Astronomy 8824: Statistics Notes 5
Hypothesis Testing

We have focused so far on the task of estimating parameter values and their errors when fitting data.

These results presume that the data are described by the model in question for *some* value of the parameters.

But once we have fit for parameters, how do we decide whether the model itself is viable, or compare two models.

We'll eventually take a Bayesian approach to this problem, but let's first look at a frequentist recipe that is often useful in practice.

Expected value of χ^2

Suppose that we have fit a model to data by minimizing χ^2.

Gould (2003) proves (a standard result) that at the minimum

$$\langle \chi^2 \rangle = N - n + \langle \Delta_k \rangle C^{-1}_{kl} \langle \Delta_l \rangle,$$

where N is the number of data points and n is the number of parameters that are fit.

But if the model is a correct description of the data for some choice of parameters, then $\langle \Delta_k \rangle = \langle y_{\rm mod}(x_k) \rangle - \langle y_k \rangle = 0$.

Thus, for a correct model, we expect χ^2 to be approximately $N - n$, the number of data points minus the number of fitted parameters, usually referred to as the number of "degrees of freedom."

Alternatively, the *reduced* χ^2, sometimes written χ^2/d.o.f., is expected to be approximately one.

This result does not assume Gaussian errors on the data, and it does not assume that the errors are uncorrelated.

Distribution of χ^2

If the errors on the data are Gaussian and the model is correct, then value of χ^2 follows a χ^2 distribution with $k = N - n$ degrees of freedom (see Ivezic §3.3.7).

The variance of this distribution is $2k$.

Alternatively, the standard deviation for χ^2/k (reduced χ^2) is $\sqrt{2/k}$.

If the number of degrees of freedom is large, then the distribution of χ^2/k approaches a Gaussian distribution with mean 1 and standard deviation $\sqrt{2/k}$.

Suppose we have 12 data points that we fit with a straight line, and we get $\chi^2 = 14.47$ for the best-fit slope and amplitude.

Then χ^2/d.o.f. $= 1.447 = 1 + \sqrt{2/10}$, so this fit is only discrepant with the data at the 1σ level.

However, if we have 120 data points and the same χ^2/d.o.f., then the discrepancy is $0.447/\sqrt{2/118} = 3.4\sigma$.

If the value of χ^2/k is much larger than $1 + \sqrt{2/k}$, then it probably indicates that either (1) the model is incorrect, or (2) the errors have been underestimated, or (3) the errors are significantly non-Gaussian, so that "outliers" are giving anomalously large contributions to χ^2.

It will generally take thought and further inspection to determine which of these is going on.

Note that these results apply unchanged for correlated (multi-variate Gaussian) errors, but the calculation of χ^2 must correctly incorporate the error covariance matrix.

Thus, a specific instance of "(2) the errors have been underestimated" is "the covariance matrix has significant off-diagonal terms that have not been accounted for when computing χ^2."

If the value of χ^2/k is much *smaller* than $1 - \sqrt{2/k}$ then it usually indicates that the errors have been underestimated.

Linear constraints

The above results are consistent with our basic intuition.

If a model is correct and the errors are correct, then data will typically scatter about the model at about the level of the 1σ error bars.

For N data points we therefore expect $\chi^2/N \approx 1$, not $\chi^2 = 0$.

Each free parameter increases our ability to "fit the noise," so we expect a lower value of χ^2. We could in principle use a free parameter to exactly fit one data point, reducing the expected χ^2 by one.

This turns out to be exactly right, as $\langle \chi^2 \rangle = N - n$.

We may also have a linear constraint on the parameters, for example that they sum to one, or that the average of the distribution is zero, or even just knowing the value of one parameter.

Gould (2003) gives formulae for the best-fit parameter values in this case.

He further shows that (if both the constraints and the model are correct) then imposing m constraints changes the expected value of χ^2 to $\langle \chi^2 \rangle + N - n + m$.

This again accords with intuition: imposing a constraint is equivalent to removing one degree of freedom.

The χ^2 hypothesis test

The frequentist version of the χ^2 test is simply this: a model should be rejected if its value of χ^2 (for the best-fit parameters) is large enough to be highly improbable.

Specifically, if the probability $P(> \chi^2)$ of obtaining a χ^2 greater than the best-fit value is q, then the model is rejected at the $1 - q$ confidence level. For example, if $P(> \chi^2) = 0.01$, then the model is rejected at 99% confidence.

The cumulative probability distribution $P(> \chi^2)$ can be found in tables or computed via python routines; it can be approximated by a complementary error function (integral of a Gaussian) if the number of degrees of freedom is large.

One can make various complaints about this test — Why integrate over values of χ^2 larger than the observed one? Why reject a model for anomalously large χ^2 values but not for anomalously small ones? — but it basically makes sense. If a model has a very small $P(>\chi^2)$ it is probably wrong, or else the errors are wrong.

An important note about χ^2 parameter constraints

The likelihood of a set of parameter values relative to the best-fit values is $\exp(-\Delta\chi^2/2)$, where $\Delta\chi^2$ is the change in χ^2 relative to its minimum value.

The 68% confidence interval on a parameter (in a one-parameter fit) corresponds to $\Delta\chi^2 = 1$, *not* to $\Delta\chi^2/\text{d.o.f.} = 1$.

More than one astronomy paper has incorrectly used the latter.

Bayesian Hypothesis Comparison

(See Ivezic §5.4.)

Bayes' Theorem gives a straightforward expression for the relative probability of two hypotheses:

$$\frac{p(H_1|DI)}{p(H_2|DI)} = \frac{p(H_1|I)}{p(H_2|I)} \times \frac{p(D|H_1I)}{p(D|H_2I)}.$$

We multiply our prior probabilities by the relative probabilities of obtaining the data under the two hypotheses. The global likelihood $p(D|I)$ cancels out of the comparison.

This ratio is called the *odds ratio.*

If the hypotheses are simple, with no free parameters, then this comparison is straightforward. However, if the hypotheses are models with parameters, we must integrate over the possible parameter values. This can be complicated, but it also has interesting effects when comparing two models with different numbers of parameters, or even with the same number of parameters but different degrees of prior predictiveness.

Example (From Loredo, §5.3)

We previously gave

$$p(D|\mu I) = (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{Ns^2}{2\sigma^2}\right] \exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\right]$$

as the probability of obtaining the data $D = \{x_i\}$ drawn from a Gaussian distribution with mean μ and dispersion σ.

Consider the competing hypotheses

H_1 = mean of distribution is a specified value μ_1

H_2 = mean of distribution is in range $\mu_{\rm min} \le \mu \le \mu_{\rm max}$, with a flat prior $p(\mu|I) = (\mu_{\rm max} - \mu_{\rm min})^{-1}$ in this range.

H_2 will *always* fit the data better, unless the mean happens to be exactly μ_1, in which case it fits equally well.

But does this mean H_2 is actually the preferred hypothesis?

$$P(D|H_1 I) = K \times \exp\left[-\frac{N}{2\sigma^2}(\overline{x} - \mu_1)^2\right],$$

where

$$K = (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{Ns^2}{2\sigma^2}\right]$$

is independent of μ_1.

$$\begin{aligned} p(D|H_2 I) &= \int_{\mu_{\rm min}}^{\mu_{\rm max}} p(D|\mu I) p(\mu|I) d\mu \\ &= K(\mu_{\rm max} - \mu_{\rm min})^{-1} \int_{\mu_{\rm min}}^{\mu_{\rm max}} d\mu \exp\left[-\frac{N}{2\sigma^2}(\overline{x} - \mu)^2\right]. \end{aligned}$$

If $\mu_{\rm max} - \overline{x}$ and $\overline{x} - \mu_{\rm min}$ are both $\gg \sigma/\sqrt{N}$, then the integral is just $(2\pi\sigma^2/N)^{1/2}$, since a Gaussian $(2\pi\sigma^2)^{-1/2}\exp(-x^2/2\sigma^2)$ integrates to one.

In this case

$$\frac{p(D|H_1 I)}{p(D|H_2 I)} = \frac{(\mu_{\rm max} - \mu_{\rm min})}{(2\pi\sigma^2/N)^{1/2}} \exp\left[-\frac{N}{2\sigma^2}(\overline{x} - \mu_1)^2\right].$$

If we considered the two hypotheses equally probable before hand, $p(H_1|I) = p(H_2|I)$, then this ratio is also the ratio of posterior probabilities.

Model 2 is "penalized" for having less predictive power than Model 1, and the amount of the penalty depends on the ratio of $(\mu_{\rm max} - \mu_{\rm min})$ to the actual uncertainty in the mean $\sigma/\sqrt{N}$.

Model 1 is penalized because it doesn't fit the data as well as the best fit versions of Model 2. If it is nonetheless fairly close, then it may win out as the more probable hypothesis, otherwise it won't.

For another example, see Ivezic §5.4.2.

More generally, we can see from the structure of the integral $\int p(\theta|I)p(D|\theta I)d\theta$ that a model with a free parameter θ will gain to the extent that its best fit value $\hat{\theta}$ yields a greater likelihood $p(D|\hat{\theta}I)$, but will lose to the extent that $p(\theta|I)$ is broad and "spreads out" the predictive power.

The Bayesian expression for hypothesis comparison thus yields Occam's razor as a *result*: the preferred model is the one that fits the data adequately with the least freedom to be adjusted to do so.

In principle, this provides a well defined way to decide whether a more complicated model is "worth it."

In general cases, the integrals over parameter values may be impossible to do analytically, though they can probably be done numerically.

Note that while we have used a Gaussian example here, the analysis is not restricted to any particular probability distribution.

Indeed, one could use these ratio tests to compare the hypothesis that the data have Gaussian errors with a fixed dispersion to the hypothesis that there is an additional "outlier" population drawn from a broader Gaussian, or that the error distribution is exponential instead of Gaussian.

Rules of thumb

Leaving aside the Bayesian approach, we should also mention the $\Delta\chi^2$ rule of thumb: an additional parameter should reduce χ^2 by $\Delta\chi^2 > 1$ to be considered significant.

Roughly, you can think of this rule as saying that one parameter can be chosen to perfectly explain one data point, so it should typically reduce $\Delta\chi^2$ by one even if the more complicated model has no more explanatory power than the simpler model.

This rule can be justified more rigorously in terms of the expected value of χ^2 in linear model fits, where adding n parameters reduces the expected value of χ^2 by n.

A $\Delta\chi^2 = 1$ is enough to prefer one parameter value over another at 1σ, but it would be an undemanding criterion for accepting a model that was actually more complicated.

The Aikake information criterion (AIC, Ivezic §4.3.2) is a popular choice for frequentist comparison of models with different numbers of parameters.

In terms of the Bayesian odds ratio, a ratio > 10 might be taken as interesting evidence for one hypothesis over another.

For equal priors (so that the odds ratio equals the likelihood ratio) and Gaussian errors, an odds ratio of 10 corresponds to $\Delta\chi^2 = -2\ln 0.1 = 4.6$ or a 2.1σ difference.

An odds ratio of 100 corresponds to $\Delta\chi^2 = 13.8$ or a 3.7σ difference, which might be taken as "decisive" evidence.

The Bayesian Information Criterion (BIC, Ivezic §5.4.3) is an approximate method of estimating the odds ratio from the maximum values of the data likelihood, without marginalizing over the full parameter space.

The preferred model is the one with the smaller value of

$$\mathrm{BIC} \equiv -2\ln\left[L^0(M)\right] + k\ln N$$

where $L^0(M)$ is the likelihood of the model with best-fit parameter values, k is the number of model parameters, and N is the number of data points.

Absolute model assessment

In a Bayesian approach, there is really no such thing as an absolute model assessment.

If one has an exhaustive set of possible hypotheses, H_1, H_2, ... H_N, then one can ask about the probability that any one of those hypotheses is correct

$$p(H_i|DI) = p(H_i|I)\frac{p(D|H_iI)}{p(D|I)},$$

where

$$p(D|I) = \sum_{i=1}^{N} p(D|H_i I)$$

is computed by summing over all of the hypotheses.

But there isn't a Bayesian way to assess a hypothesis in isolation without specifying alternatives.

The traditional way to do an absolute model assessment in the frequentist approach is to compute some statistic, say χ^2, that increases for worse fits, then ask how often one would expect to get a value that large *or larger* if the hypothesis were true.

If this probability α is small, then the model is rejected at the $1 - \alpha$ confidence level.

There are some problems with this approach: the answer depends on what statistic you choose, it may depend on what you think the alternative "data sets" are, and there is sometimes ambiguity about what "tail" of the distribution one should consider. For example, low χ^2 values can be as improbable as high χ^2 values — should a model be rejected because it fits the data too well?

Despite these problems, these frequentist assessments seem to make good sense in some cases, and choices among seemingly ambiguous alternatives (e.g., whether to reject low χ^2 values) can often be made sensibly in the context of a specific problem.