Science involves the creation of hypothesis (or theories), and the testing of those theories by comparing their predictions with experimental observations.

In many cases, the conclusions of an experiment are obvious – the theory is supported or disproven.

In other cases, results are much more marginal. e.g. How big a sample size do we need to distinguish a successful drug from placebo effect?

Given the uncertainties of the data, we cannot usually determine whether a hypothesis is right or wrong – only how likely it is to be right: The probability of the hypothesis.

In order to do this, our hypothesis must be detailed enough for us to work out how likely we would have been to get the results we observe, assuming that the hypotheses is true.

We then use Bayes' theorem to determine the probability.

In order to correctly understand the impact of our experiments on our theories, we need some knowledge of statistics.

This is especially necessary in crystallography, since we have:

- a very weak signal: (the observed magnitudes)
- a great deal of noise: (the missing phases + measurement errors)
- from which we are trying to test a very detailed hypothesis: (the position of every atom in the unit cell)

Examples:

- Hypothesis: The observed X-ray data arises from a molecule which is roughly homologous to this known molecule in this orientation in the cell. (Molecular replacement – how probable is a given model)
- Hypothesis: The position of this atom (and its neighbors better explains the X-ray data when moved in this direction. (Refinement – what is the relative probability of two very similar models. Includes heavy-atom refinement.)

In fact all problems come down to the comparison of different hypotheses.
Understanding Likelihood

Example:

- We have a cat. She has a kitten. We don't know who the father is, but there are four possibilities in the neighborhood.
- What can we say about the color of the father from the color of the kitten?

Cat genetics is complex: we will simplify.

One black kitten: which is the father?

Actually, it could be any one, since they may all carry the appropriate genes. But they are not all equally probable.

We need some more information.

An extremely clever transformation has occurred:

\[ P(\text{kitten color} | \text{father color}) \]

\[ P(\text{father color} | \text{kitten color}) \]

\[ P(x | y): \text{The probability of } x, \text{ given } y. \]
This is a simple experiment:

- The result of the experiment is our observed data: the color of the kitten.
- The hypothesis is concerning the color of the cat. We make 4 hypotheses about the father (orange, black, cream, grey) and calculate the probability of each.
- We can work out \( P(\text{data} | \text{hypothesis}) \)
- We want to know \( P(\text{hypothesis} | \text{data}) \)

How do we do this sort of maths for a general problem?

- Use Bayes' theorem:

\[
P(x | y) = \frac{P(y | x) P(x)}{P(y)}
\]

Therefore:

\[
P(\text{hypothesis} | \text{data}) = \frac{P(\text{data} | \text{hypothesis}) P(\text{hypothesis})}{P(\text{data})}
\]

What if the population of male cats is non-uniform?

i.e. \( P(\text{father}_{\text{color}}) \) is non-uniform

Assumes \( x \) and \( y \) independent!
Understanding Likelihood

The elements of Bayes' theorem have names:

- **Likelihood:** The probability of getting the observed data, given the current hypothesis.
- **Prior:** The probability of the current hypothesis, before we even do an experiment.
- **Normalization:** So that all the possibilities add up to 1. Usually ignore.
- **Posterior:** Our final measure of confidence in the hypothesis having tested it.

As applied to the cats:

- **Likelihood:** If we suppose that a given cat is the father, we can work out the probable kitten colors.
- **Prior:** We know the distribution of male cats before the kitten is born.
- **Normalization:** Since we only want to compare relative probabilities, this doesn’t matter.
- **Posterior:** We can determine the probably father color from the results.

Likelihood and Crystallography

- How do we apply this to crystallography?
  - **Hypothesis:** The observed X-ray data arises from a molecule which is roughly homologous to this known molecule in this orientation in the cell. (Molecular replacement — how probable is a given model)
  - **Hypothesis:** The position of this atom (and its neighbors better explains the X-ray data when moved in this direction. (Refinement — what is the relative probability of two very similar models. Includes heavy-atom refinement.)
- Each hypothesis leads to a set of predicted structure factors: \( F_c(h) \). How well these explain the observed \( |F_{obs}(h)| \) determines the likelihood.

For most purposes, we treat each reflection as an independent observation. Therefore we can consult each reflection separately to determine how well it agrees with the model. Then, we multiply all the resulting likelihoods together.

- **Problem:** the product of 10,000s of small numbers gives an underflow on a computer.
- **Solution:** Take the log of all the likelihoods and sum them:
  \[
  \sum_i \log(x_i) = \log\left(\prod_i x_i\right)
  \]
  Usually minimize \(-\log(\text{likelihood})\) because it is +ve.

Likelihood and Crystallography

- Each hypothesis leads to a set of predicted structure factors: \( F_c(h) \). How well these explain the observed \( |F_{obs}(h)| \) determines the likelihood.
- But: we have a continuum of hypotheses. We can rotate an MR model or move a refinement atom continuously to improve the model.
- We refine the parameters of the model (e.g. rotation of MR model, position of refinement atoms) in order to best explain the observed data, i.e. to give the highest value of the likelihood, hence:
  \[\text{Maximum Likelihood}\]
Likelihood and Crystallography

- But actually we want to maximize the posterior, e.g. in refinement:
  - Prior gives the probability of the model on the basis of its agreement with stereo-chemical restraints.
  - Likelihood gives the probability of the model on the basis of the observed X-ray data.
- If we just maximize the likelihood, we get lousy geometry.
- But people call it 'maximum likelihood' anyway.

Likelihood and Crystallography

- Each hypothesis leads to a set of predicted structure factors: \( F_c(h) \). How well these explain the observed \( |E_{\text{obs}}(h)| \) determines the likelihood.
- Note: To calculate a probability we must also estimate the error associated with the \( F_c(h) \).
- The error estimation is a vital part of the model or hypothesis.

Likelihood and Crystallography

- How do we estimate the errors? Surely as the error estimate increases, the model always becomes a better description of the data?

Likelihood and Crystallography

- Error estimation is in terms of a parameter \( \sigma_A \), where \( \sigma_A \) is the fraction of the normalised structure factor \( E_c(h) \) which is correct, and \((1 - \sigma_A ^2)\) is the variance of the noise signal.
- Typically estimated as a function of resolution.
- Read (1986) Acta Cryst A42, 140-149

Likelihood and Crystallography

- We can calculate the Likelihood Function for \( E_{\text{obs}} \) given \( E_c \):

\[
P(E_{\text{obs}}|E_c) \propto \exp \left( \frac{|E_{\text{obs}} - \sigma_A E_c|}{\epsilon (1 - \sigma_A ^2)} \right)
\]
Likelihood and Crystallography

- But we don’t know $E_{\text{obs}}$!
  - The data are the observed magnitudes: $|E_{\text{obs}}|$
- We want $P(\text{data} | \text{model})$
- Therefore, sum (integrate) the likelihood over all the unknown phases: Rice fn (i.e. eliminate nuisance variable)

$$P\left(|E_{\text{obs}}| | E_c\right) \propto \exp \left( \frac{E_{\text{obs}}^2 + \sigma_a^2 E_c^2}{\epsilon (1 - \sigma_a^2)} \right) \int_0^L \frac{2|E_{\text{obs}}| \sigma_a |E_c|}{\epsilon (1 - \sigma_a^2)}$$

Likelihood and Crystallography

Steps:

- Construct a model, with some parameters: e.g.
  - MR: Rotation R, Translation T, Error term $\sigma_x$
  - Refinement: Coords $x_i$, Temp factors $B_i$, Error term $\sigma_A$
- Refine parameters $R,T$ / $x_i,B_i$, $\sigma_A$ to maximize the likelihood using the known magnitudes.
- Then use the resulting probability function for the phases to calculate an electron density map.
  - Programs will output ML map coefficients.

Likelihood and crystallography

Other details: Molecular replacement

- Programs will also use a likelihood function for unpositioned models to rank rotation function results.
- More complex likelihood functions allow combination of information from multiple fragments, even when relative position is unknown.

Likelihood and crystallography

Other details: Refinement

- Programs may also perform anisotropy correction, TLS refinement, bulk solvent correction. ML parameter refinement may be used to refine all of these parameters.
-Heavy atom refinement is similar, but is applied against multiple data sets simultaneously.

Likelihood and crystallography

Summary:

- Likelihood provides a tool for establishing the probability of a hypothesis.
- When data is weak, this is vital for describing our current state of knowledge.
- Direct benefits include improved models and weighted maps.
- Employed in:
  - phasing, MR, refinement, phase improvement, map interpretation.