Statistics made easy:

Understanding basic statistics in papers & presentations

Colette Smith

17th Annual NHIVNA conference, Leeds
Thursday 18th June 2015

"Understanding basic statistics"….

- This session will focus on understanding:
  - Hypothesis testing and P-values
  - Confidence Intervals
  - (If time) Risk Ratios and Rate Ratios
- Example:
  - Strategic Timing of AntiRetroviral Treatment (START): results released early on 27 May 2015[1]

[1]http://www.niaid.nih.gov/news/newsreleases/2015/Pages/START.aspx

Hypothesis testing, *P*-values and confidence intervals

Background

- Presentations of data in the medical world are littered with p-values - '*P*<0.05' is thought to be a magical phrase, guaranteed to ensure that your paper will be published
- But what do these *P*-values really tell us, and is a *P*-value <0.05 really that important?
- Why is it important to also show confidence intervals - what additional information do they provide?

Example

- Two drugs (A and B) are compared in a RCT. The response rates in each group are:

	(a)
Drug A	3/10
Drug B	6/10

- Assuming all other factors are similar (e.g. side effects etc.) do you believe that drug B is more effective than drug A?

Example

- Two drugs (A and B) are compared in a RCT. The response rates in each group are:

	(a)	(b)
Drug A	3/10	30/100
Drug B	6/10	60/100

- Assuming all other factors are similar (e.g. side effects etc.) do you believe that drug B is more effective than drug A?

Example

- Two drugs (A and B) are compared in a RCT. The response rates in each group are:

	(a)	(b)	(c)
Drug A	3/10	30/100	300/1000
Drug B	6/10	60/100	600/1000

- Assuming all other factors are similar (e.g. side effects etc.) do you believe that drug B is more effective than drug A?

Why do we need statistical tests?

- When sample sizes are large, it may be reasonable to assume that the results are genuine and not simply a chance finding
- However, as the sample size decreases, it is hard to know whether any observed differences are genuine
- We need a way to formally assess whether the results we see reflect a genuine difference in drug efficacy, or are simply the results of random fluctuation
- P-values are usually calculated to help us make comparisons between groups

What is the *P*-value?

- *P*-value: probability of obtaining an effect at least as big as that observed if the null hypothesis is true (i.e. there is no real effect)
- Large *P*-value – results are consistent with chance variation
  - *Insufficient evidence that effect is real*
- Small *P*-value – results are inconsistent with chance variation
  - *Sufficient evidence that effect is real*

What is large and what is small?

By convention:

$P<0.05$ – SMALL

$P>0.05$ – LARGE

The general approach to court cases

- Start by defining two hypotheses:
  - **Null hypothesis (H_0):** The suspect is innocent
  - **Alternative hypothesis (H_1):** The suspect is guilty
- Conduct trial and present evidence
- Jury weighs up evidence from the trial against the null hypothesis
- Obtain a verdict

The general approach to hypothesis testing

- Start by defining two hypotheses:
  - **Null hypothesis (H_0):** There is no real difference in viral load response rates between two regimens
  - **Alternative hypothesis (H_1):** There is a real difference in viral load response rates between two regimens
- Conduct trial and collect data
- Use data from that trial to perform a hypothesis test (e.g. Chi-squared test, t-test, ANOVA)
- Obtain a *P*-value

Choosing the right hypothesis test

Tests that may be used (a small selection):

Comparing proportions

- Chi-squared test
- Chi-squared test for trend
- Fisher's exact test
- Relative risk
- Odds ratio

Comparing numbers

- Unpaired *t*-test
- Paired *t*-test
- Mann-Whitney U test
- ANOVA
- Kruskal-Wallis test

Example – the Chi-squared test

	VL≤50 copies/ml	VL >50 copies/ml	Total
Regimen	N (%)	N (%)	N (%)
A	28 (52)	26 (48)	54 (100)
B	22 (48)	24 (52)	46 (100)
Total	50 (50)	50 (50)	100 (100)

Is regimen A (new regimen) better than regimen B?

Example – i) Define hypotheses

We wish to know whether patients receiving a new treatment regimen (A) are more likely to achieve viral load suppression than those receiving standard-of-care (B)

Hypotheses:

H_0: There is no real difference in the proportion of people with a VL≤50 copies/ml between those receiving regimen A and those receiving regimen B

H_1: There is a real difference in the proportion of people with a VL≤50 copies/ml between those receiving regimen A and those receiving regimen B

Example – the Chi-squared test

	VL≤50 copies/ml	VL >50 copies/ml	Total
Regimen	N (%)	N (%)	N (%)
A	28 (52)	26 (48)	54 (100)
B	22 (48)	24 (52)	46 (100)
Total	50 (50)	50 (50)	100 (100)

Example – the Chi-squared test

- Computer output gives p-value of 0.84
- If there really was no difference in viral load response between the two groups, and we repeated the study 100 times, we would have observed a difference of this size (or greater) on 84 of the 100 occasions
- As $P>0.05$, there is insufficient evidence of a real difference in viral load response rates between the two regimens

Points to note

- We have not proven that the difference was due to chance, just that there was a reasonable probability that it might have been
- We can never prove the null hypothesis
- We take an 'innocent until proven guilty' approach

Treatment effects

- *P*-values by themselves are of limited value
- Although they give an indication of whether the findings are likely to be genuine, they do not allow you to put findings into clinical context
- Should provide an estimate of the effect of interest (i.e. some comparative effect) as well as an indication of the precision of the estimate (i.e. its 95% confidence interval)

Treatment effects

- The 'treatment effect' ('risk difference' or 'absolute risk reduction') is the additional benefit that the new drug/regimen provides compared to 'standard of care'
- Example:
  - Drug A (new regimen) 80% response
  - Drug B (standard of care) 68% response
- The treatment effect is 12% (= 80% - 68%)

How do we interpret trial outcomes?

- Estimate of 12% was a point estimate; this is our 'best guess' but it gives no indication of variability
- Confidence intervals provide a range of additional plausible values that are supported by the results of the study – they indicate the precision of the estimate
- In a trial, the 95% CI for the treatment effect allows us to put the results from the trial into clinical context; can weigh up benefits in light of any disadvantages of drug (e.g. increased cost or worse toxicity profile)

Example

Trial number	Drug				
	A		B		
	n	n (%) responding	n	n (%) responding	Difference (B – A)
1	50	34 (68)	50	40 (80)	12%

- We believe that drug B is 12% more effective than Drug A
- The 95% CI for this estimate is: -5.0% to +29.0%
- Drug B could be up to 5% *less effective* than drug A, or up to 29% *more effective* than drug A
- What are your views about drug B?

Example

Trial number	Drug				
	A		B		
	n	n (%) responding	n	n (%) responding	Difference (B – A)
1	150	102 (68)	150	120 (80)	12%

- We believe that drug B is 12% more effective than Drug A
- The 95% CI for this estimate is: 2.2% to 21.8%
- Drug B could be as little as 2% *more effective* or as much as 22% *more effective* than drug A
- What are your views about drug B?

Precise vs imprecise estimates

- First confidence interval was too wide to allow us to judge whether drug B was better, worse or the same as drug A
- The estimate was imprecise, or lacked precision
- Second confidence interval was narrower, allowing us to conclude that drug B was likely to be better than drug A
- The estimate from this trial was more precise
- Major determinant of width of CI is the sample size

Other points

- Although we have focussed on confidence intervals for the difference in two proportions, they can be generated for almost every statistic
- Calculations may be tricky, but most statistical packages will generate them automatically
- Most journals now require that confidence intervals are provided for all treatment effects reported in a paper

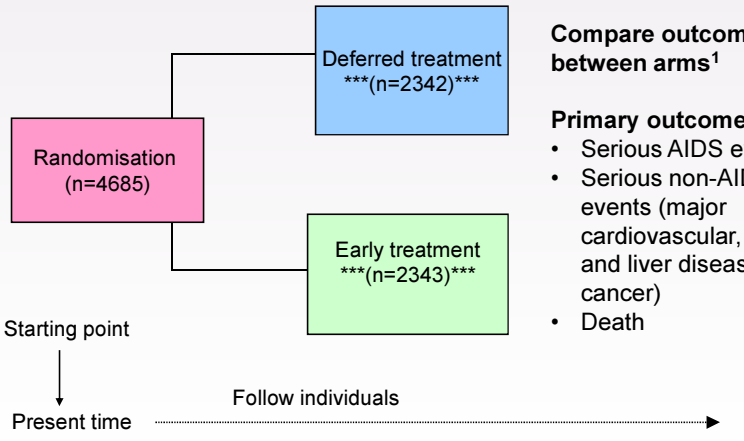
The START trial

START trial 1

- 4685 study participants:
  - HIV-positive ART-naïve (215 sites in 35 countries)
  - Aged 18 years or older
  - Two CD4 cell counts above 500 cells/µL at least 2 weeks apart within 60 days before randomization
- Randomised to:
  - Immediate ART
  - Defer ART until CD4 cell count declines to 350 cells/mm^3

Sharma, S. et al. Demographic and HIV-specific characteristics of participants enrolled in the INSIGHT START trial. HIV Medicine, 16: 30–36

START trial 2



Compare outcome between arms[1]

Primary outcome:
- Serious AIDS events
- Serious non-AIDS events (major cardiovascular, renal and liver disease and cancer)
- Death

[1] http://www.niaid.nih.gov/news/newsreleases/2015/Pages/START.aspx

***Assumption/estimate: "roughly equal numbers" in each study arm

START trial 3

- On average, participants were followed for 3 years[1]
- "In 2013, researchers thought 213 events would be needed to see a clear difference between the groups"[2]
- Based on data up until March 2015, DSMB found[2]:

Table 1a. Number of primary endpoints in each arm (15 May 2015)

	Number of events	
	Early arm (A)	Later arm (B)
Category 1:AIDS, serious non-AIDS, or death (primary).	41	86
Category 2:AIDS or AIDS death.	14	46
Category 3:Serious non-AIDS or non-AIDS death.	28	41

[1]http://www.niaid.nih.gov/news/newsreleases/2015/Pages/START.aspx
[2]http://i-base.info/i-base-qa-on-the-start-study-results/

Describing the risk of an event

(Incidence) Risk of an event

$$= \frac{\text{Number of new cases over study period}}{\text{Total population at risk at the start of the study period}}$$

Example – START trial

- Over an average of 3 years follow-up:

Regimen/ Intervention	Experienced event	Did not experience event*	Total
Immediate ART	86 (1.8%)	2256 (96.3%)	2342
Deferred ART	41 (3.7%)	2301 (98.2%)	2343
Total	127	4557	4685

*Estimated

Example – START trial

- Over an average of 3 years follow-up:

Regimen/ Intervention	Experienced event	Did not experience event*	Total
Immediate ART	86 (1.8%)	2256 (96.3%)	2342
Deferred ART	41 (3.7%)	2301 (98.2%)	2343
Total	127	4557	4685

P<0.0001 (chi-squared test)

*Estimated

Comparing the risk of an event in two groups

Relative risk (RR) of an event

$$= \frac{\text{Risk of event in intervention arm}}{\text{Risk of event in control arm}}$$

Absolute risk reduction (ARR) of an event

= Risk in intervention arm – Risk in control arm

Example – START trial

- Over an average of 3 years follow-up:

Regimen/ Intervention	Experienced event	Did not experience event*	Total
Immediate ART	86 (1.8%)	2256 (96.3%)	2342
Deferred ART	41 (3.7%)	2301 (98.2%)	2343
Total	127	4557	4685

Risk difference: 1.8% - 3.7% = -1.9%

95% CI: -2.9% to -1.0%

*Estimated

Example – START trial

- Over an average of 3 years follow-up:

Regimen/ Intervention	Experienced event	Did not experience event*	Total
Immediate ART	86 (1.8%)	2256 (96.3%)	2342
Deferred ART	41 (3.7%)	2301 (98.2%)	2343
Total	127	4557	4685

Risk ratio (relative risk): 1.75% ÷ 3.67% = 0.48

95% CI: 0.33 to 0.69

*Estimated

But……..

- "On average, participants were followed for 3 years"
- Recruitment started March 2011 and completed 2014
- Data analysed up until March 2015
- Our analysis methods assumed everyone was followed for the same amount of time (which is clearly not true)
- We can use rates instead of risk, which allow us to account for different follow-up times

http://www.niaid.nih.gov/news/newsreleases/2015/Pages/START.aspx

Results of START study (reported results)

Table 1b. Relative rates of primary endpoints in each arm (15 May 2015)

	Rate per 100 PY		Hazard Ratio
	Early arm (A)	Late arm (B)	Arm A/B (95% CI)
Category 1:AIDS, serious non-AIDS, or death (primary).	0.60	1.25	0.47(0.32 to 0.68)
Category 2:AIDS or AIDS death.	0.20	0.66	0.30(0.17 to 0.55)
Category 3:Serious non-AIDS or non-AIDS death.	0.41	0.59	0.670.42 to 1.09) NS **

* PY = patient years, ** NS = not statistically significant

http://i-base.info/i-base-qa-on-the-start-study-results/

Hazard ratio = "Hazard rate ratio" or "Relative rate"

Results of START study (reported results)

Table 1b. Relative rates of primary endpoints in each arm (15 May 2015)

	Rate per 100 PY		Hazard Ratio
	Early arm (A)	Late arm (B)	Arm A/B (95% CI)
Category 1:AIDS, serious non-AIDS, or death (primary).	0.60	1.25	0.47(0.32 to 0.68)
Category 2:AIDS or AIDS death.	0.20	0.66	0.30(0.17 to 0.55)
Category 3:Serious non-AIDS or non-AIDS death.	0.41	0.59	0.67(0.42 to 1.09) NS **

* PY = patient years, ** NS = not statistically significant

http://i-base.info/i-base-qa-on-the-start-study-results/

Hazard ratio = "Hazard rate ratio"

Summary

- P-values are used to give an indication of whether we believe an observed difference in treatment response between treatment groups is likely to be a chance finding or not
- Confidence intervals are useful for providing us with an estimate of how sure we are of our results
- Risk ratios and rate ratios can be used to summarise the results of RCTs. However, the absolute risk of events occurring should also be considered