EACS HIV Summer School 2018

Plenary 6:
p-values and hypothesis testing

Valentina Cambiano

UCL Institute of Global Health

31st August 2018

Slides developed by Colette Smith, Caroline Sabin & Valentina Cambiano

Conflict of Interests

No conflict of interests to declare.

Outline

- The role of chance
- Defining and interpreting p-values
- Commonly used hypothesis tests
- Limitations of p-values

Outline

- The role of chance
- Defining and interpreting p-values
- Commonly used hypothesis tests
- Limitations of p-values

Hypothesis tests – background

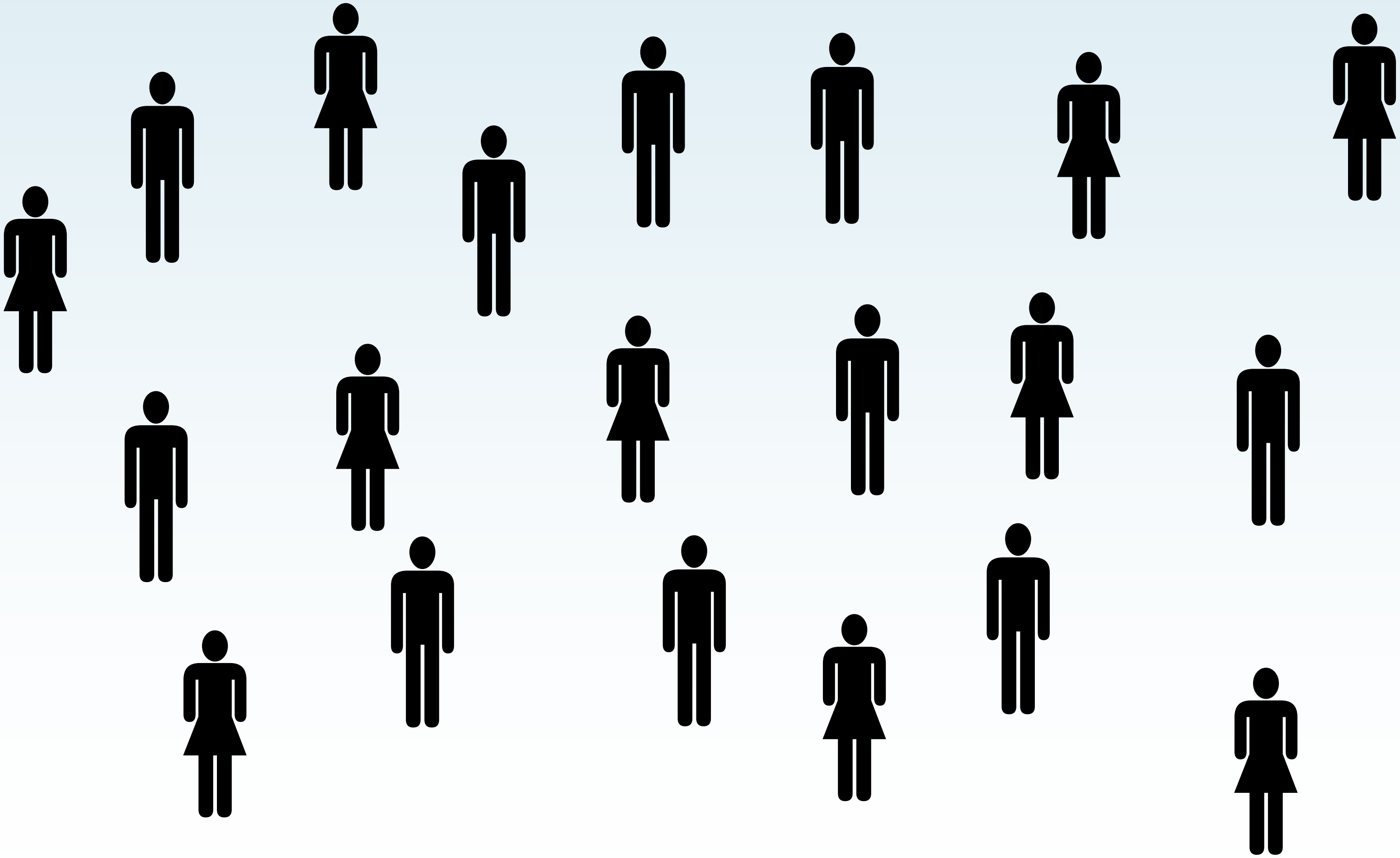
‘p<0.05’

- Presentations of data in the medical world are littered with p-values - ‘p<0.05’. It 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?

Example – baseline imbalance in trials

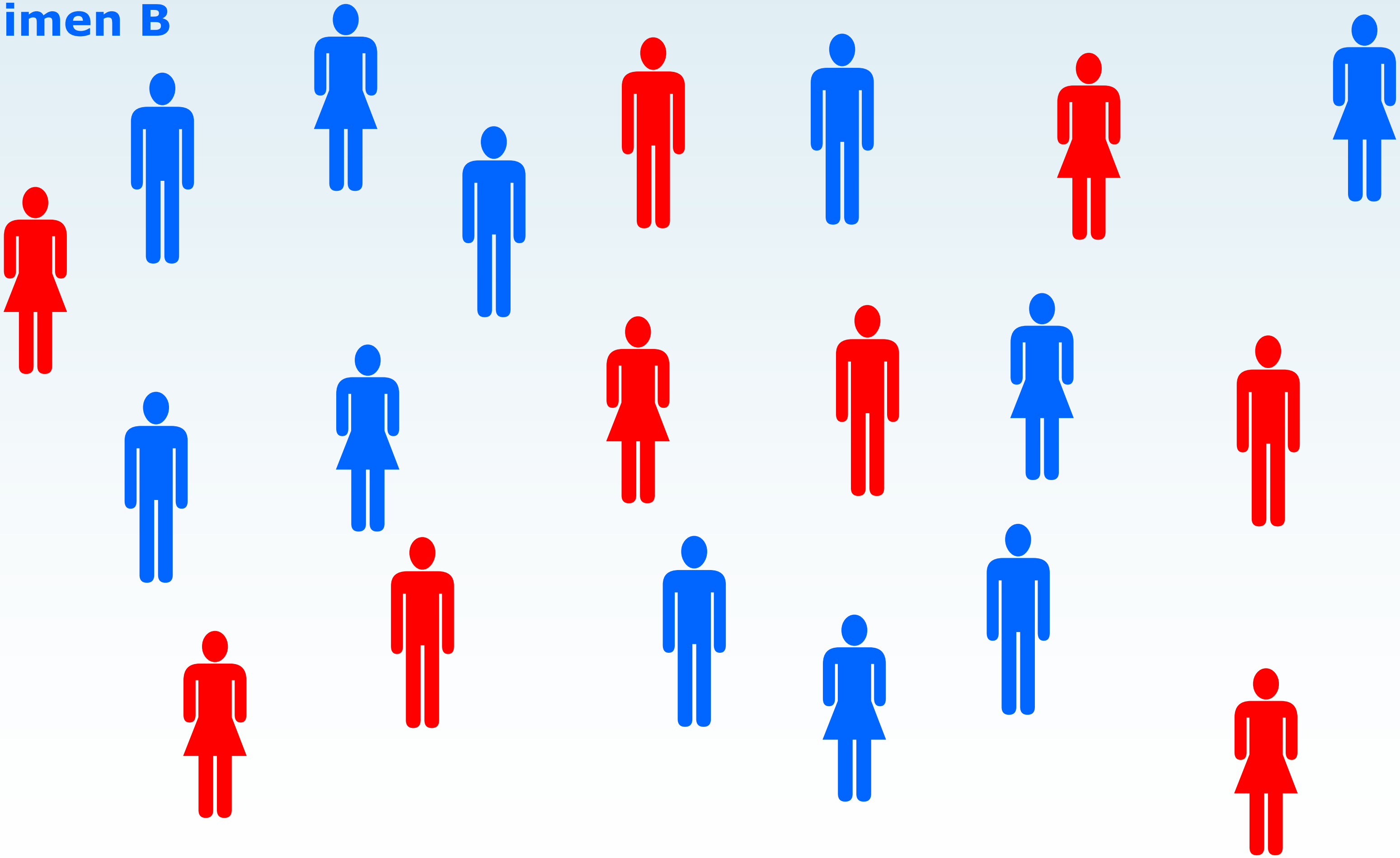
- Imagine 20 participants in a trial, 50% of whom are female
- We randomise the group in a 1:1 manner to receive one of two regimens, A (red) or B (blue)
- We should end up with approximately 10 patients allocated to regimen A and 10 patients to regimen
- What happens in practice?

20 trial participants

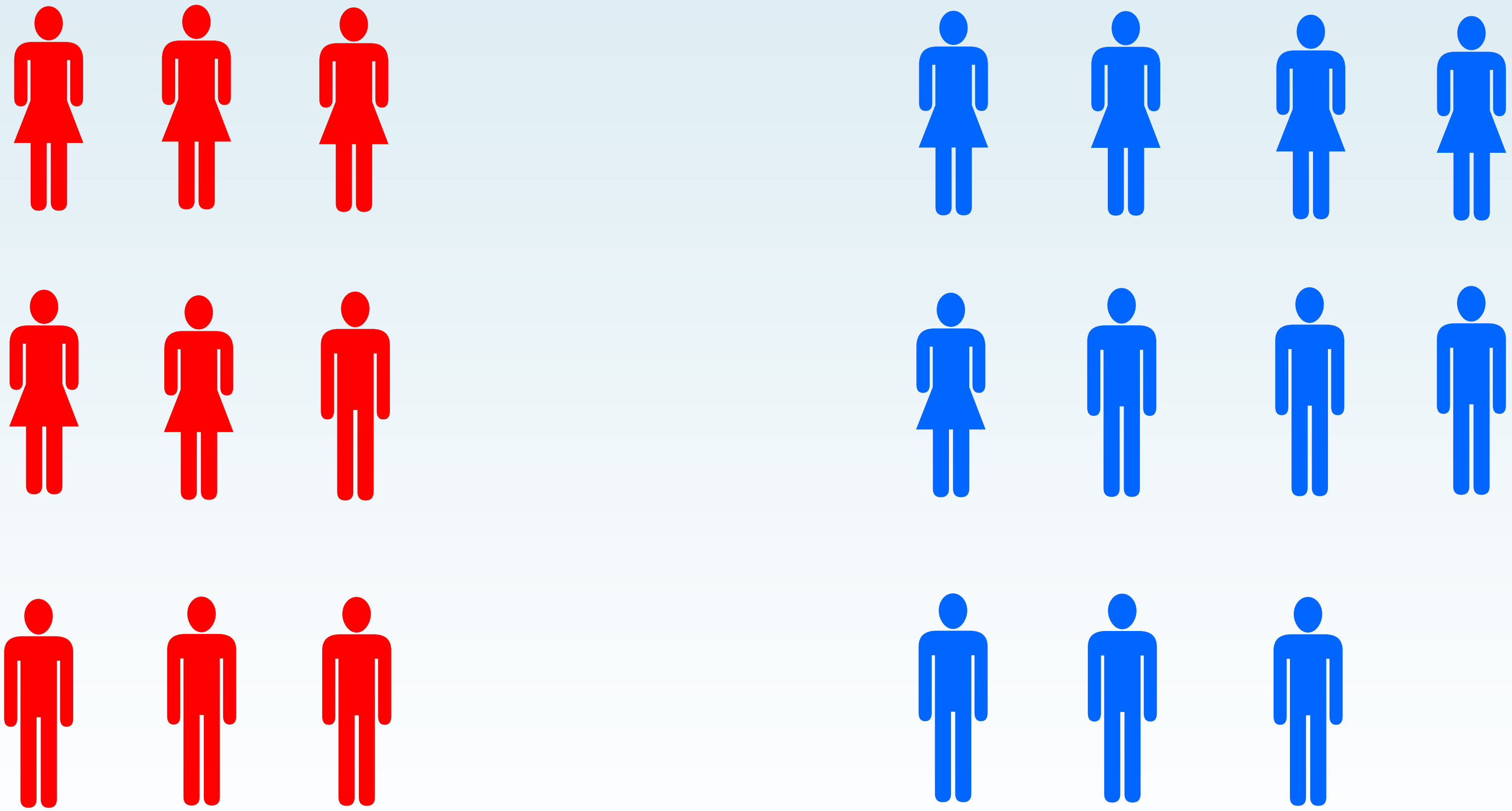


20 trial participants

Regimen A
Regimen B



20 trial participants



Regimen A

Regimen B

20 trial participants - % female

	Regimen				Overall
	A		B		
Trial number	N	N (%) female	N	N (%) female	N
1	9	5 (55.6)	11	5 (45.5)	20

20 trial participants - % female

	Regimen				Overall
	A		B		
Trial number	N	N (%) female	N	N (%) female	N
1	9	5 (55.6)	11	5 (45.5)	20
2	10	5 (50.0)	10	5 (50.0)	20
3	7	3 (42.9)	13	7 (53.8)	20
4	15	7 (46.7)	5	3 (60.0)	20
5	8	5 (62.5)	12	5 (41.7)	20
6	8	4 (50.0)	12	6 (50.0)	20
7	10	5 (50.0)	10	5 (50.0)	20
8	10	6 (60.0)	10	4 (40.0)	20
9	11	7 (63.6)	9	3 (33.3)	20
10	10	3 (30.0)	10	7 (70.0)	20

20 trial participants - % female

	Regimen				Overall
	A		B		
Trial number	N	N (%) female	N	N (%) female	N
1	9	5 (55.6)	11	5 (45.5)	20
2	10	5 (50.0)	10	5 (50.0)	20
3	7	3 (42.9)	13	7 (53.8)	20
4	15	7 (46.7)	5	3 (60.0)	20
5	8	5 (62.5)	12	5 (41.7)	20
6	8	4 (50.0)	12	6 (50.0)	20
7	10	5 (50.0)	10	5 (50.0)	20
8	10	6 (60.0)	10	4 (40.0)	20
9	11	7 (63.6)	9	3 (33.3)	20
10	10	3 (30.0)	10	7 (70.0)	20

20 trial participants - % female

	Regimen				Overall
	A		B		
Trial number	N	N (%) female	N	N (%) female	N
1	9	5 (55.6)	11	5 (45.5)	20
2	10	5 (50.0)	10	5 (50.0)	20
3	7	3 (42.9)	13	7 (53.8)	20
4	15	7 (46.7)	5	3 (60.0)	20
5	8	5 (62.5)	12	5 (41.7)	20
6	8	4 (50.0)	12	6 (50.0)	20
7	10	5 (50.0)	10	5 (50.0)	20
8	10	6 (60.0)	10	4 (40.0)	20
9	11	7 (63.6)	9	3 (33.3)	20
10	10	3 (30.0)	10	7 (70.0)	20

100 trial participants - % female

	Regimen				Overall
	A		B		
Trial number	N	N (%) female	N	N (%) female	N
1	54	28 (51.9)	46	22 (47.8)	100
2	53	24 (45.3)	47	26 (55.3)	100
3	61	30 (49.2)	39	20 (51.3)	100
4	51	25 (49.0)	49	25 (51.0)	100
5	57	29 (50.9)	43	21 (48.8)	100
6	50	24 (48.0)	50	26 (52.0)	100
7	51	22 (43.1)	49	28 (57.1)	100
8	54	30 (55.6)	46	20 (43.5)	100
9	57	28 (49.1)	43	22 (51.2)	100
10	47	20 (42.6)	53	30 (56.6)	100

The role of ‘chance’

- So even if we randomly subdivide patients into two groups, their characteristics may be imbalanced
- The size of the imbalance generally gets smaller as the trial increases in size
- Random baseline covariate imbalance is not usually a problem in a trial (unless it is big) as statistical methods can deal with this
- However, if we are describing outcomes rather than baseline covariates, then there is more cause for concern

Trial participants - % viral load <50 cps/ml

	Regimen			
	A		B	
Trial number	N	N (%) VL<50 copies/ml	N	N (%) VL<50 copies/ml
1	54	28 (51.9)	46	22 (47.8)
2	53	24 (45.3)	47	26 (55.3)
3	61	30 (49.2)	39	20 (51.3)
4	51	25 (49.0)	49	25 (51.0)
5	57	29 (50.9)	1	21 (48.8)
6	50	24 (48.0)	50	26 (52.0)
7	51	22 (43.1)	49	28 (57.1)
8	54	30 (55.6)	46	20 (43.5)
9	57	28 (49.1)	43	22 (51.2)
10	47	20 (42.6)	53	30 (56.6)

Trial participants - % viral load <50 cps/ml

	Regimen			
	A		B	
Trial number	N	N (%) VL<50 copies/ml	N	N (%) VL<50 copies/ml
1	54	28 (51.9)	46	22 (47.8)
2	53	24 (45.3)	47	26 (55.3)
3	61	30 (49.2)	39	20 (51.3)
4	51	25 (49.0)	49	25 (51.0)
5	57	29 (50.9)	43	21 (48.8)
6	50	24 (48.0)	50	26 (52.0)
7	51	22 (43.1)	49	28 (57.1)
8	54	30 (55.6)	46	20 (43.5)
9	57	28 (49.1)	43	22 (51.2)
10	47	20 (42.6)	53	30 (56.6)

Trial participants - % viral load <50 cps/ml

	Regimen			
	A		B	
Trial number	N	N (%) VL<50 copies/ml	N	N (%) VL<50 copies/ml
1	54	28 (51.9)	46	22 (47.8)
2	53	24 (45.3)	47	26 (55.3)
3	61	30 (49.2)	39	20 (51.3)
4	51	25 (49.0)	49	25 (51.0)
5	57	29 (50.9)	43	21 (48.8)
6	50	24 (48.0)	50	26 (52.0)
7	51	22 (43.1)	49	28 (57.1)
8	54	30 (55.6)	46	20 (43.5)
9	57	28 (49.1)	43	22 (51.2)
10	47	20 (42.6)	53	30 (56.6)

14% difference in outcome

Outline

- The role of chance
- Defining and interpreting p-values
- Commonly used hypothesis tests
- Limitations of p-values

The general approach to hypothesis testing

- Investigator may want to conduct a study to address a certain theory (study hypothesis)
  - e.g. % viral load <50 cps/ml is higher in people receiving regimen B compared to regimen A

1. Start by defining two hypotheses:
   - **Null hypothesis (H_0):** There is no real difference in viral load response rates between the two regimens
   - **Alternative hypothesis (H_1):** There is a real difference in viral load response rates between the two regimens

Null hypothesis (H0)

- E.G. The difference in % viral load <50 cps/ml between the population receiving regimen A and regimen B is 0%

- Can’t look at whole population who could receive regimen A and B!!

- Use a sample to make inferences about wider population

- Is there any evidence from our sample against the null hypothesis?

The general approach to hypothesis testing

1. Definition of two hypotheses
2. Conduct trial and collect data
3. Use data from that trial (sample) to calculate a test statistic (e.g. Chi-squared test, t-test, ANOVA). Type of test statistic depends upon type of data (e.g. quantitative or categorical)
4. Test statistic can then be ‘looked up’ in tables and a p-value obtained

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
  - *Insufficient evidence that effect is real*
- Small p-value
  - *Evidence that effect is real*

What is large and what is small?

By convention:

P<0.05 – SMALL

P>0.05 – LARGE

Outline

- The role of chance
- Defining and interpreting p-values
- Commonly used hypothesis tests
- Limitations of p-values

Choosing the right hypothesis test

All statistical tests will generate a *P*-value - the choice of statistical test will be based on a number of factors, including:

- The hypothesis being studied
- The variables of particular interest
- The distribution of their values
- The number of individuals who will be included in the analysis
- The number of ‘groups’ being studied
- The relationship (if any) between these groups

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

Comparing numbers

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

Example – the Chi-squared test

- Two groups
- Interested in whether the proportion of individuals with an outcome differs between these groups
- Measurement of interest is categorical
- Can draw up a table of responses in the groups
- Expected numbers in each cell of the table are >5

Example – Define hypotheses

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

Hypotheses:

H_0: There is **no** real difference in the proportion of patients with a VL$\leq$50 copies/ml between those receiving regimen A and those receiving regimen B

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

Example – Collect data

	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 – Interpret *P*-value

- p-value associated with this test value =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
- So we would conclude that 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

Outline

- The role of chance
- Defining and interpreting p-values
- Commonly used hypothesis tests
- Limitations of p-values

Limitation of p-values

- Although p-values are helpful in telling us which effects are likely to be real, they also suffer from limitations
- An estimate of the size of the effect and its corresponding confidence interval provides complementary information
- The limitations of p-values, as well as the use of confidence intervals, will be seen in the next session