Why use a Bayesian approach?

• We often know how are measurements are related to our model...

• The Bayesian approach gives us the probability of our model once we have made a measurement

• It is useful for dealing with cases where there are errors (uncertainties) in the model specification (missing parts of model)
Introduction to Bayesian methods in macromolecular crystallography

Basics of the Bayesian approach

- Working with probability distributions
- Prior probability distributions
- How do we go from distributions to the value of “x”?
- Bayesian view of making measurements
- Example: from “400 counts” to a probability distribution for the rate
- Bayes' rule
- Applying Bayes' rule
- Visualizing Bayes' rule

Marginalization: Nuisance variables and models for errors

- How marginalization works
- Repeated measurements with systematic error

Applying the Bayesian approach to any measurement problem
Basics of the Bayesian approach
Working with probability distributions

Representing what we know about x as a probability distribution

$p(x)$ tells us the relative probability of different values of $x$

$p(x)$ does not tell us what $x$ is...
...just the relative probability of each value of $x$
I am sure $x$ is at least 2.5
Prior probability distributions
What we know before making measurements

All values of $x$ are equally probable
Prior probability distributions
What we know before making measurements

\[ p(x) \]

\[ x \text{ is less than about 2 or 3} \]
We don't know exactly what “x” is...

but we can calculate a weighted estimate:

\[ \langle x \rangle = A \int x \, p(x) \, dx \]

Weight each value of x by its relative probability \( p(x) \)

\[ A = 1 / \int p(x) \, dx \]

A is normalization factor
A Bayesian view of making measurements

A crystal is in diffracting position for a reflection
The beam and crystal are stable...

We measure 400 photons hitting the corresponding pixels in our detector in 1 second

What is the probability that the rate of photons hitting these pixels is actually less than 385 photons/sec?
A Bayesian view of making measurements

A crystal is in diffracting position for a reflection
The beam and crystal are stable...

We measure 400 photons hitting the corresponding pixels in our detector
in 1 second: $N_{\text{obs}} = 400$

A good guess for the actual rate $k$ of photons hitting these pixels is 400:
$k \sim 400$

What is the probability that $k$ is actually $< 385$ photons/sec?

What is $p( k<385 \mid N_{\text{obs}} = 400)$
A Bayesian view of making measurements

Start with prior knowledge about which values of $k$ are probable: $p_o(k)$

Make measurement $N_{\text{obs}}$

For each possible value of parameter $k$ (385...400...)

Calculate probability of observing $N_{\text{obs}}$ if $k$ were correct:

$p(N_{\text{obs}} \mid k)$

Use Bayes' rule to get $p(k)$ from $p_o(k)$, $N_{\text{obs}}$ and $p(N_{\text{obs}} \mid k)$:

$$p(k) \propto p_o(k) p(N_{\text{obs}} \mid k)$$
A Bayesian view of making measurements

What is the probability that we would measure $N_{\text{obs}}$ counts if the true rate were $k$?

$$p(N_{\text{obs}} | k)$$

$p(N_{\text{obs}} | k)$

$k=385$  
k=400
Bayes' rule

\[ p(k) \propto p_o(k) \ p(N_{obs} \mid k) \]

The probability that \( k \) is correct is proportional to...

the probability of \( k \) from our prior knowledge

multiplied by...

the probability that we would measure \( N_{obs} \) counts if the true rate were \( k \)
Bayes' rule

\[ p(k) \propto p_o(k) p \left( N_{obs} \mid k \right) \]

The probability that \( k \) is correct is proportional to...

the probability of \( k \) from our prior knowledge (prior)

multiplied by...

the probability that we would measure \( N_{obs} \) counts if the true rate were \( k \) (likelihood)
Application of Bayes' rule

\[ p(k) \propto p_o(k) p(N_{obs}|k) \]

No prior knowledge:

\[ p_o(k) = 1 \]

Poisson dist. for \( N_{obs} \) (large \( k \))

\[ p(N_{obs}|k) \propto e^{-\frac{[N_{obs} - k]^2}{2k}} \]
Application of Bayes' rule

Probability distribution for $k$ given our measurement $N_{obs} = 400$:

$$p(k) \propto e^{-\left[ N_{obs} - k \right]^2 / (2k)}$$

Probability that $k < 385$:

$$p(k < 385) = A \int_{-\infty}^{385} p(k) \, dk$$

$$A = 1 / \int_{-\infty}^{\infty} p(k) \, dk$$

$p = 22\%$
Visualizing Bayes' rule

\[ p(x \mid y_{obs}) \propto p_o(x) \, p(y_{obs} \mid x) \]

Where does Bayes' rule come from?

Using a graphical view to show how \( p(x \mid y) \) is related to \( p(y \mid x) \)
Visualizing Bayes' rule: $p(x) and p(y)$

$$p(x \mid y_{obs}) \propto p_o(x) \ p(y_{obs} \mid x)$$

$p(x) \ dx$ is the fraction of all drops from $x$ to $x+dx$

- $A$ is all the drops in the box
- $B$ is the drops in the vertical strip
- $C$ is drops in horizontal strip
- $D$ is the intersection of $B$ and $C$
Considering only drops from $x$ to $x+dx$, $p(y|x)dy$ is the fraction of drops from $y$ to $y+dy$.

Visualizing Bayes' rule: $p(y|x)$ and $p(x|y)$.
Visualizing Bayes' rule: \( p(x,y) \)

\[
p(x,y)dydx = p(y|x)dy \quad p(x)dx
\]

\[
\begin{bmatrix}
D
\end{bmatrix} = \begin{bmatrix}
D/B
\end{bmatrix} \begin{bmatrix}
B
\end{bmatrix}
\]

\( p(x,y)dydx = p(y|x)dy \quad p(x)dx \)

\( p(x,y)dydx \) is the fraction of all drops inside the box from \( x \) to \( x+dx \) and \( y \) to \( y+dy \)
Visualizing Bayes' rule: $p(x,y)$

$p(x,y)dxdy$ is the fraction of all drops inside the box from $x$ to $x+dx$ and $y$ to $y+dy$

$\int p(x)dx = B$

$\int p(x\mid y)dx = D/C$

$\int p(y)dy$

$\int p(x,y)dxdy = \int p(x\mid y)dx \cdot \int p(y)dy$
Visualizing Bayes' rule

\[ D = \frac{D}{B} \times B \]
\[ D = \frac{D}{C} \times C \]

\[ p(x)dx = B \]
\[ p(x)dx = B \]
\[ p(x, y)dxdy = p(y|x) \times p(x)dx dy \]
\[ p(x, y)dxdy = p(y|x) \times p(x)dy dx \]
\[ p(x, y)dxdy = p(x|y) \times p(y)dy dx \]
\[ p(y|x)dy = \frac{D}{B} \]
\[ p(y)dy = C \]
An identity we will need now and later....

\[ p(y) = \int p(y|x) p(x) \, dx \]
Visualizing Bayes' rule

$p(x,y)$ written two ways

\[ p(x \mid y) \, p(y) = p(y \mid x) \, p(x) \]

rearrangement...

\[ p(x \mid y) = p(y \mid x) \, p(x) / p(y) \]

An identity

\[ p(y) = \int p(y \mid x) \, p(x) \, dx \]

Substitution...Bayes' rule:

\[ p(x \mid y) = p(y \mid x) \, p(x) / \int p(y \mid x) \, p(x) \, dx \]
Bayes' rule as a systematic way to evaluate truth-tables

$p(x)\, dx$ is the fraction of all drops from $x$ to $x+dx$
Bayes' rule as a systematic way to evaluate truth-tables

We toss a coin twice and get at least one “heads”. What is the probability that the first toss was a “head?”

<table>
<thead>
<tr>
<th>First toss</th>
<th>Second toss</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>H H</td>
</tr>
<tr>
<td></td>
<td>H T</td>
</tr>
<tr>
<td>T</td>
<td>T H</td>
</tr>
<tr>
<td></td>
<td>T T</td>
</tr>
</tbody>
</table>
Bayes' rule as a systematic way to evaluate truth-tables

We toss a coin twice and get at least one “heads”. What is the probability that the first toss was a “head?”

FS=head on first or second toss
H= heads first toss       T= tails first toss

Bayes' rule:

\[ p(H) = A \cdot p(H) \cdot p(FS|H) \]
\[ A = 1 / [ p(H) \cdot p(FS|H) + p(T) \cdot p(FS|T) ] \]

\[ p_o(H) = 1/2 \]
\[ p(FS|H) = 1 \]
\[ p(FS|T) = 1/2 \]

\[ A = 1 / [1/2 + 1/2 \times 1/2] = 4/3 \]
\[ p(H) = 4/3 \times 1/2 = 2/3 \]
Quick Review of Bayes' rule

\[ p(x \mid y_{obs}) \propto p_o(x) \cdot p(y_{obs} \mid x) \]

- \( p(x \mid y_{obs}) \): Probability of \( x \) given our observations
- \( p_o(x) \): What we knew beforehand about \( x \)
- \( p(y_{obs} \mid x) \): Probability of measuring these observations if \( x \) were the correct value
Marginalization

What if the observations depend on $z$ as well as $x$? (Maybe $z$ is model error)

$$p(y_{obs} | x)$$

What we want to use in Bayes' rule

$$p(y_{obs} | x) = \int p(y_{obs} | x, z) \, p(z) \, dz$$

“Integrate over the nuisance variable $z$, weighting by $p(z)$”
Marginalization

\( y_{\text{obs}} = \text{observations} \)

\[
p(y_{\text{obs}}) = \int p(y_{\text{obs}} | z) p(z) \, dz
\]

Identity we saw earlier

\[
p(y_{\text{obs}} | x) = \int p(y_{\text{obs}} | z, x) p(z | x) \, dz
\]

The whole equation can be for a particular value of \( x \)

If \( z \) does not depend on \( x \),

\[
p(z) = p(z | x)
\]

"Integrate over the nuisance variable \( z \), weighting by \( p(z) \)"
Marginalization with Bayes' rule

We want to get $p(x)$ using $p(y_{obs}|x)$ in Bayes' rule...

$y_{obs}$ is an experimental measurement of $y$

$$p(y_{obs}|y) \propto e^{-(y_{obs} - y)^2/2 \sigma^2}$$

$y$ depends on $x$ and $z$ (perhaps $z$ is model error)

$$y = y(z, x)$$

...then we can integrate over $z$ to get $p(y_{obs}|x)$:

$$p(y_{obs}|x) = \int p(y_{obs}|y(z, x)) \, p(z) \, dz$$
Repeated measurements with systematic error

We want to know on average how many drops $D_{\text{avg}}$ of rain hit a surface per 100 cm² per minute.

The rain does not fall uniformly: $D(x) = D_{\text{avg}} + E(x)$ where the SD of $E(x)$ is $e$. However we only sample one place

We count the drops $N$ falling in 1 minute into a fixed bucket with top area of 100 cm² $m$ times ($N_1, N_2...$) with a mean of $n$.

What is the weighted mean estimate $<D_{\text{avg}}>$? What is the uncertainty in $<D_{\text{avg}}>$?
Repeated measurements with systematic error

We want to get $p(D_{avg})$ using $p(N_{obs}|D_{avg})$ in Bayes' rule...but the rate into our bucket $D$ depends on $D_{avg}$ and $E$:

$$D = D_{avg} + E$$

$$p(E) \propto e^{-E^2/2 \sigma^2}$$

$N_{obs}$ is the number of drops we count with SD of $n^{1/2}$:

$$p(N_{obs}|D_{avg}, E) \propto e^{-(N_{obs}-(D_{avg}+E))^2/2s^2}$$

Including all $m$ measurements $N_1, N_2...$

$$p(N_1, N_2...|D_{avg}, E) \propto e^{-\sum_i (N_i-(D_{avg}+E))^2/2s^2}$$
From previous slide

\[ p(N_1, N_2...|D_{avg}, E) \propto e^{-\sum_i(N_i-(D_{avg}+E))^2/2s^2} \]
\[ p(E) \propto e^{-E^2/2 e^2} \]

We have \( p(N_1, N_2...|D_{avg}, E) \). We want \( p(N_1, N_2...|D_{avg}) \). Integrate over the nuisance variable \( E \):

\[ p(N_1, N_2...|D_{avg}) = \int p(N_1, N_2...|D_{avg}, E) \ p(E) \ dE \]

Yielding (where \( n \) is the mean value of \( N \): \( <N_1,N_2...> \))

\[ p(N_1, N_2...|D_{avg}) \propto e^{-(D_{avg}-n)^2/2(e^2+s^2/m)} \]

Now we have \( p(N_1, N_2...|D_{avg}) \) and we are ready to apply Bayes' rule
We have the probability of the observations given $D_{\text{avg}}$, 

$$p(N_1, N_2 \ldots | D_{\text{avg}}) \propto e^{-\frac{(D_{\text{avg}}-n)^2}{2(e^2+s^2/m)}}$$

Bayes' rule gives us the probability of $D_{\text{avg}}$ given the observations:

$$p(D_{\text{avg}} | N_1, N_2 \ldots) \propto p_o(D_{\text{avg}}) e^{-\frac{(D_{\text{avg}}-n)^2}{2(e^2+s^2/m)}}$$

If the prior $p_o(D_{\text{avg}})$ is uniform:

$$p(D_{\text{avg}} | N_1, N_2 \ldots) \propto e^{-\frac{(D_{\text{avg}}-n)^2}{2(e^2+s^2/m)}}$$

$$\langle D_{\text{avg}} \rangle = n = \langle N \rangle \quad \sigma^2 = e^2 + s^2 / m$$
Summary: How to apply a Bayesian analysis to any measurement problem

1. Write down what you really want to know: \( p(D_{avg}) \)

2. Write down prior knowledge: \( p_o(D_{avg})=1 \)

3. Write down how the true value of the thing you are measuring depends on what you really want to know and any other variables: \( D=D_{avg}+E \)

4. Write down probability distributions for errors in measurement and for the variables you don't know: \( p(N_{obs}|D) \) and \( p(E) \)
How to apply a Bayesian analysis of any measurement problem

5. Use 3&4 to write probability distribution for measurements given values of what you want to know and of nuisance variables: \( p(N_1, N_2 \ldots | D_{\text{avg}}, E) \)

6. Integrate over the nuisance variables \( (E) \), weighted by their probability distributions \( p(E) \) to get probability of measurements given what you want to know:
\[
p(N_1, N_2 \ldots | D_{\text{avg}})
\]

7. Apply Bayes' rule to get the probability distribution for what you want to know, given the measurements:
\[
p(D_{\text{avg}} | N_1, N_2 \ldots) = p_o(D_{\text{avg}}) \ p(N_1, N_2 \ldots | D_{\text{avg}})
\]
Applications of the Bayesian approach in macromolecular crystallography

- Correlated MIR phasing (errors due to non-isomorphism are correlated among heavy-atom derivatives)
- Correlated MAD phasing (errors in heavy-atom model are correlated among wavelengths)
- Bayesian difference refinement (errors in model of macromolecular structure correlated between two structures)
- Macromolecular refinement (phase unknown and model errors present)
• Working through simple Bayesian exercises from handout in a group
• PHENIX demo and discussion
• Density modification and model-building theory and discussion
• Discussion of individual challenging examples and questions from students