Astronomy Statistics Notes

David Weinberg, Fall 2007

Some References

Bayesian Methods

From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics, by T. Loredo

The Promise of Bayesian Inference for Astrophysics, by T. Loredo

both of these are available at `http://www.astro.cornell.edu/staff/loredo/bayes/tjl.html`

There is substantial overlap between the two articles. The first is more "philosophical" in orientation, the second more "practical."

Fitting lines to data

Fits, and especially linear fits, with errors on both axes, extra variance of the data points, and other complications, by G. D'Agostini, astro-ph/0511182

Monte Carlo Markov Chains

First-Year WMAP Observations: Parameter Estimation Methodology, by L. Verde et al., ApJS 148, 195. See §3, especially §3.3.

Analyze This! A cosmological constraint package for CMBEASY, by M. Doran & C. M. Müller, astro-ph/0311311

General Statistics

Statistics in Theory and Practice, by R. Lupton, Princeton University Press

Probability and Statistics, by M. DeGroot, Addison-Wesley

Statistical Tasks in Astrophysics

Four common statistical tasks:

Parameter estimation

Comparison of hypotheses

"Absolute" evaluation of a hypothesis

Forecasting of errors

Simple example: Data points with error bars.

Parameter estimation: What are slope and amplitude of a power-law fit?

Assumes that power-law description is valid.

Hypothesis comparison: Is a double power-law better than a single power-law?

Hypothesis comparisons are trickier when the number of parameters is different, since one must decide whether the fit to the data is *sufficiently* better given the extra freedom in the more complex model.

A simpler comparison would be single power-law vs. two constant plateaus with a break at a specified location, both with two parameters.

Absolute evaluation: Are the data consistent with a power-law?

Absolute assessments of this sort are generally much more problematic than hypothesis comparisons.

Forecasting of errors: How many more objects, or what reduction of uncertainties, would allow single and double power-law models to be clearly distinguished?

Need to specify goals, and assumptions about data.

Common need for observing proposals, grant proposals, satellite proposals ...

Complicated example: CMB power spectrum with errors.

Parameter estimation: In a "vanilla" ΛCDM model, what are the best values of Ω_m, Ω_b, h, n, and τ?

Often want to combine CMB with other data to break degeneracies, get better constraints.

Hypothesis comparisons: Are data consistent with $\Omega_m = 1$? Do they favor inclusion of space curvature, or gravity waves?

Typically involves comparison of models with different numbers of parameters.

Absolute assessment: Can the restricted, "vanilla" ΛCDM model be rejected?

Forecasting: What constraints or tests could be achieved with a new experiment?

This kind of analysis played a key role in the design and approval of WMAP and Planck.

There is now lots of work along these lines for SNAP, Sunyaev-Zel'dovich surveys.

Warmup

Some Definitions

If $p(x)$ is the probability distribution function (pdf) of a random variable x, then $p(x)dx$ is the probability that x lies in a small interval dx.

The expectation value of a function $y(x)$ is $\langle y(x)\rangle = \int_{-\infty}^{\infty} y(x)p(x)dx$.

The distribution mean is $\mu = \langle x\rangle = \int_{-\infty}^{\infty} xp(x)dx$.

The variance is $V(x) = \left\langle (x-\mu)^2 \right\rangle \equiv \sigma^2$.

The standard deviation is $\sigma = \sqrt{\sigma^2}$. This is also called the dispersion.

For *independent* random variables y_1, y_2, ... y_N (drawn from the same distribution or different distributions), the variance of the sum is the sum of the variances:

$$V(y_1 + y_2 + ...y_N) = \sum_{i=1,N} V(y_i).$$

This can be proved by induction.

Some Simple Estimators

Suppose we have N independent data points drawn from an unknown distribution $p(x)$.

The obvious estimator for the mean of the distribution is the sample mean, $\overline{x} = \frac{1}{N}\sum x_i$.

$$\langle \overline{x} \rangle = \left\langle \frac{1}{N}\sum x_i \right\rangle = \frac{1}{N}\sum \langle x_i \rangle = \mu.$$

Thus, the sample mean is an *unbiased* estimator of μ.
The variance of this estimator is

$$\left\langle (\overline{x}-\mu)^2 \right\rangle = V\left(\frac{1}{N}\sum x_i\right) = \frac{1}{N^2}V\left(\sum x_i\right) = \frac{1}{N^2}\sum V(x_i) = \frac{1}{N^2}\times N\sigma^2 = \frac{\sigma^2}{N},$$

where σ^2 is the variance of the underlying distribution.
We have used the fact that $\langle \overline{x} \rangle = \mu$, and we have used the assumed independence of the x_i to go from the variance of a sum to a sum of variances.

An alternative estimator for the mean is the value of the third sample member, x_3.
Since $\langle x_3 \rangle = \mu$, this estimator is unbiased, but $V(x_3) = \sigma^2$, so this estimate is noisier than the sample mean by $\sqrt{N}$.

A more reasonable estimator is the sample *median*, though this is a biased estimator if $p(x)$ is asymmetric about the mean.
If $p(x)$ is Gaussian, then the variance of the sample median is $\frac{\pi}{2}\frac{\sigma^2}{N}$, so it is a less *efficient* estimator than the sample mean.
However, if $p(x)$ has long non-Gaussian tails, then the median may be a much *more* efficient estimator of the true mean (i.e., giving a more accurate answer for a fixed number of data points), since it is not sensitive to rare large or small values.
Estimators that are insensitive to the extremes of a distribution are often called *robust* estimators.

The obvious estimator for the variance of the distribution is the sample variance

$$s^2 = \frac{1}{N}\sum (x_i - \overline{x})^2 = \frac{1}{N}\sum x_i^2 - \overline{x}^2.$$

However, a short derivation shows that

$$\left\langle s^2 \right\rangle = \frac{N-1}{N}\sigma^2,$$

biased low because we had to use the sample mean rather than the true mean, which on average drives down the variance.
An unbiased estimator is therefore

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum (x_i - \overline{x})^2.$$

If you compute the mean of a sample, or of data values in a bin, the estimated *standard deviation of the mean* is

$$\hat{\sigma}_\mu = \left[\frac{1}{N(N-1)}\sum(x_i - \overline{x})^2\right]^{1/2}.$$

Note that this is smaller by $N^{-1/2}$ than the estimate of the dispersion within the bin. You should always be clear which quantity (dispersion or standard deviation of the mean) you are plotting.

If $p(x)$ is Gaussian, then the distribution of $\overline{x}/\sigma$ is a Gaussian of width $N^{-1/2}$. However, the distribution of $\overline{x}/\hat{\sigma}$ is broader (a Student's t distribution).

What is wrong with this plot?

Bayesian Statistics

Suppose we have measured the mean mass of a sample of G stars, by some method, and say: at the 68% confidence level the mean mass of G stars is $a \pm b$. What does this statement mean?

Bayesian answer: There is some true mean mass α of G stars, and there is a 68% probability that $a - b \leq \alpha \leq a + b$.

More pedantically: The hypothesis that the true mean mass α of G stars lies in the range $a - b$ to $a + b$ has a 68% probability of being true.

The probability of the hypothesis is a real-numbered expression of the degree of belief we should have in the hypothesis, and it obeys the axioms of probability theory.

In "classical" or "frequentist" statistics, a probability is a statement about the frequency of outcomes in many repeated trials. With this restricted definition, one can't refer to the probability of a hypothesis — it is either true or false. One can refer to the probability of data if a hypothesis is true, where probability means the fraction of time the data would have come out the way it did in many repeated trials.

So the statement means something like: if $\alpha = a$, we would have expected to obtain a sample mean in the range $a \pm b$ 68% of the time.

This is the fundamental conceptual difference between Bayesian and frequentist statistics.

Bayesian: Evaluate the probability of a hypothesis in light of data (and prior information). Parameter values or probability of truth of a hypothesis are random variables, data are not.

Frequentist: Evaluate the probability of obtaining the data — more precisely, the fraction of times a given *statistic* (such as the sample mean) applied to the data would come out the way it did in many repeated trials — given the hypothesis, or parameter values. Data are random variables, parameter values or truth of hypotheses are not.

My opinion: The Bayesian formulation corresponds better to the way scientists actually think about probability, hypotheses, and data. It provides a better conceptual basis for figuring out what to do in a case where a standard recipe does not neatly apply. But frequentist methods sometimes seem easier to apply, and they clearly capture *some* of our intuition about probability.

Bottom line: One should be a Bayesian in principle, but maybe not always in practice.

Probability Axioms and Bayes' Theorem

Probabilities are real numbers $0 \leq p \leq 1$ obeying the axioms

$$p(A|C) + p(\overline{A}|C) = 1.$$

$$p(AB|C) = p(A|BC)P(B|C)$$

Here $\overline{C}$ means "not C" and AB means "A and B" and is thus equivalent to BA.

From this equivalence we see that

$$p(AB|C) = p(A|BC)p(B|C) = p(BA|C) = p(B|AC)p(A|C).$$

From the 2nd and 4th entries above, we arrive at *Bayes' Theorem*

$$p(A|BC) = p(A|C)\frac{p(B|AC)}{p(B|C)}.$$

In application to scientific inference, this theorem is usually written

$$p(H|DI) = p(H|I)\frac{p(D|HI)}{p(D|I)},$$

where

$H =$ hypothesis, which might be a statement about a parameter value, e.g., the population mean lies in the range $x \rightarrow x + dx$.

$D =$ data

$I =$ background information, which may be minimally informative or highly informative.

$p(H|I) =$ "prior" probability, i.e., before data are considered

$p(D|HI) =$ "likelihood" of data given H and I

$p(D|I) =$ "global likelihood"

$p(H|DI) =$ "posterior" probability, the probability of the hypothesis after consideration of the data

Thus, Bayes' Theorem tells us how to update our estimate of the probability of a hypothesis in light of new data.

It can be applied sequentially, with the posterior probability from one experiment becoming the prior for the next, as more data become available.

Calculation of likelihood, $P(D|HI)$, is sometimes straightforward, sometimes difficult. I may specify assumptions like Gaussian error distribution, independence of data points.

Important aspect of Bayesian approach: only the actual data enter, not hypothetical data that could have been taken.

All the evidence of the data is contained in the likelihood.

The global likelihood of the data, $P(D|I)$ is the sum (or integral) over “all” hypotheses. This can be a slippery concept.

Often $P(D|I)$ doesn’t matter: in comparing hypotheses or parameter values, it cancels out.

When needed, it can often be found by requiring that $p(H|DI)$ integrate (or sum) to one, as it must if it is a true probability.

The Bayesian approach forces specification of alternatives to evaluate hypotheses.

Frequentist assessment tends to do this implicitly via the choice of statistical test.

Criticism of Bayesian approach

The incorporation of priors makes Bayesian methods seem subjective, and it is the main source of criticism of the Bayesian approach.

If the data are compelling and the prior is broad, then the prior doesn’t have much effect on the posterior. But if the data are weak, or the prior is narrow, then it can have a big effect.

Sometimes there are well defined ways of assigning an “uninformative” prior, but sometimes there is genuine ambiguity.

Bayesian methods sometimes seem like a lot of work to get to a straightforward answer.

In particular, we sometimes want to carry out an “absolute” hypothesis test without having to enumerate all alternative hypotheses.

Criticism of frequentist approach

Doesn’t correspond as well to scientific intuition. We want to talk about the probability of hypotheses or parameter values.

The choice of which statistical test to apply is often arbitrary. There is not a clear way to go from the result of a test to an actual scientific inference about parameter values or validity of a hypothesis.

Bayesians argue (and I agree) that frequentist methods obtain the appearance of objectivity only by sweeping priors under the rug, making assumptions implicit rather than explicit.

Frequentist approach relies on hypothetical data as well as actual data obtained. Choice of hypothetical data sets is often ambiguous, e.g., in the “stopping” problem.

Sometimes we *do* have good prior information. It is straightforward to incorporate this in a Bayesian approach, not so in frequentist.

Frequentist methods are poorly equipped to handle “nuisance parameters,” which in Bayesian approach are easily handled by marginalization.

Parameter Estimation

Hypothesis is "true value of parameter is $\theta_{\rm true} = \theta$" (discrete) or "true value of parameter is $\theta \le \theta_{\rm true} \le \theta + d\theta$" (continuous).

$$p(\theta|DI) = p(\theta|I)\frac{p(D|\theta I)}{p(D|I)}.$$

If θ is continuous, then, technically, $p(\theta|DI)$ and $p(\theta|I)$ both have a $d\theta$ attached.

A Bayesian searches for the parameter value with *maximum posterior probability* $p(\theta|DI)$.

If $p(\theta|I)$ is flat, then this is also the value with *maximum likelihood* $p(D|\theta I)$.

Maximum likelihood estimators play a major role in both Bayesian and classical approaches.

Example: Estimate mean from N measurements x_i, when dispersion σ is known, and x_i are Gaussian distributed and independent. (Following Loredo, §5.2.2)
Flat prior: $p(\mu|I) = (\mu_{\rm max} - \mu_{\rm min})^{-1}$.
Likelihood:

$$\begin{aligned} p(\{x_i\}|\mu I) &= \prod_i (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= (2\pi\sigma^2)^{-N/2}\exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right] \\ &= (2\pi\sigma^2)^{-N/2}\exp\left[-\frac{Ns^2}{2\sigma^2}\right]\exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\right], \end{aligned}$$

where $\overline{x} = \frac{1}{N}\sum x_i$ is sample mean and $s^2 = \frac{1}{N}\sum(x_i - \overline{x})^2$ is sample variance.
Global likelihood: $p(\{x_i\}|I) = \int_{\mu_{\rm min}}^{\mu_{\rm max}} p(\{x_i\}|\mu I) d\mu$.

Final result is

$$p(\mu|\{x_i\}I)d\mu = K\left(\frac{N}{2\pi\sigma^2}\right)^{1/2}\exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\right], \quad \mu_{\rm min} \le \mu \le \mu_{\rm max},$$

a Gaussian with mean $\overline{x}$ and dispersion $\sigma/\sqrt{N}$, truncated at $\mu_{\rm min}$ and $\mu_{\rm max}$, with K a normalization constant such that the probability integrates to one.

As long as prior range is big compared to $\sigma/\sqrt{N}$, prior doesn't matter, otherwise it does, by truncation and normalization $K > 1$.

If new measurements come in, they can be incorporated by taking output of this result as *prior* for new analysis.

At least at informal level, this is often done, e.g., H_0 priors on CMB analyses.

To have $p(\theta|DI) \propto p(D|\theta I)$, we need the prior $p(\theta|I)$ to be flat in the range allowed by the data, not universally.

For example, we may know that $\mu > 0$ on physical grounds. If $\bar{x} \gg \sigma/\sqrt{N}$, then $p(\mu|I)$ is approximately flat in the allowable range if it is "broad" compared to $\sigma/\sqrt{N}$. But if $\bar{x} \sim \sigma/\sqrt{N}$, then a flat prior cannot be a good approximation.

For a positive-definite parameter where we have essentially no prior knowledge about its value, a common choice of prior is $p(\theta|I) \propto 1/\theta$, i.e., flat in $\ln\theta$ instead of θ itself.

Another example: slightly more complicated, but standard.

Determine best values of a and b in linear fit $y = ax + b$, given data points with known errors on y, assuming Gaussian error distribution:

$$p(\hat{y}_i|y_i) = (2\pi\sigma_{y,i}^2)^{-1/2} \exp\left[\frac{-(\hat{y}_i - y_i)^2}{2\sigma_{y,i}^2}\right],$$

where y_i is the true value and $\hat{y}_i$ is the observed value.

Likelihood

$$\begin{aligned}\mathcal{L} = p(\{\hat{y}_i\}|a,b) &= \prod_i p(\hat{y}_i|ax_i + b) \\ &= \prod_i (2\pi\sigma_{y,i}^2)^{-1/2} \exp\left[-\frac{(\hat{y}_i - ax_i - b)^2}{2\sigma_{y,i}^2}\right].\end{aligned}$$

It is often convenient to work with the logarithm of the likelihood

$$\ln\mathcal{L} = -\frac{1}{2}\sum \frac{(\hat{y}_i - ax_i - b)^2}{2\sigma_{y,i}^2} + C,$$

where C depends on the (known) errors $\sigma_{y,i}$ but is independent of a and b.

The maximum likelihood solution is thus the solution with minimum

$$\chi^2 = \sum \frac{(\hat{y}_i - ax_i - b)^2}{\sigma_{y,i}^2},$$

and $\ln\mathcal{L} = \exp(-\chi^2/2) + C$.

For this problem, one can find standard analytic expressions for a and b in terms of the data and error bars by solving the equations that define the maximum of the likelihood function,

$$\frac{\partial \ln\mathcal{L}}{\partial a} = 0, \qquad \frac{\partial \ln\mathcal{L}}{\partial b} = 0.$$

Another example: Very useful, but non-standard.

Now consider a more complicated variation of this problem: fit $\overline{y} = ax + b$, with measurement errors in x and y *and* intrinsic scatter in the relation between y and x.

A model with intrinsic scatter (here assumed constant from point to point and denoted σ) is usually more realistic than the commonly adopted, perfect correlation model.

If all of the scatters are Gaussian distributed, we have

$$\begin{aligned} p(y_i|x_i) &= (2\pi\sigma^2)^{-1/2}\exp\left[\frac{-(y_i - ax_i - b)^2}{2\sigma^2}\right] \\ p(\hat{y}_i|y_i) &= (2\pi\sigma_{y,i}^2)^{-1/2}\exp\left[\frac{-(\hat{y}_i - y_i)^2}{2\sigma_{y,i}^2}\right] \\ p(\hat{x}_i|x_i) &= (2\pi\sigma_{x,i}^2)^{-1/2}\exp\left[\frac{-(\hat{x}_i - x_i)^2}{2\sigma_{x,i}^2}\right]. \end{aligned}$$

In this case we want to maximize

$$\mathcal{L} = \prod_i p(\hat{y}_i|\hat{x}_i) \qquad \Longrightarrow \qquad \ln\mathcal{L} = \sum_i \ln p(\hat{y}_i|\hat{x}_i).$$

So we need the expression for $p(\hat{y}_i|\hat{x}_i)$.

$$\begin{aligned} p(\hat{y}_i|\hat{x}_i) &= \int_{-\infty}^{\infty} dy_i\, p(\hat{y}_i|y_i)\, p(y_i|\hat{x}_i) \\ &= \int_{-\infty}^{\infty} dy_i\, p(\hat{y}_i|y_i) \int_{-\infty}^{\infty} dx_i\, p(y_i|x_i)\, p(x_i|\hat{x}_i). \end{aligned}$$

Now assume a flat prior on x_i, $p(x_i)$ =const., so that $p(x_i|\hat{x}_i) = p(\hat{x}_i|x_i)$ (by Bayes' theorem and the requirement that probabilities integrate to one). This assumption is non-trivial, but usually OK because we only require flatness over the range allowed by $\hat{x}_i$.

We can now substitute our expressions for the probabilities, and several pages of algebra and integrals lead eventually to the expression

$$p(\hat{y}_i|\hat{x}_i) = (2\pi)^{-1/2}(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)^{-1/2}\exp\left[-\frac{(\hat{y}_i - a\hat{x}_i - b)^2}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)}\right].$$

This expression looks eminently sensible. For $\sigma_{x,i} = 0$, we get a Gaussian whose width is the quadrature sum of the intrinsic and observational scatter in y. Non-zero $\sigma_{x,i}$ increases the probability of larger deviation between observed and predicted y_i by allowing the true value of $ax_i + b$ to be closer to $\hat{y}_i$ than $a\hat{x}_i + b$.

A deviation $\Delta y_i/\sigma_{y,i}$ has similar weight to a deviation $a\Delta x_i/\sigma_{x,i}$. If you think of x and y as having different units, then it is obvious that a factor of a is needed to give $\sigma_{y,i}$ and $a\sigma_{x,i}$ the same dimensions.

The maximum likelihood solution requires maximizing

$$\sum_i \ln p(\hat{y}_i|\hat{x}_i) = -\frac{1}{2}\sum_i \ln(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2) - \sum_i \frac{(\hat{y}_i - a\hat{x}_i - b)^2}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)} + \text{constant},$$

and thus solving the equations

$$\frac{\partial \ln\mathcal{L}}{\partial a} = 0, \qquad \frac{\partial \ln\mathcal{L}}{\partial b} = 0, \qquad \frac{\partial \ln\mathcal{L}}{\partial \sigma} = 0.$$

There is a straightforward algebraic solution for b,

$$b = \frac{-\sum_i (a\hat{x}_i - \hat{y}_i)W_i}{\sum_i W_i},$$

where the weights are

$$W_i = \frac{1}{2(\sigma^2 + \sigma_{y,i}^2 + a^2\sigma_{x,i}^2)} \ .$$

This is just an inverse-variance weighted average of the individual estimates of b.

I couldn't find algebraic solutions for a and σ, but it is straightforward to search a grid of (a, σ), finding the best b for each (a, σ) from the above equation and evaluating the overall likelihood.

There are a couple of points worth noting about the likelihood expression.

First, you might naively have thought that with intrinsic scatter as a free parameter, the maximum likelihood solution would be to have a very large intrinsic scatter, since then each deviation would contribute very little to χ^2.

However, while the second term in the likelihood always rewards large σ^2, the first term penalizes it, basically because the prediction $ax + b$ is diluted by being spread over a large range, so it doesn't get much "credit" when it is close.

If a significant fraction of points have deviations that put them on the exponential tail of the Gaussian, then raising σ will increase the likelihood, but once the typical deviation falls to $\sim 1\sigma$, raising σ will decrease the likelihood.

This is, of course, what ought to happen. If the prediction is a scatterplot (as happens in the limit of large intrinsic scatter), then it is unlikely to actually have the points lie close to a line.

Second, if we reverse the roles of y and x, letting the intrinsic scatter be on x rather than y, then the solution for a and b (especially a) will be different.

Intrinsic scatter on y is a *different* hypothesis from intrinsic scatter on x, and the corresponding best fit slopes and intercepts are different.

The difference goes away if σ is small compared to the observational errors.

Confidence intervals, nuisance parameters, and marginalization

From a Bayesian point of view, the end result of a parameter estimation calculation *is* the posterior probability distribution $p(\theta|DI)$. For a flat prior $p(\theta|I)$, this is just proportional to the likelihood.

If you give an expression for, table of, or plot of the likelihood function, then you have presented all of the evidence of the data, and others can apply prior probabilities or frequentist assessments as they wish. Thus, if statistics are important to your answer, there is much to be said for presenting things this way if you can.

We often *summarize* the results of a calculation with an estimate and a confidence interval.

Typically, one would quote the maximum likelihood (or maximum posterior probability) value as the estimate, though if the likelihood function is far from Gaussian people sometimes quote the likelihood weighted mean.

The confidence interval is a region of highest likelihood (or posterior probability) and is characterized by the fraction of the probability that it contains.

For a 1-dimensional problem (1 parameter), this is usually straightforward, though even here a complicated likelihood function may have multiple maxima.

For a Gaussian likelihood function,

$$\ln\mathcal{L} = \ln\mathcal{L}_{\rm max} - \frac{1}{2}\Delta\chi^2, \qquad \mathcal{L} = \mathcal{L}_{\rm max}e^{-\Delta\chi^2/2}.$$

The regions $\Delta\chi^2 \leq 1$, $\Delta\chi^2 \leq 4$, and $\Delta\chi^2 \leq 9$ contain 68.3%, 95.4%, and 99.73% of the probability.

For a non-Gaussian likelihood function, it can be useful instead to quote the values where $\mathcal{L}$ falls to some fraction of its maximum value, say 0.1, in which case the parameter value is 10 times less probable than its most probable value. This particular fraction corresponds in the Gaussian case to 2.14σ, since $e^{-2.14^2/2} = 0.1$.

If there are multiple parameters, then confidence intervals are defined by contours in a multi-dimensional parameter space.

If the likelihood function is a multi-variate Gaussian, then these contours are ellipses, with the direction of the ellipse axes depending on the covariance of the errors in the parameters.

For the 2-d case, the contours $\Delta\chi^2 = 2.30$, 6.17, and 9.21 enclose 68.3%, 95.4%, and 99% of the probability. See the *Numerical Recipes* chapter on "Modeling of Data" for higher dimensions and more discussion.

In some cases, a sensible choice of parameters will eliminate or minimize covariance, making results easier to interpret. An obvious case is the slope and intercept of a linear fit. These are usually highly correlated, but the covariance can be eliminated by defining the intercept at an appropriate "pivot point," fitting $y = a(x - x_p) + b$ instead of $y = ax + b$.

Suppose we have fit the slope and intrinsic scatter of a relation as in the previous section (and we for some reason know the intercept without fitting).

What if we are only interested in the slope and its errors, and σ is something we have to include in the fit but don't care about?

In this case, σ is what is called a "nuisance parameter," and we get rid of it by integrating over its probability distribution, a procedure called "marginalization."

$$p(a|DI) = \int_0^\infty d\sigma p(a,\sigma|DI).$$

This procedure can be used to go from any number of parameters to any smaller subspace, e.g., if we included b in our fits:

$$\begin{aligned} p(ab|DI) &= \int_0^\infty d\sigma p(ab\sigma|DI)d\sigma \\ p(a|DI) &= \int_{-\infty}^\infty db \int_0^\infty d\sigma p(ab\sigma|DI). \end{aligned}$$

Hypothesis Comparison

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.

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 $\mu_{\rm min} \leq \mu \leq \mu_{\rm max}$, 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?

$$P(D|H_1I) = K \times \exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu_1)^2\right],$$

where

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

is independent of μ_1.

$$\begin{aligned} p(D|H_2I) &= \int_{\mu_{\rm min}}^{\mu_{\rm max}} p(D|\mu I)p(\mu|I)d\mu \\ &= K(\mu_{\rm max}-\mu_{\rm min})^{-1}\int_{\mu_{\rm min}}^{\mu_{\rm max}} d\mu \exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\right]. \end{aligned}$$

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

$$\frac{p(D|H_1I)}{p(D|H_2I)} = \left(\frac{2\pi\sigma^2}{N}\right)^{-1/2}(\mu_{\rm max}-\mu_{\rm min})\exp\left[-\frac{N}{2\sigma^2}(\overline{x}-\mu_1)^2\right].$$

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.

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 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.

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 the case of linear models, where the y_i are linear functions of the model parameters (but not necessarily of the independent variables x_i, so $y_i = ax_i + bx_i^2$ is a linear model). Adding N parameters to a linear model that already adequately describes the data reduces the expected value of χ^2 by N.

This reduction in $\langle\chi^2\rangle$ applies whether or not the observational errors are Gaussian, but the variance of χ^2 is different (generally higher) if the errors are non-Gaussian, so the values of χ^2 that correspond to different levels of "absolute" goodness-of-fit are different.

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

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

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.

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.

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, a model that predicts a mean of 0.001 objects 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).

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.

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:

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

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.

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,

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

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.

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

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.

Whichever approach one uses, 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 (see Andy's lectures), noisy estimates of the covariance matrix can cause misleading conclusions about best-fit parameter values, parameter uncertainties, or relative merit of models.

Monte Carlo Markov Chains

A fairly common statistical problem is estimating the probability distribution of parameters in a high-dimensional parameter space.

For example, we might be trying to determine the constraints from a CMB data set D on the set of cosmological parameters $\vec{\theta} = (\Omega_m, h, \Omega_b, A, n, \tau)$ that determines the CMB spectrum in the simplest current cosmological scenario.

There are tools for calculating $p(D|\vec{\theta}I)$, but this calculation might take a few seconds, or minutes, for each model in the parameter space.

Since the parameter space is six-dimensional, even a relatively coarse grid with 10 points along each parameter direction over the plausible range requires 10^6 evaluations of $p(D|\vec{\theta}I)$, and if we add another two parameters then 10^6 becomes 10^8.

Thus, a naive grid-based evaluation of the likelihood to find best-fit parameters and error bars may be prohibitively expensive.

Monte Carlo Markov Chains (MCMC) are a useful tool for this kind of problem, and this approach has taken rapid hold in the cosmology literature.

For details, see the references listed at the top and the things that they in turn refer to, but in brief the idea is as follows.

The goal is to map the posterior probability distribution $p(\vec{\theta}|DI) \propto p(\vec{\theta}|I)p(D|\vec{\theta}I)$, in the neighborhood of its maximum value.

If the prior $p(\vec{\theta}|I)$ is flat, then we just have $p(\vec{\theta}|DI) \propto \mathcal{L}$.

Procedure:

1. Start from a randomly chosen point in the parameter space, $\vec{\theta} = \vec{\alpha}_1$.

2. Take a random step to a new position $\vec{\alpha}_2$.

3. If $p(\vec{\alpha}_2|DI) \geq p(\vec{\alpha}_1|DI)$, "accept" the step: add $\vec{\alpha}_2$ to the chain, and substitute $\vec{\alpha}_2 \rightarrow \vec{\alpha}_1$. Return to step 2.

4. If $p(\vec{\alpha}_2|DI) < p(\vec{\alpha}_1|DI)$, draw a random number x from a uniform distribution from 0 to 1. If $x < p(\vec{\alpha}_2|DI)/p(\vec{\alpha}_1|DI)$, "accept" the step and proceed as in 3. If $x \geq p(\vec{\alpha}_2|DI)/p(\vec{\alpha}_1|DI)$, reject the step. Save $\vec{\alpha}_1$ as another link on the chain, and go back to 2.

The chain takes some time to "burn in," i.e., to reach the neighborhood of the most likely solutions.

However, once this happens, a "long enough" chain will have a density of points that is proportional to $p(\vec{\theta}|DI)$.

To get, for example, the joint pdf of a pair of parameters, one can just make contours of the density of points in the space of those two parameters. Other "nuisance" parameters are marginalized over automatically, because the points sample the full space.

If you want to calculate the posterior distribution of some *function* of the parameters (e.g., the age of the Universe, given parameter estimates from the CMB), you can just calculate that function for all points in the chain, then plot the pdf of the result.

There are numerous technical issues related to determining whether a chain has "converged" (i.e., is fairly sampling the probability distribution), and to choosing steps in a way that produces fast convergence and good "mixing" (sampling the distribution fairly with a relatively small number of points).

These issues are discussed briefly in the references provided, and I gather that they are the subject of an extensive literature.

But with a few basic tricks described in these references, an effective version of MCMC is relatively straightforward to implement.