



浙江大学爱丁堡大学联合学院

**ZJU-UoE Institute**

# Power and sample size

ADS2, Lecture 2.4

Dr Rob Young – [robert.young@ed.ac.uk](mailto:robert.young@ed.ac.uk)

Semester 2, 2023/24

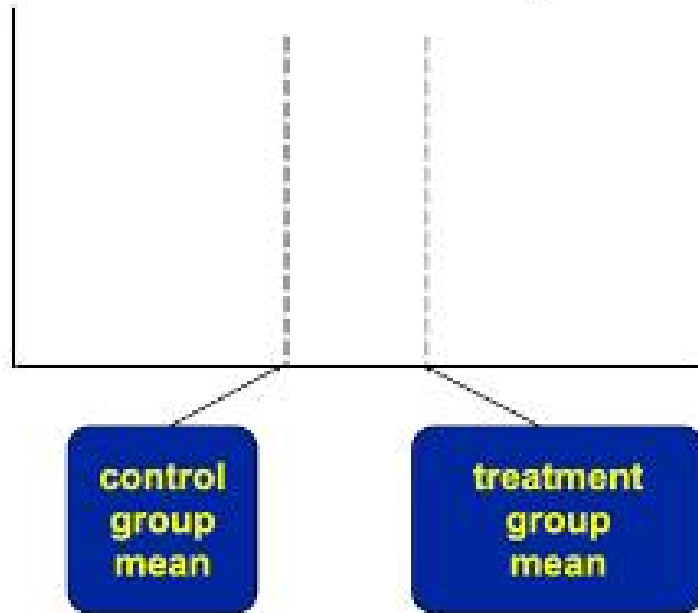
# Learning objectives

After this lecture, you should be able to:

- Define **statistical power**
- Explain how power relates to **sample size**
- Discuss **ethical issues** around power and sample size
- Use a simulation-based approach to **compute power**

# Why do we need to perform statistical tests?

## Statistical Analysis

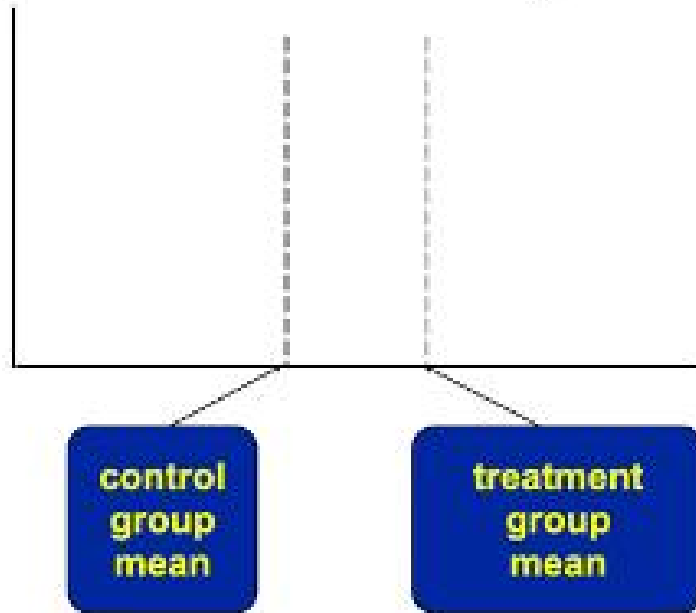


Without \_\_\_\_\_, there is no need for statistics.

Is there a *difference*?

# Why do we need to perform statistical tests?

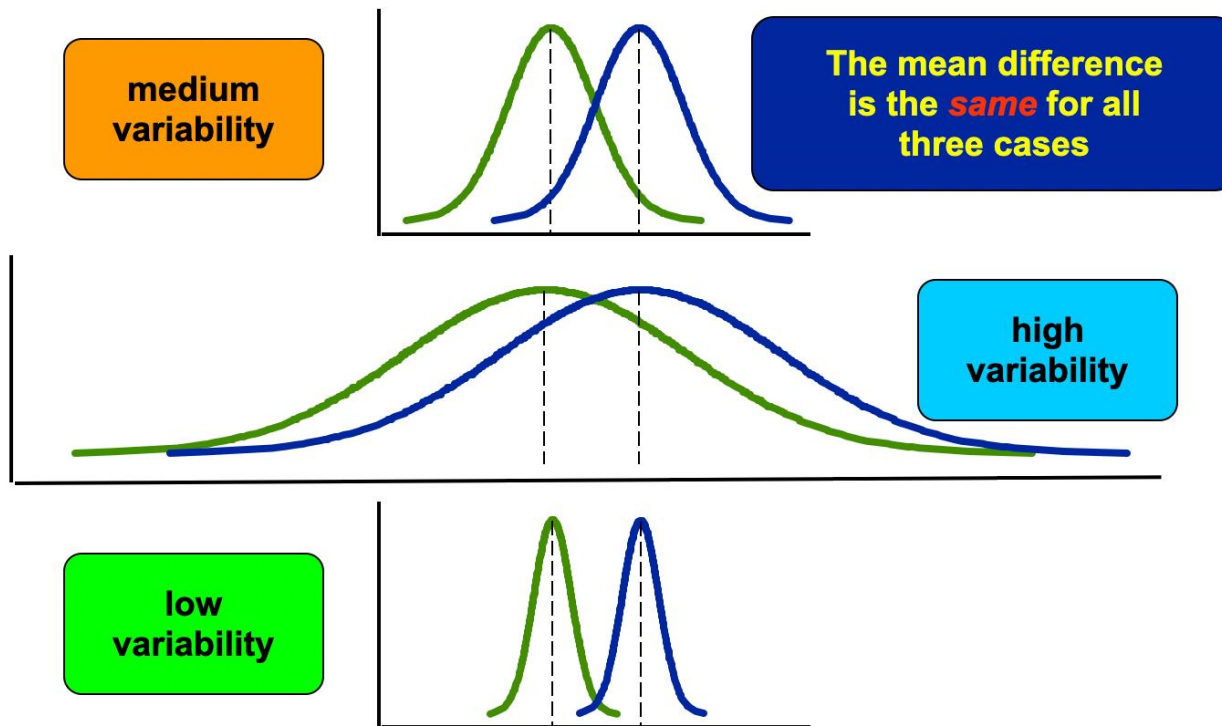
## Statistical Analysis



Without variability, there is no need for statistics.

Is there a *difference*?

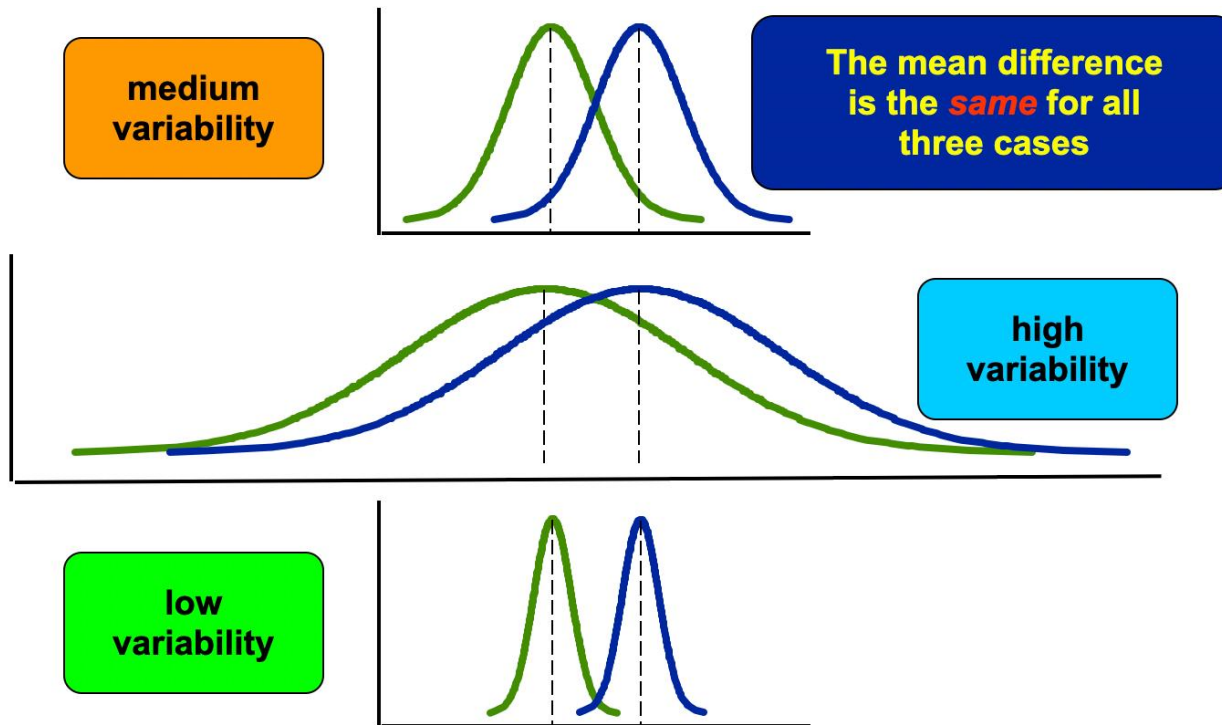
# Why do we need to perform statistical tests?



Without variability, there is no need for statistics.

# Why do we need to perform statistical tests?

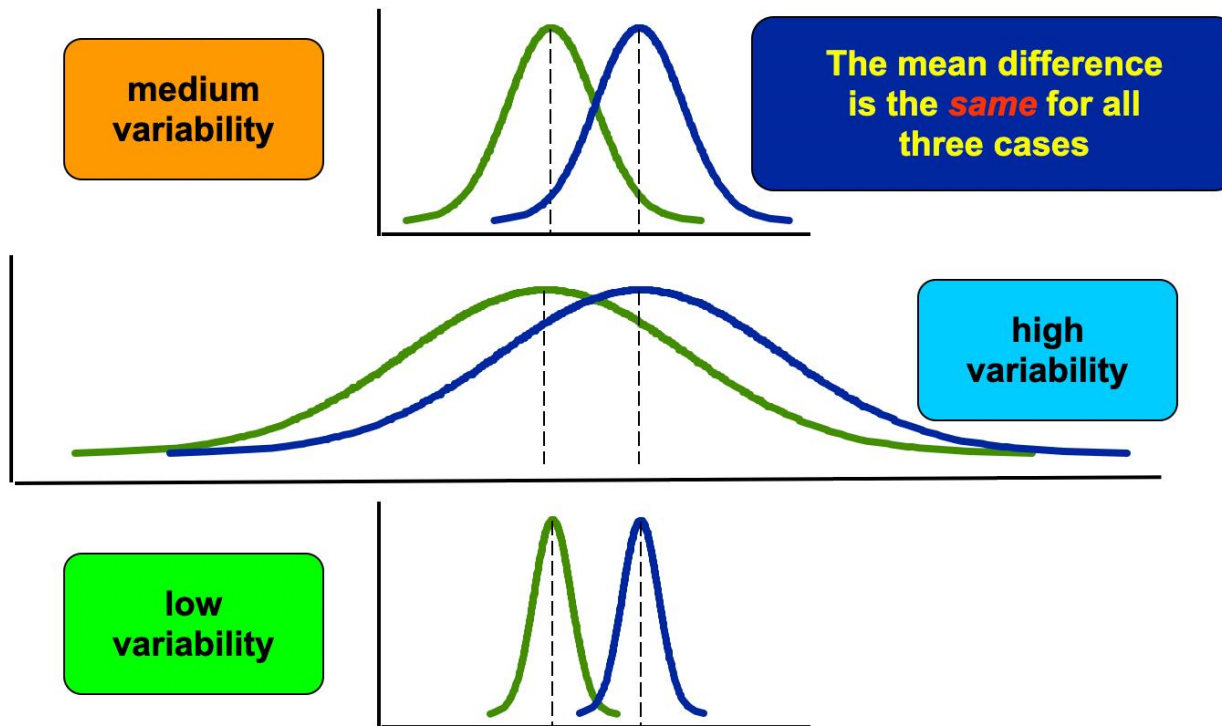
Same mean values, same sample sizes, different variabilities



In which case are we most likely to reject the null hypothesis?

# Why do we need to perform statistical tests?

Same mean values, same sample sizes, different variabilities



In which case are we most likely to reject the null hypothesis?

The lower the variability, the higher probability that we will reject the null hypothesis.

# What is statistical power?

- Statistical power is the **probability** that the statistical test will **reject a false null hypothesis**.
- ...or, in plain English:
- Statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected.





# Why do we care about statistical power?

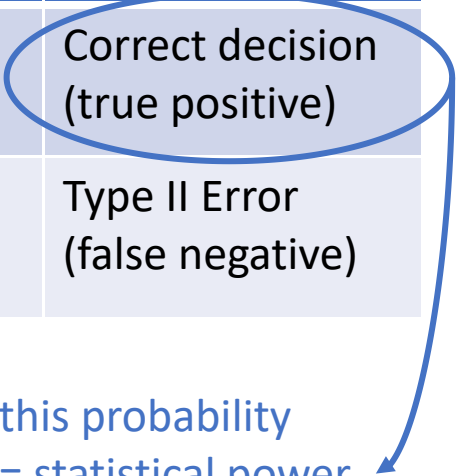
- **Lack of statistical significance** does not prove that there is no difference.
- Instead, it may be a consequence of **low power**.
- Do you remember what we call cases when we **reject a true null hypothesis**?

# Statistical power and $\beta$

## Review – Type I and Type II error

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I Error (false positive)	Correct decision (true positive)
Do not reject $H_0$	Correct decision (true negative)	Type II Error (false negative)

this probability  
= statistical power



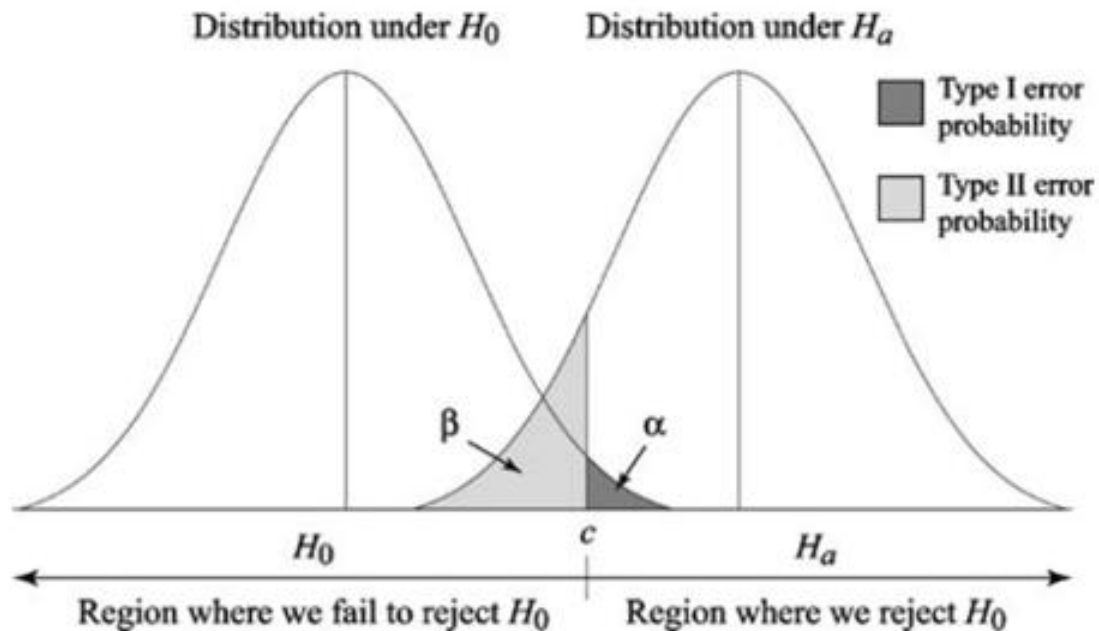
### Type I error

- A Type I Error is rejecting the null hypothesis when it is true.
- $\text{Prob}(\text{Type I Error}) = \text{Significance level } \alpha = P(\text{reject } H_0 | H_0 \text{ true})$

### Type II error

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

# Statistical power and $\beta$



What does power equal here?

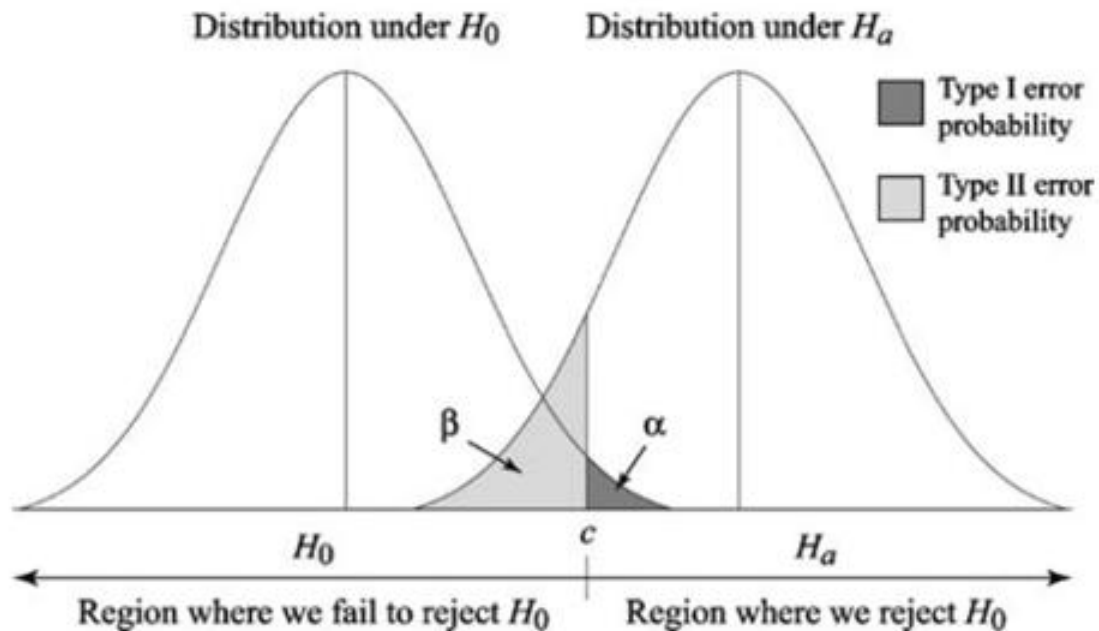
## Type I error

- A Type I Error is rejecting the null hypothesis when it is true.
- Prob(Type I Error) = Significance level  $\alpha$  =  $P(\text{reject } H_0 | H_0 \text{ true})$

## Type II error

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

# Statistical power and $\beta$



## Power of a hypothesis test

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

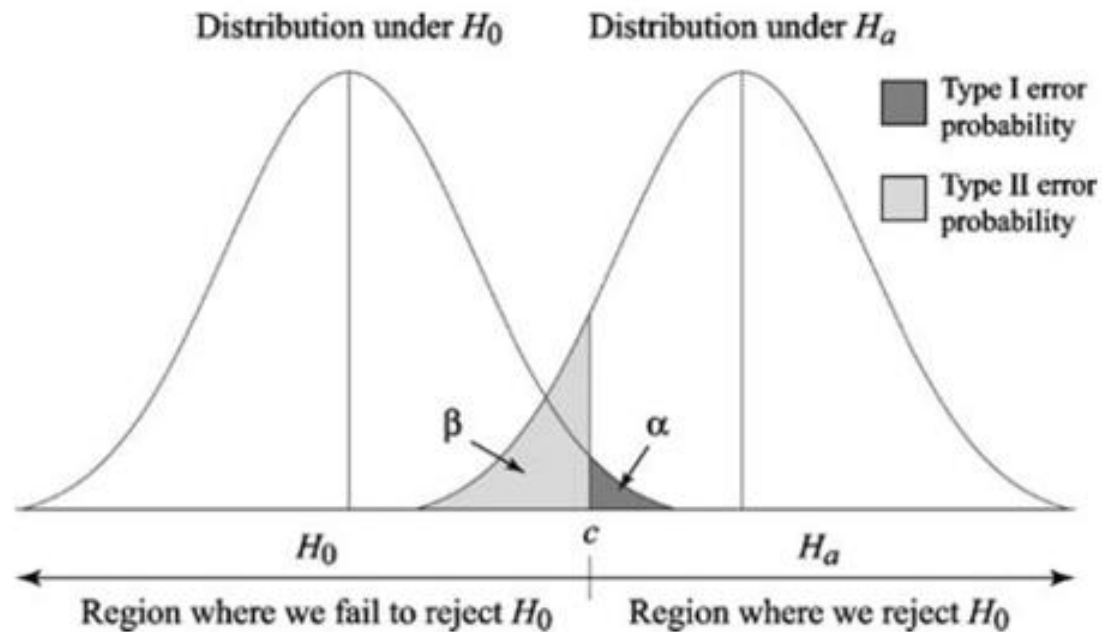
## Type I error

- A Type I Error is rejecting the null hypothesis when it is true.
- **Prob(Type I Error) = Significance level  $\alpha$  =  $P(\text{reject } H_0 | H_0 \text{ true})$**

## Type II error

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

# So what power do we need?



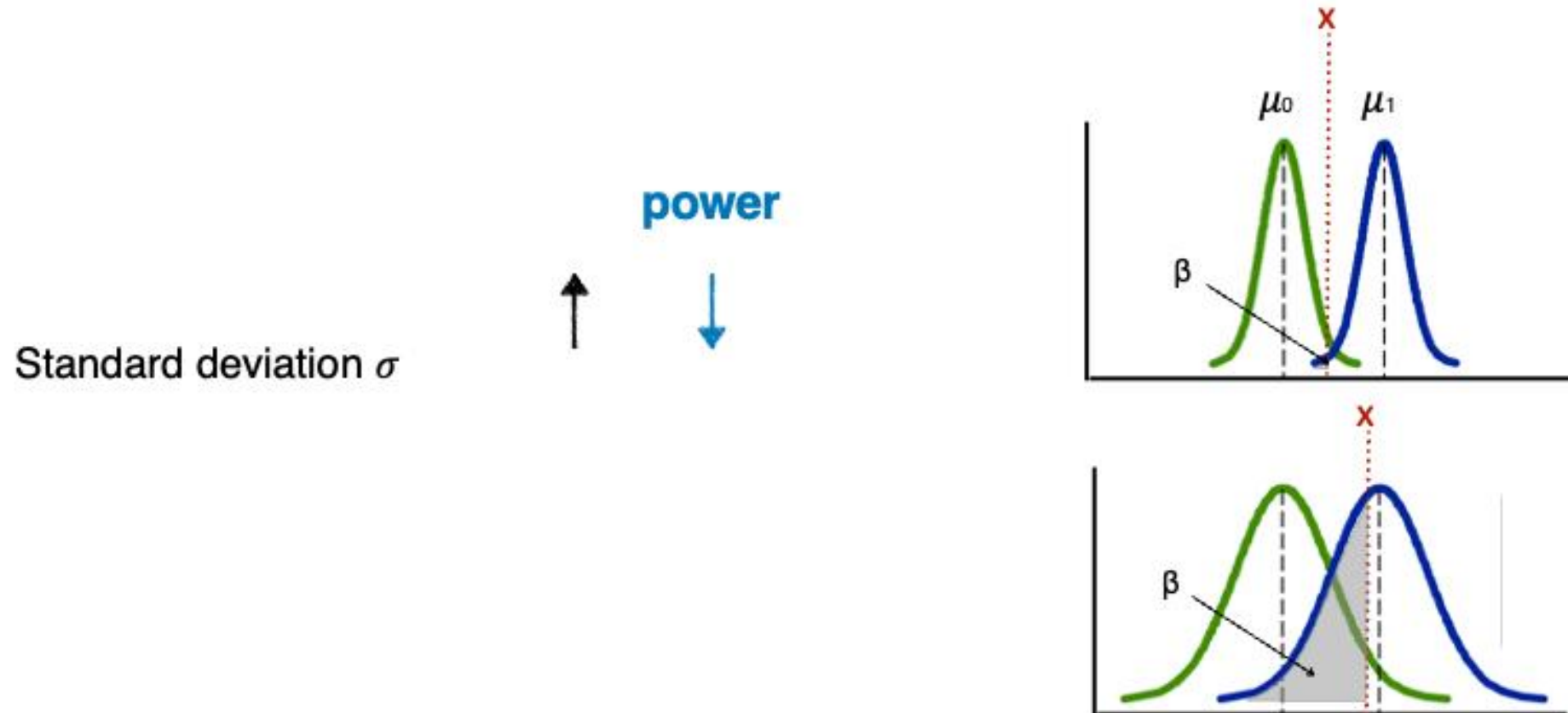
## Power of a hypothesis test

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

**It depends on the type of study!**

- In clinical trials, Phase III: industry minimum=80%
- Some say Type I Error=Type II Error
- Omics studies: aim for high power, because you want to minimise Type II error

# Factors affecting (statistical) power



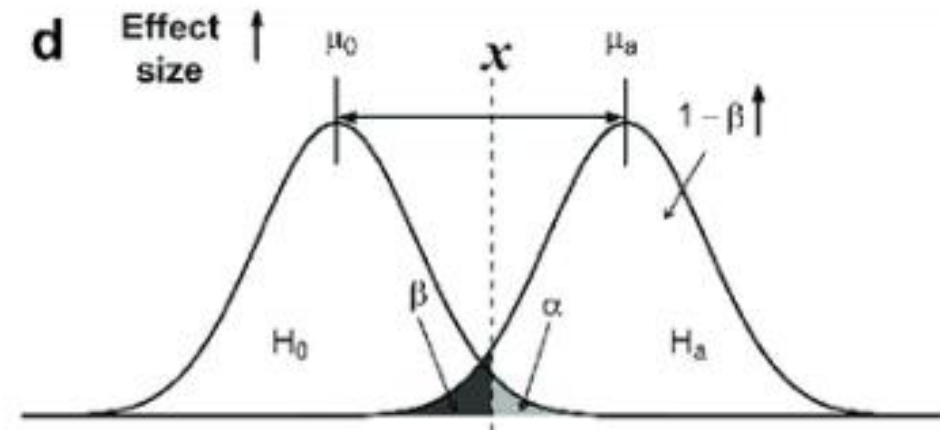
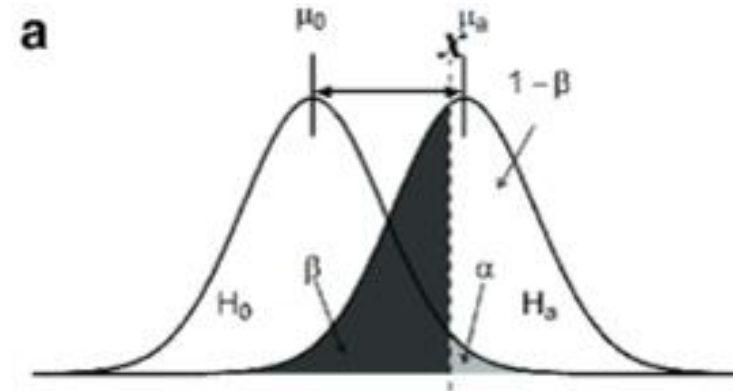
# Factors affecting (statistical) power



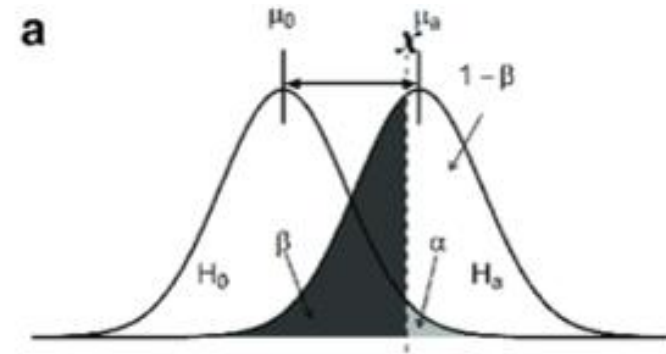
Effect size  $\Delta\mu$



power



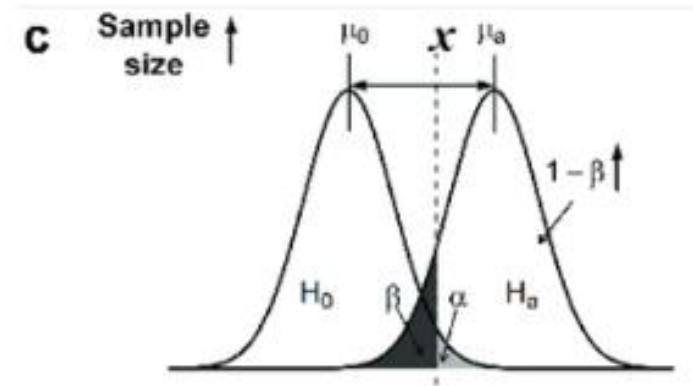
# Factors affecting (statistical) power



Sample size  $n$



power





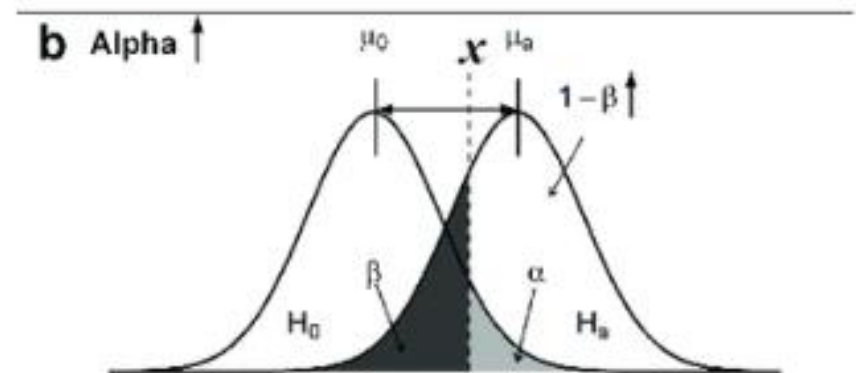
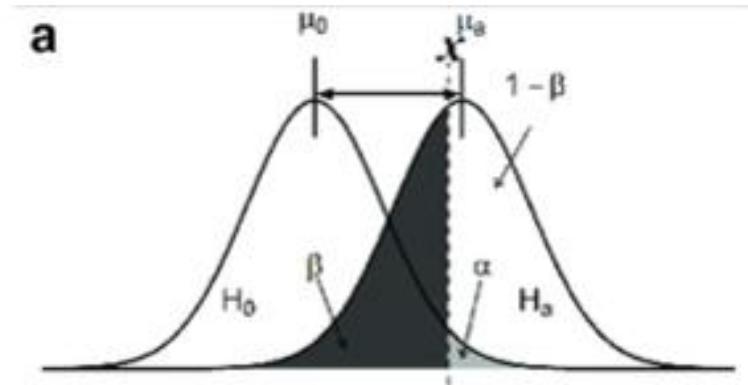
# Factors affecting (statistical) power



Significance level desired  $\alpha$



power



# Factors affecting (statistical) power



- power
- |                                        |   |   |
|----------------------------------------|---|---|
| 1) Effect size $\Delta\mu$             | ↑ | ↑ |
| 2) Standard deviation $\sigma$         | ↑ | ↓ |
| 3) Sample size $n$                     | ↑ | ↑ |
| 4) Significance level desired $\alpha$ | ↑ | ↑ |

In reality, which of the four factors can we change to increase the statistical power?

# Power and sample size



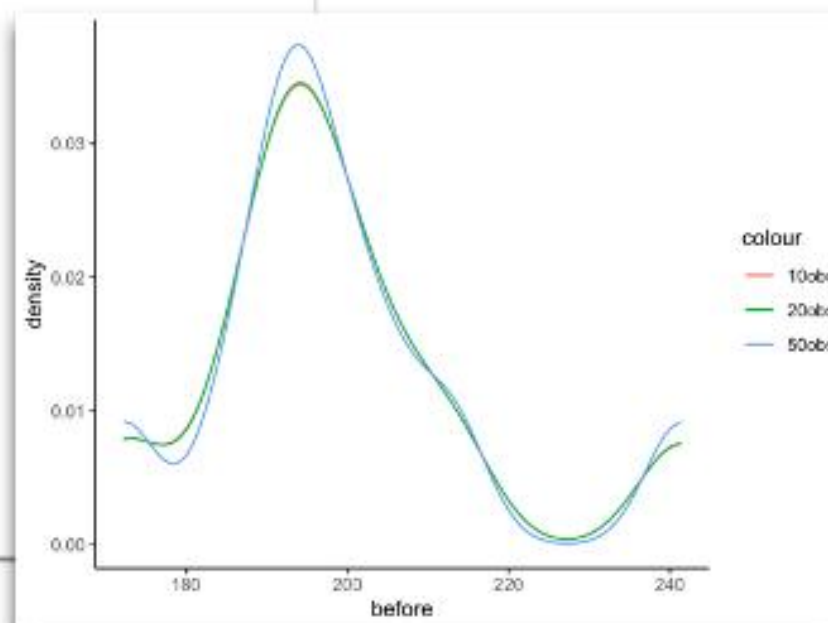
- In order to have **enough statistical power**, we need to have an **adequate sample size**.

## 'Mice weights' example of the 'Shapiro test'

```
mice_data= read.csv('Desktop/on_going_folder/teaching/ADS2/ADS2_9/mice_weights.txt')
before<- mice_data$before
before2<- rep(mice_data$before, 2)
before5<- rep(mice_data$before, 5)
```

```
ggplot() +
  geom_line(aes(x=before, col='10obs'), stat='density') +
  geom_line(aes(x=before2,col='20obs'), stat='density') +
  geom_line(aes(x=before5,col='50obs'), stat='density') +
  theme_classic()
```

```
shapiro.test(before)
shapiro.test(before2)
shapiro.test(before5)
```



# Power and sample size



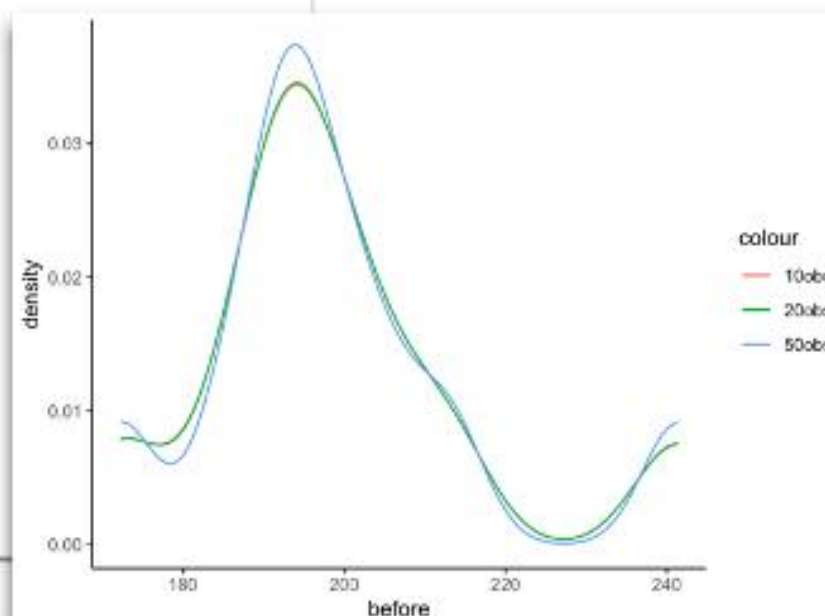
- In order to have **enough statistical power**, we need to have an **adequate sample size**.

## 'Mice weights' example of the 'Shapiro test'

```
mice_data = read.csv('Desktop/on_going_folder/teaching/ADS2/ADS2_9/mice_weights.txt')
before <- mice_data$before
before2 <- rep(mice_data$before, 2)
before5 <- rep(mice_data$before, 5)
```

```
ggplot() +
  geom_line(aes(x=before, col='10obs'), stat='density') +
  geom_line(aes(x=before2, col='20obs'), stat='density') +
  geom_line(aes(x=before5, col='50obs'), stat='density') +
  theme_classic()
```

```
shapiro.test(before)
shapiro.test(before2)
shapiro.test(before5)
```



```
> shapiro.test(before)

Shapiro-Wilk normality test

data: before n=10
W = 0.90938, p-value = 0.2768

> shapiro.test(before2)

Shapiro-Wilk normality test

data: before2 n=20
W = 0.8809, p-value = 0.01836

> shapiro.test(before5)

Shapiro-Wilk normality test

data: before5 n=50
W = 0.86849, p-value = 5.097e-05
```

# Choosing your power level – ethical issues

- What problems do you think we will have?
- **Too small a sample size with an underpowered study:**
- **Too big a sample size with an overpowered study:**

# Choosing your power level – ethical issues

- What problems do you think we will have?
- **Too small a sample size with an underpowered study:**
  - Waste resources; can't reject  $H_0$
  - Misleading conclusions if results are nonsignificant
  - Unethical if the conclusions lead to inferior treatment clinically
- **Too big a sample size with an overpowered study:**
  - Waste resources; especially for needless sacrifice of animals
  - Pick up essentially trivial results which are meaningless
  - Cost of collecting data > benefits

# Stopping rules in practice

- Stop after the power analysis, even before you collect samples:
  - If, to reach a decent power (~80%), you need to collect sample numbers that are impossible to achieve (money, time, ethical issues, etc).
- Stop at different stages of clinical trials:
  - Clinical trials are unusual in that enrolment of subjects is a continual process staggered in time.

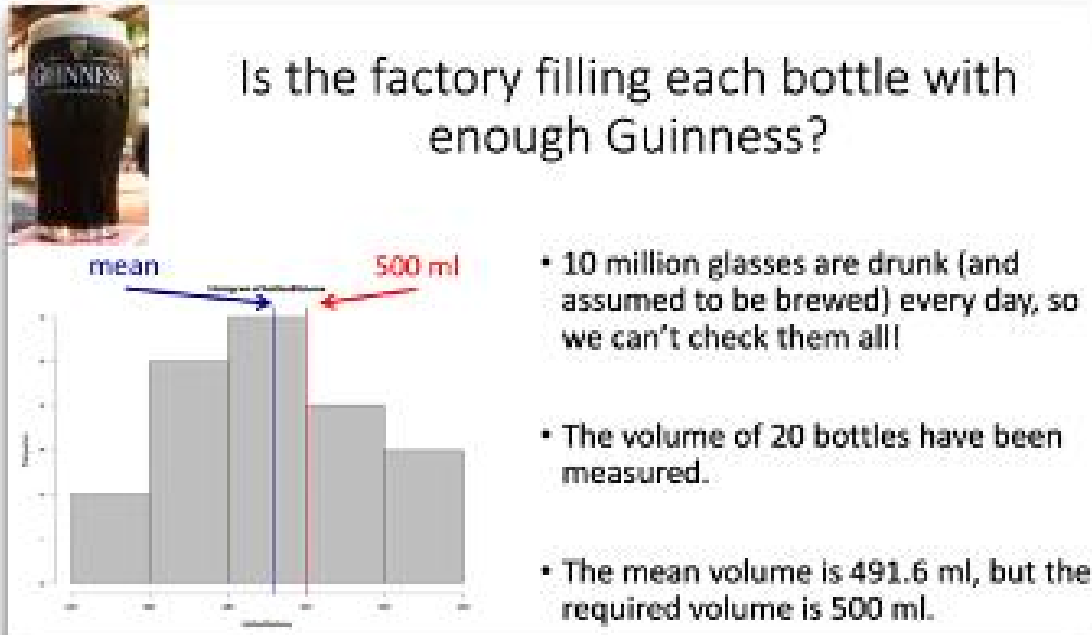
	Aim	Number	When to stop?
Phase I	Drug resistance, metabolism, side-effects, etc.	< 100 (normally 20-30)	If a treatment can be proven to be clearly <b>beneficial</b> or <b>harmful</b> compared to the concurrent <b>control</b> , or to be obviously futile, based on a predefined analysis of an incomplete data set while the study is ongoing, the investigators may <b>stop the study early</b> .
Phase II	Effectiveness in a defined group compared to the placebo	At least 100	
Phase III	Effective and safety compared to the existing drug	At least 300	
Phase IV	Safety	> 2,000	

# Choosing your power level

- It's all about a balance between risks ( $\alpha$  and  $\beta$ )
- Generally, a Type I error is considered worse so  $\alpha$  is rarely  $> 0.05$
- Generally, if we can tolerate a 5% type I error, we can tolerate a 20% type II error
- Large clinical trials use 0.9 or 0.95 (90-95% power,  $\beta = 0.1 - 0.05$ ).
- Animal studies usually use 0.8 (80% power,  $\beta = 0.2$ ).



# How to calculate power



$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{491.6 - 500}{24.8/\sqrt{20}} = -1.52$$

With df= 19, P-val= 0.07

The factory is NOT filling statistically significantly less Guinness.

Now you are probably wondering if this conclusion is due to a lack of statistical power...

...but...

Power calculations:

- Depend on the study design
- May not be hard, but can be very algebra intensive
- May want to use a computer program or consult a statistician (or ADS2 student!)

# Using simulation to calculate power

Our data:

$X = 491.6$

$\sigma = 24.8$

$n = 20$

$\mu = 500$

Power.t.test built-in function  
power.20=0.49581

```
power.t.test(n=20, delta = 9.4, sd =24.8, sig.level = 0.05, type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

```
      n = 20
  delta = 9.4
     sd = 24.8
sig.level = 0.05
   power = 0.4958127
alternative = one.sided
```

Which is more reliable, in  
your opinion?

```
ps.20 <- replicate(1e5, t.test(rnorm(20, 491.6,24.8), mu=500, alternative = "less")$p.value)
power.20<- length(which(ps.20 <= 0.05))/1e5
```

Simulation result in R  
power.20=0.42669

# Calculating the minimum sample size

Our data:  
 $X = 491.6$   
 $\sigma = 24.8$   
 $\text{power} = 0.8$   
 $n = 20$   
 $\mu = 500$

What is the minimum sample size that would allow us to have power of 80%, with  $\alpha = 0.05$ ?

```
power.t.test(power = 0.8, delta = 9.4, sd = 24.8, sig.level = 0.05, type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

```
n = 44.41896
delta = 9.4
sd = 24.8
sig.level = 0.05
power = 0.8
alternative = one.sided
```

```
> power.t.test(n=44, delta = 9.4, sd = 24.8, sig.level = 0.05, type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

```
n = 44
delta = 9.4
sd = 24.8
sig.level = 0.05
power = 0.7965842
alternative = one.sided
```

```
> power.t.test(n=45, delta = 9.4, sd = 24.8, sig.level = 0.05, type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

```
n = 45
delta = 9.4
sd = 24.8
sig.level = 0.05
power = 0.8046528
alternative = one.sided
```

Confirm by checking  $n=44$  and  $n=45$

# Learning objectives

Now, you should be able to:

- Define **statistical power**
  - The probability of rejecting the null hypothesis when it is false
- Explain how power relates to **sample size**
  - The bigger the sample size, the higher the power
- Discuss **ethical issues** around power and sample size
  - Under-powered or over-powered studied can be unethical, due to sample size-choices
- Use a simulation-based approach to **compute power**
  - For a given effect size (and standard deviation), power, and alpha, you can simulate to find the minimum 'n'



浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

# Power and sample size

## Any questions?

ADS2, Lecture 2.4

Dr Rob Young – [robert.young@ed.ac.uk](mailto:robert.young@ed.ac.uk)

Semester 2, 2023/24