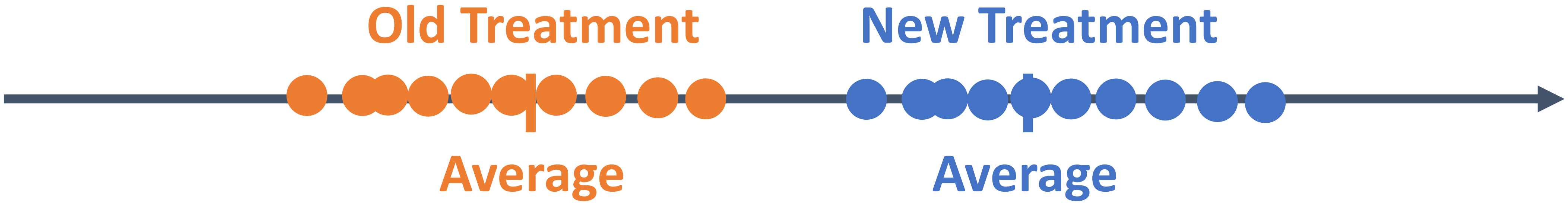
Why care about sample size and power

- Power = **probability** of getting a statistically significant result, when in fact there is a ‘clinically’ meaningful difference (unknown to us)
- By definition, studies with low power are less likely to produce statistically significant results, even when a clinically meaningful effect does exist
- Lack of statistical significance does not prove that there is no treatment effect, but instead may be a consequence of small sample size (i.e. low power)
- Therefore, it is **important** to have **enough power** and an adequate sample size

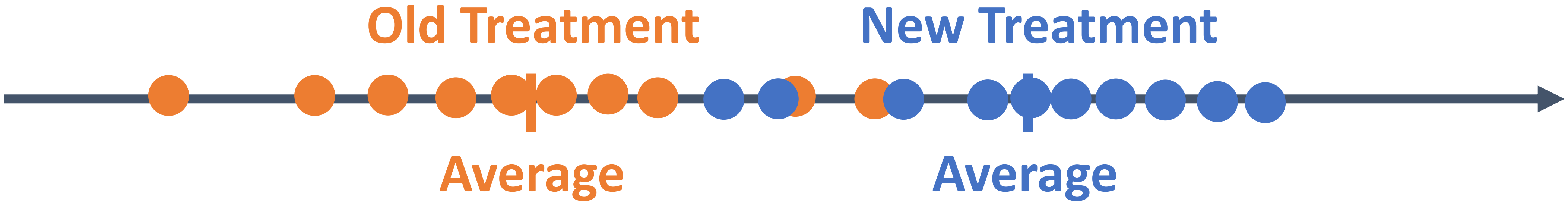
Question

Without variability__________ , there is no need for Statistics

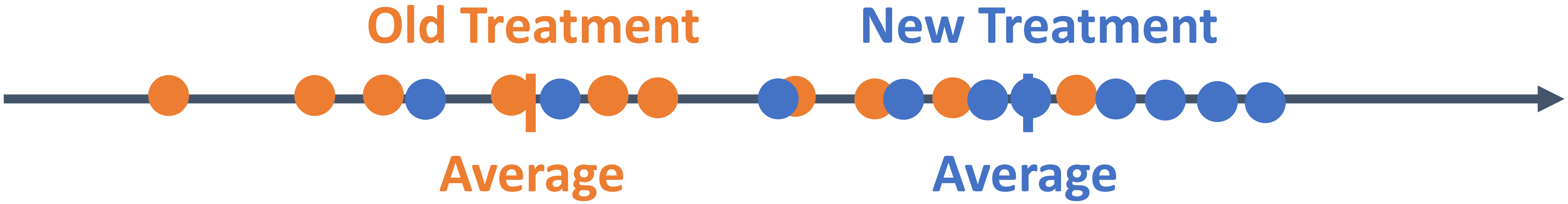
Variability



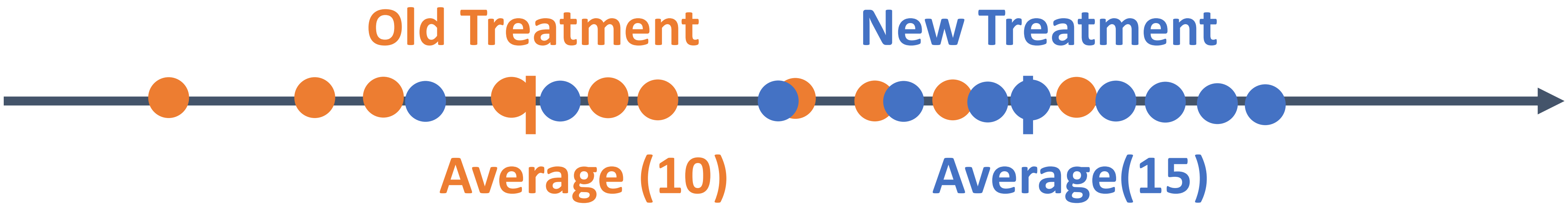
Variability



Variability



Sample size & Effect size



- **What’s the sample size for old/new treatment?**
  - n(old)=10, n(new)=10
- **What’s the effect size?**
  - Effect size= the difference to be detected $(\boldsymbol{\delta})$
  - In this case δ = 15 – 10 = 5
- **Variation** $(\boldsymbol{\sigma^2})$
  - Sample size ↑, variation(σ^2) ↓

Review – Type I and Type II error

	H_0 is True	H_0 is false
Reject H_0	Error Type I (False Positive)	Correct decision (True positive)
Do not reject H_0	Correct decision (True negative)	Error Type II (False Negative)

Type I error

- A Type I Error is rejecting the null hypothesis when it is true.

Prob(Type I error) = Significance level α = $P(reject\ H_0|H_0\ true)$

Type II error

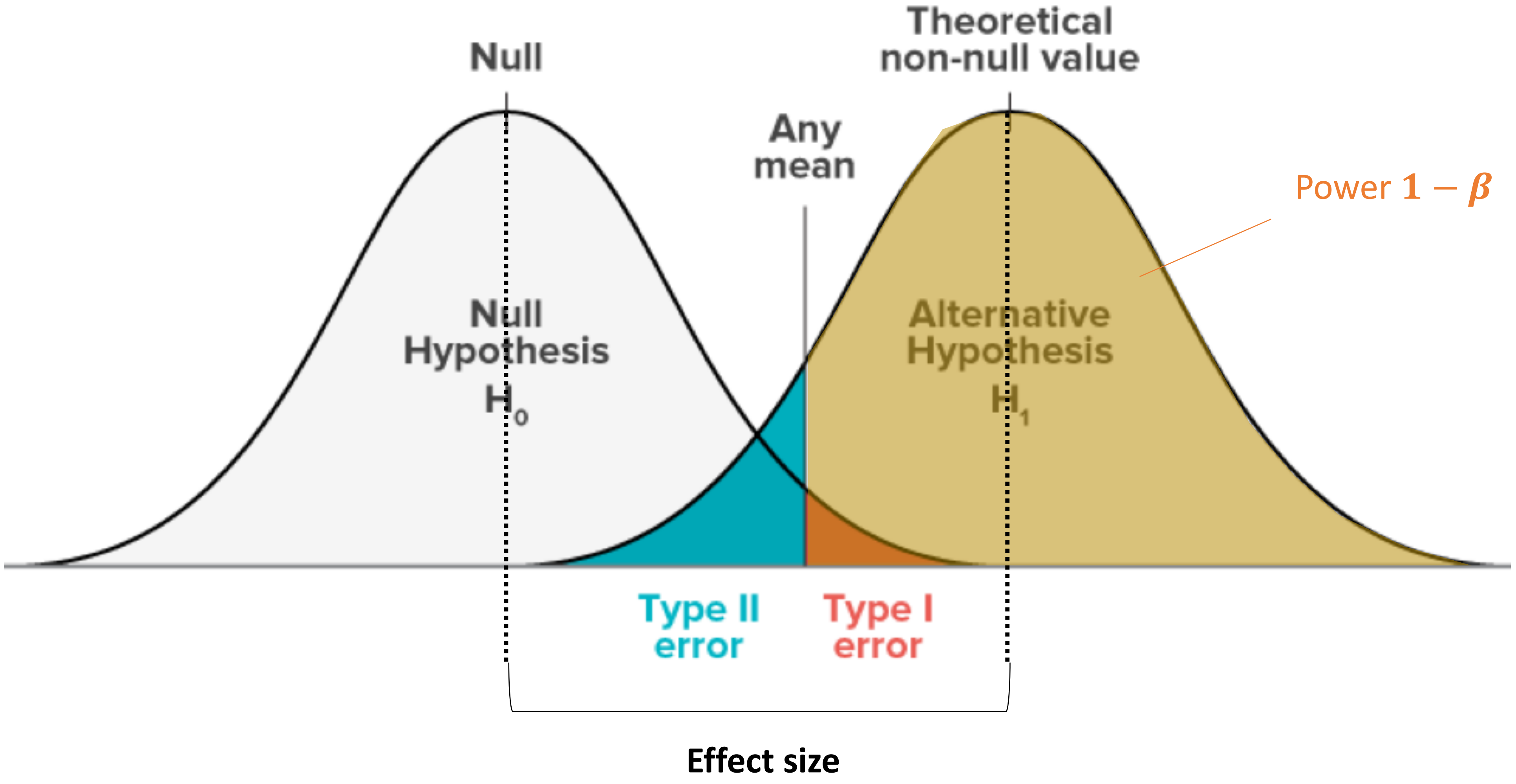
- A Type II Error is not rejecting a null hypothesis when it is false.
- **Prob(Type II error) = β = $P(accept\ H_0|H_1\ true)$**
- value of β typically depends on which particular alternative hypothesis is true.

Power of a hypothesis test

- **Power =1 - $\beta = P(reject\ H_0|H_1\ true)$**
- Probability of rejecting the null hypothesis if the alternative hypothesis is true.

$\alpha, \beta, 1-\beta$

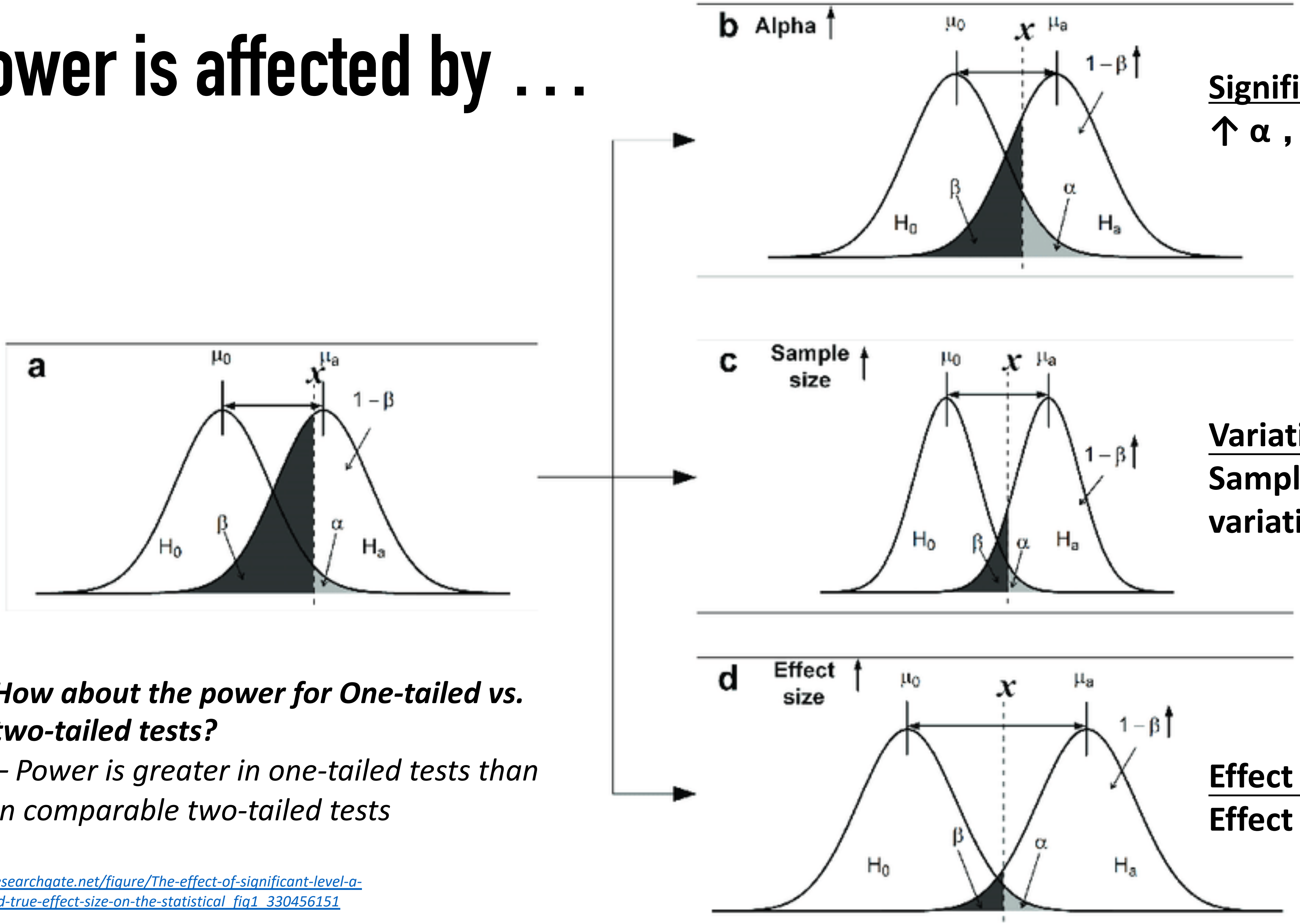
What is **power** in this figure?



Power $1-\beta$

https://www.abtasty.com/blog/type-1-and-type-2-errors/

Power is affected by . . .



Significance level (α)

↑ α ,   power

Variation/sample size

Sample size ↑ ,
variation   , power

Effect size

Effect size ↑ ,   power

How about the power for One-tailed vs. two-tailed tests?

– Power is greater in one-tailed tests than in comparable two-tailed tests

https://www.researchgate.net/figure/The-effect-of-significant-level-a-sample-size-and-true-effect-size-on-the-statistical_fig1_330456151

Power should be … ?

- Phase III: industry minimum = 80%
- Some say Type I error = Type II
- Many large “definitive” studies have power around 99.9%
- Omics studies: aim for high power because Type II error like a bear!

Power Formula

- Depends on study design
- Not hard, but can be VERY algebra intensive
- May want to use a computer program or statistician

Sample size – distribution of response

- **Nominal/binary (Binomial)**
  - Dead, alive
  - N is a function of probability of response in control and probability of response in treated animals
- **Ordinal (Non-parameteric)**
  - Inflammation (mild, moderate, severe)
- **Continuous (Normal)**
  - Blood pressure
  - N is a function of difference in means and standard deviation

Comparing Two Means

- The formula for the total sample size required to compare two population means (under normal distribution, two.sided), μ_0 and μ_1 with common variance, σ^2 is:

$$2n = \frac{4\left(z_{1-\alpha/_2} + z_{1-\beta}\right)^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}$$

Effect size

Control sample size +
treatment sample size

Standard deviation

Calculate n – Example

A clinical trail is planned to compare cognitive behavioral therapy (CBT) and a drug therapy for the treatment of depression. The primary outcome measure is the HoNOS scale, which is a measure of impairment due to psychological distress. From published data, the within group standard deviation of HoNOS is estimated to be 5.7 units.

Calculate the sample size required for each treatment to detect a treatment effect of 2 units on the HoNOS scale with 80% power and a two group t-test with a 0.05 two sided significance level.

$\alpha =$

$\beta =$

$z_{1-\alpha/2} =$

$z_{1-\beta} =$

$\sigma =$

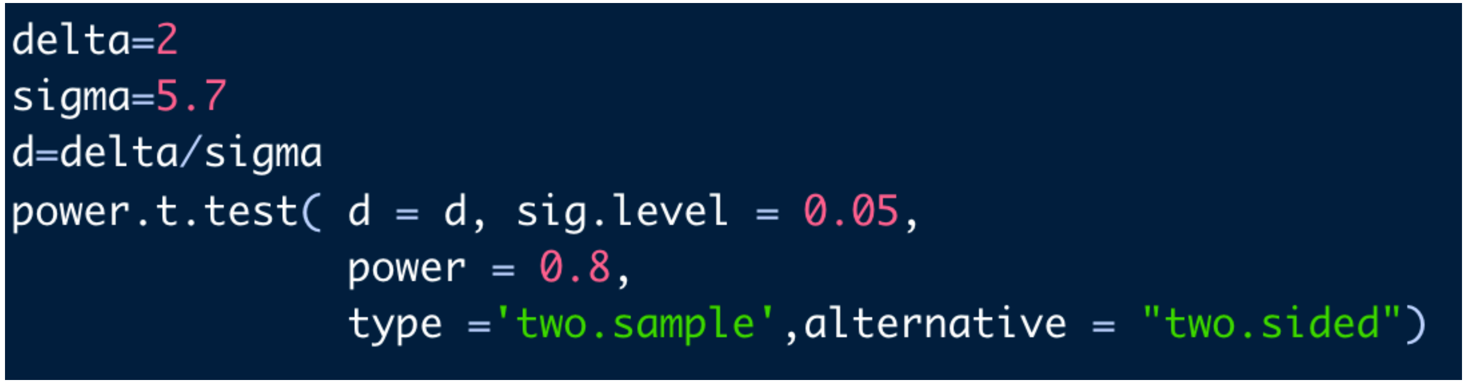
$\mu_0 - \mu_1 =$

So sample size per sample is

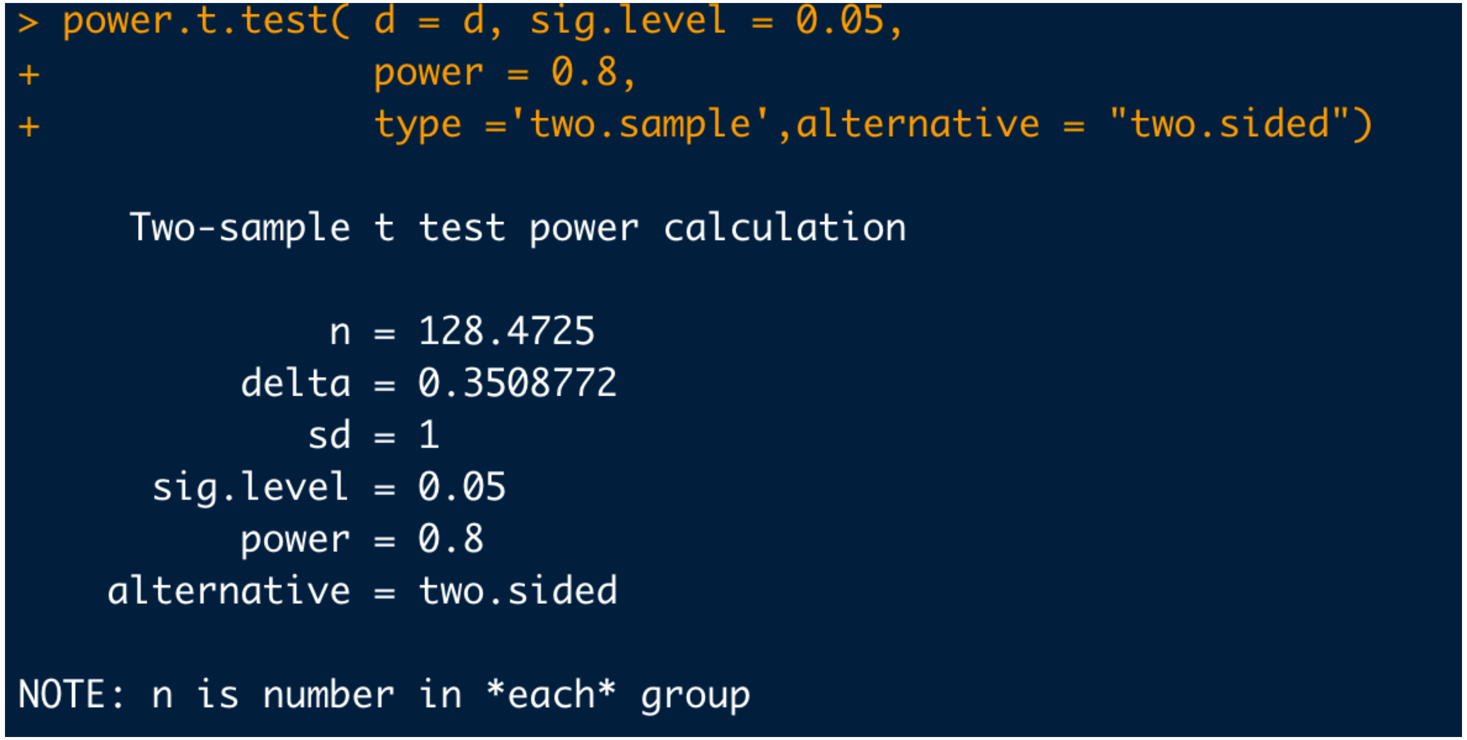
$$2n = \frac{4\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2} =$$

A more accurate estimation would be use t-distribution.

Calculate n in R



```r
delta=2
sigma=5.7
d=delta/sigma
power.t.test( d = d, sig.level = 0.05,
              power = 0.8,
              type ='two.sample',alternative = "two.sided")
```



```
> power.t.test( d = d, sig.level = 0.05,
+               power = 0.8,
+               type ='two.sample',alternative = "two.sided")

     Two-sample t test power calculation 

              n = 128.4725
          delta = 0.3508772
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Choosing Power Level – ethical issue in animal/human studies

- **Underpowered study**
  - Waste resources; can't reject H_0
  - Can misdirect future studies if results are NS (potential misleading conclusion, unnecessary experimentation)
  - Unethical if subjecting individual to inferior treatment
- **Overpowered study**
  - Waste resources? (some animal needlessly sacrificed)
  - Pick up essentially trivial results – meaningless?
  - Costs of collecting data > benefits

Choosing Power Level

What does 80% power means?

- **20% chance missing true difference!**

- Balance between risks
- Large clinical trails use 0.9 or 0.95 ; animal studies usually use 0.8 (80% power).
- Generally Type I error is considered worse
- If can tolerate 5% α, can tolerate 20% β
- Meant as a guideline in considering competing risks, but taken as more absolute these days.

Six rules of thumb for determining power and sample size

- **Rule of Thumb #1**

A larger sample **increases** the statistical power of the evaluation.

- **Rule of Thumb #2**

If the **effect size** of a program is **small**, the **evaluation** needs a **larger** sample to achieve a given level of power.

- **Rule of Thumb #3**

An evaluation of a program with **low take-up** needs a **larger sample**.

- **Rule of Thumb #4**

If the underlying population has **high variation** in outcomes, the evaluation needs a **larger sample**.

- **Rule of Thumb #5**

For a given sample size, power is maximized when the sample is **equally split** between the treatment and control group.

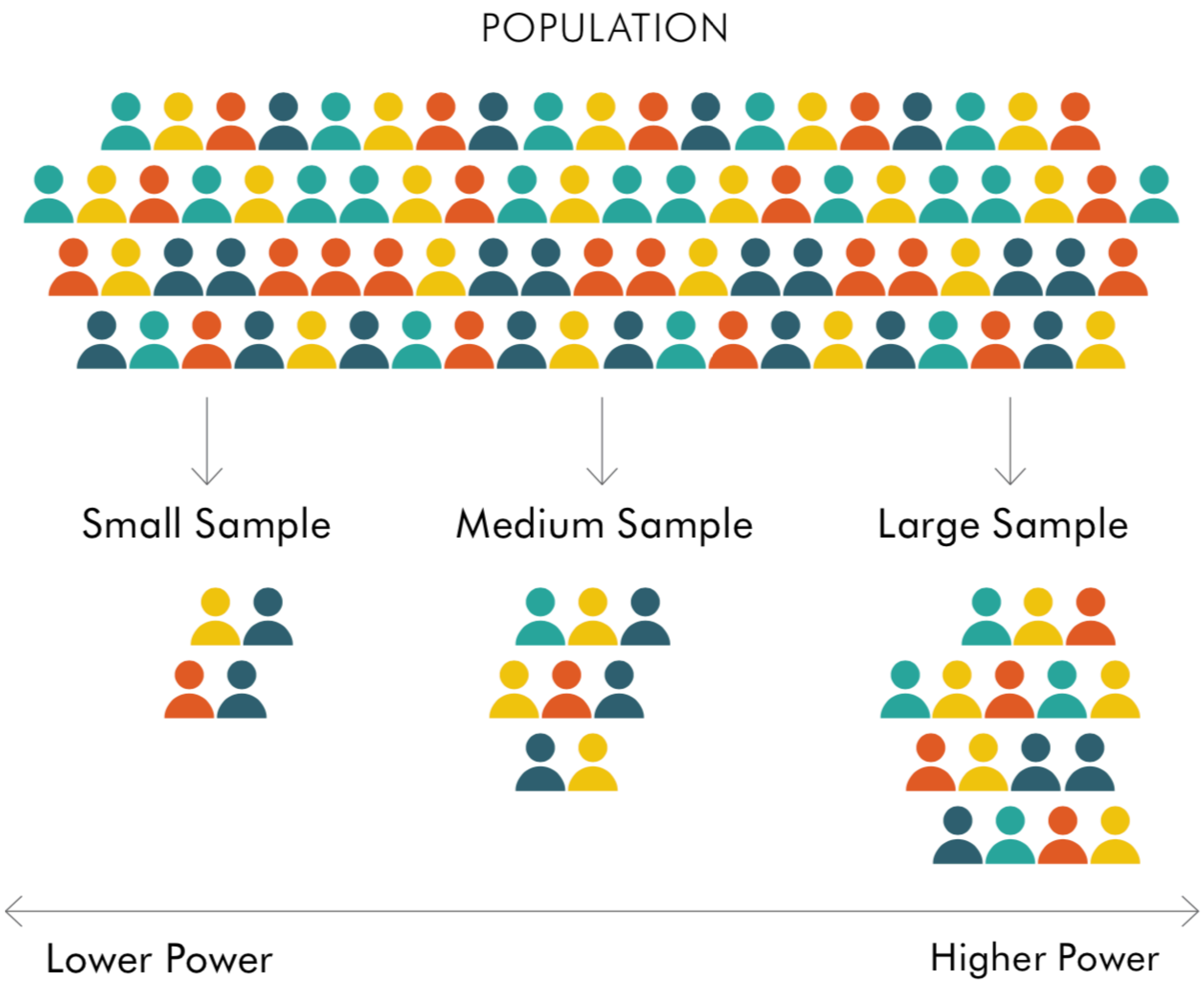
- **Rule of Thumb #6**

For a given sample size**, randomizing at the cluster level** as opposed to the individual level **reduces** the **power of the evaluation**. The more similar the outcomes of individuals within clusters are, the larger the sample needs to be.

https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #1: a larger sample increases the statistical power of the evaluation

- Researchers run evaluations on samples that are selected from a larger population. Need to decide the sample size.
- Ideally, Include the whole population.
- Larger samples are more likely to be representative of the original population and are more likely to capture impacts that would occur in the population.

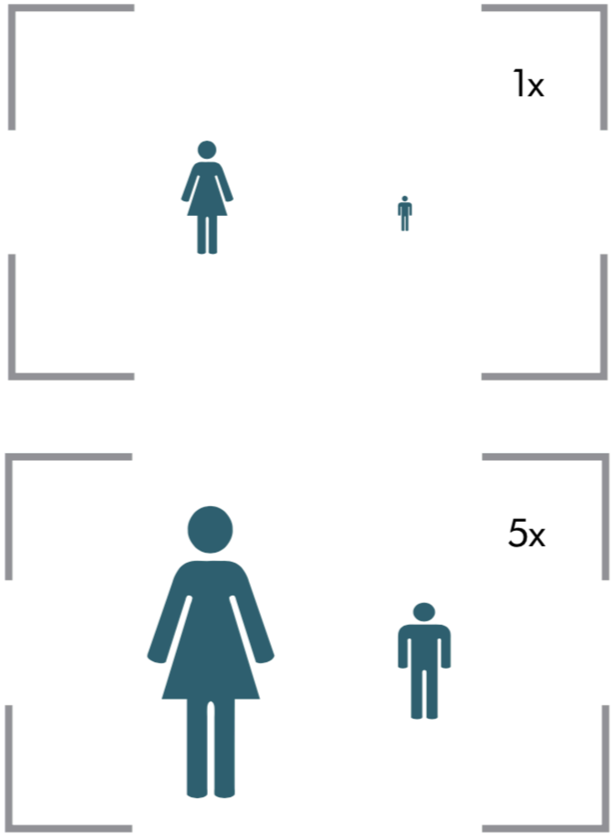


https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #2: if the effect size is small, the evaluation needs a larger sample to achieve a given level of power

- When designing an evaluation, the research team wants to ensure that they are able to identify the effect of the program with precision.
- When an evaluation has sufficient power, impact estimates are precise.
- Both the effect size and sample size affect precision.

- The size of the images represents the effect size, and the level of zoom represents the sample size of the evaluation.
- For a given level of power, large effects can be precisely detected with a smaller sample size, while smaller effects can only be precisely detected with larger sample sizes.



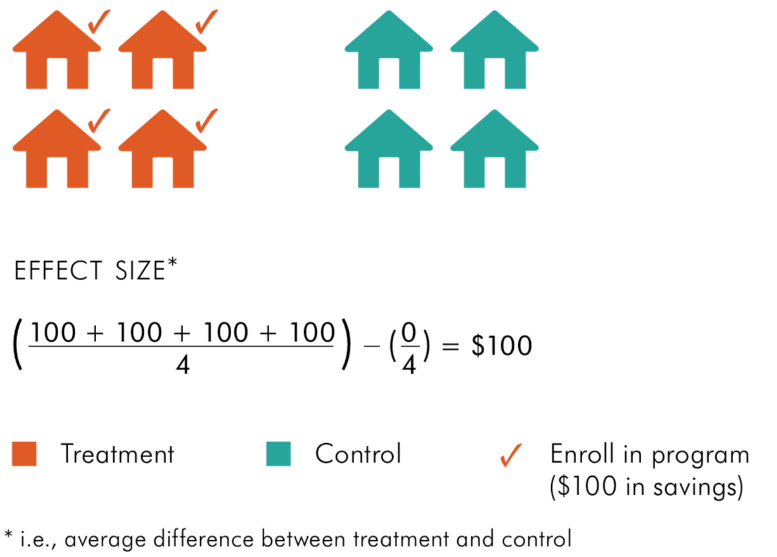
https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #3: an evaluation of a program with low take-up needs a larger sample

- **Randomized evaluations** are designed to detect the average effect of a program over the entire sample that is assigned to the treatment group.

- Therefore, **lower take-up decreases** the magnitude of the average effect of the program.

- Since a larger sample is required to detect a smaller effect, it is important to **plan ahead** if low take-up is anticipated and run the evaluation with a larger sample.

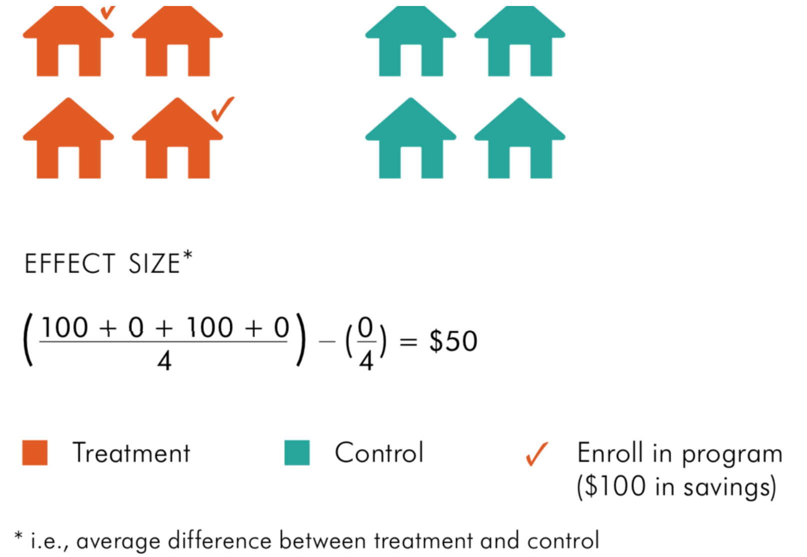


EFFECT SIZE*

$$\left(\frac{100 + 100 + 100 + 100}{4}\right) - \left(\frac{0}{4}\right) = \$100$$

🟧 Treatment 🟩 Control ✓ Enroll in program ($100 in savings)

* i.e., average difference between treatment and control



EFFECT SIZE*

$$\left(\frac{100 + 0 + 100 + 0}{4}\right) - \left(\frac{0}{4}\right) = \$50$$

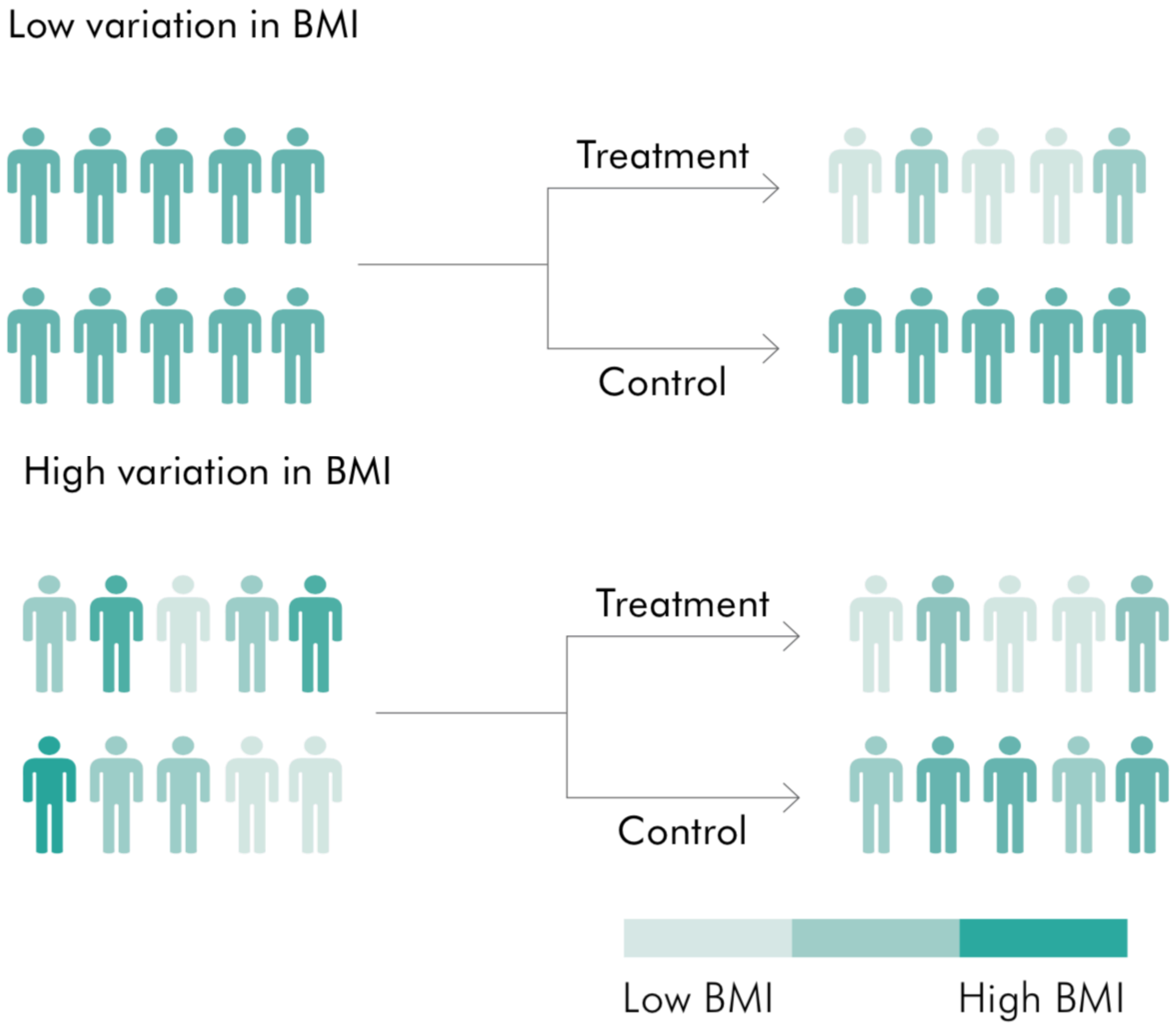
🟧 Treatment 🟩 Control ✓ Enroll in program ($100 in savings)

* i.e., average difference between treatment and control

https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #4: if the population has high variation in outcomes, need a larger sample

- In a population with **high variation** in key outcome measures (e.g., BMI), it is challenging to disentangle the effect of the program from the effect of random variation in these outcome measures.

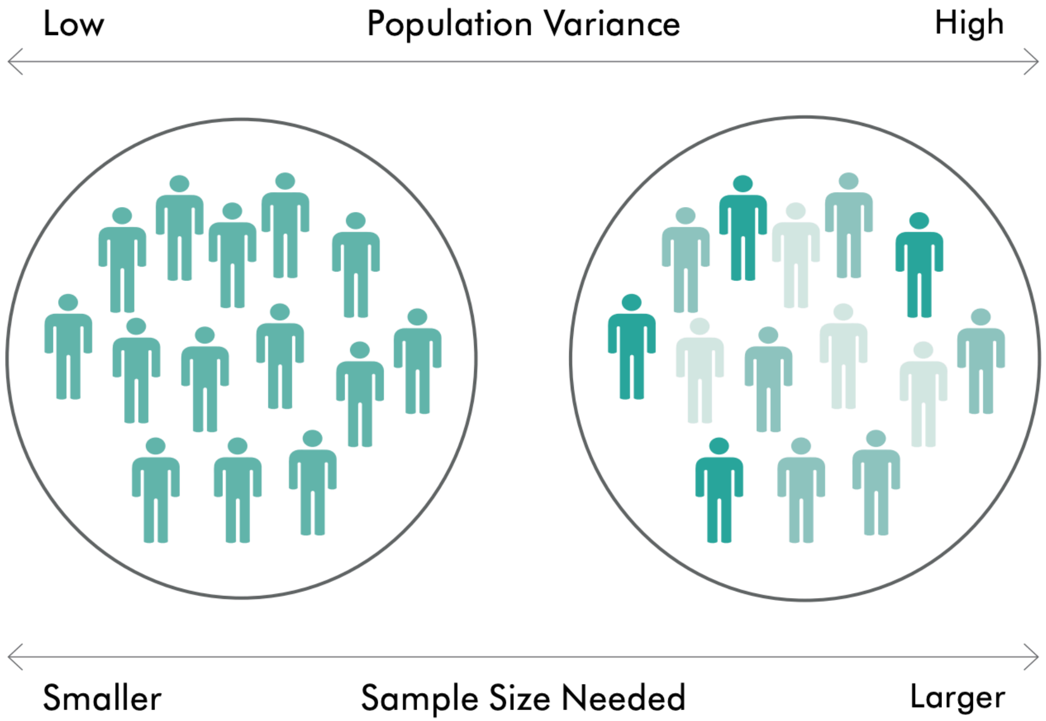


https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #4: if the population has high variation in outcomes, need a larger sample

- Especially when running an evaluation on a population with **high variance**, selecting a **larger sample increases** the **likelihood** that you will be able to distinguish the impact of the program from the impact of naturally occurring variation in key outcome measures.

- Larger samples in the presence of high variance make it easier to identify the causal impact of a program.



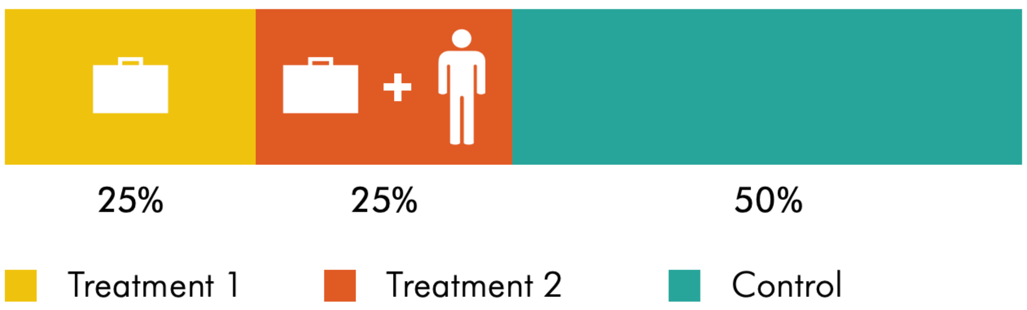
https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #5: for a given sample size, power is maximized when the sample is equally split between the treatment and control group

- To achieve maximum power for a given sample size, the sample should be **evenly divided** between the treatment group and control group.

- Taking resource constraints, intervention costs, data collection costs, and multiple treatment arms into account, research teams **may decide on an uneven ratio** of treatment to control participants.

- Evaluations with **multiple treatment arms** (i.e., different versions or combinations of treatments) help researchers to disentangle mechanisms, determine which aspect of a treatment bundle drives impact, and identify whether the components of the treatment bundle are complements or substitutes.



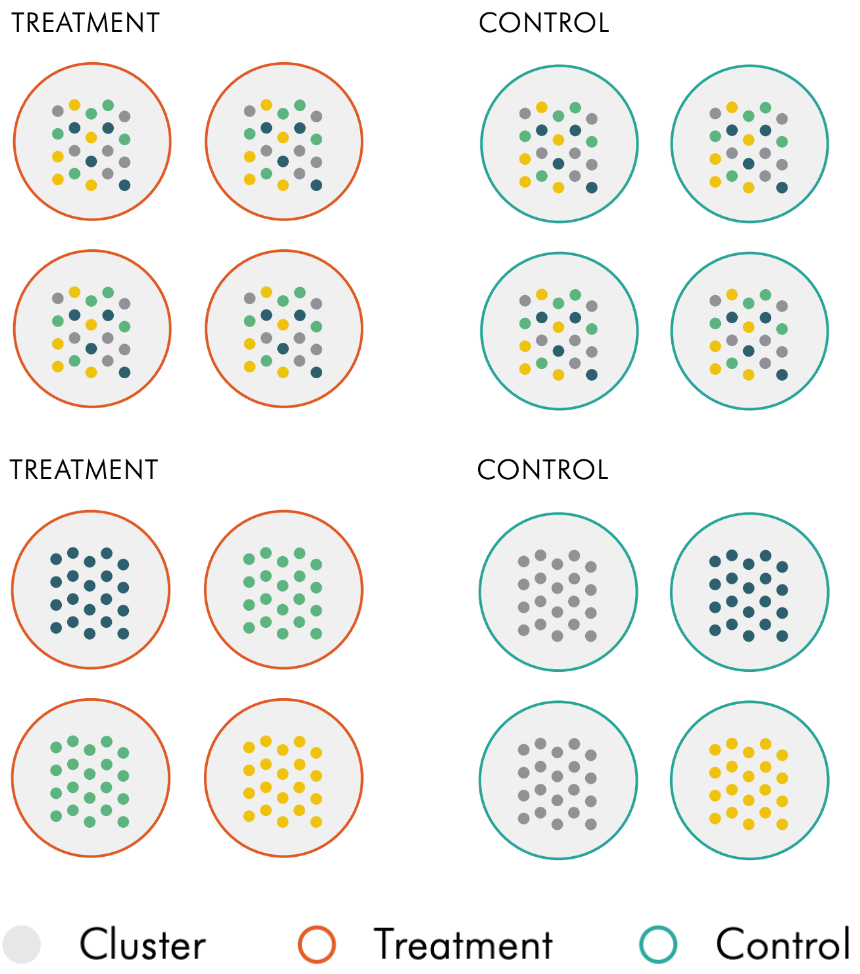
ALLOCATION OF SAMPLE TO STUDY ARMS

What's the alternative way of splitting your sample in this case?

https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

rule of thumb #6: the more similar the outcomes of individuals within clusters are, the larger the sample needs to be.

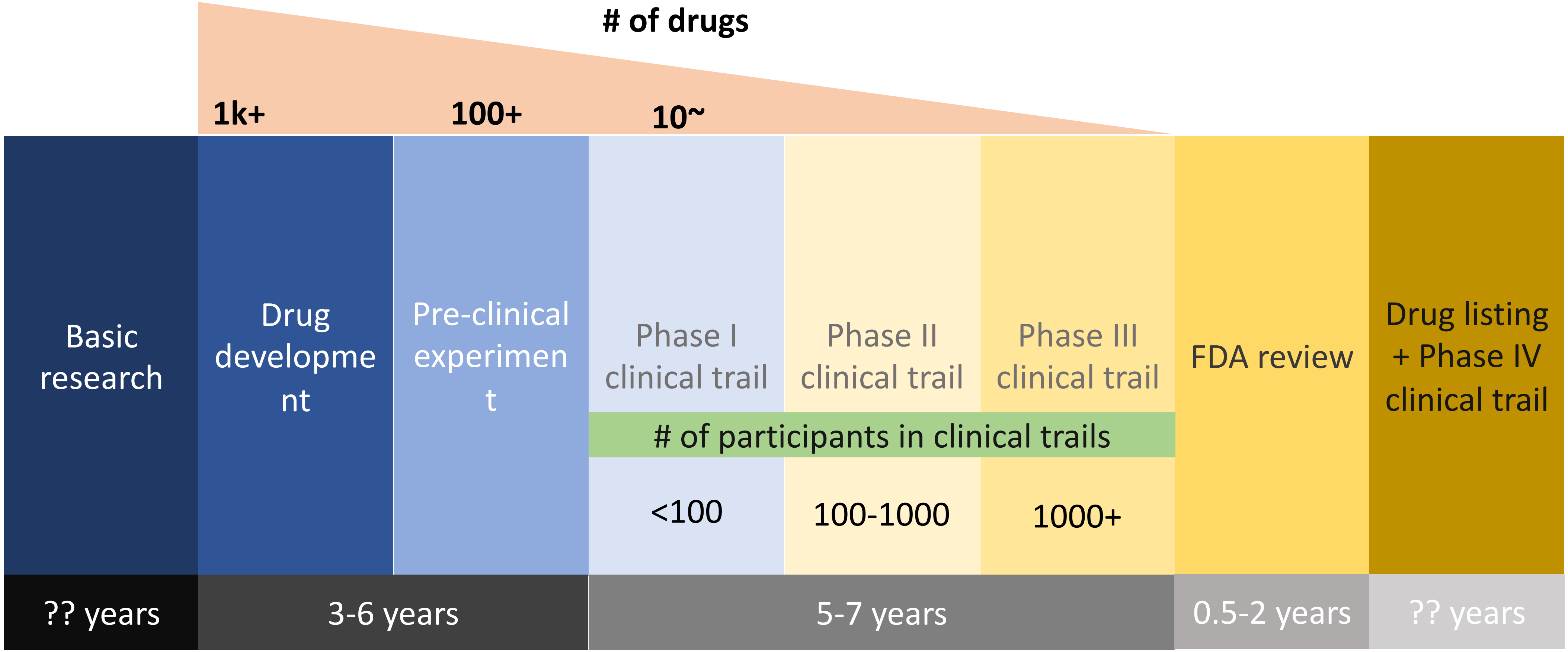
- When designing an evaluation, the research team must choose the **unit of randomization**.
- For example, individuals can be randomly assigned to the treatment group or control group.
- Alternatively, randomization can be done by “**clusters**.”
- By this method, groups of individuals are treated as units, whether they are households, classrooms, schools, or neighborhoods, and each **cluster** is randomly assigned to the treatment group or the control group.
- For a given sample size, **randomizing clusters** as opposed to individuals **decreases** the power of the study. The reason for this relates to how similar the outcomes of individuals



https://www.povertyactionlab.org/sites/default/files/resources/2018.03.21-Rules-of-Thumb-for-Sample-Size-and-Power.pdf

How FDA approve a drug

What's the purpose of phase I/II/III/IV clinical trails?



# of drugs

1k+

100+

10~

Basic research

Drug development

Pre-clinical experiment

Phase I clinical trail

Phase II clinical trail

Phase III clinical trail

FDA review

Drug listing + Phase IV clinical trail

# of participants in clinical trails

<100

100-1000

1000+

?? years

3-6 years

5-7 years

0.5-2 years

?? years

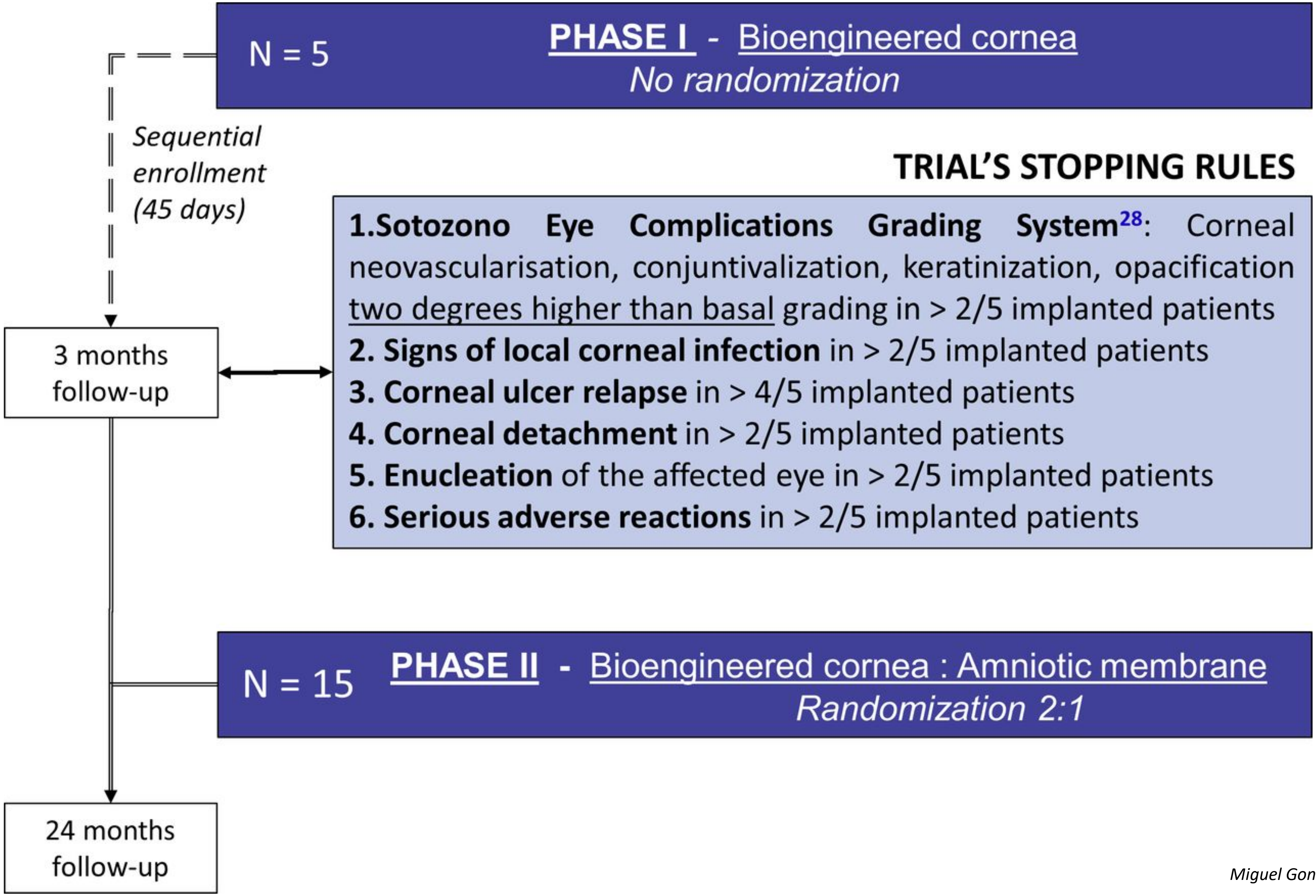
Stopping rules (phase I as example)

- Clinical trials are unusual in that enrollment of subjects is a **continual** process staggered in time.

- If a treatment can be proven to be clearly **beneficial** or **harmful** compared to the concurrent **control**, or to be obviously futile, based on a predefined analysis of an incomplete data set while the study is ongoing, the investigators may **stop the study early**.

- **Interim analysis** is an analysis of data that is conducted before data collection has been completed.

Stopping rules (phase I as example)



N = 5

PHASE I - Bioengineered cornea

No randomization

Sequential enrollment (45 days)

TRIAL'S STOPPING RULES

1.Sotozono Eye Complications Grading System[28]: Corneal neovascularisation, conjuntivalization, keratinization, opacification two degrees higher than basal grading in > 2/5 implanted patients

2. Signs of local corneal infection in > 2/5 implanted patients

3. Corneal ulcer relapse in > 4/5 implanted patients

4. Corneal detachment in > 2/5 implanted patients

5. Enucleation of the affected eye in > 2/5 implanted patients

6. Serious adverse reactions in > 2/5 implanted patients

3 months follow-up

N = 15

PHASE II - Bioengineered cornea : Amniotic membrane

Randomization 2:1

24 months follow-up

Miguel Gonzalez-Andrades et al., 2013, BMJ Open

Summary

- Understand intuition behind power calculations
  - Why we care about power? Overpower? Underpower?
- Know how to perform power/sample size analysis with formula or in R
  - Formula and which function to use in R?
- Reveal the relationship among significance level, power, effect size and sample size
  - How power changes when others changes?
- Demonstrate different stages in clinical trails and stopping rules