Astronomy 8824: Statistics Notes 6
Estimating Errors From Data

Where does the error bar go?

Suppose you measure the average depression of flux in a quasar caused by absorption from the Lyman-alpha forest. You find that 30% of the flux is absorbed, $D_A = 0.3$.

You have two models that predict $D_A = 0.32$ and $D_A = 0.4$, respectively. Which do the data favor? Is either ruled out?

To answer, we need an error bar, and this may be different for the two models.

If the first model predicts $D_A = 0.32$ on average and an rms variation of 0.002 from one quasar to another, then the predicted $D_A = 0.32 \pm 0.002$ is strongly inconsistent with the observed $D_A = 0.3$, unless the predicted distribution of variations is highly non-Gaussian.

If the second model predicts $D_A = 0.4$ on average and an rms variation of 0.05 from one quasar to another, then its prediction $D_A = 0.4 \pm 0.05$ is marginally inconsistent with your measurement. The data favor this model even though its mean prediction is further from the observed value.

This example illustrates the Bayesian insistence that error bars really belong *on the model*, not on the data, since different models may predict different error bars for the same data set.

But suppose we measure the decrement for 20 quasars instead of one, and we find a mean of 0.3 and an rms variation about the mean of 0.05.

Here it seems legitimate to say that the uncertainty on the mean is $0.05/\sqrt{20} = 0.01$, and that our measurement implies $D_A = 0.3 \pm 0.01$.

What allows us to attach an error bar to the data, and to implicitly claim that it is model independent?

In effect, this procedure relies on the assumption (which should be good in this case) that any model that will fit the data must also predict an rms variation similar to the value 0.05 that you measured, and that it will therefore predict an error on the mean for a sample of 20 quasars that is close to $0.05/\sqrt{20}$.

A model that predicts a mean $D_A = 0.35$ and an rms quasar-to-quasar variation of 0.3 gives $D_A = 0.35 \pm 0.06$ for a sample of 20 quasars. But although its mean prediction is consistent with the measured mean within its expected error, the rms variation for this model is inconsistent with the measured rms variation of 0.05, so the model is ruled out, or at least disfavored, on other grounds. (To decide just how inconsistent the model is, we would need to calculate the error bar on the rms variation.)

For quantities that are well measured (i.e., determined to fairly high fractional precision), it is usually OK to "transfer" the error bar in this way, because the data have sufficient power to constrain the variation within the sample and yield an estimated error bar that must be close to that of any model that would be consistent with the data.

However, you should be very cautious about "transferring" the error bar in any case where

the estimated fractional uncertainty is large. In these cases, the error bar is often highly model dependent.

Small number statistics

The extreme, and often relevant, example is a survey that turns up one object of a certain class.

It is tempting to say that the measured number of objects is 1 ± 1 and therefore consistent with zero.

It is true that a model that predicts a mean of one object in a survey of this size predicts (assuming Poisson statistics) that a fraction $e^{-1} = 0.37$ of such surveys would detect no objects, and that a model that predicts a mean of two objects predicts that $2^1 e^{-2}/1! = 0.27$ of such surveys should yield one object and is therefore consistent with the data.

However, if we have a model that predicts a mean of 0.001 objects, then it predicts that only $0.001^1 e^{-0.001}/1! = 0.001$ of such surveys should yield one object, so it is ruled out (or at least strongly disfavored).

A model that predicts a mean of 10 objects is also strongly disfavored, as $P(k = 1|\mu = 10) = 10^1 e^{-10} = 4.5 \times 10^{-4}$.

Even with very small numbers of objects, one can make statistically interesting statements about some models.

Another example, relevant, e.g., to *Chandra* data.

Suppose that the background is very low, e.g., 10^{-4} counts/pixel in a 50 ksec exposure.

If you have 10^6 pixels, there will be ~ 100 background counts, so a single-photon detection isn't significant.

However, the probability of getting 2 background photons in a pixel is 10^{-8}, so there should be only 0.01 pixels out of 10^6 with two background photons by chance.

Therefore, 2 counts in a single pixel would be a statistically significant detection of an object.

Also, if you knew ahead of time where you were going to look (e.g., at a recent supernova) to within a pixel, then even a single photon detection would be significant at the 10^{-4} level, and would rule out a model that predicted only 10^{-3} source counts in a 50 ksec exposure.

Moral: Don't automatically discount small number statistics, though you should use them with caution.

Estimating error bars from the data: subsample, jackknife, bootstrap

In a case like the average quasar flux decrement above, it is obvious how to estimate the error bar from the data using the rms variation.

But suppose we are doing something more complicated, e.g., measuring the power spectrum of the flux (a 1-d function) after fitting a continuum to each spectrum, removing metal lines, and subtracting photon noise. We have done some complicated processing, and the

signal-to-noise of the measurement in each individual spectrum may be quite different from one quasar to another.

Subsample

One way to proceed in a complicated case like this is to divide the data into subsamples, say five groups of four quasars each. You can then apply your measurement separately to each subsample and estimate errors from the subsample-to-subsample variation.

For example, you now have $N = 5$ estimates ${}^kP_i$ of the power spectrum on spatial scale i, where $k = 1, ...N$. You can estimate the error bar σ_{ii} on P_i as

$$\sigma_{ii}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\frac{({}^kP_i - \bar{P}_i)^2}{N},$$

where $\bar{P}_i$ is your estimate from the full sample of all quasars.

You might also want to estimate the covariance of errors from two different length scales i and j:

$$\sigma_{ij}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\frac{({}^kP_i - \bar{P}_i)({}^kP_j - \bar{P}_j)}{N}.$$

Jackknife

This approach can run into problems if you really need something close to your full sample size to get a usable measurement in the first place, so that the estimates from your much smaller subsamples are wildly varying. (This is especially problematic if, for instance, you know that the quantity you are measuring is positive-definite but noise means that you can get negative values in individual measurements.)

An attractive, more robust alternative is jackknife error estimation, where you *omit* each subsample in turn, and apply your measurement to all of the *remaining* data. The error estimate in this case is

$$\sigma_{ii}^2 = \frac{N-1}{N}\sum_{k=1}^{N}({}^kP_i - \bar{P}_i)^2,$$

where ${}^kP_i$ now represents the estimate of P_i after subsample k is *omitted* from the data sample.

The pre-factor is larger by a factor of $(N-1)^2$, but the variation $({}^kP_i - \bar{P}_i)$ is smaller because each subset k is now close to the full sample.

In the case we are considering, the individual subsamples could now be single quasars, so we could set $N = 20$ and omit each quasar in turn.

Where the jackknife and subsample error estimates would give different answers, I think the jackknife estimate is generally preferable.

Bootstrap

A widely used variant on the same theme is bootstrap resampling. Here you create new samples the same size as the original data sample by drawing from that sample "with replacement."

Each of the N bootstrap samples has $M = 20$ quasars randomly drawn from the original set, but in an individual sample quasar 1 may appear three times, quasar 2 twice, and quasar 3 not at all.

The error bars are simply computed from the dispersion among the M bootstrap samples,

$$\sigma_{ii}^2 = \sum_{k=1}^{N} \frac{({}^k P_i - \bar{P}_i)^2}{N},$$

where $\bar{P}_i$ is the estimate from the full sample, not the mean of the bootstrap samples. There is no pre-factor because now each bootstrap sample is the same size as the full sample.

Bootstrapping seems to be moderately preferred by the cognoscenti over subsampling or jackknife, but the fact that involves replacement (and a bootstrap sample therefore has some identical elements) can cause problems in some cases, so think about what you are doing.

General Remarks

All of these approaches are implicitly "transferring" the error bars as discussed above.

The idea behind all three is that the data are drawn from some distrbution and that we can estimate that distribution from the data themselves. Each subsample (or jackknife sample, or bootstrap sample) is drawn from this distribution, so we get an internal estimate of what variation is expected in data drawn from this distribution.

Again, we are implicitly assuming that any model that would actually fit the data would have a similar distribution and would therefore predict similar errors.

A critical assumption for any of these methods is that the individual subsamples are *independent*.

For the quasar case described above, this assumption is probably fine, since the regions of the universe sampled by different quasars are far enough apart that they are uncorrelated.

Suppose we are instead trying to estimate uncertainties in the galaxy correlation function measured from a redshift survey.

Each galaxy is a separate data point, but they are highly correlated because they trace the same underlying structure (e.g., the same clusters and superclusters).

Using subsamples, jackknife, or bootstrap with individual galaxies would severely underestimate the errors.

For this case, we need to define subsamples that are spatially contiguous volumes, large enough that the estimates of the correlation function in each subsample are independent. Roughly speaking, we want to be sure that the spatial size of each subsample is large compared to the largest coherent structures that are found in the universe.

For an example of this approach, see Zehavi et al. 2005, ApJ, 630, 1

Error bars from artificial data sets

If we have a model that we want to test, and we can generate complete, independent artificial data sets from the model, then it is better to estimate errors and covariances from large numbers of mock data sets instead of using these "internal" techniques.

For example, it is now common practice to create artificial mock galaxy catalogs from cosmological simulations to estimate errors and covariances for galaxy clustering measurements.

This can be very computationally demanding, and developing efficient tools for creating simulated data sets that are "accurate enough" for evaluating errors can be a research problem in itself.

In principle one should generate different sets of mock catalogs for all models being tested, or evaluate the dependence of the covariance matrix on model parameters.

The accuracy required to estimate errors is usually lower than the accuracy required to evaluate parameters by fitting a model to the data.

In practice this approach is usually applied for a fiducial model that is expected to represent the properties of the data reasonably well.

Noise and bias in covariance matrices

Whether one is using an internal method or artificial data sets, one should be aware of the potential problem of noise in the estimated covariance matrix, since one may be estimating large numbers of σ_{ij}.

Even if the individual estimates are unbiased, noise may cause some of them to be artificially large. Since it is the inverse of the covariance matrix that gets used in evaluating the likelihood, noisy estimates of the covariance matrix can cause misleading conclusions about best-fit parameter values, parameter uncertainties, or relative merit of models.

If you have an idea of what the general structure of the covariance matrix should be, you can impose "regularization" constraints to reduce noise. See Padmanabhan et al. (2016, MNRAS 460, 1567, `arXiv:1512.01241`) for a recent discussion and references therein.

A somewhat separate problem is that the inverse covariance matrix estimated from a finite number of artificial data sets or subsamples can be systematically biased. This problem (and a partial solution) is discussed by Hartlap & Schneider (2007, A&A 464, 399) and more recently by Paz & Sanchez (2015, MNRAS 454, 4326).