P-values and hypothesis testing

Caroline Sabin

UCL Institute for Global Health

Conflicts of interest

I have received funding for the membership of Data Safety and Monitoring Boards, Advisory Boards and for the preparation of educational materials from:

- Gilead Sciences
- ViiV Healthcare
- Janssen-Cilag

Outline

- The role of chance
- Interpreting *P*-values
- Commonly used hypothesis tests
- Limitations of *P*-values

The role of chance

Hypothesis tests – background

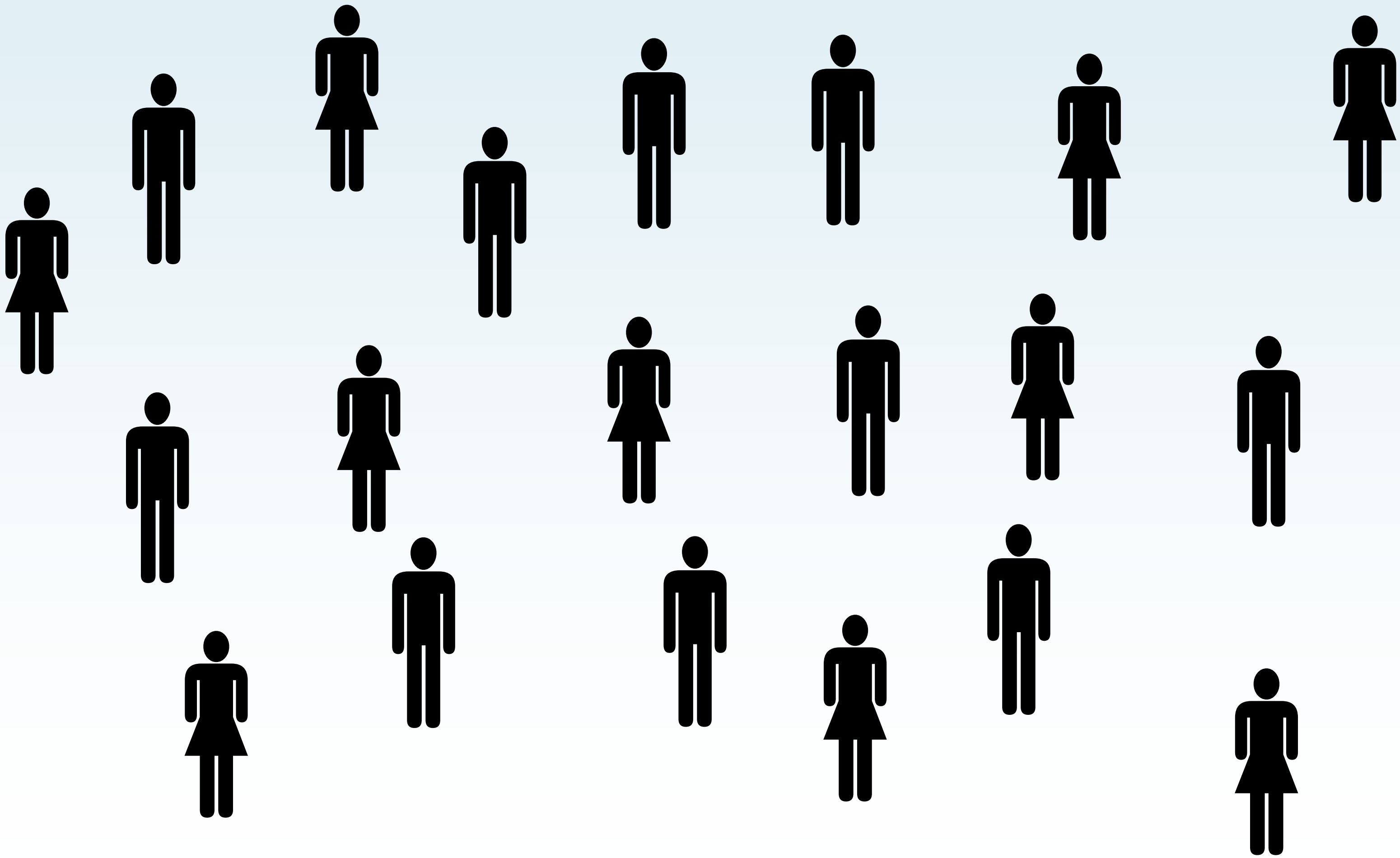
- Presentations of data in the medical world are littered with *P*-values - ‘$P<0.05$’ is thought to be a magical phrase, guaranteed to ensure that your paper will be published
- But what do these *P*-values really tell us, and is a *P*-value <0.05 really that important?

P-values – what do they tell us?

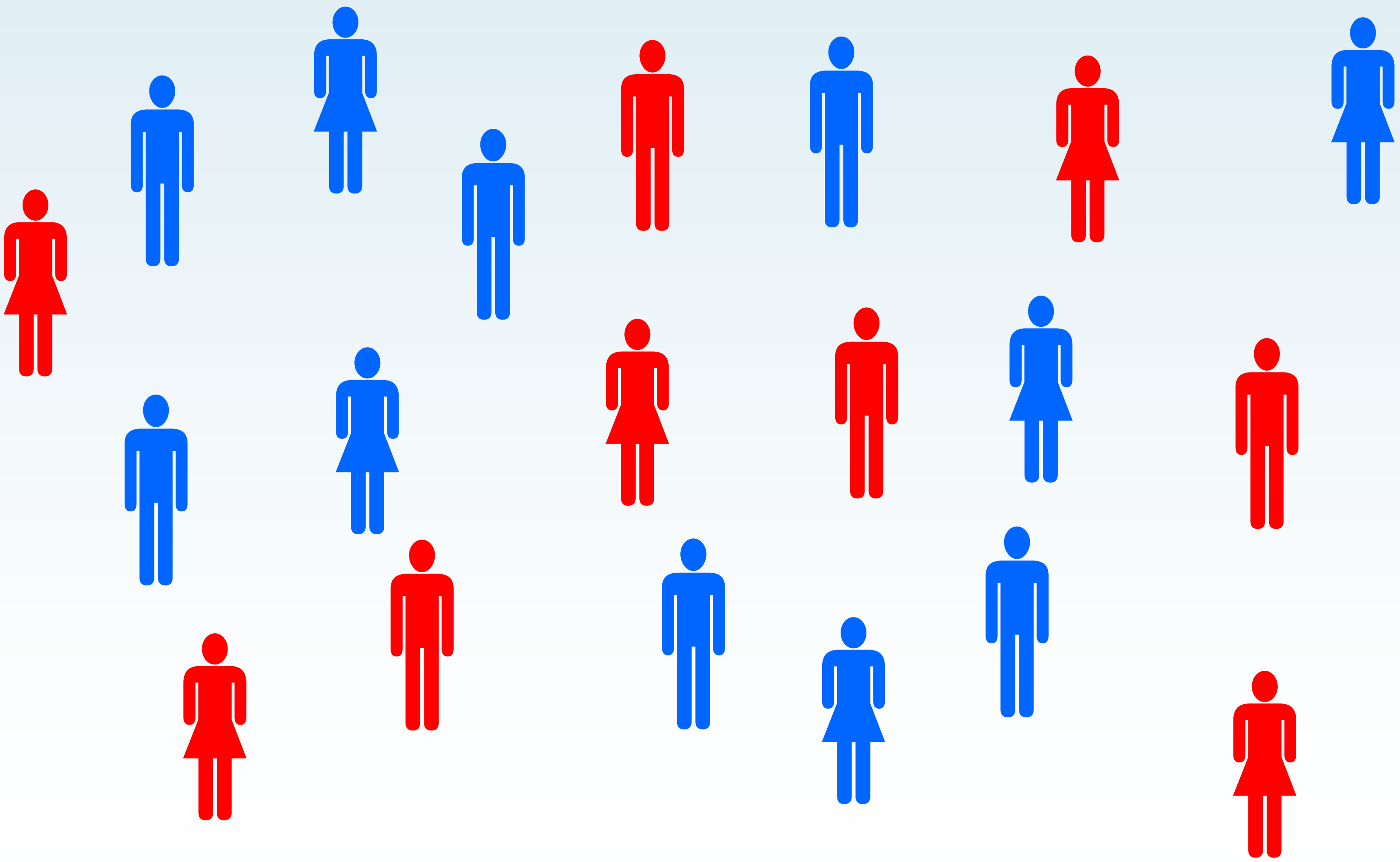
Example – baseline imbalance in trials

- 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?

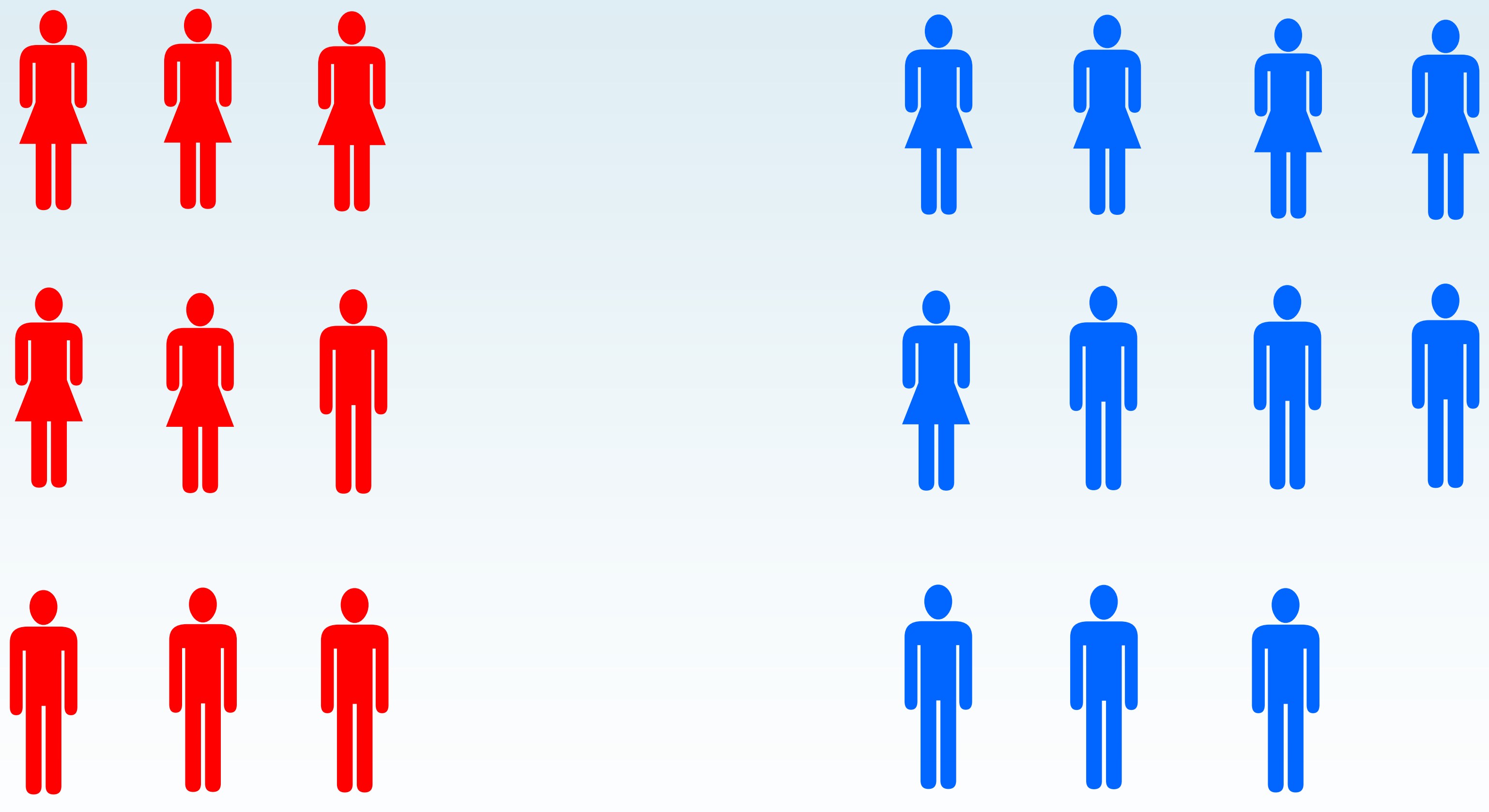
20 trial participants



20 trial participants



20 trial participants



Regimen A **Regimen B**

20 trial participants - % female

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

20 trial participants - % female

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

20 trial participants - % female

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

20 trial participants - % female

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

100 trial participants - % female

	Regimen			
	A		B	
Trial number	N	N (%) female	N	N (%) female
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)

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)	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

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

Hypothesis testing - how do we obtain a *P*-value?

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 the two regimens
  - Alternative hypothesis (H_1): There is a real difference in viral load response rates between the 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

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

- The hypotheses 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 (A) are more/less 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 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

Regimen	VL≤50 copies/ml N (%)	VL >50 copies/ml N (%)	Total 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
- 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

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 discussed in Plenary 7