PH 458/231
Evidence & Policy
Lecture 4

John Worrall
LAK 3.02
Office Hours:
Monday 13.30-14.30
Thursday 12.30 – 13.30

A Scientific Example

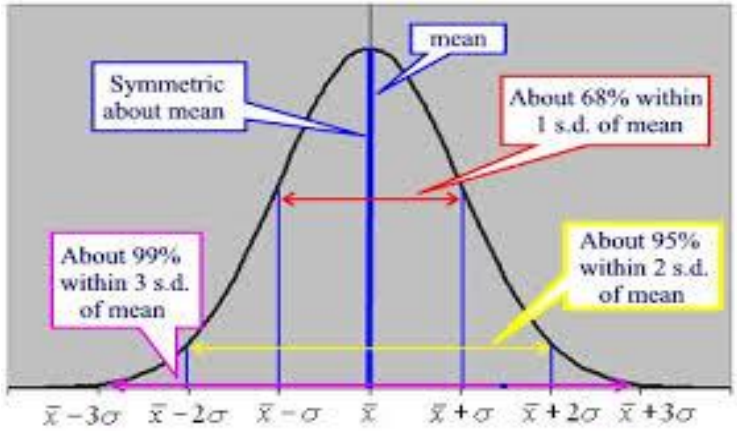
- Suppose a new drug is being tested
- Study population divided into two: experimental and control
- Some statistic is measured for each member of each group: for example reduction in pain level
- Standardly variation in response in both groups
- We would like to characterise situations in which we can say that in light of evidence the treatment ‘works’ (or at least to be able to say that with some degree of confidence)

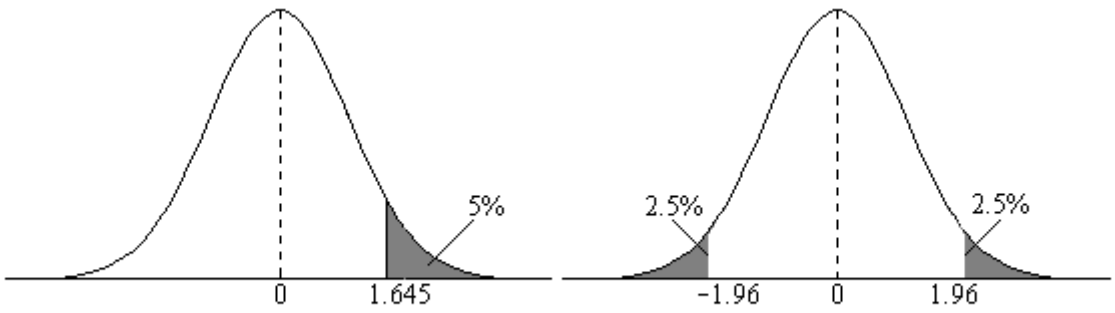
A Scientific Example

- Classical statistics advises: think of yourself as testing the hypothesis that what you have are *two samples from the same population*
- "Null hypothesis": no difference in average pain reduction (and none in variation of response) between the two groups
- Measure average pain reduction level in treatment group: X_t
- and similarly in the placebo group average: X_p
- The parameter of interest $X_t - X_p$ will vary from trial to trial (?)
- But Null hypothesis predicts the observed values will be distributed around a mean of 0
- And with a dispersion, related to, but smaller than the population dispersion - in fact for sample size n, the SD of the various samples will be the SD of the population $\sigma/\sqrt{n}$ (sample SD usually called 'standard error')
- Also tricky because you generally have to estimate the population variance from the sample

A Scientific Example

- Various assumptions lead to different mathematics and so to different tests
- But underlying logic is always the same
- E.g in the ubiquitous ‘t-test’, the null is “rejected” if the actually observed value of $X_t - X_p$ is 'sufficiently far' away from zero
- Meaning, as we saw, that the probability of the observed value when added to the probabilities of even ‘more extreme’ values carries 5% or less of the total probability. (Assuming a 5% significance level.)
- In the simplest case this will be when the observed value $X_t - X_p$ lies outside the interval 0 +/- 1.96 S.E.


mean
Symmetric about mean
About 68% within 1 s.d. of mean
About 99% within 3 s.d. of mean
About 95% within 2 s.d. of mean
$\bar{x}-3\sigma$
$\bar{x}-2\sigma$
$\bar{x}-\sigma$
$\bar{x}$
$\bar{x}+\sigma$
$\bar{x}+2\sigma$
$\bar{x}+3\sigma$



(a) One-tailed test

(b) Two-tailed test

Is the underlying logic of significance testing sound?

- At first glance, significance testing fits well with the intuitive idea about a 'severe test' that we gathered earlier
- Suppose the test reveals a positive difference in the means: $X_t - X_p > 0$
- This is of course compatible with the treatment being better than placebo
- But it supplies little support for that claim unless that same outcome 'rules out' the rival theory
- (here that the treatments are equivalent and the improved outcome in the treatment group was 'due to chance')
- And a 'SS' result at least 'tells against' the 'chance' (= 'null') hypothesis
- However, when we examine the underlying logic of the method in more detail, any number of conceptual issues arise.

Is the underlying logic of significance testing sound?

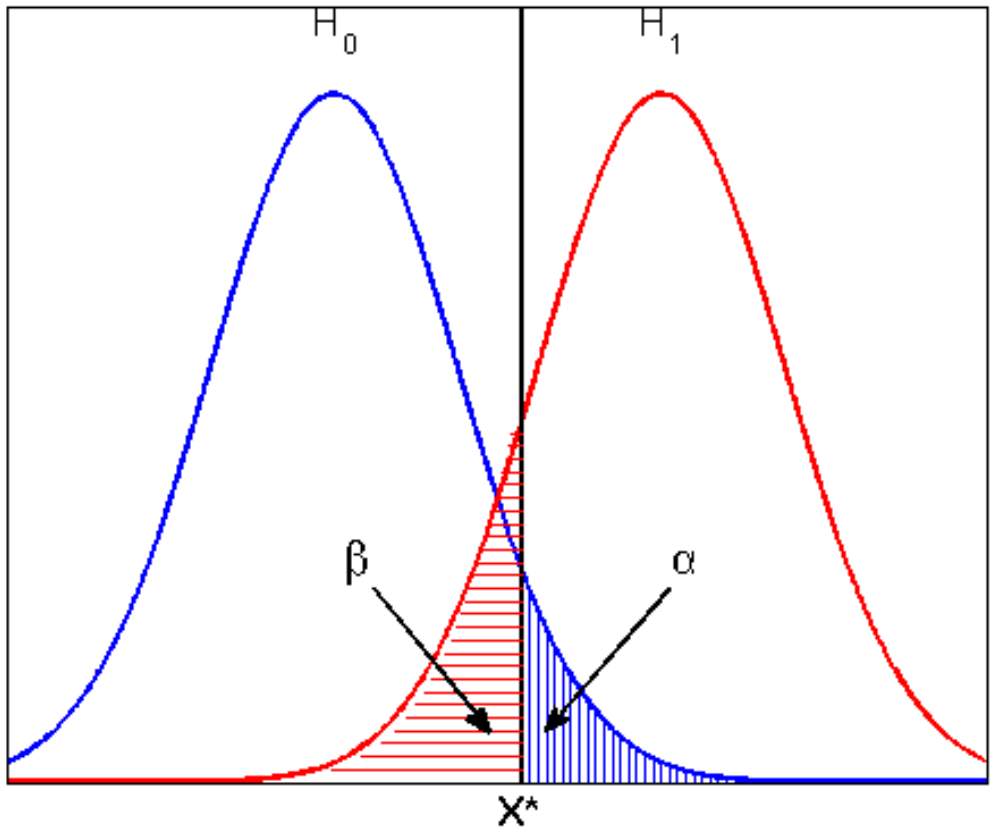
- For a start, Fisher actively resisted the suggestion that we should think about an alternative hypothesis as involved in such tests
- For him, you are always just testing the null hypothesis
- And the only thing you are entitled to say – if the evidence warrants it (i.e outcome is 'SS') is something negative about the null hypothesis
- And *nothing* positive about any other hypothesis

Is the underlying logic of significance testing sound?

- Moreover (deep conceptual problems)
- 1. The most straightforward negative thing to say (usually contra Fisher!) would be that, although the null is of course not strictly refuted, we are entitled reasonably to think of it as false.
- 'Cut off point'
- BUT ... The Lottery Paradox
- 2. Why should other possible results that might have but did not occur play any role in our evaluation of the evidential impact of some observed outcome e?
- 3. This actually has more definite consequences: it means that it is possible for a result to be 'SS' for h relative to one protocol for testing and yet the very same result fails to be SS for the very same h relative to another protocol

Is the underlying logic of significance testing sound?

- 4. Contrary to Fisher, it does seem that an alternative hypothesis is always in fact involved (otherwise why do one tail rather than two tail tests?). **And** that we do (at least want) to say something positive about the alternative if the null is ‘refuted’.


H0
H1
β
α
X*

Is the underlying logic of significance testing sound?

- 4. Contrary to Fisher, it does seem that an alternative hypothesis is always in fact involved (otherwise why do one tail rather than two tail tests?). **And** that we do (at least want) to say something positive about the alternative if the null is 'refuted'.
- But that (Neyman-Pearson) move
- (a) as Fisher emphasised, involves a very different 'logic' and
- (b) produces conceptual 'oddities' of its own:
- If we are testing one theory against another it shouldn't surely matter which we regard as the 'null' (really 'the theory under test') and which the 'alternative'
- BUT it is straightforward that there are cases where an observed result e is SS (and hence leads to rejection of one theory and 'acceptance' of another) when one of a pair of theories is regarded as the null
- ANY YET is 'statistically insignificant' if we think of the other as the null!

Is the underlying logic of significance testing sound?

- 5. ‘Statistically significant’ by no means entails ‘significant’
- 6. (Most fundamentally) a significance test is not telling an experimenter what she really wants to know:
- Tempting to think that e is significant at the p=0.05 level for h, means ‘e shows that there is a 95% chance that h is false’
- IT DOESN’T: all it means is that e (together with ‘more extreme’ possible, but non-actual alternatives) carries less than 5% of the total probability, **if h is true**
- In order to get from this to what you would really like, you would need to invoke Bayes’ theorem
- And that approach Fisher unambiguously rejected as completely unsound.

Examples of worrying conceptual problems

- E.g. 3:
- Hypothesis: P(H)= ½
- Test 1: toss the coin 20x
- Test 2: toss the coin until the point at which 6H have been observed
- Outcome:
- (6H,14T)

Examples of worrying conceptual problems

- The statistician doing test 1 will find the result ‘insignificant’ and hence ‘accept’ the hypothesis that it is fair
- The statistician doing test 2 will find the result ‘significant’ and hence ‘reject’ the null
- Yet it is the *very same* result
- Moreover, you might not be able to tell which test was being done
- [Difference is in the ‘at least equally extreme’ outcomes that *might have* ***but did not*** occur.
- Possible outcomes for test 1 are (20,0), (19,1) ….. (1,19), (0,20)
- Possible outcomes for test 2 are (6,0), (6,1), (6,2) …..
- On trial 2 the ‘at least as extreme’ outcomes are (6,14), (6,15), (6,16) ..
- Their probabilities add to 0.0319]

Examples of worrying conceptual problems: ‘Statistically significant’ does not entail ‘significant’

- One reason is that it would be remarkable if there were really *zero* difference between two ‘treatments’
- That is , if a ‘null hypothesis’ were literally true
- But if there is a difference, no matter how tiny, then a statistically high-powered (large sample) test is very likely to “find it”
- The result would be SS but might be of no real significance at all
- Effect size (or ‘likely effect size’) and statistical significance are two quite separate things

‘Statistically significant’ does not entail ‘significant’

- And neither does non-SS entail ‘not significant’!
- Suppose you have very good reason to believe ahead of the test/trial that some new treatment is really effective
- Yet the test fails to refute the null
- That surely, again *pace* Fisher, gives you some reason to think that the null might be wrong
- And since that would be surprising, the result is surely significant in the general sense
- (Need for judgment)

Lessons for ‘evidence savvy’ Policy makers

- 1. Don’t be fooled by “impressive” (i.e. small) p values, you want to know whether the evidence is really significant , not just statistically significant.
- 2. Look at *all* the evidence, don’t ignore ‘negative’ (“insignificant”!) evidence.
- (so you should worry about meta-analysis)
- 3. Don’t fall for the tempting ‘ the evidence is significant at the x% level’ therefore ‘it’s x% likely that the null is false (and the ‘alternative’ true)’
- (a) there are lots of alternatives!
- (b) it’s a fallacy anyway

Lessons for ‘evidence savvy’ Policy makers

- 4.However, it would also be a mistake to be *too* negative about the impact of SS results
- Suppose we accept
- (i) that ‘all’ an SS result tells you that p(e/h) is low; and
- (ii) that you really want to know p(h/e); and
- (iii) that you need further information to get to what you want to know
- STILL: that further information may be available
- Indeed you may reasonably think you have it!
- Blue cabs/yellow cabs again
- (Reason why Kahnemann and Tversky stuff may be questionable and the ‘base rate fallacy’ may not always be a fallacy)
- But still it is something else you MUST as an evidence savvy policy person think about.

Ethics & evidence can get intertwined: a case study

- Suppose you are Chair of the Ethics Committee at some University Hospital
- A researcher is seeking permission to recruit patients for an RCT on some new treatment
- She appears before the committee and says ‘ We are really excited about this trial because we are sure that this new treatment marks a big advance.’
- Reaction 1: ‘Hold on, if you are sure that the treatment is a big advance, how about those patients in your trial who would be randomized to the control arm? Wouldn’t you be consigning them to a treatment you were sure was inferior? And isn’t that contrary to the Hippocratic Oath?’
- Reaction 2: ‘will you tell the patients that you are sure that the new treatment is a big advance (‘informed consent’)? If so, you may have recruitment problems!’