Confidence Intervals
Lecture 2

First ICFA Instrumentation School/Workshop
At Morelia, Mexico, November 18-29, 2002

Harrison B. Prosper
Florida State University
Recap of Lecture 1

- To interpret a confidence level as a relative frequency requires the concept of a set of ensembles of experiments.

- Each ensemble has a coverage probability, which is simply the fraction of experiments in the ensemble with intervals that contain the value of the parameter pertaining to that ensemble.

- The confidence level is the minimum coverage probability over the set of ensembles.
What’s Wrong With Ensembles?

- Nothing, if they are objectively real, such as:
  - The people in Morelia between the ages of 16 and 26
  - Daily temperature data in Morelia during the last decade

- But the ensembles used in data analysis typically are not
  - There was only a single instance of Run I of DØ and CDF. But to determine confidence levels with a frequency interpretation we must embed the two experiments into ensembles, that is, we must decide what constitutes a repetition of the experiments.

- The problem is that reasonable people sometimes disagree about the choice of ensembles, but, because the ensembles are not real, there is generally no simple way to resolve disagreements.
Outline

• Deductive Reasoning
• Inductive Reasoning
  – Probability
  – Bayes’ Theorem
• Example
• Summary
Consider the propositions: \( A = \text{(this is a baby)} \)
\[ B = \text{(she cries a lot)} \]

Major premise: If \( A \) is TRUE, then \( B \) is TRUE
Minor premise: \( A \) is TRUE
Conclusion: Therefore, \( B \) is TRUE

Major premise: If \( A \) is TRUE, then \( B \) is TRUE
Minor premise: \( B \) is FALSE
Conclusion: Therefore, \( A \) is FALSE

Aristotle, \( \sim 350 \) BC

\[ AB = A \]
Deductive Reasoning - ii

**A** = (this is a baby)
**B** = (she cries a lot)

Major premise: If **A** is TRUE, then **B** is TRUE
Minor premise: **A** is FALSE
Conclusion: Therefore, **B** is ?

Major premise: If **A** is TRUE, then **B** is TRUE
Minor premise: **B** is TRUE
Conclusion: Therefore, **A** is ?

AB = A
Inductive Reasoning

$A = \text{(this is a baby)}$

$B = \text{(she cries a lot)}$

Major premise: If $A$ is TRUE, then $B$ is TRUE

Minor premise: $B$ is TRUE

Conclusion: Therefore, $A$ is more plausible

Major premise: If $A$ is TRUE, then $B$ is TRUE

Minor premise: $A$ is FALSE

Conclusion: Therefore, $B$ is less plausible

Can “plausible” be made precise?
In 1946, the physicist Richard Cox showed that inductive reasoning follows rules that are isomorphic to those of probability theory.
The Rules of Inductive Reasoning: Boolean Algebra + Probability

Boolean Algebra

\[
\begin{align*}
A + 0 &= A \\
A + \overline{A} &= 1 \\
A + B &= B + A \\
A + B \cdot C &= (A + B) \cdot (A + C)
\end{align*}
\]

\[
\begin{align*}
A \cdot 1 &= A \\
A \cdot \overline{A} &= 0 \\
A \cdot B &= B \cdot A \\
A \cdot (B + C) &= A \cdot B + A \cdot C
\end{align*}
\]
Probability

Conditional Probability

\[ P(A \mid B) = \frac{P(AB)}{P(B)} \]

\[ P(B \mid A) = \frac{P(AB)}{P(A)} \]

A theorem

\[ P(A + B) = P(A) + P(B) - P(AB) \]
Probability - ii

**Product Rule**

\[ P(AB) = P(B \mid A)P(A) \]
\[ = P(A \mid B)P(B) \]

**Sum Rule**

\[ P(A \mid B) + P(\bar{A} \mid B) = 1 \]

**Bayes’ Theorem**

\[ P(B \mid A) = P(A \mid B)P(B) / P(A) \]

These rules together with Boolean algebra are the foundation of **Bayesian Probability Theory**
Bayes’ Theorem

\[ P(C_i D_j \mid A) = P(A \mid C_i D_j)P(C_i D_j) / P(A) \]

if \( C_i D_j \) are exhaustive propositions, i.e., \( \sum_{i,j} P(C_i D_j \mid A) = 1 \),

then we can write Bayes' Theorem as

\[ P(C_i D_j \mid A) = \frac{P(A \mid C_i D_j)P(C_i D_j)}{\sum_{i,j} P(A \mid C_i D_j)P(C_i D_j)} \]

We can sum over propositions that are of no interest

\[ P(C_i \mid A) = \sum_j P(C_i D_j \mid A) \]
Bayes’ Theorem: Example 1

- **Signal/Background Discrimination**
  - \( S = \text{Signal} \)
  - \( B = \text{Background} \)

\[
P(S \mid \text{Data}) = \frac{P(\text{Data} \mid S)P(S)}{P(\text{Data} \mid S)P(S) + P(\text{Data} \mid B)P(B)}
\]

- The probability \( P(S \mid \text{Data}) \), of an event being a signal given some event \( \text{Data} \), can be approximated in several ways, for example, with a feed-forward neural network.
Bayes’ Theorem: Continuous Propositions

\[ P(\theta, \lambda | x, I) = \frac{P(x | \theta, \lambda, I)P(\theta, \lambda | I)}{\int P(x | \theta, \lambda, I)P(\theta, \lambda | I) \, \theta, \lambda} \]

\[ I \text{ is the prior information} \]

\[ P(\theta | x, I) = \int P(\theta, \lambda | x, I) \, \lambda \]

\[ P(\theta, \lambda | x, I) \equiv f(\theta, \lambda | x) \, d\theta d\lambda \]
Bayes’ Theorem: Model Comparison

For each model $M$, we integrate over the parameters of the theory.

$$L(x \mid M, I) = \int P(x \mid \theta, M, I) P(\theta \mid M, I)$$

$$P(M \mid x, I) = \frac{L(x \mid M, I)P(M \mid I)}{\sum_{M} L(x \mid M, I)P(M \mid I)}$$

$P(M \mid x, I)$ is the probability of model $M$ given data $x$. 
Bayesian measures of uncertainty

\[ \sigma^2 = \int_{\theta}(\theta - \hat{\theta})^2 P(\theta | x, I) \theta \]
\[ \theta = u(x) \]

\[ \beta = \int_{\theta} P(\theta | x, I) \theta = l(x) \]

where

\[ \hat{\theta} = \int_{\theta} \theta P(\theta | x, I) \theta \]
Example 1: Counting Experiment

•实验:
  – 为了测量UHECRs的平均速率$\theta$，高于$10^{20}$ eV每单位立体角。

• 假设$N$事件的概率由泊松分布给定

$$P(n | \theta) = \frac{e^{-\theta} \theta^n}{n!}, \quad \theta = < n >, \quad \sigma = \sqrt{\theta}$$
Example 1: Counting Experiment - ii

- **Apply Bayes’ Theorem**

- **What should we take for the prior probability** $P(\theta | I)$?

- **How about**
  - $P(\theta | I) = d\theta$?

\[
P(\theta | n, I) = \frac{P(n | \theta, I)P(\theta | I)}{\int_\theta P(n | \theta, I)P(\theta | I)}
\]

\[
P(\theta | n, I) = \frac{e^{-\theta} \theta^n}{n!}
\]
But why choose $P(\theta|l) = d\theta$?

- Why not $P(\theta^2|l) = d\theta$?
- Or $P(\tan(\theta)|l) = d\theta$?

- Choosing a prior probability in the absence of useful prior information is a difficult and controversial problem.
  - The difficulty is to determine the variable in terms of which the prior probability density is constant.
  - In practice one considers a class of prior probabilities and uses it to assess the sensitivity of the results to the choice of prior.
Credible Intervals With $P(\theta|I) = d\theta$

Bayesian

$N \pm \sqrt{N}$

Bayesian Interval - Poisson Distribution

Count

Parameter

$\theta$
Credible Interval Widths

Bayesian

\[ N \pm \sqrt{N} \]
Coverage Probabilities: Credible Intervals

Even though these are Bayesian intervals nothing prevents us from studying their frequency behavior with respect to some ensemble!
Bayes’ Theorem: Measuring a Cross-section

Model
\[ \theta = \sigma l + b \]

Data
\[ n \]

Prior information
\[ (\hat{l}, \delta l), (\hat{b}, \delta b) \]

Likelihood
\[
P(n \mid \sigma, b, l, I) = \frac{e^{-\theta} \theta^n}{n!}
\]

\( l \) is the efficiency times branching fraction times integrated luminosity
\( b \) is the mean background count
\( \hat{l}, \hat{b} \) are estimates
Measuring a Cross-section - ii

Apply Bayes’ Theorem:

\[ P(\sigma, b, l | n, I) = \frac{P(n | \sigma, b, l, I) P(\sigma, b, l | I)}{\int \int \int P(n | \sigma, b, l, I) P(\sigma, b, l | I) \, \sigma, b, l} \]

Then marginalize: that is, integrate over nuisance parameters

\[ P(\sigma | n, I) = \int \int P(\sigma, b, l | n, I) \, b, l \]
What Prior?

\[ P(\sigma, l, b \mid I) = P(l \mid \sigma, b, I)P(b \mid \sigma, I)P(\sigma \mid I) \]

But usually, \( P(l \mid \sigma, b, I) = P(l \mid I) \) and \( P(b \mid \sigma, I) = P(b \mid I) \)

so we can write

\[ P(\sigma, l, b \mid I) = P(l \mid I)P(b \mid I)P(\sigma \mid I) \]

We can usually write down something sensible for the luminosity and background priors. But, again, what to write for the cross-section prior is controversial. In DØ and CDF, as a matter of convention, we set \( P(\sigma \mid I) = d \sigma \)

Rule: Always report the prior you have used!
Summary

• Confidence levels are probabilities; as such, their interpretation depends on the interpretation of probability adopted.

• If one adopts a frequency interpretation the confidence level is tied to the set of ensembles one uses and is a property of that set. Typically, however, these ensembles do not objectively exist.

• If one adopts a degree of belief interpretation the confidence level is a property of the interval calculated. A set of ensembles is not required for interpretation. However, the results necessarily depend on the choice of prior.