Statistics:

Model Hypothesis Testing

Aneta Siemiginowska

Harvard-Smithsonian Center for Astrophysics

Steps in the X-ray Data Analysis:

1/ Obtain the data (observe or archive)

2/ Reduce Data => standard processing or reprocessed,
extract an image or a spectrum

3/ Analysis – fit the data

4/ Conclude – Hypothesis Testing!

5/ Reflect

How do we compare the two different models?

Steps in Hypothesis Testing

1/ Set up 2 possible exclusive hypotheses:

M0 – null hypothesis – formulated to be rejected

M1 – an alternative hypothesis, research hypothesis

each has associated terminal action

2/ Specify a priori the significance level α

choose a test which:

- approximates the conditions
- finds what is needed to obtain the sampling distribution and the region of rejection, whose area is a fraction of the total area in the sampling distribution

3/ Run test: reject ***M0*** if the test yields a value of the statistics whose probability of occurance under ***M0*** is $<\alpha$

4/ Carry on terminal action

A model ***M*** has been fit to dataset ***D*** :

- » the maximum of the likelihood function L_{max},
- » the minimum of the χ^2 statistic χ^2_{min},
- » or the mode of the posterior distribution $p(\hat{\theta} \mid D)$

Model Comparison. The determination of which of a suite of models (***e.g.,*** blackbody, power-law, ***etc.***) best represents the data.

Parameter Estimation. The characterization of the sampling distribution for each best-fit model parameter (***e.g.,*** blackbody temperature and normalization), which allows the errors (***i.e.,*** standard deviations) of each parameter to be determined.

STEPS AGAIN

Two models, M_0 and M_1, have been fit to ***D.*** M_0, the "simpler" of the two models (generally speaking, the model with fewer free parameters) is the ***null hypothesis.***

A frequentist would compare these models by:

- constructing a test statistic ***T*** from the best-fit statistics of each fit (***e.g.,*** $\Delta\chi^2 = \chi_0^2 - \chi_1^2$);
- determining each sampling distributions for T, $p(T \mid M_0)$ and $p(T \mid M_1)$;
- determining the ***significance***, or Type I error, the probability of selecting M_1 when M_0 is correct:

$$\alpha = \int_{T_{\rm obs}}^{\infty} dT p(T \mid M_0);$$

- and determing the ***power***, or Type II error, which is related to the probability β of selecting M_0 when M_1 is correct:

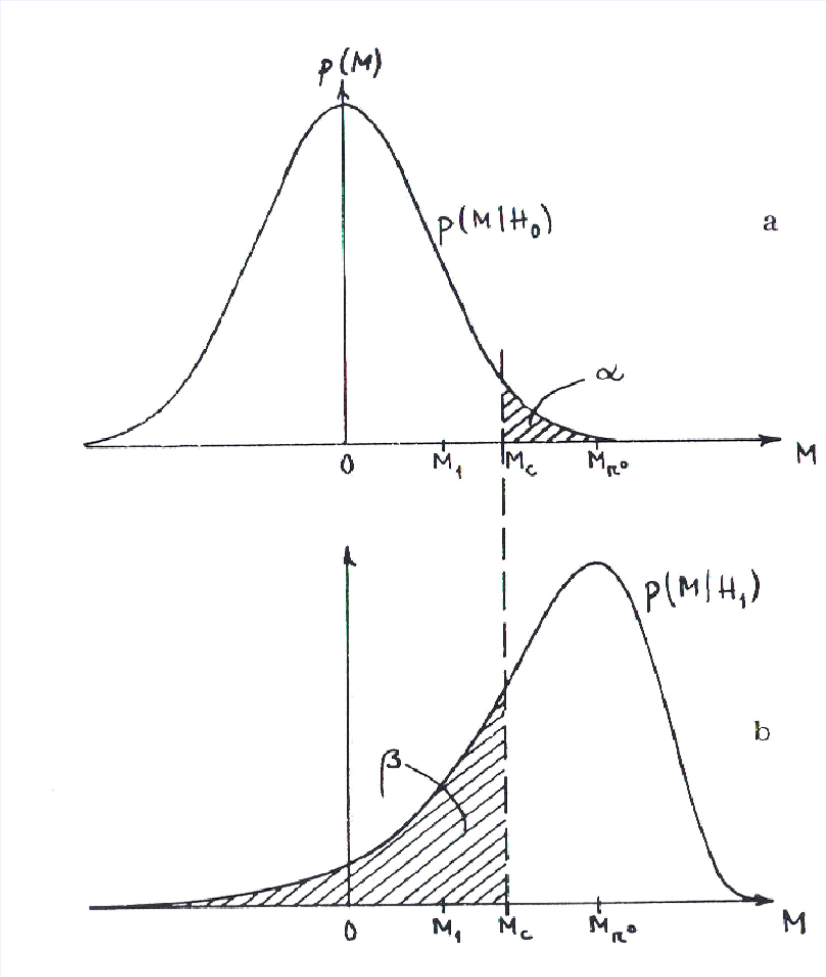
$$1 - \beta = \int_{T_{\rm obs}}^{\infty} dT p(T \mid M_1).$$

⇒ If α is smaller than a pre-defined threshold (≤ 0.05, or $\leq 10^{-4}$, ***etc.***, with smaller thresholds used for more controversial alternative models), then the frequentist rejects the null hypothesis.

⇒ If there are several model comparison tests to choose from, the frequentist uses the most powerful one!

α- significance

1- β – power of test



Comparison of distributions ***p***(***T*** | ***M₀***) (from which one determines the significance ***α***) and ***p***(***T*** | ***M₁***) (from which one determines the power of the model comparison test 1 – ***β***) (Eadie ***et al.*** 1971, p.217)

Frequentist Model Comparison

Standard frequentist model comparison tests include:

The χ^2 Goodness-of-Fit (GoF) test:

$$\alpha_{\chi^2} = \int_{\chi^2_{\min,0}}^{\infty} d\chi^2 \, p(\chi^2 \,|\, N - P_0) = \frac{1}{2\Gamma(\frac{N-P_0}{2})} \int_{\chi^2_{\min,0}}^{\infty} d\chi^2 \left(\frac{\chi^2}{2}\right)^{\frac{N-P_0}{2}-1} e^{-\frac{\chi^2}{2}} .$$

The Maximum Likelihood Ratio (MLR) test:

$$\alpha_{\chi^2 MLR} = \int_{\Delta\chi^2}^{\infty} d\chi^2 \, p(\Delta\chi^2 \,|\, \Delta P),$$

where ΔP is the number of additional freely varying model parameters in model $\boldsymbol{M}_1$

The F-test: $$F = \frac{\Delta\chi^2}{\Delta P} \Big/ \frac{\chi_1^2}{(N-P_1)} .$$

where $\boldsymbol{P}_1$ is the total number of thawed parameters in model $\boldsymbol{M}_1$

These are standard tests because they allow estimation of the significance without time-consuming simulations!

Model Comparison Tests:

Notes and caveats regarding these standard tests:

- The GoF test is an “alternative-free” test, as it does not take into account the alternative model $\boldsymbol{M}_1$. It is consequently a ***weak*** (***i.e.,*** not powerful) model comparison test and should not be used!
- Only the version of ***F***–test which generally has the greatest power is shown above: in principle, one can construct three ***F*** statistics out of $\chi_0^2, \chi_1^2, \text{and } \Delta\chi^2$
- The MLR ratio test is generally the most powerful for detecting emission and absorption lines in spectra.

But the most important caveat of all is that…

The *F* and *MLR* tests are commonly misused by astronomers!

There are two important conditions that must be met so that an estimated derived value ***α*** is actually correct, ***i.e.,*** so that it is an accurate approximation of the tail integral of the sampling distribution (Protassov ***et al.*** 2001):

- ***M_0 must be nested within M_1, i.e.,*** one can obtain M_0 by setting the extra ΔP parameters of M_1 to default values, often zero; and
- ***those default values may not be on a parameter space boundary.***

The second condition may not be met, ***e.g.,*** when one is attempting to detect an emission line, whose default amplitude is zero and whose minimum amplitude is zero. Protassov ***et al.*** recommend Bayesian posterior predictive probability values as an alternative,

---

If the conditions for using these tests are not met, then they can still be used, but the significance must be computed via Monte Carlo simulations.

Bayesian Model Comparison

we showed how Bayes' theorem is applied in model fits. It can also be applied to model comparison:

$$p(M \mid D) = p(M)\,\frac{p(D \mid M)}{p(D)}.$$

p(***M***) is the prior probability for ***M***;

p(***D***) is an ignorable normalization constant; and

p(***D*** | ***M***) is the average, or global, likelihood:

$$p(D \mid M) = \int d\theta\, p(\theta \mid M)\, p(D \mid M,\theta)$$

$$= \int d\theta\, p(\theta \mid M)\, L(M,\theta).$$

In other words, it is the (normalized) integral of the posterior distribution over all parameter space. Note that this integral may be computed numerically, by brute force, or if the likelihood surface is approximately a multi-dimensional Gaussian ($i.e.$ if $L \;\alpha\; \exp[-\chi^2/2]$), by the ***Laplace approximation:***

$$p(D \mid M) = p(\hat{\theta} \mid M)(2\pi)^{P/2}\sqrt{\det C}\, L_{\max},$$

where C is the covariance matrix (estimated numerically at the mode).

Bayesian Model Comparison

To compare two models, a Bayesian computes the odds, or odd ratio:

$$\begin{aligned} O_{10} &= \frac{p(M_1 \mid D)}{p(M_0 \mid D)} \\ &= \frac{p(M_1)\, p(D \mid M_1)}{p(M_0)\, p(D \mid M_0)} \\ &= \frac{p(M_1)}{p(M_0)} B_{10}\,, \end{aligned}$$

where ***B***$_{10}$ is the ***Bayes factor***. When there is no ***a priori*** preference for either model, ***B***$_{10}$ = 1 of one indicates that each model is equally likely to be correct, while ***B***$_{10}$ ≥ 10 may be considered sufficient to accept the alternative model (although that number should be greater if the alternative model is controversial).