Astronomy 8824: Statistics Notes 1
Some High-Level Background

Reading: Chapter 3 of Ivezic et al. See the *Reader's Guide* on the course web page for outline and advice on where to focus your attention.

Statistical Tasks in Astrophysics

Four common statistical tasks:

Parameter estimation

Comparison of hypotheses

"Absolute" evaluation of a hypothesis

Forecasting of errors

Another task, slightly less common: Prediction of values from a model fit to some set of data, when the parameters of the model are uncertain.

Simple example: Data points with error bars.

Parameter estimation: What are slope and amplitude of a power-law fit? What are the uncertainties in the parameters?

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 future cosmological surveys and CMB experiments, for example.

PDF, Mean, Variance

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.

If random variables x and y are independent, then $p(x,y) = p(x)p(y)$ and

$$\mathrm{Cov}(x,y) \equiv \langle (x-\mu_x)(y-\mu_y)\rangle = 0.$$

The second statement can be proved from the first.

Estimators

An estimator is a mathematical function of data that estimates a quantity of interest.

Ideally one wants an estimator to be

unbiased – even with a small amount of data, the expectation value of estimator is equal to the quantity being estimated

efficient – makes good use of the data, giving a low variance about the true value of the quantity

robust – isn't easily thrown off by data that violate your assumptions about the pdf, e.g., by non-Gaussian tails of the error distribution

consistent – in the limit of lots of data, it converges to the true value

These four desiderata sometimes pull in different directions.

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).

Snap-judging Error Bars

What is wrong with this plot?

Bayesian vs. Frequentist 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* (though they are drawn from a pdf).

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 $\overline{A}$ means "not A" 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*

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

Bayesian Inference

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. The background information I may specify assumptions like a 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.

Global Likelihood and Absolute Assessment

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.

For example, the marginal distribution of a parameter x

$$p(x) = \int p(x|a, b, c) da\, db\, dc$$

can only exist if x is a random variable.