Statistical Issues in Discovery

Richard Lockhart

Simon Fraser University

Banff Workshop, July 11 – 16, 2010
I see the following major issues

- In data analysis the null and alternative hypotheses specify both scientific assertions and assumptions about the experimental procedure.

- It is eminently possible that both the null and alternative statistical hypotheses are false even when that is not true of the physics hypotheses.

- Compelling evidence of discovery demands compelling modelling of systematics.

- You cannot expect to maximize a vector valued objective function.

- In drug trials a data analysis protocol is required; protocols make frequency theory analyses and calculations relevant and credible.
More issues

- It looks to me like physicists doing data analysis are just like statisticians doing data analysis – they tune things after seeing the data.
- Re-analysis of data is not generally convincing.
- When a $P$-value of $10^{-8}$ gets called back you have some obligation to understand the error!
- We, the statisticians, need to understand how much of the preprocessing we need to understand.
- No peak might be rejected if the background model is not right. We need to understand how badly we might exaggerate a small $P$-value by mild, not statistically significant, underfitting of the background.
- Model data as Poisson Process of events in time.
- At each event measure a response $X$ – the marks.
- Given times of events, marks are nearly independent and identically distributed (iid).
- Collapse data over time to get sample of $N$ values of $X_i$.
- Poisson process on the mark space; intensity $\lambda(x)$ (or $\lambda(x, t)$ if not collapsed over time).
Hypotheses

- Null hypothesis is
  *There is no such thing as a Higgs particle*

- Or perhaps “The Standard Model” including Higgs.

- Alternative hypothesis is some other model of physics.

- My own view (remark targetted at statisticians who disagree)
  *There is always an alternative hypothesis.*
Null hypothesis is $\lambda = \lambda_0$ recast as

$$N \sim \text{Poisson}(\Lambda_0 = \int \lambda_0(x)dx).$$

and

$$X \mid N \sim \text{iid } f = \frac{\lambda_0}{\Lambda_0}.$$

Alternative is $N$ has Poisson($\Lambda_0 + M$) distribution and given $N$ the $X_i$ are iid with some density $g$ given by

$$g = \frac{\Lambda_0}{\Lambda_0 + M} f + \frac{M}{\Lambda_0 + M} f^*$$

with $f^* \neq f$.

The density $f^*$ is the density of the marks in events which produce Higgs particles.
This is a mixture model problem.

The main issue is to distinguish $g$ from $f$ NOT to distinguish $\Lambda_0 + M$ from $\Lambda_0$; if $g = f$ then there is no effective way to make cuts and do triggering.

Lots is known about $f^*$; this should definitely be used in hypothesis testing.

I am conflicted about how much is known about $f$. In the pentaquark example $f$ restricted to area surviving the cuts is fitted just from the data.
On-off problem is prototypical.

\[ N \sim \text{Poisson}(a_{\text{off}} \lambda) \] \quad \text{and} \quad \[ M \sim \text{Poisson}(a_{\text{on}} \lambda + s) \]

- \( H_0 : s = 0. \)
- \( a_{\text{off}} \) and \( a_{\text{on}} \) are not known precisely.
- Uncertainties are not purely statistical – not data dominated.
- Similar problems in HEP.
- I want to be re-assured these systematics are indeed constant over the course of the measurements.
- If not Poisson model is in doubt – overdispersed model better?
For random effects which are really constant over all data I see no way out of integrating out the uncertainty.

So this is real Bayes.

The prior matters and **must** be informative so doubt concerning $P$-values will probably focus here.

Can statisticians help with prior selection?

One graph. $H_0 : N \text{ Pois}(\lambda = 100)$ with systematic standard error 10.
### Systematics P-value vs Poisson P-value

The graph depicts the relationship between Systematics P-value and Poisson P-value for different sample sizes. The nominal mean is 100, and the systematics standard error is 10. The graph shows the effect of sample size on the significance of the results.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Systematics P-value</th>
<th>Poisson P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 190</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 220</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Graph: Nominal mean 100, systematics SE 10

- The x-axis represents the Poisson P-value.
- The y-axis represents the Systematics P-value.
- Different lines indicate various sample sizes (n).
- The significance levels are indicated by different colors and line styles.
Exclusion defined

- Fix an interesting mass, \( m \).
- Test \( H_o(m) \): the particle does not exist at this mass.
- And test \( H^*_o(m) \): the particle does exist at this mass.
- First null is “exclusion”.
- Possible to test because specific mass implies lower limit on cross section.
- The two hypotheses are separate in sense of Cox (1961, 1962).
- It looks like one of the two hypotheses must be true.
- But this is not true about the statistical hypotheses; those hypotheses include assertions about the measuring process. They are hypotheses about Poisson rates.
- Also of great interest: \( H_o([m_L, m_H]) \): \( H_o(m) \) is true for each \( m_L \leq m \leq m_B \).
Multiple comparisons arise when you have several hypotheses which could be false – so that you could make several Type 1 errors.

But for $H_0(m)$ to be false the particle must exist at the given mass.

So at most one of these hypotheses can be false.

Louis argues that if both hypotheses are rejected there is a multiple comparisons problem.

The problem is that the physics dichotomy cannot be wrong but the statistical models, describing the behaviour of detectors, can both be wrong.

And both $P$-value calculations can be wrong. So I agree that a double rejection gives no scientific conclusion.
Gaussian peak on 3 parameter Gamma background

Invariant mass
Frequency

0 5 10 15 20 25 30 35

1.5 1.6 1.7 1.8 1.9
Pentaquarks

3 Parameter Gamma Background only

Frequency

Invariant Mass

0 5 10 15 20 25 30 35

1.5 1.6 1.7 1.8 1.9
Their analysis

- Fit model to $N_i$, $i$th cell count:
  \[ E(N_i) = \text{narrrrow Gaussian} + \text{broad Gaussian} + \text{constant} \]

- Count points under narrow peak ($\pm 2\sigma$)
- Split into background + peak = 54 + 43.
- Test statistic is $43/\sqrt{54} = 5.8$.
- $P$ value from Poisson is $8.9 \times 10^{-8}$
- $P$ value from Normal is $2.4 \times 10^{-9}$.
- I don’t approve.
Their Graph
Lessons to learn

- The conclusions are sensitive to the statistical model for the background.
- This is a hypothesis test for a missing component in a mixture. Large sample theory perilous.
- The method used makes no allowance for uncertainty in the fit. No allowance for estimation of location of peak.
- Test statistic is

\[
\frac{\text{Count in some range} - \text{area under background in range}}{\sqrt{\text{area under background in range}}}
\]

- I fitted 3 parameter gamma plus gaussian.
- Got \(2\Delta \log \ell \approx 12.3\) with 3 fewer parameters.
- Invalid approximate \(P\)-value about 0.006.
Bayes Factors

- $X \sim N(0, 1)$ vs $X \sim N(\mu, 1)$.
- $N(0, \sigma^2)$ prior on $\mu$.
- Log Bayes Factor is

$$\frac{x^2\sigma^2}{2(1 + \sigma^2)} - \frac{\log(1 + \sigma^2)}{2}.$$ 

- So for each fixed $x$ as $\sigma \to \infty$ this goes slowly to $-\infty$. (But of course $-5$ is very big in this scale.)
Bayes Factors

Bayes Factor Contours

Richard Lockhart
Statistical Issues in Discovery
Pentaquarks

Gauss + Gauss
Gauss + Gamma
Gamma Back
Gauss Back

Invariant Mass
Frequency
1.5 1.6 1.7 1.8 1.9
0 5 10 15 20 25 30 35
Pentaquarks

Emprical vs Fitted Distributions with Peak Background 3 parameter Gamma

Cumulative probability

Some physics thing
Emprical vs Fitted Distributions with no Peak
Background is 3 Parameter Gamma

Some physics thing
Cumulative probability

Richard Lockhart
Statistical Issues in Discovery

The Issues
Model
Systematics
Exclusion
Pentaquarks
Bayes Power
Post-discovery
A Bayesian trapped in frequentist world

- Must carry out fixed level $\alpha$ test.
- Must publish a protocol.
- Wants to reject $H_0$.
- Uses prior on alternative to design Neyman-Pearson test.
- Maximizes expected power.

A frequentist can use the idea to design tests.
I have used this to develop goodness-of-fit tests; same idea can be used in this mixture problem.

It looks to me like you have lots of knowledge about $f^*$ and the mixing proportions; I think that should be used even by frequentists.

Frequency theorists have a depressing tendency to do worst case analysis and to maximize or minimize everything in sight.

This leads, for instance, to all the pathologies of likelihood in mixture models.

I concede that some work is needed to compute $P$-values. My goodness of fit method (approximate contiguity calculation) gives linear combinations of non-linear chi-squares.
Having discovered one, you discover many

- Want to use the discovered population (of exoplanets, say) to describe the whole, undiscovered population.
- Know some discoveries false.
- Others have measurement errors – deconvolution needed.
- And probability of discovery depends on true properties and some measured values are not possible.
- Need to mix survey sampling non-response ideas with deconvolution and mixture modelling for the false discoveries.
- I hope someone here knows something about that.
This Δ-chi-squared stuff is a problem – the model is wrong.
I look forward to the talks without any current understanding.
Combining $P$-values

- Is this for meta-analysis – several different experiments?
- Typical situation. Each $P$ value is an upper tail probability from either normal, $t$ or linear combination of $\chi^2$ statistic.
- Each such has its own, possibly non zero, mean or non-centrality parameter.
- If all these shifts and so on depend on the same parameter of interest you really want the original analyses to put together.
- Otherwise why are you putting them together? How many nulls are likely to be false?
- Lack of associativity represents information loss in collapse to $P$-values.
The probability that both of two estimates are on the same side of the parameter begin estimated is not so small.

The fear of a combination which is not between the two estimates arises from fear the model is wrong?

Regression estimate: $X$ estimates $\mu$ and $Y$ estimates 0 and is correlated with $X$. So you pick $a$ to minimize $\text{Var}(X + aY)$.

Here $X$ is, say, high precision estimate and $Y$ is difference between the two estimates.
Some things I have yet to see

- Estimating equations.
- Admissibility and Bayes.
- Note to me: say something about independence in periodograms.
- Note to me: stop talking.