Lecture 7 – inference in various forms

- Parameter estimation:
  - Bayesian approach.
  - Maximum likelihood.
  - χ^2 & assumptions.
  - Application to fitting *y*=*mx*+*c*; weighted data.

- Goodness of fit & hypothesis testing:
  - χ^2 & its applications.

- Non-parametric statistics.

How should we estimate parameters from our data?

- Will use the example of fitting a straight line to our data:
  - We would like to estimate *m* (slope) and *c* (intercept).
- This is one example of the specific case of having data values y_1, y_2, y_3…, and trying to estimate the value of some parameter vector $\underline{a}$ that describes some model:
  - $\underline{a}$ is a two parameter vector, with components *m* & *c* here.
- Note that what we are actually doing here is: making measurements of y, taking it as a given that $\underline{a}$ has some true value.
- If each datum is assumed independent, then we can write:
  - Probability of getting our dataset = $p(y_1|\underline{a})\ dy_1 \times p(y_2|\underline{a})\ dy_2 \times \ldots\ p(y_N|\underline{a})\ dy_N$ $=\prod_i p(y_i|\underline{a})\ dy_i \equiv L(y_1\ldots y_N|\underline{a})$ = "Likelihood".

The “maximum likelihood” approach

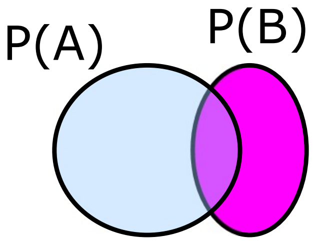
- One approach – the “maximum likelihood” method – for estimating $\underline{a}$ says: “find the value of $\underline{a}$ that maximizes $L(y_1...y_N|\underline{a})$:
  - This sounds plausible and is often done.
  - BUT there is a catch.

- $L(y_1...y_N|\underline{a})$ measures $p(\text{data}|\underline{a})$, and what we really want to maximize is $p(\underline{a}|\text{data})$:
  - This is a subtle but important difference.

- We can get to what we want by using Bayes’ Theorem:
  - This relates $p(A|B)$ to $p(B|A)$ where A and B correspond, in our example, to:
    - A = the parameter vector we are interested in, i.e. $\underline{a}$.
    - B = the data that we have measured.

Bayes’ theorem

- Consider the Venn diagram associated with the two independent “propositions” A and B:



  - Intersection = p(A and B) = p(A|B)p(B) = p(B|A)p(A)
  - Rearranging we get p(A|B) = p(B|A) × p(A) ÷ p(B)

$$\text{So,} \Rightarrow p(\underline{a}\,|\,Data) = p(Data\,|\,\underline{a}) \times \frac{p(\underline{a})}{p(Data)}.$$

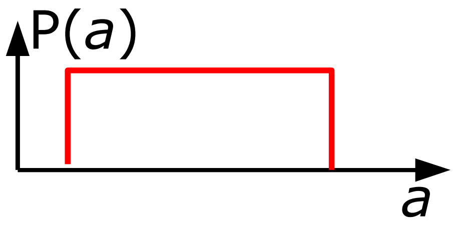
- Now, p(Data) is just a normalization, so we can ignore it here, but p(*a*) matters.
  - This is called the “prior” probability for *a*. It incorporates our knowledge of *a*, prior to performing the experiment.
  - P(a|data) is called the “posterior” probability for *a*. It captures our knowledge of *a*, after performing the experiment.

We will consider three cases of prior knowledge today

$$p(\underline{a} \mid data) = p(data \mid \underline{a}) \cdot \frac{p(\underline{a})}{p(data)}.$$

1. We have no idea what the magnitude of our parameter *a* is:
   - E.g. 40 years ago, no one knew what the magnitude of the anisotropy of the 2.7K microwave background was: ΔT = mK, μK, nK?
   - If you have no idea of scale, then a sensible choice of prior is one that is uniform in log space: p(*a*)da $\propto$ 1/a da.

2. We assume that within some range, all values of *a* are equally likely:
   - □ This is a common situation in experiments.
   - □ In this case p(*a*) is uniform, and so p(*a*|*data*) $\propto$ p(*data*|*a*) and we are actually just maximizing the likelihood.
   - □ We will assume this case hereafter – but BEWARE.



The third case – where the prior is informative – should act as a warning

3. Consider this situation you might face if doing jury service:
   - A car owner says the person stealing his car last night was definitely a man wearing a pink shirt.
   - The owner is known to have excellent vision, so much so that at night he identifies pink shirts correctly 98% of the time.
   - The police round up all known male car thieves immediately the theft is reported. One is wearing a pink shirt – and is arrested.

You have to assess whether the police have arrested the right person – if so, they will be convicted.

□ The naïve response would be "sure" – the owner is so good at recognizing colors that the police must have the correct person in the cells.

□ This is incorrect: what matters here is how many male car thieves wear pink, i.e. our prior for this.

We can assess this in a Bayesian way, once we know the relevant priors

- Let H be the hypothesis that the shirt seen by the witness was indeed pink, and $\overline{H}$ be its complement (i.e. the shirt was in fact not pink & the witness mis-identified it).

- Let D be the data, i.e. the information the witness provides – "I'm absolutely sure the thief was wearing a pink shirt"
- What we are interested in is p(H|D), and we know the following:

$$p(H|D) + p(\overline{H}|D) = 1$$
$$p(D|H) = 0.98.\ \text{This is the likelihood for H.}$$
$$p(D|\overline{H}) = 0.05.$$

NB: the final line implies that the witness will misidentify a shirt as pink (when it isn't) 5% of the time at night.

$$\text{Bayes says: } p(H \mid D) = \frac{p(D \mid H)\,p(H)}{p(D)} \quad \& \quad p(\overline{H} \mid D) = \frac{p(D \mid \overline{H})\,p(\overline{H})}{p(D)}.$$

- Now assume that only 1% of male car thieves wear pink shirts.

The story of the pink shirt continued

- That means:

$$p(H) = 0.01 \;\; \& \;\; p(\bar{H}) = 0.99 \quad \text{These are our priors}$$

$$\text{So } p(H|D) = 0.98 \times 0.01/p(D) \;\; \& \; p(\bar{H}|D) = 0.05 \times 0.99/p(D)$$

$$\Rightarrow p(H|D) = \frac{0.0098}{0.0098+0.0495} = 0.165$$

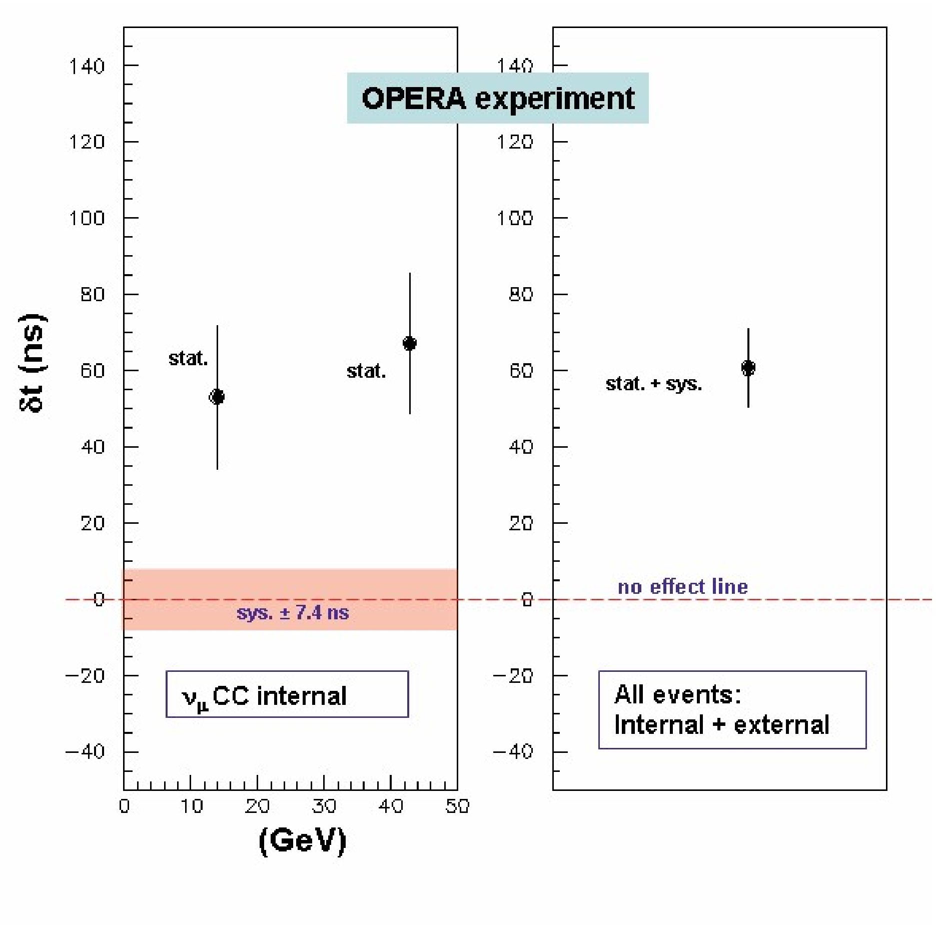
So, despite the “excellent” witness – the “posterior” probability that the shirt worn by the car thief was pink is only 17%

- What this means is that it’s much more likely that the witness made a mistake and the thief was not wearing a pink shirt (i.e. the police have detained the wrong car thief).
- The key point is that if you have strong a priori evidence for some hypothesis then you need very strong data to alter that position.

Faster than light neutrinos in 2011?

- This is why no one was surprised when a systematic error was uncovered in the OPERA results.

- 731km.   2.4… ms travel time at light speed c. Looking for neutrino mass which would reduce speed from c. But instead measured faster….
- Tricky experiment needs exact distance and timing.



Let’s continue looking at our example of straight-line fitting (flat prior)

$$L(y_1, y_2, y_3, \ldots y_n \mid a) = \prod_1^n p(y_i \mid a).$$

- For each data value, y_i, there will be:
  - An error σ_i;
  - A value x_i of a controllable quantity with no error (if not, there are methods to cope with errors (not discussed here));
  - The model or theoretical value $f(x_i \mid \underline{a})$.
- We will assume the errors are Gaussian:
  - May be so intrinsically;
  - May be many causes, in which case CLT $\Rightarrow$ Gaussian;
  - May be wrong.

- Under these conditions for the ith measurement:

$$p(y_i \mid \underline{a}) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-[y_i - f(x_i|\underline{a})]^2 / 2\sigma_i^2}$$

  - So the likelihood, $L(y_i \ldots \mid \underline{a})$ will be a product of such Gaussians and we want to find the value of $\underline{a}$ that maximizes L.

In this case, it is easier to maximize ln(L):

$$\ln(L) = -\frac{1}{2}\sum_i \left[\frac{y_i - f(x_i\,|\,\underline{a})}{\sigma_i}\right]^2 - \sum \ln(\sigma\sqrt{2\pi}).$$

- So to maximize L, we need to minimize:

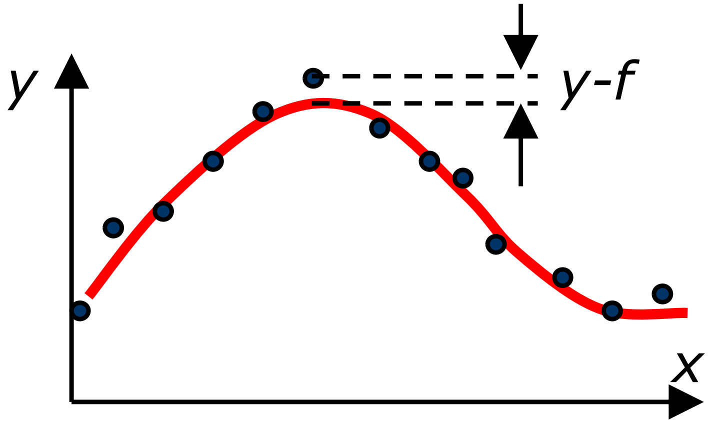
$$\boxed{\sum_i \left[\frac{y_i - f(x_i\,|\,\underline{a})}{\sigma_i}\right]^2 \equiv \chi^2 \qquad \text{"chi-squared".}}$$

- i.e. we have an equation $\partial\chi^2 / \partial a_j = 0$ for each of the parameters, and we solve these to get the best estimates of the parameters.

- Let’s examine χ^2: It’s the sum of:
  - Squares of the deviations, $\Delta = y_{\text{actual}} - y_{\text{model}}$. This is sensible:
    - Sign of deviation is unimportant.
    - Big deviations matter much more.
  - Weighted inversely as the error^2 in y_{actual}, so that Δ^2 is compared with the expected average deviation^2, σ_i^2.

Chi-squared cont$^{d.}$

- So χ^2 seems like a sensible thing to minimize, and it is, provided:
  - The prior is uniform;
  - The errors are only in y_i;
  - The y_i's are Gaussian distributed about their model values.



- Now, let's apply χ^2 minimization to straight line fitting – $y=mx+c$:
  - First, assume the σ_i are the same for all i. Then:

$$\chi^2 = \sum_i \left[\frac{y_i - (mx_i + c)}{\sigma_i} \right]^2 \equiv \frac{1}{\sigma^2} \sum_i \left(y_i - (mx_i + c) \right)^2 .$$

$$\partial\chi^2 \big/ \partial c \quad \rightarrow \quad \sum_i -2\left(y_i - (mx_i + c) \right) = 0.$$

Chi-squared straight-line fitting cont$^{\text{d.}}$

$$\boxed{\chi^2 = \frac{1}{\sigma^2}\sum_i \left(y_i - m x_i - c\right)^2 .}$$

best estimates

$$\bar{y} - \overset{\downarrow}{\hat{m}}\bar{x} - \overset{\downarrow}{\hat{c}} = 0. \ (1)$$

- If we divide $\partial\chi^2 / \partial c = 0$ by N (# of data) we get:
- Likewise: $\dfrac{\partial\chi^2}{\partial m} = 0 \rightarrow \overline{xy} - \hat{m}\,\overline{x^2} - \hat{c}\,\overline{x} = 0. \ (2)$

□ We can use (1) and (2) to show that:

$$\boxed{\hat{m} = \frac{\overline{(xy)} - (\bar{x})(\bar{y})}{\overline{(xx)} - (\bar{x})(\bar{x})} \equiv \frac{\text{Covariance (x,y)}}{\text{Variance (x)}} \quad (3)}$$

□ And that

$$\boxed{\hat{c} = \frac{\overline{(xx)}\,\overline{(y)} - \overline{(x)}\,\overline{(xy)}}{\overline{(xx)} - (\bar{x})(\bar{x})} \equiv \bar{y} - \hat{m}\bar{x} \ (4)}$$

□ In this case, the best fit line goes through the center of gravity of the points.

Error estimates in Chi-squared straight-line fitting

- If we denote the error on $\hat{m}$ as σ_m and use the error quadrature formula, we find

$$\sigma_m^2 = \sum_i \left(\frac{\partial \hat{m}}{\partial y_i} \right)^2 \hat{\sigma}^2,$$

Here $\hat{\sigma}^2$ quantifies the deviation of the actual data from the best-fit model and is given by:

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_i \left(y_i - (\hat{m} x_i + \hat{c}) \right)^2.$$

- This and the expression for $\hat{m}$ given in equation 3 imply:

$$\boxed{\sigma_m^2 = \frac{\hat{\sigma}^2}{N \times [\overline{(xx)} - (\bar{x})(\bar{x})]}.}$$

- Similarly, we can show that the error on $\hat{c}$ is given by:

$$\boxed{\sigma_c^2 = \frac{\hat{\sigma}^2 \overline{(xx)}}{N \times [\overline{(xx)} - (\bar{x})(\bar{x})]}.}$$

What happens if the errors, σ_i, are not all equal? "Weighted means"

- The sums to be minimized are:

$$\chi^2 = \sum_i \left[\frac{y_i - mx_i - c}{\sigma_i} \right]^2$$

- This implies the same equations for $\hat{m},\ \hat{c}$ as before BUT $\bar{x},\ \bar{y}$ etc are not simple averages but ones that are:
  - Weighted by $1/\sigma_i^2$;
  - Normalized by $\Sigma\ 1/\sigma_i^2$ not by N.

□ So, e.g. $\bar{y} \neq \frac{\sum_i y_i}{N}$ but becomes $= \frac{\sum_i y_i / \sigma_i^2}{\sum_i 1/\sigma_i^2}$ with variance $\frac{1}{\sum_i 1/\sigma_i^2}$,

and importantly, $\hat{\sigma}^2 \Rightarrow \mathbf{1}$.

What about hypothesis testing though?

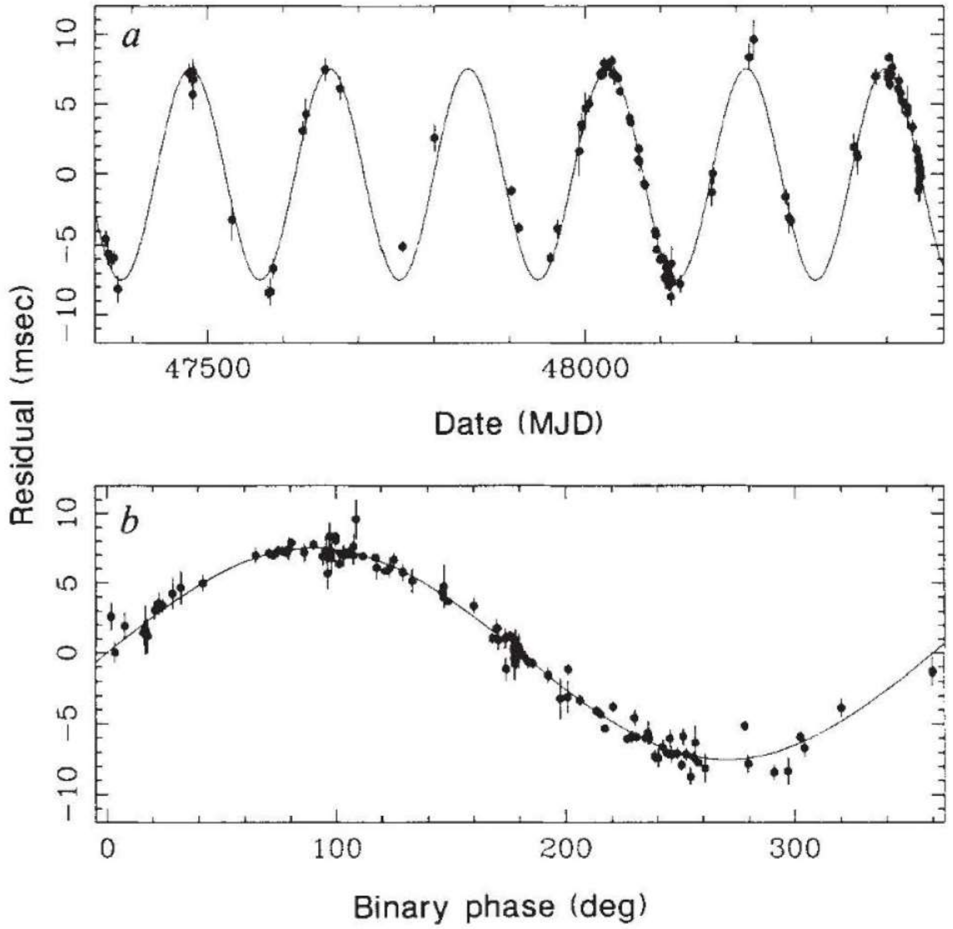
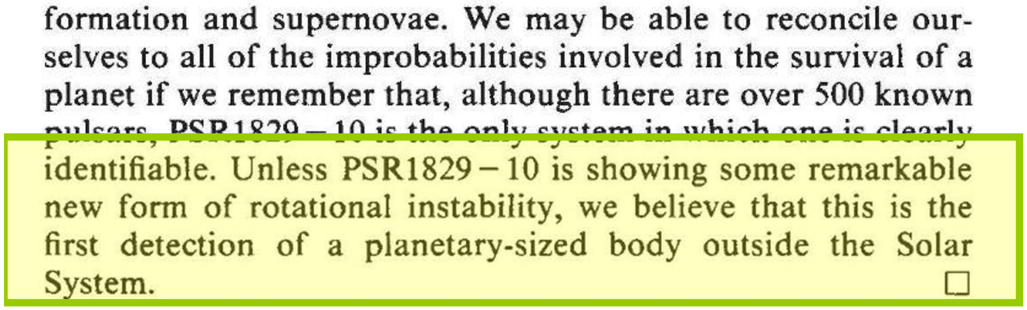


FIG. 1 The timing residuals for PSR1829 – 10. *a*, Plotted relative to a simple model of the slow down. The smooth curve represents the solution for a binary system with the parameters given in Table 1. *b*, As *a*, but plotted against orbital phase.

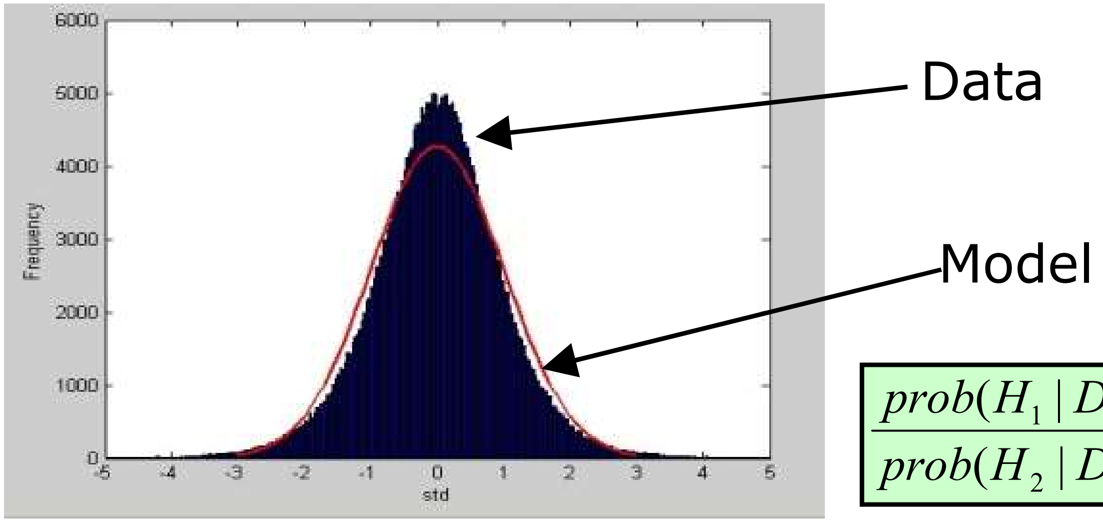
Is there a planet?

Bailes, Lyne and Shemar, Nature 352, 311 - 313 (25 July 1991)



formation and supernovae. We may be able to reconcile ourselves to all of the improbabilities involved in the survival of a planet if we remember that, although there are over 500 known pulsars, PSR1829 – 10 is the only system in which one is clearly identifiable. Unless PSR1829 – 10 is showing some remarkable new form of rotational instability, we believe that this is the first detection of a planetary-sized body outside the Solar System. □

How do we “test” hypotheses?



Do these data make my preferred theory H_1 more probable than its competing theory H_2?

$$\frac{prob(H_1 \mid Data, I)}{prob(H_2 \mid Data, I)} = \frac{prob(Data \mid H_1, I)}{prob(Data \mid H_2, I)} \times \frac{prob(H_1 \mid I)}{prob(H_2 \mid I)}$$

- However, at this level, usually one asks a different and somewhat easier question: what is the significance of the mismatch between H_1 and the data?

□ To establish this we evaluate χ^2 again, the point being that we know something about its expected distribution.

Data values Model values

$$\chi^2 = \sum_i \left[\frac{y_i - f(x_i)}{\sigma_i} \right]^2$$

(NB Formally, the χ^2 test is for binned data where the model predicts the population of each bin.)

The chi-squared distribution

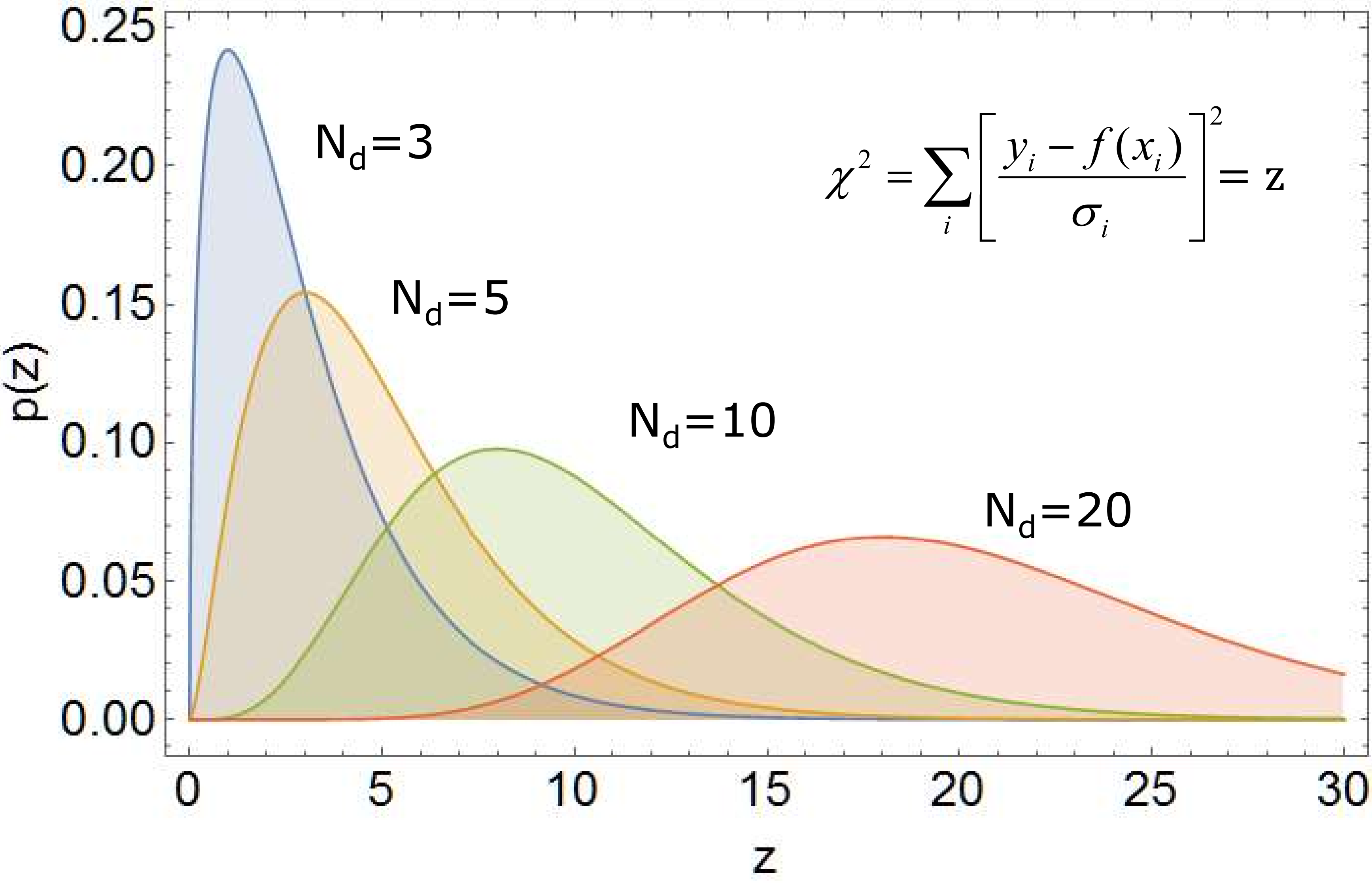
$$\chi^2 = \sum_i \left[\frac{y_i - f(x_i)}{\sigma_i}\right]^2$$

- If *f* really does model the observed data, then $|y_i\text{-}f(x_i)|$ should on average equal σ_i, i.e. χ^2 should approximate N_{data}:
  - This approximation becomes an equality as $N \rightarrow \infty$.
  - If χ^2 is >>*N* (for large *N*), it is likely the model is wrong.
  - If χ^2 is << *N*, then one should be suspicious – are the estimates of σ_i too large?
  - As $N \rightarrow \infty$, the χ^2 distribution tends to Normal with variance = 2*N*.
  - Actually, the *N* referred to in the previous line is the number of degrees of freedom, ie $N_{\text{data}}\text{-}N_{\text{param}}$.

- Note that within this framework we focus on the likelihood function and not on the posterior probability distribution for our hypothesis H_1:
  - Just because χ^2 has an unlikely value is not a sufficient reason to reject a hypothesis in comparison with another.
  - But it should force us to consider better alternatives.

The χ^2 distribution $p(z) \propto z^{(N_d/2\ -\ 1)}e^{-z/2}$

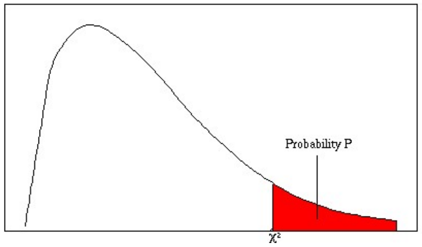
(with a denominator complicated function of N_d)



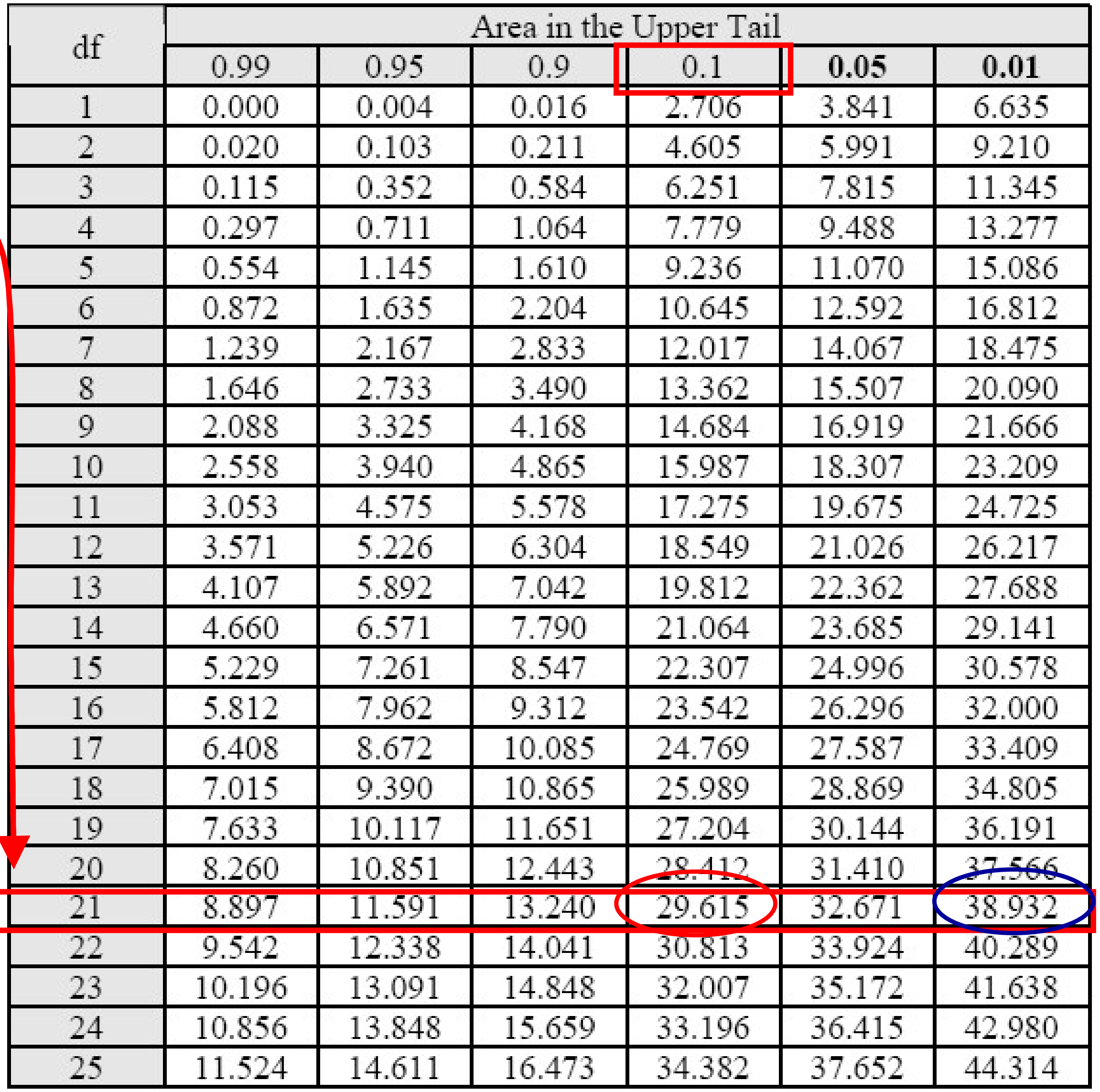
$$\chi^2 = \sum_i \left[\frac{y_i - f(x_i)}{\sigma_i}\right]^2 = \mathrm{z}$$

Using critical values of the χ^2 distribution

$$\chi^2 = \sum_i \left[\frac{y_i - f(x_i)}{\sigma_i} \right]^2$$

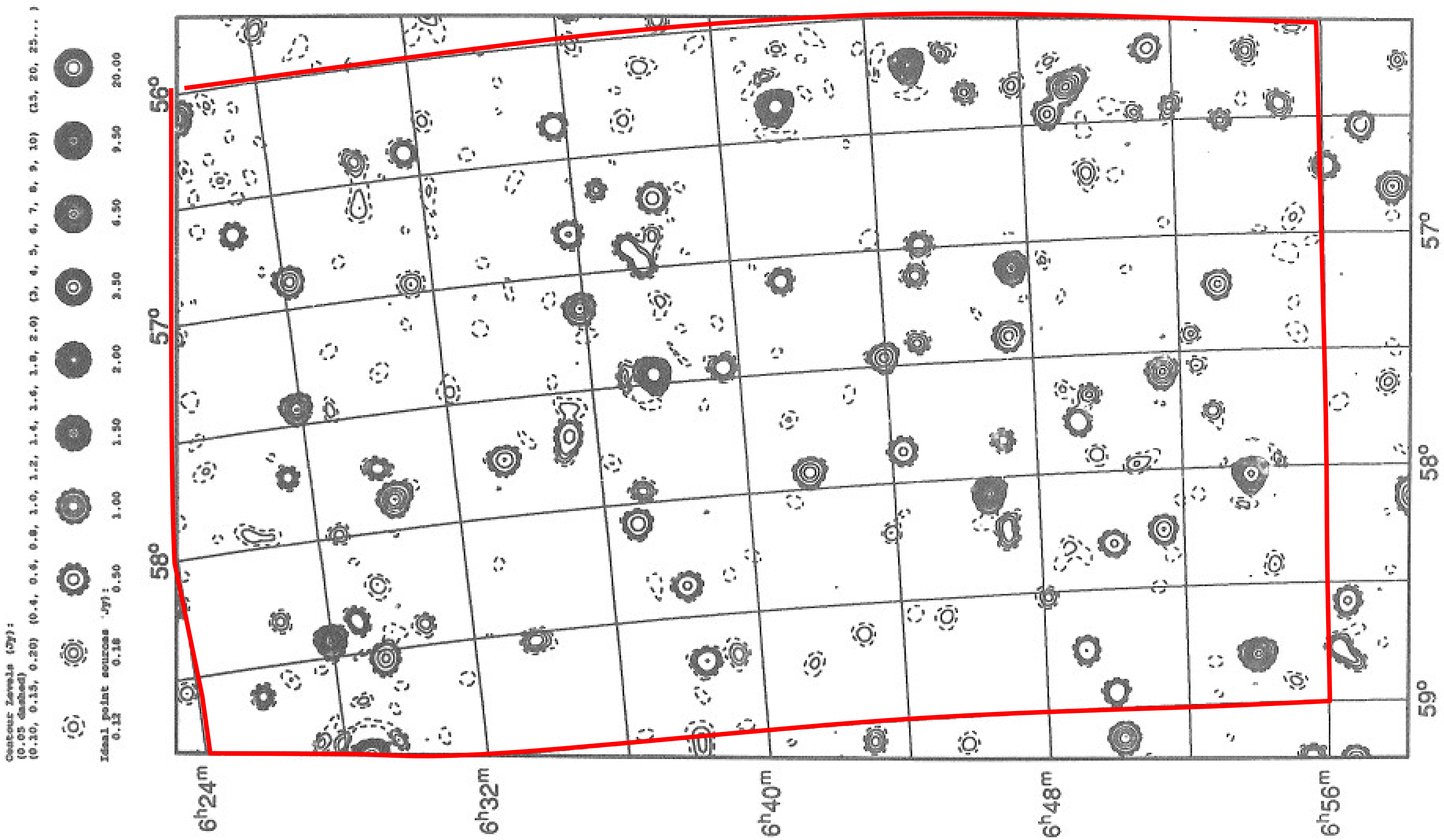


- If your value for χ^2 is, say, 29.6;
- And df – the number of degrees of freedom – is 21;
- Then, if your model is true, there is a 10% chance that the value of χ^2 is as large as 29.6 or larger.
- If χ^2 were as high as 39, the chance of that would be <1%, and you might reconsider the model.



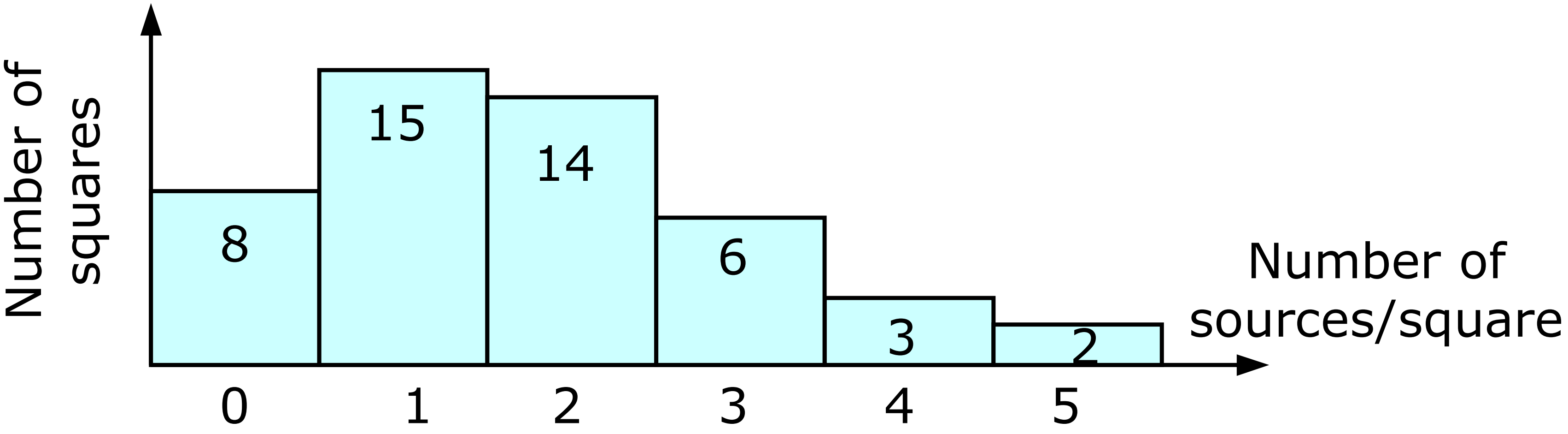
df	Area in the Upper Tail					
	0.99	0.95	0.9	0.1	**0.05**	**0.01**
1	0.000	0.004	0.016	2.706	3.841	6.635
2	0.020	0.103	0.211	4.605	5.991	9.210
3	0.115	0.352	0.584	6.251	7.815	11.345
4	0.297	0.711	1.064	7.779	9.488	13.277
5	0.554	1.145	1.610	9.236	11.070	15.086
6	0.872	1.635	2.204	10.645	12.592	16.812
7	1.239	2.167	2.833	12.017	14.067	18.475
8	1.646	2.733	3.490	13.362	15.507	20.090
9	2.088	3.325	4.168	14.684	16.919	21.666
10	2.558	3.940	4.865	15.987	18.307	23.209
11	3.053	4.575	5.578	17.275	19.675	24.725
12	3.571	5.226	6.304	18.549	21.026	26.217
13	4.107	5.892	7.042	19.812	22.362	27.688
14	4.660	6.571	7.790	21.064	23.685	29.141
15	5.229	7.261	8.547	22.307	24.996	30.578
16	5.812	7.962	9.312	23.542	26.296	32.000
17	6.408	8.672	10.085	24.769	27.587	33.409
18	7.015	9.390	10.865	25.989	28.869	34.805
19	7.633	10.117	11.651	27.204	30.144	36.191
20	8.260	10.851	12.443	28.412	31.410	37.566
21	8.897	11.591	13.240	29.615	32.671	38.932
22	9.542	12.338	14.041	30.813	33.924	40.289
23	10.196	13.091	14.848	32.007	35.172	41.638
24	10.856	13.848	15.659	33.196	36.415	42.980
25	11.524	14.611	16.473	34.382	37.652	44.314

Let's revisit an earlier question: are radio sources distributed randomly on the sky?



Let’s reasonably assume a Poisson model

- Take 6C survey of sky at 0.15 Ghz:
  - Angular resolution = 4 minutes of arc.
  - Important because of “confusion”.
- Look at patch of sky well away from the plane of the galaxy.
- Count sources in 48 squares each of area ≈0.5°×0.5°.



- Then, “mean number of events/interval” = “mean number of sources/square”

  = (15+28+18+12+10)/48 = 1.73.

Now compare data with Poisson prediction with a "rate", λ=1.73 sources per patch

- Poisson probability for λ=1.73 = p(r|1.73)= $\dfrac{(1.73)^r}{r!}e^{-1.73}$

Probability	Expected #	Likely fluctⁿ	Actual #
r=0: 0.177	8	±3	8
r=1: 0.307	15	±4	15
r=2: 0.265	13	±4	14
r=3: 0.153	7	±3	6
r>3: 0.098	5	±2	5

- So, $$\chi^2 = \left(\frac{8-8}{3}\right)^2 + \left(\frac{15-15}{4}\right)^2 + \left(\frac{14-13}{4}\right)^2 + \left(\frac{6-7}{3}\right)^2 + \left(\frac{5-5}{2}\right)^2 = 0.2$$

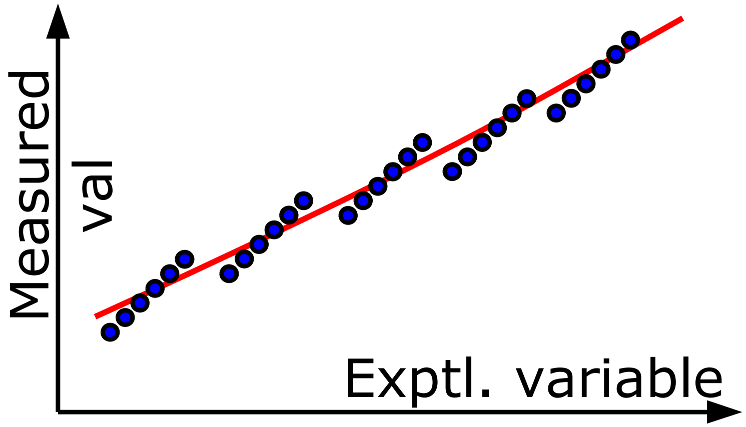
- Now, N_{dof}=5-2 (because of normalization and λ): table ⇒ ~97% of time Poisson predicts somewhat higher value of χ^2.
  - Hmm – just about OK – agreement is almost too good!

What happens if the underlying probability distribution is not known?

- Need to use non-parametric statistics
  - Look at your data!

□ Given the knowledge you have now, you will be able to extract these from books (see feedback notes for lec 7).

- E.g. χ^2 may well say the fit is good. Yet merely plotting the points shows that the model is inadequate. The "runs test" would cope here.
  - Runs test – a test for detecting non-randomness.



□ Other important, easy-to-use non-parametric tests include:

- Sign test: is the dist. of *x* the same as the dist. of *y*?
- Mann-Whitney test: do 2 samples come from the same dist.?
- Kolmogorov-Smirnov test: do 2 probability distributions differ?

Summary

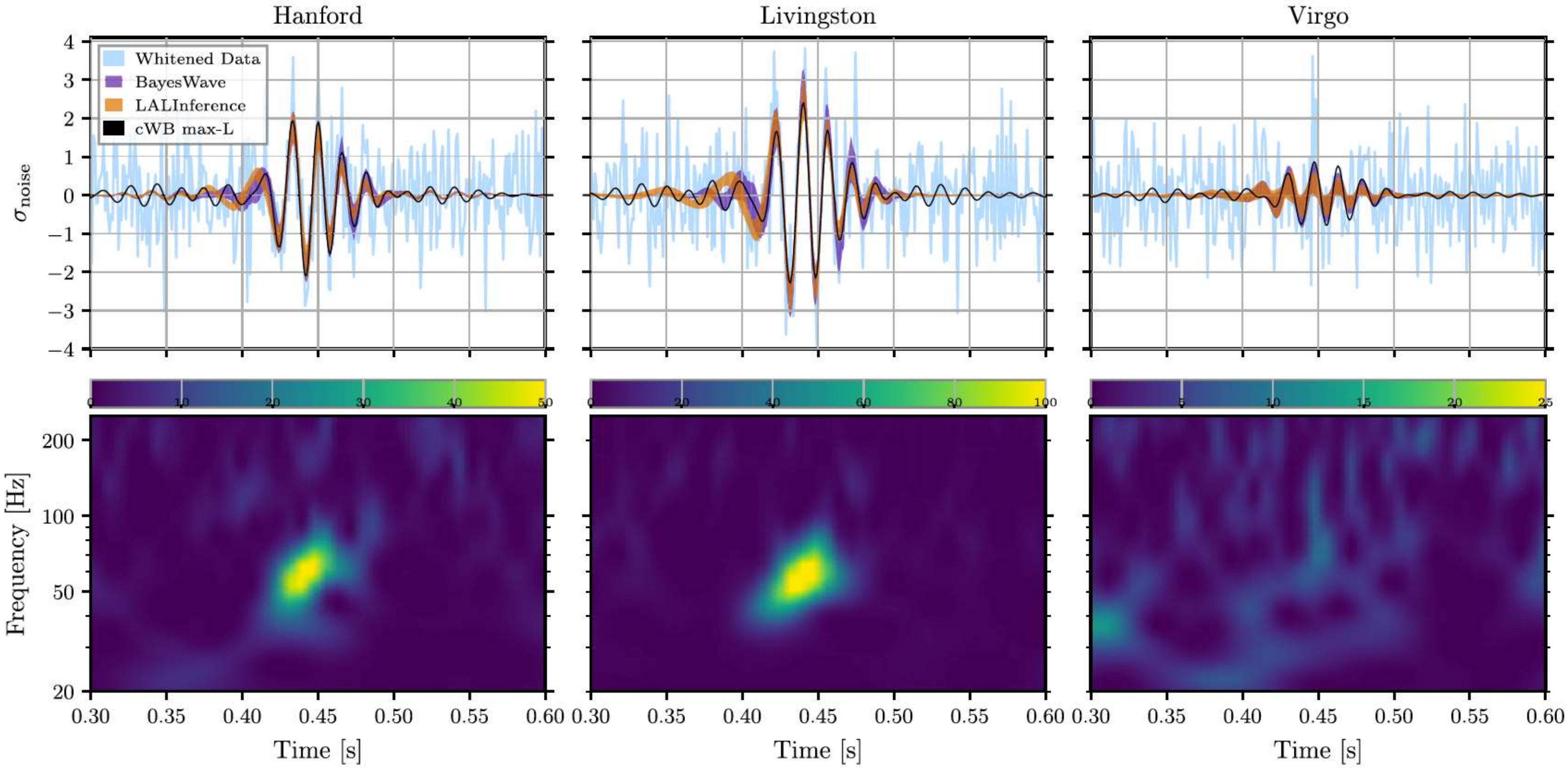
- Inference:
  - Bayes' theorem;
  - Maximum Likelihood – when priors are uniform;
  - The importance of strong priors.

- Least squares fitting:
  - A maximum likelihood approach – first view of χ^2;
  - Example of fitting to a straight line:
    - Unequal errors and weighed means.

- Hypothesis testing and goodness of fit:
  - χ^2 revisited in actual problems;
  - Warnings and caveats;
  - Non-parametric approaches.

Review of the whole course - what was it about?

- GW19052: A BBH merger with a total mass of 150 M_{sun} (Abbot et al, PRL, 125, 101102, 2020)

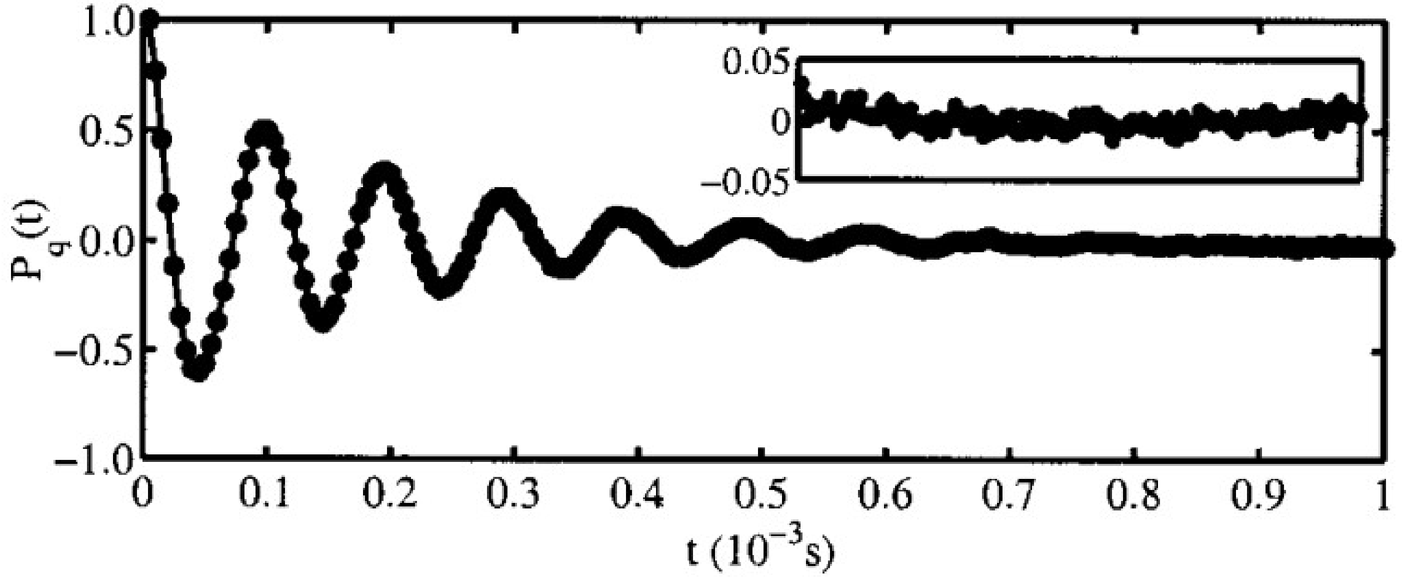


- Physics is not just "equation juggling".
- Making measurements like this needs judgement, skill and care.

Measurement underpins progress in physics.
Measurements that matter are tough to make.

J. Chem. Phys. 114, 8659 (2001)

Why does this matter?



$$D(\omega)=[\epsilon q^2+i\omega\eta(q+m)]\left[\gamma q^2+i\omega\eta(q+m)-\frac{\rho\omega^2}{q}\right]-[i\omega\eta(m-q)]^2 \qquad m=\sqrt{q^2+i\frac{\omega\rho}{\eta}},\quad \mathrm{Re}(m)>0,$$

$$P_q(\omega)=\frac{k_BT}{\pi\omega}\mathrm{Im}\left[\frac{i\omega\eta(m+q)+\epsilon q^2}{D(\omega)}\right]$$

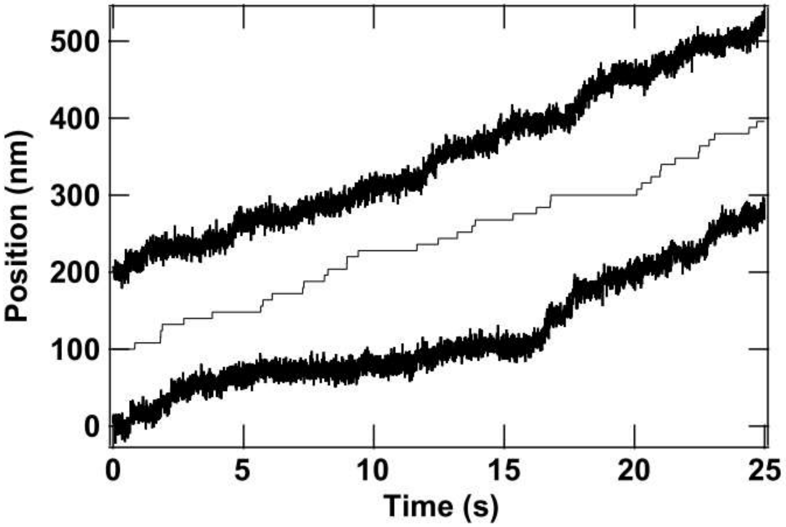
Data of P(t), plotted, is fitted with the Fourier transform of this P_q(w), multiplied by a function that has 3 more instrumental parameters. It’s crazy – the simple oscillation ends up fitted by over 10 parameters…. Of these, only ε is of interest here.

A long time ago in a galaxy far away - Pietro’s PhD

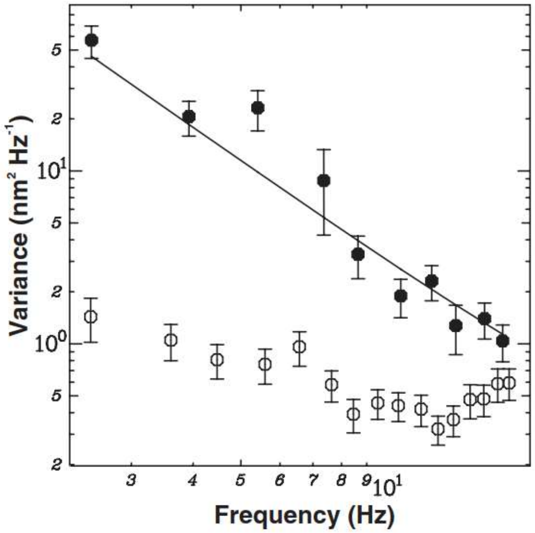
Knowing what to look for – but is it there?

J. Phys.: Condens. Matter 17 S3811 (2005)

Molecular motors “walk” along DNA, in steps. Can we hope to see this?



Experiment. Fourier transform of the variance, fitted with a model with step size. Empty markers are a null-model.



Simulations of single-molecule position recordings (*x(t)*) are generated by adding Gaussian noise to a stochastic stepping trace. Individual step transitions are masked by the noise. Two noisy traces are shown along with one trace without noise. Traces are displaced on the y axis for clarity. The average velocity *v* for each trace is obtained by a line fit to x(t). The simulated data had a step size 8 nm, Gaussian noise with σ = 8 nm, and a stepping rate of 1.2 s^{-1}. Analysis was performed on 100 traces. The average velocity is 9.4 ± 0.2 nm s^{-1}. Measured step size obtained: 7.8 ± 0.3 nm. *Wow*.

Some experiments really face statistical challenges.

Some words from Pl, Op and Fe

- An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer.

- There are children playing in the streets who could solve some of my top problems in physics, because they have modes of sensory perception that I lost long ago.

- The principle of science, the definition, almost, is the following: *The test of all knowledge is experiment.* Experiment is the *sole judge* of scientific "truth."

  But what is the source of knowledge? Where do the laws that are to be tested come from? Experiment, itself, helps to produce these laws, in the sense that it gives us hints.

  But also needed is imagination to create from these hints the great generalizations – to guess at the wonderful, simple, but very strange patterns beneath them all, and then to experiment to check again whether we have made the right guess.