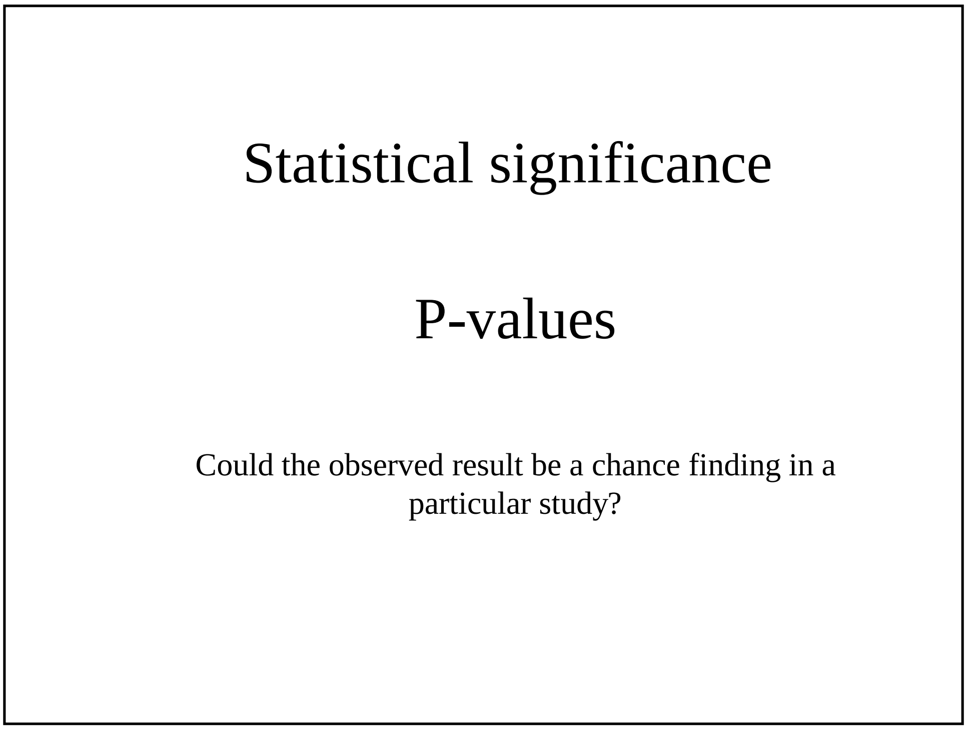
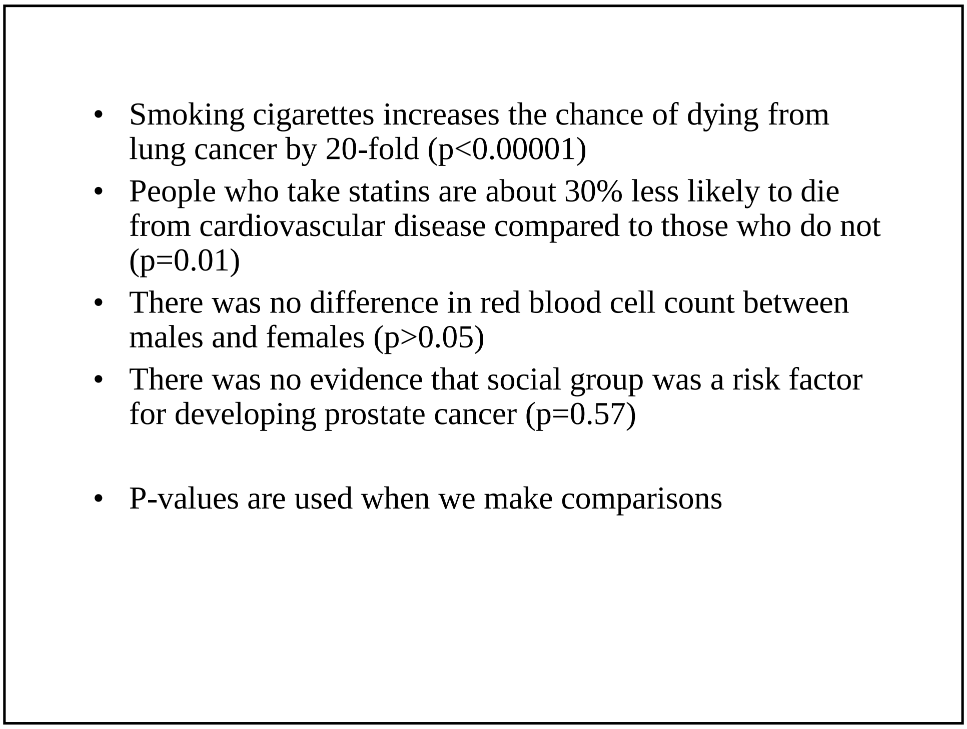


Statistical significance

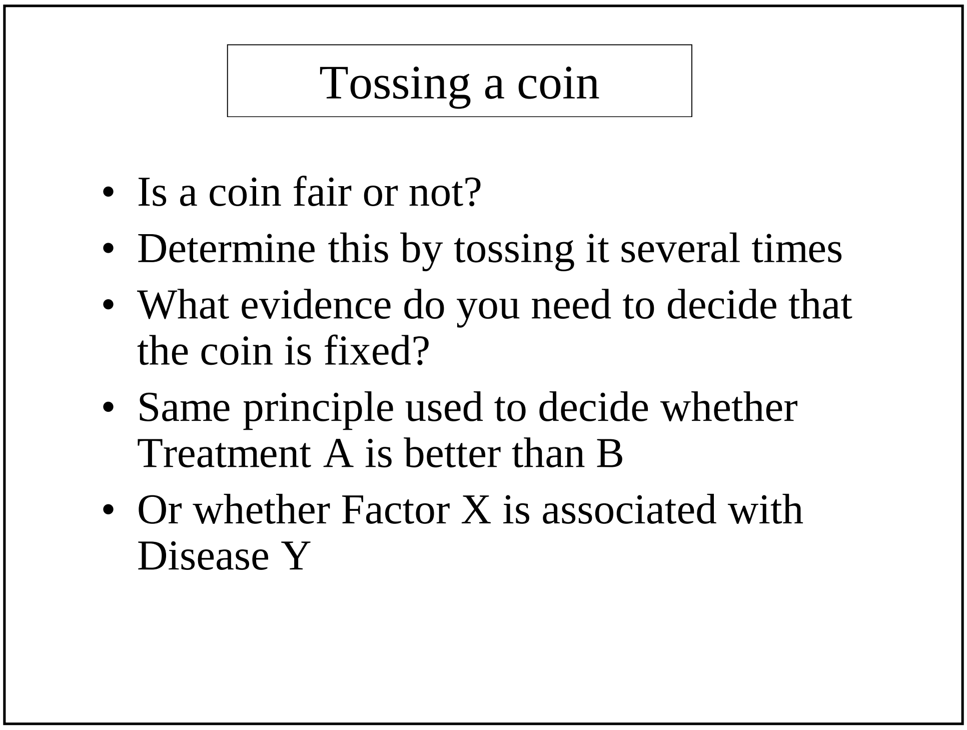
P-values

Could the observed result be a chance finding in a particular study?



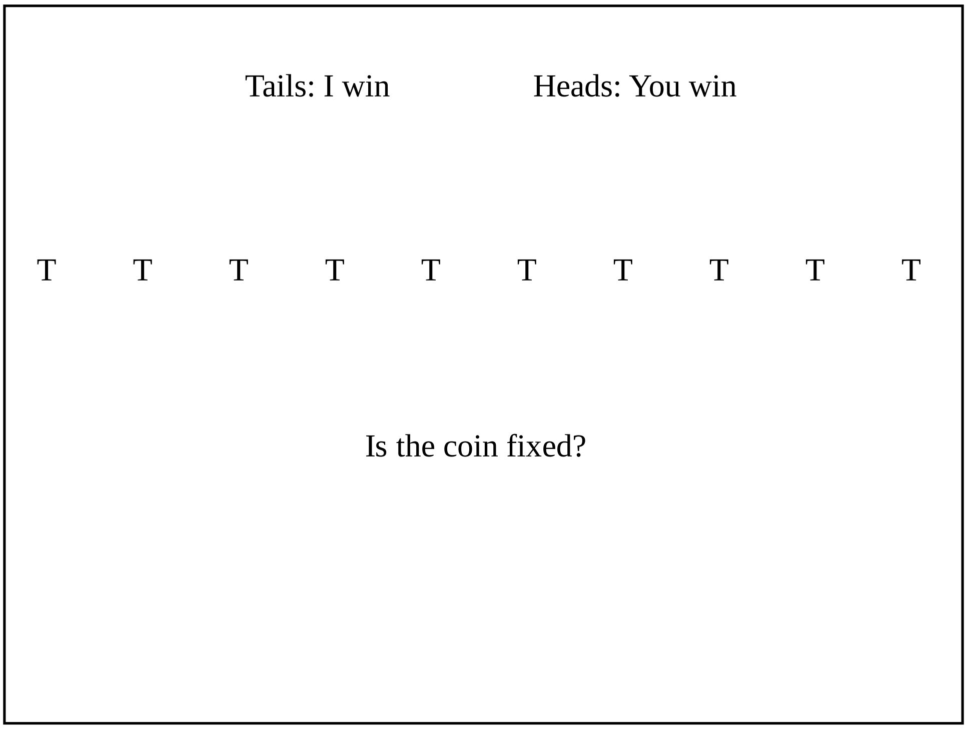
- Smoking cigarettes increases the chance of dying from lung cancer by 20-fold (p<0.00001)
- People who take statins are about 30% less likely to die from cardiovascular disease compared to those who do not (p=0.01)
- There was no difference in red blood cell count between males and females (p>0.05)
- There was no evidence that social group was a risk factor for developing prostate cancer (p=0.57)

- P-values are used when we make comparisons



Tossing a coin

- Is a coin fair or not?
- Determine this by tossing it several times
- What evidence do you need to decide that the coin is fixed?
- Same principle used to decide whether Treatment A is better than B
- Or whether Factor X is associated with Disease Y



Tails: I win Heads: You win

T T T T T T T T T T

Is the coin fixed?

Tails: I win Heads: You win

T T T T H T H T T T

Is the coin fixed?

A: patient is alive at end of study D: Patient is dead

New Treatment

A A D A A A D A A A

Standard treatment

D D A A D D D A D A

Is the new treatment different to the standard?

Each “experiment” involves throwing the coin 10 times

Number of heads	Number of tails	Probability of this occurring
0	10	0.0010
1	9	0.0097
2	8	0.0440
3	7	0.1172
4	6	0.2051
5	5	0.2460
6	4	0.2051
7	3	0.1172
8	2	0.0040
9	1	0.0097
10	0	0.0010

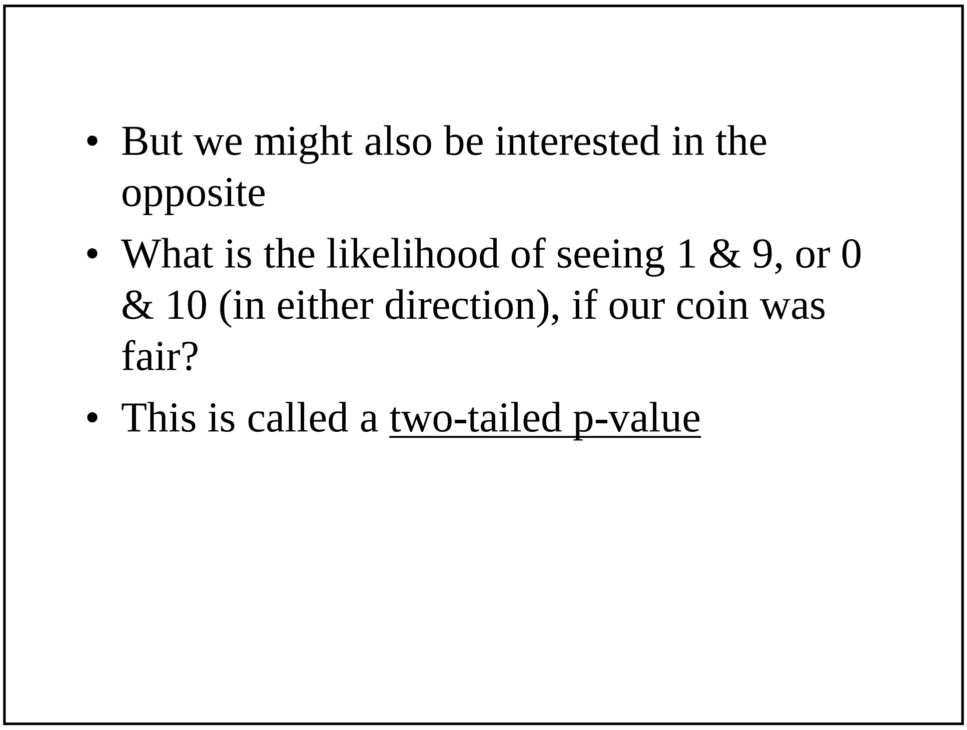
- If we see 0 Heads and 10 Tails, the p-value is 0.001
- It is possible to see this just by chance alone, ie if the coin were fair.
- But we would have to throw the coin 10 times, and do this 1000 times, and we only expect to see the observed results (ie 0 Heads and 10 Tails) for one of the 1000 times
- So the observed results, while not impossible to get by chance, are highly unlikely if the coin were fair

- If we observe 1 Heads and 9 Tails, we are also interested in anything more extreme than this
- What is the likelihood of seeing
  - – 1 Heads & 9 Tails or
  - – 0 Heads & 10 Tails
- if our coin was fair?
- This is called a one-tailed p-value

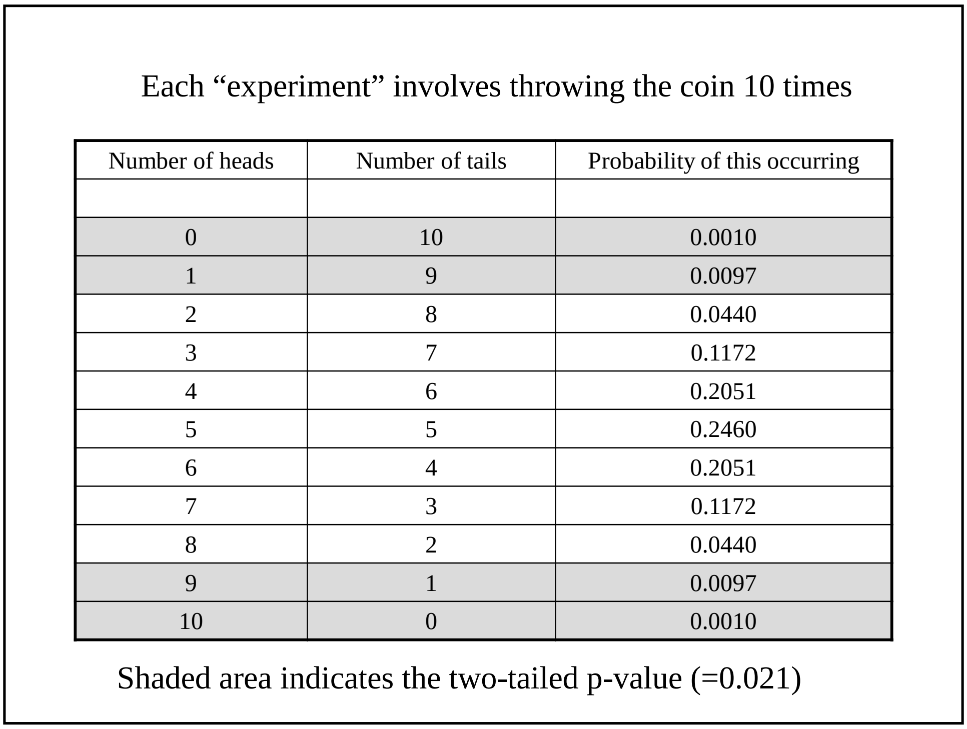
Each “experiment” involves throwing the coin 10 times

Number of heads	Number of tails	Probability of this occurring
0	10	0.0010
1	9	0.0097
2	8	0.0440
3	7	0.1172
4	6	0.2051
5	5	0.2460
6	4	0.2051
7	3	0.1172
8	2	0.0440
9	1	0.0097
10	0	0.0010

Shaded area indicates the one-tailed p-value (=0.0107)



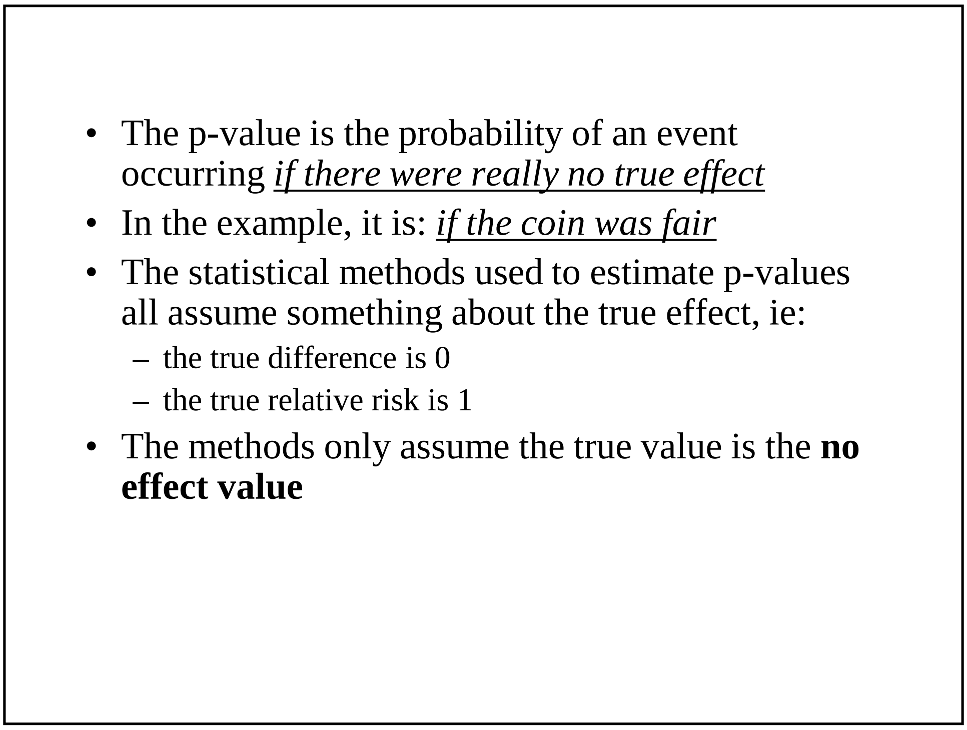
- But we might also be interested in the opposite
- What is the likelihood of seeing 1 & 9, or 0 & 10 (in either direction), if our coin was fair?
- This is called a <u>two-tailed p-value</u>



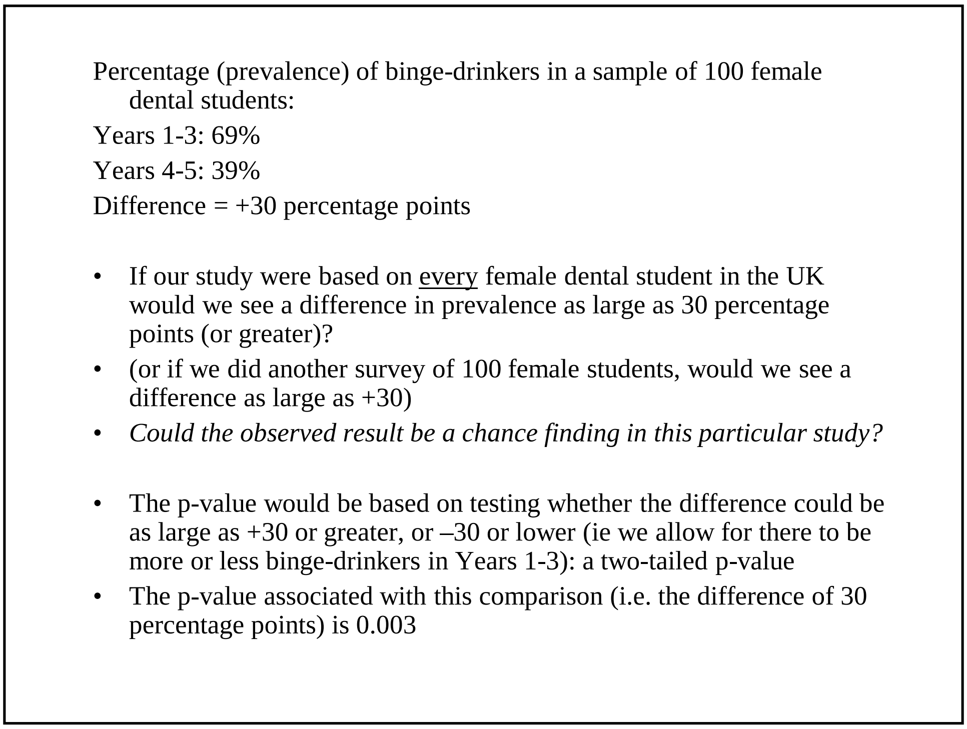
Each “experiment” involves throwing the coin 10 times

Number of heads	Number of tails	Probability of this occurring
0	10	0.0010
1	9	0.0097
2	8	0.0440
3	7	0.1172
4	6	0.2051
5	5	0.2460
6	4	0.2051
7	3	0.1172
8	2	0.0440
9	1	0.0097
10	0	0.0010

Shaded area indicates the two-tailed p-value (=0.021)



- The p-value is the probability of an event occurring *if there were really no true effect*
- In the example, it is: *if the coin was fair*
- The statistical methods used to estimate p-values all assume something about the true effect, ie:
  - the true difference is 0
  - the true relative risk is 1
- The methods only assume the true value is the **no effect value**



Percentage (prevalence) of binge-drinkers in a sample of 100 female dental students:

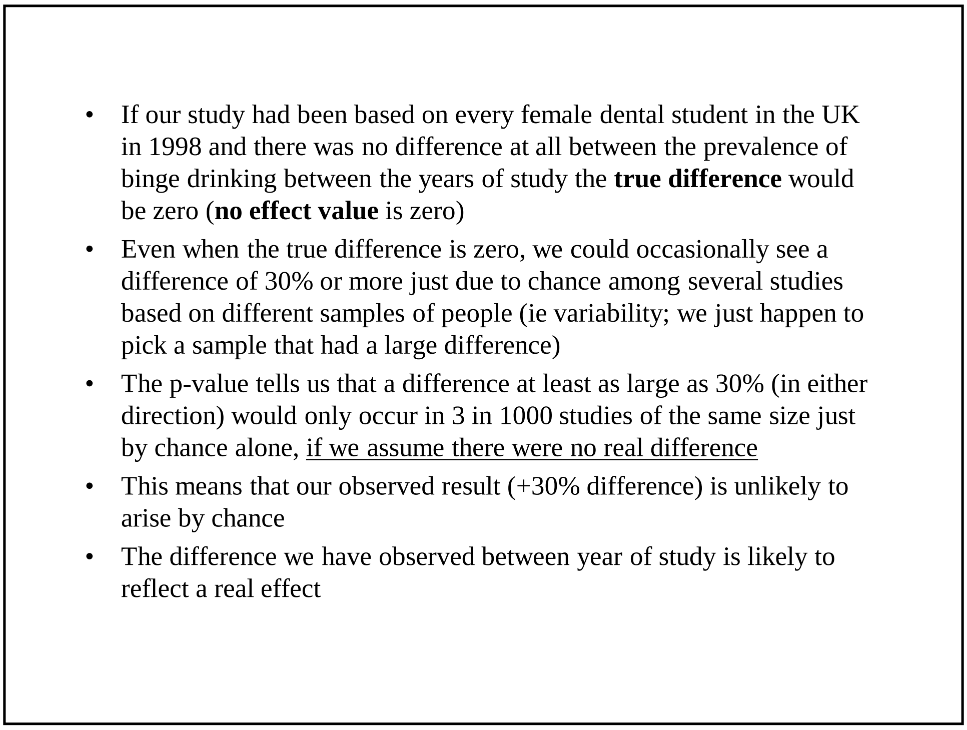
Years 1-3: 69%

Years 4-5: 39%

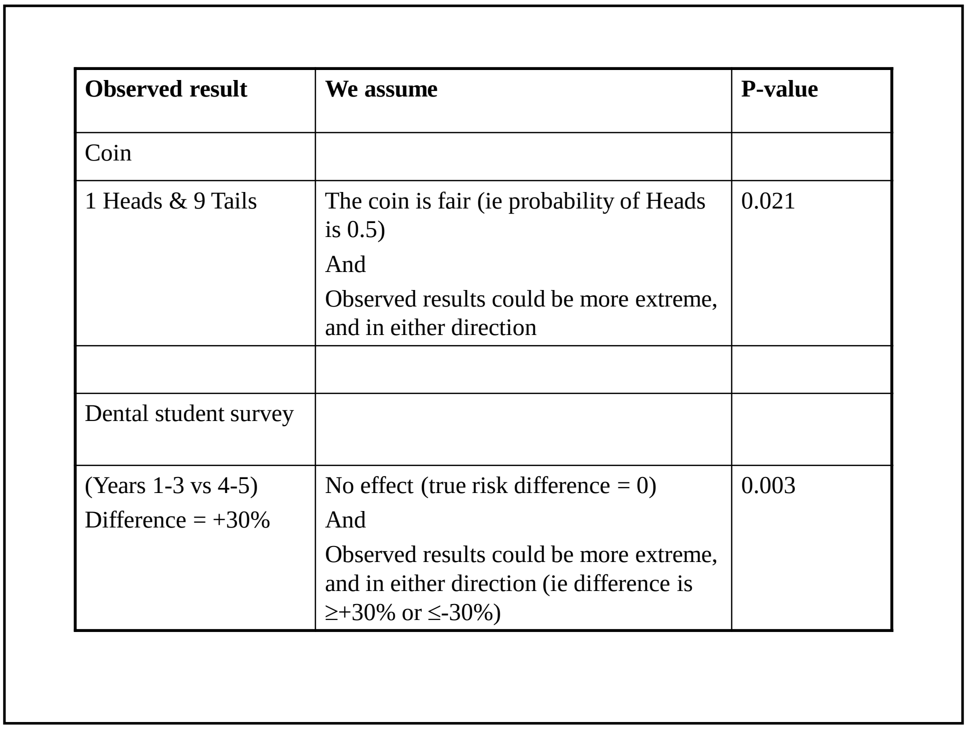
Difference = +30 percentage points

- If our study were based on every female dental student in the UK would we see a difference in prevalence as large as 30 percentage points (or greater)?
- (or if we did another survey of 100 female students, would we see a difference as large as +30)
- *Could the observed result be a chance finding in this particular study?*

- The p-value would be based on testing whether the difference could be as large as +30 or greater, or –30 or lower (ie we allow for there to be more or less binge-drinkers in Years 1-3): a two-tailed p-value
- The p-value associated with this comparison (i.e. the difference of 30 percentage points) is 0.003



- If our study had been based on every female dental student in the UK in 1998 and there was no difference at all between the prevalence of binge drinking between the years of study the **true difference** would be zero (**no effect value** is zero)
- Even when the true difference is zero, we could occasionally see a difference of 30% or more just due to chance among several studies based on different samples of people (ie variability; we just happen to pick a sample that had a large difference)
- The p-value tells us that a difference at least as large as 30% (in either direction) would only occur in 3 in 1000 studies of the same size just by chance alone, if we assume there were no real difference
- This means that our observed result (+30% difference) is unlikely to arise by chance
- The difference we have observed between year of study is likely to reflect a real effect

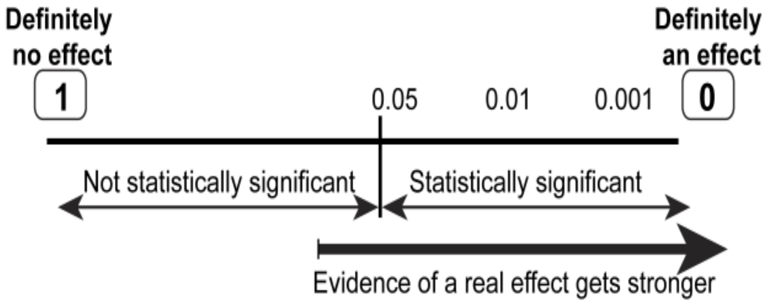


Observed result	**We assume**	**P-value**
Coin		
1 Heads & 9 Tails	The coin is fair (ie probability of Heads is 0.5) And Observed results could be more extreme, and in either direction	0.021
Dental student survey		
(Years 1-3 vs 4-5) Difference = +30%	No effect (true risk difference = 0) And Observed results could be more extreme, and in either direction (ie difference is ≥+30% or ≤-30%)	0.003

- P-values are calculated using a **statistical test**
- For example:
  - when comparing 2 proportions, the test is a chi-squared test
  - when comparing 2 means, it is a t-test
- The type of statistical test used depends on the type of outcome measure
- The main value of these tests is that they produce a p-value
- Details of how the p-values are calculated are not important here (the computer does it)
- You just need to know which test to use

Definition: The probability that an effect size as large as that observed, or more extreme, is due to chance **if** there really were no effect

All p-values are between 0 and 1



Definitely no effect

1

0.05 0.01 0.001

Definitely an effect

0

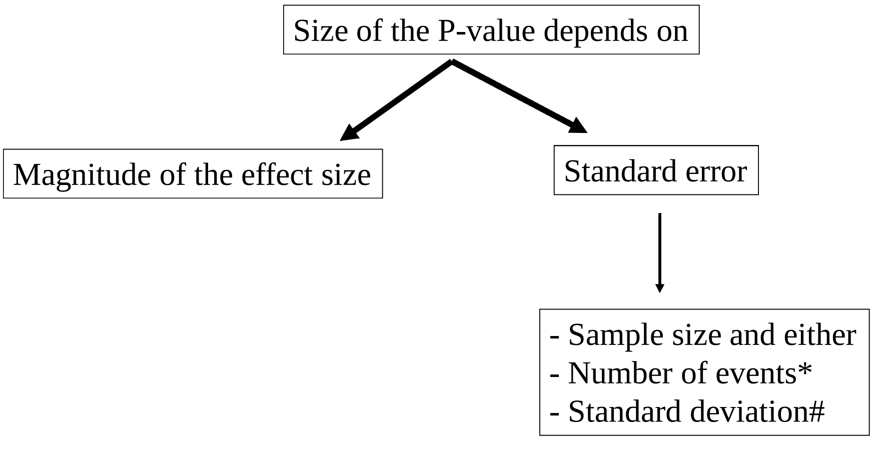
Not statistically significant

Statistically significant

Evidence of a real effect gets stronger

- Traditionally, 0.05 is taken to be the cut-off to conclude statistical significance (ie <0.05)
- It means that we allow ourselves to get the wrong answer 5% of the time
- But there is nothing scientific or special about 0.05
- It is a level of error that is generally judged to be acceptably low

- <u>Never use</u>: ‘p>0.05’, ‘p<0.05’, or ‘NS’ (not stat sig)
- Always provide the actual p-value to 2 or 3 decimal places (eg 0.23, 0.01, 0.004, or if very small <0.001), because the following provides very different levels of evidence:
  - P=0.049 versus p<0.001
  - P=0.058 versus p=0.75



Size of the P-value depends on

Magnitude of the effect size

Standard error

- Sample size and either
- Number of events*
- Standard deviation#

*if using counting people or time-to-event endpoints
#if using taking measurements on people

- Interpret p-values carefully, considering sample size as well as
  - the number of events, if outcome measure is based on ‘counting people’ or time-to-event data
  - variability (ie standard deviation), if outcome measure is based on ‘taking measurements on people’
- A p-value of eg 0.04 should not always be considered strong evidence of a real effect
- A p-value of eg 0.06 should not be used to dismiss a real effect

- We could get a big p-value (>0.05) if:
- A large effect size is seen in a small study (or there were too few events)

- We could get a small p-value (<0.05) if:
- A small effect is seen in a large study

One or two-tailed p-values

- One-tailed p-value is usually half a two-tailed value
- Always use two-tailed, because this gives the most conservative estimate, and reduces the chance of concluding an effect when there might not be one (eg two-tailed p-value of 0.08 is a one-tailed value of 0.04)
- One-tailed should only be used if there is clear justification that the effect can only go in one direction (eg mammography screening vs no screening in reducing breast cancer deaths: it is implausible that mammography would increase the number of deaths)

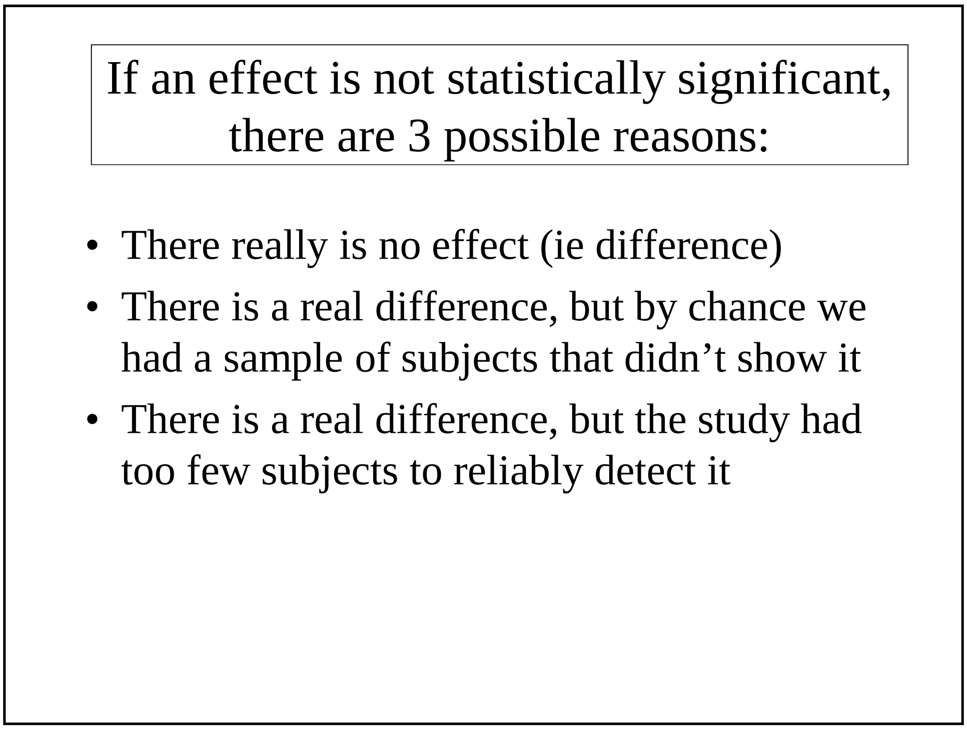
Effect size = difference between 2 things No effect value=0	If $p<0.05$ then the result is statistically significant	The true difference is unlikely to be 0; there is likely to be a real effect
	If $p\geq0.05$ then the result is not statistically significant	We do not have enough evidence to say that there is an effect
Effect size = ratio between 2 things No effect value=1	If $p<0.05$ then the result is statistically significant	The true ratio is unlikely to be 1; there is likely to be a real effect
	If $p\geq0.05$ then the result is not statistically significant	We do not have enough evidence to say that there is an effect

- A p-value of 0.05: we incorrectly conclude there is an effect when there really isn’t one 5% of the time
- A 95% CI: we expect to include the true effect 95% of the time, but get it wrong 5% of the time
- Both allow a 5% error rate, hence there is a relationship between p-value and 95% CI

Common misinterpretation of p-values

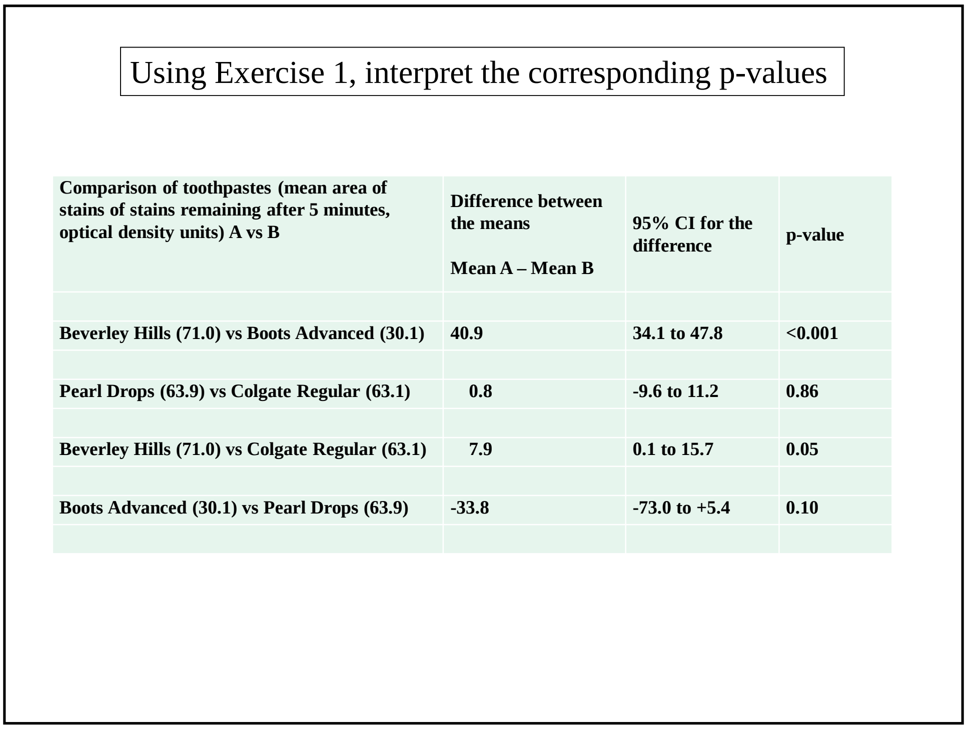
- A p-value greater than 0.05, eg p=0.25, is often used to conclude that

  “*there is no effect*”
- This is incorrect
- Such p-values only tell you that you don’t have enough evidence to say there is an effect
- You cannot use p-values to conclude “no effect”



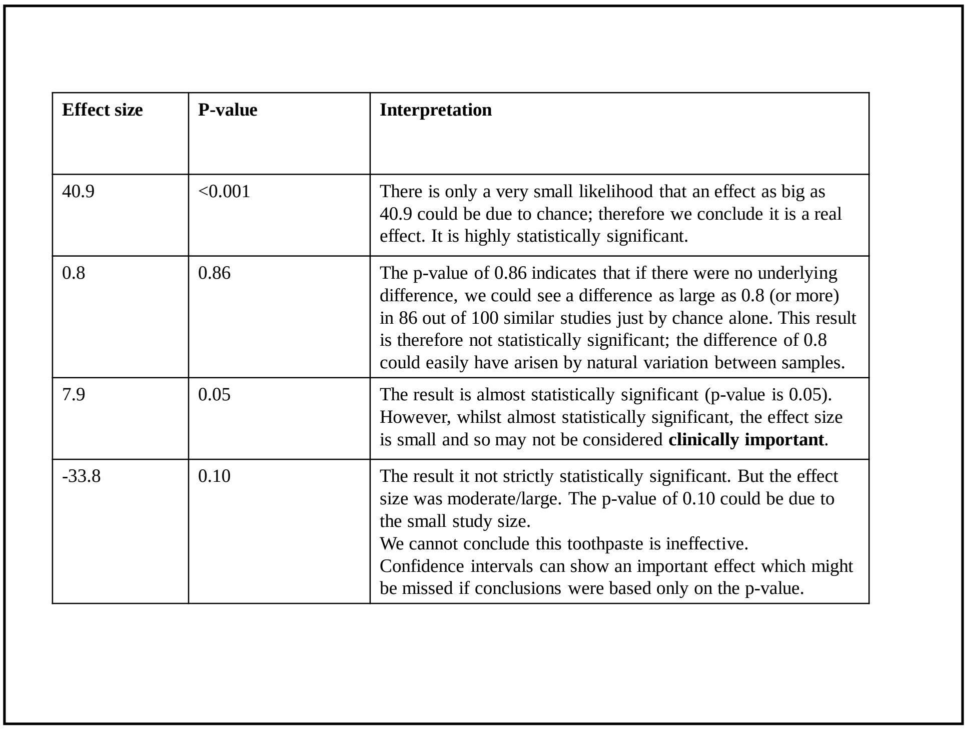
If an effect is not statistically significant, there are 3 possible reasons:

- There really is no effect (ie difference)
- There is a real difference, but by chance we had a sample of subjects that didn’t show it
- There is a real difference, but the study had too few subjects to reliably detect it



Using Exercise 1, interpret the corresponding p-values

Comparison of toothpastes (mean area of stains of stains remaining after 5 minutes, optical density units) A vs B	Difference between the means Mean A – Mean B	95% CI for the difference	p-value
Beverley Hills (71.0) vs Boots Advanced (30.1)	40.9	34.1 to 47.8	<0.001
Pearl Drops (63.9) vs Colgate Regular (63.1)	0.8	-9.6 to 11.2	0.86
Beverley Hills (71.0) vs Colgate Regular (63.1)	7.9	0.1 to 15.7	0.05
Boots Advanced (30.1) vs Pearl Drops (63.9)	-33.8	-73.0 to +5.4	0.10



Effect size	P-value	Interpretation
40.9	<0.001	There is only a very small likelihood that an effect as big as 40.9 could be due to chance; therefore we conclude it is a real effect. It is highly statistically significant.
0.8	0.86	The p-value of 0.86 indicates that if there were no underlying difference, we could see a difference as large as 0.8 (or more) in 86 out of 100 similar studies just by chance alone. This result is therefore not statistically significant; the difference of 0.8 could easily have arisen by natural variation between samples.
7.9	0.05	The result is almost statistically significant (p-value is 0.05). However, whilst almost statistically significant, the effect size is small and so may not be considered **clinically important**.
-33.8	0.10	The result it not strictly statistically significant. But the effect size was moderate/large. The p-value of 0.10 could be due to the small study size. We cannot conclude this toothpaste is ineffective. Confidence intervals can show an important effect which might be missed if conclusions were based only on the p-value.