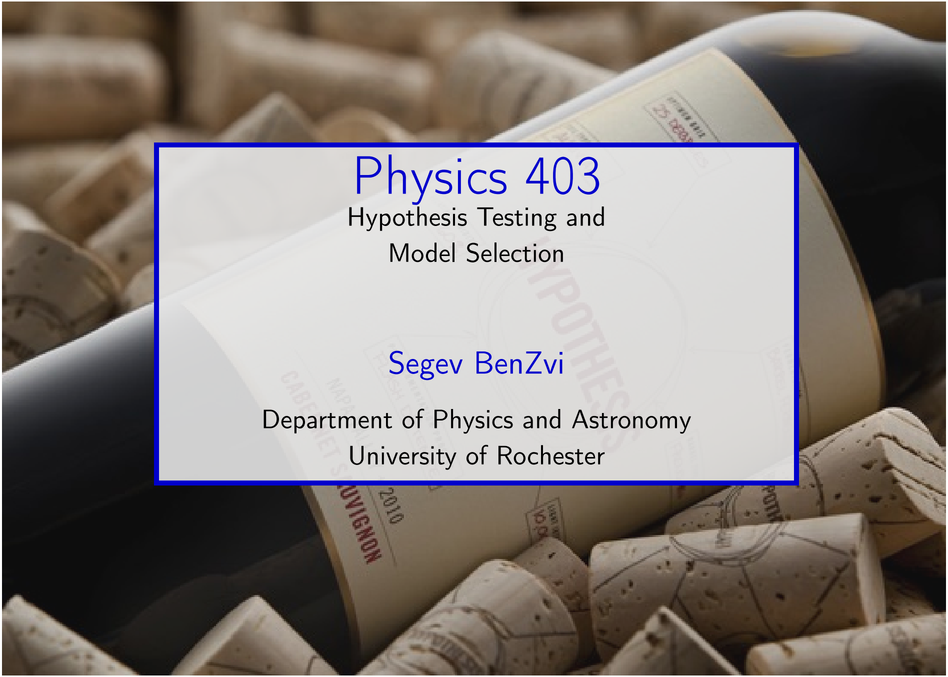


Physics 403

Hypothesis Testing and Model Selection

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

Table of Contents

1. Review of Systematic Effects
2. Model Selection
   - Hypothesis Testing
   - Posterior Odds
   - The Ockham Factor
   - Comparing Several Models with Free Parameters
   - Hypothesis Testing vs. Parameter Estimation
   - Classical Statistics: Type I and Type II Errors
   - Classical Statistics: Neyman-Pearson Lemma
   - Using p-Values

Last Time: Systematic Uncertainties

Bayesian approach to systematic uncertainties: marginalize them

Frequentist approach: different terminology used ("nuisance parameters") but the approach is a quasi-Bayesian marginalization

To propagate systematic uncertainties, there are three methods typically used:

1. Monte Carlo: simulate your analysis by generating different random values for your nuisance parameters. Very popular technique
2. Covariance Method: add systematics as common covariance terms to your error matrix, carry out ML/LS method. Perfectly correct, not typically done
3. Pull Method: calculate pull distributions for physical parameters and nuisance parameters. Very common, because it tells you which parameters contribute most to your error budget

Table of Contents

1 Review of Systematic Effects

2 Model Selection
- Hypothesis Testing
- Posterior Odds
- The Ockham Factor
- Comparing Several Models with Free Parameters
- Hypothesis Testing vs. Parameter Estimation
- Classical Statistics: Type I and Type II Errors
- Classical Statistics: Neyman-Pearson Lemma
- Using p-Values

Model Selection

- For the past month we have discussed parameter estimation, which gets us the "best estimate" of a model parameter given some measurement
- In today's class we will cover the topic of model selection, also known as hypothesis testing
- In model selection, you don't find a best fit parameter given a model. Rather, you test whether or not the model is itself a good fit to the data
- While the question you are asking of the data is different, the techniques used for parameter estimation and model selection are essentially identical (at least in the Bayesian framework)
- As usual, we don't evaluate a hypothesis or model in isolation, but in the context of several competing and sometimes mutually exclusive models. You'll see how this works with some simple examples, but it's pretty intuitive

Hypothesis Testing

A cute framing device used in Sivia [1]:

> *Mr. A has a theory; Mr. B also has a theory, but with an adjustable parameter λ. Whose theory should we prefer on the basis of data D?*

Example

Suppose D represents noisy measurements y as a function of x.

- Mr. A: the data are described by $y = 0$
- Mr. B: the data are described by $y = a$, with $a =$ constant
- Mr. C: the data are described by $y = a + bx$
- Mr. F: the data are described by $y = a + bx + cx^2 + dx^3 + \dots$

Are the data best fit by a constant? A line? A high-order polynomial? How do we choose?

Posterior Odds Ratio

- As in parameter estimation, we choose between two models or hypotheses using the ratio of posterior PDFs

$$\text{posterior ratio} = O_{AB} = \frac{p(A|D,I)}{p(B|D,I)}$$

- Recall the criteria for making a decision about which model to favor [2]

O_{AB}	**Strength of Evidence**
$< 1:1$	negative (supports B)
$1:1$ to $3:1$	barely worth mentioning
$3:1$ to $10:1$	substantial support for A
$10:1$ to $30:1$	strong support for A
$30:1$ to $100:1$	very strong support for A
$> 100:1$	decisive evidence for A

The Bayes Factor and Prior Odds

- Applying Bayes' Theorem to the numerator and denominator of the odds ratio gives

$$O_{AB} = \frac{p(A|D,I)}{p(B|D,I)} = \frac{p(D|A,I)}{p(D|B,I)} \times \frac{p(A|I)}{p(B|I)}$$

where the normalizing term $p(D|I)$ cancels out

- Recall that the likelihood ratio is called the Bayes Factor.
- The second term is the prior odds ratio. It describes how much you favor model A over B before taking data
- Normally one might like to treat the models in an unbiased manner and set $p(A|I) = p(B|I)$, so that the odds ratio is completely given by the likelihood ratio (or "Bayes Factor"). But can you think of any situations where this might not be the case?

When to use Nontrivial Prior Odds

Example

You are conducting a medical trial to determine if a treatment is effective. A says it's effective; B says it's ineffective but otherwise harmless, i.e., $B = \overline{A}$. It might be both ethical and economical to set $p(A|I) > p(B|I)$.

Example

You are a particle physicist looking for new physics, e.g., a signature of supersymmetry, with A saying the new physics is real and B saying it's not ($B = \overline{A}$). The outcome of a false claim supporting A could be harmful – colleagues' time wasted on analysis or designing new experiments, public embarrassment for the field, etc. – so you might be justified starting your experiment with the prior belief $p(A|I) < p(B|I)$, or perhaps even $p(A|I) \ll p(B|I)$.

Computing the Likelihood Ratio

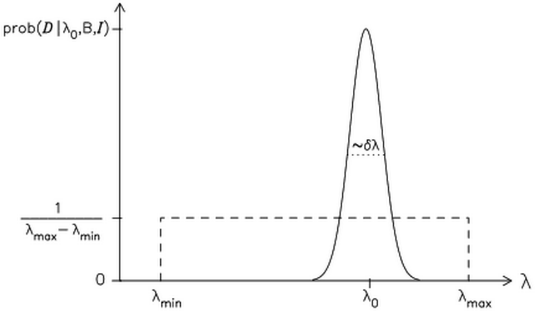
- Let's get back to the original problem of Mr. A and Mr. B, where B proposal a model with an adjustable parameter λ
- Since λ is adjustable and unknown *a priori* we marginalize the likelihood $p(D|B,I)$:

$$p(D|B,I) = \int p(D,\lambda|B,I)\ d\lambda = \int p(D|\lambda,B,I)\ p(\lambda|B,I)\ d\lambda$$

- The first term is an ordinary likelihood function parameterized in terms of λ
- The second term contains any prior knowledge about λ
- It is the responsibility of Mr. B to provide some PDF describing the state of knowledge of λ. As usual for priors, it could be a previous measurement, a theoretical calculation, or a personal opinion (hopefully well-motivated)

Computing the Marginal Likelihood

- Suppose that B can only say that $\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]$. In this case
$$p(\lambda|B, I) = \frac{1}{\lambda_{\text{max}} - \lambda_{\text{min}}}$$
for λ inside the limits and zero otherwise
- Also suppose there is a best value $\hat{\lambda}$ (or λ_0) that yields the closest agreement with the measurements, such that $p(D|\hat{\lambda}, B, I)$ is a maximum there



Combining the Likelihood and Prior for B

- Without much loss of generality, let's assume that $p(D|\lambda, B, I)$ is approximately Gaussian for $\lambda = \hat{\lambda} \pm \delta\lambda$:

$$p(D|\lambda, B, I) = p(D|\hat{\lambda}, B, I) \times \exp\left[-\frac{(\lambda - \hat{\lambda})^2}{2\ \delta\lambda^2}\right]$$

- Since the prior does not depend on λ, the marginal likelihood of B is

$$p(D|B, I) = \frac{1}{\lambda_{\text{max}} - \lambda_{\text{min}}} \int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} p(D|\lambda, B, I)\ d\lambda$$

- As long as the limits of integration do not significantly truncate the Gaussian in λ, the integral is approximately

$$\int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} p(D|\lambda, B, I)\ d\lambda \approx p(D|\hat{\lambda}, B, I) \times \delta\lambda\sqrt{2\pi}$$

Combining the Likelihood and Prior for B

- Putting all the pieces together, the odds ratio of A and B is

$$O_{AB} = \frac{p(A|I)}{p(B|I)} \, \frac{p(D|A,I)}{p(D|\hat{\lambda},B,I)} \, \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\delta\lambda\sqrt{2\pi}}$$

- **First term**: the usual prior odds ratio
- **Second term**: the likelihood ratio or Bayes factor. Because λ is an adjustable parameter we expect this term will definitely favor B over A
- **Third term**: the Ockham (or Occam) factor. We expect that $\lambda_{\text{max}} - \lambda_{\text{min}}$ will be larger than the small range $\delta\lambda$ allowed by the data, so this term favors A over B
- The Ockham factor penalizes over-constrained fits:

  It is vain to do with more what can be done with fewer

Comments about the Uniform Prior

- Issue: isn't it a problem if λ_{min} and λ_{max} are allowed to go to $\pm\infty$?
- In this case there would be an infinite penalty on model B and we would never favor it, no matter what the data say
- In practice this pretty much never happens; claiming absolute ignorance is just not realistic and wilfully ignores lots of physical insight

Example

Suppose we are looking for deviations of Newtons Law of Gravitation in the form

$$\frac{1}{r^2} \rightarrow \frac{1}{r^{2+\epsilon}}$$

We would never claim a prior on ϵ of $\pm\infty$. From below we expect $\epsilon > 0$, and from above we know that $\epsilon \ll 2$; if it weren't we would have already observed a large effect

Results Dominated by the Priors or the Ockham Factor

- In pretty much every decent experiment you tend to be in a situation where the data (in the form of the Bayes Factor) dominates the prior odds
- The Ockham factor becomes important if model B does not give a much better result with more data. In this case $\delta\lambda$ becomes increasingly narrow, leading to bigger and bigger penalites against B
- This does not happen when the data are of bad quality, or irrelevant, or you have low statistics. I.e., you've designed a bad experiment for the physics you are trying to accomplish
- If the the data are poor then you expect

$$\delta\lambda \gg \lambda_{\text{max}} - \lambda_{\text{min}}$$

$$p(D|\hat{\lambda}, B, I) \approx p(D|A, I)$$

$$O_{AB} \approx \frac{p(A|I)}{p(B|I)}$$

Two Models with Free Parameters

- Let's add a complication and suppose that A also has an adjustable parameter μ. For example, A could predict a Gaussian peak and B a Lorentzian peak, and λ and μ are the FWHM of the predictions
- In this case the posterior odds ratio is

$$O_{AB} = \frac{p(A|D,I)}{p(B|D,I)} = \frac{p(A|I)}{p(B|I)} \times \frac{p(D|\hat{\mu},A,I)}{p(D|\hat{\lambda},B,I)} \times \frac{\delta\mu(\lambda_{\text{max}} - \lambda_{\text{min}})}{\delta\lambda(\mu_{\text{max}} - \mu_{\text{min}})}$$

- If we set $p(A|I) = p(B|I)$ and choose a similar prior range for μ and λ, then

$$O_{AB} \approx \frac{p(D|\hat{\mu},A,I)}{p(D|\hat{\lambda},B,I)} \times \frac{\delta\mu}{\delta\lambda}$$

- For data of good quality, the best-fit likelihood ratio dominates. But, if both models give similar agreement with the data then the one with the larger error bar $\delta\mu$ or $\delta\theta$ will be favored
- Wait, **what**? How can the less discriminating theory do better? In the context of model selection, a larger uncertainty means that more parameter values are consistent with a given hypothesis

Two Models with Free Parameters

- There is another case: A and B have the same physical theory but different prior ranges on μ and λ
- In this case, we imagine that A and B set limits that are large enough that they incorporate all parameter values fitting reasonably to the data
- Assuming equal *a priori* weighting towards A and B, the odds ratio is

$$O_{AB} = \frac{p(A|D,I)}{p(B|D,I)} = \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\mu_{\text{max}} - \mu_{\text{min}}}$$

  because we expect $\hat{\lambda} = \hat{\theta}$ and $\delta\lambda = \delta\mu$
- The analysis will support the model with a narrow prior range, which it should if B has a good reason to predict the value of his parameter mor accurately than A

Comparison with Parameter Estimation

- Note how this differs from parameter estimation, where we assume that a model is correct and calculate the best parameter given that model
- To infer the value of λ from the data, given that B is correct, we write

$$p(\lambda|D,B,I) = \frac{p(D|\lambda,B,I)\ p(\lambda|B,I)}{p(D|B,I)}$$

- To estimate λ we want to **maximize the likelihood** over the range $[\lambda_{\text{min}}, \lambda_{\text{max}}]$. As long as the range contains enough of $p(D|\lambda,B,I)$ around $\hat{\lambda}$, its particular bounds do not matter for finding $\hat{\lambda}$
- To calculate the odds ratio of A and B we are basically comparing the **likelihoods averaged over the parameter space**
- Therefore, in model selection the Ockham factor matters because there is a cost to averaging the likelihood over a larger parameter space

Hypothesis Testing

- You have seen that parameter estimation and model selection are quite similar; we are just asking different questions of the data
- In model selection we calculate the probability that some hypothesis H_0 is true, starting from Bayes' Theorem:

$$p(H_0|D,I) = \frac{p(D|H_0,I)\ p(H_0|I)}{p(D|I)}$$

- The marginal evidence $p(D|I)$ can be ignored if we are calculating the odds ratio of H_0 with some other hypothesis H_1
- If we actually want to know $p(H_0|D,I)$ we need to calculate $p(D|I)$. This requires the alternative hypothesis. Using marginalization and the produce rule,

$$p(D|I) = p(D|H_0,I)\ p(H_0|I) + p(D|H_1,I)\ p(H_1|I)$$

Hypothesis Testing

- It's very nice when the alternative hypothesis and H_0 completely exhaust all the possibilities, i.e., $H_1 = \overline{H_0}$. However, this need not be the case

Example

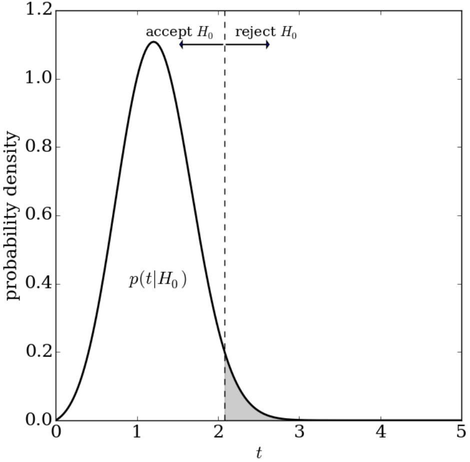
Suppose we're looking for a peak in some data. H_0 could be "the shape of the peak is Gaussian," and H_1 could be "the shape of the peak is Lorentzian."

Clearly $H_1 \neq \overline{H_0}$, but we can still define $p(H_0|D, I)$ using the specific set of possibilities $\{H_0, H_1\}$.

Still, defining a generic alternative hypothesis $H_1 = \overline{H_0}$ is possible if we're willing to work hard at it. Consider the example of binned data where the expected count λ_i in bin i is given by a flat backround and Gaussian signal in H_0. What could $\overline{H_0}$ look like?

Hypothesis Testing in Classical Statistics

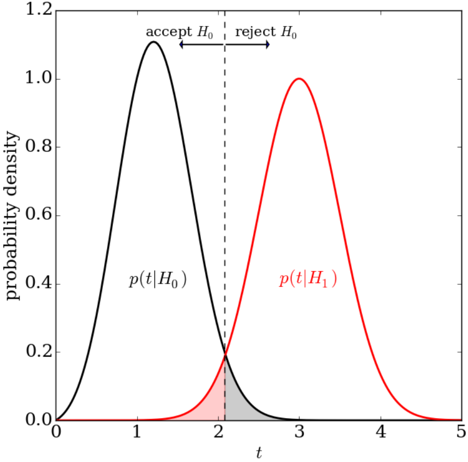
Type I Errors



- Construct a test statistic t and use its value to decide whether to accept or reject a hypothesis
- The statistic t is basically a summary of the data given the hypothesis we want to test
- Define a cut value t_{cut} and use that to accept or reject the hypothesis H_0 depending on the value of t measured in data
- Type I Error: reject H_0 even though it is true with tail probability α (shown in gray)

Hypothesis Testing in Classical Statistics

Type II Errors



- You can also specify an alternative hypothesis H_1 and use t to test if it's true
- Type II Error: accept H_0 even though it is false and H_1 is true. This tail probability β is shown in pink

$$\alpha = \int_{t_{\text{cut}}}^{\infty} p(t|H_0)\ dt$$

$$\beta = \int_{-\infty}^{t_{\text{cut}}} p(t|H_1)\ dt$$

Statistical Significance and Power

- As you can see there is some tension between α and β. Increasing t_{cut} will increase β and reduce α, and vice-versa
- Significance: α gives the significance of a test. When α is small we disfavor H_0, known as the **null hypothesis**
- Power: $1 - \beta$ is called the power of a test. A powerful test has a small chance of wrongly accepting H_0

Example

It's useful to think of the null hypothesis H_0 as a less interesting default/status quo result (your data contain only background) and H_1 as a potential discovery (your data contain signal). A good test will have high significance and high power, since this means a low chance of incorrectly claiming a discovery and a low chance of missing an important discovery.

The Neyman-Pearson Lemma

The Neyman-Pearson Lemma is used to balance signifance and power. It states that the acceptance region giving the highest power (and hence the highest signal "purity") for a given significance level α (or selection efficiency $1 - \alpha$) is the region of t-space such that

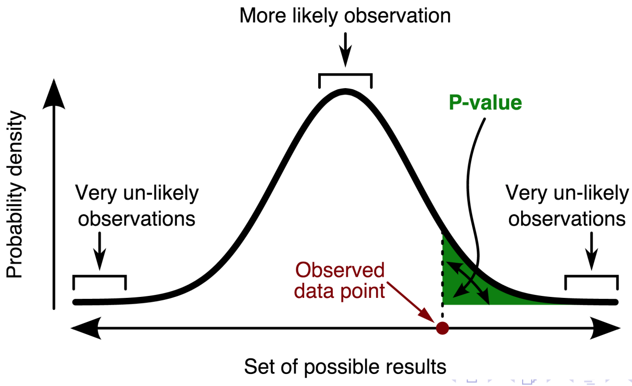
$$\Lambda(\boldsymbol{t}) = \frac{p(\boldsymbol{t}|H_0)}{p(\boldsymbol{t}|H_1)} > c$$

Here $\Lambda(t)$ is the likelihood ratio of the test statistic $\boldsymbol{t}$ under the two hypotheses H_0 and H_1. The constant c is determined by α. Note that $\boldsymbol{t}$ can be multidimensional.

In practice, one often estimates the distribution of $\Lambda(t)$ using Monte Carlo by generating $\boldsymbol{t}$ according to H_0 and H_1. Then use the distribution to determine the cut c that will give you the desired significance α.

Hypothesis Testing in Classical Statistics: χ^2 p-Value

- We have already seen a bit of model selection when discussing the goodness of fit provided by the χ^2 statistic
- If a model is correct, and the data are subject to Gaussian noise, then we expect $\chi^2 \approx N$. Deviations from the expectation by more than a few times $\sqrt{2N}$ would be surprising
- So, should we reject a hypothesis if χ^2 is too large?



Hypothesis Testing in Classical Statistics

- When we calculate a χ^2 probability, we are calculating a one-sided *p*-value:

$$\int_{\chi^2_{\text{obs}}}^{\infty} p(\chi^2|N, H_0, I)\ d\chi^2$$

- There is an assumption baked into this *p*-value; it assumes that H_0 is true by definition
- To test a theory, we need the posterior probability $p(H_0|D, I)$, not $p(D|H_0, I)$. So we are missing $p(H_0|I)$ and $p(D|I)$
- While rejecting H_0 on the basis of a small *p*-value can be done, it's risky because we are only testing the probability that the data fluctuated away from the predictions of the model H_0, not the probability that H_0 is correct given the data
- **Consquence**: using a *p*-value can overstate the evidence against H_0, leading to a Type-I error – the rejection of H_0 when it is true

Summary

- Hypothesis testing and parameter estimation are quite similar in terms of the calculations we need to do, but they ask different things of the data
- Parameter estimation: we use the maximum likelihood. Hypothesis testing: we use the average likelihood
- Frequentist approach is to minimize Type I errors (rejecting a true H_0) and Type II errors (rejecting a true H_1) using a likelihood ratio test. This is justified by the Neyman-Pearson lemma
- A p-value and a Type I error rate α are not the same thing
- If you use a p-value to choose between two hypotheses, you're asking for trouble unless you demand very strong evidence against the null hypothesis

References I

[1] D.S. Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press, 1998.

[2] Harold Jeffreys. *The Theory of Probability*. 3rd ed. Oxford, 1961.