18 Sample Size

Estimation, Type I & II errors, power, effect size

COMS10011

Luluah Albarrak

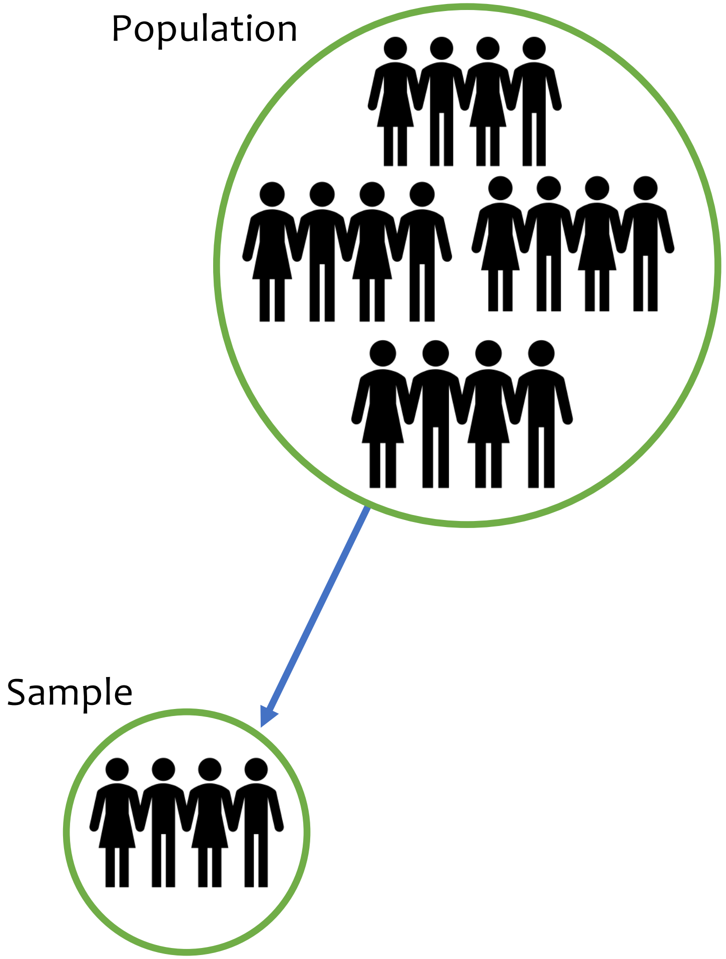
https://github.com/coms10011/

Random Sample

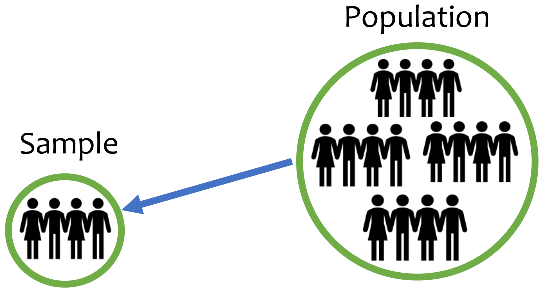
Population: is a set of all units of interest.

Sample: a subset of the population.

Random sample: a sample collected in such a way that every member of the population is equally likely to be selected.



Random Sample



The **goal** is to make estimates and predictions about a population based on information from a sample. In particular, we want to estimate the population mean μ , and the population standard deviation σ.

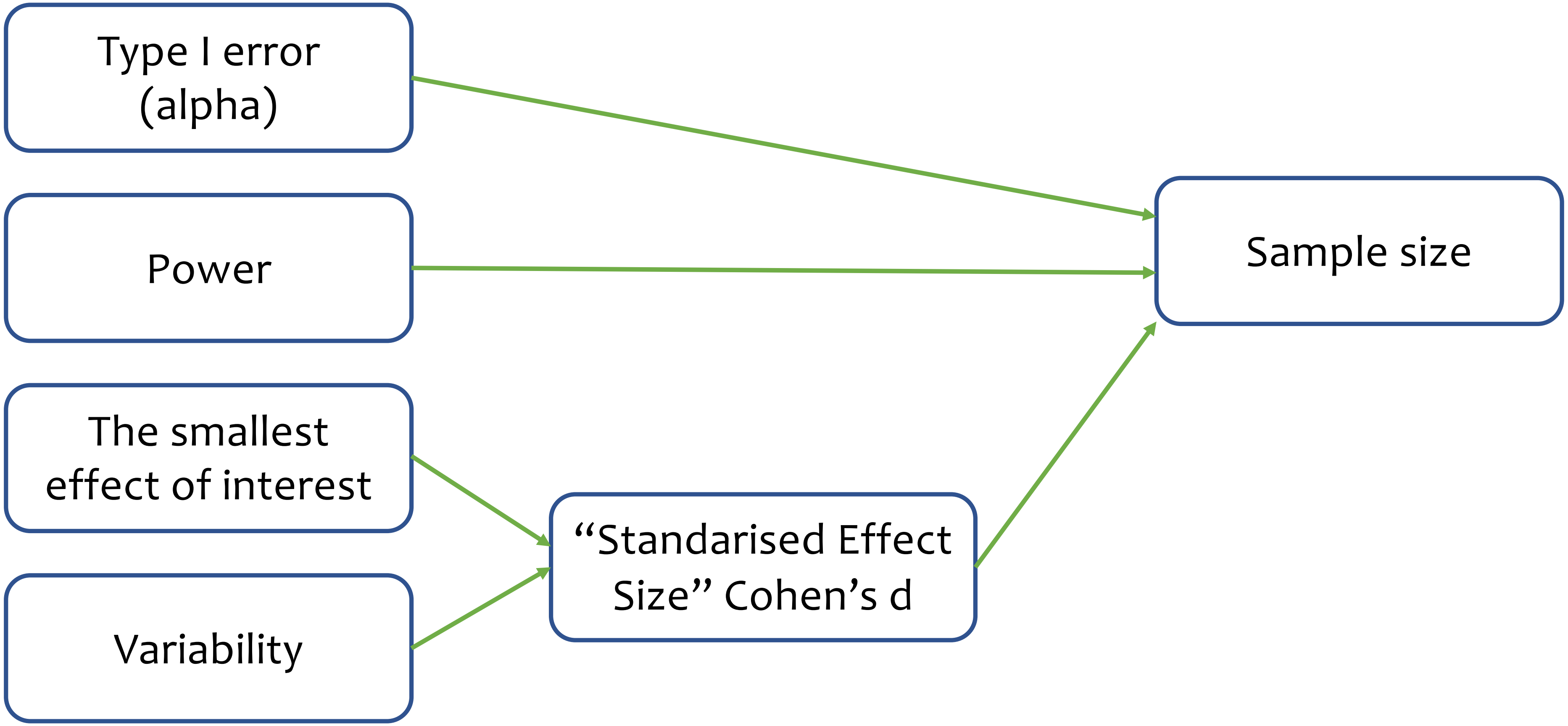
Sample **S**ize **C**alculation

The main aim of a sample size calculation is to determine the number of participants needed to detect a scientifically relevant treatment effect.

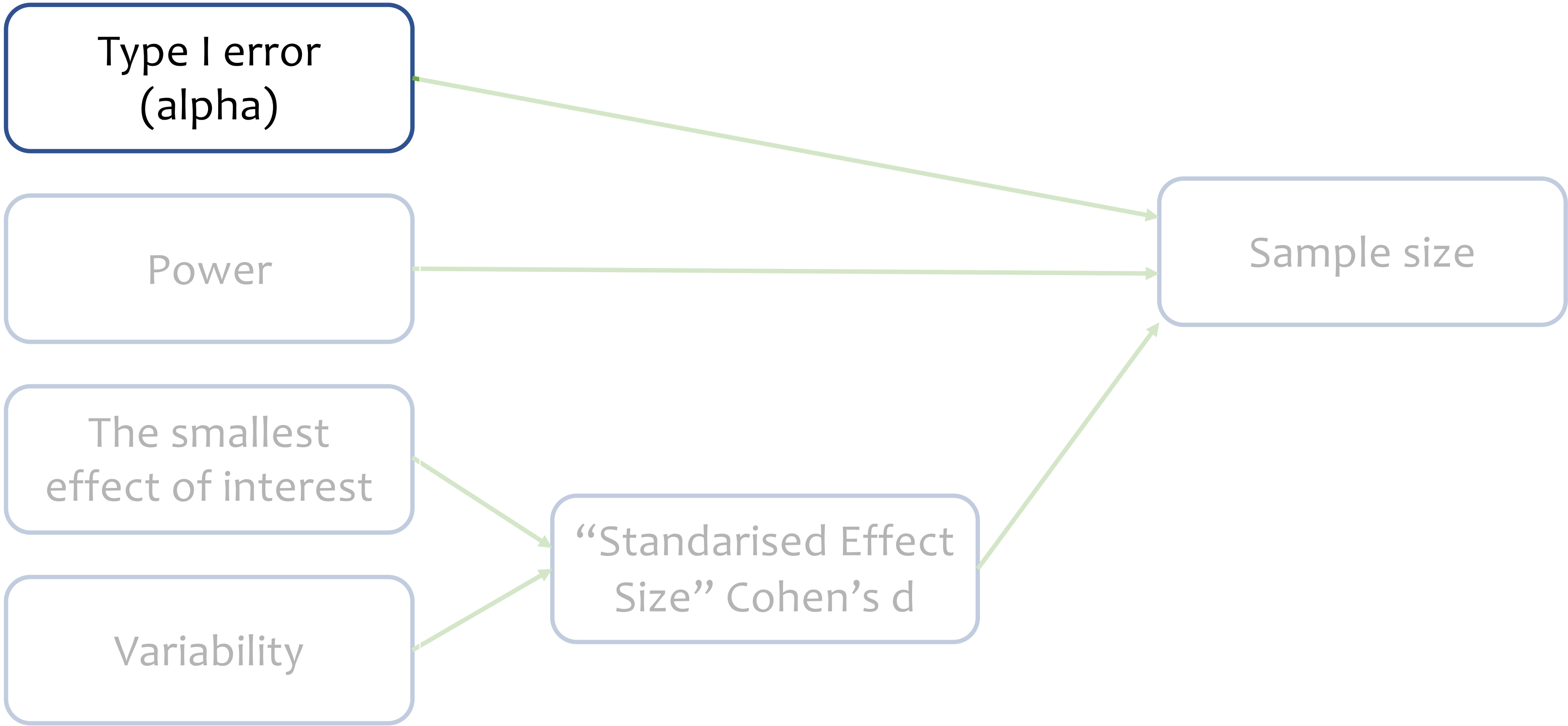
Sample size is too small → one may not be able to detect an important existing effect.

Sample size is too large → waste of time, resources and money.

Components of sample size calculations


Type I error (alpha)
Power
The smallest effect of interest
Variability
“Standarised Effect Size” Cohen’s d
Sample size

Components of sample size calculations



Type I error (alpha)

Power

The smallest effect of interest

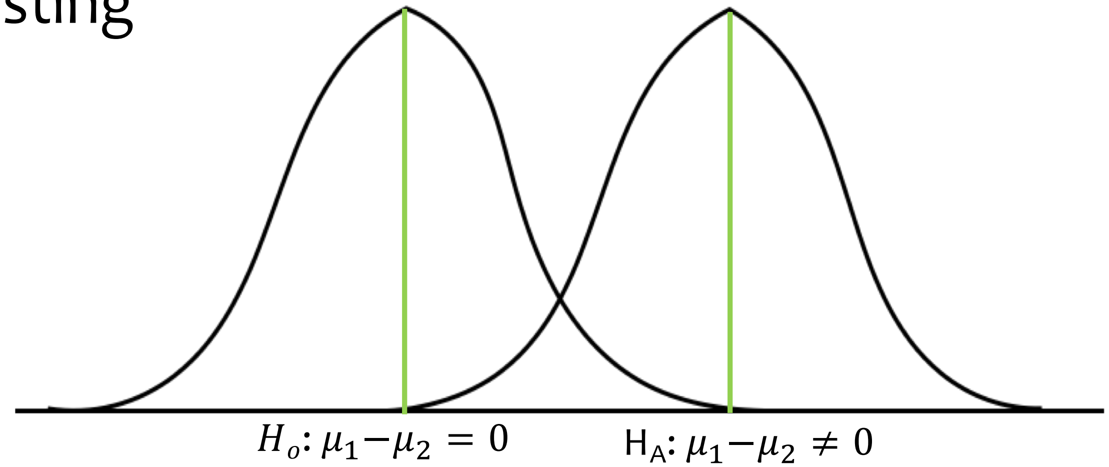
Variability

"Standarised Effect Size" Cohen's d

Sample size

Type I & **T**ype II **E**rrors

Hypothesis testing



When we conduct a test of any hypothesis regardless of the test used we make one of two possible decisions:

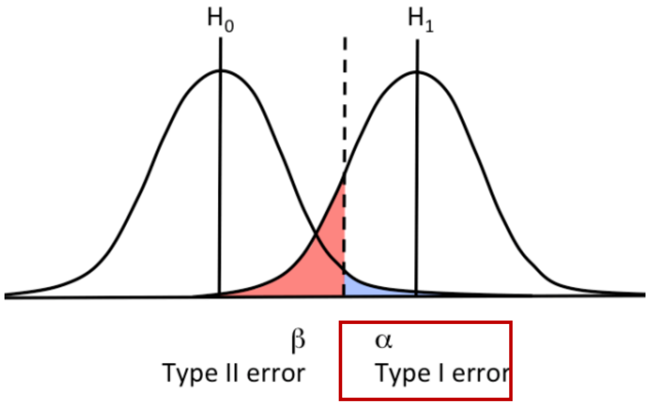
Reject the null (H_o) in favor of the alternative (H_A)

OR

Fail to reject the null hypothesis (H_o)

Type I Errors

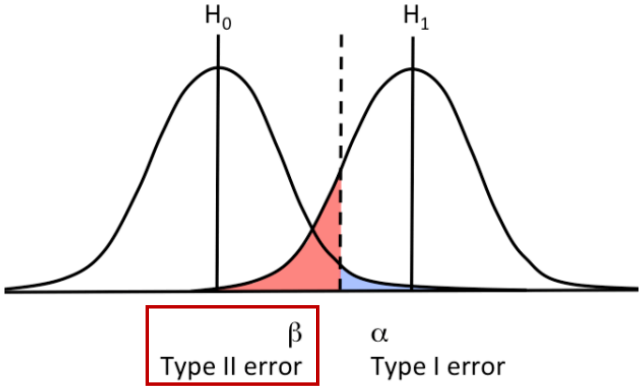
	Truth about the population/reality	
TEST DECISION	H_o TRUE	H_a TRUE
Reject the Null Hypothesis	TYPE I ERROR (α)	Power (1-β) CORRECT DECISION
Fail to Reject the Null Hypothesis	CORRECT DECISION	TYPE II ERROR (β)



Type I error: the chance of detecting a statistically significant difference when there is no real difference between treatments (The risk of a false positive result).

Type II Errors

	Truth about the population/reality	
TEST DECISION	H_o TRUE	H_a TRUE
Reject the Null Hypothesis	TYPE I ERROR (α)	Power (1-β) CORRECT DECISION
Fail to Reject the Null Hypothesis	CORRECT DECISION	TYPE II ERROR (β)



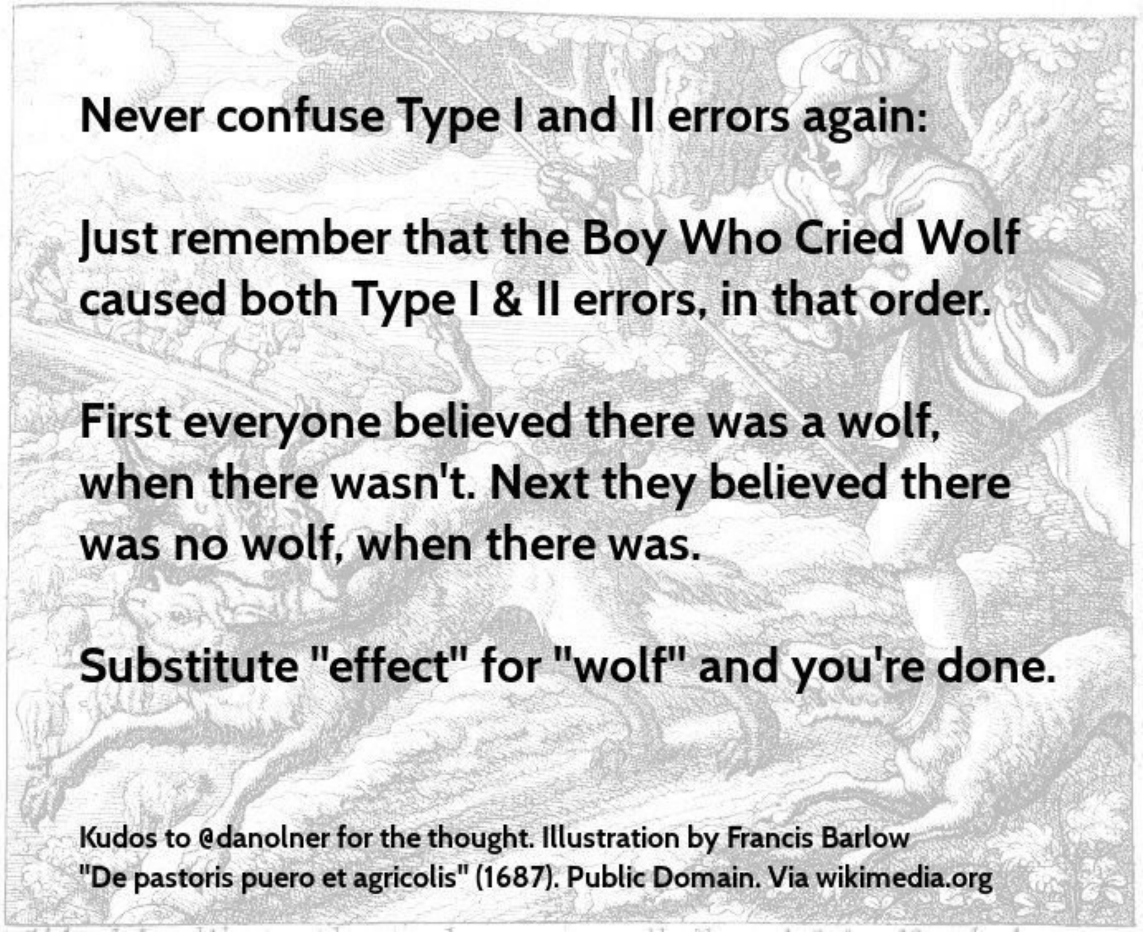
Type II error: the chance of not detecting a significant difference when there really is a difference (The risk of a false negative result).

Type I & **T**ype II **E**rrors

We choose P(Type I Error) = α
Typically α = .05 or .01 or .10

However we do NOT directly choose
β = P(Type II Error)

Type I & **T**ype II **E**rrors



Never confuse Type I and II errors again:

Just remember that the Boy Who Cried Wolf caused both Type I & II errors, in that order.

First everyone believed there was a wolf, when there wasn't. Next they believed there was no wolf, when there was.

Substitute "effect" for "wolf" and you're done.

Kudos to @danolner for the thought. Illustration by Francis Barlow "De pastoris puero et agricolis" (1687). Public Domain. Via wikimedia.org

Question

It has been shown many times that on a certain memory test, recognition is substantially better than recall. However, the probability value for the data from your sample was 0.12, so you were unable to reject the null hypothesis that recall and recognition produce the same results. What type of error did you make?

Type I or Type II Error?

Question

It has been shown many times that on a certain memory test, recognition is substantially better than recall. However, the probability value for the data from your sample was 0.12, so you were unable to reject the null hypothesis that recall and recognition produce the same results. What type of error did you make?

Type I or Type II Error?

Type II Error

Question

In the population, there is no difference between men and women on a certain test. However, you found a difference in your sample. The probability value for the data was 0.03, so you rejected the null hypothesis. What type of error did you make?

Type I or Type II Error?

Question

In the population, there is no difference between men and women on a certain test. However, you found a difference in your sample. The probability value for the data was 0.03, so you rejected the null hypothesis. What type of error did you make?

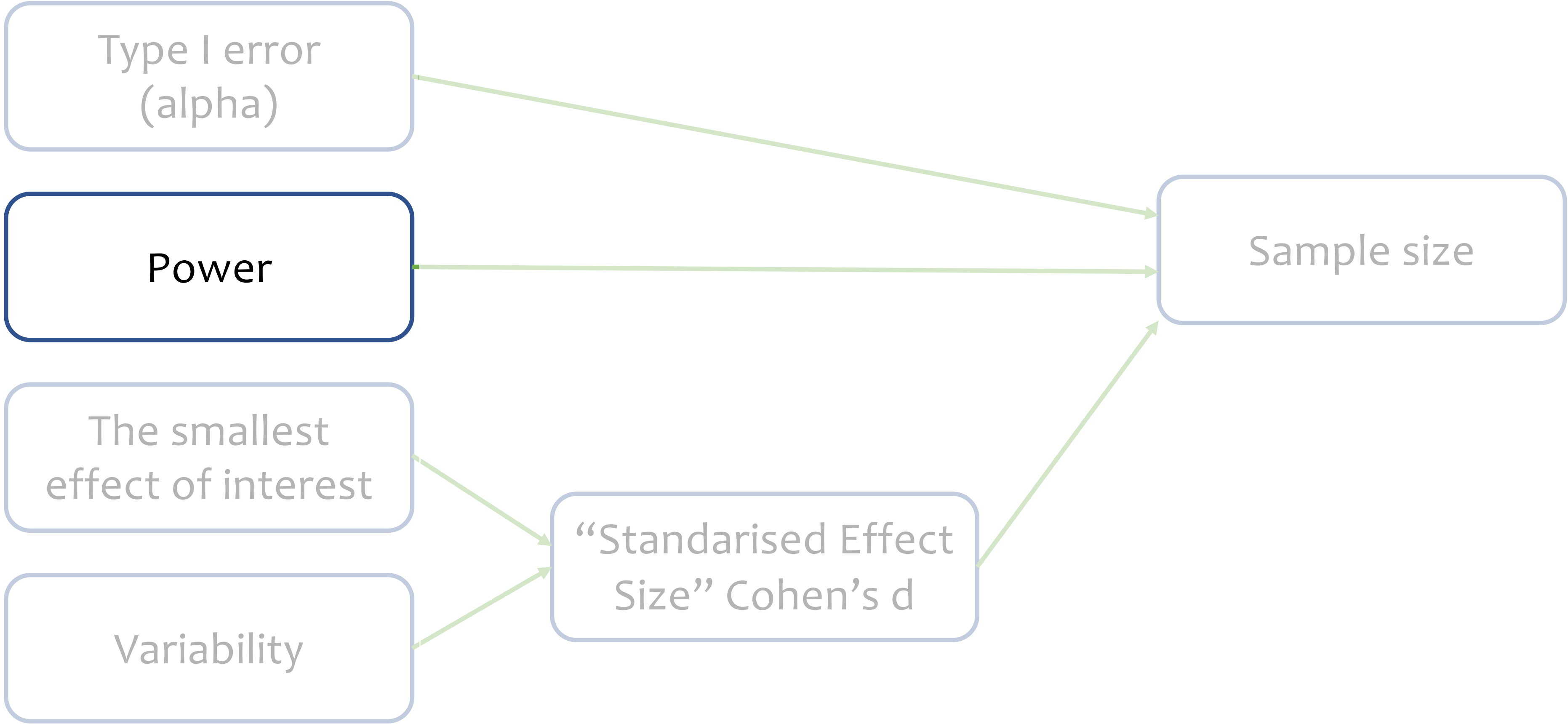
Type I or Type II Error?

Type I Error

How p-value affects sample size?

As the p-value **decreases,** the necessary sample size **increases.**

Components of sample size calculations



Type I error (alpha)

Power

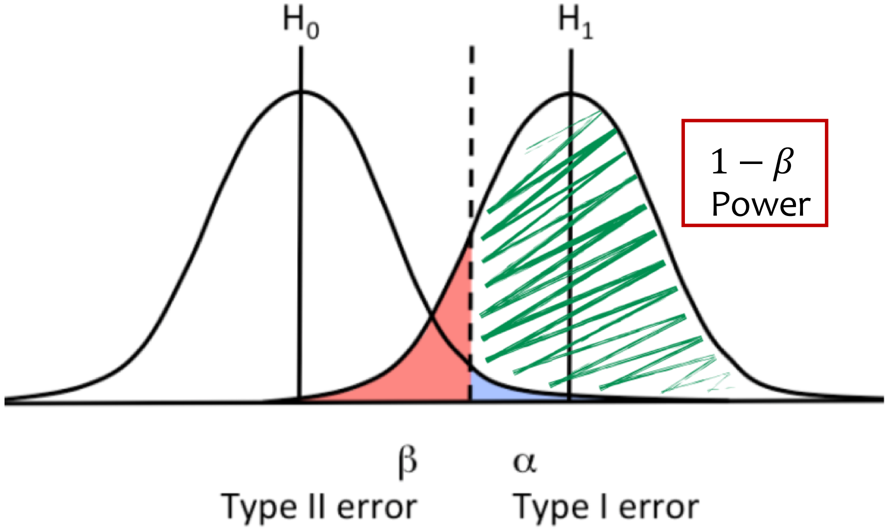
The smallest effect of interest

Variability

"Standarised Effect Size" Cohen's d

Sample size

Power



The **power** of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the **power** of a hypothesis test is the probability of rejecting the null hypothesis H_o when the alternative hypothesis H_A is true.

Power

Higher power is better (the closer the power is to 1.0 or 100%).

The ideal power is considered to be 80% → we are accepting that one in five times (20%) we will miss the difference.

Power

To increase power (and hence decrease type 2 error rate):

Increase the sample size

Decrease the standard deviation of the sample

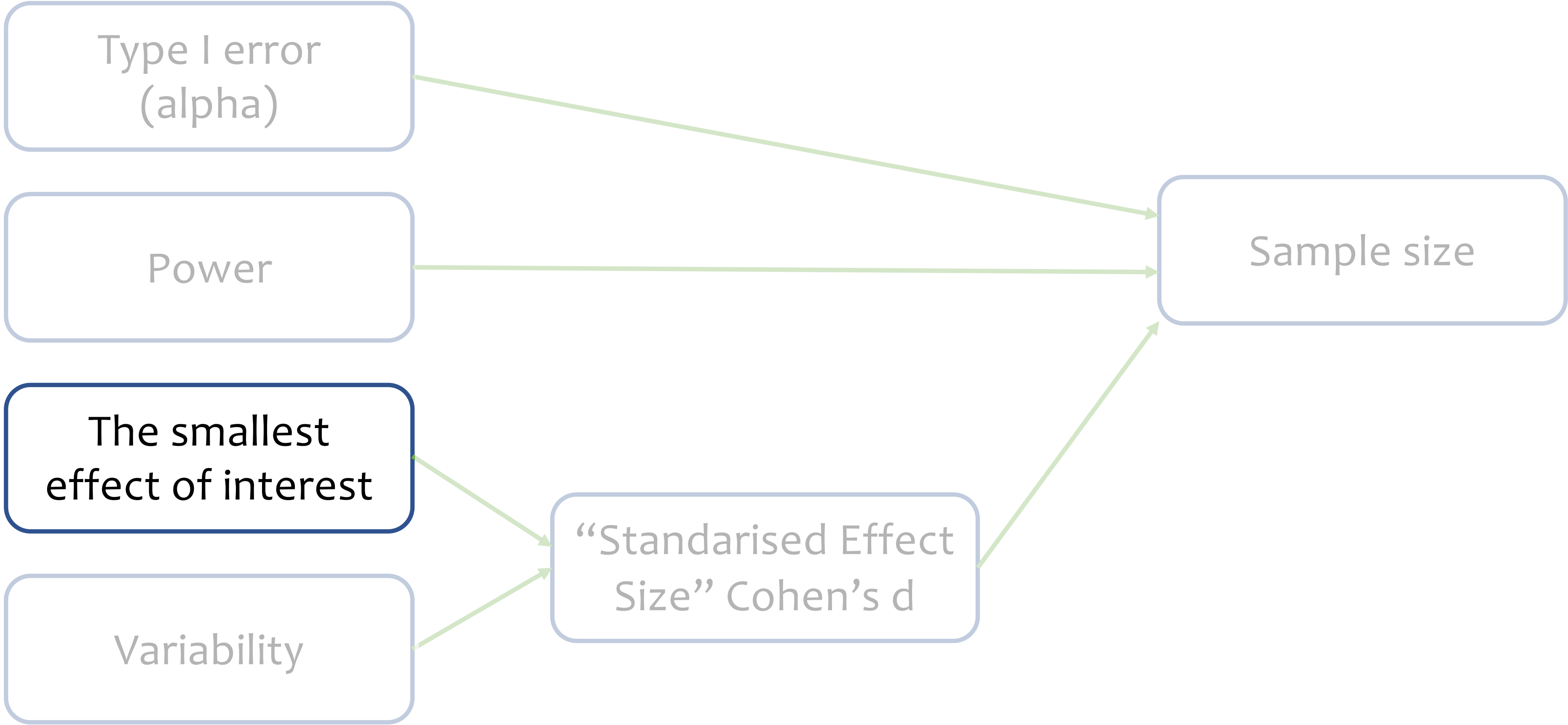
Increase α

Consider a larger effect size

How power affects sample size?

As the power **increases,** the necessary sample size **increases**.

Components of sample size calculations



Type I error (alpha)

Power

The smallest effect of interest

Variability

"Standarised Effect Size" Cohen's d

Sample size

Smallest **E**ffect of **I**nterest

The smallest effect of interest is the minimal difference between the studied groups that the investigator wishes to detect.

Smallest **E**ffect of **I**nterest

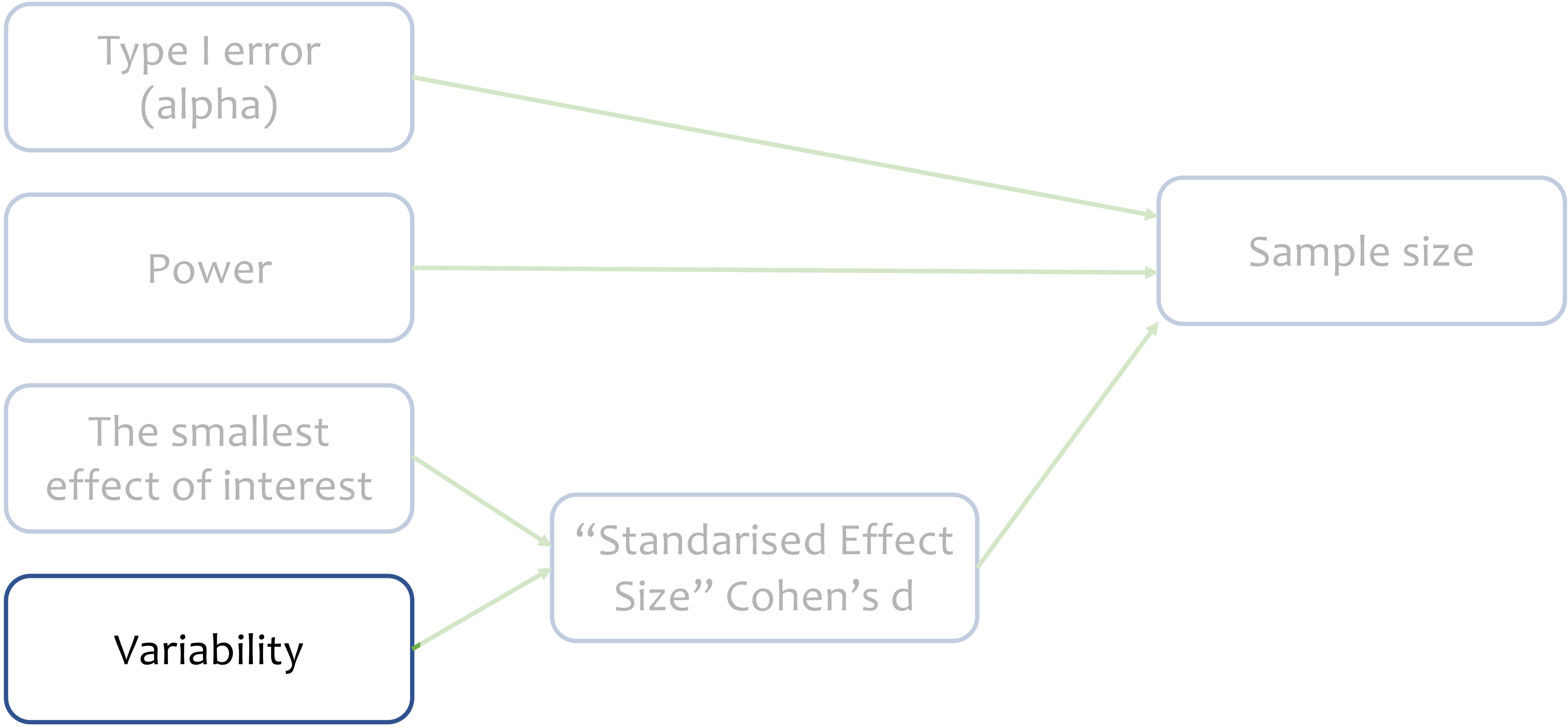
For **continuous** outcome variables, the minimal scientifically relevant difference is a **numerical** difference. For example, if body weight is the outcome of a trial, an investigator could choose a difference of 5 kg as the minimal scientifically relevant difference.

For **binary** outcome variables, the minimal difference is expressed in **rates**. For example, in the case of studying the effect of a drug on weight loss (yes/no), an investigator could choose a difference of 10% between the treatment group and control group as the minimal scientifically relevant difference .

How effect of interest affects sample size?

As the effects of interest between the study groups **increases,** the necessary sample size **decreases.**

Components of sample size calculations



Type I error (alpha)

Power

The smallest effect of interest

Variability

"Standarised Effect Size" Cohen's d

Sample size

Variability

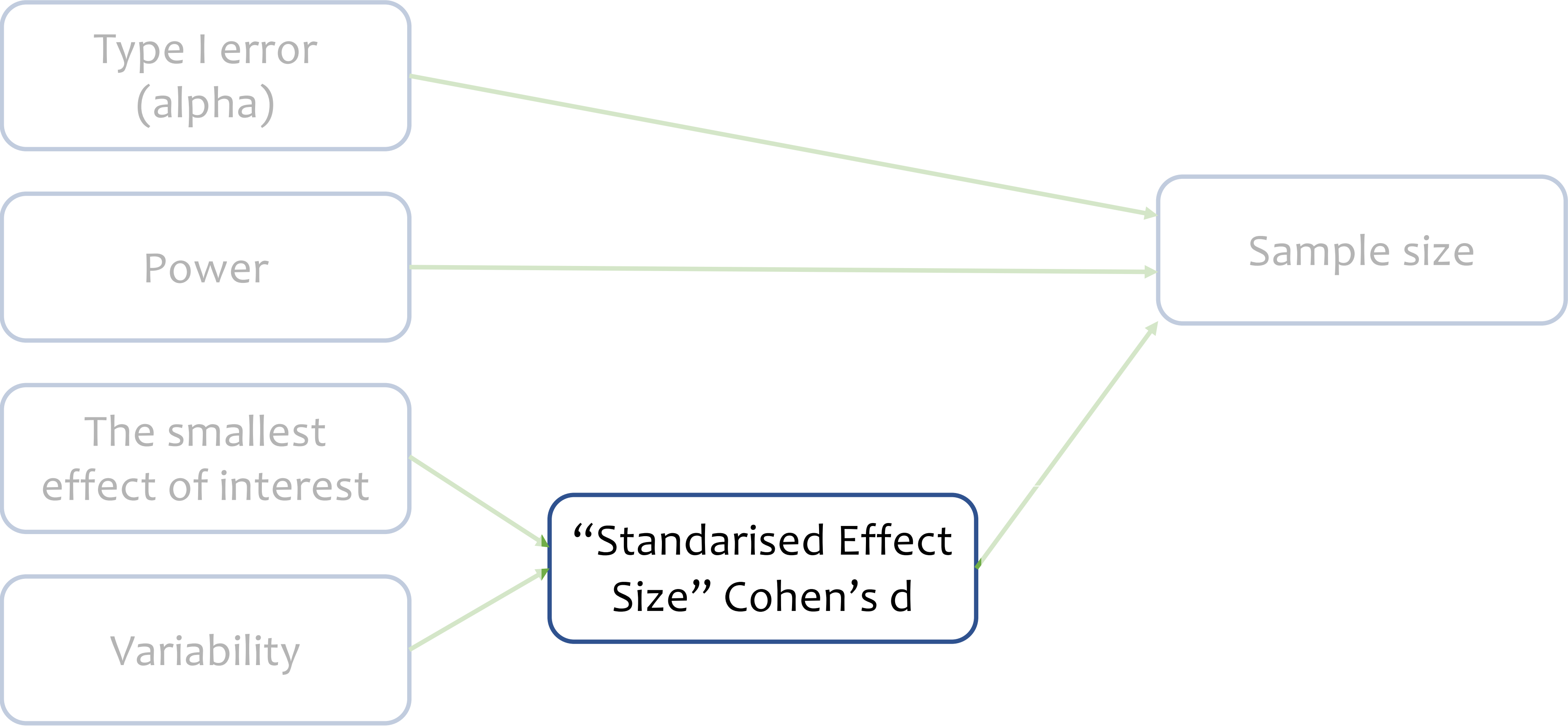
Sample size calculation is based on using the population variance of a given outcome variable that is estimated by means of the standard deviation (SD) in case of a continuous outcome.

Because the variance is usually an unknown quantity, investigators often use an estimate obtained from a pilot study or use information from a previous study.

How variance affects sample size?

As the variance **increases,** the necessary sample size **increases.**

Components of sample size calculations



Type I error (alpha)

Power

The smallest effect of interest

Variability

“Standarised Effect Size” Cohen’s d

Sample size

Effect Size

The smallest effect of interest and the variability are combined and expressed as a multiple of the SD of the observations; known as the standardised difference.

The standardised difference is also referred to as the <u>effect size</u>.

$$\text{Effect Size} = \frac{\text{Difference between the means in the two treatment groups}}{\text{Standard Deviation}}$$

Effect **S**ize

Effect size is a way of quantifying the difference between two or more groups, or a measure of the difference in the outcomes of the experimental and control groups.

For example, if one group has a new treatment and the other has not (control group), then the effect size is a measure of the effectiveness of the treatment.

Effect **S**ize

A statistically significant result does not mean it is substantive in effect. For example, two treatments could be shown to be significantly different, but their clinical effects may be so small as to be unimportant.

Cohen's d Effect Size

Depending on the type of study, effect size is estimated with different measures.

The most straightforward effect size measure is the difference between two means.

Cohen's d is a good example of a standardized effect size measurement.

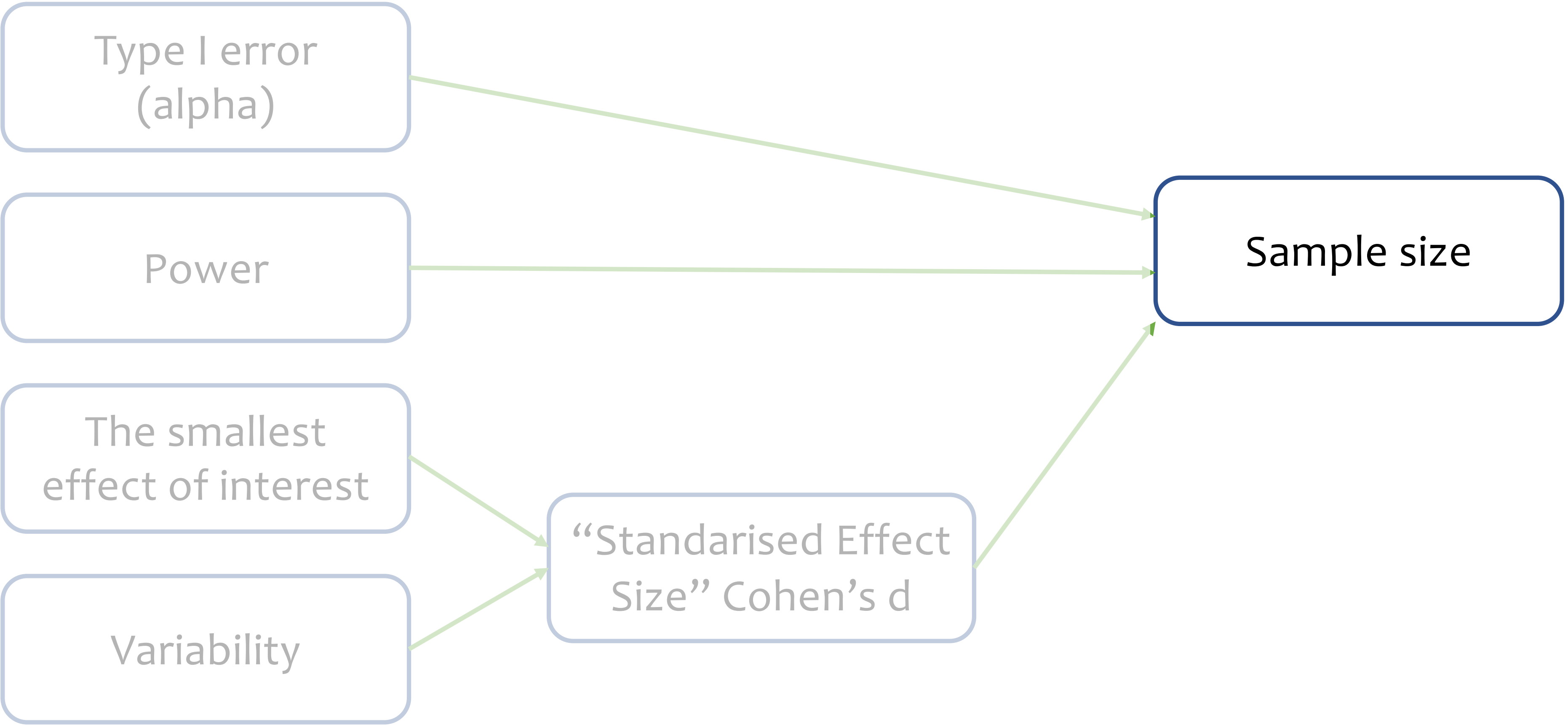
Cohen (1988) proposed a simple categorisation of small, moderate and large effect size.

$$\boxed{d = \frac{\mu_1 - \mu_2}{SD_{pooled}}}$$

Effect	Cohen's d
Small	0.20
Medium	0.50
Large	0.80

Sample **S**ize **E**stimation

Components of sample size calculations



Type I error (alpha)

Power

The smallest effect of interest

Variability

“Standarised Effect Size” Cohen’s d

Sample size

Sample **S**ize for **C**ontinuous **D**ata

There are several methods used to calculate the sample size depending on the type of data or study design.

The sample size for continuous data when comparing two means (independent) is calculated using the following formula:

$$\boxed{n = \frac{2\left(Z_{\alpha} + Z_{1-\beta}\right)^{2}\sigma^{2}}{\delta^{2}}}$$

The number, n, is the sample size required in each group.

Example

In a sample of inactive overweight men aged between 50 and 60, suppose we wish to compare the mean blood pressure of Group 1 who underwent a calorie-controlled diet to Group 2 undertook the exercise-training programme.

Let,

$\mu_1 = mean$ blood pressure of group 1

$\mu_2 = mean$ blood pressure of group 2

Example

Then,

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

Suppose researchers would like to have a power of 80% to detect a difference of 5 mmHg between these two population means at the α=.05 level. The standard deviation (based on data in a published paper) would be approximately 20 mmHg.

What samples sizes (n_1) and (n_2) should they use?

Example

Then,

$H_0: \mu_1 = \mu_2$

$H_a: \mu_1 \neq \mu_2$

Suppose researchers would like to have a power of 80% to detect a difference of 5 mmHg between these two population means at the α=.05 level. The standard deviation (based on data in a published paper) would be approximately 20 mmHg.

What samples sizes (n_1) and (n_2) should they use?

Example

$\delta^2 = (\mu_1 - \mu_2)^2 = (5)^2 = 25$

β=.20 , $\alpha = .05$

SD= 20

α-error	5%	1%
2-sided	1.96	2.5758
1-sided	1.65	2.33

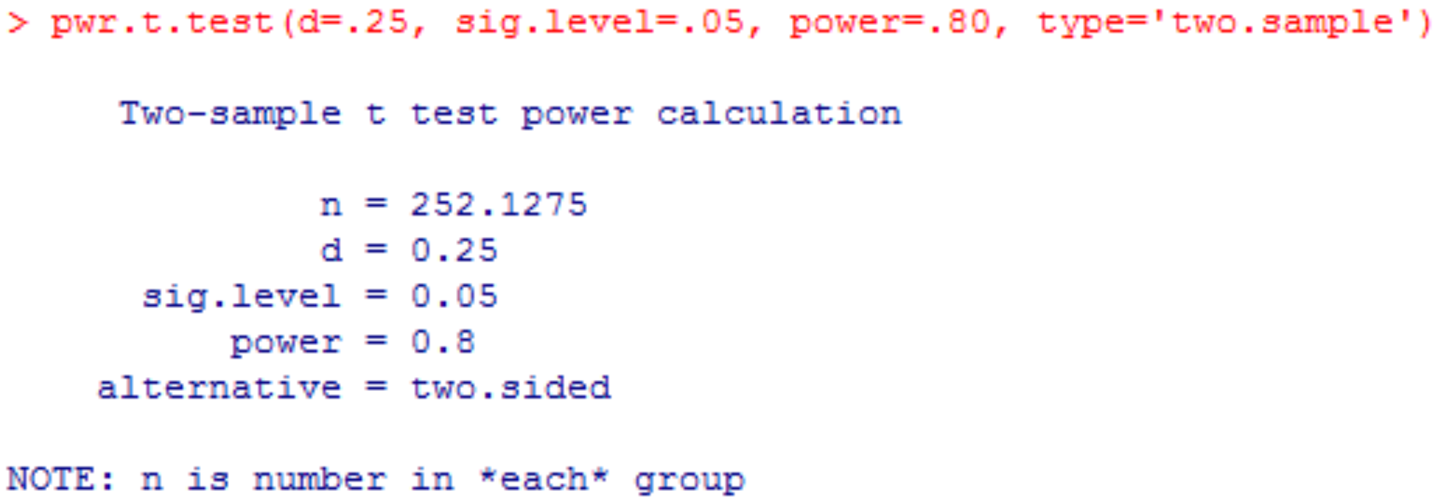
Power	80%	85%	90%	95%
Value	0.8416	1.0364	1.2816	1.6449

$$n = \frac{2\left(Z_\alpha + Z_{1-\beta}\right)^2 \sigma^2}{\delta^2} = \frac{2(1.96+0.84)^2\,(20)^2}{25} = 250.88 = 251$$

Hence we would use equal samples sizes of 251 for n_1 and n_2.

Solving in R

Install pwr package



```
> pwr.t.test(d=.25, sig.level=.05, power=.80, type='two.sample')

     Two-sample t test power calculation

              n = 252.1275
              d = 0.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

Sample **S**ize for **C**ategorical **D**ata

The sample size for categorical data when comparing two proportions is calculated using the following formula:

$$\boxed{n = \left(Z_{\alpha} + Z_{1-\beta}\right)^{2} \frac{p1(1-p1)+p2(1-p2)}{(p1-p2)^{2}}}$$

The number, n, is the sample size required in each group.

Example

A new treatment has been developed for patients who've had a heart attack. It is known that 10% of people who've suffered from a heart attack die within one year. It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect.

Let,

P_1 = proportion of deaths in placebo group = 0.1

P_2 = proportion of deaths in treatment group = 0.05.

Example

Then,

H_0: $P_1 = P_2$

H_a: $P_1 \neq P_2$

It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect. Using α = 0.05 and β = 0.10, What samples sizes (n_1) and (n_2) should they use ?

Example

Then,

H_0: $P_1 = P_2$

H_a: $P_1 \neq P_2$

It is thought that a reduction in deaths from 10% to 5% would be clinically important to detect. Using $\alpha = 0.05$ and $\beta = 0.10$, What samples sizes (n_1) and (n_2) should they use ?

Example

P1= 0.10

P2= 0.05

α-error	5%	1%
2-sided	1.96	2.5758
1-sided	1.65	2.33

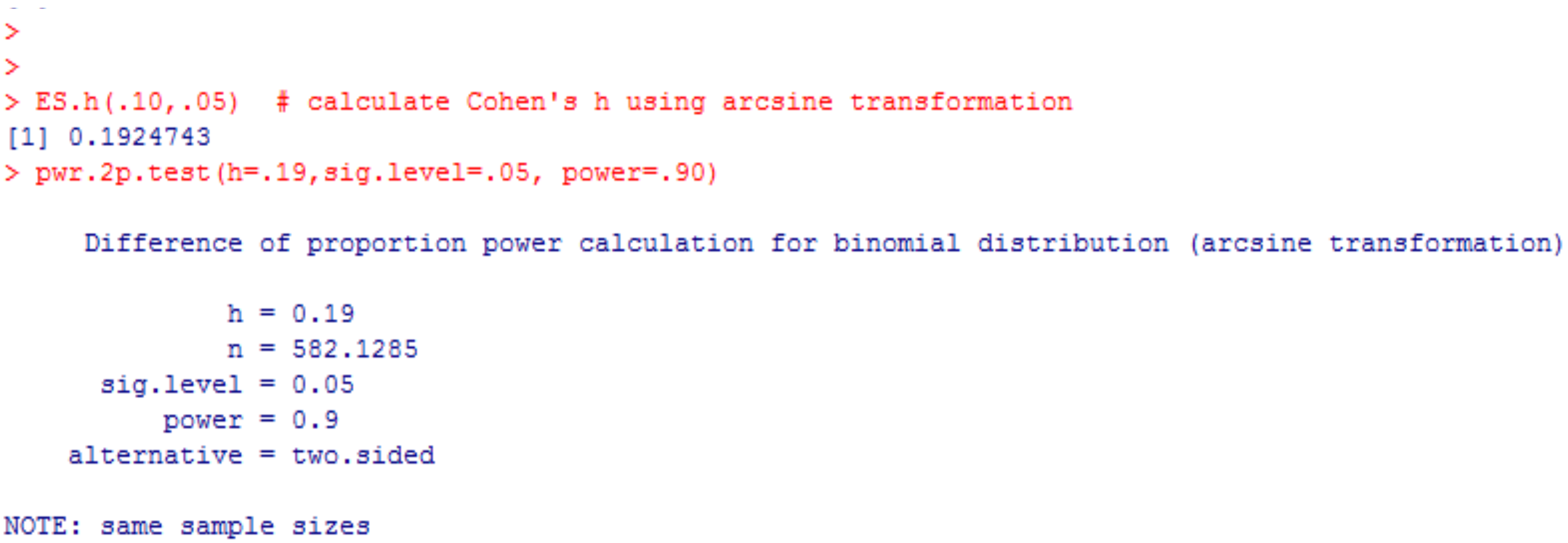
Power	80%	85%	90%	95%
Value	0.8416	1.0364	1.2816	1.6449

$$n = \left(Z_{\alpha} + Z_{1-\beta}\right)^{2} \frac{p1(1-p1)+p2(1-p2)}{(p1-p2)^{2}}$$

$$= (1.96 + 1.28\,)^{2} \frac{0.1\,(0.9)+0.05(0.95)}{(0.1-0.05)^{2}} = 10.5\frac{.09+.048}{.0025} = 579.6 = 580$$

580 patients would be needed in each group to be 90% sure of being able to detect a reduction in mortality of 5% as significant at the 5% level.

Solving in R



```
>
>
> ES.h(.10,.05)   # calculate Cohen's h using arcsine transformation
[1] 0.1924743
> pwr.2p.test(h=.19,sig.level=.05, power=.90)

     Difference of proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.19
              n = 582.1285
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: same sample sizes
```

Other **O**utcome **T**ypes

In many studies, the outcomes may not be continuous or categorical. In these cases, the details of calculation differ, but using the four aforementioned components, persist through calculations with other types of outcomes.

Summary

Definitions

α	Type I error :The risk of a false positive result. i.e. the chance of detecting a statistically significant difference when there is no real difference between treatments.
β	Type II error :The risk of a false negative result i.e. the chance of not detecting a significant difference when there really is a difference.
Power (1-β)	Power The chance of not getting a false negative result. i.e. the chance of spotting a difference as being statistically significant if there really is a difference.
The smallest effect of interest	The minimal difference between the groups that the investigator considers scientifically plausible and clinically relevant.
Variance	The variability of the outcome measure, expressed as the standard deviation.

Take Away

1. Explain the components of sample size calculations and how they affect the sample size
2. Explain the difference between Type I error, Type II error and power.
3. Explain effect size.
4. Calculate sample size for continuous and categorical data.