Hypothesis Testing and Confidence Intervals

Terascale Statistics School

Roger Barlow

7th March 2017

University of HUDDERSFIELD
International Institute
for Accelerator Applications

Contents

- Hypothesis Testing: signal or background?
- Hypothesis testing: Is there a discovery?
- χ^2 and $p-$values
- The need for 5 σ
- Confidence Intervals
- Credible intervals

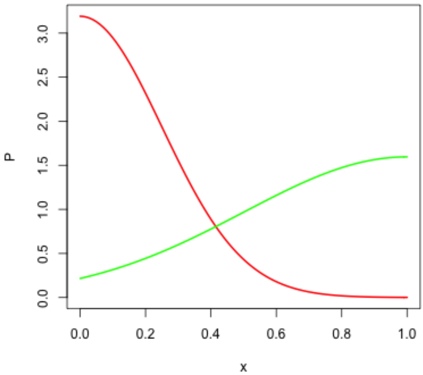
Making Choices ('Hypothesis Testing')

Classification problem: Separation into two (or more) classes: e.g. Signal and Background events

Hypothesis is 'This is a signal(red) not background(green)'

Indicator variable x - may be multidimensional and/or output from ML system

Behaviour known - from simulations or from control samples



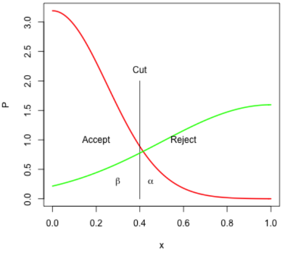
Put a cut somewhere and accept all that pass Accept that some signal events will be lost (efficiency) and some background events accepted (purity)

Type I and Type II errors

Two sorts of errors possible:

Type I: reject a signal event. Lowers efficiency

Type II: accept a background event. Lowers purity, increases contamination



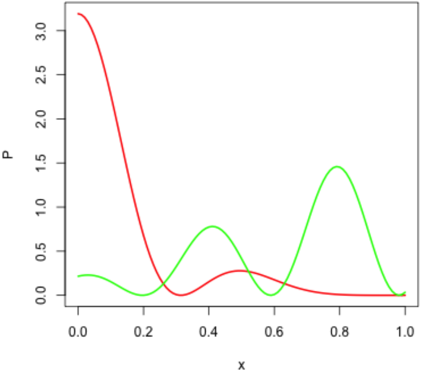
α is the probability of a type I error $\int S(x)dx$ over the rejection region. Called 'Significance'

β is the probability of a type II error $\int B(x)dx$ over acceptance region. $1-\beta$ called 'Power'

Neyman-Pearson Lemma

How to choose the best acceptance/rejection region even in cases where $S(x)$ and $B(x)$ have lots more structure.

α is imposed. You want to minimise β.



Add regions of x with the highest $S(x)/B(x)$ to the acceptance region until α is achieved. β is then minimised, as replacing any Δx with an equivalent region from the rejection region would bring in more $B(x)$

Composite Hypothesis

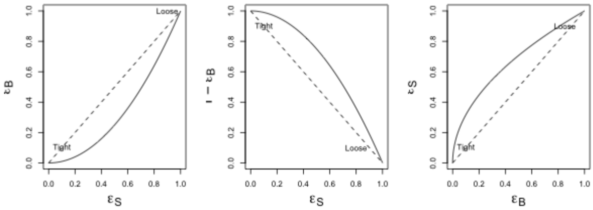
Suppose you have several different sources of background - and their relative sizes are not known so you can't just lump them together

Or a hypothesis containing an adjustable parameter, or parameters

Doesn't affect α

Need to quote worst (i.e. largest) β. Then can say background contamination probabilty$\leq \beta$

ROC plots



Varying cut value varies α, or $1 - E_S$ and β, or E_B

Plot E_S against E_B, or $1 - E_S$ against E_B, or...

'ROC plots' : also called efficiency plots, efficiency/contamination plots, and (wrongly!) efficiency/purity plots Show how good your separation technique is by how far they bend from diagonal. ROC stands for Receiver Operating Characteristic (jargon but sounds good)

The Null Hypothesis H_0

You want to test a theory. Do an experiment. Data may show the theory is wrong - but how do you show it's right using frequentist tools?

‘The data are compatible with the theory being true’ says nothing.

Have to test 'The theory are not compatible with the theory being false’

Construct Null Hypothesis H_0 which is that your theory is false.

Examples: There are no dark matter events in your sample
Web advertising does not increase sales
There are no bumps in the mass plot
New Treatment has no effect on cure rate
.....

‘Every experiment may be said to exist only to have the chance to disprove the null hypothesis’ - Fisher

Type I and Type II errors revisited

Type I: reject a signal event → Probability of rejecting the H_0 when it is true. 'False positive' probability

Description of α as 'significance' makes sense. Set $\alpha = 0.05$ and get a result in the reject region. Hooray! Null hypothesis rejected at significance level of 0.05.

Also quote as Confidence level $1 - \alpha$. Here 95%.

If H_0 passes the test I say, with 95% confidence 'There is no effect'. Statement is either true or false, but is a member of an ensemble of statements, 95% true.

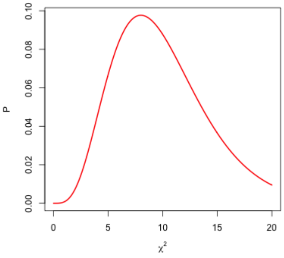
Legal point: I say with 95% confidence that a statement is true if it belongs to an ensemble of which *at least* 95% of statements are true. Extra two words mean (i) a statement true with 96% confidence is also true with 95% confidence and (ii) composite hypotheses are covered by taking the worst case.

Type II: accept a background event → fail to reject H_0 given the chance

χ^2 and Goodness of Fit

Tool often used in Hypothesis testing

Data $\{x_i, y_i\}$ and H_0 is that $y = f(x)$



Define $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2$

Expectation $\chi^2 = N$. (broadly obvious, and can be proved)

Figure shows $N = 20$

$P(\chi^2; N) = \frac{2^{-N/2}}{\Gamma(N/2)} \chi^{N-2} e^{-\chi^2/2}$

Comes from integrating over multiple Gaussians.

Use χ^2 as indicator variable for H_0 (i.e. does the curve $y = f(x)$ go through the data points $\{x_i, y_i \pm \sigma_i\}$)

More about χ^2

You will also see $\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{y_i}$. This is a simplified version and applies only to histograms

If $f(x)$ actually $f(x;a)$ and a adjusted to minimise χ^2, replace N by N_D: number of data points minus number of adjusted parameters. (Nice property of χ^2: assumes fitting acts to reduce dimensionality of χ space.)

If χ^2 is large, reject H_0. Calculate $p = \int_{\chi^2}^{\infty} P(\chi'^2; N) d\chi'^2$ - probability of getting a χ^2 value this large or worse.

If χ^2 is small - be suspicious and check your error calculations.

χ^2 is a really nice tool: minimising it gives estimates (and their errors) *and* its value gives goodness-of-fit. Contrast likelihood, where the actual value is meaningless.

p values

p values are not just for χ^2: - probability of getting a value this extreme or worse for any measure of agreement.

The Difference between p and α

In one sense, none. Both are $\int_x^\infty P(x')\,dx'$

In another sense, plenty. α is computed *before* you see the data and is a property of the test. p is a property of the data.

Wilks' Theorem

Getting a lot of use recently

Start with variant of χ^2 test. Suppose 2 models H_0 and H_1, where H_1 is similar to H_0 but has more parameters (e.g. H_0 is straight line fit, H_1 is quadratic)

The difference $\Delta\chi^2$ between the two is also χ^2 distributed, with N_D the number of new parameters.

Use to answer question 'Do the extra parameters really help?' Sort of decoupled from total χ^2 values

Wilks extends this to likelihoods: $-2\Delta lnL$ has a χ^2 distribution. (Hence name 'Likelihood Ratio test')

Proviso (1) applies in large n limit and (2) Models must be nested: for particular parameter value H_1 reverts to H_0

Five sigma

$p-$ values often expressed in terms of Gaussian σ.

E.g. probability of Gaussian exceeding 3 σ is 0.13%. (1-tailed). Call a $p-$value of 0.13% or below '3 sigma'

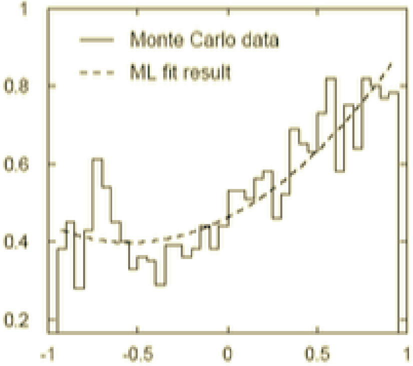
```
Do these calculations in R with 1-pnorm(Nsigma) and
qnorm(1-p)
```

For a 'discovery' claim we need 5 sigma. (Probability below 3×10^{-7})

Harsh? Yes, but justified by history

Look Elsewhere effect

Multiply number of particle physicists by number of plots they can draw per day times days in year - many thousand. There will be 3 sigma effects!



If you histogram random numbers, you see lots of apparently significant bumps

Blind Analysis

Cuts must be performed without looking at the final sample

Otherwise the analyst will tweak cut values to enhance the signal size, or the effect they are looking for.

The story of the split A_2 meson is a cautionary tale

We are better today but not immune - as the Z(750) shows

Statistics texts warn against publication bias - if 20 studies are done, one will (probably) report a 5% significance effect. OK, except it may be the only one that gets published.

We probably err in the reverse direction. If a measurement agrees with the Standard Model, we publish. If it doesn't we check and wait for more data.

Confidence Intervals

Given Gaussian measurement (say) $M = 123 \pm 4.5 GeV$.

We know this was sampled from a distribution $\propto \exp^{-\frac{1}{2}(\frac{M-\mu}{4.5})^2}$

But μ is unknown and we want to say something about it

Argument:
M will be within 4.5 GeV of μ in 68% of all cases
μ will be within 4.5 GeV of M in 68% of all cases
If I say 'μ lies between 118.5 and 127.5 GeV' I have a 68% chance of being correct. (The statement is a member of an ensemble of statements of which 68% are true)

118.5 to 125.5 GeV is a 68% confidence interval

Frequentists manage to say something useful about uncertainty on measurement without violating their probability definition!

Choices

The choice (68%, 95%, whatever) is up to the author. Tradeoff between precision and confidence.

The ends of the interval do not have to be symmetric, provided integral is chosen value. Other choices include lower limit, upper limit, shortest interval and central (probability of both tails equal)

Gaussians and Likelihood

Measurement may not come labeled as $\mu \pm \sigma$

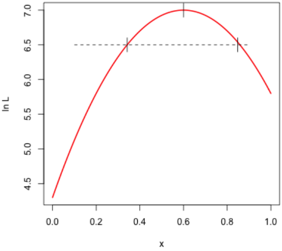
For true μ, log likelihood is $-\frac{1}{2}\left(\frac{M-\mu}{\sigma}\right)^2 +$ constant term

This is a parabola. $\Delta \ln L = -\frac{1}{2}$ at $x = \pm\sigma$

$\frac{1}{\sigma^2}$ is the second derivative.

We do not have access to the true *lnL* but we do have access to *lnL* at $\hat{\mu}$ and the second derivative is (hopefully) not that different

So estimate true σ from our $\frac{d^2 lnL}{d\mu^2}$ and estimate that from points where $\Delta lnL = -\frac{1}{2}$



This defines a 68% confidence interval.

χ^2 again

$\Delta \ln L = -\frac{1}{2}$ translates to $\Delta\chi^2 = +1$

In 1-D, probability of $\chi^2 < 1$ is 68%

In 2-D, probability of $\chi^2 < 1$ is 39% (`in R, pchisq(1.0,2)`)

If you want to find a region for which the chance is 68.3%, need $\Delta\chi^2 = 2.30$ (`qchisq(0.683,2)`)

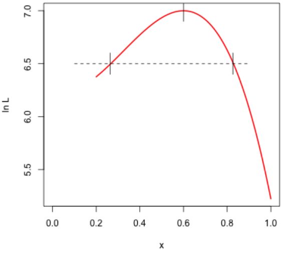
For general n-D, and other probability values, use `pchisq` and `qchisq`

Asymmetric Statistical Errors

Maybe your ln L is not a nice parabola

Δ ln L points are then not symmetric.

Argument for quoting them is: could transform to some μ' for which ln L was parabolic. Extract symmetric errors. Then transform back.



OK but what later? How should you use combination of errors when errors are asymmetric?

Standard procedure is to combine positive and negative errors separately
This is WRONG as it violates the Central Limit Theorem

For a better method, see RB 'Asymmetric Statistical Errors'
arXiv:physics/0406120

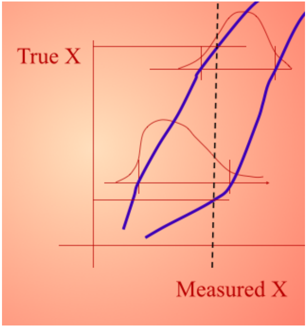
Confidence bands

Constructed Horizontally, read vertically

Gaussian measures are easy to handle because of their symmetry. What about more general $P(M; \mu)$ giving probability that a true μ will result in a measured M

Consider M, μ plane. For each μ, select the range of M according to value and choice (i.e. 68% or 95%, central or shortest...)

This defines the confidence belt.



Then take your measurement of M. We say with 68% (or whatever) confidence that (M, μ) lies inside the belt. So we read off the limits on μ

Note that the upper/lower limits in construction give the lower/upper limits in readout

Poisson confidence intervals

Extra complication as $x - axis$ is discrete. $\int_{M \epsilon region} P(M, \mu)\, dM$ becomes $\sum_{r \epsilon region} P(r, \mu)$

May not be possible to choose region to satisfy probability exactly: inclusion may bring it over the value - but that's OK due to the 'at least'

Most Poisson studies are interested in setting one-sided *upper limits*. Numbers are small. (If they were large we could use the Gaussian approximation).

So if you see N you want $\sum_{r=0}^{N} e^{-\mu} \frac{\mu^r}{r!} = 0.05$ (or whatever)

If the true value is μ, or higher, the probability of only seeing N, or fewer, is only 0.05, or less.

Magic number: for $N = 0$, 95% confidence, $\mu = 3.00$. If you see zero events, you know with 95% confidence that the true value is at most 3.0

The big problem

What do you do if you see 0 events - and your expected background is 3.1?

See CL_S, Feldman-Cousins and other topics in later lectures

Bayesian credible intervals

Everything is easy.

Take prior. Multiply by likelihood, and normalise to get your posterior.

Find regions for which $\int_{\mu \epsilon region} P(\mu|data)\, d\mu = C$ where C is the desired probability content and the region my be symmetric, central, whatever

Poisson Upper limit (i.e. credible interval$[0, \mu]$) with uniform prior

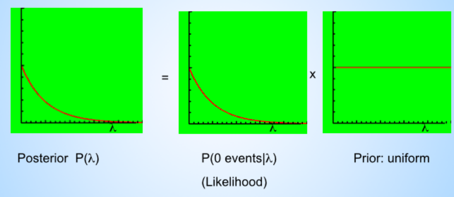
$P(\mu|r) = \frac{e^{-\mu}\mu^r/r!}{\int e^{-\mu'}\mu'^r/r! d\mu'}$

Integral can be done by parts and you end up with $\sum_{r=0}^{N} e^{-\mu}\frac{\mu^r}{r!} = 0.05$ – same as Frequentists!

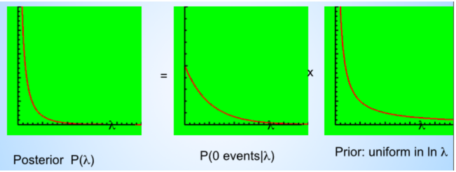
This is just a coincidence - doesn't hold for upper limits. But has saved a lot of arguments.

Sensitivity to Priors

Consider case of zero observed events



95% upper limit is 3.00



95% upper limit much smaller than 3.00

Coverage

Another use for Toy Monte Carlos

Check out that your confidence interval procedure makes sense by
1) Choose some true value μ
2) Generate lots of random pseudo-experimental results
3) Construct the confidence interval for each
4) Count how many times the interval encloses the true value
Ideally the result of (4) should be the same as you specified when you chose the constructor.

If more ('over-coverage') this is OK, thanks to the 'at least' clause, but inefficient. If less ('under-coverage') then something is wrong,

Note: coverage is a function of μ, not a single number. Typical sawtooth plots.

Bayesians do not care about coverage - but check it anyway

Tau decay puzzle

Here is just one page of PDG limits on ‘forbidden’ tau decays, all quoted at 90% Confidence

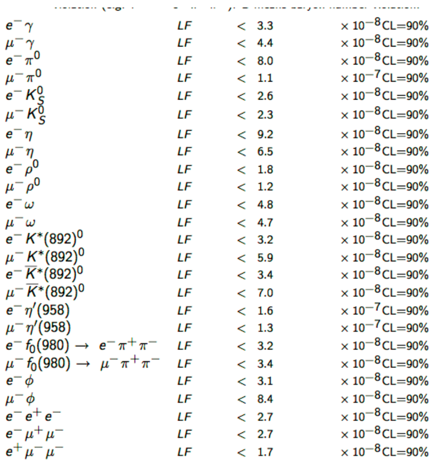
10% of these results are allowed to be wrong.

Frankly, I don’t think any of them are.

What’s going on?

A frequentist can invoke the ’at least’ clause in the definition of confidence

But what about the Bayesians?



$e^- \gamma$	*LF*	< 3.3	$\times 10^{-8}$ CL=90%
$\mu^- \gamma$	*LF*	< 4.4	$\times 10^{-8}$ CL=90%
$e^- \pi^0$	*LF*	< 8.0	$\times 10^{-8}$ CL=90%
$\mu^- \pi^0$	*LF*	< 1.1	$\times 10^{-7}$ CL=90%
$e^- K^0_S$	*LF*	< 2.6	$\times 10^{-8}$ CL=90%
$\mu^- K^0_S$	*LF*	< 2.3	$\times 10^{-8}$ CL=90%
$e^- \eta$	*LF*	< 9.2	$\times 10^{-8}$ CL=90%
$\mu^- \eta$	*LF*	< 6.5	$\times 10^{-8}$ CL=90%
$e^- \rho^0$	*LF*	< 1.8	$\times 10^{-8}$ CL=90%
$\mu^- \rho^0$	*LF*	< 1.2	$\times 10^{-8}$ CL=90%
$e^- \omega$	*LF*	< 4.8	$\times 10^{-8}$ CL=90%
$\mu^- \omega$	*LF*	< 4.7	$\times 10^{-8}$ CL=90%
$e^- K^*(892)^0$	*LF*	< 3.2	$\times 10^{-8}$ CL=90%
$\mu^- K^*(892)^0$	*LF*	< 5.9	$\times 10^{-8}$ CL=90%
$e^- \overline{K}^*(892)^0$	*LF*	< 3.4	$\times 10^{-8}$ CL=90%
$\mu^- \overline{K}^*(892)^0$	*LF*	< 7.0	$\times 10^{-8}$ CL=90%
$e^- \eta'(958)$	*LF*	< 1.6	$\times 10^{-7}$ CL=90%
$\mu^- \eta'(958)$	*LF*	< 1.3	$\times 10^{-7}$ CL=90%
$e^- f_0(980) \to e^- \pi^+ \pi^-$	*LF*	< 3.2	$\times 10^{-8}$ CL=90%
$\mu^- f_0(980) \to \mu^- \pi^+ \pi^-$	*LF*	< 3.4	$\times 10^{-8}$ CL=90%
$e^- \phi$	*LF*	< 3.1	$\times 10^{-8}$ CL=90%
$\mu^- \phi$	*LF*	< 8.4	$\times 10^{-8}$ CL=90%
$e^- e^+ e^-$	*LF*	< 2.7	$\times 10^{-8}$ CL=90%
$e^- \mu^+ \mu^-$	*LF*	< 2.7	$\times 10^{-8}$ CL=90%
$e^+ \mu^- \mu^-$	*LF*	< 1.7	$\times 10^{-8}$ CL=90%