BAYES and FREQUENTISM: The Return of an Old Controversy

Louis Lyons and Lorenzo Moneta

Imperial College & Oxford University

CERN
Topics

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

Will discuss mainly in context of PARAMETER ESTIMATION. Also important for GOODNESS of FIT and HYPOTHESIS TESTING
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)
WHAT IS 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”

LEGAL PROBABILITY
Bayesian versus Classical

Bayesian

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

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

\[ B = \) event contains W boson \]

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

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

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

Completely uncontroversial, provided…. 
Bayesian

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

\[ p(param \mid data) \propto p(data \mid param) \times p(param) \]

Problems: \( p(param) \) Has particular value

“Degree of belief”

Credible Intervals

Prior

What functional form?

Coverage
Prior: What functional form?

Uninformative prior: Flat?

Cannot be normalised

Ranges 0-1 and 1089-1090 equally probable

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

\[ \frac{dp}{dm} = \frac{dp}{d(ln m)} \times \frac{d(ln m)}{dm} = \frac{1}{m} \times \frac{dp}{d(ln m)} \]

Even more problematic with more params

Unimportant if “data overshadows prior”

Important for limits

Subjective or Objective prior?

Priors might be OK for parametrising prior knowledge, but not so good for prior ignorance.
Mass of Z boson (from LEP)

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

Prior = zero in unphysical region

Fred James: “Is it a reindeer?”
Bayes: Specific example

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

Observe 1 decay at time \( t_1 \): \( \mathcal{L}(\tau) = \frac{1}{\tau} \exp\left(-\frac{t_1}{\tau}\right) \)

Choose prior \( \pi(\tau) \) for \( \tau \)

e.g. constant up to some large \( \tau \)

Then posterior \( p(\tau) = \mathcal{L}(\tau) \times \pi(\tau) \)

has almost same shape as \( \mathcal{L}(\tau) \)

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

Sensitivity study: Compare with using different prior

e.g. Prior constant in decay rate \( \lambda = 1/\tau \) \( \rightarrow \) different range

Contrast frequentist method for same situation later.
Bayesian posterior $\rightarrow$ intervals

- Upper limit
- Lower limit
- Central interval
- Shortest

UL $\rightarrow$ includes 0; LL $\rightarrow$ excludes 0; Central $\rightarrow$ usually excludes 0; Shortest is metric dependent
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% 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”
P (Data;Theory) ≠ P (Theory;Data)

Theory = Murderer or not

Data = Eats bread for breakfast or not

P (eats bread ; murderer) ~ 99%

but

P(murderer; eats bread) ~ 10^{-6}
\[ P (\text{Data};\text{Theory}) \neq P (\text{Theory};\text{Data}) \]

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

\[ P(\text{female};\text{pregnant}) \gg 3\% \]
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. $0 \leq h \leq 1 \text{ km}$
- Dog can swim across river - Statement a) still true

If dog at –101 m, Peasant cannot be within 100m of dog
Statement b) untrue
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 $P(\text{data}|\text{theory})$

Theoretical Parameter $\mu$

Observation $x \rightarrow$

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

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

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

Other methods have different behaviour at negative \( x \)

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

at 90% confidence

**Frequentist**

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

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

**Bayesian**

- \( \mu_l \) and \( \mu_u \) known, and fixed
- \( \mu \) unknown, and random

Probability/credible statement about \( \mu \)
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 \( t_1 / 1.8 \rightarrow t_1 / 0.17 \)
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
\[
L(\mu) = e^{-\mu} \frac{\mu^n}{n!}
\]
and range of \( \mu \) from
\[
\ln\left\{\frac{L(\mu_{\text{best}})}{L(\mu)}\right\} < 0.5
\]
For each \( \mu_{\text{true}} \) calculate coverage \( C(\mu_{\text{true}}) \), and compare with nominal 68%
Coverage: $\mathcal{L}$ approach
(Not Neyman construction)

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

$-2 \ln \lambda < 1 \quad \lambda = P(n, \mu) / P(n, \mu_{\text{best}})$  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 \( \mu \to \infty \)
Wants to avoid empty classical intervals

Uses “$\mathcal{L}$-ratio ordering principle” to resolve ambiguity about “which 90% region?”

[Neyman + Pearson say $\mathcal{L}$-ratio is best for hypothesis testing]

No ‘Flip-Flop’ problem
Feldman-Cousins 90% Confidence Interval for Gaussian

\[ X_{\text{obs}} = -2 \text{ now gives upper limit} \]
Flip-flop

Black lines  Classical 90% central interval

Red dashed: Classical 90% upper limit
Not good to let $x_{\text{obs}}$ determine how result will be presented. 

F-C: Move smoothly from 1-sided to 2-sided interval
Features of Feldman-Cousins

Almost no empty intervals
Unified 2-sided and 1-sided intervals
Eliminates flip-flop
No arbitrariness of interval
Less over-coverage than ‘x% at both ends’
‘Readily’ extends to several dimensions
  ‘x% at each end’ or ‘Max prob density’ problematic

Neyman construction time-consuming (esp in n-dimensions)
Minor pathologies: Occasional disjoint intervals
  Wrong behaviour wrt background
Tight limits when $b > n_{\text{obs}}$
  e.g. $n_{\text{obs}}$ bgd 90% UL
  0 3.0 1.08
  0 0.0 2.44

Exclusion of $s=0$ at lower $x$
Taking Systematics into account

Result for physics param s depends on systematic param ν
e.g. Mass of $H \rightarrow \gamma\gamma$ depends on energy scale for $\gamma$
Subsidiary measurement/info about $\nu$

1) Bayesian:
Post(s;n) = $\int Post(s,\nu;n) \, d\nu$ MARGINALISE
   where Post(s,ν;n) = $\mathcal{L}(n;s,\nu) \pi(s) \pi(\nu) / \int \mathcal{L}(n;s,\nu) \pi(s) \pi(\nu) \, ds \, d\nu$
   $\pi(\nu)$ from subsidiary expt. Maybe Gaussian $N(\nu_0, \sigma_\nu)$

2) $\mathcal{L}_{\text{prof}}(s) = \mathcal{L}(s,\nu_{\text{best}}(s))$ PROFILE
Then use $\mathcal{L}_{\text{prof}}(s)$ in likelihood, Frequentist, Bayesian approach

3) Frequentist:
Region in (s,ν) space for which measured values in main and subsid expts were likely.
   Problematic computationally

4) Mixed (Highland-Cousins):
Frequentist for main expt, but Bayesian smearing over $\nu$

Usually (many) more than one nuisance parameter
Reminder of PROFILE $\mathcal{L}$

Contours of $\ln\mathcal{L}(s,\nu)$

$s$ = physics param
$\nu$ = nuisance param

Stat uncertainty on $s$ from width of $\mathcal{L}$ fixed at $\nu_{\text{best}}$

Total uncertainty on $s$ from width of $\mathcal{L}(s,\nu_{\text{prof}(s)}) = \mathcal{L}_{\text{prof}}$
$
u_{\text{prof}(s)}$ is best value of $\nu$ at that $s$
$
u_{\text{prof}(s)}$ as fn of $s$ lies on green line

Total uncert $\geq$ stat uncertainty

Contrast with MARGINALISE
Integrate over $\nu$
$\mathcal{L}(s, \nu)$ for different fixed $\nu$
### Bayesian versus Frequentism

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

53
## 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 $\rightarrow$ 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>
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”
Recommended to use both Frequentist and Bayesian approaches for parameter determination

If agree, that’s good

If disagree, see whether it is just because of different approaches
Goodness of Fit: Kolmogorov-Smirnov

Compares data and model cumulative plots (or 2 sets of data)
Uses largest discrepancy between dists.
Model can be analytic or MC sample

Uses individual data points
Not so sensitive to deviations in tails
   (so variants of K-S exist)
Not readily extendible to more dimensions
Distribution-free conversion to p; depends on n
   (but not when free parameters involved – needs MC)