Asymptotic formulae for likelihood-based tests of new physics

A UW Journal Club review of the namesake paper, available at:

https://arxiv.org/pdf/1007.1727.pdf

Outline

OVERVIEW OF THE GENERAL METHOD OF HYPOTHESIS TESTING

HYPOTHESIS TESTING IN A COUNTING EXPERIMENT

EXAMPLES

I. Overview of the general method of hypothesis testing

Hypothesis Testing

Type of Test	H_0	H_1
Discovery	Background	Signal
Exclusion Limit	Signal	Background

- In physics experiments, we often conduct statistical inference by means of a hypothesis test. We compare two hypotheses: a **background hypothesis** and a **signal hypothesis**. For the purposes of the test, one hypothesis is considered the **null hypothesis** H_0, while the other is considered the **alternative hypothesis** H_1. The test determines to what extent we can **reject the null hypothesis**.
- Depending on our labelling of the background/signal hypotheses as null/alternative, we attempt to either verify the **discovery** of a new model or **limit** the parameters of that model.

Hypothesis Testing (continued)

- Hypothesis testing proceeds as follows:
  1. Define H_0 (the null hypothesis) and H_1 (the alternative hypothesis).
  2. Choose a **test statistic** q that can quantify the level of agreement between H_0 and the data.
  3. Calculate $q_{\mathbf{obs}}$ (q of the observed data).
  4. Use $q_{\mathbf{obs}}$ to calculate a p-value, and determine whether the p-value is sufficiently small to reject H_0.

Test Statistic

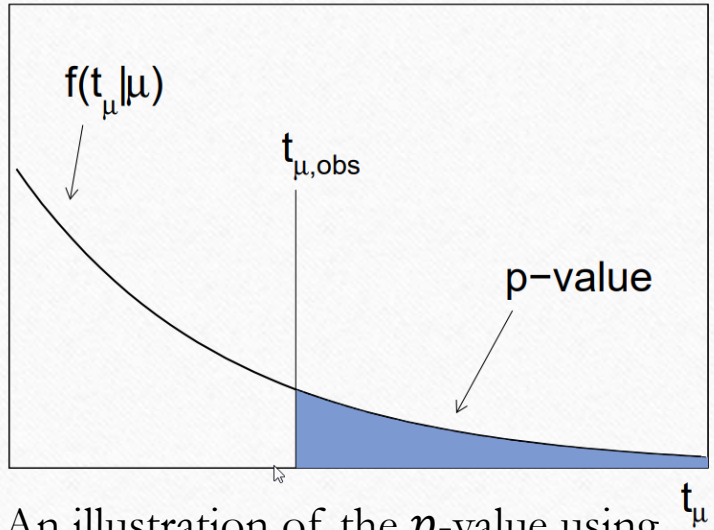
- An appropriate test statistic quantifies the degree to which the observed data follows the predictions of one hypothesis more strongly than the other.
- Most test statistics utilize a **ratio of likelihoods**. The likelihood of a hypothesis H given a dataset $\boldsymbol{x}$ is the probability of observing that dataset given the hypothesis. Mathematically, $L(H|x) = P(x|H)$. If you parametrize H in terms of some parameters $\boldsymbol{\theta}$, then you can speak of a **likelihood function** $\boldsymbol{L(\theta|x) = P(x|\theta)}$.[1]

[1] Note that the dependence of L on $\boldsymbol{x}$ is often left implied, such that $\boldsymbol{L(\theta) = L(\theta|x)}$.

p-value

- With basic probability, we can answer the question "How likely is the observed data given our null hypothesis?"
- However, what we'd *really* like to know is "How likely is data (given our null hypothesis) that is *at least* as incompatible with our null hypothesis as the observed data?" The p-value answers this question.
  - A test statistic q quantifies the notion of incompatibility. All we need to do is find $f(q|H_0)$, the pdf (probability density function) of q given our null hypothesis.

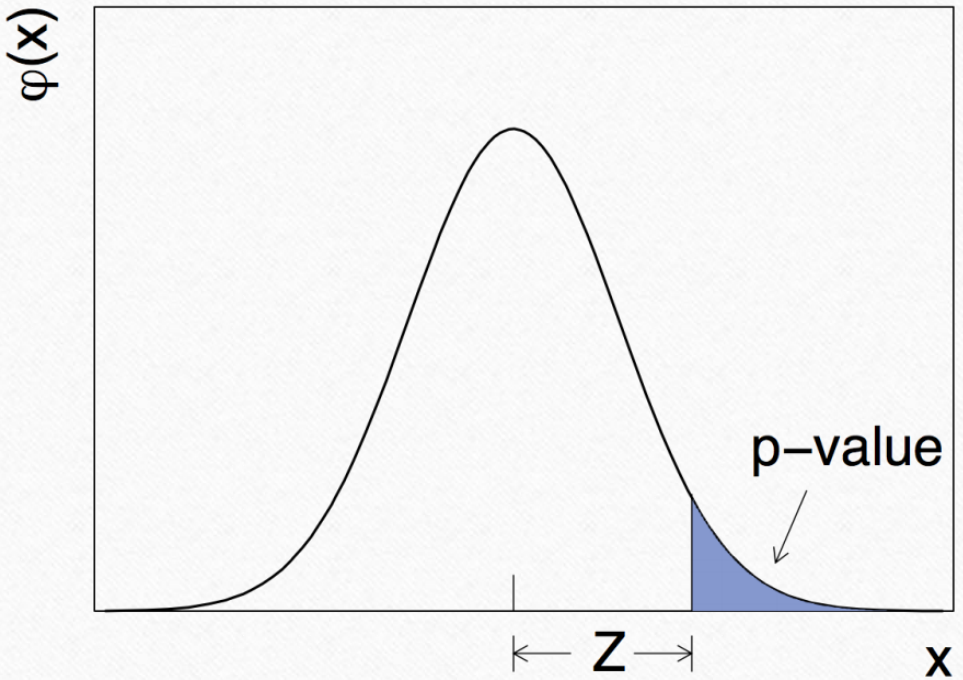
$$p_{H_0} = \int_{q_{\mathbf{obs}}}^{\infty} f\,(q|H_0)dq$$



An illustration of the p-value using the test statistic t_μ and hypothesis μ (this will make sense later).

Equivalent Significance

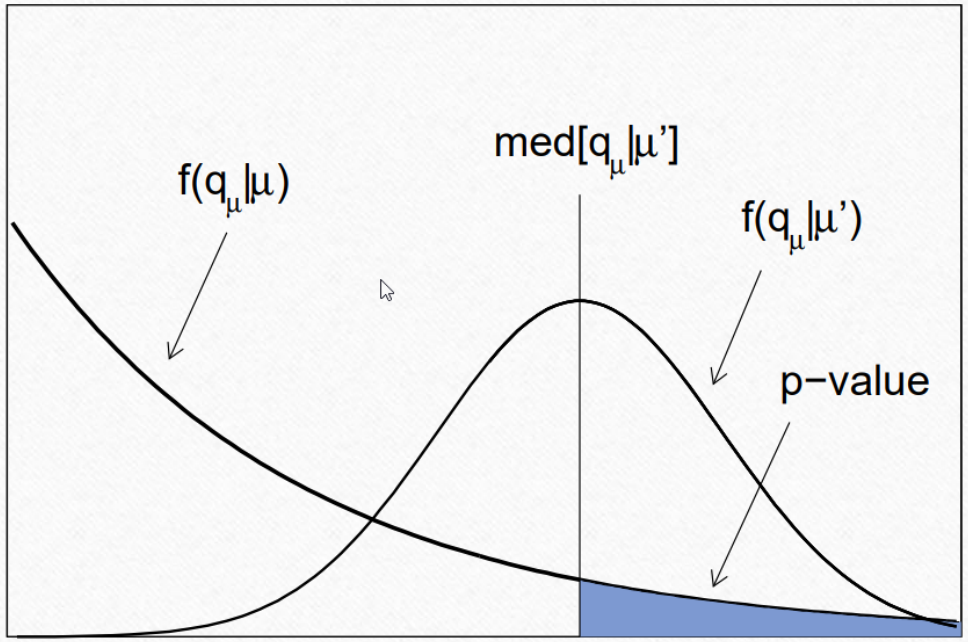
- When conducting a search, it is often convenient to discuss the **equivalent significance** $Z = \Phi^{-1}(1-p)$ of the p-value.
  - The probability of observing a Gaussian-distributed variable Z σs above its mean is equal to p.
- When testing for discovery, $Z \geq 5$ or $p \leq 2.87 \times 10^{-7}$ is the usual threshold for H_0 rejection. For exclusion, $Z \geq 1.64$ or $p \leq 0.05$ is generally considered sufficient.



Experimental Sensitivity (expected significance)

- In addition to calculating the significance for the observed dataset under our null hypothesis H_0, it would be useful to know the median significance under H_1.Type equation here.
- In this way, we can estimate the **sensitivity** of the experiment. If the median significance is very low, we are unlikely to reject H_0 even if H_1 is true, so our experiment is a waste of time.

$$\text{med}[p_{H_0}] = \int_{\text{med}[q_{\text{obs}}|H_1]}^{\infty} f(q|H_1)dq$$



An example of the p-value corresponding to the median significance Z of an experiment.

Approximate Distributions

- Reminder: p_{H_0} is the probability of observing data that is at least as incompatible with our null hypothesis H_0 as the observed data.
- To calculate p_{H_0} we need:
  - $f(q|H_0)$ – the pdf of our test statistic q given H_0.
- To calculate $\mathbf{med}[p_{H_0}]$ we need:
  - $f(q|H_1)$ – the pdf of our test statistic q given our alternative hypothesis H_1.
- In practice, approximations of these distributions are used, as calculation of the exact distributions tends to be infeasible.

Control Samples

- Treat the control samples that constrain the nuisance parameters as fixed, $\pi_0(\theta)$
- Determine the distribution of q by generating the main search measurement.
- For systematic uncertainties:
  - Take control samples as the basis of Bayesian prior density $\pi(\theta)$
  - $f(q) = \int f(q|\theta)\pi(\theta)d\theta$
- Find the prior $\pi(\theta)$ by Bayes' theorem.
  - $\pi(\theta) \propto L_\theta(\theta)\pi_0(\theta)$, $\pi_0(\theta)$ is take as constant in many cases.

Control Samples(Continued)

At Tevatron:

- Determine the distribution of q by generating only the main search measurement. (at Tevatron)
- Nuisance parameters are constrained by Gaussian distributed estimates, the initial prior $\pi_0(\theta)$

MC:

- For a given assumed point in the model's parameter space
- Simulate both the control measurements and the main measurement.

II. Hypothesis testing in a counting experiment

The Statistical Model

- Consider an experiment resulting in measurements of a variable of interest x. The measurements collectively form a binned histogram, with n_i entries in bin i according to a Poisson distribution with $E[n_i] = \mu s_i + b_i$.
  - s_i – the expected number of signal samples in bin i.
  - b_i – the expected number of background samples in bin i.
  - μ – the **signal strength** of the model.
    - $\mu = 0 \rightarrow$ background hypothesis
    - $\mu = 1 \rightarrow$ signal hypothesis
- s_i and b_i are characterized by:
  - s_{tot} & b_{tot} – the total expected number of signal/background samples.
  - $f_s(x;\boldsymbol{\theta}_s)$ & $f_b(x;\boldsymbol{\theta}_b)$ – pdfs of x given nuisance parameters[1] $\boldsymbol{\theta}_s, \boldsymbol{\theta}_b$ that describe the shape of the distributions.

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x;\boldsymbol{\theta}_s)dx$$

$$b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x;\boldsymbol{\theta}_b)dx$$

- The **likelihood** $L(\boldsymbol{\mu},\boldsymbol{\theta})$ is then given by $L(\boldsymbol{\mu},\boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-\mu s_j + b_j}$.

[1]parameters that are necessary for analysis, but are not themselves of interest. Note that b_{tot} is also a nuisance parameter.

Maximum Likelihood Estimators (MLEs)

- Given data $\boldsymbol{x}$ and a hypothesis $H(\boldsymbol{\theta})$, we might wish to know which values of $\boldsymbol{\theta}$ are most likely given $\boldsymbol{x}$. These values, denoted $\widehat{\boldsymbol{\theta}}$, are known as the MLEs of $\boldsymbol{\theta}$. If we know $L(\boldsymbol{\theta})$ corresponding to $H(\boldsymbol{\theta})$, then we can find $\widehat{\boldsymbol{\theta}}$ by setting $\frac{dL}{d\boldsymbol{\theta}} = 0$, as per elementary calculus.
- MLEs are a useful way to estimate parameters of a model based on a sample. They are asymptotically (1) unbiased, (2) normally distributed, and (3) efficient estimates of the true parameters.

Profile Likelihood Ratio

- The test statistics described in the paper rely upon the **profile likelihood ratio** $\lambda(\mu) = \frac{L\left(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu)\right)}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$.
  - $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the MLEs of μ and $\boldsymbol{\theta}$.
  - $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ is the MLE of $\boldsymbol{\theta}$ conditional on μ.
  - Since $L(\hat{\mu}, \hat{\theta})$ is a maximum of L by definition, $0 \leq \lambda(\mu) \leq 1$.
  - $\lambda(\mu) \approx 1$ implies that the data supports the given μ.
  - $\lambda(\mu) \approx 0$ implies that the data does not support the given μ.

Test Statistics (continued)

- As can be seen on the right, the test statistics are variants of $\lambda(\mu)$ adapted for different use-cases.
- For some of the statistics, a modified profile likelihood ratio $\tilde{\lambda}(\mu)$ is used to avoid negative values of μ being preferred.

$$\tilde{\lambda}(\mu) = \begin{cases} \lambda(\mu), & \hat{\mu} \geq 0 \\ \dfrac{L\left(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu)\right)}{L\left(0, \hat{\hat{\boldsymbol{\theta}}}(0)\right)}, & \hat{\mu} < 0 \end{cases}$$

$$\lambda(\mu) = \frac{L\left(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu)\right)}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

Statistic	Use	Definition
t_μ	two-sided interval	$t_\mu = -2 \ln \lambda(\mu)$
$\tilde{t}_\mu$	enforces positive signal	$\tilde{t}_\mu = -2 \ln \tilde{\lambda}(\mu)$
q_0	discovery	$q_0 = \begin{cases} t_0, & \hat{\mu} \geq 0 \\ 0, & \hat{\mu} < 0 \end{cases}$
q_μ	upper limit	$q_\mu = \begin{cases} t_\mu, & \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases}$
$\tilde{q}_\mu$	enforces positive signal	$\tilde{q}_\mu = \begin{cases} \tilde{t}_\mu, & \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases}$

Asimov Dataset

- Problem: to calculate median significance, we need $\mathbf{med}[q_{\mathbf{obs}}|H_1]$.
  - Hard solution: $\int_{-\infty}^{\mathbf{med}[q|H_1]} f(q|H_1)dq = \frac{1}{2}$
  - Easy solution: use "Asimov dataset."
- The Asimov dataset is defined "such that when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values."
- What this means:
  - $n_{i,A} = E[n_i]$
  - $m_{i,A} = E[m_i]$

$$\mathbf{med}[q_{\mathbf{obs}}|H_1] \approx q_A$$

- In practice, the Asimov dataset can be found by Monte Carlo simulations of H_1.

Approximate Distributions (continued)

- Approximating the likelihood ratio $\lambda(\mu)$ is the main issue in approximating the test statistic distributions.
- It turns out that $\ln \lambda(\mu)$ follows what is called a non-central chi-square distribution.
- The approximation of $\lambda(\mu)$ can be used to approximate the test statistic distributions, which in turn give the approximate significances shown on the right.
- Replacing $q \to q_A$ in the expressions for Z gives the approximate median significance using the Asimov dataset.

Statistic	Z
t_μ	$\Phi^{-1}(2\Phi(\sqrt{t_\mu}) - 1)$
$\tilde{t}_\mu$	…
q_0	$\sqrt{q_0}$
q_μ	$\sqrt{q_\mu}$
$\tilde{q}_\mu$	…

Exclusion Limits

- Test statistic:
  - Use either q_μ or $\tilde{q}_\mu$ (they are asymptotically equivalent: q_μ is generally more convenient). It follows that $Z = \sqrt{q_\mu}$.
- Goal: exclude μ at CL $1 - \alpha$ (usually 95%).
- Use the approximations $Z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu}$ and $q_\mu = \frac{(\mu - \hat{\mu})^2}{\sigma^2}$ to solve for μ s.t. $p_\mu = \alpha$
- Solution: $\mu = \hat{\mu} + \sigma\Phi^{-1}(1 - \alpha)$
  - $\hat{\mu}$ – MLE from data.
  - σ – standard deviation of $\hat{\mu}$, approximated either from $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$ with $\sigma^2 = V_{00}$ and $V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$, or from $\sigma_A^2 = \frac{(\mu - \hat{\mu})^2}{q_{\mu,A}}$.
  - Since σ depends on μ, the equation is solved numerically.

Expected Limit

- As before, we should like to know the median limit assuming the background hypothesis.
- Using the Asimov dataset for the background hypothesis (thus $\hat{\mu} = 0$) and the approximation $\sigma_A^2 = \frac{(\mu - \hat{\mu})^2}{q_{\mu,A}}$, the expression $\mu = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha)$ reduces to $\sqrt{q_{\mu,A}} = \Phi^{-1}(1-\alpha)$. Solve numerically for μ.
- Error bands for the median limit can similarly be found by $\mathrm{band}_{N\sigma} = \sigma(\Phi^{-1}(1-\alpha) \pm N)$

III. Examples

Example: Shape Analysis

- In this example we take the case where you are searching for a peak in an invariant mass distribution
  - Invariant mass distribution: Distribution of Invariant mass, which is the mass in the "rest frame" .
- To find a peak, you test every mass in a given range – the appearance of a signal like peak could lead to rejection of the background-only hypothesis

The "Look Elsewhere Effect"

- Since we are looking at a very large range, we must take into account the "look-elsewhere" effect:
  - This is the effect that a fluctuation could occur at any mass within the range – a good analogy of this is that if you are drawing hands from a deck, you will eventually draw a royal flush (or some other good hand).
- To account for the look else-where effect you divide the threshold by the number of trials to get $p <$ threshold / number of trials
- In this case we don't have to worry about this effect because we will effectively test each mass and signal strength individually.

Scale Factor

- Signal Scale Factor: Corresponds to the strength parameter μ
- Background Scale Factor: Introduce factor called θ
  - Mean value of events given by $E[n_i] = \mu s_i + b_i$ where μ and s_i are taken to be known
  - We assume that the background terms, given by b_i can be expressed as $b_i = \theta f_{b,i}$ where $\theta f_{b,i}$ is the probability to find a background event in bin i, which is known and θ is a nuisance parameter that gives the total number of background events.

Likelihood Function

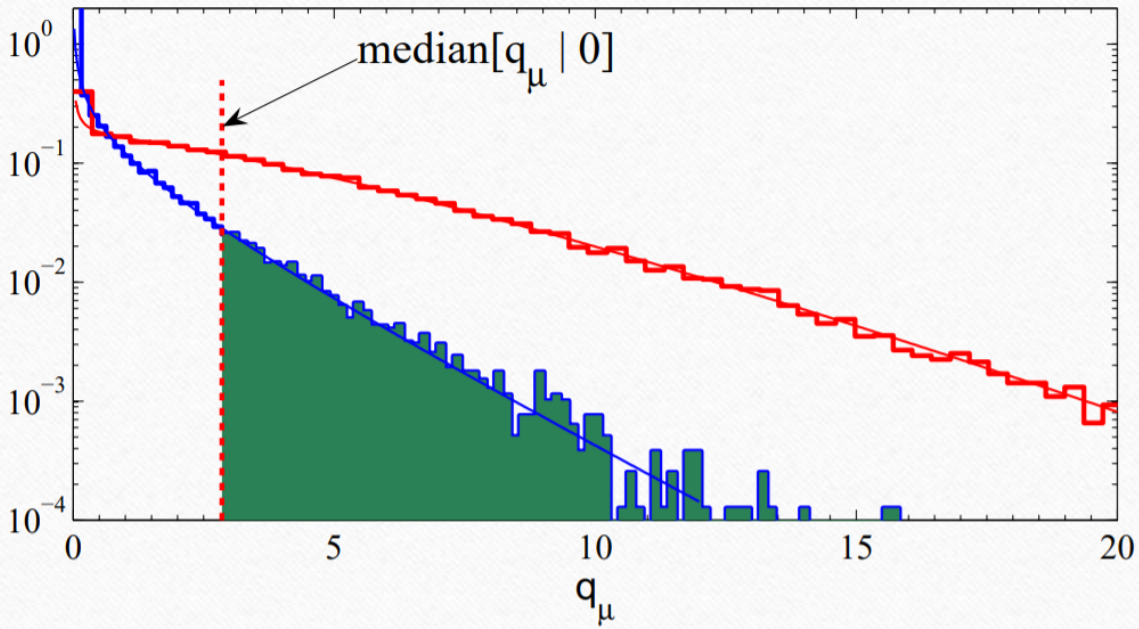
- Using the scale factors defined in the last slide, along with the likelihood function as given in equation 6 we get:

$$L(\mu, \theta) = \prod_{J=1}^{N}((\mu s_i + b_i)^{ni}/n_i!)e^{-(\mu s_i + b_i)} \Rightarrow L(\mu, \theta) = \prod_{J=1}^{N}((\mu s_i + \theta f_{b,i})^{ni}/n_i!)e^{-(\mu s_i + \theta f_{b,i})}$$

- Using this likelihood function you can evaluate it and get any of the test statistics.

Test Statistic q_μ

- The median assumes a strength parameter of μ'
- The upper limit on μ at a confidence level of $CL = 1-\alpha$ is the value of μ for which $p_\mu = \alpha$



f(q_μ |0) (red) and f(q_μ |μ) (blue) p-value of hypothesized μ shaded in green – Figure shows the value of μ that gave $p_\mu = 0.05$

Conclusion

- Today we had an overview of the general method of hypothesis testing, hypothesis testing in a counting experiment and an example of how likelihood functions can be modified to get test statistics
  - General Method:
    - Getting q and then determining, using the p-value if the hypothesis fits the data
  - Counting Experiment:
    - The use of the likelihood ratio in getting test statistics and "shortcuts" such as the Asimov data set
  - Shape analysis:
    - Application of these methods to find just one of the test statistics, q_μ
- Most importantly using these methods eliminates the need to perform lengthy MC calculation, which for the case of a discovery at 5σ significance could require the simulation of around 10^8 measurements.

Thank You

- Questions?