Problems with p-values and NHST and how to overcome them using likelihood ratios

Scott Glover

Outline

- Scientific inference
- Problems with NHST
- Problems with *p*-values
- Likelihood ratios
- How LRs address the problems that arise with NHST and *p*-values
- Summary

Scientific Inference

Science is the process of developing or updating beliefs about the world based on empirical evidence

We do experiments, we gather evidence, and we develop/update our beliefs

Statistics are a valuable tool we use in this process, but they are not infallible

Scientific Inference

Stroop effect

RED GREEN YELLOW BLUE

Scientific Inference

Why do we believe a Stroop effect exists?

Scientists have compared two opposing models of the world (i.e., either there is a Stroop effect or there isn't), and the data have been consistently against the model assuming no effect

Therefore, we infer that a Stroop effect exists and is real

Modelling the world

As scientists we try to provide and test scientific models of the world

There are different ideas of what these models should do

e.g., Popper view vs. Putnam view

Statistics are used to provide us with a guide to deciding between competing models

The use of statistics is valuable but can cause problems when the statistical method used is poorly devised, poorly implemented, or poorly understood

Null hypothesis significance testing

NHST evaluates the evidence in terms of the probability of the obtained result occurring if one model (the null or ‘Ho’) is true, and then making inferences about the truth of the other (alternative or ‘Ha’)

To do so, we calculate a *p*-value

If $p < 0.05$, we “reject the Ho” (and implicitly accept the Ha)

If $p >= 0.05$, we “fail to reject the Ho” (the results are inconclusive)

Problems with NHST

1. Asymmetry: We can only ever find evidence for one of the two models of interest (and this only implicitly by rejecting the other model)

	Model 1 Ho A does not affect B	Model 2 Ha A affects B
$P < 0.05$	rejected	(supported)
$P > 0.05$	fail to reject	(inconclusive)

Problems with NHST

2. Tries to apply black/white logic to grey probabilities

Either Ho or Ha is true

If the Ho is true, then $p < 0.05$ should not occur (except rarely)

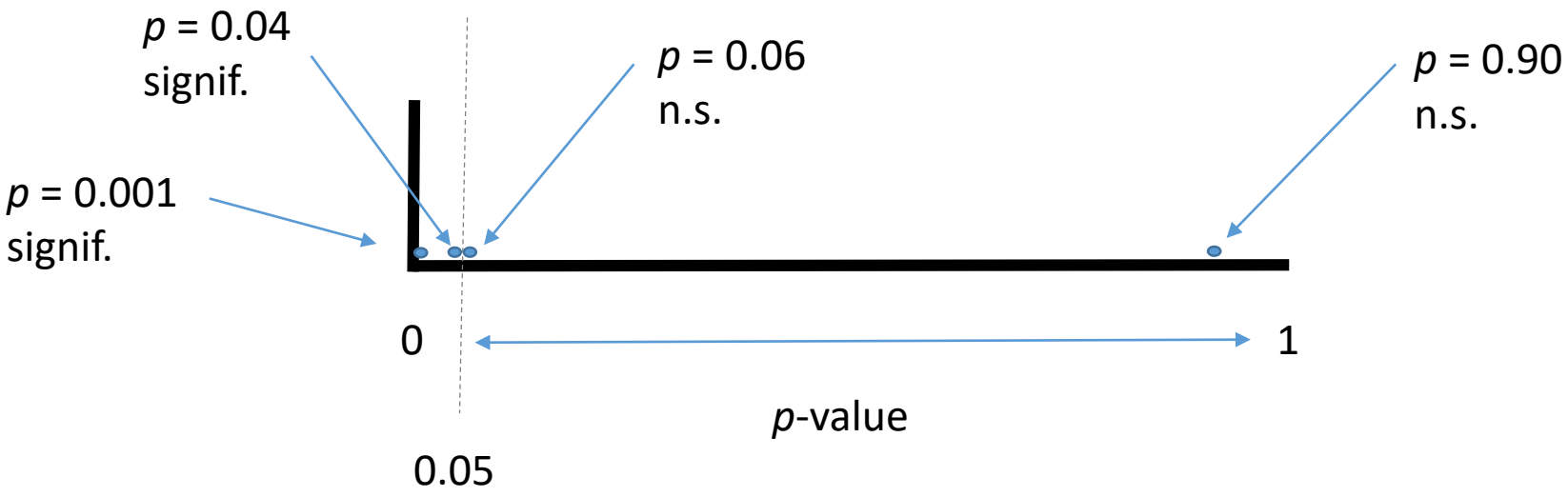
$p < 0.05$ occurred

Therefore, Ho is false (or, this is just one of those rare occurrences)
Therefore, Ha is true (or, this is just one of those rare occurrences)

Problems with NHST

3. Uses a binary decision process with an arbitrary decision criterion

Values very similar can have opposite interpretations, whereas values far from each other can have identical interpretations



Problems with NHST

4. Accepts fairly weak evidence as the standard

Even Fisher argued that $p < 0.05$ was only weak evidence for an effect

As we'll see later, this criterion allows an unacceptable number of false positives

Problems with NHST

5. Removes the process of scientific inference from a scientist who *can* think and puts it in the hands of a statistic which *can't*

Arguably, anyone who lets a statistic *alone* tell them what to believe isn't thinking very hard

Statistics only provide *estimates* about the existence and/or size of an effect (and sometimes not very reliable ones, and sometimes just plain wrong ones)

Summary: Problems with NHST

1. Asymmetry
2. Tries to apply logic to probabilities
3. Uses an arbitrary, binary decision process
4. Accepts fairly weak evidence as the standard
5. Takes the scientist's judgment out of the process of scientific inference

Problems with *p*-values

1. They are non-intuitive and thus misleading

We are taught that a *p*-value represents the probability of making a Type I error (finding an effect where none exists)

So, if $p < 0.05$ there must be a very strong chance the effect exists

Problems with *p*-values

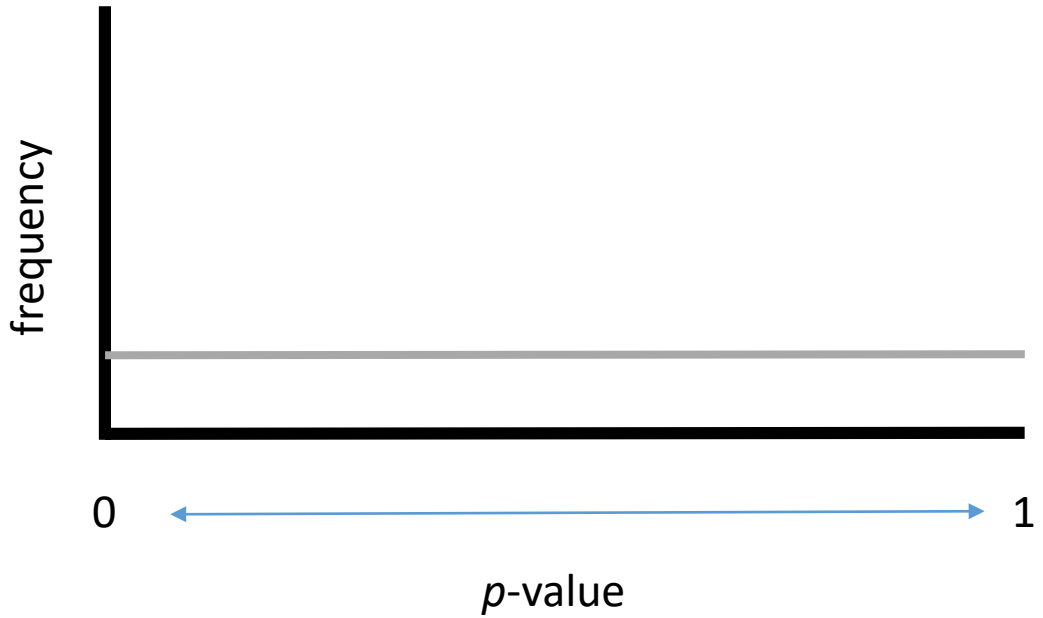
But it doesn't work this way…

The problem comes back to the asymmetry of NHST

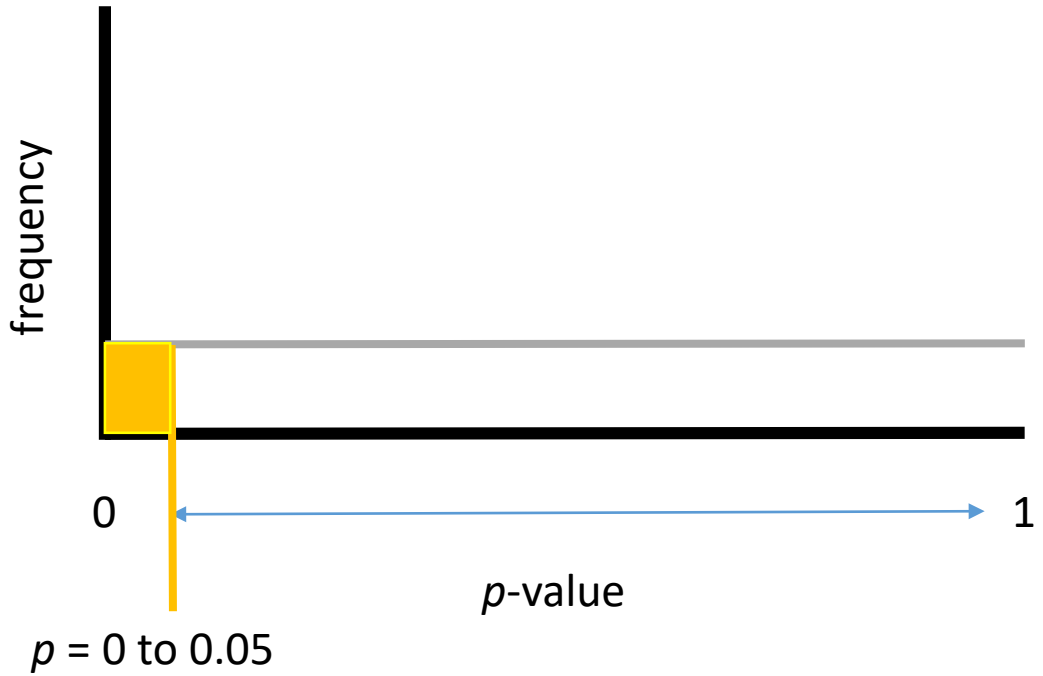
We are only testing one of the models

Thus, the *p*-value tells us how likely a result that or more extreme is *if and only if the Ho is true*

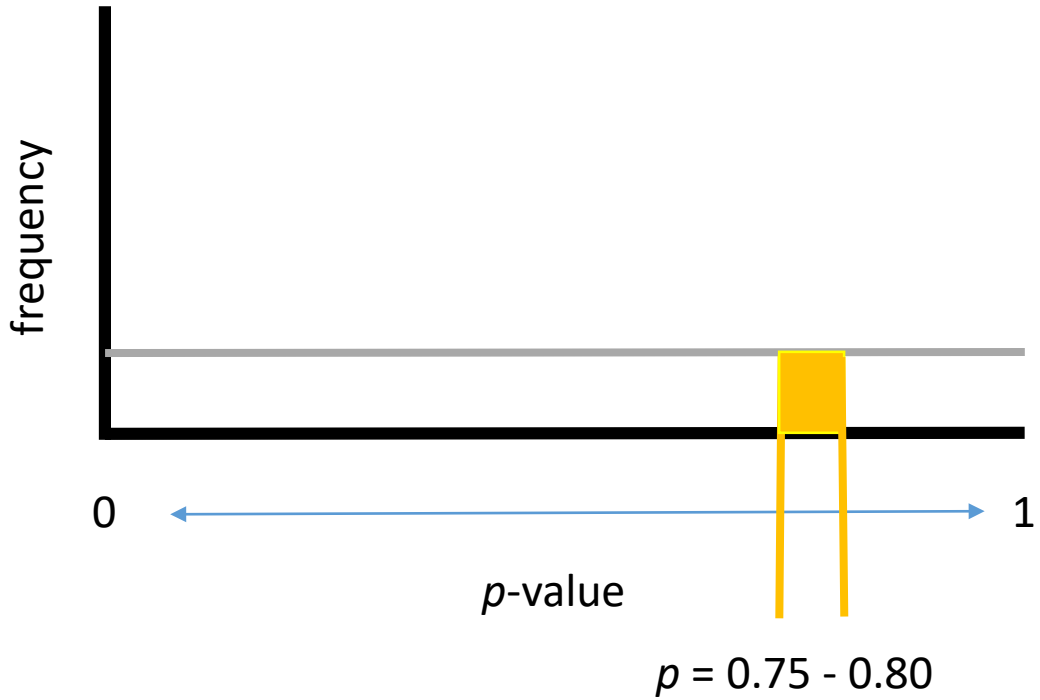
The distribution of *p* when the Ho is true



# *p* < 0.05 is meaningless *when the Ho is true*



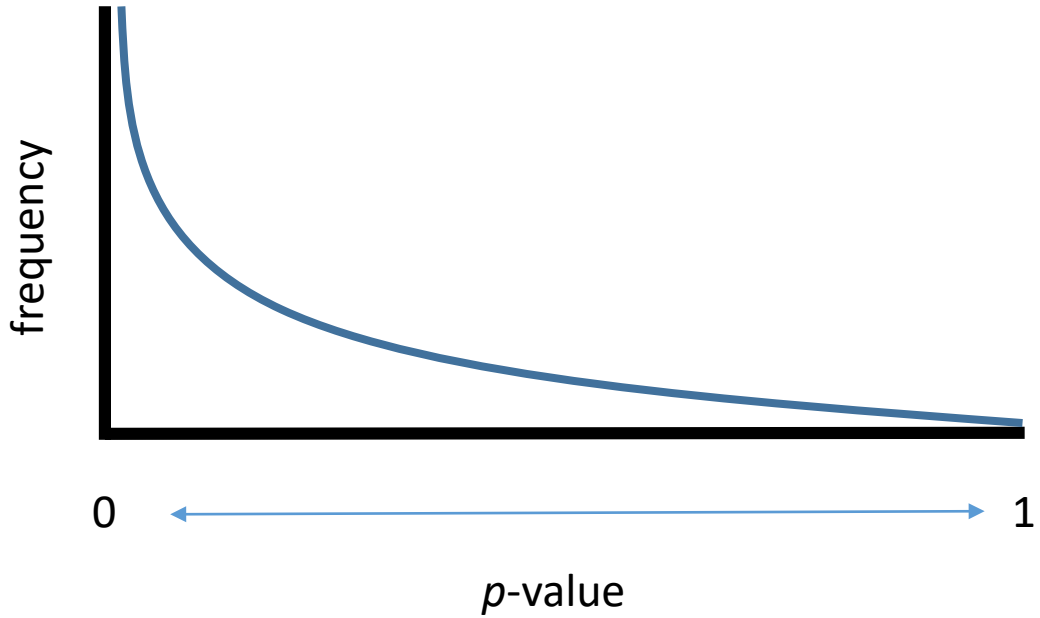
Any other result within an equally small range of *p*-values is equally rare *when the Ho is true*



The distribution of *p* when the Ho Is true

- Thus, the *p*-value when the Ho is true is *uninformative*
- So why use *p*-values at all?
- The distribution of *p*-values is *not* uniform when the Ha is true

The distribution of p when the Ha is true *



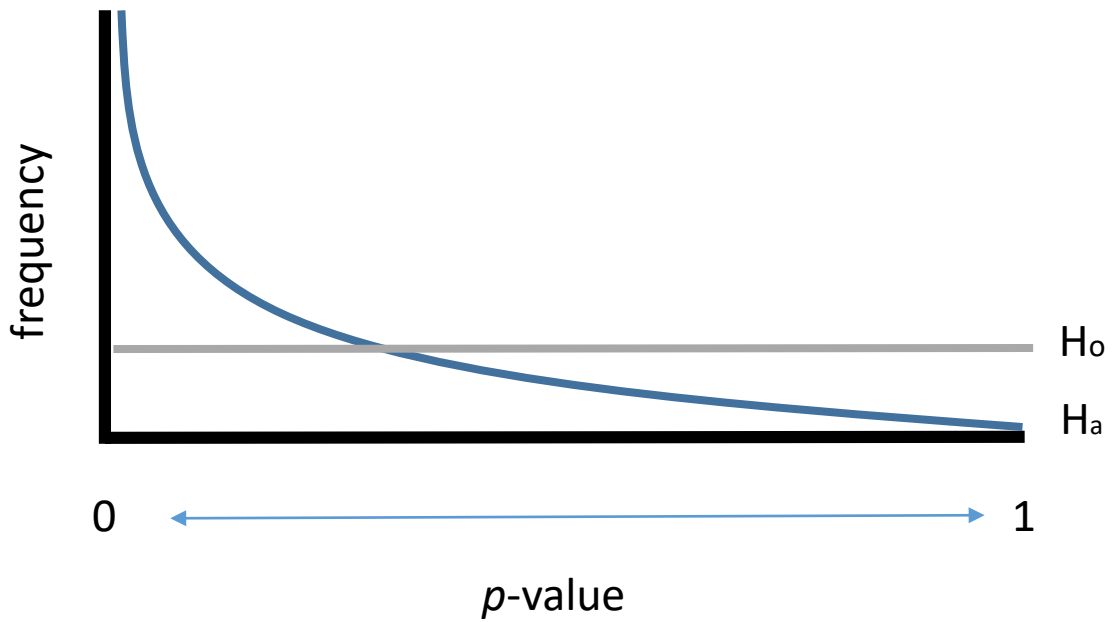
(* The shape of this distribution will vary depending on power)

The critical factor

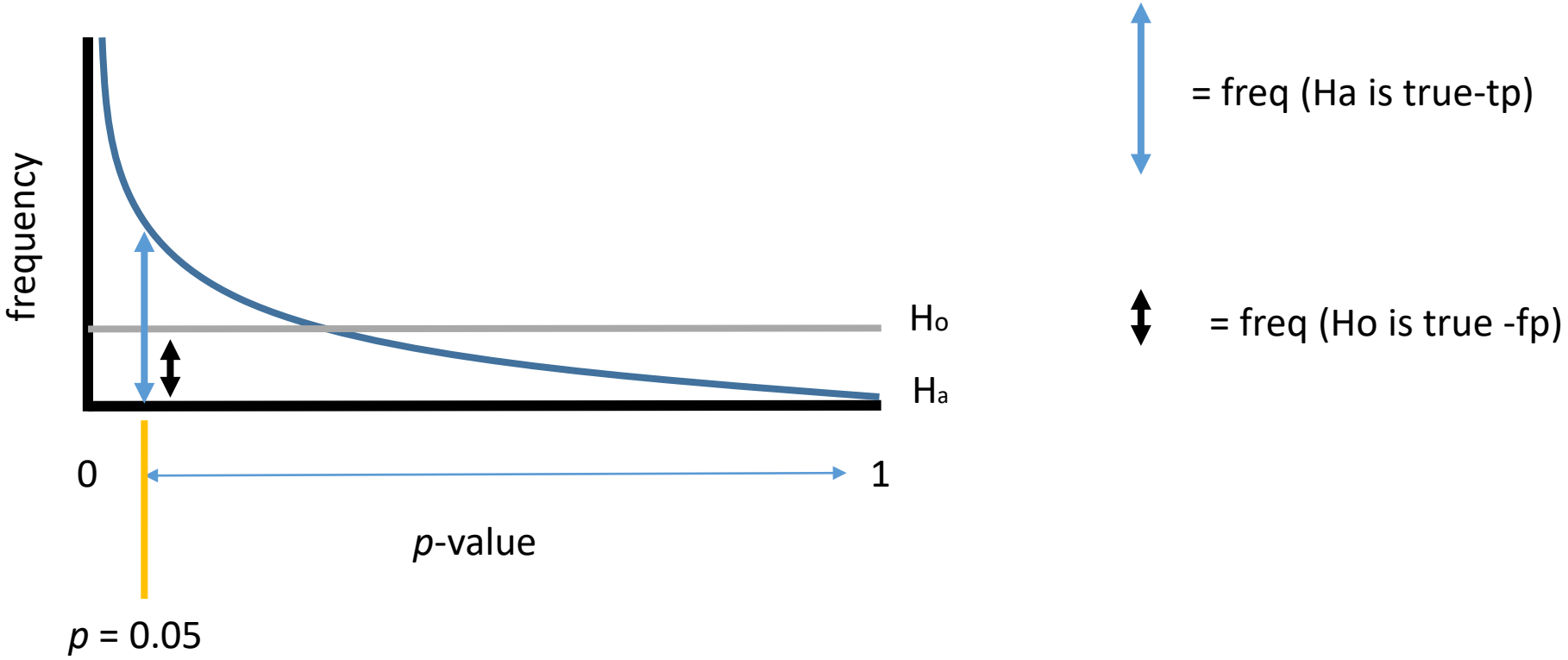
- The critical factor in deciding between two models is not simply *p*

- The critical factor in evaluating *p* is the *relative likelihood* of *p* depending on whether the Ho or Ha are true

Comparing distributions of p



Comparing distributions of p



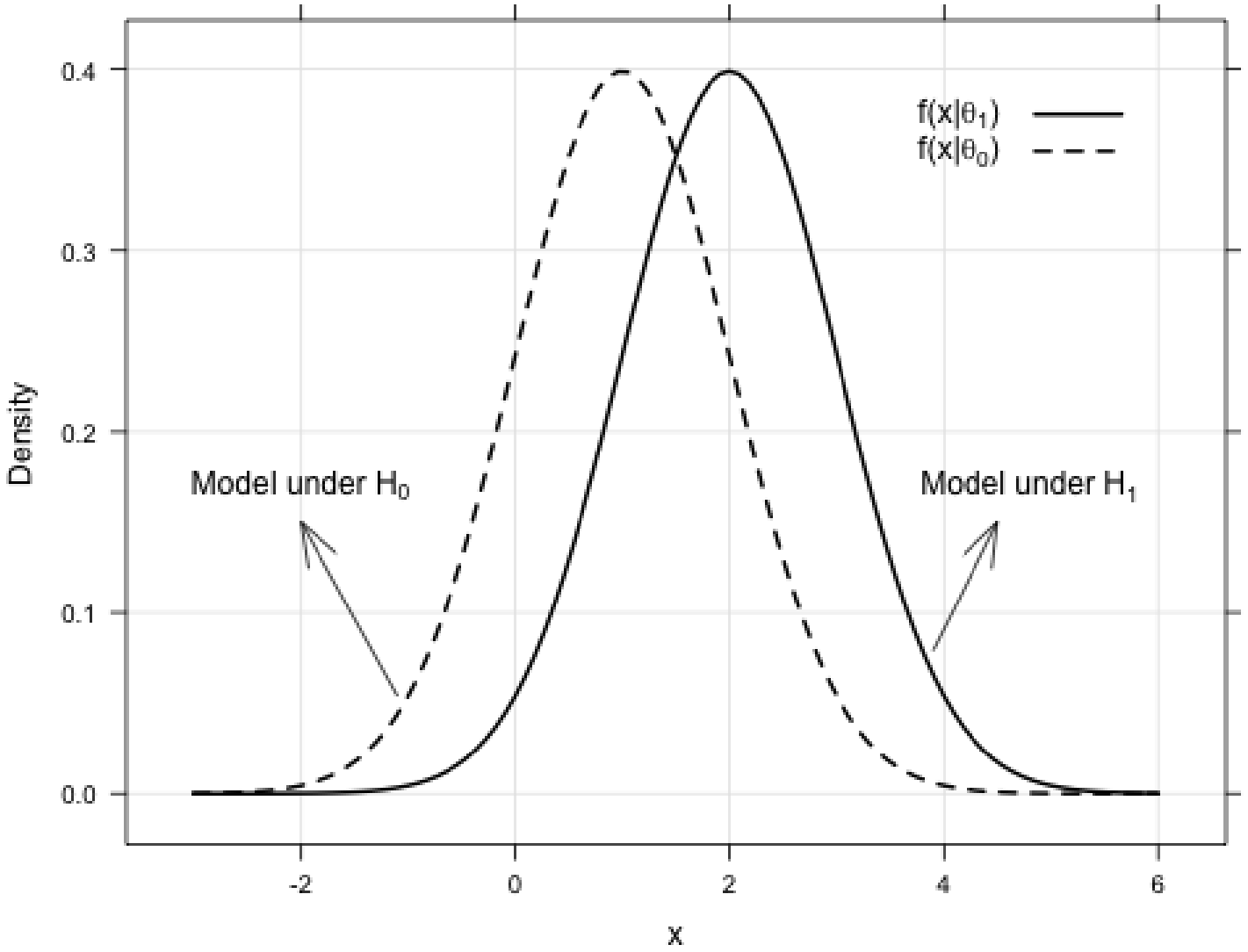
Some lessons from these distributions

- A *p*-value of 0.05 is only about 3x as likely to occur when the Ha is true as when the Ho is true
- $P < 0.05$ is probably too weak a criterion (as Fisher argued)
- The *p*-value is clearly not = the false positive rate – the chance of finding a signif. result when no effect exists
- More generally, a *p*-value is an indirect, non-intuitive and potentially misleading index of the strength of the evidence in favour of the Ha

Likelihood ratios

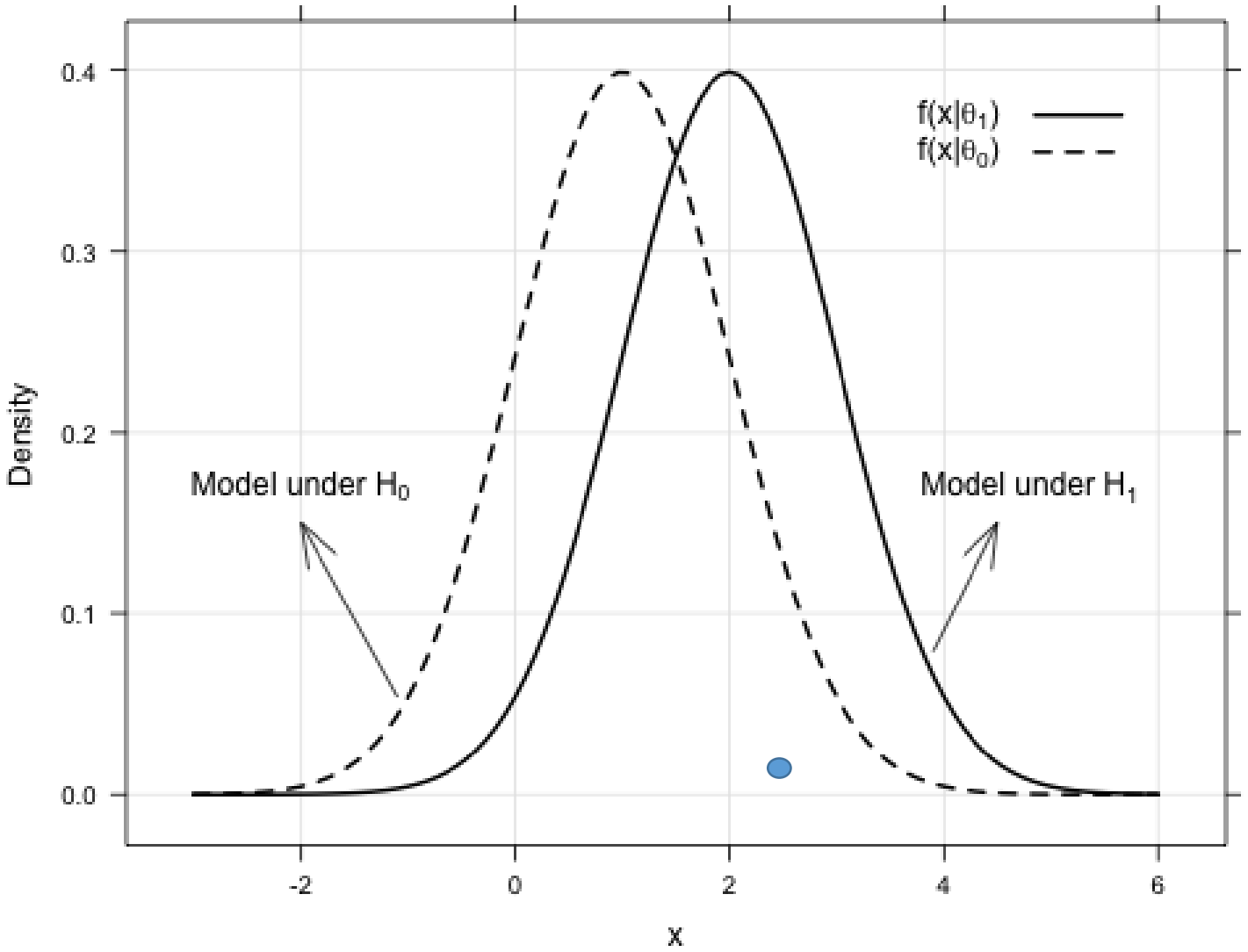
- Compare the relative likelihood of the data given two models, expressed as a ratio
- Can be set up to compare any two models
- A number of ways this could theoretically be done
  - E.g., comparing prob. freq. dist. for *p*, or another statistic such as t, etc.
  - Such methods generally give comparable answers to LRs, but are difficult to implement for various reasons
  - Likelihood theory holds that the data frequency distributions themselves should be compared

Likelihood ratios



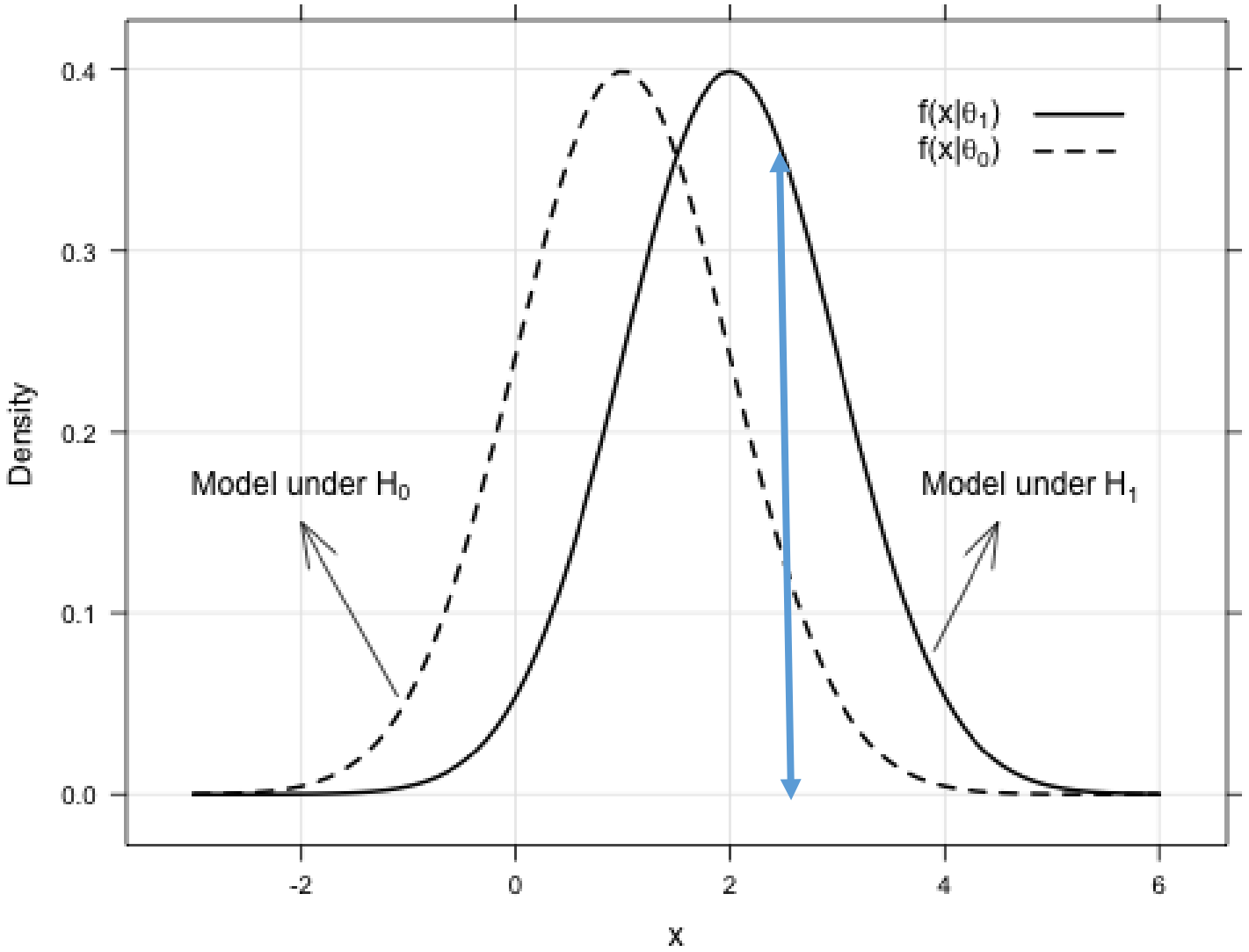
Likelihood ratios

X = 2.5



Likelihood ratios

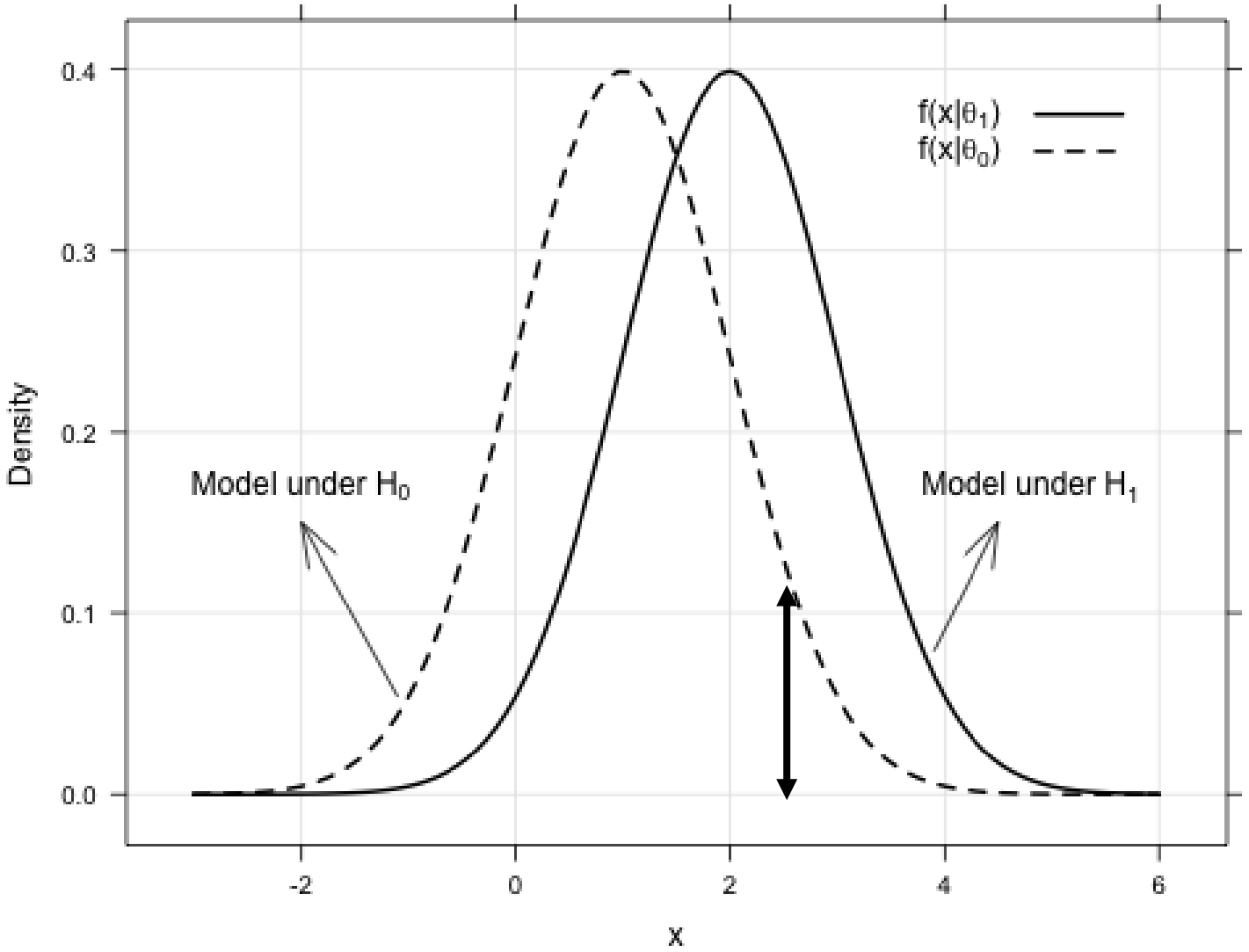
X = 2.5



Freq (Ha) = 0.33

Likelihood ratios

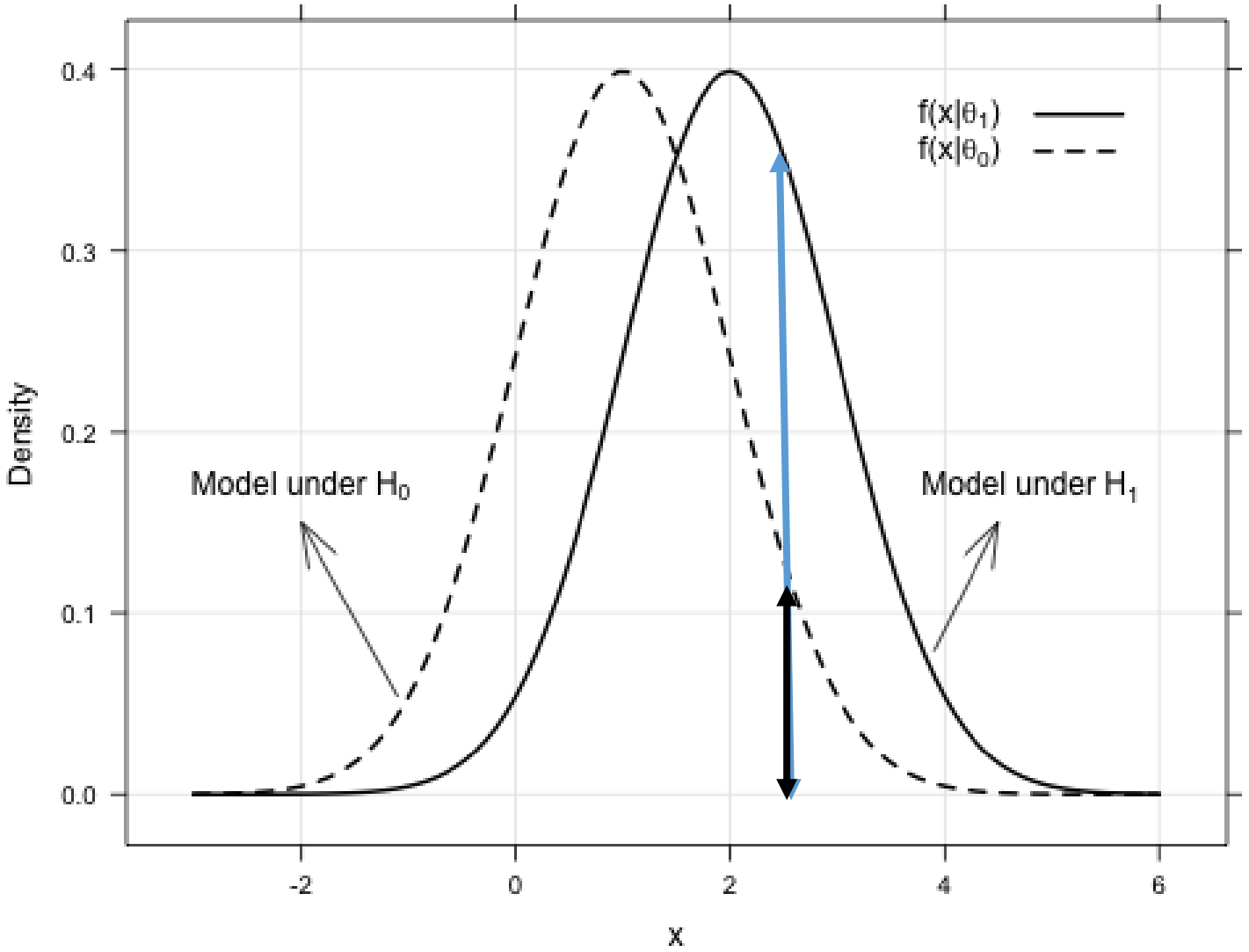
X = 2.5



Freq (Ho) = 0.11

Likelihood ratios

X = 2.5



:

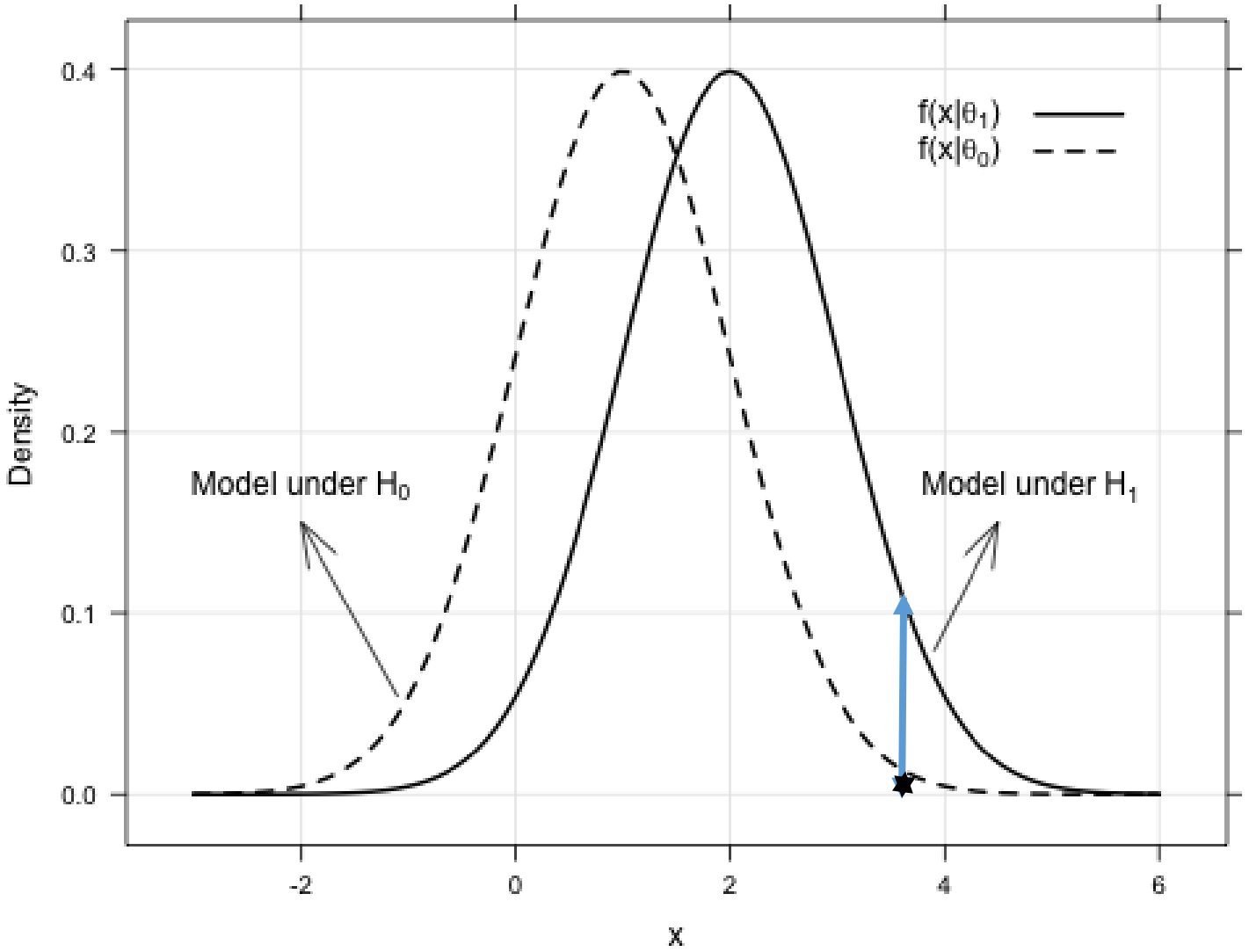
= likelihood ratio

λ = 0.33 : 0.11

λ = 3:1

Likelihood ratios

X = 3.7



: ★ = likelihood ratio

λ = 0.12 : 0.02

λ = 6:1

Calculating Likelihood ratios

Likelihood ratios are related to *p*-values, and in typical situations, can be approximated as:

$$\lambda_{adj} = \frac{1}{7p}$$

Comparing *p* and LR

From this the following translations are approximately correct

p	λ_{adj}
0.10	1.4: 1
0.05	2.7: 1
0.03	4.8: 1
0.01	13.5: 1
0.005	27: 1
0.001	135: 1

Likelihood ratios

- Can be precisely calculated from many common statistics

- from t- score $\lambda_{adj} = (1 + \frac{t^2}{n-2})^{\frac{n}{2}}$ [exp [$k_1 (\frac{n}{n\_k1\_1}) - k_2 (\frac{n}{n\_k2\_1})$]]

- from ANOVA $\lambda_{adj} = (\frac{unexplainedvariance_{M1}}{unexplainedvariance_{M2}})^{\frac{n}{2}}$ [exp [$k_1 (\frac{n}{n\_k1\_1}) - k_2 (\frac{n}{n\_k2\_1})$]]

Likelihood ratios

- from linear regression $\lambda_{adj} = \left(\frac{1-R^2_{M1}}{1-R^2_{M2}} \right)^{\frac{n}{2}} [\exp [k_1 (\frac{n}{n\_k1\_1}) - k_2 (\frac{n}{n\_k2\_1})]]$

- the second part of each equation is a correction for number of parameters in the model and leads to the “adjusted” LR

- this is done to penalise the model with more parameters (which will almost always fit the data better)

Likelihood ratios in practice: Hypothetical Experiment 1

- A researcher is interested in the effects of a particular distractor on RT, and if this interacts with quality of sleep

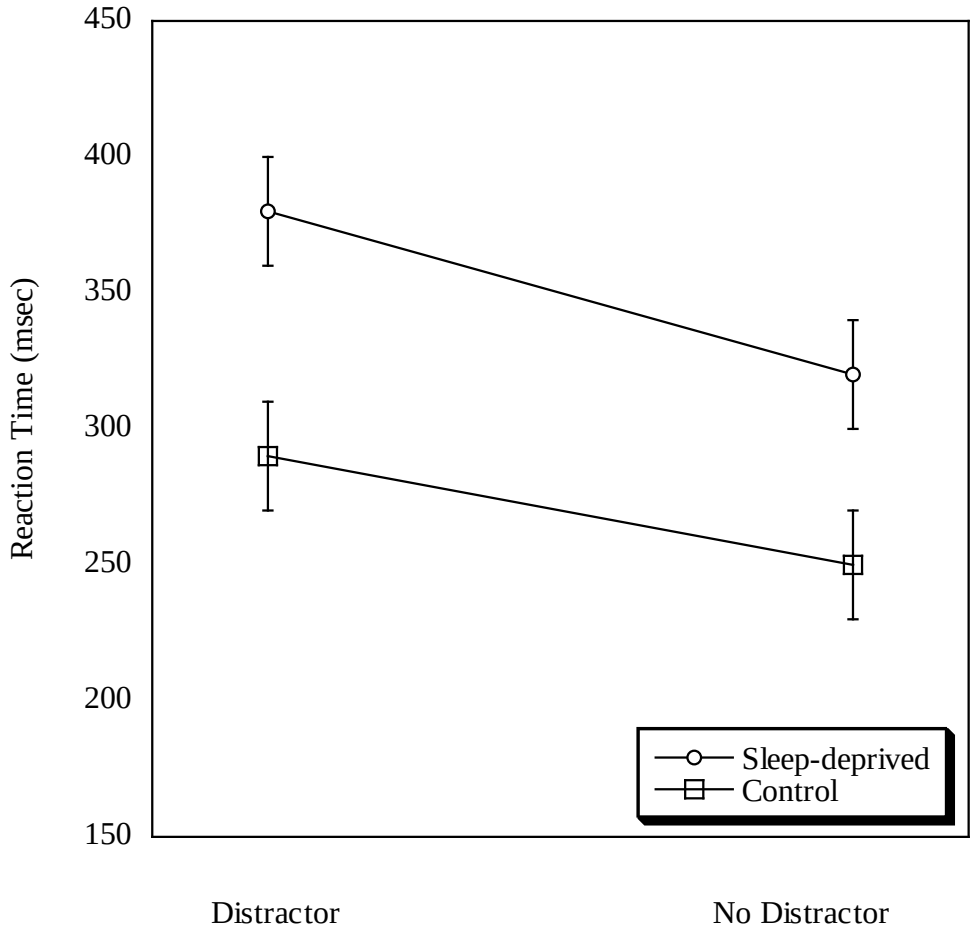
- Their hypotheses are:
- 1) RT will be greater when the distractor is present
- 2) RT will be greater when the participant is sleep-deprived
- 3) The effects will be superadditive (an interaction between distractor and sleep)

- Sets up an experiment with 25 ss/group

- Each participant reacts to a stimulus as quickly as possible

- Each is in either the distractor/no distractor and sleep-deprived/not sleep deprived group (2x2 b/w ss design)

Hypothetical Experiment 1: Results



Hypothetical Experiment 1: ANOVA output

Source	df	SS	MS	F	*p*
Distractor	1	260	260	11.30	< 0.01
Sleep	1	160	160	6.96	< 0.05
D X S	1	100	100	4.35	< 0.05
Error	21	483	23		
Total	24				

All hypotheses supported

Hypothetical Experiment 1: Likelihood ratios

- From ANOVA $\lambda_{adj} = \left(\frac{unexplainedvariance_{M1}}{unexplainedvariance_{M2}}\right)^{\frac{n}{2}} [\exp [k_1 \left(\frac{n}{n\_k1\_1}\right) - k_2 \left(\frac{n}{n\_k2\_1}\right)]]$
- Unexplained variance is SS not included in model
- For Hdist, Hsleep, and Hdxs, the unex. var. is the SSError = 483
- For the Ho, unex. var. in each case adds the SS for the effect (SSdist, SSsleep, or SSdxs, respectively)
- E.g., for comparing with Hdist, the unex. var. for the Ho is SSError+SSdist = 743

Hypothetical Experiment 1: Likelihood ratios

- For Hdist, $\lambda_{adj} = (\frac{743}{483})^{\frac{25}{2}}$ (exp (2*(25/22) – 3*(25/21)))

$\lambda_{adj} = 59.4$

So, the data are about 60 times as likely given an effect of distractor than no such effect

Hypothetical Experiment 1: Likelihood ratios

- For Hsleep, $\lambda_{adj} = (\frac{643}{483})^{\frac{25}{2}}$ (exp (2*(25/22) – 3*(25/21)))

$\lambda_{adj} = 9.8$

So, the data are about 10 times as likely given an effect of sleep than no such effect

Hypothetical Experiment 1: Likelihood ratios

- For Hdxs, $\lambda_{adj} = (\frac{583}{483})^{\frac{25}{2}}$ (exp (2*(25/22) – 3*(25/21)))

$\lambda_{adj} = 2.9$

So, the data are about 3 times as likely given an interaction between distractor x sleep than no such effect

Comparing NHST and LR

- NHST told us to reject the Ho (and implicitly, to accept the Ha) for all three tests
- An intuitive understanding of *p* would also suggest the Ho is very unlikely to be true (< 1% for the effects of distractor, < 5% for sleep and the interaction d x s)
- The LR tells us the evidence is truly compelling only for the effect of distractor (59.4: 1), good but not overwhelming for sleep (9.8: 1), and rather equivocal for the interaction d x s (2.9: 1)
- By evaluating the data in terms of both models rather than just the Ho, the LR gives a more intuitive appraisal of the strength of the evidence than NHST and *p*-values

Comparing *p* and LR

We tested the accuracy of the values provided by *p* and LRs using Monte Carlo simulations

Drew 100k samples from a Ho and 100k samples from a Ha distribution

Count the number of times we observe *p*-values in a given range

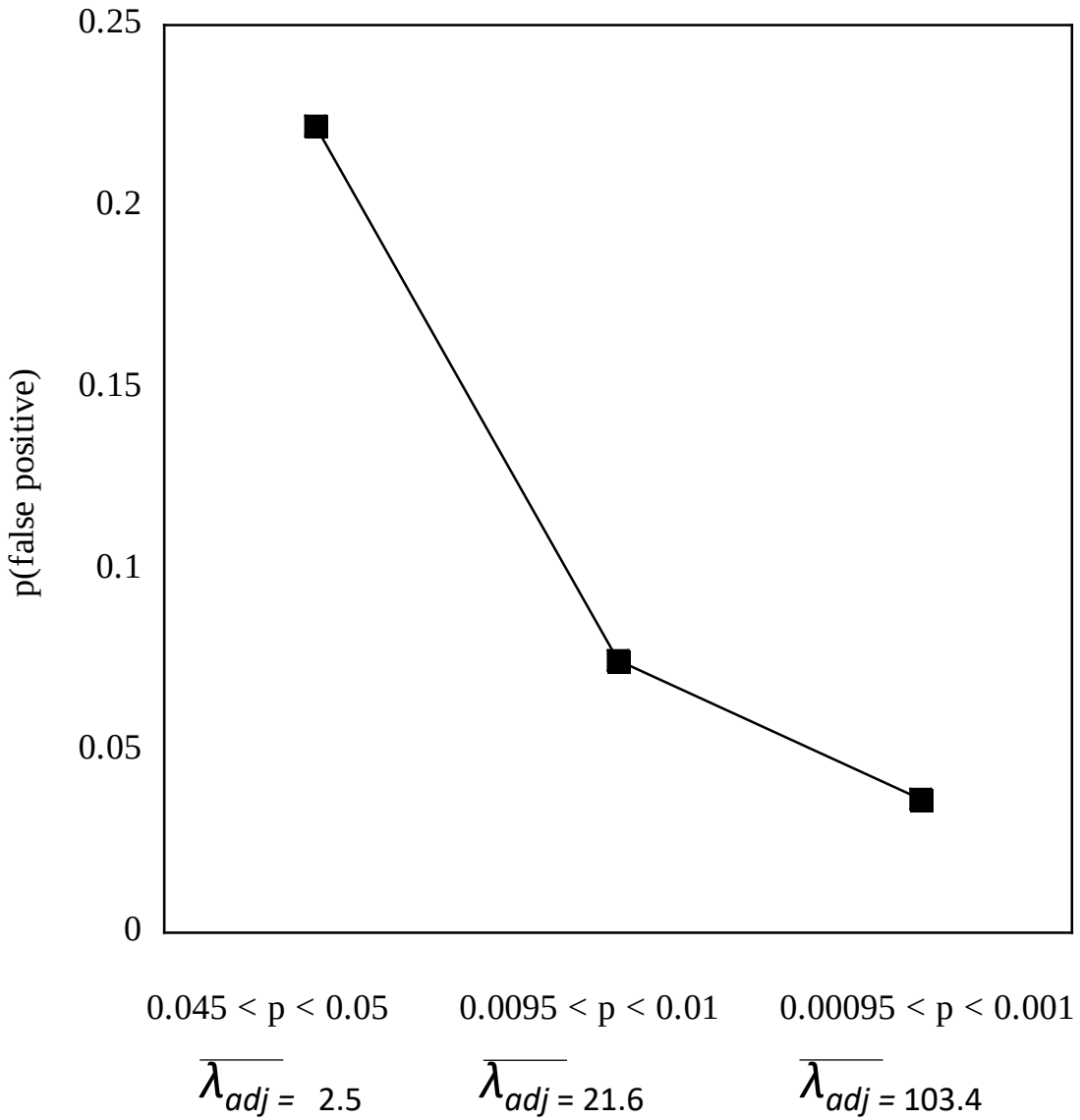
p = 0.045 to 0.05; *p* = 0.0095 to 0.01; and *p* = 0.00095 to 0.001

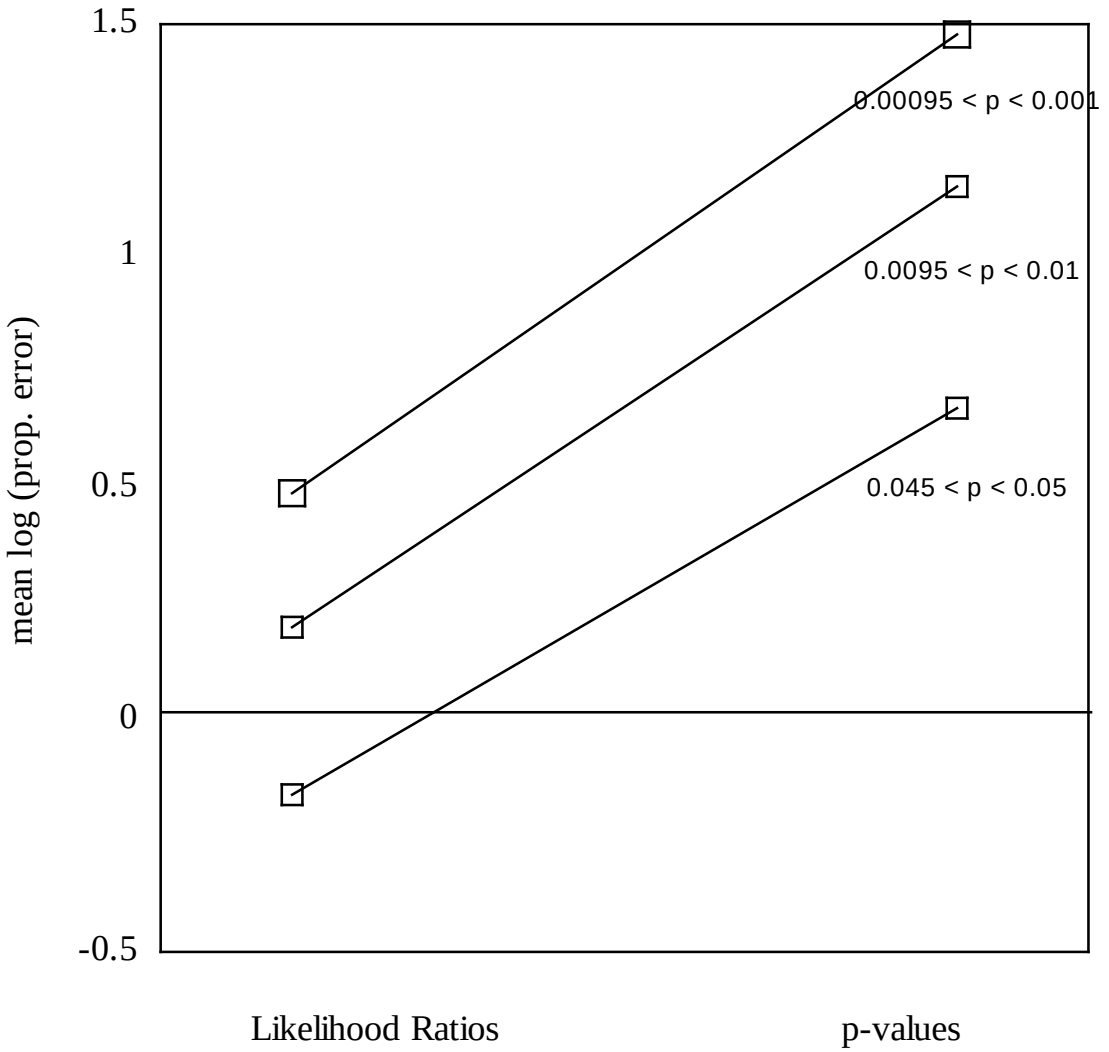
Count how many of these came from the Ho distribution (false positive) and how many from the Ha (true positive)

- compute a false positive rate

- calculate how accurate *p* and LR are in reflecting the fp rate

Comparing *p* and LR





Comparing *p* and LR

LRs are slightly conservative when *p* is close to 0.05, and become more liberal as *p* gets smaller

P-values are consistently liberal as an index of the false positive rate, overestimating the strength of the evidence in all cases

Likelihood ratios consistently provide a more accurate index of the strength of the evidence (measured through false positive rates) than *p*-values

Solving other Problems with using p-values and NHST

- NHST is asymmetrical – i.e., we can only ever find evidence to refute the Ho, never to support it
- Another problem is that it implies that a statistically significant effect is also an important effect
- A related issue is that any sized effect, with enough ss, can be statistically significant
- Using LRs and model comparison allows one to avoid all of these problems

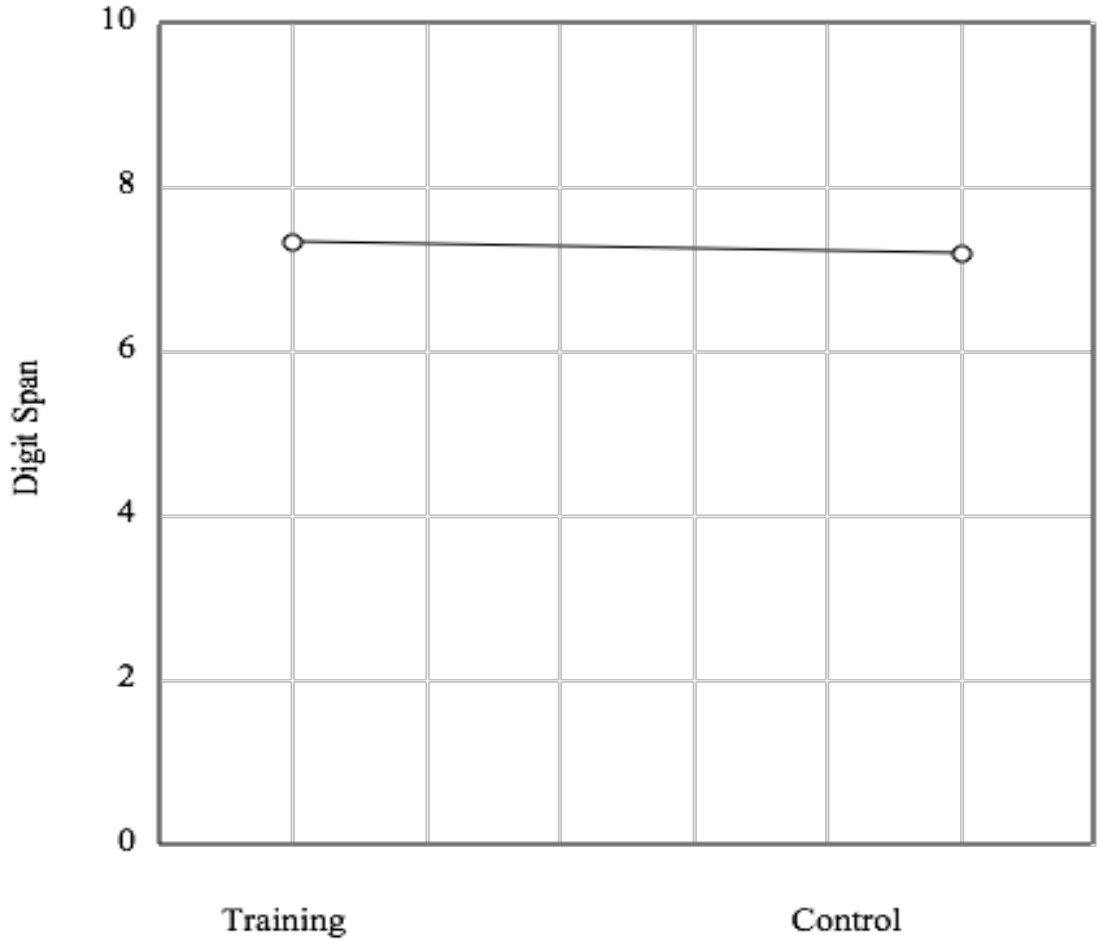
Hypothetical Experiment 2

- A researcher sets up an online experiment testing the effects of “brain training” on intelligence
- 5,000 ss work on various puzzles one hour per day for a month, another 5,000 do no training
- Digit span is compared between groups after the training period

Hypothetical Experiment 2

- The researcher finds a “significant” effect of brain training, $p < 0.05$
- Q: From this result, should we all do brain training?
- A: According to NHST, absolutely
- But let’s look at the data more closely…

Hypothetical Experiment 2: Results



Hypothetical Experiment 2

	Training	No training
Mean	7.344	7.20

t (9998) = 2.0

$p = 0.046$ (reported as $p < 0.05$)

$\lambda_{adj} = 2.7$

This small effect (0.144 digits, or a 2% difference between groups) was significant almost entirely because of the bigly no. of ss (5,000/group)

Hypothetical Experiment 2

- What we really want to know is whether there is evidence for an effect that is large enough to be theoretically interesting (Htie)?
- I.e., is there evidence the effect is large enough to cause a change in policy or practice?
- We cannot do this using NHST because it is only set up to test the fit of the Ho

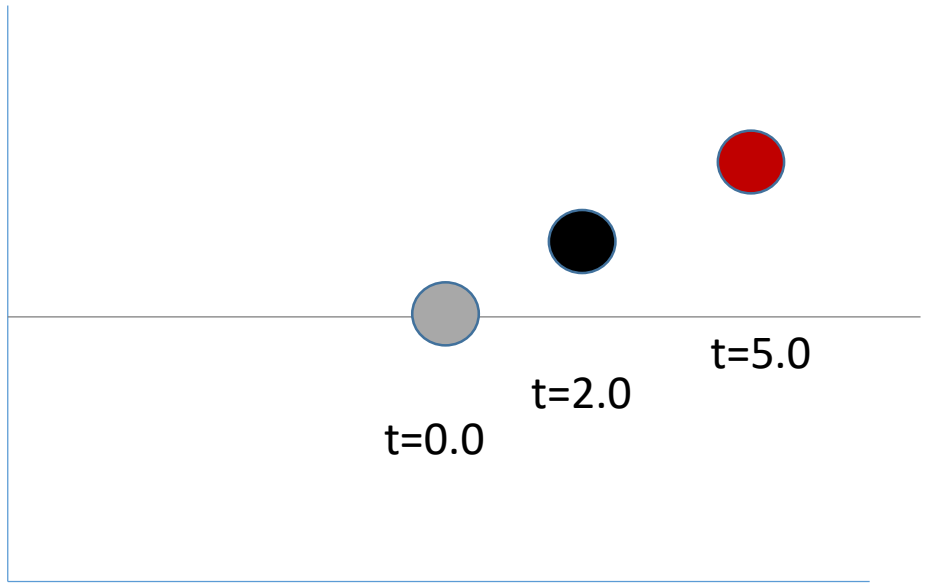
Hypothetical Experiment 2

- Likelihood ratios allow us to get around the limitations of NHST and answer the theoretically interesting question
- We first decide what constitutes a theoretically interesting effect (TIE)
- In this case, we might argue that a minimum of 5% increase in digit span is needed to be a TIE
- Then, we compare the fits of the Ho and the Htie to the observed data:
- $\lambda_{adjHo:Htie} = \frac{\lambda_{adj(obs:Htie)}}{\lambda_{adj(obs:Ho)}}$

Hypothetical Experiment 2

- Step 1: Calculate the $\lambda_{adj(obs:Htie)}$
- The t for the obs effect vs. the null of a 2% difference was 2.0
- A 5% difference (the TIE) would give a t score of 5.0
- The difference between the t-tie and t-obs is 5.0 - 2.0 = 3

Hypothetical Experiment 2



0% 2% 5%

Ho Hobs Htie

Hypothetical Experiment 2

- t-obs vs. t-tie = 3.0
- $\lambda_{adj(obs:Htie)} = 33.06$
- The data are about 33 times as likely given the observed effect size than given the TIE

Hypothetical Experiment 2

- Step 2: Calculate the $\lambda_{adj(obs:Ho)}$
- T –obs vs. t-tie = 2.0
- We already found this $\lambda_{adj(obs:Ho)} = 2.7$
- Step 3: Compute the $\lambda_{adj(Ho:Htie)}$
- $\lambda_{adj(Ho:Htie)} = 33.06/2.7 = 12.2$
- The data are 12.2 times as likely given no effect exists than given an effect large enough to be theoretically interesting

Hypothetical Experiment 2: testing for a theoretically interesting effect

- According to LR and model comparison approach, we should probably not be doing brain training

- This is in stark contrast to the policy decision implied by NHST, wherein a significant effect would also be assumed to be an important effect

Hypothetical Experiment 2: LR vs. NHST

- This experiment highlighted three problems inherent to NHST:

  1) Asymmetry

  2) With enough ss, any-sized effect can be statistically significant

  3) A statistically significant result implies it is also an important result

Hypothetical Experiment 2: LR vs. NHST

Likelihood ratios allowed us to solve each of these problems by posing the statistical question symmetrically

-allows us to find evidence either for or against the Ho

We did this by comparing the Ho to a Htie to determine whether the evidence supported an effect that was large enough to be theoretically interesting (and found good evidence for the Ho)

Summary: NHST and *p* vs. LR and model comparison

NHST is a flawed system to use for scientific inference

It seeks to determine which of two models is superior by testing the fit of only one model to the data!

It involves the use of a decision criterion ($p < 0.05$) that is both arbitrary and rather liberal

p-values are a non-intuitive index of the strength of the evidence, and can be particularly misleading if one believes that *p* = false positive rate

Summary: NHST and *p* vs. LR and model comparison

Some of these problems are currently being addressed by (e.g.,) arguing for a more stringent *p*-value criterion, and educating people about *p*-values, and/or insisting on more power in studies

Bayesian approaches are also being promoted– like likelihood ratios, these use model comparison rather than NHST

Likelihood ratios avoid the problems of NHST and *p*-values in the first place

Summary

The LR tells you the likelihood of the data occurring given one model relative to another

A likelihood ratio used to compare models is much more aligned with the false positive rate, has no arbitrary decision criterion attached to it, and allows one to find evidence both for and against the Ho

A likelihood ratio may not always be as easy to calculate as a *p*-value, but it is always easier to interpret

Thank you!