Ph458/213 Evidence and Policy
Lecture 3
Jan 26

John Worrall
LAK 3.02
Office Hours:
Monday 13.30-14.30
Thursday12.30-13.30

Probability Theory

- P(A∩B)?? (equivalently P(A&B))
- probability of drawing a card that is *both* a club *and* at the same time at least as high in value as a Jack?
- Only 4 cards satisfy both conditions so P (Club & ≥Jack) = 4/52
- How does this relate to P(Club) and P(≥Jack)?
- Simpler case: tossing a fair coin twice
- $P(H_1) = P(H_2) = ½$
- $P(H_1$ & $H_2) = ¼$
- So(??) P(A&B) = P(A).P(B) ??
- Gives right result also with P (Club & ≥Jack)

Probability Theory

- BUT
- P(Heart & Red) ?
- Intuitively problem is that there is a connection between ‘Heart’ and ‘Red’
- Cp 1st and 2nd tosses of coin

Probability Theory

- Probabilistic dependence/independence
- ***Correct Law for Joint Probabilities***: for any two events A and B, P(A&B) = P(A).P(B/A)
- Special case: IF A and B are *independent*, then P(A&B) = P(A).P(B)
- Since A and B are independent iff P(B/A) = P(B)
- Gives correct answers in all cases – e.g. P(Heart & Red)
- Since A & B is the same event as B & A we should also have
- for any two events A and B, P(A&B) = P(B).P(A/B)
- And that must imply P(A/B) = P(A) iff P(B/A) = P(B)
- Finally correct law yields a 'definition' of conditional probability
- P(B&A) = P(B).P(A/B) and so
- P(A/B) = P(B&A)/ P(B) = P(A&B)/ P(B) ***IF*** P(B) ≠ 0
- [For Probability axioms and some theorems – see further reading.]

“Inverse Probability”

- Back to coin tossing:
- Given: coin is fair
- Trial: toss it 4 times
- Statistic: r= number of heads out of 4
- Given that the coin is fair we can readily calculate the probabilities (“direct” probabilities) for all possible values of r
- P(4 heads) = P(0 heads) = $(½)^4$ = 1/16
- P(3 heads) = P(1 head) = 4/16
- P(2 heads) = 6/16

“Inverse Probability”

- Now suppose we have another coin which we are told is biassed
- In fact P(H) = ¾
- Same trial, same statistic
- Again easy to calculate the direct probabilities
- P(4 heads) = $(¾)^4$ = 81/256
- P(3heads) = $4.(1/4).(3/4)^3$
- P(2heads) = $6.(1/4)^2.(3/4)^2$
- P(1 head) = $4. (1/4)^3.(3/4)$
- P(0 heads) = $(1/4)^4$

“Inverse Probability”

- But ***now*** suppose we have both coins in a box, that they are physically indistinguishable, and that we draw one at random and then toss it 4 times
- Suppose the outcome is r = 1
- ***What’s the probability that it was the fair (/biased) coin that we tossed to produce that result?***
- Bayes’ Theorem allows us to answer.
- Bayes’ Theorem: P(A/B) = P(B/A).P(A)/ P(B).

“Inverse Probability”

- Bayes’ Theorem: $P(A/B) = P(B/A).P(A)/ P(B)$.
- Let A be the event that it was the fair coin that we tossed
- (So ¬A is the event of having tossed the biased coin)
- And let B be the observed event: $r = 1$
- We know $P(B/A) = ¼$
- And, given that the coin was chosen from the box ‘at random’ [??] $P(A) = ½$
- So we need one more probability to apply the theorem – viz
- P(B)
- What was the probability of getting one head out of 4 *whichever* coin was tossed?

“Inverse Probability”

- It’s a general result (‘Theorem on Total Probability’) that
- P(B) = P(A).P(B/A) + P(¬A). P(B/¬A)
- So here
- P(B) = ½.P(B/A) + ½.P(B/¬A)
- (Why this is intuitively correct)
- So P(B) = ½ . ¼ + ½ . 4. $(1/4)^3$.(3/4) = 38/256
- So, remember, Bayes’ Theorem is
- P(A/B) = P(B/A).P(A)/ P(B)
- Plugging in these values we get
- P(A/B) = ¼. ½ /(36/256) = 256/304
- So we have a ‘prior’ of ½ and a ‘posterior’ of over 2/3
- So probability has gone up that it is the fair coin even though an event occurred that was relatively unlikely to occur if it was the fair coin.
- (Why this is intuitively correct.)

The Bayesian approach to confirmation

- Rewriting Bayes' theorem to apply to the case of general interest where we have some hypothesis h and some evidence e, we have
- P(h/e) = (P(e/h).P(h))/P(e)
- P(e/h) is the "likelihood" of the evidence in the light of h
- P(h) is the "prior probability" of h
- P(e) is the "prior probability" of e
- ***Fundamental Bayesian Principle:*** e confirms h if and only if ***and to the extent that*** P(h/e) > P(e)
- ***N.B. 'Confirms' ≠ 'makes it more likely than not to be true'***

Testing Probabilistic Theories

- So in probability theory, we take certain basic probabilities (the distribution over the outcome space) as given and then work out, *purely deductively*, what the probabilities are of various other events; what the probability is of one event, conditional on another event; ...
- But in the real world, we don’t know what the basic probabilities are – we can only have theories about them
- And then we want to look for evidence that those theories are true
- So now we are doing “induction” rather than deduction
- Statistics rather than probability theory

Testing Probabilistic Theories

- Inverse probability (Bayes’ theorem) seems to be taking us in the inductive direction (though clearly need further input)
- BUT Fisher unambiguously rejected ‘inverse probability’
- Instead he held that the only defensible way to proceed was via *tests*
- Fisher’s methodology (albeit in a ‘hybrid’ form) with its ‘p values’ and ‘95% confidence intervals’ permeates medicine, psychology, and more recently a wider swathe of social science, notably including development studies
- Thousands and thousands of policy decisions – grand and small - have been based on evidence underwritten by that methodology
- So we had better have a clear idea of
- (a) what the methodology is; and
- (b) what its RATIONALE is.

Significance tests

- So back to coin-tossing
- But now we don’t take ourselves as knowing that the coin is fair
- Instead that’s a *hypothesis* (p(H)= ½) and we want to look for evidence for it.
- The obvious thing seems to be to toss the coin some reasonably large number of times and see what the outcome is
- Suppose n = 20
- One immediate problem is that if the full result is recorded (say H, H, T, H, T, T, T, H, ...) then the hypothesis assigns the very same probability to that result as to any other conceivable result viz $1/2^{20}$
- Fisher suggested that instead we should operate with a summary of the outcome – a “statistic”
- Here the natural one seems to be r = the observed frequency of heads

Significance tests

- r , unlike the total outcome statistic, is non-uniform
- But still: unlike in testing deterministic theories, *any* conceivable observed value of r is consistent with the hypothesis p(H) = ½
- Fisher argued that we should nonetheless proceed in falsificationist manner
- The underlying idea seeming to be that we will never learn anything of a probabilistic nature unless we adopt a *rejection rule* (or rather a rejection "rule")
- If we observe an outcome (value of r) that is 'is very unlikely to occur if the hypothesis were true' then we are entitled to reject the hypothesis
- PROBLEM: any observed value of r – including the most probable – is very improbable!
- p(r=10) = 0.1762 [and if the number of tosses were much higher ...]

Significance tests

- Fisher suggests:
- 1.probability of refuting outcome not just low, but low relative to other possible outcomes
- 2. we should look not just at the event that did occur but also events that might have but didn’t occur that are ‘equally or even more extreme’ [??]
- 3. The total probability of those events is the ‘p value’ of the result.

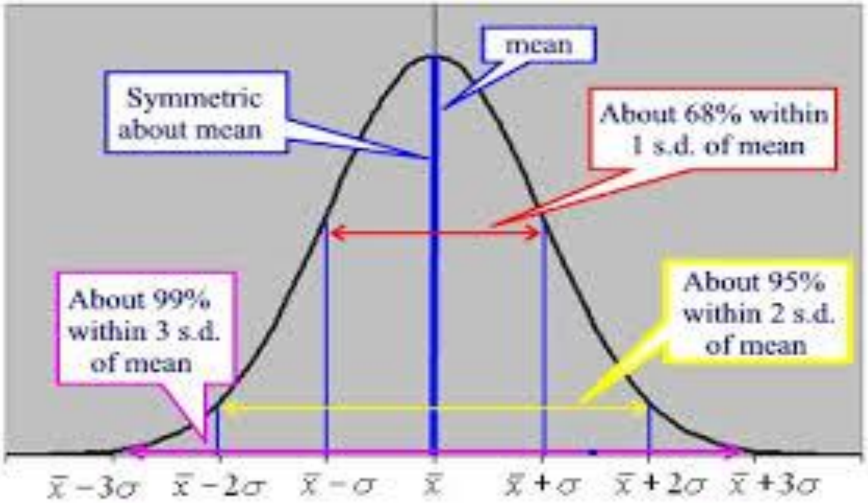
Significance tests

- So, suppose in our experiment, the outcome is r=4
- The outcomes with less than or equal probability to this are r= 4,3,2,1,0 and r= 16,17,18,19,20
- This yields p = 0.012
- Fisher actually held that just the actual ‘observed’ p value should be reported
- But the ‘logic’ that has come to be standard (and thought of as Fisherian) is
- 1. A (frankly conventional) cut-off point p*should be decided in advance
- 2. The usual conventions being either .05 or .01
- 3. The result is declared ‘statistically significant’ if the actually observed outcome gives a p value ≤ p*,
- 4. If so, the hypothesis under test is “rejected”
- In our case, r =4 would be ‘significant at the 5% level’ and so p(H)= ½ would be rejected
- (Though not if the significance level adopted had been 1% (p* = .01))

A Scientific Example

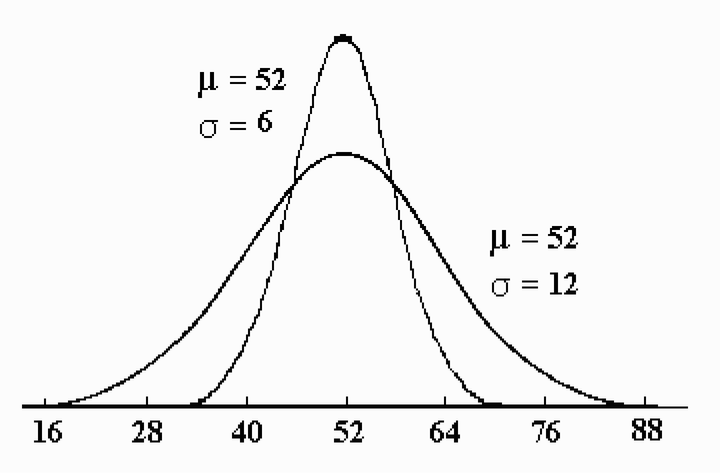
- Before investigating the logic here, let’s look at a more representative example
- Needs more work/assumptions and there are of course different cases
- But many (most?) of the cases of practical import in policy are thought of as testing for a *difference between 2 means*
- Many quantities have frequencies of occurrence that are distributed in a similar way (height, IQ, recovery time from colds)
- An average value of the quantity which is also most frequent
- And then frequencies that fall off symmetrically on either side of the mean.

A Scientific Example



A Scientific Example

- Any normal distribution is characterised by 2 parameters: the average (central) value or *mean μ* and the variance (or standard deviation, σ).



A Scientific Example

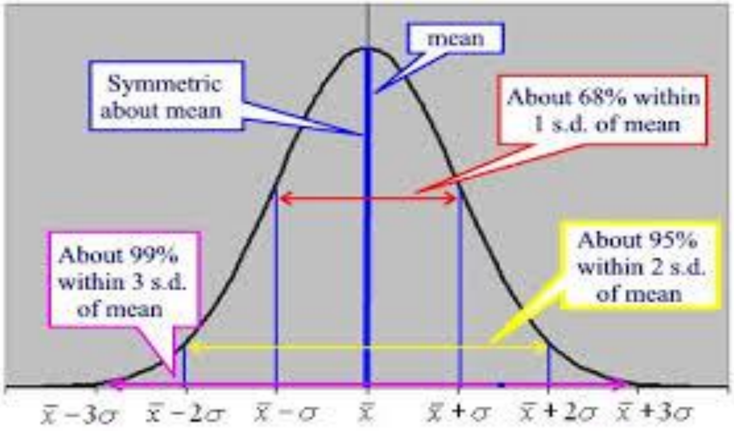
- Suppose a new drug is being tested
- Study population divided into two: experimental and control
- Some statistic is measured for each member of each group: for example reduction in pain level
- Standardly variation in response in both groups
- We would like to characterise situations in which we can 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
$\bar{x}-3\sigma$
$\bar{x}-2\sigma$
$\bar{x}-\sigma$
$\bar{x}$
$\bar{x}+\sigma$
$\bar{x}+2\sigma$
$\bar{x}+3\sigma$

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 mean
- 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 (actually 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’.
- But that (Neyman-Pearson) move 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.

Example of worrying conceptual problem

- 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)

Example of worrying conceptual problem

- 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
- 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 1 are (20,0), (19,1) ….. (1,19), (0,20)
- Possible outcomes for 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]