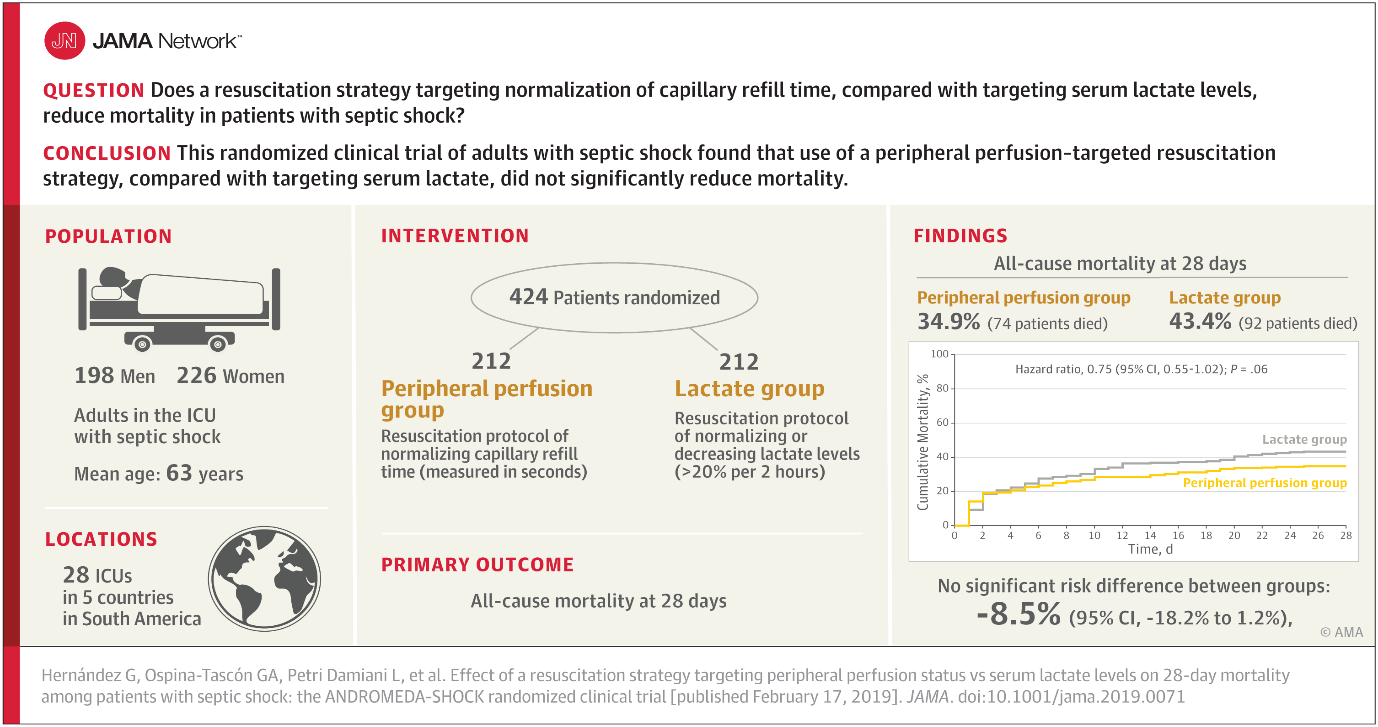
# Exercise 1

Consider the following abstract reporting a clinical trial of reducing mortality by peripheral perfusion compared to lactate among patients in intensive care with septic shock.



1. What was the mortality rate in the treatment arm? In the control arm?
2. What was the estimate for treatment effect?
3. What was its confidence interval? (check our understand of 95% CI)
4. What conclusion did the authors/journal draw? Why?
5. What conclusions do you draw from this trial?
6. If you or a loved one was in septic shock, which treatment would you prefer?
7. Would you allow yourself to be randomised to these treatments in a future trial?

The next exercises will cover sample size calculations for:

**Simple estimation**

Estimating a proportion (binary outcome, single sample)

Estimating a mean average (continuous outcome, single sample)

**Simple estimates/tests of treatment effects**

Comparing proportions between two groups (binary outcome, binary predictor)

Comparing means between groups (continuous outcome, binary predictor)

Estimating a correlation (continuous outcome and continuous predictor)

One-way ANOVA with more than two treatments (continuous outcome, multinomial predictor)

**More complex designs**

Estimating a treatment effect using a cross-over design (one group, estimating within-unit difference)

Estimating a treatment effect with extra covariates using a regression or analysis of covariance

For each situation we explore sample size calculations based on power or precision, and by either simulation or by analytic equations/functions using R.

Example code and discussion for how to conduct each of the calculations will be supplied.

# Exercise 2 – CoPS study

# Precision of the estimate of a proportion

Suppose we want to estimate the prevalence COVID in stool samples of COVID positive patients?

How many samples do we need?

Suppose we recruited 10 patients, and found two positive samples:

* Use binom.test() in R
* or the website <https://statpages.info/confint.html>

To find the estimate of prevalence and its 95% confidence interval.

What if we recruited 100 participants and found 20 positive samples?

Suppose we expect the prevalence to be about 20%.

How many samples will we need to estimate the prevalence to within +-5 percentage points.

How many samples should we collect for this study?

What do we need to think about when deciding on a sample size?

# Bonus election exercise:

How many voters would we need to sample to estimate the likely Liberal Democrat vote to within 2 percentage points in the forthcoming general election?

# Exercise 3 – The breads study

# Comparison of means of a continuous response between two groups

We are planning a clinical study of the glycaemic response of different breads.

Glycaemic response is a measure of how much blood sugar changes in response to eating a food product. Glycaemic response is measured by a single number (‘iAUC’, the area under a glucose response curve).

The iAUC is variable. It varies from person to person (some people have a naturally high glycaemic response) and it varies from occasion to occasion within the same person. That is, if I eat the same food on two different occasions my glycaemic response will differ between the two occasions just by chance.

Foods with a lower glycaemic response are generally considered to be healthier. We have developed a bread that we believe to have a lower GI than standard white bread while still being just as nice to eat.

We want to design a study to test that:

* The test bread *has a lower* iAUC than the control bread

We will design a study to test this.

In particular we will address the question: ***What sample size do we need?***

**Parallel Groups Study**

First we’ll consider a **parallel groups study** to test the difference in glycaemic response between breads.

* Two groups of N participants each are recruited.
* Group A gets the control bread, Group B gets the test bread.
* We measure the iAUC for each participant once after eating their bread
* We will estimate the treatment difference
* And do a hypothesis test for whether the test bread is different to the control

**Approach to sample size calculation**

Similarly to exercise 1, I will this study by simulation to see whether the new bread is better than the old!

We will do this using R / Shiny app.

<https://georgemsavva.shinyapps.io/powerSimulator/>

What information do we need to simulate this experiment?

**Consider:**

What could we conclude if we recruit only one person to each group?

How precise was the estimate of effect with 10 participants?

How does this change with increasing sample size?

How reliable was the p-value across experiments?

Did the experiments agree or disagree with respect to the efficacy of the bread?

How many participants was ‘enough’?

### Sample size calculation

*If we assume a standard deviation iAUC between individuals of 60 units – how many will we need to estimate the difference to within ±20 units?*

*How many participants would we need to recruit to have an 80% chance of getting a statistically significant result, at p<0.05.*

# Exercise 4 - What is statistical power?

What does it mean if an experiment has a power of, say, 80%?

What is an ‘under-powered’ study?

What would you learn from an underpowered study?

How would you know if a study that you are reading was underpowered?

# Exercise 5 – Logic of statistical inference

### Why do scientists ignore sample size issues?

Because we often have this paradigm:

1. Do experiment
2. Do analysis
3. If p>0.05 – we have proved there is no effect,
4. if p<0.05 – we have proved there is an effect

What’s wrong with this logic? Can you correct it?

# Exercise 6 - Precision for a two-group comparison of proportions:

*Suppose we are testing whether fecal microbial transplant (FMT) can cure ME/CFS.*

*How many people do we need for a two group trial?*

*We expect the baseline remission rate to be about 20%, and the treated remission rate to be about 40%. We would like to estimate the difference in absolute remission rate to within +- 10%, with 95% confidence.*

*You can use the Precisely web-app,* [*https://malcolmbarrett.shinyapps.io/precisely/*](https://malcolmbarrett.shinyapps.io/precisely/)

*or the `precisely` R package, or the formula for the standard error of a difference in proportions:*

*Where n is the number of participants in each group. Remember that a 95% confidence interval is about +- 2\*standard error.*

*If we could only recruit 15 people per group, how precise would our estimate of the treatment effect be?*

*Would this be enough?*

# Exercise 7. Power calculations for breads study

We will use pwr.t.test() to calculate the power for the experiment to test whether test bread is healthier than the control bread (in terms of iAUC).

Look at the help for pwr.t.test().

What information do you need to know before you can calculate the required sample size?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

For standard deviation – start with 60 units.

Use the pwr.t.test() function to calculate the sample size needed to detect a difference of at least iAUC=20 between two breads (two sided, with p<0.05), using an estimate of standard deviation from our analysis earlier (or 60sd if you don’t have that) with a power of 80%.

Explore how the required sample size changes with different estimates of the standard deviation of iAUC.

What is the smallest effect size you could reliably detect with your with an experiment of only 10 participants per group?

What if you could recruit 100 patients per group?

What is the power to detect an effect size of 50 units using a two group study with 10 participants per group?

Can you make a power curve showing how the sample size determines the detectable effect size? (difficult!)

# Exercise 8 How big should the ANDROMEDA-SHOCK trial have been?

1. What is the risk of death given the usual treatment (targeting serum lactate)?

*Risk of death in lactate group:*

1. What do you think might be the smallest clinically meaningful improvement in death risk.  
   (there’s no right or wrong answer to this, discuss amongst yourselves if needed)?

*Smallest risk difference:*

*Risk of death for the peripheral perfusion group in this case:*

1. Use the function ES.h() to calculate this effect size in terms of Cohen’s H statistic.

*Risk of death (control):*

*Risk of death (intervention):*

*Effect size: h =*

1. Use *pwr.2p.test()* in R to work out the required sample size to test this effect at 80% power at p<0.05.

*Sample size (per group):*

*Sample size (total):*

1. If you have time, repeat for different effect sizes. What magnitude of differences can you *reasonably* expect to be able to detect in a trial like this?
2. **Exercise:**

Consider the clinical trial of FMT in ME/CFS we discussed above.

Use pwr.2p.test() and ES.h() to find the sample size needed to detect an expected difference in prevalence from 40% to 20% in a parallel groups study, with 90% power at p<0.05.

# Exercise 9: Cross-over study.

Recall that iAUC varies within person and between people. If we were able use the same group of people for both the control *and* the test bread we remove the variability *between people* from the estimate for treatment effect (because the comparison is made within the individual, and not between individuals).

We also change the analysis from a two-sample t-test (test of whether one group differs from another) to a one-sample t-test (test of whether the average of the differences across people is greater than zero). (This is exactly the same as a ‘paired’ t-test).

This can have dramatic implications for the sample size calculation.

*Now our estimate for the treatment effect is the average of the treatment effects measured within individuals. The standard deviation of the difference for one individual is:*

Suppose the standard deviation of iAUC measurements for a single individual is 30 units (around half what the standard deviation would be if measures were taken in different people).

**Task:**

1. Look at the help for pwr.t.test() for how to perform a paired power calc.
2. Use pwr.t.test() to calculate the required sample size to detect a difference of 20 units in iAUC between the breads, with a power of 80% and a p-value of 0.05. How does this differ from the previous study?
3. How does the required sample size change if we want to use a critical p-value of or 0.01 or 0.001 instead of 0.05?

Note we now need to know (or be able to calculate or assume) the ICC, or the within-patient standard deviation!

# Exercise / Worked example 10: Longitudinal study

### Aim

We want to know whether (and by how much) four weeks on a new special diet leads to better short chain fatty acid (SCFA) concentration in the guts of healthy adults compared to a usual diet.

### Designs

Two possible experimental designs:

Design 1: (parallel)

Recruit two groups, treat with either usual diet or special diet for four weeks, measure SCFAs before and after the treatment.

There are three possible analyses

Analysis 1: Compare outcome (post- intervention SCFA) between groups with a t-test.

Analysis 2: Compare change scores (difference between pre- and post- treatment) between groups with a t-test

Analysis 3: Compare outcome between groups, adjusting for the pre-treatment measures scores.

Design 2: (cross-over)

Recruit one group, treat each individual with either: (a) usual diet for four weeks, then special diet for four weeks or (b) special diet for four weeks first then usual diet.

Analysis 4: Calculate the difference between meals for each person, then test the average difference using a paired t-test.

### Sample size calculation

Let’s think about how many patients would we need for each design.

Getting our inputs:

1. What information will we need before we can start for each approach?

Design 1 / Analysis 1

Design 1 / Analysis 2

Design 1 / Analysis 3

Design 2 / Analysis 4

Performing our power calculation:

Lets assume:

1. We would like to detect a difference of 20 mMol/L with 90% power at p<0.05
2. the standard deviation between individuals is 30mMol/L
3. the intra-class correlation between two observations in the same person is about 0.6.

Design 1: Simple two groups t.test comparing “SCFA at outcome”:

Design 1 / Analysis 2: Simple two groups t.test, comparing “change in SCFA”:

Design 1 / Analysis 3: ANCOVA or linear regression, comparing “SCFA at outcome conditional on SCFA at baseline”: (‘Superpower’ package)

Design 2 / Analysis 4: One sample t-test, is difference in SCFA between meals different from 0

# Exercise 11: One-way ANOVA The ‘Superpower’ R package

Not every hypothesis conforms to a simple superiority test between two groups.

But whatever your hypothesis/design there will be discussions of sample size in the literature.

For example, consider testing a 1-way ANOVA with four treatments. We want to show that:

1. There is a significant difference between treatments
2. There are significant differences between each pair of treatments.

What information will we need?

What assumptions might we have to make?

***Final words***

***The following are extracts from conversations I’ve had with scientists over the past couple of years. Hopefully you can now see why these are not appropriate or correct. I don’t mean this by way of criticism, one of these was me!***

*“It’s OK if we’re underpowered, the more things we test, the more chance we’ll find something!”*

*“Surely if I want to detect a bigger effect, I should need more samples not less!”*

*“But we always use N=3!”*

*“But [previous work] got p=0.048 with 3 mice per group! Why do I need 8?”*

*“Its OK, we’ll repeat the experiment to confirm it afterwards”*

*“I don’t care \*how\* different, I just want to show they \*are\* different…”*

*“If we see an effect even in an underpowered study, then it* ***must*** *be real”*