How to interpret statistics in scientific papers

Catherine Hewitt
Professor of Trials and Statistics
Deputy Director York Trials Unit
catherine.hewitt@york.ac.uk

Overview

- Whistle stop tour of how to interpret statistics in scientific papers
- Start with p-values and confidence intervals
- How to analyse categorical and continuous data
  - Parametric and non-parametric tests
  - Accounting for multiple variables
- Brief introduction to sample size calculations in trials

Making sense of the results

Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial


OPEN ACCESS

Matthew L Costa *professor of trauma and orthopaedic surgery*[1], Juul Achten *senior research fellow*[2], Nicholas R Parsons *trial statistician*[2], Richard P Edlin *senior lecturer in health economics*[3], Pedro Foguet *consultant orthopaedic surgeon*[4], Udai Prakash *consultant orthopaedic surgeon*[4], Damian R Griffin *professor of trauma and orthopaedic surgery*[2], Young Adult Hip Arthroplasty team

[1]Warwick Clinical Trials Unit, Division of Health Sciences, University of Warwick, Coventry CV4 7AL, UK; [2]Division of Health Sciences, Warwick Medical School, University of Warwick; [3]Health Systems, School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand; [4]University Hospitals Coventry and Warwickshire NHS trust, Coventry

Study design

- Randomised controlled trial
  - 60 patients received resurfacing arthroplasty
  - 66 patients received total hip arthroplasty

Making sense of the results

Results Intention to treat analysis showed no evidence for a difference in hip function between treatment groups at 12 months (P=0.242 and P=0.070 for Oxford hip score and Harris hip score, respectively). Mean Oxford hip score was 40.4 in the resurfacing group and 38.2 in the total arthroplasty group (estimated treatment effect size 2.23 95% confidence interval −1.52 to 5.98).

Making sense of the results

Results Intention to treat analysis showed no evidence for a difference in hip function between treatment groups at 12 months (P=0.242 and P=0.070 for Oxford hip score and Harris hip score, respectively). Mean Oxford hip score was 40.4 in the resurfacing group and 38.2 in the total arthroplasty group (estimated treatment effect size 2.23 95% confidence interval −1.52 to 5.98).

What do these things mean?

Making sense of the results

- There are two methods of ‘statistical inference’:
  - Confidence intervals for an estimate
  - P value for a significance test

Making sense of the results

- The data we have are from a **sample** from a much larger **population**
- **Sample**: people in your study
- **Population**: all people of interest (e.g. with a particular disease) both present and future
- We want to use the **sample** to tell us about the **population**

Samples and populations

- **Problem:** we have a sample, would another sample give us a different answer?
  - In Matt's arthroplasty trial the difference in the Oxford hip score was found to be 2.23 95% confidence interval −1.52 to 5.98
- Would another sample give the same difference of 2.23 points?
- Might it give 2 or 5 or even -5 points?
- The **confidence interval** helps us to deal with this problem

Samples and populations

Oxford hip score difference: 2.23 95% CI–1.52 to 5.98

- In the **sample** the difference is 2.23 points
- We want to know the difference in the **population**
- We cannot know exactly what it is
- The **sample** difference is only an estimate of the difference in the **population**
- We estimate that, in the **population** the difference is somewhere between -1.52 to 5.98

Samples and populations

- Why 95% confidence interval?
- We choose the interval so that for 95% of the possible samples which we could take, of which this is just one, the interval would include the population value
- 95% of confidence intervals include their population value
  - Conversely, 5% do not

Samples and populations

Results Intention to treat analysis showed no evidence for a difference in hip function between treatment groups at 12 months (P=0.242 and P=0.070 for Oxford hip score and Harris hip score, respectively). Mean Oxford hip score was 40.4 in the resurfacing group and 38.2 in the total arthroplasty group (estimated treatment effect size 2.23 95% confidence interval −1.52 to 5.98).

The p-value is an indicator of the strength of evidence which this sample provides

P-values

- P is a probability between 0 and 1
- Small P → strong evidence = statistically significant
- Large P → weak evidence = not statistically significant
- Usual cut off for decision is: P=0.05
- If P>0.05 we say that the difference is not significant
  - This does not mean that there is no difference; we have not found evidence for a difference

P-values

- In the example, P=0.242 for the Oxford hip score
- A difference as big as 2.23 (or bigger) is likely to occur 24 in 100 times if the difference in the population were zero (i.e. quite likely)
- Usual choices:
  - P>0.05 -> not significant
  - P≤0.05 -> significant
  - P<0.01 -> highly significant
  - P<0.001 -> very highly significant

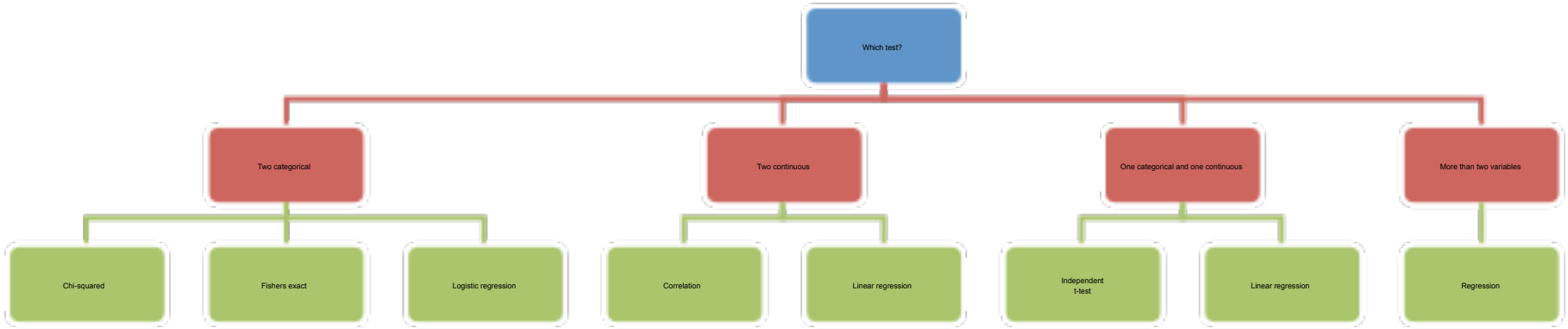
Samples and populations

- Significance tests and confidence intervals are two ways to make the link between sample and population
- If a confidence interval can be found, it conveys more information so should be reported
  - It is the approach recommended by CONSORT
- BUT, we cannot always find a confidence interval; we can almost always do a significance test though

How decide which test?

- What question are you trying to answer?
- What types of variables do you have?
  - Continuous (or quantitative)
    - Oxford hip score
  - Categorical (or qualitative)
    - Complications
- The types of variables you have determine the statistical approach you use

How do we decide which test?



Comparing outcomes in published research

- Costa et al. (2012), BMJ

Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial

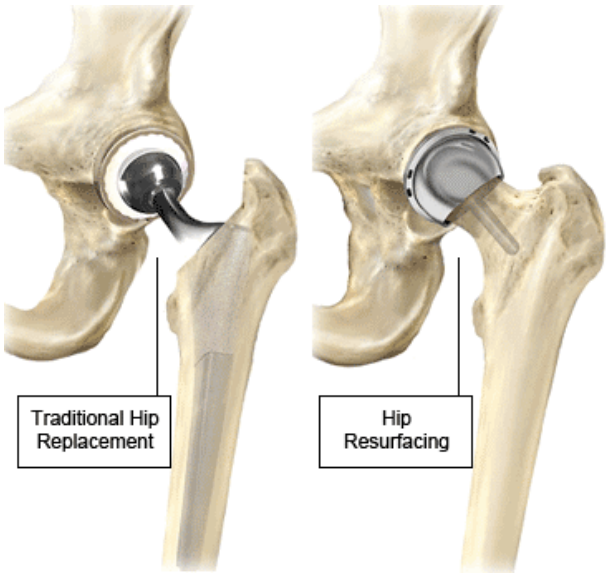

OPEN ACCESS

Matthew L Costa *professor of trauma and orthopaedic surgery*[1], Juul Achten *senior research fellow*[2], Nicholas R Parsons *trial statistician*[2], Richard P Edlin *senior lecturer in health economics*[3], Pedro Foguet *consultant orthopaedic surgeon*[4], Udai Prakash *consultant orthopaedic surgeon*[4], Damian R Griffin *professor of trauma and orthopaedic surgery*[2], Young Adult Hip Arthroplasty team

[1]Warwick Clinical Trials Unit, Division of Health Sciences, University of Warwick, Coventry CV4 7AL, UK; [2]Division of Health Sciences, Warwick Medical School, University of Warwick; [3]Health Systems, School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand; [4]University Hospitals Coventry and Warwickshire NHS trust, Coventry

Comparing proportions in published research

- Outcome
  - Complications
- Results
  - Resurfacing arthroplasty
    - Total complications: 11/60 (18.3%)
  - Total hip arthroplasty
    - Total complications: 18/66 (27.3%)
- So??



Application to research example

- Complications at 12 months
  - Resurfacing arthroplasty: 18.3%
  - Total hip arthroplasty: 27.3%
- Hypotheses (inference at the population level)
  - H0: No association between complications and treatment
  - HA: An association between complications and treatment
- Difference in proportion
  - 18.3% vs 27.3%
  - p = 0.291 (cannot reject H0)
  - *So how did we arrive at that p-value??*

Fishers exact test - reporting

- Could have used Chi-squared test but not valid in this case as sample size small
- Fisher's exact test was used:

  "Overall complications rates did not differ between treatment groups (Fisher's exact test, P=0.291; table 3). However, we saw more superficial wound complications in the total arthroplasty group (P=0.056) and more thromboembolic events in the resurfacing arthroplasty group (P=0.049)"

Comparing outcomes in published research

- Costa et al. (2014), BMJ

Percutaneous fixation with Kirschner wires versus volar locking plate fixation in adults with dorsally displaced fracture of distal radius: randomised controlled trial


OPEN ACCESS

Matthew L Costa *professor of trauma and orthopaedic surgery*[1 2], Juul Achten *senior research fellow*[1 2], Nick R Parsons *medical statistician*[3], Amar Rangan *professor of trauma and orthopaedic surgery*[4], Damian Griffin *professor of trauma and orthopaedic surgery*[1 2], Sandy Tubeuf *lecturer in health economy*[5], Sarah E Lamb *professor of rehabilitation*[6], on behalf of the DRAFFT Study Group

[1]Warwick Clinical Trials Unit, University of Warwick, Coventry CV4 7AL, UK; [2]University Hospitals Coventry and Warwickshire NHS Trust, Coventry CV2 2DX, UK; [3]Clinical Sciences Research Laboratories, University of Warwick, Coventry CV2 2DX, UK; [4]Wolfson Research Institute, School of Medicine and Health, Durham University, Queen's Campus, Stockton-on-Tees TS17 6BH, UK; [5]Leeds Institute of Health Sciences, Leeds University, Leeds LS2 9LJ, UK; [6]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Nuffield Orthopaedic Centre, Oxford OX3 7HE, UK

Comparing proportions in published research

DRAFFT trial: wires vs. plates for distal radius fractures

Outcome

- Perioperative antibiotic use

Results

- Wires
  - 165/230 (71%)
- Plates
  - 192/231 (83%)

So??

Application to research example

- Perioperative antibiotic use
  - Wires: 71%
  - Plates: 83%
- Hypotheses (inference at the population level)
  - H0: No association between antibiotic use and treatment
  - HA: An association between antibiotic use and treatment
- Difference in proportion
  - 71% vs 83%
  - Odds ratio = 3.5 (2.0 to 6.5)
  - Confidence intervals does not include 1 and $p<0.001$(reject H0)
  - *So how did we arrive at that p-value??*

Logistic regression – reporting

- *"The rate of peri-operative antibiotic use was higher in the plate group than in the Kirschner wire group; 83% v 71% of study participants were prescribed antibiotics (estimated odds ratio 3.5, 95% confidence interval 2.0 to 6.5, P<0.001"*

- Assumption: binary outcome and at least 10 observations with a 'yes' outcome and 10 observations with a 'no' outcome per variable

Comparing outcomes in published research

- Costa et al. (2012), BMJ

Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial

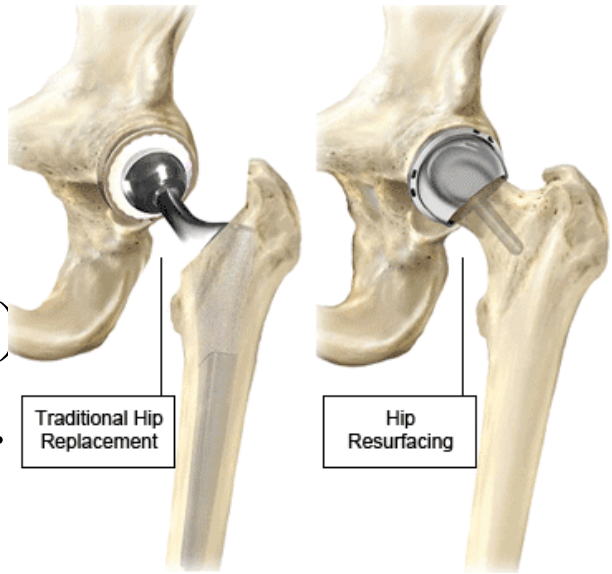

OPEN ACCESS

Matthew L Costa *professor of trauma and orthopaedic surgery*[1], Juul Achten *senior research fellow*[2], Nicholas R Parsons *trial statistician*[2], Richard P Edlin *senior lecturer in health economics*[3], Pedro Foguet *consultant orthopaedic surgeon*[4], Udai Prakash *consultant orthopaedic surgeon*[4], Damian R Griffin *professor of trauma and orthopaedic surgery*[2], Young Adult Hip Arthroplasty team

[1]Warwick Clinical Trials Unit, Division of Health Sciences, University of Warwick, Coventry CV4 7AL, UK; [2]Division of Health Sciences, Warwick Medical School, University of Warwick; [3]Health Systems, School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand; [4]University Hospitals Coventry and Warwickshire NHS trust, Coventry

Comparing means in published research

- Outcome
  - Hip functioning at 12 months after surgery (Harris hip score)
- Results
  - Resurfacing arthroplasty
    - Before surgery: Mean=48.6 (SD=14.2)
    - After 12 months: Mean=88.4 (SD=15.
  - Total hip arthroplasty
    - Before surgery: Mean=50.1 (SD=13.5)
    - After 12 months: Mean=82.3 (SD=21.4)
- So??



Application to research example

- Mean Harris Hip Score at 12 months
  - Resurfacing arthroplasty: 88.4
  - Total hip arthroplasty: 82.3
- Hypotheses (inference at the population level)
  - H0: Hip functioning is the same for the two treatments
  - HA: Hip functioning is different between the two treatments
- Group Difference
  - Observed Mean Difference = 6.04 score points
  - H0: Mean Difference = 0
  - Likelihood of 6.04 if H0 is true: p = 0.070 (cannot reject H0)
  - *So how did we arrive at that p-value??*