How Statisticians Quantify Evidence

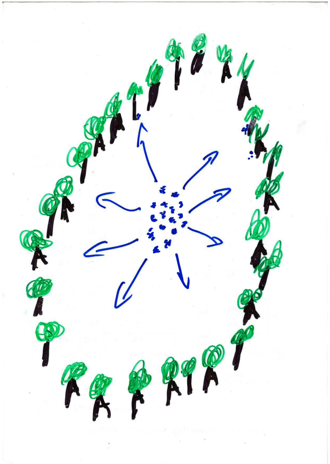
(and why they still don't agree about it)

Christian Hennig

Overview

1. Introduction: quantifying evidence
2. How significance tests work
3. Frequentist probability and p-values: what they are and what they are not
4. Bayesian probability and testing
5. Mathematical models and reality: why the statisticians don't and won't agree
6. Conclusion

1. Quantifying evidence



1. Quantifying evidence

Acacia species	not invaded	invaded
A	2	13
B	10	3

How strong is the evidence that the ants prefer species A?

1. Quantifying evidence

Acacia species	not invaded	invaded
A	2	13
B	10	3

How strong is the evidence that the ants prefer species A /one of the species?

For people who are not so interested in ants...
another example:

Coronary heart disease	CHD	non-CHD
Heavy coffee drinkers	38	752
Weak or no coffee drinkers	39	889

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?
- Does Potassium make your breakfast cereal taste better?

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?
- Does Potassium make your breakfast cereal taste better?
- Do the products of a company satisfy an industrial standard?

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?
- Does Potassium make your breakfast cereal taste better?
- Do the products of a company satisfy an industrial standard?
- Can it be a coincidence that many patients died when a particular nurse was at work?

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?
- Does Potassium make your breakfast cereal taste better?
- Do the products of a company satisfy an industrial standard?
- Can it be a coincidence that many patients died when a particular nurse was at work?
- Does homeopathy work against allergies?

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?
- Does Potassium make your breakfast cereal taste better?
- Do the products of a company satisfy an industrial standard?
- Can it be a coincidence that many patients died when a particular nurse was at work?
- Does homeopathy work against allergies?
- Is a new teaching method/therapy/fertilizer better than the old one?

Are some data evidence in favour or against a hypothesis?

- Does better street lightning reduce crime?
- Does Potassium make your breakfast cereal taste better?
- Do the products of a company satisfy an industrial standard?
- Can it be a coincidence that many patients died when a particular nurse was at work?
- Does homeopathy work against allergies?
- Is a new teaching method/therapy/fertilizer better than the old one?
- Is the spectrum of a celestial object compatible with a standard star type?

Subject matter expertise is necessary.

Possible influence of tree location?
Can it be considered independent what different ant tribes do?
How could it be explained that ants show preference for Acacia species A in this experiment
(apart from them generally preferring A)?

Subject matter expertise is necessary.

Possible influence of tree location?
Can it be considered independent what different ant tribes do?
How could it be explained that ants show preference for Acacia species A in this experiment
(apart from them generally preferring A)?
Is causal interpretation for link between coffee drinking and CHD justified (confounding, e.g., by stress on job)?

Subject matter expertise is necessary.

Possible influence of tree location?
Can it be considered independent what different ant tribes do?
How could it be explained that ants show preference for Acacia species A in this experiment
(apart from them generally preferring A)?
Is causal interpretation for link between coffee drinking and CHD justified (confounding, e.g., by stress on job)?

. . . we'd be happy about some unified logic. . .

Subject matter expertise is necessary.

Possible influence of tree location?
Can it be considered independent what different ant tribes do?
How could it be explained that ants show preference for Acacia species A in this experiment
(apart from them generally preferring A)?
Is causal interpretation for link between coffee drinking and CHD justified (confounding, e.g., by stress on job)?

. . . we'd be happy about some unified logic. . .
. . . though can all this be quantified?

2. How significance tests work

Basic idea: *Could it have happened by chance?*

2. How significance tests work

Basic idea: *Could it have happened by chance?*

Set up a probability model for “by chance”, and see whether data look “too unlikely”
(quantify how likely data would be “by chance”).

"By chance" may mean very different things:

- "Where the ants go is independent of the Acacia species."
- " The distribution of an allergy indicator is the same for homeopathy and placebo."
- "All nurses have the same probability to see patients dying."
- "All nurses' probabilities to see patients dying depend in the same way on how their work shifts are organised."

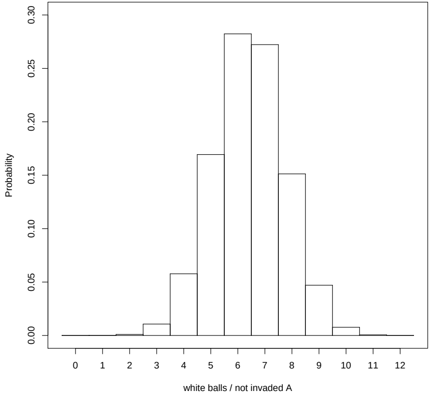
Acacia species	not invaded	invaded	sum
A	2	13	15
B	10	3	13
sum	12	16	28

Expected under independence:

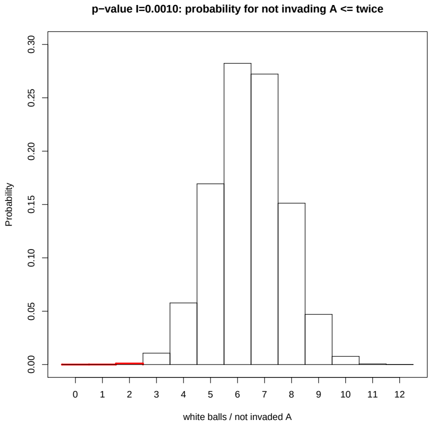
Acacia species	not invaded	invaded	sum
A	6.4	8.6	15
B	5.6	7.4	13
sum	12	16	28

Can $n_{11} = 2$ have occurred by chance?

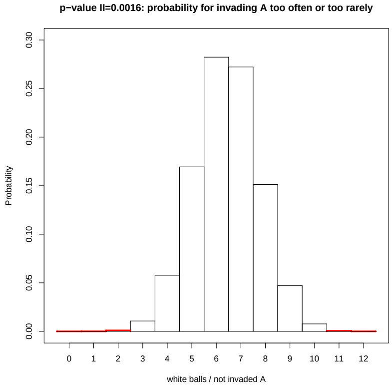
Draw 12 balls (Acacia trees)
from an urn with 15 black (A) and 13 white (B) balls.



Draw 12 balls (Acacia trees)
from an urn with 15 black (A) and 13 white (B) balls.



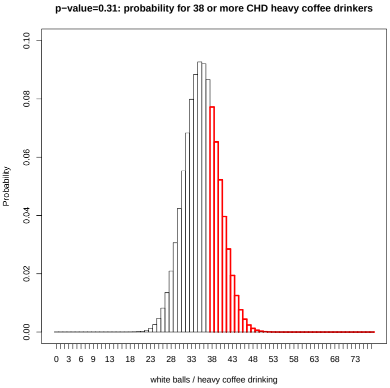
Draw 12 non invaded from 15 A and 13 B trees.



...these are very small probabilities

$\Rightarrow$ strong evidence that ants prefer Arcacia A.

Draw 77 CHD from 790 heavy coffee drinkers and 928 others.



. . . not very small probability, quite possible by chance
$\Rightarrow$ no strong evidence that coffee linked to CHD.

General principle of significance tests

Null hypothesis H_0: probability model for "chance/no effect/independence"

Alternative hypothesis H_1 (usually more than one)

Test statistic T: expected value distinguishes null and alternative

Distribution of T under H_0

p-value: probability under H_0 that T is as far or further away from what is expected under H_0 as observed value.

Small p-value: *evidence against* H_0.

Small p-value: *evidence against* H_0.

Large p-value (e.g., $>$ 0.1):
observations are compatible with H_0,
no evidence against it
(though it doesn't have to be true).

Small p-value: *evidence against H_0*.

Large p-value (e.g., > 0.1):
observations are compatible with H_0,
no evidence against it
(though it doesn't have to be true).

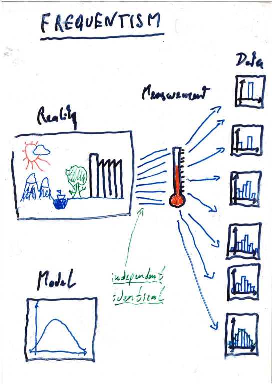
For example, dependence of CHD on coffee drinking
may be weak (at least to weak to be detected by these data),
but existent.
Or model assumptions (e.g., ant colonies independent of each other) could be violated.

3. Frequentist probability and p-values

p-value:
we imagine ant behaviour independent of Acacia species.
we imagine the experiment to be performed very, very often.

(One sided) p-value: relative frequency of having 2 or less A trees not invaded.

That's the *frequentist* interpretation of probability.


FREQUENTISM
Data
Measurement
Reality
Model
independent
identical

What can we say about the probability that ant behaviour is independent of Acacia species?

What can we say about the probability that ant behaviour is independent of Acacia species?
Nothing! They are either totally independent
or dependent (according to frequentist thinking).
There is no frequentist probability of H_0 being true.

What can we say about the probability that ant behaviour is independent of Acacia species?
Nothing! They are either totally independent
or dependent (according to frequentist thinking).
There is no frequentist probability of H_0 being true.

But isn't *this* probability what we really want to know?
Shouldn't an "evidence against independence" measure rather be a probability of/against independence?

What can we say about the probability that ant behaviour is independent of Acacia species?
Nothing! They are either totally independent or dependent (according to frequentist thinking).
There is no frequentist probability of H_0 being true.

But isn't *this* probability what we really want to know?
Shouldn't an "evidence against independence" measure rather be a probability of/against independence?
This is how p-values are often mis-interpreted.

Dependence of the p-value on unobserved events

The probability under independence of CHD and coffee
to draw *precisely*
38 CHD cases out of 790 strong coffee drinkers
is small (0.07).

Dependence of the p-value on unobserved events

The probability under independence of CHD and coffee
to draw *precisely*
38 CHD cases out of 790 strong coffee drinkers
is small (0.07).
But this alone isn't evidence against independence.

Dependence of the p-value on unobserved events

The probability under independence of CHD and coffee
to draw *precisely*
38 CHD cases out of 790 strong coffee drinkers
is small (0.07).
But this alone isn't evidence against independence.

Probability of 38 *or more* measures our evidence.
Or (depending on H_1) "38 or more or 32 or less".
But why should probability of 42 matter if we observe 38?

Dependence of the p-value on unobserved events

The probability under independence of CHD and coffee
to draw *precisely*
38 CHD cases out of 790 strong coffee drinkers
is small (0.07).
But this alone isn't evidence against independence.

Probability of 38 *or more* measures our evidence.
Or (depending on H_1) "38 or more or 32 or less".
But why should probability of 42 matter if we observe 38?
Because we need event likely under H_1 if not H_0
(unless we can compute probability of H_0 from observation).

The "garden of forking paths" (Gelman and Loken)

If you run many tests, you quite likely observe
one or few low p-values "by accident".
(Do something with 5% error probability many times,
then surely you'll make an error some time.)

The “garden of forking paths” (Gelman and Loken)

If you run many tests, you quite likely observe
one or few low p-values “by accident”.
(Do something with 5% error probability many times,
then surely you’ll make an error some time.)

People “fork” through data looking for things to test;
no wonder they find something.

Pre-defined experiment with pre-registered outcome is safer.

4. Bayesian probability and testing

The Bayesians *can* compute the probability that
ant behaviour is independent of Acacia species (0.005),
CHD is independent of coffee (0.959)

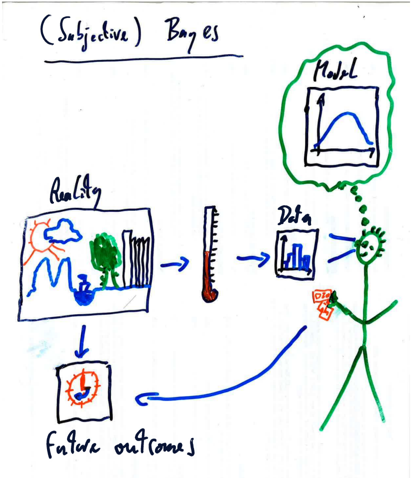
4. Bayesian probability and testing

The Bayesians *can* compute the probability that
ant behaviour is independent of Acacia species (0.005),
CHD is independent of coffee (0.959)
independently of unobserved events.

4. Bayesian probability and testing

The Bayesians *can* compute the probability that
ant behaviour is independent of Acacia species (0.005),
CHD is independent of coffee (0.959)
independently of unobserved events.

The catch is:
a different interpretation of probability is needed.


(Subjective) Bayes
Model
Reality
Data
Future outcomes

4. Bayesian probability and testing

The Bayesians *can* compute the probability that
ant behaviour is independent of Acacia species (0.005),
CHD is independent of coffee (0.959)

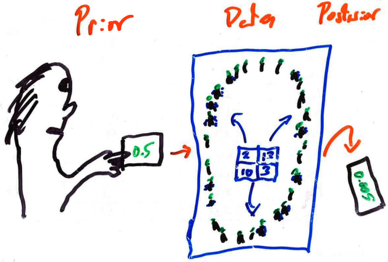
4. Bayesian probability and testing

The Bayesians *can* compute the probability that
ant behaviour is independent of Acacia species (0.005),
CHD is independent of coffee (0.959)
independently of unobserved events.

4. Bayesian probability and testing

The Bayesians *can* compute the probability that
ant behaviour is independent of Acacia species (0.005),
CHD is independent of coffee (0.959)
independently of unobserved events.

The catch is:
a different interpretation of probability is needed.
. . . and it depends on the prior distribution
(0.005 could be 0.09; 0.959 could be 0.54 or 0.03).


Prior
Data
Posterior
0.5
2
13
10
3
0.85

How to do it?

Bayes's theorem:

$$P(H_0|\text{data}) = \frac{P(\text{data}|H_0)P(H_0)}{P(\text{data}|H_0)P(H_0) + P(\text{data}|H_1)P(H_1)}$$

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Approach I: Choose $P(H_0),\ P(\text{parameters}|H_0/H_1)$ subjectively.

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Approach I: Choose $P(H_0),\ P(\text{parameters}|H_0/H_1)$ subjectively.

Distinguish *subjective* and *objective* Bayesians.

Subjectivists start with *personal* priors.
Objectivists:
there should be unique objective priors
(given some body of background knowledge),

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Approach I: Choose $P(H_0),\ P(\text{parameters}|H_0/H_1)$ subjectively.

Distinguish *subjective* and *objective* Bayesians.

Subjectivists start with *personal* priors.
Objectivists:
there should be unique objective priors
(given some body of background knowledge),
unique objective non-informative priors.

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Approach I: Choose $P(H_0),\ P(\text{parameters}|H_0/H_1)$ subjectively.

Distinguish *subjective* and *objective* Bayesians.

Subjectivists start with *personal* priors.
Objectivists:
there should be unique objective priors
(given some body of background knowledge),
unique objective non-informative priors.

Unfortunately they can't agree on these.

Needed: $P(\mathrm{data}|H_0), P(H_0), P(\mathrm{data}|H_1)$.

Approach II: Choose $P(H_0) = 0.5$, all possible table entry probabilities under H_1/H_0 uniformly.
Gives $P(H_0|\mathrm{data}) = 0.005$ (Acacia), 0.959 (CHD).

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Approach II: Choose $P(H_0) = 0.5$, all possible table entry probabilities under H_1/H_0 uniformly.
Gives $P(H_0|\text{data}) = 0.005$ (Acacia), 0.959 (CHD).

Try subjectively $P(H_0) = 0.05$ for CHD,
because we don't really believe
that coffee doesn't have any effect at all
$\Rightarrow P(H_0|\text{data}) = 0.528$.

Needed: $P(\text{data}|H_0), P(H_0), P(\text{data}|H_1)$.

Approach II: Choose $P(H_0) = 0.5$, all possible table entry probabilities under H_1/H_0 uniformly.
Gives $P(H_0|\text{data}) = 0.005$ (Acacia), 0.959 (CHD).

Try subjectively $P(H_0) = 0.05$ for CHD,
because we don't really believe
that coffee doesn't have any effect at all
$\Rightarrow P(H_0|\text{data}) = 0.528$.

$P(H_0) = 0.95$ gives $P(H_0|\text{data}) = 0.09$ (Acacia).

Approach III: actually we don't believe
that coffee doesn't have any effect at all,
and we could ask for the probability
that the effect is very weak (but possibly existing).

Approach III: actually we don't believe
that coffee doesn't have any effect at all,
and we could ask for the probability
that the effect is very weak (but possibly existing).

Choose $P(H_0) = 0$, uniform prior probabilities
for CHD rates in strong coffee drinkers (p)
and weak/no coffee drinkers (q).

Approach III: actually we don't believe
that coffee doesn't have any effect at all,
and we could ask for the probability
that the effect is very weak (but possibly existing).

Choose $P(H_0) = 0$, uniform prior probabilities
for CHD rates in strong coffee drinkers (p)
and weak/no coffee drinkers (q).
Consider *odds ratio* $r = \frac{p(1-p)}{q(1-q)}$. Close to1?

Approach III: actually we don't believe
that coffee doesn't have any effect at all,
and we could ask for the probability
that the effect is very weak (but possibly existing).

Choose $P(H_0) = 0$, uniform prior probabilities
for CHD rates in strong coffee drinkers (p)
and weak/no coffee drinkers (q).
Consider *odds ratio* $r = \frac{p(1-p)}{q(1-q)}$. Close to1?

$$P\left(0.99 < r < \frac{1}{0.99}\middle|\,\mathrm{data}\right) = 0.029,$$
$$P\left(0.8 < r < \frac{1}{0.8}\middle|\,\mathrm{data}\right) = 0.582.$$

Bayesians claim that often data dominate prior.
“Objective” approaches (from two standard books)
give quite different answers in coffee/CHD example.
Not so different in “ants and trees”.

Bayesians claim that often data dominate prior.
“Objective” approaches (from two standard books)
give quite different answers in coffee/CHD example.
Not so different in “ants and trees”.

Could do better subjectively:
CHD probability is expected to be small in both groups,
both probabilities are expected similar even under dependence.
This would likely give reasonable posteriors.

Bayesians claim that often data dominate prior.
“Objective” approaches (from two standard books)
give quite different answers in coffee/CHD example.
Not so different in “ants and trees”.

Could do better subjectively:
CHD probability is expected to be small in both groups,
both probabilities are expected similar even under dependence.
This would likely give reasonable posteriors.

Always need $P(H_0)$ first to get $P(H_0|\mathrm{data})$.
May agree on “group-intersubjective” $P(H_0)$.

Statistical approaches to quantify evidence - overview

Frequentist approaches (about 1880-1940)

Frequentist probability (Venn, von Mises, Kolmogorow)

Frequentist testing (Fisher, Neyman, Pearson)

Statistical approaches to quantify evidence - overview

Frequentist approaches (about 1880-1940)

- Frequentist probability (Venn, von Mises, Kolmogorow)
- Frequentist testing (Fisher, Neyman, Pearson)

Bayesian approaches (about 1890-1960)

- Objective Bayes (Keynes, Jeffreys, Carnap)
- Subjective Bayes (de Finetti, Ramsey, Savage, Lindley)
- "Modern" Bayes (Gelman, nowadays)

Likelihood (Edwards, Hacking, 1950-1970):
$\frac{P(\mathrm{data}|H_0)}{P(\mathrm{data}|H_1)} = \frac{1}{284}$ (ants), $\frac{1}{1.17}$ (CHD)

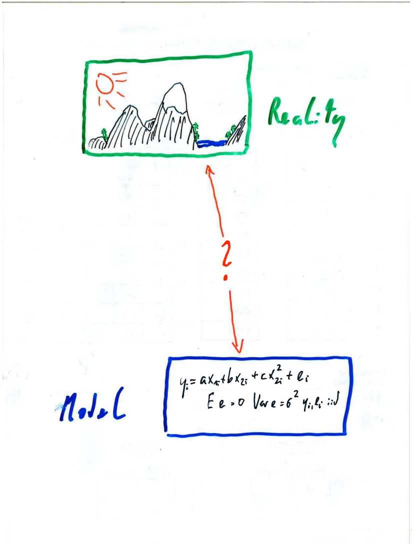
Likelihood (Edwards, Hacking, 1950-1970):
$\frac{P(\text{data}|H_0)}{P(\text{data}|H_1)} = \frac{1}{284}$ (ants), $\frac{1}{1.17}$ (CHD)
connected to both frequentist and Bayes,
but “how small is too small”?
Depends on richness of H_0/H_1.

Likelihood (Edwards, Hacking, 1950-1970):
$\frac{P(\text{data}|H_0)}{P(\text{data}|H_1)} = \frac{1}{284}$ (ants), $\frac{1}{1.17}$ (CHD)
connected to both frequentist and Bayes,
but "how small is too small"?
Depends on richness of H_0/H_1.

Interval probabilities (Fine, Walley, from 1950).

5. Mathematical models and reality

and why statisticians don’t and won’t agree.


Reality
?
Model
$y_i = a x_{1i} + b x_{2i} + c x_{2i}^2 + e_i$
$E e = 0$ $Var\, e = \sigma^2$ y_i, e_i iid

5. Mathematical models and reality

and why statisticians don't and won't agree.

It all involves idealisation:

- Ant colonies behave independently.
- People's coffee drinking is independent.
- There is identical repetition.

5. Mathematical models and reality

and why statisticians don't and won't agree.

It all involves idealisation:

- Ant colonies behave independently.
- People's coffee drinking is independent.
- There is identical repetition.
  More precisely, these are *mathematical* assumptions!

5. Mathematical models and reality

and why statisticians don't and won't agree.

It all involves idealisation:

- Ant colonies behave independently.
- People's coffee drinking is independent.
- There is identical repetition.
  More precisely, these are *mathematical* assumptions!
- The trees are only distinguished meaningfully by species.
- A single number can quantify strength of evidence.

Antony Gormley - Allotment

Idealisations involved in probability thinking

Frequentism: data generating mechanism doing infinite identical repetition.

Idealisations involved in probability thinking

Frequentism: data generating mechanism doing infinite identical repetition.
(Probability conceptualised as relative frequencies in repeated fair die throws.)

Idealisations involved in probability thinking

Frequentism: data generating mechanism doing infinite identical repetition.
(Probability conceptualised as relative frequencies in repeated fair die throws.)

Bayesian: betting analogy,
probability theory formalises rational betting,
i.e., experience enters via Bayes' formula only.

Idealisations involved in probability thinking

Frequentism: data generating mechanism doing infinite identical repetition.
(Probability conceptualised as relative frequencies in repeated fair die throws.)

Bayesian: betting analogy,
probability theory formalises rational betting,
i.e., experience enters via Bayes' formula only.
(Probability conceptualised as betting on fair dice.)

Idealisations involved in probability thinking

Frequentism: data generating mechanism doing infinite identical repetition.
(Probability conceptualised as relative frequencies in repeated fair die throws.)

Bayesian: betting analogy,
probability theory formalises rational betting,
i.e., experience enters via Bayes' formula only.
(Probability conceptualised as betting on fair dice.)

Historically, difference wasn't perceived (dice),
became apparent about 1840 (Poisson/Cournot),
long after Bernoulli (1713), Bayes (1763), Laplace (1814).

Probability is always about what could happen, other than just what happens!

It touches the essentially unobservable.

Probability is always about what could happen, other than just what happens!

It touches the essentially unobservable.

There is no objective way to model what cannot be seen.

Decisions:
model repetitive mechanisms in world outside
($\Rightarrow$ frequentist idealisation)
or model rational behaviour facing uncertainty
($\Rightarrow$ Bayesian idealisation).

Decisions:
model repetitive mechanisms in world outside
($\Rightarrow$ frequentist idealisation)
or model rational behaviour facing uncertainty
($\Rightarrow$ Bayesian idealisation).

Still need to decide model (& prior),
event/alternative,
borderline ("how small is too small?").

Decisions:
model repetitive mechanisms in world outside
($\Rightarrow$ frequentist idealisation)
or model rational behaviour facing uncertainty
($\Rightarrow$ Bayesian idealisation).

Still need to decide model (& prior),
event/alternative,
borderline ("how small is too small?").

Could try to negotiate agreement, but need subjectivity.

6. Conclusion

Some reasons for mathematical/formal thinking in spite of untestable/unobservable idealisations:

- Makes communication clearer.
- Supports agreement (by clarification, unification, decision rules).
- Supports imagination and creativity.

6. Conclusion

Some reasons for mathematical/formal thinking
in spite of untestable/unobservable idealisations:

- Makes communication clearer.
- Supports agreement
  (by clarification, unification, decision rules).
- Supports imagination and creativity.

However, don't forget

- it's not objective,
- something real (and potentially important)
  is ignored by idealisation (don't sweep under the carpet),
- it doesn't have to be done all the time.

Some practical consequences:

- Choice of approach:
  not “which one is correct”, but
  “which approach is implied by our decisions,
  and what does the approach imply?”
- “How do we *want* to think about our topic?”

Some practical consequences:

- Choice of approach:
  not "which one is correct", but
  "which approach is implied by our decisions,
  and what does the approach imply?"
- "How do we *want* to think about our topic?"
- Frequentist vs. Bayes:
  - Do we model "world outside" or "rational betting"?
  - Do we have background information nicely to be modelled as prior?
  - Do we want "$P(H_0$ given data)" to the price that we have to decide $P(H_0)$ in advance?
  - How can we get scientific agreement about prior, model, decision rule?

- Important issue about model assumptions
  is *not* whether they really hold,
  but whether we want to idealise this way,
  and whether we see reasons why this may be misleading.

- Important issue about model assumptions
  is *not* whether they really hold,
  but whether we want to idealise this way,
  and whether we see reasons why this may be misleading.
- Quantification of evidence: what is it needed for?
  Who should agree?
  Want to think of strength of evidence as single number?

- Important issue about model assumptions
  is *not* whether they really hold,
  but whether we want to idealise this way,
  and whether we see reasons why this may be misleading.
- Quantification of evidence: what is it needed for?
  Who should agree?
  Want to think of strength of evidence as single number?
- All this requires
  informed judgement about the background
  and consideration of the aim of study.

Paper related to this presentation
Hennig, C. (2009) A Constructivist View of the Statistical Quantification of Evidence. *Constructivist Foundations* 5(1): 39-54.
http://www.univie.ac.at/constructivism/journal/articles/5/1/039.hennig.pdf

This presentation:
http://www.homepages.ucl.ac.uk/ ucakche/presentations/evidencestat.pdf