Lecture 5:

- p-values
- Combined p-values as a Statistic
- Maximum Likelihood
- Neyman-Pearson Lemma
- Wilk's Theorem
- Extended Likelihood

A common quantity to compute when testing the null hypothesis:

p-value ("chance probability"):

The probability of obtaining a value of some parameter at least as extreme as that which is observed, assuming the null hypothesis is true.

"How much does this particular data set look like what is expected from the null hypothesis?"

But the p-value is **NOT** the probability of a particular hypothesis being true or false!

Example 1:

Search for Episodic X-Ray Emission

Over the course of a year, 36000 x-rays are observed to come from a particular astrophysical object. However, on one particular day, 130 events are observed. What is the statistical significance of this observed burst?

$$\langle x \rangle = \frac{36000}{365} = 98.6 \qquad \mu \simeq \langle x \rangle \qquad \sigma = \sqrt{\mu}$$

$$s \cong \frac{(130-98.6)}{\sqrt{98.6}} = 3.16\sigma$$

odds of getting at least this many events by a chance fluctuation from the average rate of emission

$$P = 8 \times 10^{-4}$$

p-value for this test, but need to look at it in the context of all other tests

Is this sufficient to claim the observation of a burst from this object?

Correct question:

What is the chance of seeing at least one burst with an excess at least as large given the number of independent tests I've done ?

Binomial !!

N Bernoulli trials where the chance of each success is **P**

$$\sum_{i=1}^{\infty} \binom{N}{i} P^i (1-P)^{N-i} = 1 - \binom{N}{0} P^0 (1-P)^{N-0}$$

$$\mathrm{P}_{\text{post-trial}} = 1 - (1-P)^N \qquad (\sim NP \quad \text{for} \quad NP << 1)$$

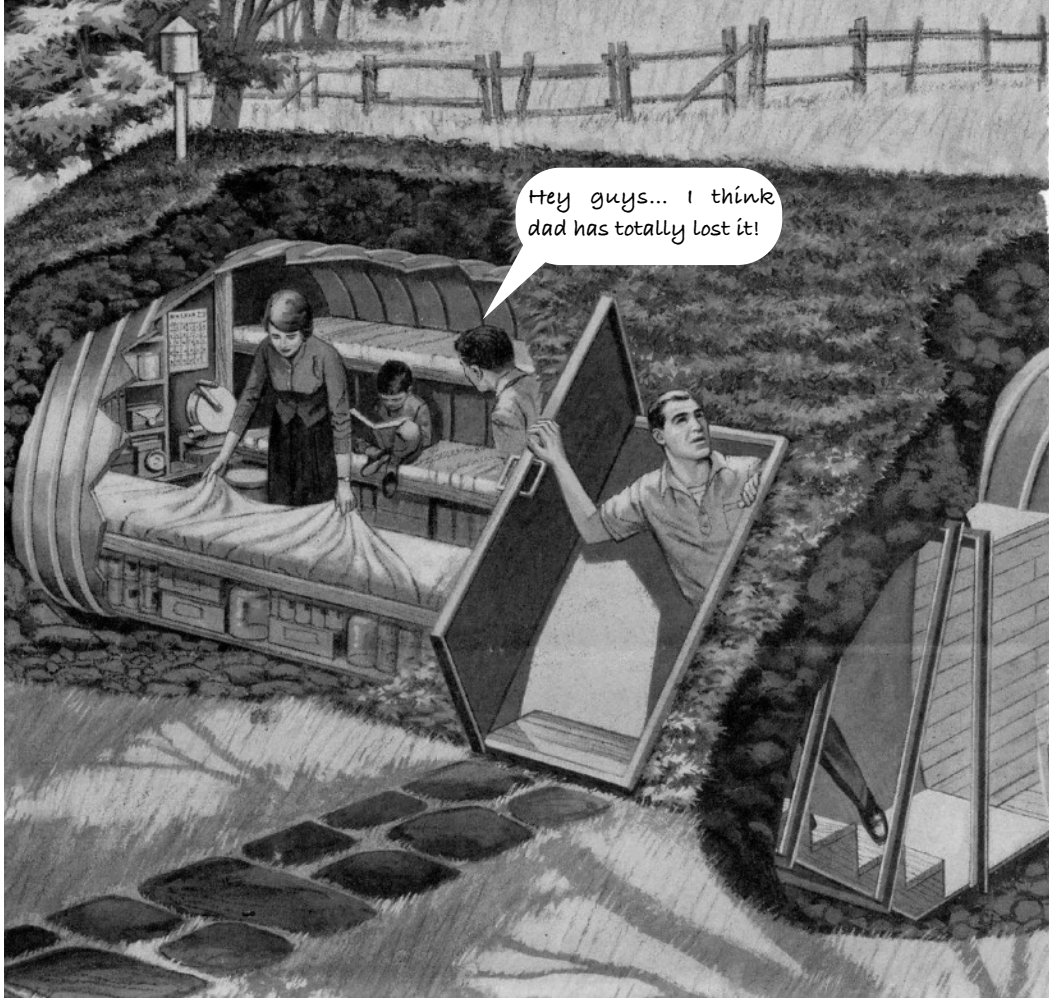
$\mathrm{P} = 8 \times 10^{-4}$, $\mathrm{N} = 365$ → $\mathrm{P}_{\text{post-trial}} = 25\%$

How many timescales were considered? How many objects examined?

Example 2:

During his year of self-isolating, Dave peered out of his bunker on six random occasions and found that it was always dark.

Assuming that the earth goes around the sun, you would expect it to be dark about half the time, averaged over the year. So the chance probability for it to be dark outside on all six occasions is:



$$P(\textit{dark all 6 times}) = (0.5)^6 = 0.0156$$

Importance of prior probabilities (more on this later)

"Gosh, That's pretty small! **Hey everyone, it looks like there's a very good chance that we're no longer going around the sun!!**"

Pragmatism!

Look carefully at context:

Very small p-values, even after careful accounting of trials, confirmed by independent observations, which could be explained by plausible alternative hypotheses…

↓

Reject H0

Combination of p-values

Two identical experiments observe evidence of the brexiton (a particle now outside of the Standard Model that inevitably then decays to a less attractive state). The first experiment assess the odds that their observation is due to chance fluctuations as being 1%, while the second assesses their observation to have a chance probability of 10%. What is the combined chance probability that these two data sets are consistent with the null hypothesis (*i.e.* there is no brexiton)?

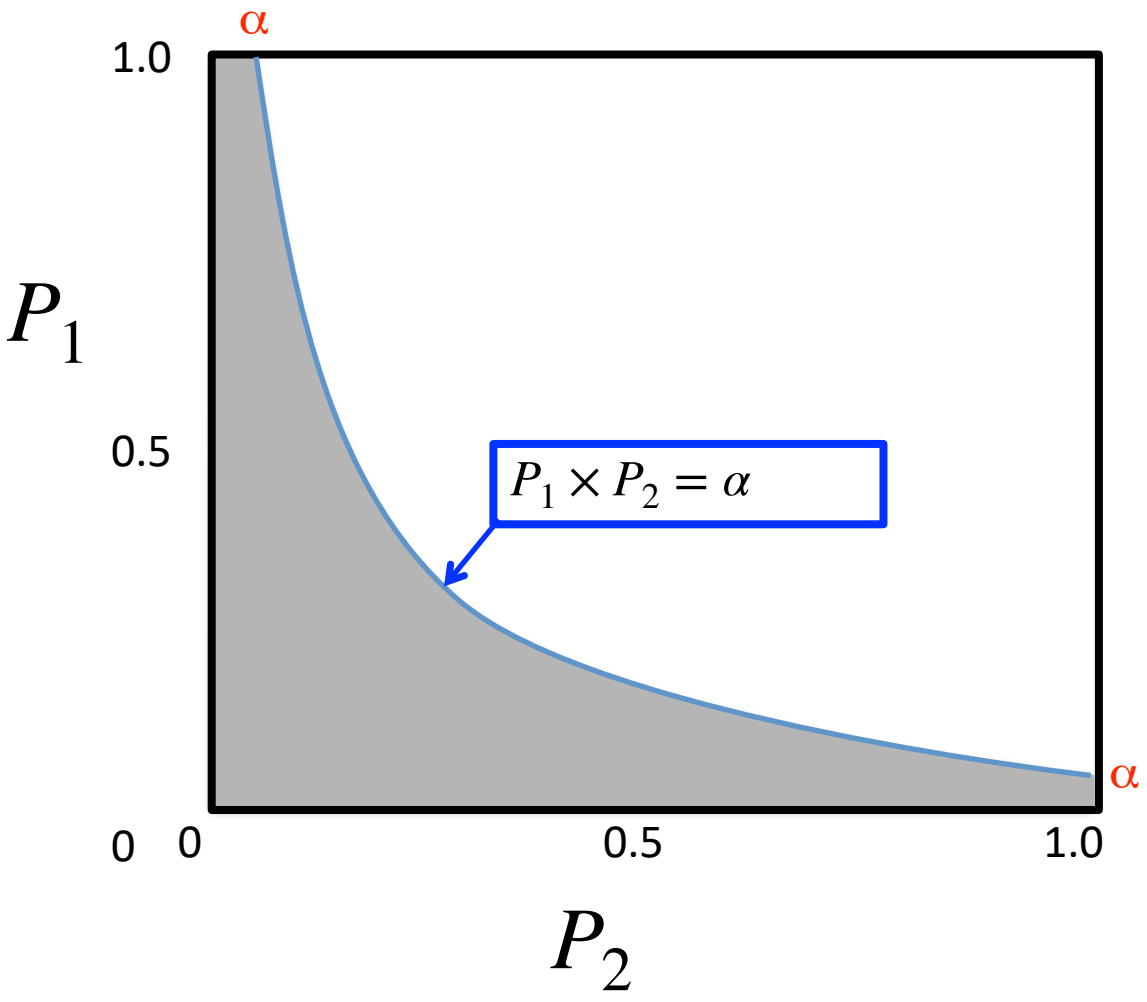
$$P_1 \times P_2 = 0.001\ ?$$

Need to look at properties of the product:

Define the statistic: $\Gamma \equiv P_1 \times P_2$

What is the chance probability for Γ

to be at least as small as some value α ?



Integrated area under the curve:

$$\alpha\,(\,1 - \ln\alpha\,)$$

$$= P(\alpha)$$

i.e. this is the chance that a background fluctuation would yield a value of Γ that is at least as small as α.

So, for the case here: $\alpha = (0.01)(0.1) = 0.001$

$$P(\leq \alpha) = 0.001(1 - \ln(0.001)) = 0.004$$

More generally…

Fisher's Method

$$F \equiv -2\ln\left(\prod_{i=1}^{n} p_i(\leq p_{obs})\right) = \sum_{i=1}^{n}\left(-2\ln p_i(\leq p_{obs})\right) \equiv \sum_{i=1}^{n} f_i$$

$$p_i(\leq p_{obs}) = e^{-\frac{f_i}{2}}$$

or $$p_i(> p_{obs}) = 1 - e^{-\frac{f_i}{2}}$$

$$p_i^{diff}(x) = \frac{1}{2}e^{-\frac{f_i}{2}}$$

Recall: $$P(\chi^2,2) = \frac{1}{2}e^{-\frac{\chi^2}{2}}$$

so f_i values are distributed like a χ^2 distribution with 2 DoF

and we can express:

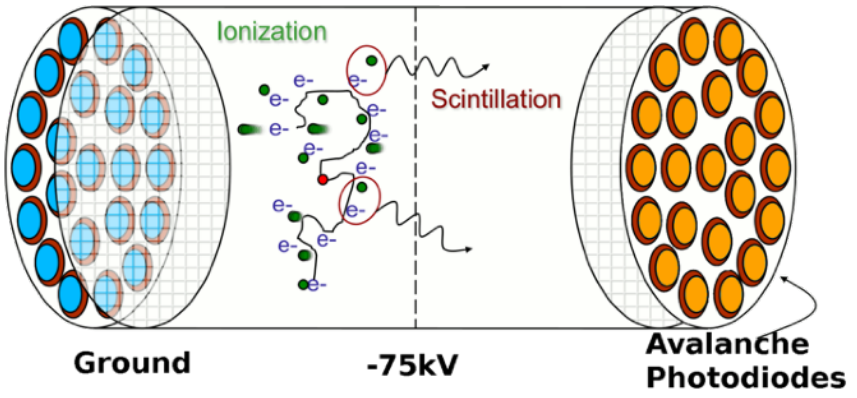
$$\chi^2 = \sum_{i=1}^{k(\equiv 2n)} g_i^2 = \sum_{i=1}^{n} \overbrace{\left(g_{2i-1}^2 + g_{2i}^2\right)}^{\chi_i^2}$$

$$P(\chi^2,2n) \equiv \sum_{i=1}^{n} P(\chi_i^2,2)$$

F is distributed like a χ^2 distribution with 2n DoF

Example:

The EXO experiment uses liquid xenon to search for evidence of neutrinoless double beta decay, which produces 2 electrons with a total energy that is well defined. The interaction produces scintillation light in the liquid xenon target, and the ionisation tracks of charged particles are also drifted to a readout plane to record the time and position of charges. Backgrounds come from radioactivity in the xenon and, to a greater extent, from the walls of the detector.



Assume that an event is observed and the chance probability for it to be background is assessed using several independent measures:

Event energy estimated from the scintillation light: $P_{scint} = 0.14$

Event energy estimated from the total charge: $P_{charge} = 0.05$

The proximity of the event to the cavity walls: $P_{charge} = 0.32$

The density of charge deposition (event topology): $P_{charge} = 0.53$

What is the overall chance probability (p-value) that this event is background?

$$-2\log(0.14 \times 0.05 \times 0.32 \times 0.53) = 13.43$$

$$P(\chi^2 > 13.43,\ DoF = 2 \times 4) = 0.10$$

Likelihood

We wish to express the likelihood for a given set of data to have resulted from a particular model of probability distributions:

conditional probability

$$L = P\left(D \,|\, H(\mathbf{q})\right)$$

likelihood — data set — assuming a particular hypothesis defined by a set of parameters $\mathbf{q}$

for independent events

$$= P(x_1|H(\mathbf{q}))P(x_2|H(\mathbf{q}))...(x_n|H(\mathbf{q})) \quad = \prod_{i=1}^{n} P(x_i|H(\mathbf{q}))$$

more practical to compute

$$\log L = \sum_{i=1}^{n} \log\left[P\left(x_i \,|\, H(\mathbf{q})\right)\right]$$

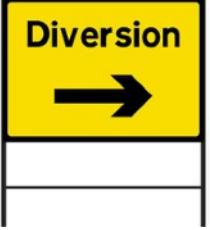
More likely data sets for $\mathrm{H}(\mathbf{q})$ will have a higher combined probability (*i.e.* likelihood)

$$\log L = \sum_{i=1}^{n} \log \left[P\left(x_i \,|\, H(\mathbf{q})\right) \right]$$

The game will then be to find the model for which the observed data set is “most likely”

Note: When used in this way, L is referred to as the “Likelihood Function” rather than a probability, because it is used to describe the *relative* probability for different models given a fixed data set… however that dependence need not be normalised to 1 over the models tested!

(the normalisation is instead defined over all possible data sets for a fixed model)



Tests of Simple vs Composite Hypotheses

Simple hypothesis: All parameters of the relevant distributions are specified. (*i.e.* PDFs can be used to completely characterise the problem)

Composite hypothesis: Where this is not the case and parameters span a range of possibilities.

This is probably a university student, because they spend £20 per week on alcohol and the average student spends more than £15 per week on this.

COMPOSITE

(exact distribution not defined)

This is probably a university student, because they spend £20 per week on alcohol and the average student spends £17 per week on this with a standard deviation of ~ £13.

COMPOSITE

(distribution of alternative not defined)

This is probably a university student, because they spend £20 per week on alcohol and the average student spends £17 per week on this with a standard deviation of ~ £13, whereas this is normally what is expected for the typical UK household with an average of 1.9 adults.

SIMPLE

Statistical Power

When comparing 2 hypotheses, H0 and H1, the "statistical power" is the fraction of times that H0 is correctly rejected when H1 is true if one were to repeat the test many times with "identical" ensembles of data subject only to statistical fluctuations

"Frequentist"

That's ridiculous… I only care whether I'VE made the right choice given THIS set of data!

Bayesian Power

When comparing 2 hypotheses, H0 and H1, the "Bayesian power" is the confidence you have in correctly rejecting H0 given the assumed probability distributions of H0 and H1 for this particular set of data

That's ridiculous… hypotheses don't have probability distributions: they are true or false!

Neyman-Pearson Lemma:

$$\Lambda \equiv \frac{L(D \mid H_0)}{L(D \mid H_1)}$$

(sometimes defined as one over this)

is

"Uniformly Most Powerful" (in a frequentist sense)
discriminate between simple hypotheses

(The exact distribution of Λ will, in general, depend on the distributions of L)

Assume that the set of possible hypotheses that describe a particular data set are distinguished only by the values of some unknown set of model parameters (*e.g.* the number of signal events, or the slope and intercept of a line, *etc.*).

Determining the best set of model parameters by comparing to find the **Maximum Likelihood** is therefore the UMP method of parameter estimation!

Simple example: You wait at a bus stop and no bus arrives for the first 10 minutes, but then 2 buses arrive in the next 10 minute interval. What is the best estimate of the mean number of buses per 10 minutes?

assume Poisson process $P(n\,|\,\mu) = \dfrac{\mu^n e^{-\mu}}{n!}$

$$L = P(0\,|\,\mu)P(2\,|\,\mu) = (e^{-\mu})(\tfrac{1}{2}\mu^2 e^{-\mu}) = \tfrac{1}{2}\mu^2 e^{-2\mu}$$

maximise the likelihood:

$$\frac{\partial L}{\partial \mu} = \mu e^{-2\mu} - \mu^2 e^{-2\mu} = 0$$

$$\rightarrow \mu_m = 1$$

(as expected)

Consider the case where uncertainties on data points are normally distributed. Assume that the mean values and variances, μ_i and σ_i, are predicted at each data point by some model. Then we have:

$$\log L = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right) \right]$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma_i} - \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}$$

$$-2 \log L = -2 \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma_i} + \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

looks like χ^2

constant

Thus, maximising L = maximising $\log L$ = minimising $-2\log L$ is equivalent to the **Method of Least Squares** in this limit !!

Can we approximate the general shape of likelihood functions?

Consider a single parameter, q, which maximises the likelihood at $q=q_m$.
Now Taylor expand around the maximum likelihood point:

$$\ln L(q) = \ln L(q_m) + \left[\frac{\partial \ln L}{\partial q}\right]_{q=q_m} (q - q_m) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial q^2}\right]_{q=q_m} (q - q_m)^2 + \ldots$$

zero by definition

can be shown to be approximately $-\frac{1}{\sigma^2_{q_m}}$ as $n \to \infty$

$$\ln L(q) \sim \ln L(q_m) - \frac{1}{2}\frac{(q - q_m)^2}{\sigma^2_{q_m}}$$

$$q \to q_m \pm \sigma_{q_m}$$

$$\ln L(q_m \pm \sigma_{q_m}) \sim \ln L(q_m) - \frac{1}{2} \quad \text{or} \quad -2[\ln L(q_m \pm \sigma_{q_m}) - \ln L(q_m)] \sim 1$$

looks like $\Delta\chi^2$

Wilks' Theorem

more generally:

$$-2[\ln L(\mathbf{q_o}) - \ln L(\mathbf{q})] = -\mathbf{2}\ln\left(\frac{L(\mathbf{q_o})}{L(\mathbf{q})}\right) \equiv -\mathbf{2}\ln L_{\mathbf{R}} \sim \chi^2_d$$

where $\mathbf{q_o}$ are the set of model parameters that define the default (null) hypothesis, and the d = DoF = the difference in the number of model parameters constrained

(*i.e.* how many extra degrees of freedom one model has compared to the other)

Legal Statement:

- *For nested hypotheses (i.e. a continuous transition from one hypothesis to the next)*
- *Away from boundaries in likelihood space*
- *In the limit of large amounts of data*

However, for example, in the case of Poisson distributions, this actually works pretty well even for small numbers of events and also near μ=0. But generally need to check. Can do this, for example, by generating simulated data sets under a given hypothesis to directly look at the distribution of likelihood estimates.

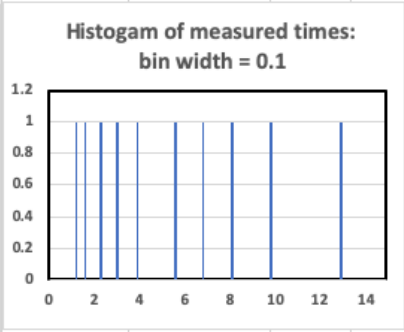
Example:

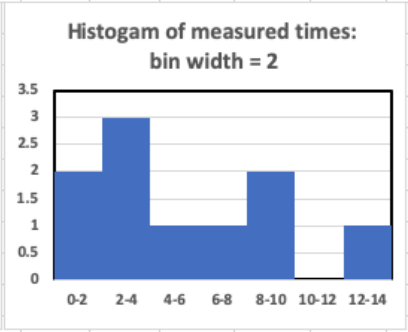
A newly commissioned underground neutrino detector sees a rate of internal radioactive contamination decreasing as a function of time. 10 events are observed over a period of 15 consecutive days. Determine the best fit mean decay time in order to determine the source of the contamination.

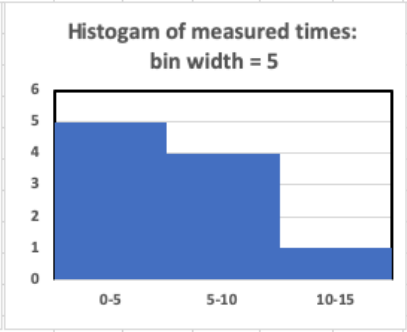
decay probability:

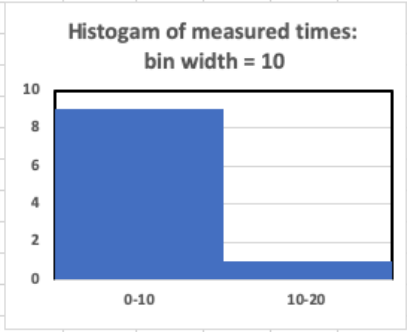
$$P(t) = \frac{1}{t_o} e^{-\frac{t}{t_o}}$$

t_o = mean decay lifetime

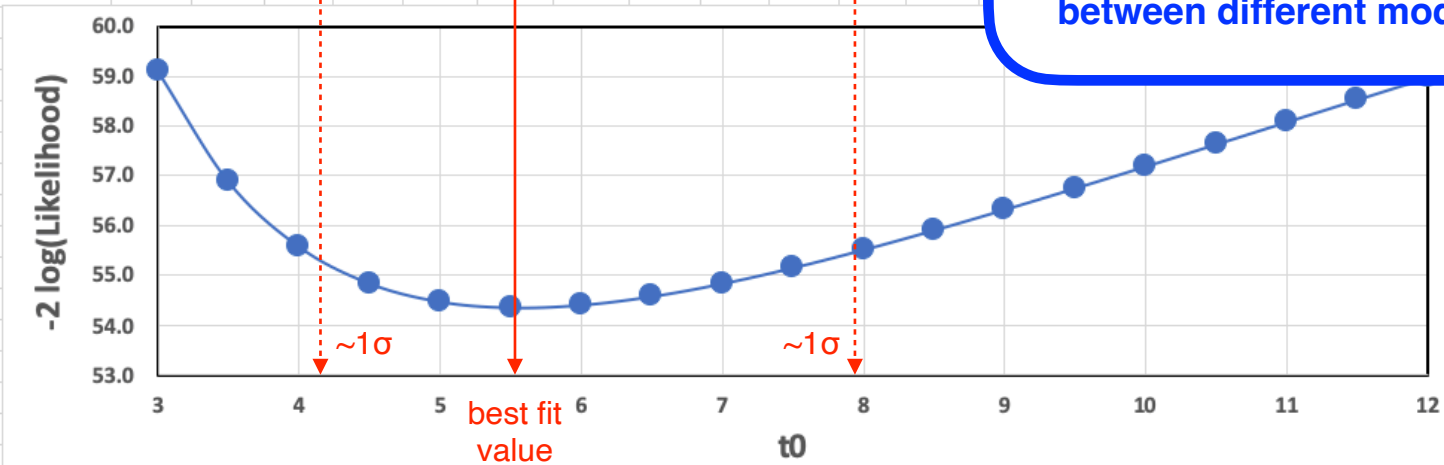




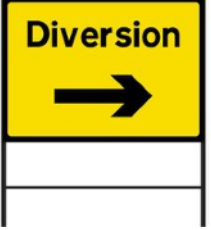




Measured Time (days)	t0:	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12
		Table of Probabilities: P(t) = (1/t0)*exp(-t/t0) for different assumed values of t0																		
5.6		0.0515	0.058	0.062	0.064	0.065	0.066	0.066	0.065	0.064	0.063	0.062	0.061	0.06	0.058	0.057	0.056	0.055	0.053	0.052
1.3		0.2161	0.197	0.181	0.166	0.154	0.144	0.134	0.126	0.119	0.112	0.106	0.101	0.096	0.092	0.088	0.084	0.081	0.078	0.075
2.4		0.1498	0.144	0.137	0.13	0.124	0.118	0.112	0.106	0.101	0.097	0.093	0.089	0.085	0.082	0.079	0.076	0.073	0.071	0.068
12.9		0.0045	0.007	0.01	0.013	0.015	0.017	0.019	0.021	0.023	0.024	0.025	0.026	0.027	0.027	0.028	0.028	0.028	0.028	0.028
6.8		0.0346	0.041	0.046	0.049	0.051	0.053	0.054	0.054	0.054	0.054	0.053	0.053	0.052	0.051	0.051	0.05	0.049	0.048	0.047
1.7		0.1891	0.176	0.163	0.152	0.142	0.133	0.126	0.118	0.112	0.106	0.101	0.096	0.092	0.088	0.084	0.081	0.078	0.075	0.072
9.8		0.0127	0.017	0.022	0.025	0.028	0.031	0.033	0.034	0.035	0.036	0.037	0.037	0.037	0.038	0.038	0.037	0.037	0.037	0.037
4		0.0879	0.091	0.092	0.091	0.09	0.088	0.086	0.083	0.081	0.078	0.076	0.073	0.071	0.069	0.067	0.065	0.063	0.061	0.06
8.1		0.0224	0.028	0.033	0.037	0.04	0.042	0.043	0.044	0.045	0.045	0.045	0.045	0.045	0.045	0.044	0.044	0.044	0.043	0.042
3.1		0.1186	0.118	0.115	0.112	0.108	0.103	0.099	0.095	0.092	0.088	0.085	0.082	0.079	0.076	0.073	0.071	0.069	0.066	0.064
Product of probabilities:		1E-13	4E-13	9E-13	1E-12	1E-12	2E-12	2E-12	1E-12	1E-12	1E-12	9E-13	7E-13							
Sum of $\log_e$(probabilities):		-29.55	-28.4	-27.8	-27.4	-27.23	-27.17	-27.2	-27.3	-27.4	-27.6	-27.8	-28							
-2 x Sum of the logs :		59.106	56.88	55.58	54.84	54.47	54.35	54.4	54.57	54.83	55.15	55.51	55.91							



No absolute goodness-of-fit, just the "relative goodness" between different models

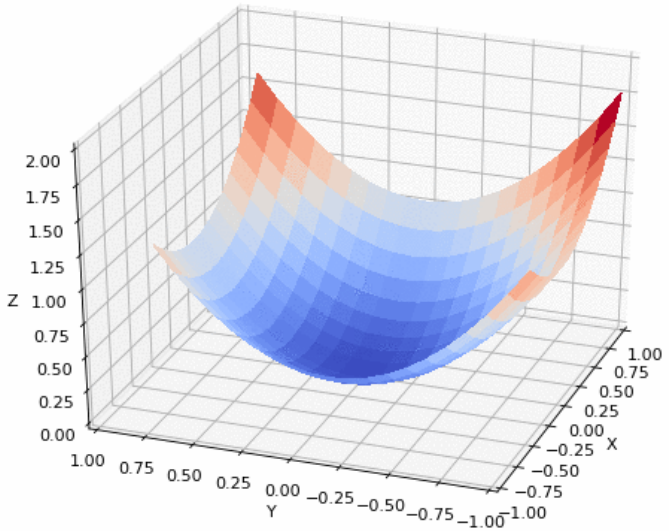


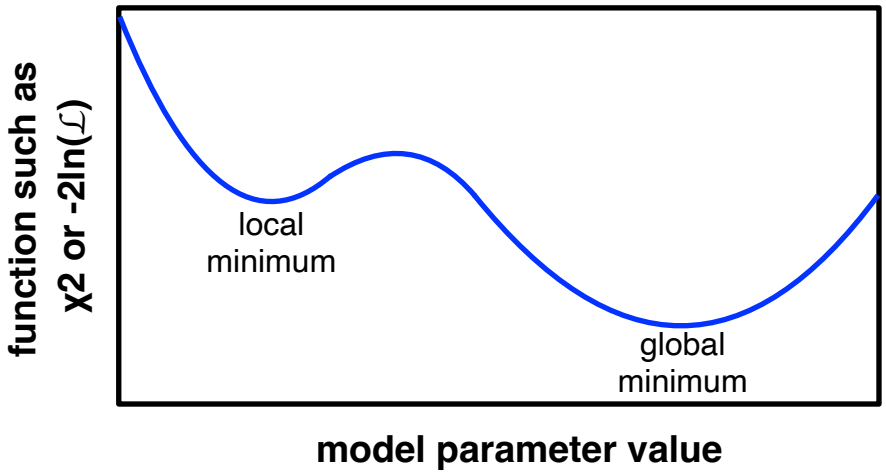
Numerical Optimisation (Minimisation/Maximisation)

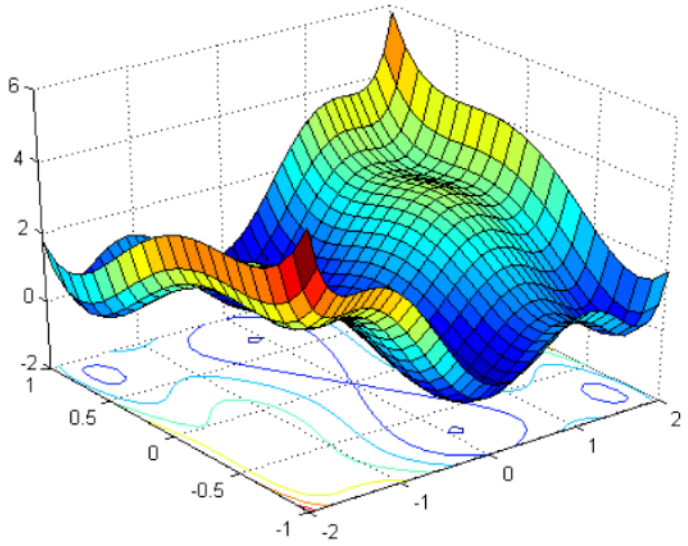
Simplest - “Grid Search”: Systematically step through possible parameter values on an n-dimensional grid of some pre-defined resolution to find the best values.

Pros: Simple and robust
Cons: Inefficient

Other approaches usually require an initial guess for the parameter values (or “seed”) and then progress though parameter space in a direction and with a variable step size based on how successive function evaluations change. These typically make use numerical function derivatives to follow a gradient decent path. There is generally some convergence criteria to specify when sufficient accuracy has been achieved and/or when the function evaluations no longer seems to be changing very much (*i.e.* second derivatives are close to zero).







Depending on the nature of the problem, the function space can be irregular and may contain local minima, particularly when dealing with multiple dimensions and parameters have correlations or degeneracies (*i.e.* where different parameter combinations can produce similar solutions). Discontinuities such as "hard" physical boundaries can cause particular problems, as can binned PDFs created with limited statistics.

Numerous algorithms exist to sample parameter space, bounce out of local minima, smooth out irregularities, deal with hard boundaries, etc. These may makes use of parallel processing, machine learning, Markov chains, simulated annealing… **THIS IS A VAST AREA!**

Always important to look at your parameter space

Joint Analysis of Multiple Data Sets

$$-2\log L = \sum_{i=1}^{n} -2\log\left[P\left(x_i \mid H(\mathbf{q})\right)\right]$$

$$= \sum_{i=1}^{k} -2\log\left[P\left(x_i \mid H(\mathbf{q})\right)\right] + \sum_{i=k+1}^{n} -2\log\left[P\left(x_i \mid H(\mathbf{q})\right)\right]$$

Likelihood for one set of data under $\mathrm{H}(\mathbf{q})$.

Likelihood for the same hypothesis, but a different set of data. Could even be from a different experiment and assessed in a completely different way, so long as it is eventually turned into a probability.

Can jointly analyse multiple data sets from multiple experiments to determine the best overall parameter estimations by adding together their likelihoods over the same parameter space

It's always good to show your likelihood space as part of the presentation of results both as an overall summary of the relevant information content of your data and to allow for such joint analyses

"Extended" Likelihood

The number of events, n, in a data set is often the result of Poisson fluctuations about the expected mean number of events. If the expected mean is itself a parameter of interest (*e.g.* the "true" flux of signal and/or background events), the associated Poisson fluctuation can then be included in the likelihood as follows:

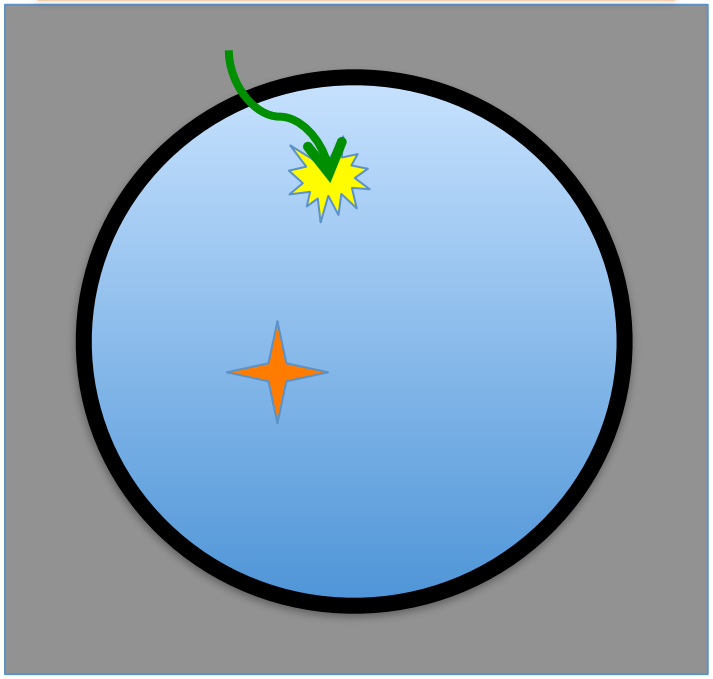
$$L = \left(\frac{\mu^n e^{-\mu}}{n!} \right) \prod_{i=1}^{n} P(x_i \,|\, H(\mathbf{q}))$$

$$\log L = n \log \mu - \mu - log(n!) + \sum_{i=1}^{n} \log \left[P\left(x_i \,|\, H(\mathbf{q})\right) \right]$$

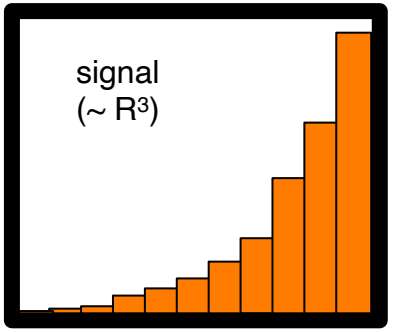
Can ignore this term, since this is a constant and we're only concerned with derivatives and differences of the likelihood

Example of a 2-component model of signal and background:

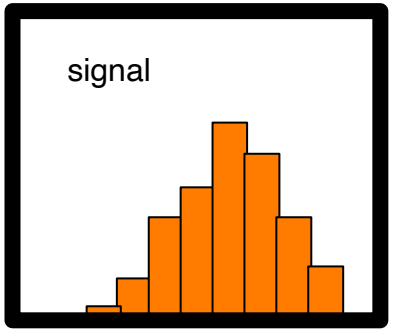
An Experiment Searching for Rare Interactions



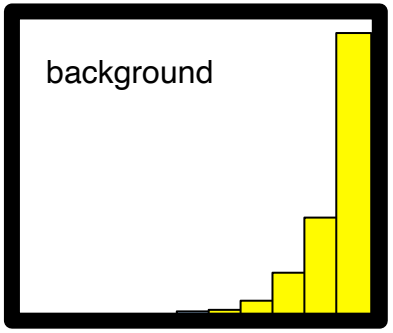
Simulation and/or Calibration Data



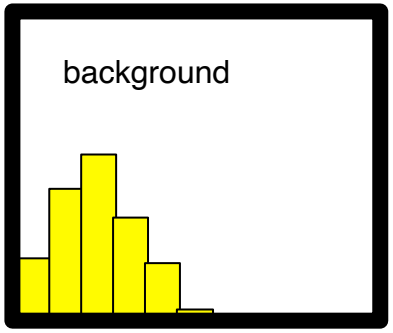
"Radius"



"Energy"



"Radius"



"Energy"

Reconstructed energy and position could be correlated (e.g. higher energy events could be easier to reconstruct accurately). So, form 2-D histograms to preserve these correlations and normalise these to one to produce PDFs for each type of event class:

$$P(\tilde{E}_i, \tilde{R}_i | S)$$

$$P(\tilde{E}_i, \tilde{R}_i | B)$$

Consider a hypothesis, H, in which a certain fraction of the data is signal and remaining fraction is background:

$$P(\tilde{E}_i, \tilde{R}_i \,|\, H) = P(\tilde{E}_i, \tilde{R}_i \,|\, S)\left(\frac{\mu_S}{\mu_S + \mu_B}\right) + P(\tilde{E}_i, \tilde{R}_i \,|\, B)\left(\frac{\mu_B}{\mu_S + \mu_B}\right)$$

where $\mu_{total} = \mu_S + \mu_B$

extended likelihood part

$$\log L = n \log(\mu_S + \mu_B) - (\mu_S + \mu_B)$$

$$+ \sum_{i=1}^{n} \log \left[P(\tilde{E}_i, \tilde{R}_i \,|\, S)\left(\frac{\mu_S}{\mu_S + \mu_B}\right) + P(\tilde{E}_i, \tilde{R}_i \,|\, B)\left(\frac{\mu_B}{\mu_S + \mu_B}\right) \right]$$

→ Maximise $log\ L$ (or minimise $-2log\ L$) over μ_S and μ_B in addition to any other parameters of the model

We're particularly interested in the value and significance of the signal, so look at the projection where the likelihood is maximised over all other free model parameters as μ_S is varied:

"nuisance parameters"

"Profile Likelihood"