Hypothesis Tests

Aneta Siemiginowska
Harvard-Smithsonian Center for Astrophysics
CHASC

April 8, 2013

This is one of many statistical topics that was important to Alanna

Outline

- Hypothesis testing: motivation and basic framework
- Methods:
  - P-values
  - Bayesian posterior predictive p-value
  - Bayes Factors
- Conclusions Further Reading

Based on David Van Dyk talk at the 2008 HEAD meeting in Los Angeles

Motivation and Basic Framework

- Last step in data modeling:
  - Choose between a simpler and a more complex model, e.g. addition of an emission or absorption line to the continuum
  - discriminate between models, e.g. power law or thermal emission
- Framework:
  - The Null Hypothesis:

    H0: the line does not exist in the data
  - The Alternative Hypothesis:

    H1: the line exists in the data

The Null is a special case of the Alternative => Line intensity equal zero.

Methods: p-values

- Assuming the Null hypothesis is true, how likely do we see the test statistics, T, as extreme or more extreme than the observed value, t_{obs}?

$$Pr(T \geq t_{obs} | H_0) = p_{\rm value}$$

- Although p-values are very popular for model selections, they raise important challenges:
  - They are often based on appropriate asymptotic results
  - They can bias inference in the direction of false detection

T - Test statistics

Methods: Test Statistics

- **Test Statistics:**

Likelihood Ratio: $$R = \frac{L(\theta_0|Y)}{L(\theta_1|Y)}$$

L - likelihood

θ_0 - fitted MLE of null model parameters

θ_1 - fitted MLE of alternative model parameters

Y - data

Important!

The distribution -2log(R) under H_0 approaches $\chi^2_{(d\text{-}do)}$ as the data sample size increases under certain assumptions.

Methods: LRT

- Assumptions of the Likelihood Ratio Test statistics:
  - The null hypothesis must be a special case of the alternative
  - The parameter space of the null must be interior of the alternative parameter space.
- The second assumption fails when testing for a spectral emission line:
  - When there is no line, the line intensity is zero, it may not be negative.
  - The line locations and width of the line do not exist when there is no line. They have no values.

See Protassov et al (2002) for details

Methods: LRT - distribution

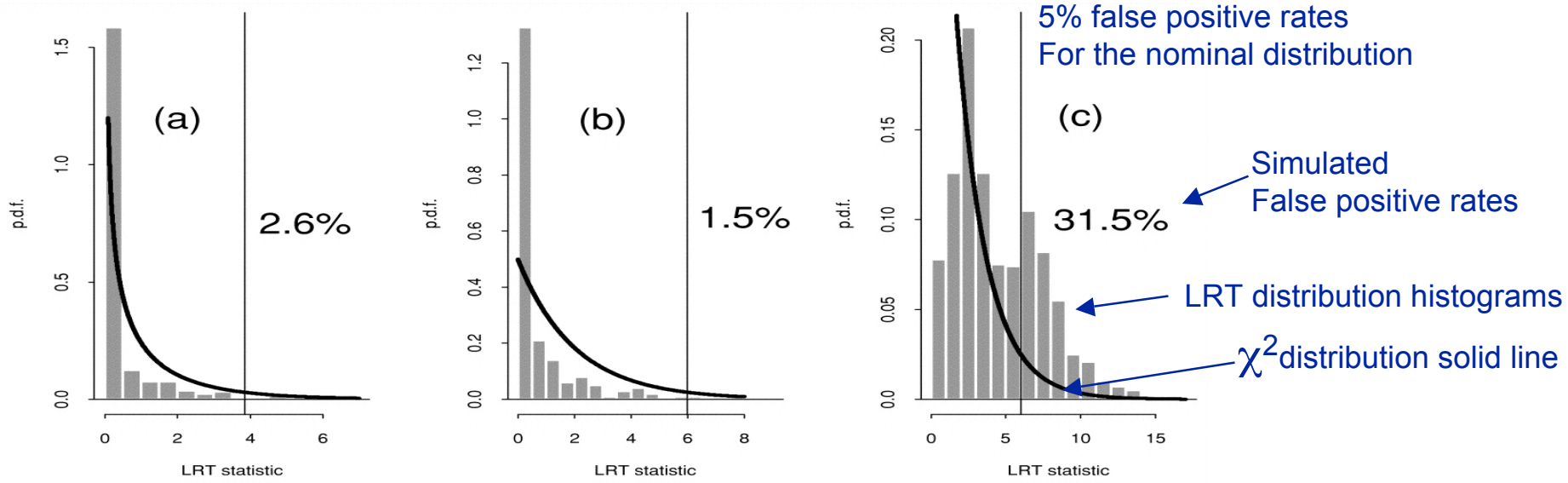
IMPORTANT! We do not know the true distribution of the test statistics.

Narrow line
at fixed location

Unknown location of the line
Fixed width

Absorption line

Protassov et al. (2002)

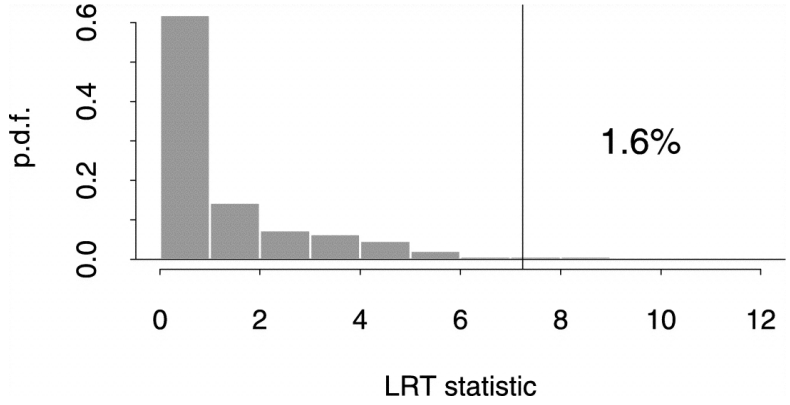


- Results of three tests compared to the nominal χ^2 distribution

Methods: Monte Carlo Calibration

- Instead of using the fixed best fit parameter values run Monte Carlo simulations to access the sampling distribution of the LRT (or other Test Statistics). This will calibrate the value of the statistic computed on the data and determine a p-value.

- **Recipe:**
  - Simulate N data sets assuming H_0
  - Compute the test statistics (LRT) for each data set.
  - Make a histogram of the simulated test statistics
  - The histogram approximates the sampling distribution of the test statistic
  - calculate the p-value => proportion of simulated test statistics larger than t_{obs}



$$Pr(T \geq t_{obs} | H_0) = p_{\text{value}}$$

Methods: Posterior Predictive Sampling

- If there are unknown parameters in the null we cannot simulate data.
- **Solutions:**
  - Fit the real data under the null model and compute fitted parameters and error bars.
  - Parametric Bootstrap - resample data with unknown parameters to account for the error bars.
  - Bayesian Posterior Predictive modeling simulates unknown parameters from their posterior distribution, which are used to simulate the data sets.

Methods: Bayes Factors

- **Motivation:**
  - P-values are based on Pr(Data|H0)
  - We are interested in Pr(H0|Data)

If we compare these two calculations P-values can vastly overstate the evidence for the H1

Solution: Quantify p(H0 |Data) directly. Bayes factors give a method to do this.

$$\frac{p(H_0|Y)}{p(H_1|Y)} = B_{01}\frac{p(H_0)}{p(H_1)} \qquad B_{01} = \frac{p(Y|H_0)}{p(Y|H_1)}$$

H - model
θ - model parameters
Y - data

posteriors

priors

Methods: Bayes Factors

- **Challenges:**
  - Computing BF can be very challenging
  - BF assumes that one of the models H0 or H1 is true
    (can be a completely different model?)
  - Priors are much more important/influential when computing BF than with parameter inference. We have to be very careful about prior specification
- We are working on methods that address these challenges in practice.

Summary

- Bayesian Posterior Predictive p-values (PPP)
  - independent on the prior, need a careful evaluation of the evidence.
- Bayes Factors
  - Challenging issues in computing, sensitive to priors.
- Conclusion:
  - use PPP, Monte Carlo Sampling or Bootstrap to calibrate test statistics
  - but there are enough issues with the PPP that we need to look at BF more seriously

Further Reading

- CHASC web page -- https://hea-www.harvard.edu/astrostat/
- Protassov, R., van Dyk, D.A., Connors, A., Kashyap, V.L., and Siemiginowska, A., (2002) *Statistics: Handle with care - detecting multiple model components with the likelihood ratio test* ApJ, 571, 545
- van Dyk, D.A., Connors, A., Kashyap, V.L., and Siemiginowska, A., (2001) *Analysis of Energy Spectra with Low Photon Counts via Bayesian Posterior Simulation*, ApJ, 548, 224
- Park, T., van Dyk, D.A., and Siemiginowska, A. (2008) *Searching for Narrow Emission Lines in X-ray Spectra: Computation and Methods* ApJ, 688, 807

Methods: Bayes Factors

- Bayesian Evidence:

  The average likelihood over the prior distribution

$$p(Y|M) = \int p(Y|M,\theta)\underline{p(\theta|M)}d\theta$$

prior

M- model

θ - model parameters

Y - data

- Bayes Factor:

  The ratio of the Bayesian Evidence

  for each model

  Large BF values (>100) are decisive.

$$B_{01} = \frac{p(Y|M_0)}{p(Y|M_1)}$$

$$\frac{p(M_0|Y)}{p(M_1|Y)} = B_{01}\frac{p(M_0)}{p(M_1)}$$

prior!

- BF and posterior probability ratios

Can be used to “compare” not-nested models, such as a power law vs thermall

Methods: Bayes Factors

- Interpretation of the BF against the Jeffreys' scale:

BF Strength of evidence (toward M0)

1 ~ 3 Barely worth mentioning
3 ~ 10 Substantial
10 ~ 30 Strong
30 ~ 100 Very strong
> 100 Decisive