A Frequentist Travels to Bayesland: Field Notes on a Late Rumspringa

Georges Monette

York University

November 21, 2016

Roots of the controversy

Cicero (106 BCE - 43 BCE) gave a definition of *probabile*:

that which usually happens or
that which is commonly believed

- Still: Two main interpretations of probability

  relative frequency
  subjective belief

What's the difference between Bayesians and Frequentists? An oversimplification:

- 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed

What's the difference between Bayesians and Frequentists? An oversimplification:

- 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs

What's the difference between Bayesians and Frequentists? An oversimplification:

- 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs
- Both attempt to incorporate uncertainty – in orthogonal directions – but nonetheless often produce similar results because of symmetries in common statistical models and asymptotically

What's the difference between Bayesians and Frequentists? An oversimplification:

- 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs
- Both attempt to incorporate uncertainty – in orthogonal directions – but nonetheless often produce similar results because of symmetries in common statistical models and asymptotically

From ideology to utility

- Until recently the debate was mainly philosophical. F methods were much easier
- With improvements in MCMC, B methods have become more feasible and surpass F methods for many complex problems

What's the difference between Bayesians and Frequentists? An oversimplification:

- 'Proper subjective Bayesians' believe that the role of empirical evidence is to *transform* beliefs in a coherent way using Bayes theorem by conditioning on what has been observed
- Frequentists believe (hope?) that it is possible to find an objective basis for the interpretation of empirical evidence, one that does not require a reference to personal beliefs
- Both attempt to incorporate uncertainty – in orthogonal directions – but nonetheless often produce similar results because of symmetries in common statistical models and asymptotically

From ideology to utility

- Until recently the debate was mainly philosophical. F methods were much easier
- With improvements in MCMC, B methods have become more feasible and surpass F methods for many complex problems

R. A. Fisher's clever idea

The Lady Tasting Tea and the p-value: a frequentist basis for inference

In 1919, Dr. Muriel Bristol at Rothampsted Experimental Station claimed she could tell whether the milk was poured in first or the tea first.

- Imagine that she was offered 12 cups of tea in random order – 6 prepared milk first and 6 tea first
- She got 10 of the 12 right

What does this tell us about her ability to tell the difference?

Rationale behind the p-value

How can we quantify the evidence that she can tell the difference?

- Pretend that she can't tell the difference: 'null hypothesis' H_0
- The probability of getting 10 out 12 right is $p(y|H_0) = 0.038961$
- But the probability of any single outcome, even one consistent with H_0, might be very small and might say nothing against H_0
- Fisher's idea: use the *tail probability*, the probability of y as or more extreme than the observed value of y

p-value:

$$\begin{aligned} \Pr\left(y^+|H_0\right) &= p(y=10|H_0) + p(y=12|H_0) \\ &= 0.038961 + 0.001082 \\ &= 0.040043 \end{aligned}$$

Proof by contradiction/implausibility

Contradiction	**Implausibility**
A implies not B	A implies B is improbable
B true	B is observed
Therefore A is false	Therefore A is unlikely

Courtroom analogy: presumption of innocence

H_0: Innocence

Consider probability of data (evidence) | innocence

If evidence inconsistent with innocence, then reject innocence and find guilt

P-values on trial: Sir Roy Meadow and Sally Clark

Sally Clark

- Young lawyer, gives birth to first son in September 1996
- son dies, apparently of SIDS, at 10 weeks
- second son born a year later
- dies, apparently of SIDS, at 8 weeks
- only evidence of trauma consistent with resuscitation attempts
- charged with two counts of murder

Sir Roy Meadow

- distinguished pediatrician
- as expert witness testifies:
  - probability of one SIDS death: $\frac{1}{8{,}500}$
  - probability of two: $\left(\frac{1}{8{,}500}\right)^2 = \frac{1}{72{,}250{,}000}$
  - 'if she's innocent, the chances of this happening are 1 in 72 million'
- jury convicts Sally Clark of murder in November 1999
- first appeal lost in October 2000
- second appeal succeeds and Sally Clark is released in January 2003
- she dies in 2007 at the age of 42

Some criticisms of the evidence:

- assumption of independence not reasonable
- for a *p-value* we need probability of two *or more*
- $\frac{1}{8{,}500}$ too small
- *p-value* is probability of the data given innocence, not innocence given the evidence – 'Prosecutor's Fallacy'
  - ok . . . so what should we be using a *p-value* for? If not here, when?
- rarely mentioned: the calculation of the *p-value* ignores the selection mechanism, i.e. this is a post hoc test selecting this mother of two from all mothers of two
- should use a Bayesian approch . . . will see later

Some criticisms of the evidence:

- assumption of independence not reasonable
- for a *p-value* we need probability of two *or more*
- $\frac{1}{8{,}500}$ too small
- *p-value* is probability of the data given innocence, not innocence given the evidence – 'Prosecutor's Fallacy'
  - ok . . . so what should we be using a *p-value* for? If not here, when?
- rarely mentioned: the calculation of the *p-value* ignores the selection mechanism, i.e. this is a post hoc test selecting this mother of two from all mothers of two
- should use a Bayesian approch . . . will see later

Aftermath:

- Sir Roy Meadow lost his license
- but it was reinstated soon afterwards . . . maybe they realized that almost anyone could have made his mistake

Proof beyond a reasonable doubt?

Proof beyond a reasonable doubt?

A very small value of the probability of innocence ‘given’ the evidence?

Proof beyond a reasonable doubt?

A very small value of the probability of innocence 'given' the evidence?

$$\text{Probability}(\ \text{Innocence} \mid \text{Evidence}\ )?$$

Proof beyond a reasonable doubt?

A very small value of the probability of innocence 'given' the evidence?

$$\text{Probability}(\ \text{Innocence} \mid \text{Evidence}\ )?$$

What did Roy Meadow learn from stats?

How to calculate:

$$\text{Probability}(\ \text{Evidence}^{+} \mid \text{Innocence}\ )$$

the *p-value*, the probability of obtaining evidence *as or more contradictory* assuming innocence.

The fundamental neurosis of statistics

- We really want $p(\theta|y)$ but we'd have to accept $p(\theta)$
- So we give the world $p(y^+|\theta))$
  - Most people quietly think it's a proxy for $p(\theta|y)$
  - if not, what in the world could it be?
- Gigerenzer:
  - the confusion created by this unresolved conflict among statisticians, which is both suppressed and inherent in statistics textbooks, leads to a systemic neurosis in science for which the ritual of NHST is a form of conflict resolution – like compulsive hand washing – which makes it resistant to logical arguments
- One is most strongly committed to the beliefs one does not understand

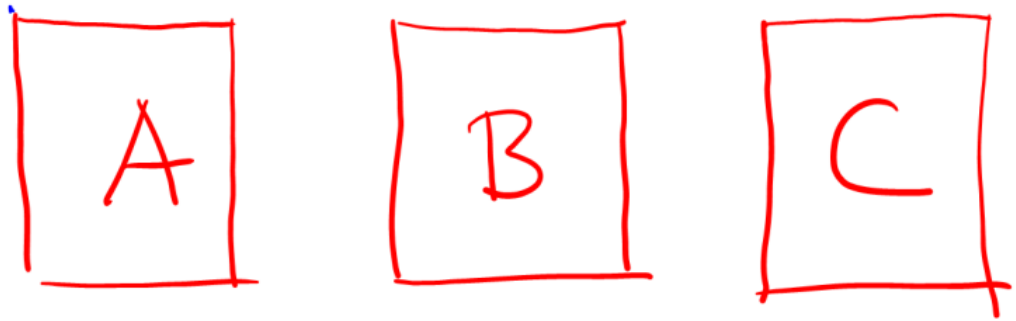
Monty Hall Example

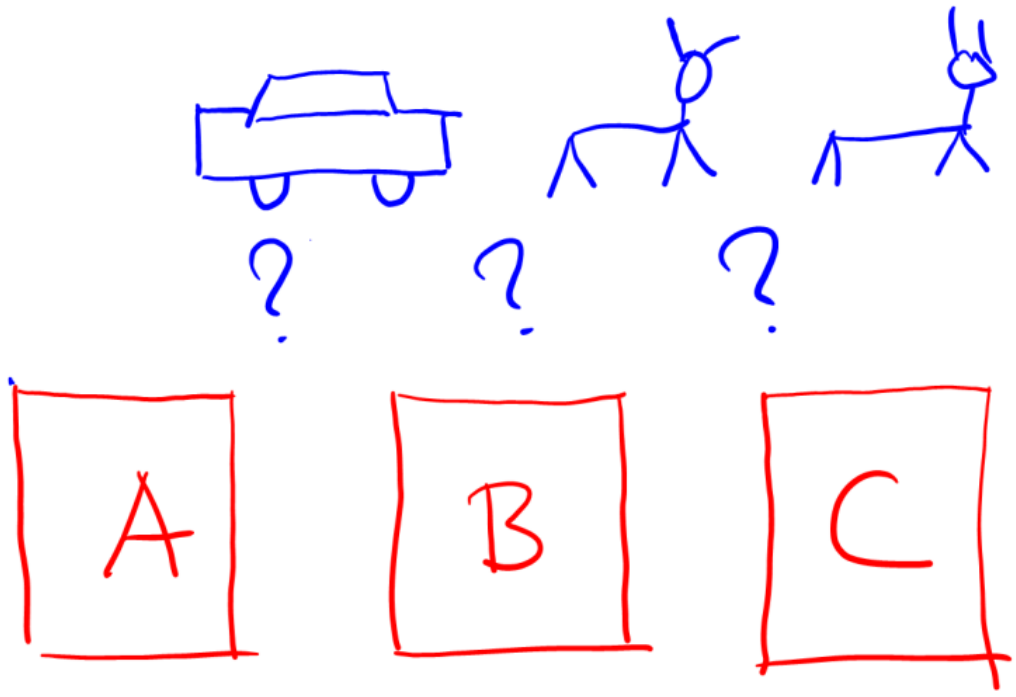
Simplest example showing
why we need **statistical** inference
– not just **logical** inference

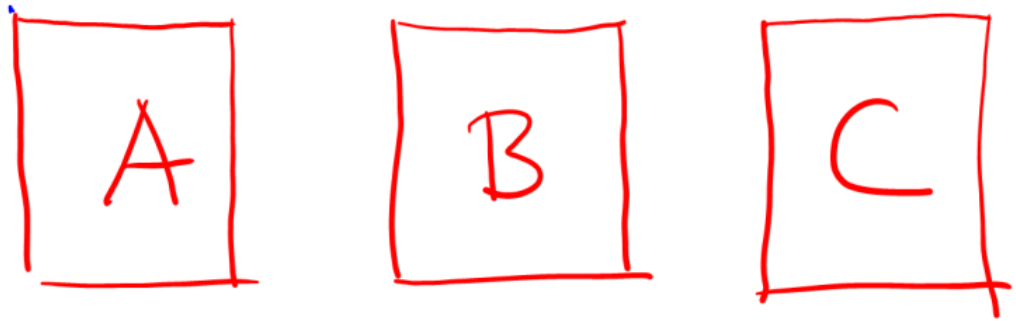
Monty Hall Example

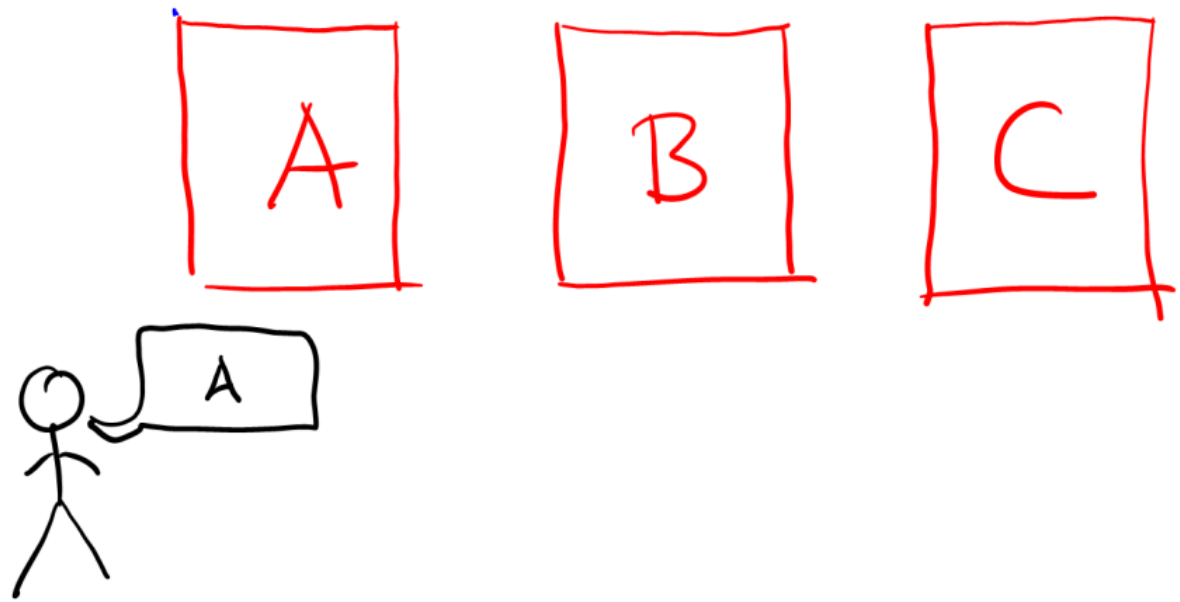
Simplest example showing
why we need **statistical** inference
– not just **logical** inference

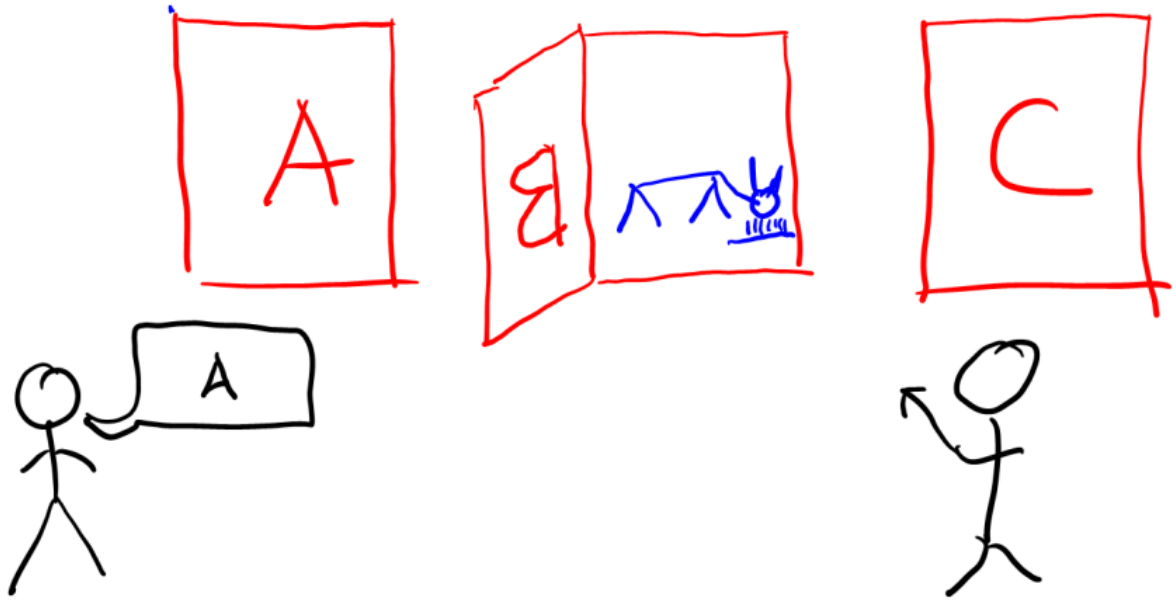
Key lesson:
If information (data) has a random component
we need a statistical model
to make correct inferences.

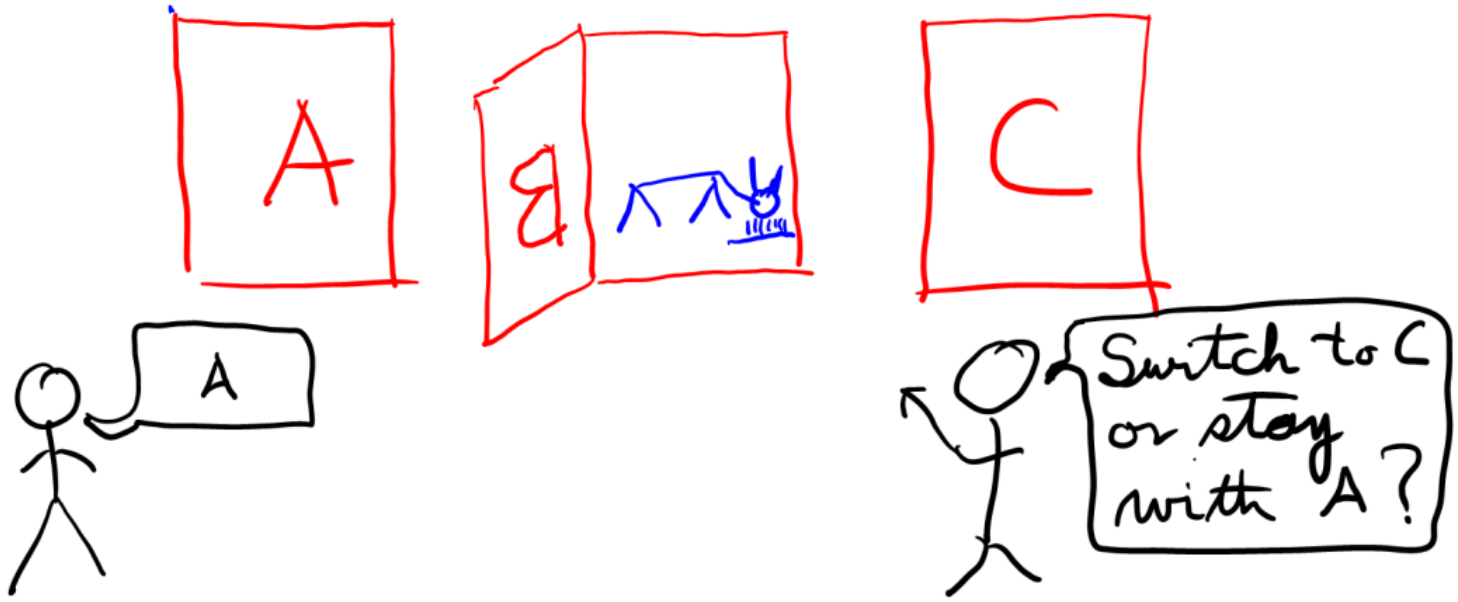

A
B
C


?
?
?
A
B
C


A
B
C


A
B
C
A


A
B
C
A


A
B
C
A
Switch to C or stay with A?

Step 1: Prior: $p(\theta)$

	Prize in	$p(\theta)$
θ	A	$1/3$
	B	$1/3$
	C	$1/3$
	sum	1

Step 1: Prior: $p(\theta)$

	Prize in	$p(\theta)$
	A	$1/3$
θ	B	$1/3$
	C	$1/3$
	sum	1

Step 2: Model: $p(y|\theta)$

	Opens B	Opens C	sum
A	$1/2$	$1/2$	1
B	0	1	1
C	1	0	1

Step 1: Prior: $p(\theta)$

	Prize in	$p(\theta)$
	A	1/3
θ	B	1/3
	C	1/3
	sum	1

Step 2: Model: $p(y|\theta)$

	Opens B	Opens C	sum
A	1/2	1/2	1
B	0	1	1
C	1	0	1

Step 3: Observed joint:
$p(\theta) \times p(Opens\ B|\theta)$

	Opens B
A	1/6
B	0
C	1/3
p(OpensB)	1/2

Step 1: Prior: $p(\theta)$

	Prize in	$p(\theta)$
	A	1/3
θ	B	1/3
	C	1/3
	sum	1

Step 2: Model: $p(y|\theta)$

	Opens B	Opens C	sum
A	1/2	1/2	1
B	0	1	1
C	1	0	1

Step 3: Observed joint:
$p(\theta) \times p(Opens\ B|\theta)$

	Opens B
A	1/6
B	0
C	1/3
p(OpensB)	1/2

Step 4: Posterior:
$p(\theta|y_{obs}) = p(y_{obs}, \theta)/p(y_{obs})$

Prizein	$p(\theta|$Opens B$)$
A	1/3
B	0
C	2/3
sum	1

What kind of analysis was this?

- Bayesian?
  - Uses conditional distribution for unknown given what's oberved
  - Does not use any part of model for information not observed
- Frequentist?
  - Uses only the relative frequency interpretation of probability

Lessons:

- The moment the model has anything other than 0s and 1s in it, you need statistical inference, i.e. if information is not deterministically deterrmined by the unknown, then we need statistical inference
- Nothing paradoxical or weird about Monty Hall problem. It's one of the simplest examples for inference

Bayesian workflow

Step 1: Prior: $p(\theta)$
Formulate a prior on **some basis**

θ_1	$p(\theta_1)$
θ_2	$p(\theta_2)$
θ_3	$p(\theta_3)$
sum	$1(\text{or } \infty!)$

Bayesian workflow

Step 1: Prior: $p(\theta)$

Formulate a prior on **some basis**

θ_1	$p(\theta_1)$
θ_2	$p(\theta_2)$
θ_3	$p(\theta_3)$
sum	1(or ∞!)

Step 2: Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p(y_1\|\theta_1)$	$p(y_2\|\theta_1)$	1
θ_2	$p(y_1\|\theta_2)$	$p(y_2\|\theta_2)$	1
θ_3	$p(y_1\|\theta_3)$	$p(y_1\|\theta_3)$	1

Bayesian workflow

Step 1: Prior: $p(\theta)$
Formulate a prior on **some basis**

θ_1	$p(\theta_1)$
θ_2	$p(\theta_2)$
θ_3	$p(\theta_3)$
sum	1(or ∞!)

Step 2: Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p(y_1\|\theta_1)$	$p(y_2\|\theta_1)$	1
θ_2	$p(y_1\|\theta_2)$	$p(y_2\|\theta_2)$	1
θ_3	$p(y_1\|\theta_3)$	$p(y_1\|\theta_3)$	1

Step 3: Observed joint:
Observe y_{obs}
$p(y_{obs}, \theta) = p(\theta) \times p(y_{obs}|\theta)$

θ_1	$p(y_{obs}, \theta_1)$
θ_2	$p(y_{obs}, \theta_2)$
θ_3	$p(y_{obs}, \theta_3)$
sum	$p(y_{obs})$ or c or ∞

Bayesian workflow

Step 1: Prior: $p(\theta)$
Formulate a prior on **some basis**

θ_1	$p(\theta_1)$
θ_2	$p(\theta_2)$
θ_3	$p(\theta_3)$
sum	1(or ∞!)

Step 2: Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p(y_1\|\theta_1)$	$p(y_2\|\theta_1)$	1
θ_2	$p(y_1\|\theta_2)$	$p(y_2\|\theta_2)$	1
θ_3	$p(y_1\|\theta_3)$	$p(y_1\|\theta_3)$	1

Step 3: Observed joint:
Observe y_{obs}
$p(y_{obs}, \theta) = p(\theta) \times p(y_{obs}|\theta)$

θ_1	$p(y_{obs}, \theta_1)$
θ_2	$p(y_{obs}, \theta_2)$
θ_3	$p(y_{obs}, \theta_3)$
sum	$p(y_{obs})$ or c or ∞

Step 4: Posterior:
$p(\theta|y_{obs}) = p(y_{obs}, \theta)/p(y_{obs})$

θ_1	$p(\theta_1\|y_{obs})$
θ_2	$p(\theta_2\|y_{obs})$
θ_3	$p(\theta_3\|y_{obs})$
sum	1 or ?

Frequentist workflow

Step 1: Formulate hypotheses; plan comparisons and estimates

Step 2:
Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p\left(y_1\middle\vert\theta_1\right)$	$p\left(y_2\middle\vert\theta_1\right)$	1
θ_2	$p\left(y_1\middle\vert\theta_2\right)$	$p\left(y_2\middle\vert\theta_2\right)$	1
θ_3	$p\left(y_1\middle\vert\theta_3\right)$	$p\left(y_2\middle\vert\theta_3\right)$	1

Frequentist workflow

Step 1: Formulate hypotheses; plan comparisons and estimates

Step 2:
Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p\left(y_1\middle\vert\theta_1\right)$	$p\left(y_2\middle\vert\theta_1\right)$	1
θ_2	$p\left(y_1\middle\vert\theta_2\right)$	$p\left(y_2\middle\vert\theta_2\right)$	1
θ_3	$p\left(y_1\middle\vert\theta_3\right)$	$p\left(y_2\middle\vert\theta_3\right)$	1

Step 3:
1. Observe y
2. Do something clever with $p(y|\theta)$
3. Estimate θ in a way that would work well on average – if you were repeat the process and get more y's

Frequentist workflow

Step 1: Formulate hypotheses; plan comparisons and estimates

Step 2:
Model: $p(y|\theta)$

	y_1	y_2	sum
θ_1	$p\left(y_1\middle\vert\theta_1\right)$	$p\left(y_2\middle\vert\theta_1\right)$	1
θ_2	$p\left(y_1\middle\vert\theta_2\right)$	$p\left(y_2\middle\vert\theta_2\right)$	1
θ_3	$p\left(y_1\middle\vert\theta_3\right)$	$p\left(y_2\middle\vert\theta_3\right)$	1

Step 3:
1. Observe y
2. Do something clever with $p(y|\theta)$
3. Estimate θ in a way that would work well on average – if you were repeat the process and get more y's

Step 4:
Insist how important it is that your results not be confused with $p(\theta|y)$ because that would require a subjective prior and science should be objective

Does it matter?

Powerful flexible techniques for complex data use MCMC. But do you need to be a Bayesian to use Bayesian methods?
Growing awareness of problems with frequentist approach but the Bayesian 'remedy' comes at a cost
Can't bake the Bayesian omelette (get a posterior) without breaking the Bayesian egg (prior)?

- Some believe that fiducial inference tries and structural inference succeeds

Where can we find priors?

- Orthodox Bayesians would say they have to be personal
- Others try to generate 'objective' standards
- The greatest benefits – i.e. not worrying about multiple comparisons or stopping time – only come at the greatest cost: true personal priors
- priors must precede the data, just like planned comparisons
- recipes for prior, e.g. uniform, Jeffreys, may sound good but they are largely arbitrary
- weakly informative priors require judgment but might be reasonable compromise

ASA statement on p-values

1. P-values can indicate how incompatible the data are with a specified statistical model.

ASA statement on p-values

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

ASA statement on p-values

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Fisher's caveat

In the 1950s Fisher warned that *p-values* should be used only if little else is known about a problem, thus excluding situations where there is other information about the plausibility of H_0

Efron's vision

Brad Efron has said that the 21st century might belong to new – and revived – approaches such a Fisher's fiducial infence

Does it matter? "Asymptotically they agree."

- The illusion of bigness:
  - Usually: as $N \uparrow$, p also $\uparrow$
  - Asymptotics require the number of independent observation $N \uparrow$, with p bounded
  - With big complex data, only a small number of top-level clusters might be independent
- The kind of large data set I see:
  - 28,000 records of data on 10+ variables:
    - 40 therapist/client interactions per session
    - 10 sessions per client
    - 70 clients
    - seeing 9 therapists
    - 5 using EFT and 4 CBT
- Can't count on asymptotics for intesting problems!
- The biggest promise of big data is being able to work on many small data problems:
  - small subpopulations, moderators with small cells
  - rare events, e.g. rare side effect

Does it matter? For many models they 'agree'

With continuous group transformation models, pivotal/fiducial/structural inference agree with Bayesian inference with right invariant Haar measure – a close cousin of Jeffreys prior.

An example where $y|\theta$ consistent with $\theta|y$

$$x = (x_1, ..., x_n) \tilde{\mathrm{N}} (\mu, 1)$$

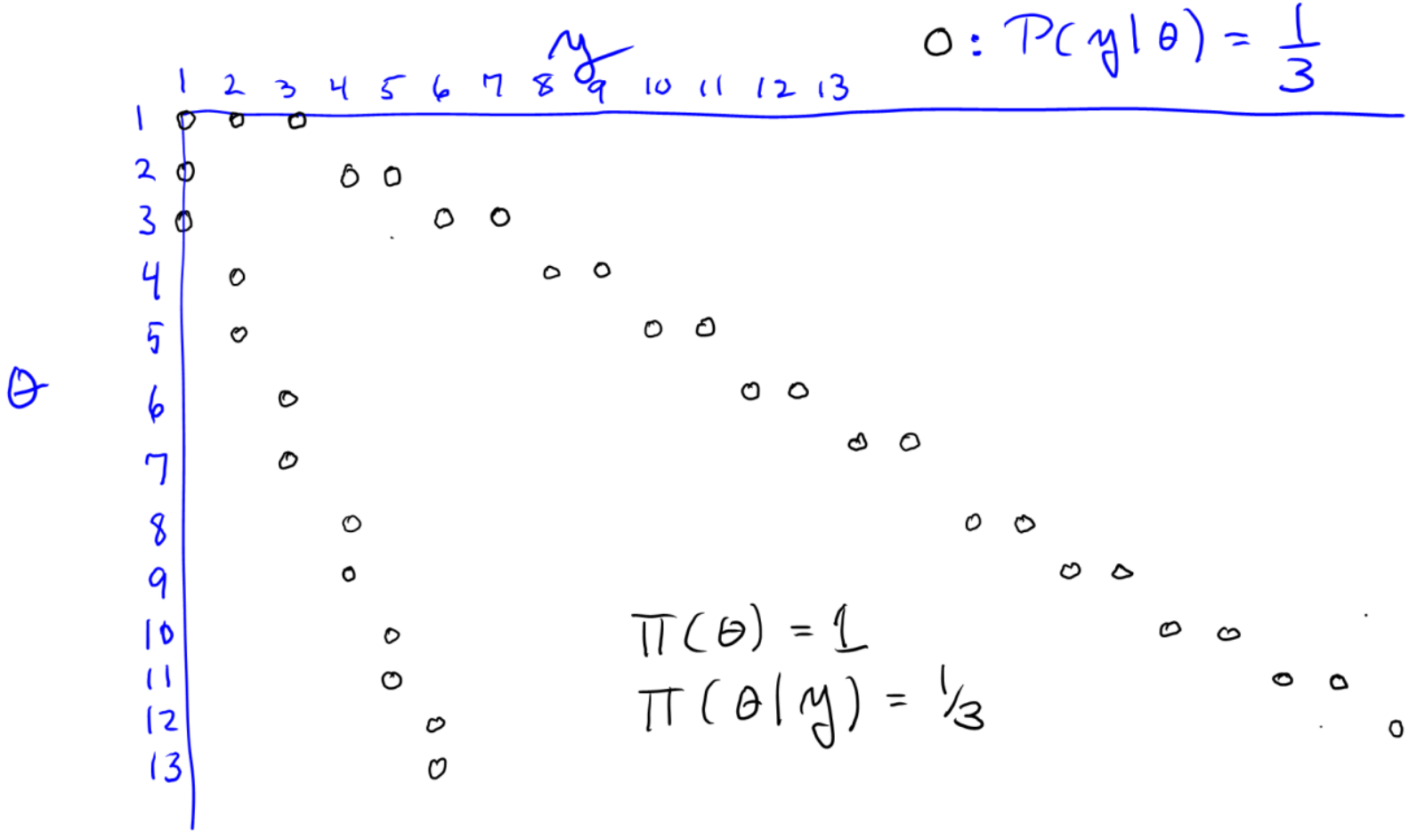
$$\text{prior: } \mu \text{ uniform on } \mathbb{R}$$

$$\text{posterior: } \theta \sim \mathrm{N}\left(\bar{x}, \frac{1}{\sqrt{n}}\right)$$

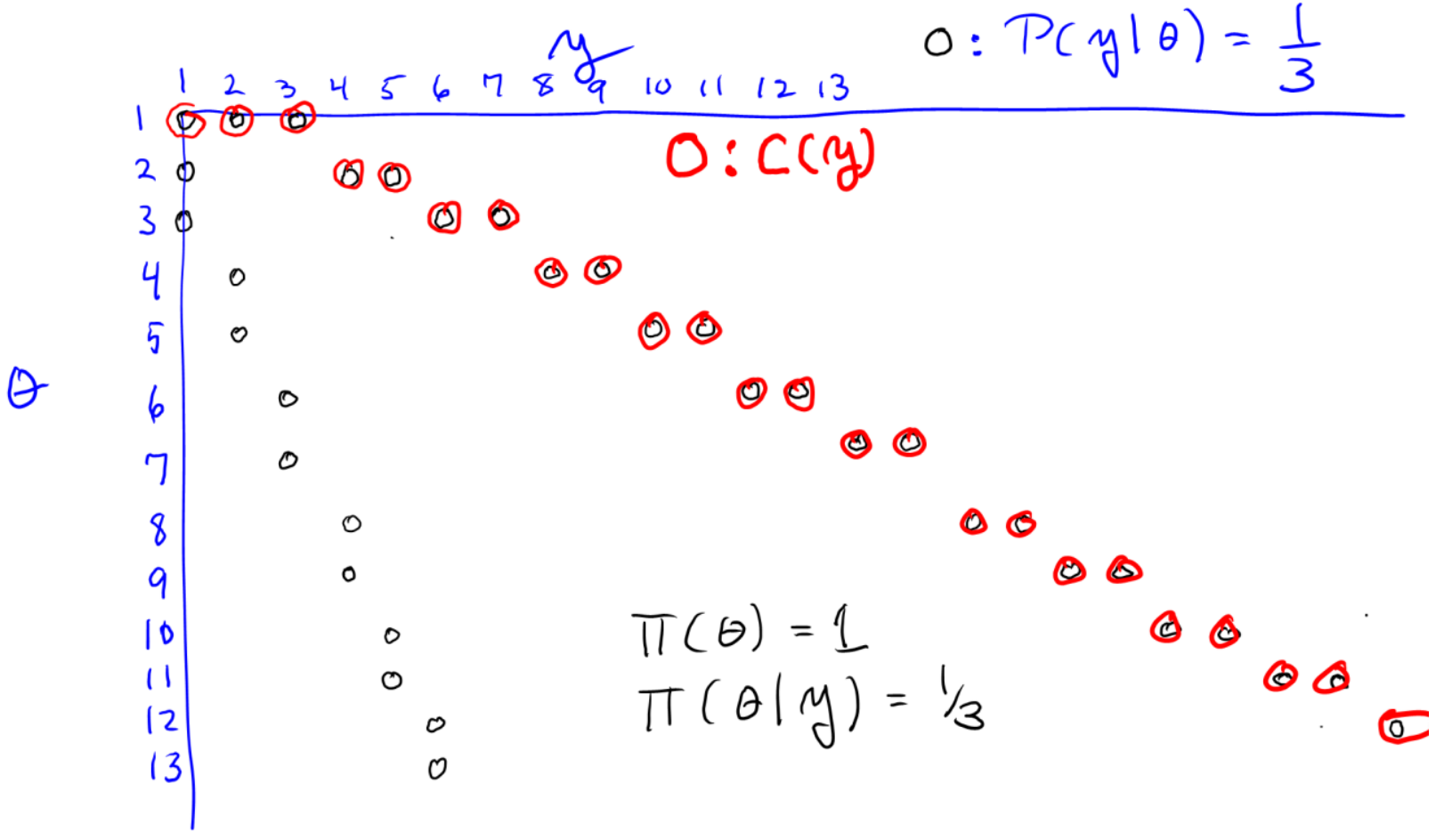
$$C(x) = \bar{x} \pm \frac{1.96}{\sqrt{n}}$$

$$\Pr_{x|\theta} \{\theta \in C(x)\} = \Pr_{\theta|x} \{\theta \in C(x)\}$$

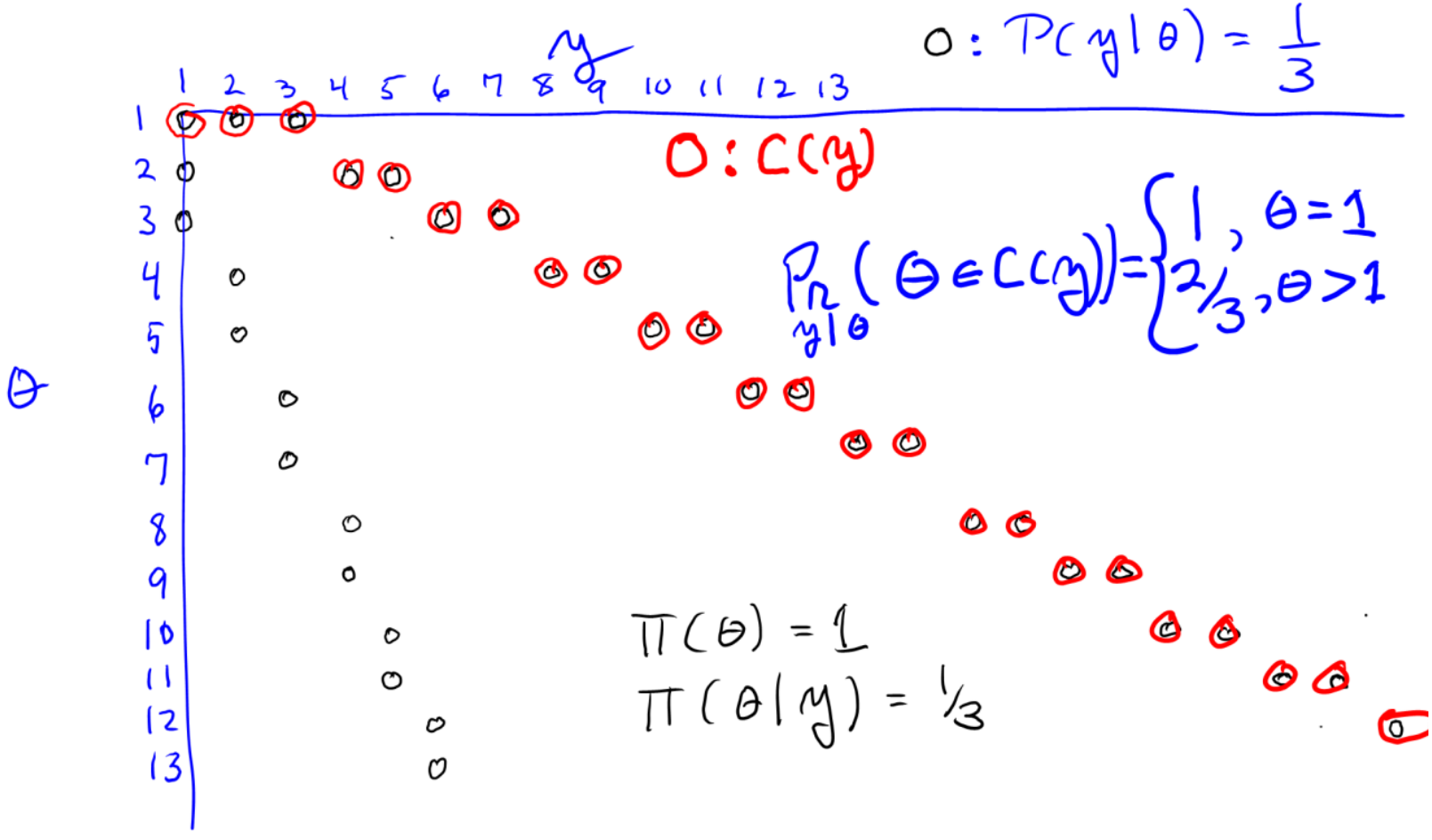
Can uniform priors go wrong: the Xmas tree example



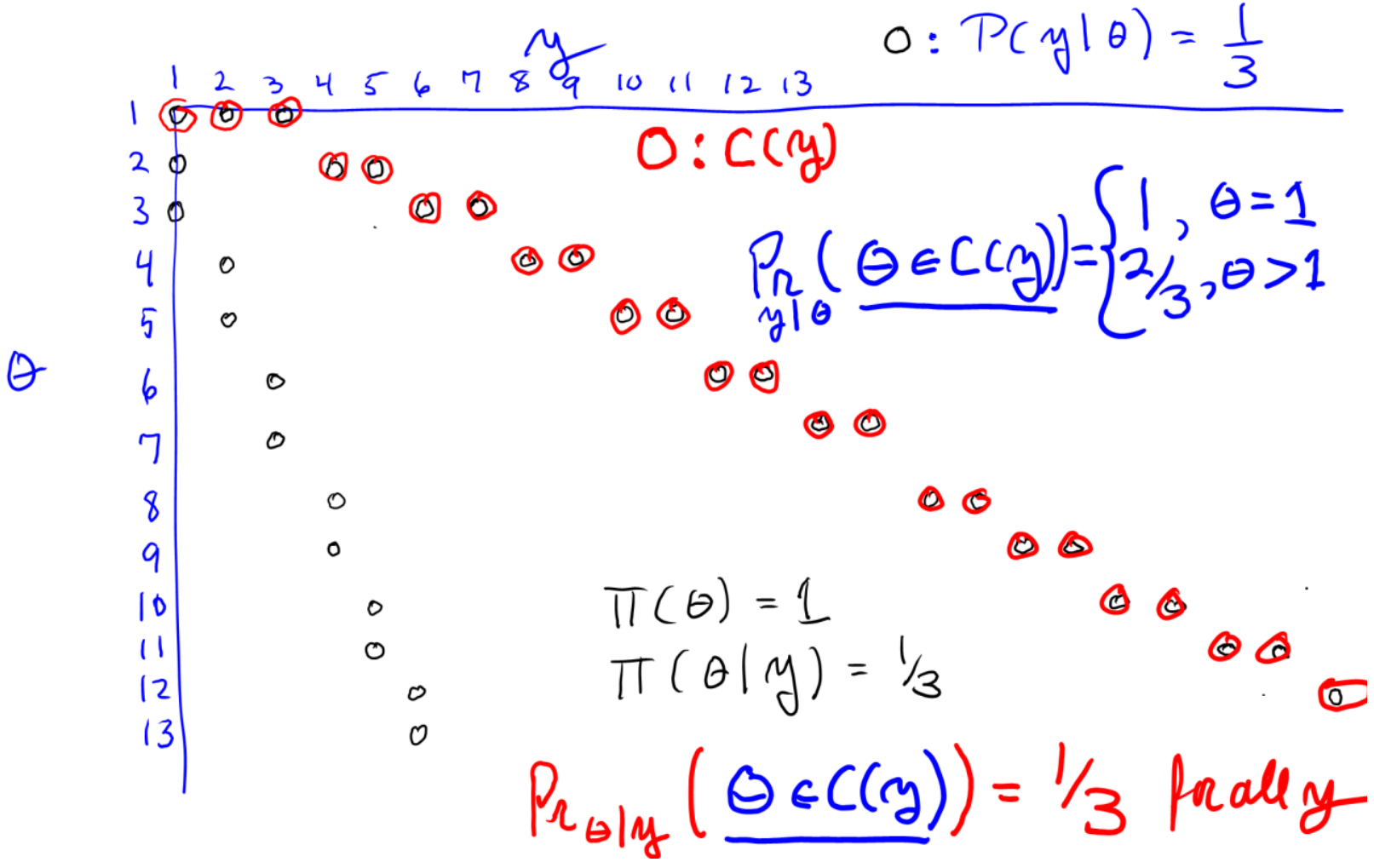
Can uniform priors go wrong: the Xmas tree example



But there's no guarantee: the Xmas tree example



Can uniform priors go wrong: the Xmas tree example



The source of the problem

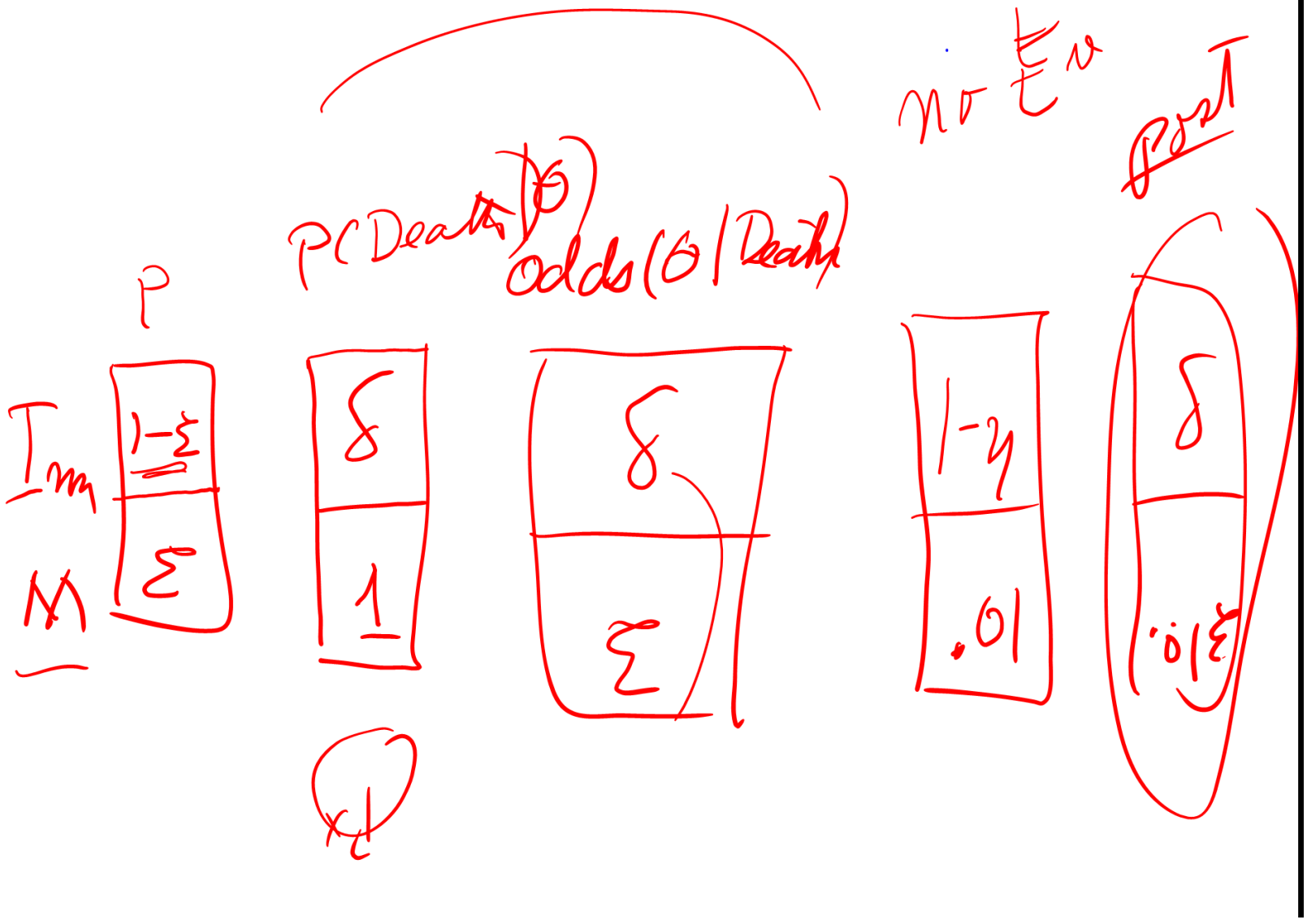
- Science deals with questions that too hard to answer by asking easy proxy questions for which answers are easier
- Hard Q: Inference under uncertainty
- Proxy Q: Find methods that have optimal properties under resampling
- Proxy Q: Use a prior that has some claim to objectivity
- We don't yet have a way to deal with hardest questions.
- It's good to remember that we are dealing with mere proxies. Stay alert to their possible shortcomings

A Bayesian approach to Sally Clark

All I've seen written on this considers the posterior probability of guilt given that two children are dead. Typical values depending on a variety of assumptions might range from 0.1 to 0.9.

However, I have not yet seen someone take into account the absence of forensic evidence of murder.

Most murders would, one presumes, leave incriminating evidence. If we take the Bayesian analysis a further step to incorporate the absence of forensic evidence, I believe that the posterior probability of guilt would be more reasonably in the range of 0.001 to 0.009 – barely a suspicion.


P
1-ξ
ξ
Im
M
P(Death)
odds(6|Death)
δ
1
δ
ξ
not ξ
1-η
.01
post
δ
.01η