What is Probability?
Bayes and Frequentist Approaches

Louis Lyons
Imperial College and Oxford

Valencia
June 2018
Probability in everyday language:

What is prob that it will rain in London tomorrow?

that signs of ETI will be discovered within 15 years?

of my winning >1M euros in lottery?

that O J Simpson murdered his ex-wife?

that I would meet the woman who is now my wife?
Probability conversation with my Barber

Barber: To improve my chance of winning the lottery, I’m going to choose my numbers at random like lottery does.

LL: I try to convince him this won’t work
I ask him whether he would choose the numbers “1, 2, 3, 4, 5 and 6”

Barber: Not a good idea. The chance of that combination is very low

LL: No lower than any other specific choice. And because people think that it is very unlikely, if you win, you are less likely to have to share the prize.

Actual fact: “1, 2, 3, 4, 5 and 6” is the most frequently chosen combination (because enough people know it is no less likely than any other), so you would probably have only a small share of the winnings if that was the winning set of numbers
Conclusion: Don’t choose “1, 2, 3, 4, 5 and 6”
Topics

• What is probability?
• Who cares?
• Bayesian approach
• Examples
• Frequentist approach
• Summary
WHAT IS PROBABILITY?

LEGAL PROBABILITY

MATHEMATICAL

Formal

Based on Axioms

FREQUENTIST

Ratio of frequencies as \( n \to \infty \)

Repeated “identical” trials

Not applicable to single event or physical constant

BAYESIAN Degree of belief

Can be applied to single event or physical constant

(even though these have unique truth)

Varies from person to person ***

Quantified by “fair bet”
BAYES and FREQUENTISM: The Return of an Old Controversy
It is possible to spend a lifetime analysing data without realising that there are two very different fundamental approaches to statistics: Bayesianism and Frequentism.
How can textbooks not even mention Bayes / Frequentism?

For simplest case \((m \pm \sigma) \leftarrow \text{Gaussian}\) with no constraint on \(\mu_{\text{true}}\), then

\[ m - k\sigma < \mu_{\text{true}} < m + k\sigma \]

at some probability, for both Bayes and Frequentist (but different interpretations)

See Bob Cousins “Why isn’t every physicist a Bayesian?” Amer Jrnl Phys 63(1995)398
We need to make a statement about **Parameters**, given **Data**

The basic difference between the two:

**Bayesian**: $\text{Prob}(\text{parameter, given data})$  
(an anathema to a Frequentist!)

**Frequentist**: $\text{Prob}(\text{data, given parameter})$  
(a likelihood function)
Bayesian versus Classical

Bayesian

\[ P(A \text{ and } B) = P(A;B) \times P(B) = P(B;A) \times P(A) \]

e.g. \( A = \text{event contains } t \text{ quark} \)

\[ B = \text{event contains } W \text{ boson} \]

or \( A = \text{I am in Valencia} \)

\[ B = \text{I am giving a lecture} \]

\[ P(A;B) = P(B;A) \times P(A) / P(B) \]

Completely uncontroversial, provided….
Bayesian

\[ P(A; B) = \frac{P(B; A) \times P(A)}{P(B)} \]

Problems: \( p(\text{param})\) Has particular value

“Degree of belief”

Prior What functional form?

Coverage
P(parameter)  Has specific value

“Degree of Belief”

Credible interval

Prior:  What functional form?

Uninformative prior:  flat?

In which variable?  e.g. m, m^2, ln m, ....?

Even more problematic with more params

  Unimportant  if “data overshadows prior”

  Important  for limits

Subjective or Objective prior?
Mass of Z boson (from LEP)

Data overshadows prior
Even more important for UPPER LIMITS
Mass-squared of neutrino

Prior = zero in unphysical region
Bayesian posterior $\rightarrow$ intervals

$p_{post}$

$\beta \rightarrow$

Upper limit

Lower limit

Central interval

Shortest
Ilya Narsky, FNAL CLW 2000

Upper Limits from Poisson data

Expect $b = 3.0$, observe $n$ events

Upper Limits important for excluding models
P (Data; Theory) \neq P (Theory; Data)
\[ P(\text{Data};\text{Theory}) \neq P(\text{Theory};\text{Data}) \]

Theory = male or female

Data = pregnant or not pregnant

\[ P(\text{pregnant} ; \text{female}) \sim 3\% \]
\[ P(\text{Data};\text{Theory}) \neq P(\text{Theory};\text{Data}) \]

**Theory** = male or female

**Data** = pregnant or not pregnant

\[ P(\text{pregnant};\text{female}) \sim 3\% \]

**but**

\[ P(\text{female};\text{pregnant}) >>> 3\% \]
P (Data;Theory) \neq P (Theory;Data)

HIGGS SEARCH at CERN

Is data consistent with Standard Model?
or with Standard Model + Higgs?

End of Sept 2000: Data not very consistent with S.M.
Prob (Data ; S.M.) < 1% \textit{valid frequentist statement}

Turned by the press into: Prob (S.M. ; Data) < 1%
and therefore Prob (Higgs ; Data) > 99%
i.e. “It is almost certain that the Higgs has been seen”
Example 1: Is coin fair?

Toss coin: 5 consecutive tails

What is \( P(\text{unbiased}; \text{data}) \) ? i.e. \( p = \frac{1}{2} \)

Depends on Prior(\( p \))

If village priest: \( \text{prior} \sim \delta(p = 1/2) \)

If stranger in pub: \( \text{prior} \sim 1 \text{ for } 0 < p < 1 \)

(also needs cost function)
Example 2: Particle Identification

Try to separate π’s and protons
(or: healthy people from those with disease)

probability (p tag; real p) = 0.95
probability (π tag; real p) = 0.05
probability (p tag; real π) = 0.10
probability (π tag; real π) = 0.90

Particle gives proton tag. What is it?
(or: Test is positive. Is person diseased?)

Depends on prior = fraction of protons (or: prevalence of disease)

If proton beam, very likely
If general secondary particles, more even (or: mostly healthy population)
If pure π beam, ~ 0
1) Dog $d$ has 50% probability of being 100 m. of Peasant $p$

2) Peasant $p$ has 50% probability of being within 100m of Dog $d$?
Given that:  

a) Dog $d$ has 50% probability of being 100 m. of Peasant,

is it true that:  

b) Peasant $p$ has 50% probability of being within 100m of Dog $d$ ?

**Additional information**

- Rivers at zero & 1 km. Peasant cannot cross them.  
  $0 \leq h \leq 1$ km
- Dog can swim across river - Statement a) still true

If dog at –101 m, Peasant cannot be within 100m of dog
Statement b) untrue
1) More specific on statement 1:

\[
\text{Prob}(d-h) = \begin{cases} 
\text{Const} & \text{for } |d-h| < 200 \text{m} \\
0 & \text{for } |d-h| > 200 \text{m} \end{cases} \quad [L'Hop]
\]

2) Hunter h uniform in 0 \rightarrow 1 \text{km} [\text{prior}]

![Graphs showing probability distributions for h and d-h with various conditions](image)
Classical Approach

Neyman “confidence interval” avoids pdf for $\mu$

Uses only $P( x; \mu )$

Confidence interval $\mu_1 \rightarrow \mu_2$

$P( \mu_1 \rightarrow \mu_2 \text{ contains } \mu_t ) = \alpha$ True for any $\mu_t$

Varying intervals fixed

from ensemble of experiments

Gives range of $\mu$ for which observed value $x_0$ was “likely” ($\alpha$)

Contrast Bayes: Degree of belief $= \alpha$ that $\mu_t$ is in $\mu_1 \rightarrow \mu_2$
Classical (Neyman) Confidence Intervals

Uses only $P(\text{data|theory})$

$\mu \geq 0$

No prior for $\mu$
90% Classical interval for Gaussian

\[ \sigma = 1 \quad \mu \geq 0 \]

e.g. \( m^2(\nu_e) \), length of small object

\( x_{\text{obs}} = 3 \) Two-sided range
\( x_{\text{obs}} = 1 \) Upper limit
\( x_{\text{obs}} = -1 \) No region for \( \mu \)

Other methods have different behaviour at negative \( x \)
Frequentism: Specific example

Particle decays exponentially: \[ \frac{dn}{dt} = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) \]

Observe 1 decay at time \( t_1 \):

\[ \mathcal{L}(\tau) = \frac{1}{\tau} \exp\left(-\frac{t_1}{\tau}\right) \]

Construct 68% central interval

68% conf. int. for \( \tau \) from \( t_1 / 1.8 \to t_1 / 0.17 \)
Bayes: Specific example

Particle decays exponentially: \[ \frac{dn}{dt} = \frac{1}{\tau} \exp(-t/\tau) \]

Observe 1 decay at time \( t_1 \):
\[ L(\tau) = \frac{1}{\tau} \exp(-t_1/\tau) \]

Choose prior \( \pi(\tau) \) for \( \tau \)
- e.g. constant up to some large \( \tau \)

Then posterior \( p(\tau) = L(\tau) \times \pi(\tau) \)
- has almost same shape as \( L(\tau) \)

Use \( p(\tau) \) to choose interval for \( \tau \) in usual way

Contrast frequentist method for same situation later.
\[ \mu_l \leq \mu \leq \mu_u \]  at 90% confidence

**Frequentist**

- \( \mu_l \) and \( \mu_u \) known, but random
- Unknown, but fixed

Probability statement about \( \mu_l \) and \( \mu_u \)

**Bayesian**

- \( \mu_l \) and \( \mu_u \) known, and fixed
- Unknown, and random

Probability/credible statement about \( \mu \)
Coverage

* What it is:
For given statistical method applied to many sets of data to extract confidence intervals for param $\mu$, coverage $C$ is fraction of ranges that contain true value of param. Can vary with $\mu$

* Does not apply to your data:
It is a property of the statistical method used
It is NOT a probability statement about whether $\mu_{\text{true}}$ lies in your confidence range for $\mu$

* Coverage plot for Poisson counting expt
Observe $n$ counts
Estimate $\mu_{\text{best}}$ from maximum of likelihood $\mathcal{L}(\mu) = e^{-\mu} \mu^n / n!$ and range of $\mu$ from $\ln\{\mathcal{L}(\mu_{\text{best}}) / \mathcal{L}(\mu)\} < 0.5$
For each $\mu_{\text{true}}$ calculate coverage $C(\mu_{\text{true}})$, and compare with nominal 68%
Coverage: $\mathcal{L}$ approach
(Not Neyman construction)

$$P(n, \mu) = e^{-\mu} \frac{\mu^n}{n!}$$  
(Joel Heinrich CDF note 6438)

$$-2 \ln \lambda < 1 \quad \lambda = \frac{P(n, \mu)}{P(n, \mu_{\text{best}})} \quad \text{UNDERCOVERS}$$
Frequentist central intervals, NEVER undercovers

(Conservative at both ends)
Feldman-Cousins Unified intervals

Neyman construction, so NEVER undercovers

Coverage (C) vs $\mu$: Unified Intervals  \( C \to 0.6827 \) as $\mu \to \infty$
Standard Frequentist

Pros:
Coverage
Widely applicable

Cons:
Hard to understand
Small or empty intervals
Difficult in many variables (e.g. systematics)
Needs ensemble
Bayesian

Pros:
Easy to understand
Physical interval

Cons:
Needs prior
Coverage not guaranteed
Hard to combine
## Bayesian versus Frequentism

<table>
<thead>
<tr>
<th></th>
<th>Bayesian</th>
<th>Frequentist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basis of method</td>
<td>Bayes Theorem → Posterior probability distribution</td>
<td>Uses pdf for data, for fixed parameters</td>
</tr>
<tr>
<td>Meaning of probability</td>
<td>Degree of belief</td>
<td>Frequentist definition</td>
</tr>
<tr>
<td>Prob of parameters?</td>
<td>Yes</td>
<td>Anathema</td>
</tr>
<tr>
<td>Needs prior?</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Choice of interval?</td>
<td>Yes</td>
<td>Yes (except F+C)</td>
</tr>
<tr>
<td>Data considered</td>
<td>Only data you have</td>
<td>….+ other possible data</td>
</tr>
<tr>
<td>Likelihood principle?</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
## Bayesian versus Frequentism

<table>
<thead>
<tr>
<th></th>
<th>Bayesian</th>
<th>Frequentist</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ensemble of experiment</strong></td>
<td>No</td>
<td>Yes (but often not explicit)</td>
</tr>
<tr>
<td><strong>Final statement</strong></td>
<td>Posterior probability distribution</td>
<td>Parameter values → Data is likely</td>
</tr>
<tr>
<td><strong>Unphysical/empty ranges</strong></td>
<td>Excluded by prior</td>
<td>Can occur</td>
</tr>
<tr>
<td><strong>Systematics</strong></td>
<td>Integrate over prior</td>
<td>Extend dimensionality of frequentist construction</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>Unimportant</td>
<td>Built-in</td>
</tr>
<tr>
<td><strong>Decision making</strong></td>
<td>Yes (uses cost function)</td>
<td>Not useful</td>
</tr>
</tbody>
</table>

**Notes:**
- Unphysical/empty ranges are excluded by the prior in Bayesian analysis but can occur in Frequentism.
- Systematics are integrated over the prior in Bayesian analysis, extending the dimensionality of the frequentist construction.
Bayesianism versus Frequentism

“Bayesians address the question everyone is interested in, by using assumptions no-one believes”

“Frequentists use impeccable logic to deal with an issue of no interest to anyone”
Approach used at LHC

Recommended to use both Frequentist and Bayesian approaches

If agree, that’s good

If disagree, see whether it is just because of different approaches
CONCLUSION

Hope you have an understanding of Bayesian and Frequentist approaches, and that if asked to explain the difference, probably you would give a good explanation