4th IMPRS Astronomy Summer School

Drawing Astrophysical Inferences from Data Sets

William H. Press
The University of Texas at Austin

Lecture 4

Goodness of Fit

Until now, we have assumed that, for some value of the parameters $\mathbf{b}$ the model $y(\mathbf{x}_i|\mathbf{b})$ is correct.

That is a very Bayesian thing to do, since Bayesians start with an EME set of hypotheses. It also makes it difficult for Bayesians to deal with the notion of a model's goodness of fit.

So we must now become frequentists for a while!

Suppose that the model $y(\mathbf{x}_i|\mathbf{b})$ does fit. This is the "null hypothesis".

Then the "statistic" $$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(\mathbf{x}_i|\mathbf{b})}{\sigma_i} \right)^2$$ is the sum of N t^2-values.

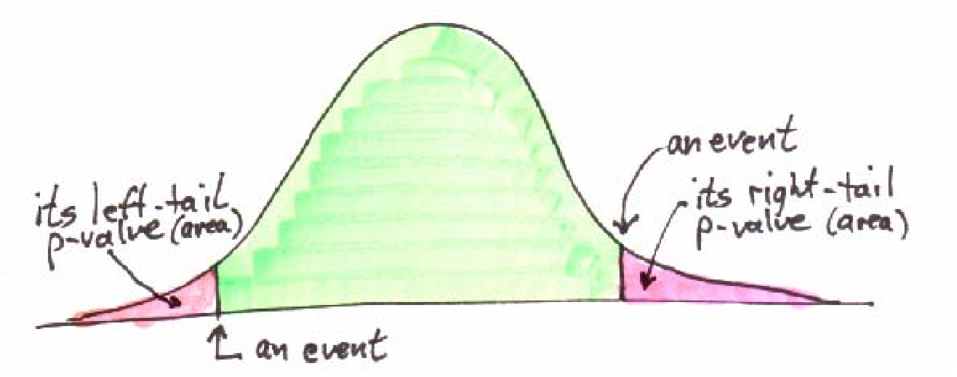
(not quite)

So, if we imagine repeated experiments (which Bayesians refuse to do), the statistic should be distributed as Chisquare(N).

If our experiment is <u>very unlikely</u> to be from this distribution, we consider the model to be disproved.

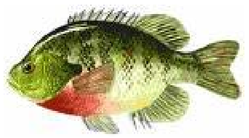
In general, **the idea of p-value (tail) tests** is to see how extreme is the observed data relative to the distribution of hypothetical repeats of the experiment under some “null hypothesis” H_0.

If the observed data is too extreme, the null hypothesis is <u>dis</u>proved. (It can never be proved.)



If the null hypothesis is true, then p-values are uniformly distributed in (0,1), <u>in principle exactly so</u>.

There are some fishy aspects of tail tests, but they have one <u>big</u> advantage over Bayesian methods: You don’t have to enumerate all the alternative hypotheses (“the unknown unknowns”).



Degrees of Freedom: Why is χ^2 with N data points "not quite" the sum of N t^2-values? Because d.o.f.'s are reduced by constraints.

First consider a hypothetical situation where the data has linear constraints:

$$t_i = \frac{y_i - \mu_i}{\sigma_i} \sim \mathrm{N}(0,1)$$

joint distribution on all the t's, if they are independent

$$p(\mathbf{t}) = \prod_i p(t_i) \propto \exp\left(-\tfrac{1}{2}\sum_i t_i^2\right)$$

$$\chi^2 \text{ is squared distance from origin } \sum t_i^2$$

Linear constraint:

$$\sum_i \alpha_i y_i = C = \langle C \rangle = \sum_i \alpha_i \mu_i$$

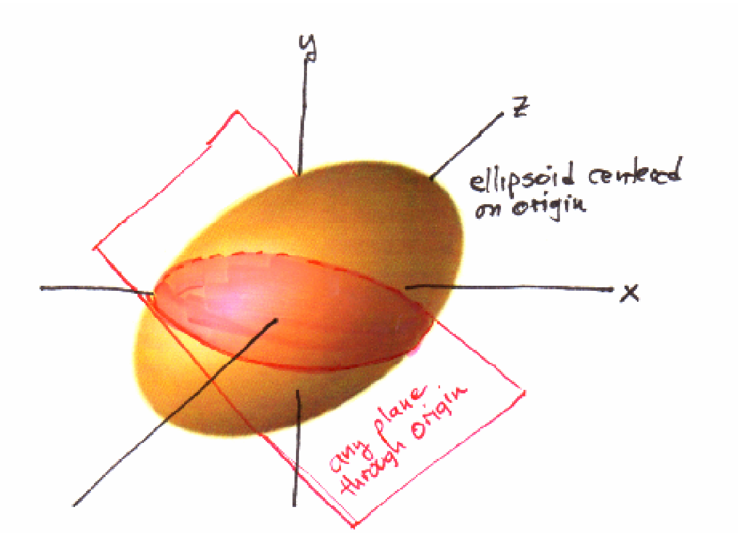
$$C = \sum_i \alpha_i(\sigma_i t_i + \mu_i)$$

$$= \sum_i \alpha_i \sigma_i t_i + C$$

$$\text{So, } \sum_i \alpha_i \sigma_i t_i = 0$$

a hyper plane through the origin in t space!

Constraint is a plane cut.
Any cut through an ellipsoid is an ellipse; any cut through a sphere is a circle.



So the distribution of distance from origin is the same as a multivariate normal "ball" in the lower number of dimensions. Thus, each linear constraint reduces ν by exactly 1.

We don't have explicit constraints on the y_i's. But as the y_i's wiggle around (within their errors) we do have the constraint that we want to keep the MLE estimate $\mathbf{b}_0$ fixed.

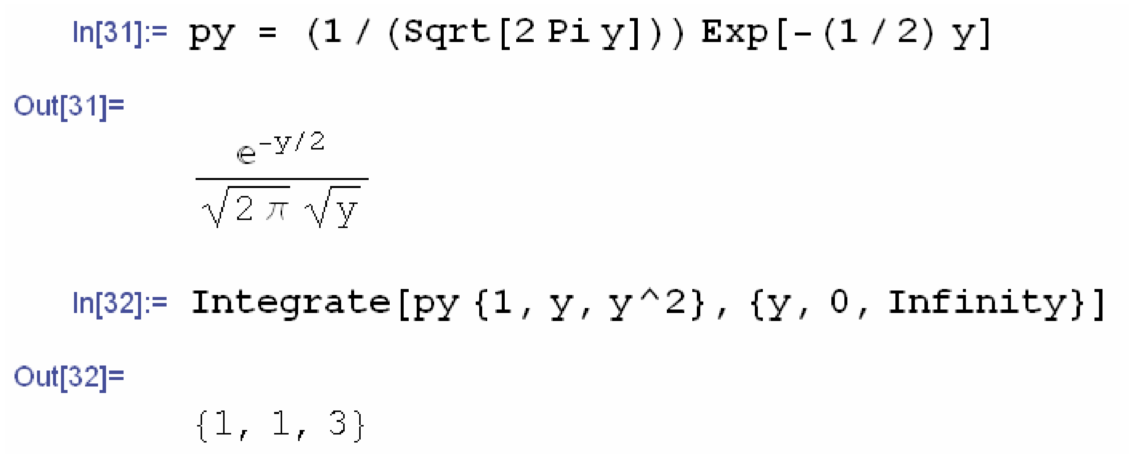
So by the implicit function theorem, there are M (number of parameters) approximately linear constraints on the y_i 's. So $\nu = N - M$, the so-called number of degrees of freedom (d.o.f.).

The Poisson-count pitfall

You can get a statistic that is “accurately” chi-square either by summing (any number of) terms that are accurately squares of normal t-values,

or by summing a large number of terms that individually have the correct mean and variance. This uses the CLT, so the exactness of chi-square is no better than its normal approximation.

Compute moments of chi-square with 1 d.f.:



```
In[31]:= py = (1 / (Sqrt[2 Pi y])) Exp[-(1 / 2) y]

Out[31]=
```

$$\frac{e^{-y/2}}{\sqrt{2\,\pi}\,\sqrt{y}}$$

```
In[32]:= Integrate[py {1, y, y^2}, {y, 0, Infinity}]

Out[32]=
         {1, 1, 3}
```

$$\text{So,}\quad \mu = 1, \qquad \sigma^2 = 3 - 1 = 2$$

$$\text{Hence,}\quad \text{Chisquare}(\nu) \to \text{Normal}(\nu, \sqrt{2\nu}) \text{ as } \nu \to \infty$$

If you are going to rely on the CLT and sum up lots of not-exactly-t bins, it is really important that they have the expected mean and variance.

Example: Chi-square test with small numbers of Poisson counts in some or all bins. (People often get this wrong!)

Recall Poisson: $$p(n) = e^{-\mu} \frac{\mu^n}{n!}$$

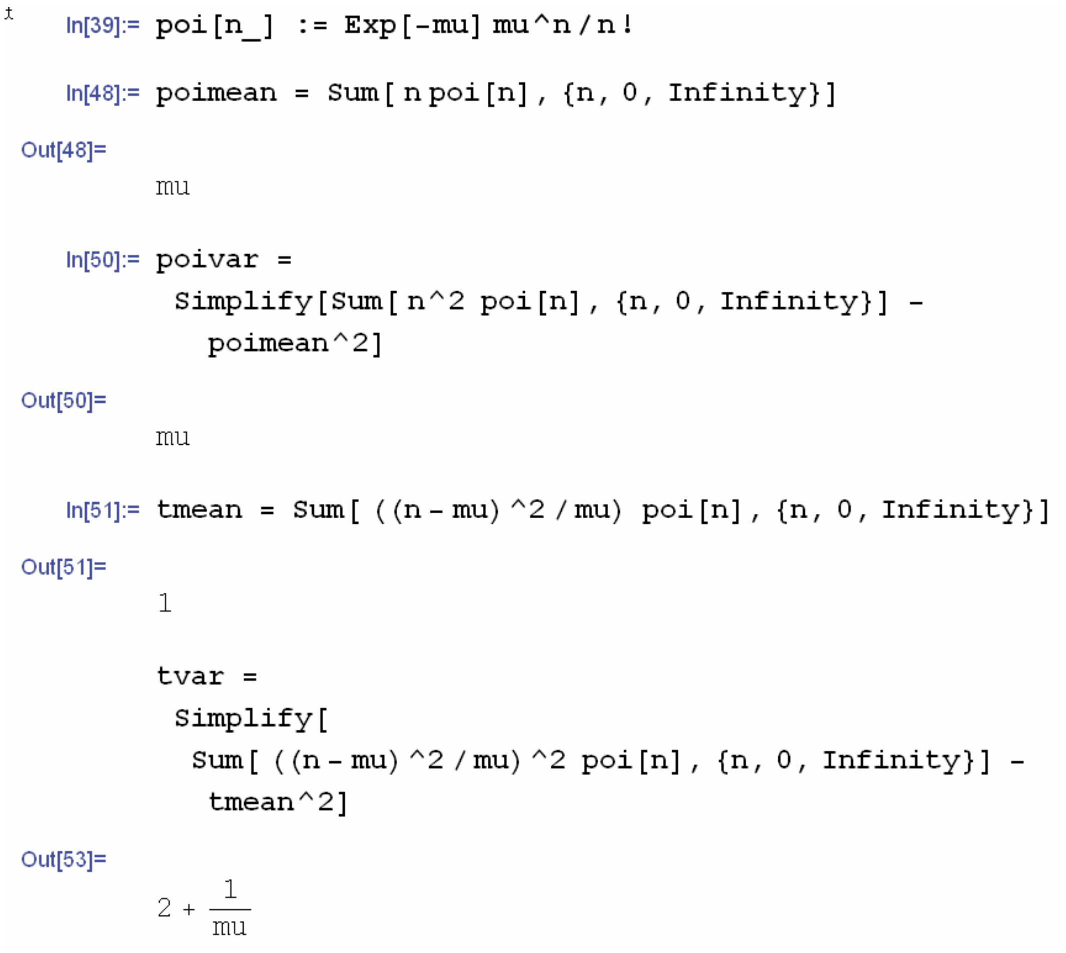
Mean and variance are both = μ

So, given a set of Poisson counts and expected values (x_i, μ_i) it is very tempting to write

$$\chi^2 = \sum_i \frac{(x_i - \mu_i)^2}{\mu_i}$$

The problem is that this is not Chisquare distributed.

And, it is not $\sim \mathrm{Normal}(\nu, \sqrt{2\nu})$ for <u>any</u> value of ν, even as the number of data points becomes large! Let's see why.



```
In[39]:= poi[n_] := Exp[-mu] mu^n/n!

In[48]:= poimean = Sum[n poi[n], {n, 0, Infinity}]

Out[48]=
         mu

In[50]:= poivar =
          Simplify[Sum[n^2 poi[n], {n, 0, Infinity}] -
            poimean^2]

Out[50]=
         mu

In[51]:= tmean = Sum[((n - mu)^2/mu) poi[n], {n, 0, Infinity}]

Out[51]=
         1

         tvar =
          Simplify[
           Sum[((n - mu)^2/mu)^2 poi[n], {n, 0, Infinity}] -
            tmean^2]

Out[53]=
             1
         2 + --
             mu
```

So this χ^2 is not Chi-square distributed! Rather, asymptotically,

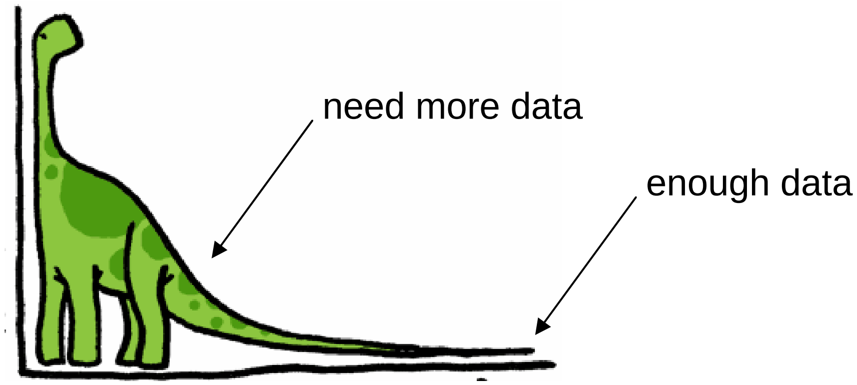
$$\chi^2 \sim \text{Normal}\left(\nu, 2\nu + \sum_i \mu_i^{-1}\right)$$

What about bins with μ near zero? (Decide in advance!)

Tips on tail tests:

Don’t sweat a p-value like 0.06. If you really need to know, the only real test is to get significantly more data. Rejection of the null hypothesis is exponential in the amount of data.

In principle, p-values from repeated tests s.b. exactly uniform in (0,1). In practice, this is rarely true, because some “asymptotic” assumption will have crept in when you were not looking. All that really matters is that (true) extreme tail values are being computed with moderate fractional accuracy. You can go crazy trying to track down not-exact-uniformity in p-values. (I have!)



The χ^2 versus $\Delta\chi^2$ pitfall

Goodness-of-fit with $\nu = N - M$ degrees of freedom:

we expect $\chi^2_{\text{min}} \approx \nu \pm \sqrt{2\nu}$

Confidence intervals for parameters **b**:

we expect $\chi^2 \approx \chi^2_{\text{min}} \pm 1$

How can ± 1 have any meaning in the presence of $\pm\sqrt{2\nu}$?

Answer: χ^2 and $\Delta\chi^2$ are different concepts!

χ^2 increases linearly with $\nu = N - M$

$\Delta\chi^2$ increases as N (number of terms in sum), but also decreases as $(N^{-1/2})^2$, since **b** becomes more accurate with increasing N :

$$\Delta\chi^2 \propto N(\delta b)^2, \quad \delta b \propto N^{-1/2} \quad \Rightarrow \quad \Delta\chi^2 \propto \text{const}$$

quadratic, because at minimum

What is the uncertainty in quantities other than the fitted coefficients:

Method 1: Linearized propagation of errors

$$
\begin{aligned}
f \equiv f(\mathbf{b}) &= f(\mathbf{b}_0) + \mathbf{f}' \cdot \mathbf{b}_1 \\
\langle f \rangle &= \langle f(\mathbf{b}_0) \rangle + \mathbf{f}' \cdot \cancel{\langle \mathbf{b}_1 \rangle} = f(\mathbf{b}_0) \\
\left\langle f^2 \right\rangle - \langle f \rangle^2 &= 2 f(\mathbf{b}_0)(\mathbf{f}' \cdot \cancel{\langle \mathbf{b}_1 \rangle}) + \left\langle (\mathbf{f}' \cdot \mathbf{b}_1)^2 \right\rangle \\
&= \mathbf{f}' \cdot \langle \mathbf{b}_1 \otimes \mathbf{b}_1 \rangle \cdot \mathbf{f}' \\
&= \mathbf{f}' \cdot \boldsymbol{\Sigma} \cdot \mathbf{f}'
\end{aligned}
$$

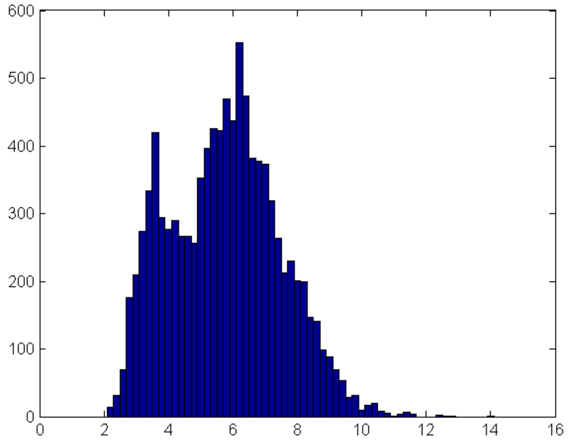
Method 2: Sample from the posterior distribution

1. Generate a large number of (vector) **b**'s

$$\mathbf{b} \sim \mathrm{MVNormal}(\mathbf{b}_0, \boldsymbol{\Sigma}_b)$$

2. Compute your $f(\mathbf{b})$ separately for each **b**

3. Histogram



Note that **b** is typically (close to) m.v. normal because of the CLT, but your (nonlinear) *f* may not, in general, be anything even close to normal!

Method 3: Bootstrap resampling of the data

- We applied some end-to-end process to a data set and got a number f out
- The data set was drawn from a population
  - which we don't get to see, unfortunately
  - we see only a sample of the population
- We'd like to draw new data sets from the population, reapply the process, and see the distribution of answers
  - this would tell us how accurate the original answer was
  - but we can't: we don't have access to the population
- However, the data set itself is an estimate of the population pdf!
  - in fact, it's the only estimate we've got!
- We draw – with replacement – from the data set and carry out the proposed program
  - Bootstrap theorem [glossing over technical assumptions]: The distribution of any resampled quantity around its full-data-set value estimates (naively: "has the same histogram as") the distribution of the data set value around the population value.

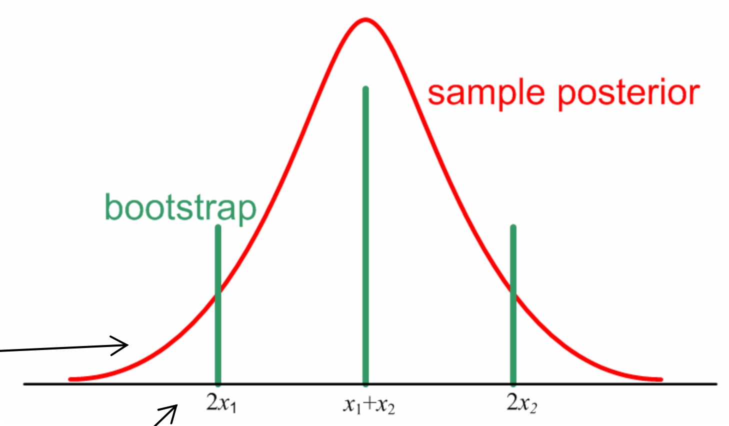
Sampling the posterior “honors” the stated measurement errors. Bootstrap doesn’t. That can be good!

Suppose (toy example) the “statistic” is

$$s = x_1 + x_2$$

then the posterior probability is

$$P(s) \propto \exp\left[-\frac{1}{2}\frac{(s - x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2}\right]$$

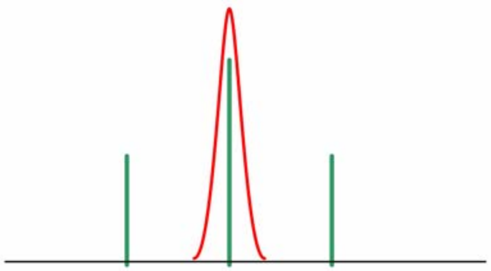


Note that this depends on the σ’s!

The bootstrap (here noticeably discrete) doesn’t depend on the σ’s. In some sense it estimates them, too.

So, if the errors were badly underestimated, sampling the posterior would give too small an uncertainty, while bootstrap would still give a valid estimate.

If the errors are right, both estimates are valid. Notice that the model need not be correct. Both procedures give valid estimates of the statistical uncertainty of parameters of even a wrong (badly fitting) model. But for a wrong model, your interpretation of the parameters may be misleading!



Compare and contrast bootstrap resampling and sampling from the posterior

Both have same goal: Estimate the accuracy of fitted parameters.

- Bootstrap is frequentist in outlook
  - draw from "the population"
  - even if we have only an estimate of it (the data set)
- Easy to code but computationally intensive
  - great for getting your bearings
  - must repeat your basic fitting calculation over all the data100 or 1000 times
- Applies to both model fitting and descriptive statistics
- Fails completely for some statistics
  - e.g. (extreme example) "harmonic mean of distance between consecutive points"
  - how can you be sure that your statistic is OK (without proving theorems)?
- Doesn't generalize much
  - take it or leave it!

- Sampling from the posterior is Bayesian in outlook
  - there is only one data set and it is never varied
  - what varies from sample to sample is the goodness of fit
  - we don't just sit on the (frequentist's) ML estimate, we explore around
- In general harder to implement
  - we haven't learned how yet, except in the simple case of an assumed multivariate normal posterior
  - will come back to this next, when we do Markov Chain Monte Carlo (MCMC)
  - may or may not be computationally intensive (depending on whether there are shortcuts possible in computing the posterior)
- Rich set of variations and generalizations are possible

(patients not polyps)