Practical Statistics for Particle Physics

Kyle Cranmer, New York University
Lecture 2
Outline

Lecture 1: Building a probability model

- preliminaries, the marked Poisson process
- incorporating systematics via nuisance parameters
- constraint terms
- examples of different “narratives” / search strategies

Lecture 2: Hypothesis testing

- simple models, Neyman-Pearson lemma, and likelihood ratio
- composite models and the profile likelihood ratio
- review of ingredients for a hypothesis test

Lecture 3: Limits & Confidence Intervals

- the meaning of confidence intervals as inverted hypothesis tests
- asymptotic properties of likelihood ratios
- Bayesian approach
Hypothesis Testing
Hypothesis testing

One of the most common uses of statistics in particle physics is Hypothesis Testing (e.g. for discovery of a new particle)

- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, $H_0$: eg. background-only
  - Alternate-Hypothesis $H_1$: eg. signal-plus-background

- one makes a measurement and then needs to decide whether to reject or accept $H_0$
Hypothesis testing

Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error $\alpha$
  - Rate of Type II $\beta$
  - Power $= 1 - \beta$

Treat the two hypotheses asymmetrically

- the Null is special.
  - Fix rate of Type I error, call it “the size of the test”

Now one can state “a well-defined goal”

- Maximize power for a fixed rate of Type I error
Hypothesis testing

The idea of a “5σ” discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually $5\sigma$ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy

![Graph showing probability distribution and events observed.](image)
Hypothesis testing

The idea of a “5σ” discovery criteria for particle physics is really a conventional way to specify the size of the test

› usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$

• eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

› but in higher dimensions it is not so easy
**Hypothesis testing**

The idea of a “5σ” discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually $5\sigma$ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy

[G. Cowan]
The Neyman-Pearson Lemma

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:
- the Null Hypothesis $H_0$ (background only)
- the Alternate Hypothesis $H_1$ (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W | H_0)$$

(Convention: if data falls in $W$ then we accept $H_0$)

Find the region $W$ such that we minimize the probability of wrongly accepting the $H_0$ (when $H_1$ is true)

$$\beta = P(x \in W | H_1)$$
The Neyman-Pearson Lemma

The region $W$ that minimizes the probability of wrongly accepting $H_0$ is just a contour of the Likelihood Ratio

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

Any other region of the same size will have less power

The likelihood ratio is an example of a Test Statistic, eg. a real–valued function that summarizes the data in a way relevant to the hypotheses that are being tested
A short proof of Neyman-Pearson

Consider the contour of the likelihood ratio that has size a given size (eg. probability under $H_0$ is $1-\alpha$)

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$
Now consider a variation on the contour that has the same size.
A short proof of Neyman-Pearson

Now consider a variation on the contour that has the same size (eg. same probability under $H_0$)
A short proof of Neyman-Pearson

Because the new area is outside the contour of the likelihood ratio, we have an inequality

\[ \frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \]

\[ P(\underbrace{\ldots}_{H_1}) < P(\underbrace{\ldots}_{H_0})k_\alpha \]
A short proof of Neyman-Pearson

\[
\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \quad \text{and} \quad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha
\]

\[
P(\bigcup H_0) = P(\bigcup H_0)
\]

\[
P(\bigcup H_1) < P(\bigcup H_0)k_\alpha \quad \text{and} \quad P(\bigcup H_1) > P(\bigcup H_0)k_\alpha
\]

And for the region we lost, we also have an inequality

Together they give...
A short proof of Neyman-Pearson

The new region has less power.
Often one uses the output of a neural network or multivariate algorithm in place of a true likelihood ratio.

- That’s fine, but what do you do with it?
- If you have a fixed cut for all events, this is what you are doing:

\[ q = \ln Q = -s + \ln \left( 1 + \frac{s f_s(x, y)}{bf_b(x, y)} \right) \]
Experiments vs. Events

Ideally, you want to cut on the likelihood ratio for your experiment

- equivalent to a sum of log likelihood ratios

Easy to see that includes experiments where one event had a high LR and the other one was relatively small
**LEP Higgs**

A simple likelihood ratio with no free parameters

\[
Q = \frac{L(x|H_1)}{L(x|H_0)} = \prod_{i}^{N_{\text{chan}}} \text{Pois}(n_i|s_i + b_i) \prod_{j}^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i + b_i} \\
\prod_{i}^{N_{\text{chan}}} \text{Pois}(n_i|b_i) \prod_{j}^{n_i} f_b(x_{ij})
\]

\[
q = \ln Q = -s_{\text{tot}} + \sum_{i} \sum_{j} \ln \left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)
\]

---

**Graphs:**

- **Graph (a)**: Observed, Expected for background, Expected for signal plus background.

- **Graph**: Observed, Expected for background, Expected for signal plus background.
The “test statistic” is a single number that quantifies the entire experiment, it could just be number of events observed, but often its more sophisticated, like a likelihood ratio. What test statistic do we choose?

And how do we build the distribution? Usually “toy Monte Carlo”, but what about the uncertainties… what do we do with the nuisance parameters?
Recall our marked Poisson model

- **observables**: \( n \) events each with some value of discriminating variable \( m \)
- **auxiliary measurements**: \( a_i \)
- **parameters**: \( \alpha \)

\[
P(m, a | \alpha) = \text{Pois}(n | s(\alpha) + b(\alpha)) \prod_{j} \frac{s(\alpha) f_s(m_j | \alpha) + b(\alpha) f_b(m_j | \alpha)}{s(\alpha) + b(\alpha)} \times \prod_{i \in \text{syst}} G(a_i | \alpha_i, \sigma_i)
\]

Useful to separate parameters into \( \alpha = (\mu, \nu) \)

- **parameters of interest** \( \mu \): cross sections, masses, coupling constants, ...
- **nuisance parameters** \( \nu \): reconstruction efficiencies, energy scales, ...

- note: not all of the nuisance parameters need to have constraint terms
Our number counting example

From our general model

\[ P(m, a|\alpha) = \text{Pois}(n|s(\alpha) + b(\alpha)) \prod_j^n \frac{s(\alpha)f_s(m_j|\alpha) + b(\alpha)f_b(m_j|\alpha)}{s(\alpha) + b(\alpha)} \times \prod_{i \in \text{syst}} G(a_i|\alpha_i, \sigma_i) \]

Consider a simple number counting model with \( s(\alpha) \rightarrow s, b(\alpha) \rightarrow b \), and replace the constraint \( G(a|\alpha, \sigma) \rightarrow \text{Pois}(n_{\text{off}} | \tau b) \) with \( \tau \) known.

\[ P(n_{\text{on}}, n_{\text{off}}|s, b) = \text{Pois}(n_{\text{on}}|s + b) \text{Pois}(n_{\text{off}}|\tau b). \]

We could simply use \( n_{\text{on}} \) as our test statistic, but to calculate the p-value we need to know distribution of \( n_{\text{on}} \).

\[ p = \sum_{n_{\text{on}}=n_{\text{obs}}}^{\infty} \text{Pois}(n_{\text{on}}|s + b) \times \underbrace{\text{Pois}(n_{\text{off}}|\tau b)}_{\text{constant}} \]

Observations:

- The distribution of \( n_{\text{on}} \) explicitly depends on both \( s \) and \( b \).
- The distribution of \( n_{\text{off}} \) is independent of \( s \)
- If \( \tau b \) is very different from \( n_{\text{off}} \), then the data are not consistent with the model parameters. However, the p-value derived from \( n_{\text{on}} \) is not small.
With nuisance parameters: Hybrid Solutions

Goal of Bayesian-frequentist hybrid solutions is to provide a frequentist treatment of the main measurement, while eliminating nuisance parameters (deal with systematics) with an intuitive Bayesian technique.

\[
P(n_{on}|s) = \int db \text{Pois}(n_{on}|s + b) \pi(b), \quad p = \sum_{n_{on}=n_{obs}}^{\infty} P(n_{on}|s)
\]

Tracing back the origin of \(\pi(b)\)

- clearly state prior \(\eta(b)\); identify control samples (sidebands) and use:

\[
\pi(b) = P(b|n_{off}) = \frac{P(n_{off}|b)\eta(b)}{\int db P(n_{off}|b)\eta(b)}.
\]

In a purely Frequentist approach we must need a test statistic that depends on both \(n_{on}\) and \(n_{off}\) and we must consider both random (eg. when generating toy Monte Carlo)

\[
P(n_{on}, n_{off}|s, b) = \text{Pois}(n_{on}|s + b) \text{Pois}(n_{off}|\tau b).
\]
**Does it matter?**

This on/off problem has been studied extensively.

- instead of arguing about the merits of various methods, just go and check their rate of Type I error
- Results indicated large discrepancy in “claimed” significance and “true” significance for various methods
- eg. 5σ is really ~4σ for some points

So, yes, it does matter.

\[ P(n_{on}, n_{off}|s, b) = \text{Pois}(n_{on}|s + b) \text{Pois}(n_{off}|\tau b). \]

---

Figure 7. A comparison of the various methods critical boundary \( x_{crit}(y) \) (see text). The concentric ovals represent contours of \( L_G \) from Eq. 15.

http://www.physics.ox.ac.uk/phystat05/proceedings/files/Cranmer_LHCStatisticalChallenges.ps
Does it matter?

This on/off problem has been studied extensively.

- instead of arguing about the merits of various methods, just go and check their rate of Type I error
- Results indicated large discrepancy in “claimed” significance and “true” significance for various methods
- eg. 5σ is really ~4σ for some points

So, yes, it does matter.

Follow-up work by Bob Cousins & Jordan Tucker, [physics/0702156]

\[
P(n_{on}, n_{off}| s, b) = \text{Pois}(n_{on}| s + b) \text{Pois}(n_{off}| \tau b).
\]

http://www.physics.ox.ac.uk/phystat05/proceedings/files/Cranmer_LHCStatisticalChallenges.ps
The Profile Likelihood Ratio

Consider our general model with a single parameter of interest $\mu$

- let $\mu=0$ be no signal, $\mu=1$ nominal signal

In the LEP approach the likelihood ratio is equivalent to:

$$Q_{\text{LEP}} = \frac{P(m|\mu = 1, \nu)}{P(m|\mu = 0, \nu)}$$

- but this variable is sensitive to uncertainty on $\nu$ and makes no use of auxiliary measurements $a$

Alternatively, one can define profile likelihood ratio

$$\lambda(\mu) = \frac{P(m, a|\mu, \hat{\nu}(\mu; m, a))}{P(m, a|\hat{\mu}, \hat{\nu})}$$

- where $\hat{\nu}(\mu; m, a)$ is best fit with $\mu$ fixed (the constrained maximum likelihood estimator, depends on data)

- and $\hat{\nu}$ and $\hat{\mu}$ are best fit with both left floating (unconstrained)

- Tevatron used $Q_{\text{Tev}} = \lambda(\mu=1)/\lambda(\mu=0)$ as generalization of $Q_{\text{LEP}}$
An example

Essentially, you need to fit your model to the data twice: once with everything floating, and once with signal fixed to 0

\[
\lambda(\mu = 0) = \frac{P(m, a|\mu = 0, \hat{\nu}(\mu = 0; m, a))}{P(m, a|\hat{\mu}, \hat{\nu})}
\]

Figure 14 shows the fit to the signal candidates for a \(m_H = 120\) GeV Higgs with (a,c) and without (b,d) the signal contribution. It can be seen that the background shape and normalizations are trying to accommodate the excess near \(m = 120\) GeV, but the control samples are constraining the variation.

Table 13 shows the significance calculated from the profile likelihood ratio for the \(ll\)-channel, the \(lh\)-channel, and the combined fit for various Higgs boson masses.
Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

\[ \lambda(\mu) = \frac{P(m, a|\mu, \hat{\nu}(\mu; m, a))}{P(m, a|\hat{\mu}, \hat{\nu})} \]

one can see the function is independent of true values of \( \nu \)

- though its distribution might depend indirectly

Wilks’s theorem states that under certain conditions the distribution of \(-2 \ln \lambda(\mu=\mu_0)\) given that the true value of \( \mu \) is \( \mu_0 \) converges to a chi-square distribution

- more on this tomorrow, but the important points are:
  - “asymptotic distribution” is known and it is independent of \( \nu \)
    - more complicated if parameters have boundaries (eg. \( \mu \geq 0 \))

Thus, we can calculate the p-value for the background-only hypothesis without having to generate Toy Monte Carlo!
Toy Monte Carlo

Explicitly build distribution by generating “toys” / pseudo experiments assuming a specific value of $\mu$ and $\nu$.

- randomize both main measurement $m$ and auxiliary measurements $a$
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate $-2\ln \lambda(\mu)$ and add to histogram

Choice of $\mu$ is straightforward: typically $\mu=0$ and $\mu=1$, but choice of $\nu$ is less clear

- more on this tomorrow

This can be very time consuming. Plots below use millions of toy pseudo-experiments on a model with ~50 parameters.
What makes a statistical method

To describe a statistical method, you should clearly specify

• choice of a test statistic
  • simple likelihood ratio (LEP) $Q_{LEP} = L_{s+b}(\mu = 1)/L_b(\mu = 0)$
  • ratio of profiled likelihoods (Tevatron) $Q_{TEV} = L_{s+b}(\mu = 1, \hat{\nu})/L_b(\mu = 0, \hat{\nu}')$
  • profile likelihood ratio (LHC) $\lambda(\mu) = L_{s+b}(\mu, \hat{\nu})/L_{s+b}(\hat{\mu}, \hat{\nu})$

• how you build the distribution of the test statistic
  • toy MC randomizing nuisance parameters according to $\pi(\nu)$
    • aka Bayes-frequentist hybrid, prior-predictive, Cousins-Highland
  • toy MC with nuisance parameters fixed (Neyman Construction)
  • assuming asymptotic distribution (Wilks and Wald, more tomorrow)

• what condition you use for limit or discovery
  • more on this tomorrow
Experimentalist Justification

So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio?

It won’t unless the parametrization is sufficiently flexible.

So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square.

Here it is pretty stable, but it’s not perfect (and this is a log plot, so it hides some pretty big discrepancies).

For the distribution to be independent of the nuisance parameters your parametrization must be sufficiently flexible.
A very important point

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically.

The p-values are usually predicated on the assumption that the true distribution is in the family of functions being considered:

- eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
- but we don’t believe we simulate everything perfectly
- ..and when we parametrize our models usually we have further approximated our simulation.

- nature -> simulation -> parametrization

At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!