Statistics for Particle Physics Lecture 1



Helmholtz Alliance



Terascale Statistics School https://indico.desy.de/event/46667/

DESY, Hamburg 24-28 Feb 2025



Glen Cowan Physics Department Royal Holloway, University of London g.cowan@rhul.ac.uk www.pp.rhul.ac.uk/~cowan

Outline

Lectures/tutorials from me:

→ Monday 15:30 Hypothesis testing
 Monday 16:30 Hypothesis testing tutorial
 Tuesday 10:30 Parameter estimation tutorial
 Tuesday 14:00 Setting limits
 Wednesday 9:00 Bayesian parameter estimation
 Wednesday 11:00 Errors on errors

More resources in the University of London course: https://www.pp.rhul.ac.uk/~cowan/stat_course.html

G. Cowan / RHUL Physics

Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers x.

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

 $P(\mathbf{x}|H)$ = the likelihood of H

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$$P(\mathbf{x}|oldsymbol{ heta}) = L(oldsymbol{ heta})$$
 = the "likelihood function"

Note:

1) For the likelihood we treat the data x as fixed.

2) The likelihood function $L(\theta)$ is not a pdf for θ .

Frequentist hypothesis tests

Suppose a measurement produces data x; consider a hypothesis H_0 we want to test and alternative H_1

 H_0 , H_1 specify probability for x: $P(x|H_0)$, $P(x|H_1)$

A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α , assuming H_0 is correct, to observe the data there, i.e.,

$$P(\mathbf{x} \in w \mid H_0) \le \alpha$$

Need inequality if data are discrete.

 α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0 .



Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same size $\langle . \rangle$

Use the alternative hypothesis H_1 to motivate where to place the critical region.

Roughly speaking, place the critical region where there is a low probability (α) to be found if H_0 is true, but high if H_1 is true:



Classification viewed as a statistical test

Suppose events come in two possible types:

s (signal) and b (background)

For each event, test hypothesis that it is background, i.e., $H_0 = b$.

Carry out test on many events, each is either of type s or b, i.e., here the hypothesis is the "true class label", which varies randomly from event to event, so we can assign to it a frequentist probability.

Select events for which where H_0 is rejected as "candidate events of type s". Equivalent Particle Physics terminology:

background efficiency

$$\varepsilon_{\mathbf{b}} = \int_{W} f(\mathbf{x}|H_0) \, d\mathbf{x} = \alpha$$

signal efficiency

$$\varepsilon_{\mathbf{s}} = \int_{W} f(\mathbf{x}|H_1) \, d\mathbf{x} = 1 - \beta = \text{power}$$

G. Cowan / RHUL Physics

Particle Physics context for a hypothesis test

A simulated SUSY event ("signal"):



G. Cowan / RHUL Physics

Background events



This event from Standard Model ttbar production also has high $p_{\rm T}$ jets and muons, and some missing transverse energy.

→ can easily mimic a signal event.

Classification of proton-proton collisions

Proton-proton collisions can be considered to come in two classes: signal (the kind of event we're looking for, y = 1) background (the kind that mimics signal, y = 0)

For each collision (event), we measure a collection of features:

 $x_1 =$ energy of muon $x_4 =$ missing transverse energy $x_2 =$ angle between jets $x_5 =$ invariant mass of muon pair $x_3 =$ total jet energy $x_6 = ...$

The real events don't come with true class labels, but computersimulated events do. So we can have a set of simulated events that consist of a feature vector x and true class label y (0 for background, 1 for signal):

$$(x, y)_1, (x, y)_2, ..., (x, y)_N$$

The simulated events are called "training data".

G. Cowan / RHUL Physics

Distributions of the features

If we consider only two features $\mathbf{x} = (x_1, x_2)$, we can display the results in a scatter plot (red: y = 0, blue: y = 1).

For real events, the dots are black (true type is not known).

For each real event test the hypothesis that it is background.

(Related to this: test that a sample of events is *all* background.)

The test's critical region is defined by a "decision boundary" – without knowing the event type, we can classify them by seeing where their measured features lie relative to the boundary.



Decision function, test statistic



A surface in an *n*-dimensional

Different values of the constant t_c result in a family of surfaces.

Problem is reduced to finding the best decision function or test statistic t(x).



Distribution of t(x)

By forming a test statistic t(x), the boundary of the critical region in the *n*-dimensional *x*-space is determined by a single single value t_c .



Types of decision boundaries

So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic t(x)?

First find best t(x), later address issue of optimal size of test.

Remember *x*-space can have many dimensions.



Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way', in particular if the data space is multidimensional?

Neyman-Pearson lemma states:

For a test of H_0 of size α , to get the highest power with respect to the alternative H_1 we need for all x in the critical region W

"likelihood ratio (LR)"
$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \ge c_{\alpha}$$

inside W and $\leq c_{\alpha}$ outside, where c_{α} is a constant chosen to give a test of the desired size.

Equivalently, optimal scalar test statistic is

$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this is leads to the same test.

G. Cowan / RHUL Physics

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs f(x|s), f(x|b), so for a given x we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate
$$\boldsymbol{x} \sim f(\boldsymbol{x}|s) \rightarrow \boldsymbol{x}_1, \dots, \boldsymbol{x}_N$$

generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathbf{b}) \rightarrow \boldsymbol{x}_1, \dots, \boldsymbol{x}_N$

This gives samples of "training data" with events of known type.

 Use these to construct a statistic that is as close as possible to the optimal likelihood ratio (→ Machine Learning).

G. Cowan / RHUL Physics

Approximate LR from histograms

Want t(x) = f(x/s)/f(x/b) for x here



One possibility is to generate MC data and construct histograms for both signal and background.

Use (normalized) histogram values to approximate LR:

$$t(x) \approx \frac{N(x|s)}{N(x|b)}$$

Can work well for single variable.

Approximate LR from 2D-histograms Suppose problem has 2 variables. Try using 2-D histograms:



Approximate pdfs using N(x,y/s), N(x,y/b) in corresponding cells. But if we want M bins for each variable, then in n-dimensions we have M^n cells; can't generate enough training data to populate. \rightarrow Histogram method usually not usable for n > 1 dimension.

G. Cowan / RHUL Physics

Strategies for multivariate analysis

Neyman-Pearson lemma gives optimal answer, but cannot be used directly, because we usually don't have f(x|s), f(x|b).

Histogram method with M bins for n variables requires that we estimate M^n parameters (the values of the pdfs in each cell), so this is rarely practical.

A compromise solution is to assume a certain functional form for the test statistic t(x) with fewer parameters; determine them (using MC) to give best separation between signal and background.

Alternatively, try to estimate the probability densities $f(\mathbf{x}|s)$ and $f(\mathbf{x}|b)$ (with something better than histograms) and use the estimated pdfs to construct an approximate likelihood ratio.

Multivariate methods (Machine Learning)

Many new (and some old) methods:

Fisher discriminant

(Deep) Neural Networks

Kernel density methods

Support Vector Machines

Decision trees

Boosting

Bagging

More in the lectures by Tilman Plehn

Testing significance / goodness-of-fit

Suppose hypothesis *H* predicts pdf f(x|H) for a set of observations $x = (x_1,...,x_n)$.

We observe a single point in this space: x_{obs} .

How can we quantify the level of compatibility between the data and the predictions of *H*?

Decide what part of the data space represents equal or less compatibility with H than does the point x_{obs} . (Not unique!)



p-values

Express level of compatibility between data and hypothesis (sometimes 'goodness-of-fit') by giving the *p*-value for *H*:

 $p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{obs})|H)$

- probability, under assumption of H, to observe data
 with equal or lesser compatibility with H relative to the
 data we got.
- probability, under assumption of H, to observe data as discrepant with H as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then *H* is "disfavoured by the data".

If the *p*-value is below a user-defined threshold α (e.g. 0.05) then *H* is rejected (equivalent to hypothesis test of size α as seen earlier).



The *p*-value of H is not the probability that *H* is true!

In frequentist statistics we don't talk about P(H) (unless H represents a repeatable observation).

If we do define P(H), e.g., in Bayesian statistics as a degree of belief, then we need to use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H) \, dH}$$

where $\pi(H)$ is the prior probability for H.

For now stick with the frequentist approach; result is p-value, regrettably easy to misinterpret as P(H). The Poisson counting experiment Suppose we do a counting experiment and observe *n* events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

$$P(n|s,b) = \frac{(s+b)^n}{n!}e^{-(s+b)}$$

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about *s*, e.g.,

test s = 0 (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero *s* (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

G. Cowan / RHUL Physics

Poisson counting experiment: discovery *p*-value Suppose b = 0.5 (known), and we observe $n_{obs} = 5$.

Should we claim evidence for a new discovery?

Give *p*-value for hypothesis s = 0, suppose relevant alt. is s > 0.

$$p$$
-value = $P(n \ge 5; b = 0.5, s = 0)$
= $1.7 \times 10^{-4} \ne P(s = 0)!$



G. Cowan / RHUL Physics

Significance from *p*-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_{Z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

 $Z = \Phi^{-1}(1-p)$

in ROOT: in pyt
p = 1 - TMath::Freq(Z) p = 1
Z = TMath::NormQuantile(1-p) Z = no

in python (scipy.stats): p = 1 - norm.cdf(Z) = norm.sf(Z) Z = norm.ppf(1-p)

Result Z is a "number of sigmas". Note this does not mean that the original data was Gaussian distributed.

G. Cowan / RHUL Physics

Poisson counting experiment: discovery significance Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1-p) = 3.6$ Often claim discovery if Z > 5 ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: *p*-value intended to quantify probability of a signallike fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

Extra slides

Example of a test for classification



For each event in a mixture of signal (s) and background (b) test

 H_0 : event is of type b

using a critical region *W* of the form: $W = \{x : x \le x_c\}$, where x_c is a constant that we choose to give a test with the desired size α .

Classification example (2)

Suppose we want $\alpha = 10^{-4}$. Require:

$$\alpha = P(x \le x_{c}|b) = \int_{0}^{x_{c}} f(x|b) \, dx = \frac{4x^{4}}{4} \Big|_{0}^{x_{c}} = x_{c}^{4}$$

and therefore $x_{
m c} = lpha^{1/4} = 0.1$

For this test (i.e. this critical region W), the power with respect to the signal hypothesis (s) is

$$M = P(x \le x_{\rm c}|{\rm s}) = \int_0^{x_{\rm c}} f(x|{\rm s}) \, dx = 2x_{\rm c} - x_{\rm c}^2 = 0.19$$

Note: the optimal size and power is a separate question that will depend on goals of the subsequent analysis.

G. Cowan / RHUL Physics

Classification example (3)

Suppose that the prior probabilities for an event to be of type s or b are:

 $\pi_{\rm s} = 0.001$ $\pi_{\rm b} = 0.999$

The "purity" of the selected signal sample (events where b hypothesis rejected) is found using Bayes' theorem:

$$P(\mathbf{s}|x \le x_{\mathbf{c}}) = \frac{P(x \le x_{\mathbf{c}}|\mathbf{s})\pi_{\mathbf{s}}}{P(x \le x_{\mathbf{c}}|\mathbf{s})\pi_{\mathbf{s}} + P(x \le x_{\mathbf{c}}|\mathbf{b})\pi_{\mathbf{b}}}$$

= 0.655

Proof of Neyman-Pearson Lemma

Consider a critical region W and suppose the LR satisfies the criterion of the Neyman-Pearson lemma:

 $P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0) \geq c_{\alpha} \text{ for all } \boldsymbol{x} \text{ in } W,$ $P(\boldsymbol{x}|H_1)/P(\boldsymbol{x}|H_0) \leq c_{\alpha} \text{ for all } \boldsymbol{x} \text{ not in } W.$

Try to change this into a different critical region W' retaining the same size α , i.e.,

$$P(\mathbf{x} \in W'|H_0) = P(\mathbf{x} \in W|H_0) = \alpha$$

To do so add a part δW_+ , but to keep the size α , we need to remove a part δW_- , i.e.,

$$W \to W' = W + \delta W_+ - \delta W_-$$

$$P(\mathbf{x} \in \delta W_+ | H_0) = P(\mathbf{x} \in \delta W_- | H_0)$$





Proof of Neyman-Pearson Lemma (2)

But we are supposing the LR is higher for all x in δW_{-} removed than for the x in δW_{+} added, and therefore

$$P(\mathbf{x} \in \delta W_+ | H_1) \le P(\mathbf{x} \in \delta W_+ | H_0) c_\alpha$$

$$\frac{\delta W_{+}}{\delta W_{-}}$$

$$P(\mathbf{x} \in \delta W_{-}|H_{1}) \ge P(\mathbf{x} \in \delta W_{-}|H_{0})c_{\alpha}$$

The right-hand sides are equal and therefore

 $P(\mathbf{x} \in \delta W_+ | H_1) \le P(\mathbf{x} \in \delta W_- | H_1)$

Proof of Neyman-Pearson Lemma (3)

We have

$$W \cup W' = W \cup \delta W_+ = W' \cup \delta W_-$$

Note W and δW_+ are disjoint, and W' and δW_- are disjoint, so by Kolmogorov's 3rd axiom,

$$\frac{\delta W_{+}}{\delta W_{-}}$$

$$P(\mathbf{x} \in W') + P(\mathbf{x} \in \delta W_{-}) = P(\mathbf{x} \in W) + P(\mathbf{x} \in \delta W_{+})$$

Therefore

$$P(\mathbf{x} \in W'|H_1) = P(\mathbf{x} \in W|H_1) + P(\mathbf{x} \in \delta W_+|H_1) - P(\mathbf{x} \in \delta W_-|H_1)$$

Proof of Neyman-Pearson Lemma (4)

And therefore

$$P(\mathbf{x} \in W'|H_1) \le P(\mathbf{x} \in W|H_1)$$

i.e. the deformed critical region W' cannot have higher power than the original one that satisfied the LR criterion of the Neyman-Pearson lemma.

Extending s/\sqrt{b} to case where b uncertain

The intuitive explanation of s/\sqrt{b} is that it compares the signal, s, to the standard deviation of n assuming no signal, \sqrt{b} .

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b .

A reasonable guess is to replace \sqrt{b} by the quadratic sum of \sqrt{b} and σ_b , i.e.,

$$\operatorname{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Profile likelihood with *b* uncertain

This is the well studied "on/off" problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

 $n \sim \text{Poisson}(s+b)$ (primary or "search" measurement) $m \sim \text{Poisson}(\tau b)$ (control measurement, τ known) The likelihood function is

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct profile likelihood ratio (*b* is nuisance parameter): $L(0, \hat{\hat{b}}(0))$

$$\lambda(0) = \frac{L(0, b(0))}{L(\hat{s}, \hat{b})}$$

G. Cowan / RHUL Physics

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\begin{split} \hat{s} &= n - m/\tau \ , \\ \hat{b} &= m/\tau \ , \\ \hat{b}(s) &= \frac{n + m - (1 + \tau)s + \sqrt{(n + m - (1 + \tau)s)^2 + 4(1 + \tau)sm}}{2(1 + \tau)} \end{split}$$

and in particular to test for discovery (s = 0),

$$\hat{\hat{b}}(0) = \frac{n+m}{1+\tau}$$

G. Cowan / RHUL Physics

Asymptotic significance

Use profile likelihood ratio for q_0 , and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0}$$
$$= \left[-2\left(n\ln\left[\frac{n+m}{(1+\tau)n}\right] + m\ln\left[\frac{\tau(n+m)}{(1+\tau)m}\right]\right) \right]^{1/2}$$

for $n > \hat{b}$ and Z = 0 otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480– 501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

G. Cowan / RHUL Physics

Asimov approximation for median significance

To get median discovery significance, replace *n*, *m* by their expectation values assuming background-plus-signal model:

$$n \to s + b$$

$$m \to \tau b$$

$$Z_{\rm A} = \left[-2\left((s+b) \ln\left[\frac{s+(1+\tau)b}{(1+\tau)(s+b)}\right] + \tau b \ln\left[1+\frac{s}{(1+\tau)b}\right] \right) \right]^{1/2}$$
Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \frac{b}{\tau}$, to eliminate τ :

$$Z_{\rm A} = \left[2\left((s+b) \ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right] - \frac{b^2}{\sigma_b^2} \ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right] \right) \right]^{1/2}$$

Limiting cases

Expanding the Asimov formula in powers of s/b and $\sigma_b^2/b~(=1/\tau)$ gives

$$Z_{\rm A} = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

So the "intuitive" formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set. Testing the formulae: s = 5



G. Cowan / RHUL Physics

Using sensitivity to optimize a cut



Figure 1: (a) The expected significance as a function of the cut value x_{cut} ; (b) the distributions of signal and background with the optimal cut value indicated.

G. Cowan / RHUL Physics