Practical Statistics

Jochen Ott

June 5, 2014

Why "Practical Statistics"?

Statistics is often taught and discussed on simple cases such as a counting experiment with a Poisson distribution and no systematics.

In practice, however, we deal with shape analyses and systematics – beyond what can easily extrapolated from a counting experiment.

Therefore, I'll try to give an statistics introduction covering both: The basic statistical concepts, and at the same time show how to actually apply them in practice in a realistic analysis, as they are e.g. common in Higgs analyses at CMS.

My talk has three parts

1. Hypothesis Tests, Monte-Carlo Methods, Model Building
2. Limits and Intervals (Frequentist only)
3. Brief Introduction to statistical tools, esp. "theta"

Part I

Hypothesis Tests

Contents

Introduction

Statistical hypothesis testing is a formal method for decision making using data from a random process.

It is an attempt to disprove a null hypothesis H_0, which is rejected if the probability to observe the data that have actually been observed – or even more extreme data – is very low for H_0.

This probability is called the *p*-value. If it is below some (small) pre-defined threshold α, the hypothesis H_0 is rejected in favor of the alternative H_1.

I'll use two examples: A counting experiment and a (more realistic) shape analysis. Complications such as systematic uncertainties are added later.

Statistical Model

A statistical model specifies the probability to observe certain data d as a function of the (real-valued) model parameters θ.

A simple statistical model is a counting experiment with known background mean $b = 5.2$ and unknown signal $s \geq 0$ we want to "discover". The data comprises only the number of observed events n, which has a Poisson distribution around $\lambda = b + s$. $b = 5.2$ is constant and $s \geq 0$ is the (only) model parameter.

This statistical model can be summarized as:

$$p(n|s) = \text{Poisson}(n|\lambda = b + s) = \frac{e^{-b-s}(b+s)^n}{n!}$$

Hypothesis Test

The null hypothesis – we would like to reject – is $s = 0$. The alternative hypothesis is a positive signal, $s > 0$.

Given the observed number of events n_{obs}, the p-value is the probability to observe as least as many events for the null hypothesis $s = 0$:

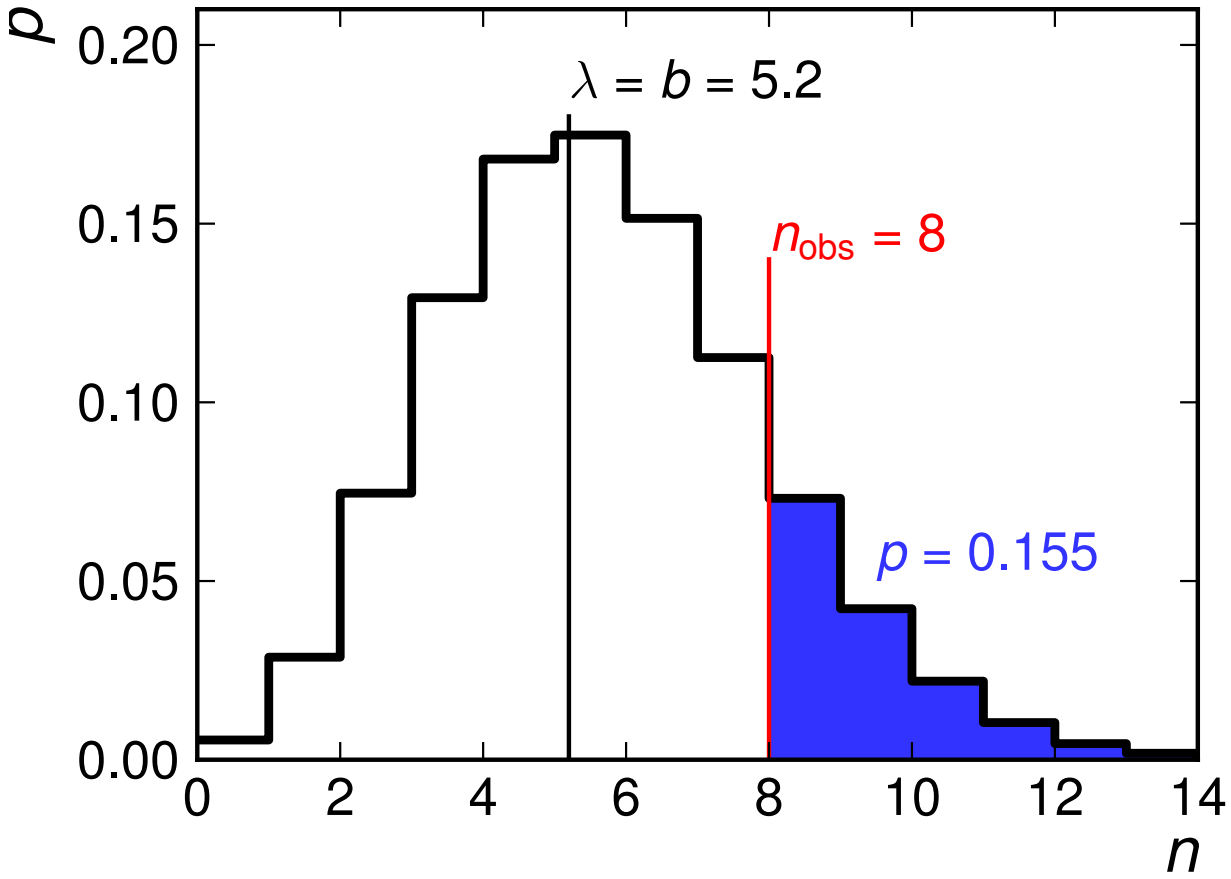
$$p(n_{\text{obs}}) = \sum_{n=n_{\text{obs}}}^{\infty} \text{Poisson}(n|\lambda = b = 5.2)$$

Remarks:

- The p-value itself is a random variable.
- Definition implies: p-value follows a uniform distribution on the interval $[0, 1]$ for H_0 (or approximately if data is discrete)

Result

For the example of $b = 5.2$, the p-value is the probability to measure at least as many events as observed for $s = 0$. This can be evaluated directly numerically or by making toys.



n_{obs}	p
6	0.419
8	0.155
10	0.040
12	0.0073

Possible Outcomes of a Hypothesis Test

There are two kinds of errors that can be made in the hypothesis test:

1. Rejecting H_0 although it is true. This is the type-I error (or "error of the first kind").
2. Not rejecting H_0 although it is false; type-II error (or "error of the second kind").

The first is usually regarded more severe. The probability for a type-I error is the ("small") threshold α used in the hypothesis test.

The type-II error is denoted β.
The probability to (correctly) reject H_0 if the alternative H_1 is true is the power $(1 - \beta)$.

Remarks

- For a given fixed type-I error rate α, one would prefer the test with high power $(1 - \beta)$; this can be used to choose "optimal" test statistic.
- The *p*-value is not simply "the probability to observe the data as observed", but to "observe such data or more extreme data". This distinction is important; also, it implies that a comparison is needed between 2 datasets (real-valued test statistic).
- The *p*-value is not the probability that the null hypothesis is true – such a statement carries no meaning in frequentist statistics, where probability always refers to (random) data and derived quantities, never to model parameters (but: Bayesian view differs, see below).
- Rejecting the null hypothesis does *not* proof that the alternative hypothesis is true: Usually, many alternatives to the null hypothesis exist which are compatible with the observed observed data.

Contents

Introduction

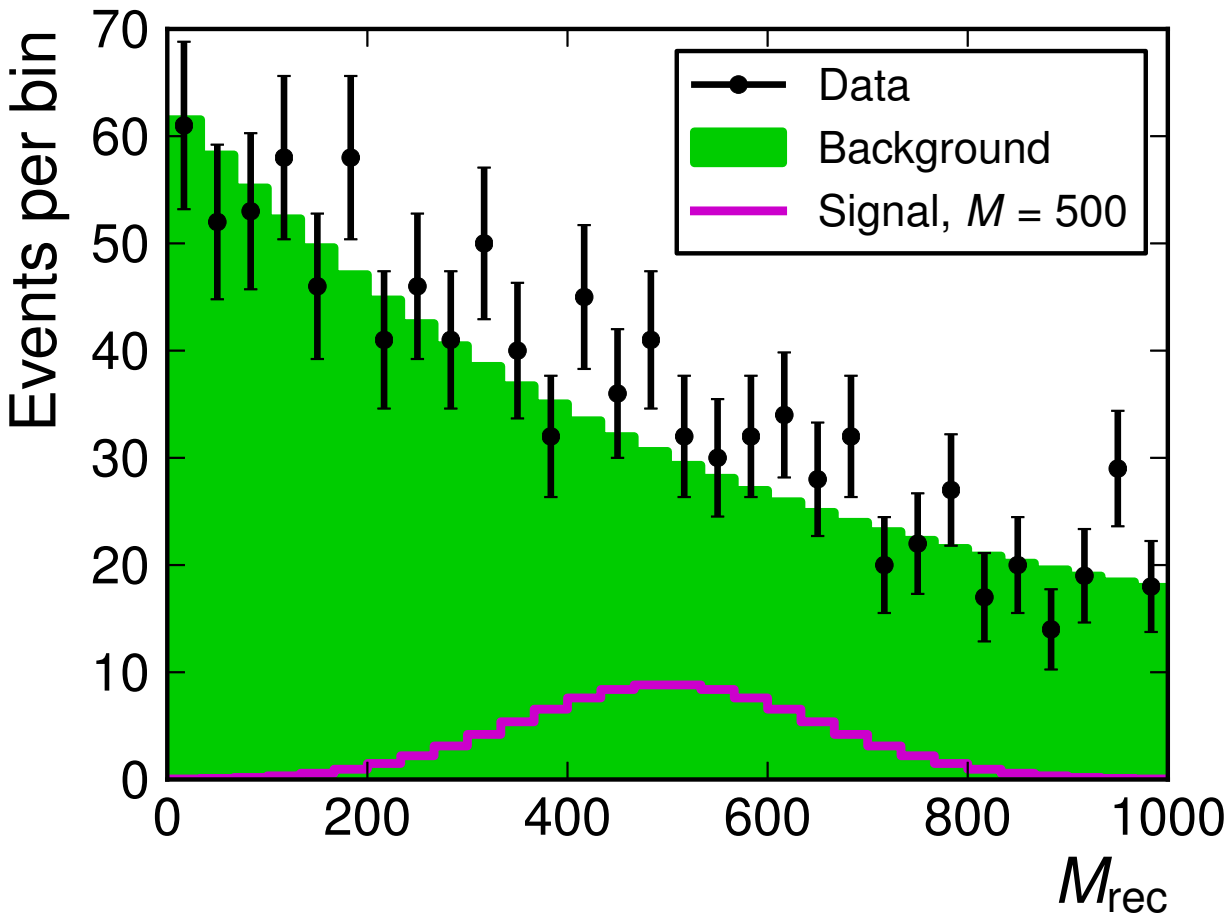
So far: simplistic Poisson model with only one event count ("counting experiment").

Now: Generalize to a more realistic analysis in which the (binned) shape of some reconstructed mass distribution (M_{rec}) is analyzed to search for a resonance of unknown mass over some falling background.

Versions of such models are used in many channels of the Higgs boson search at the LHC, but also many other searches.

Shape Model: Plot

For example, the expected Poisson means for background and data might look like this:



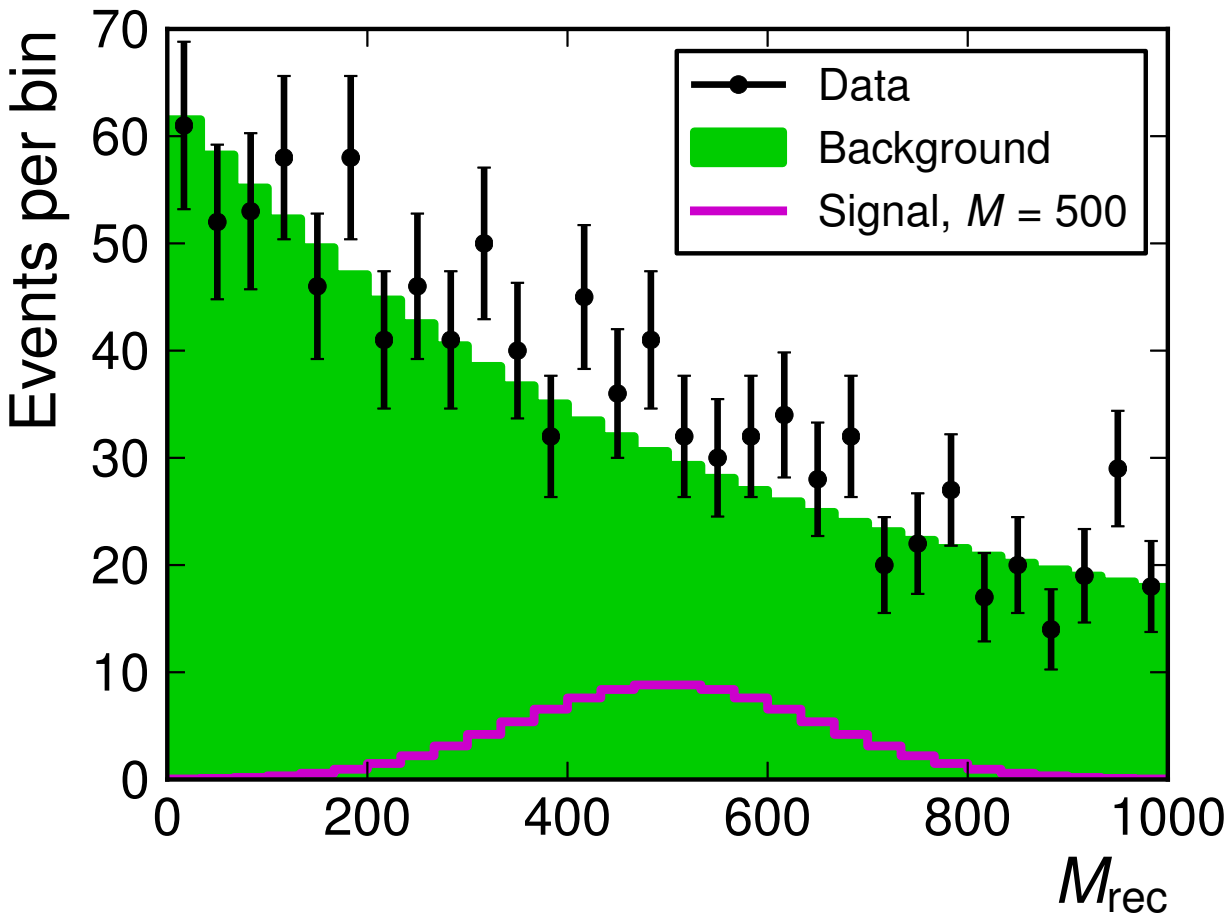
Can this data be seen as evidence against the background-only model?

Can we exclude $H_0 =$ background only in favor of $H_1 =$ background + (scaled) $M = 500$ signal?

Next steps: statistical model, hypothesis test!

Shape Model: Plot

For example, the expected Poisson means for background and data might look like this:



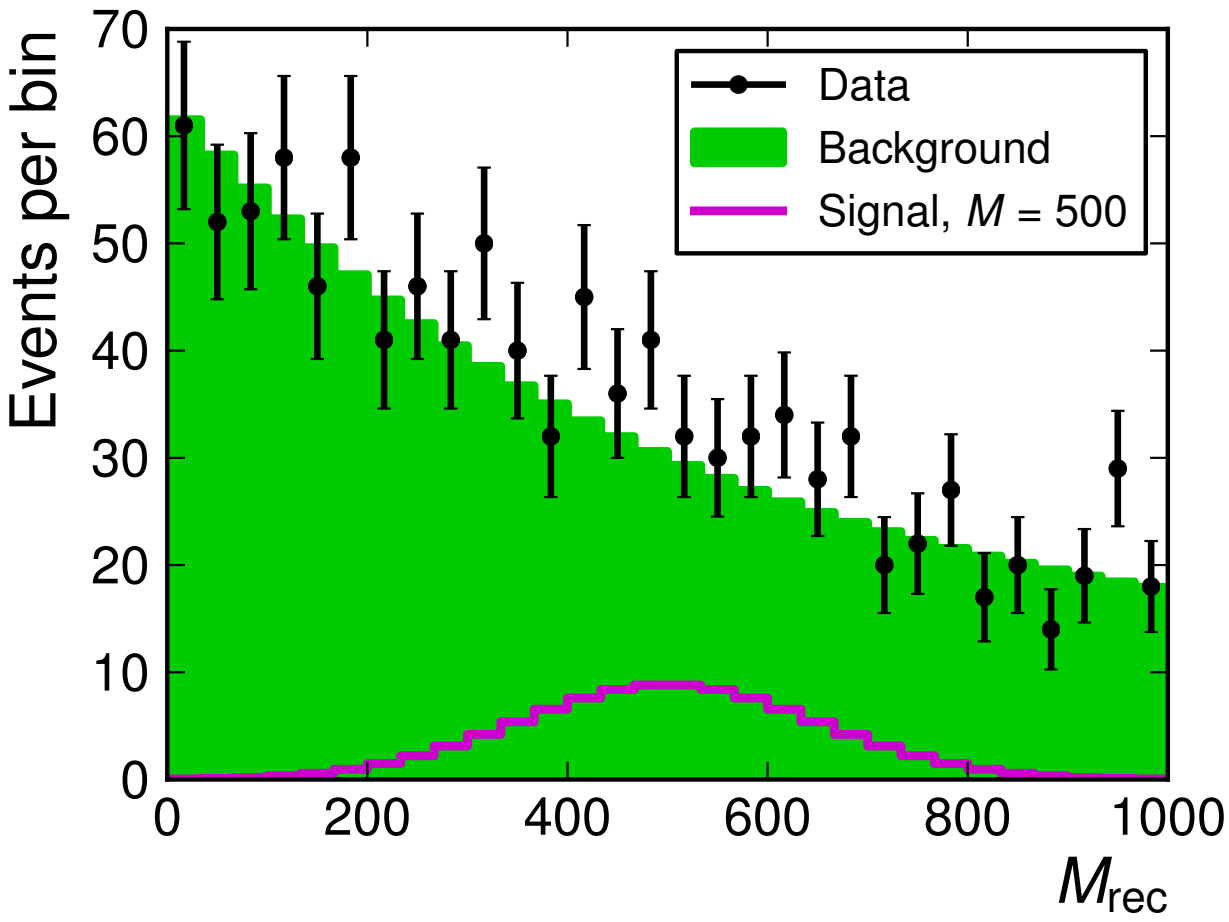
Can this data be seen as evidence against the background-only model?

Can we exclude H_0 = background only in favor of H_1 = background + (scaled) $M = 500$ signal?

Next steps: statistical model, hypothesis test!

Shape Model: Plot

For example, the expected Poisson means for background and data might look like this:



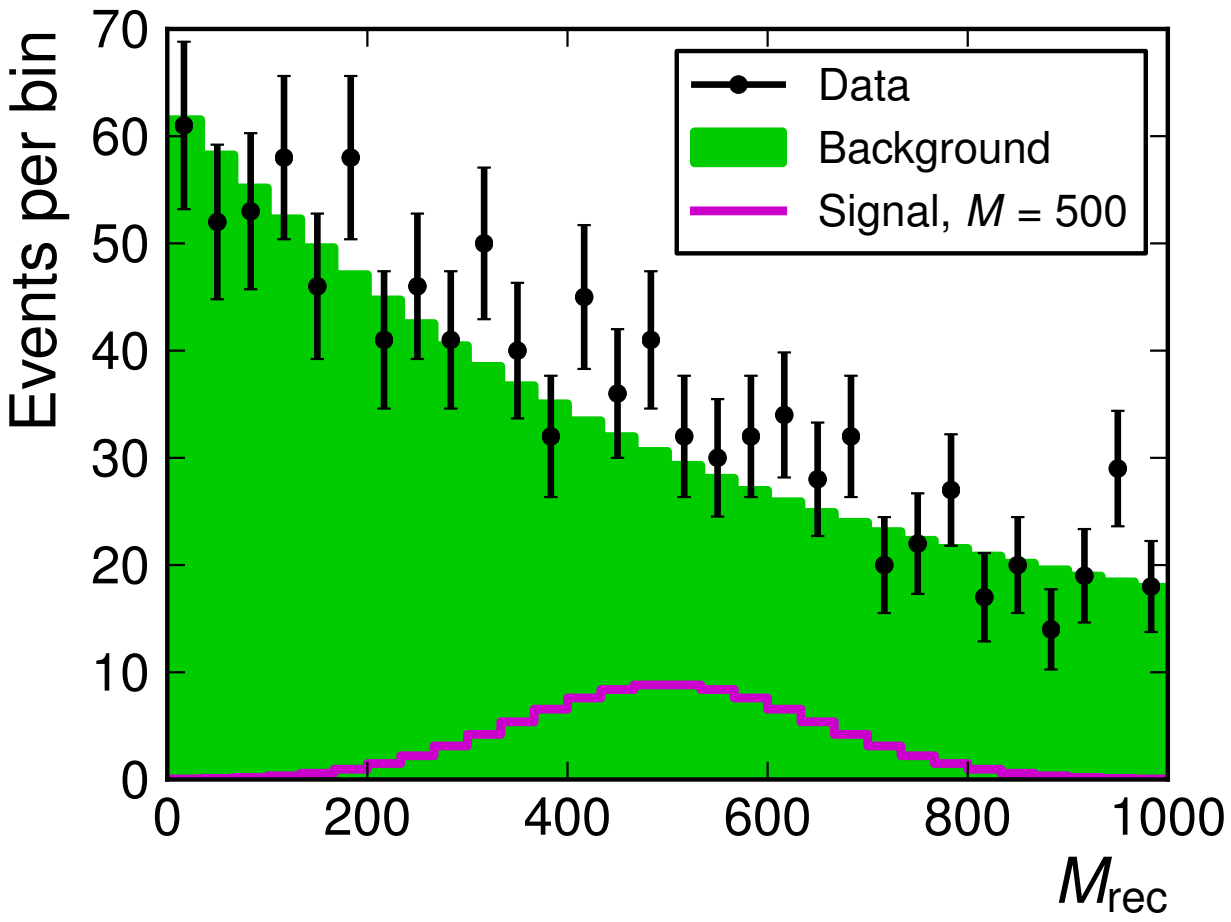
Can this data be seen as evidence against the background-only model?

Can we exclude H_0 = background only in favor of H_1 = background + (scaled) $M = 500$ signal?

Next steps: statistical model, hypothesis test!

Shape Model: Plot

For example, the expected Poisson means for background and data might look like this:



Can this data be seen as evidence against the background-only model?

Can we exclude H_0 = background only in favor of H_1 = background + (scaled) $M = 500$ signal?

Next steps: statistical model, hypothesis test!

Shape Model: Statistical Model

Probability to observe event counts $\vec{n} = (n_1, n_2, \ldots)$ is a product of Poisson probabilities in each bin:

$$\begin{aligned} p(\vec{n}|\mu) &= \prod_i \text{Poisson}(n_i|\lambda_i(\mu)) \qquad \text{with} \\ \lambda_i(\mu) &= \mu s_i + b_i \end{aligned}$$

where $i = 1, \ldots, N_{\text{bins}}$ denotes the bin index. s_i and b_i are the signal and background templates, resp., which are typically derived from Monte-Carlo simulation or from a background-enriched sideband.

$\mu \geq 0$ scales the signal template; it is the signal strength parameter. Apart from a (known constant) factor, it is the signal cross section.

Contents

The role of the Test Statistic t

Reminder: The p-value is defined as the probability to observe data at least as extreme (signal-like) as the one actually observed.

For a counting experiment, more events are more "extreme", more "signal-like". In general, one has to summarize the "signal-likeness" in one real number, this is the test statistic.

A possible choice is the (profile) likelihood ratio:

$$t = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)}$$

where we assume that the hypotheses H_0 and H_1 correspond to parameter sub-spaces of a common stat. model; for searches: H_0 is $\mu = 0$ and H_1 is $\mu > 0$.

Again: t measures the compatibility with H_0; large values mean incompatibility with H_0, favoring H_1.

The role of the Test Statistic t

Reminder: The p-value is defined as the probability to observe data at least as extreme (signal-like) as the one actually observed.

For a counting experiment, more events are more "extreme", more "signal-like". In general, one has to summarize the "signal-likeness" in one real number, this is the test statistic.

A possible choice is the (profile) likelihood ratio:

$$t = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)}$$

where we assume that the hypotheses H_0 and H_1 correspond to parameter sub-spaces of a common stat. model; for searches: H_0 is $\mu = 0$ and H_1 is $\mu > 0$.

Again: t measures the compatibility with H_0; large values mean incompatibility with H_0, favoring H_1.

The role of the Test Statistic t

Reminder: The p-value is defined as the probability to observe data at least as extreme (signal-like) as the one actually observed.

For a counting experiment, more events are more "extreme", more "signal-like". In general, one has to summarize the "signal-likeness" in one real number, this is the test statistic.

A possible choice is the (profile) likelihood ratio:

$$t = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)}$$

where we assume that the hypotheses H_0 and H_1 correspond to parameter sub-spaces of a common stat. model; for searches: H_0 is $\mu = 0$ and H_1 is $\mu > 0$.

Again: t measures the compatibility with H_0; large values mean incompatibility with H_0, favoring H_1.

p-value definition via t

The p-value is the probability to observe $t \geq t_{\text{obs}}$ if H_0 is true:

$$p = \Pr(t \geq t_{\text{obs}} | H_0).$$

This suggests using a Monte-Carlo method for calculating the p-value:

1. Generate a large number of toy data distributed according to H_0.
2. For each toy data, calculate test statistic t.
3. For the observed data, calculate test statistic t_{obs}.
4. The p-value is given by the fraction of toys with $t \geq t_{\text{obs}}$; if $p < \alpha$, reject H_0.

The values of t_{obs} for which H_0 is rejected at level α is known as critical region in t.

p-value definition via t

The p-value is the probability to observe $t \geq t_{\text{obs}}$ if H_0 is true:

$$p = \Pr(t \geq t_{\text{obs}} | H_0).$$

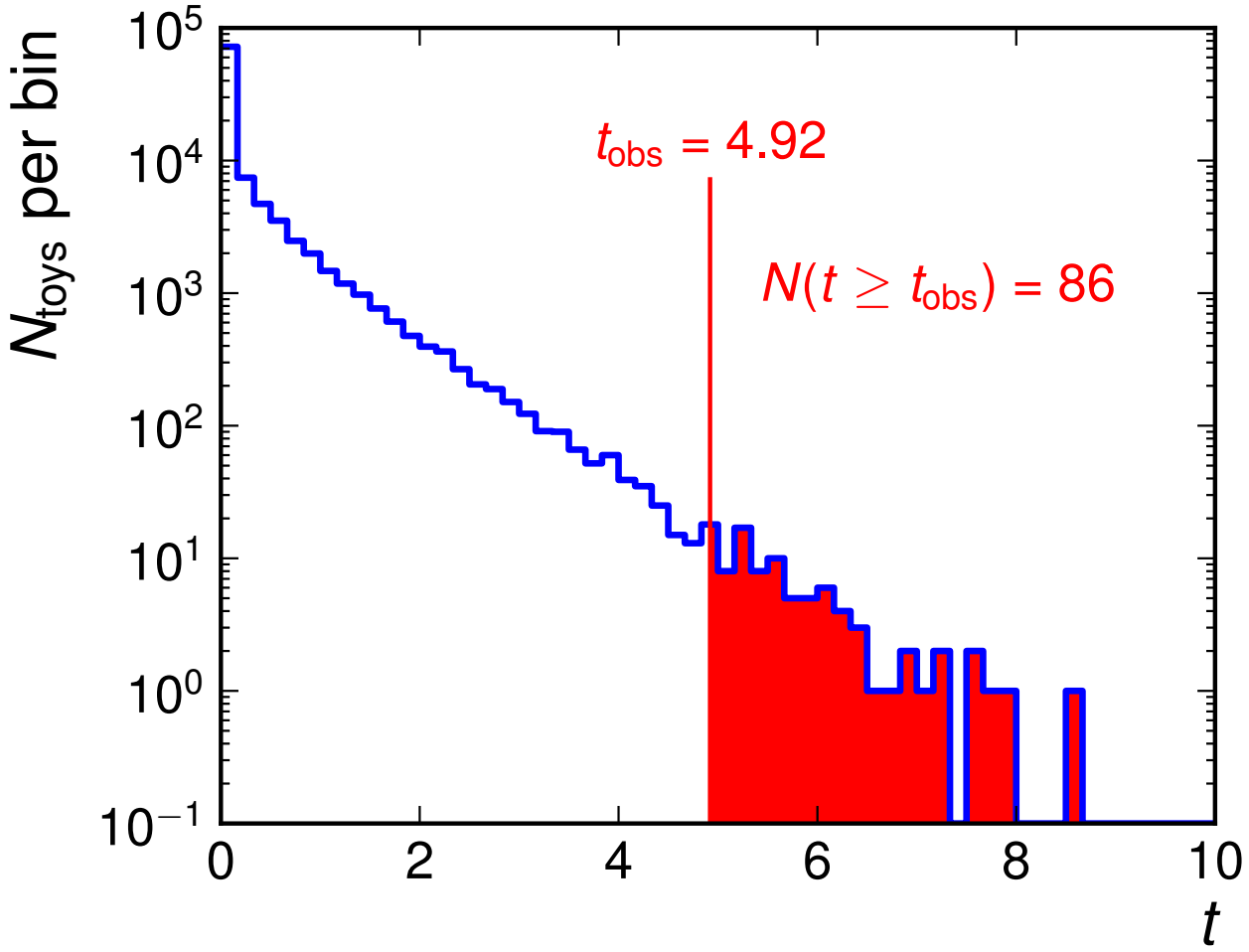
This suggests using a Monte-Carlo method for calculating the p-value:

1. Generate a large number of toy data distributed according to H_0.
2. For each toy data, calculate test statistic t.
3. For the observed data, calculate test statistic t_{obs}.
4. The p-value is given by the fraction of toys with $t \geq t_{\text{obs}}$; if $p < \alpha$, reject H_0.

The values of t_{obs} for which H_0 is rejected at level α is known as critical region in t.

MC method for p: Example II

Result of 100,000 background-only toys: $\hat{p} = 86/10^5 \rightsquigarrow \hat{Z} = 3.13$



Contents

Introduction

In the statistical model, each uncertainty is included as an additional parameter, called nuisance parameter.

Often, there is some external knowledge about the possible values of those nuisance parameters.

Introducing systematic uncertainties requires:

- Changes of the statistical model, i.e. how the nuisance parameters typically affect the probability.
- Changes of the significance calculation, i.e. how to include knowledge about nuisance parameters in the inference.

Those items will be discussed separately in the following slides.

Example

In the counting experiment, assume that the expected background $b = 5.2$ has some uncertainty (e.g. from limited statistics in a sideband). This can be included by changing the statistical model to:

$$p(n|s, b) = \text{Poisson}(n|\lambda = b + s)$$

where b now is a nuisance parameter, not a constant.

We assume that there is external knowledge about b (e.g. from a sideband measurement) suggesting that b is around $b_0 = 5.2$ with some uncertainty $\Delta b = 2.6$.

Changes to Significance Evaluation

We assume there is external knowledge about the nuisance parameters, which has to be incorporated in the procedure. Possible methods include:

1. Make it internal to the model, i.e., fit the nuisance parameter simultaneously with the parameter of interest (e.g. include sideband in likelihood model).
2. Use Bayesian priors for the nuisance parameters and take prior-average
3. Include auxiliary measurements in the statistical model in an approximate way and use bootstrapping.

Here, only item 2. is covered (see Backup for 3.).

Changes to Significance Evaluation

We assume there is external knowledge about the nuisance parameters, which has to be incorporated in the procedure. Possible methods include:

1. Make it internal to the model, i.e., fit the nuisance parameter simultaneously with the parameter of interest (e.g. include sideband in likelihood model).
2. Use Bayesian priors for the nuisance parameters and take prior-average
3. Include auxiliary measurements in the statistical model in an approximate way and use bootstrapping.

Here, only item 2. is covered (see Backup for 3.).

Changes to Significance Evaluation

We assume there is external knowledge about the nuisance parameters, which has to be incorporated in the procedure. Possible methods include:

1. Make it internal to the model, i.e., fit the nuisance parameter simultaneously with the parameter of interest (e.g. include sideband in likelihood model).
2. Use Bayesian priors for the nuisance parameters and take prior-average
3. Include auxiliary measurements in the statistical model in an approximate way and use bootstrapping.

Here, only item 2. is covered (see Backup for 3.).

Bayesian vs. Frequentist "Probability"

- Frequentist: "Probability is the relative frequency of a certain outcome in an ensemble of (imaginary or real) repetitions of a random process."
  Probability is only assigned to "data" and derived quantities (test statistic, p-value, etc.), but not to model parameters, which have only one true (unknown) value; the concept of probability does not apply.
- Bayesian: "Probability can also be used to express the current state of knowledge."
  In particular, it is valid to talk about the probability that a model parameter takes certain values.

Here, we discuss a "mixed" / "hybrid" method: Significance and p-values are a frequentist concept, but the use of nuisance parameter priors is Bayesian.

Bayesian vs. Frequentist "Probability"

- Frequentist: "Probability is the relative frequency of a certain outcome in an ensemble of (imaginary or real) repetitions of a random process."
  Probability is only assigned to "data" and derived quantities (test statistic, p-value, etc.), but not to model parameters, which have only one true (unknown) value; the concept of probability does not apply.
- Bayesian: "Probability can also be used to express the current state of knowledge."
  In particular, it is valid to talk about the probability that a model parameter takes certain values.

Here, we discuss a "mixed" / "hybrid" method: Significance and p-values are a frequentist concept, but the use of nuisance parameter priors is Bayesian.

Bayesian vs. Frequentist "Probability"

- Frequentist: "Probability is the relative frequency of a certain outcome in an ensemble of (imaginary or real) repetitions of a random process."
  Probability is only assigned to "data" and derived quantities (test statistic, p-value, etc.), but not to model parameters, which have only one true (unknown) value; the concept of probability does not apply.
- Bayesian: "Probability can also be used to express the current state of knowledge."
  In particular, it is valid to talk about the probability that a model parameter takes certain values.

Here, we discuss a "mixed" / "hybrid" method: Significance and p-values are a frequentist concept, but the use of nuisance parameter priors is Bayesian.

Prior-Averaging

Modify the Monte-Carlo method for p-value calculation: To generate toy data,

1. Draw a random value for each nuisance parameter from its prior.
2. Draw random data from the probability according to the stat. model evaluated for those (random) parameter values.

Then, as usual: p-value is the fraction of toys in which $t \geq t_{\text{obs}}$.

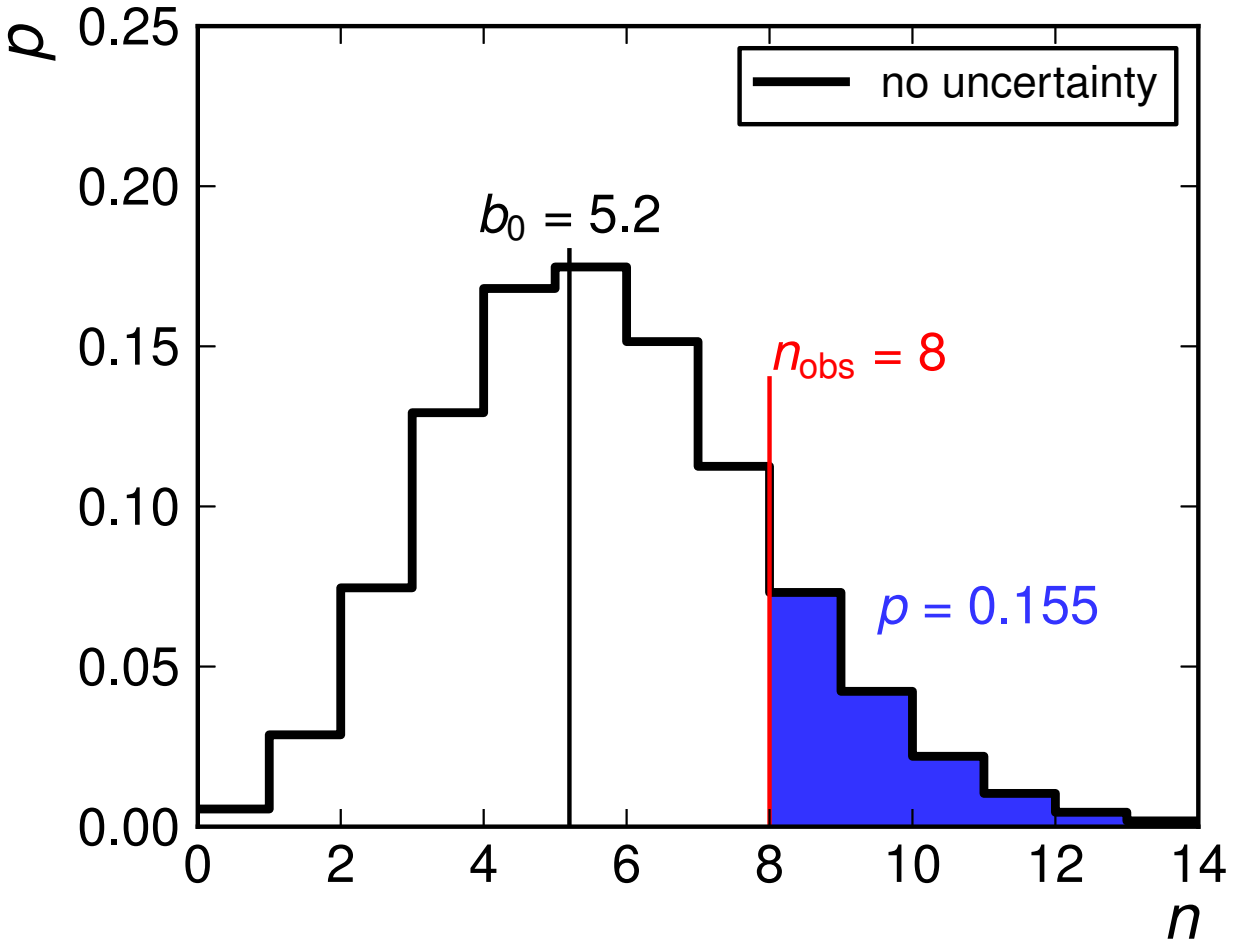
Formally, the resulting p-value is:

$$p_a = \int_\theta P(t > t_{\text{obs}}|\theta)\pi(\theta)\mathrm{d}\theta$$

where $\pi(\theta)$ is the prior for the nuisance parameters θ.

Example I: Counting Experiment

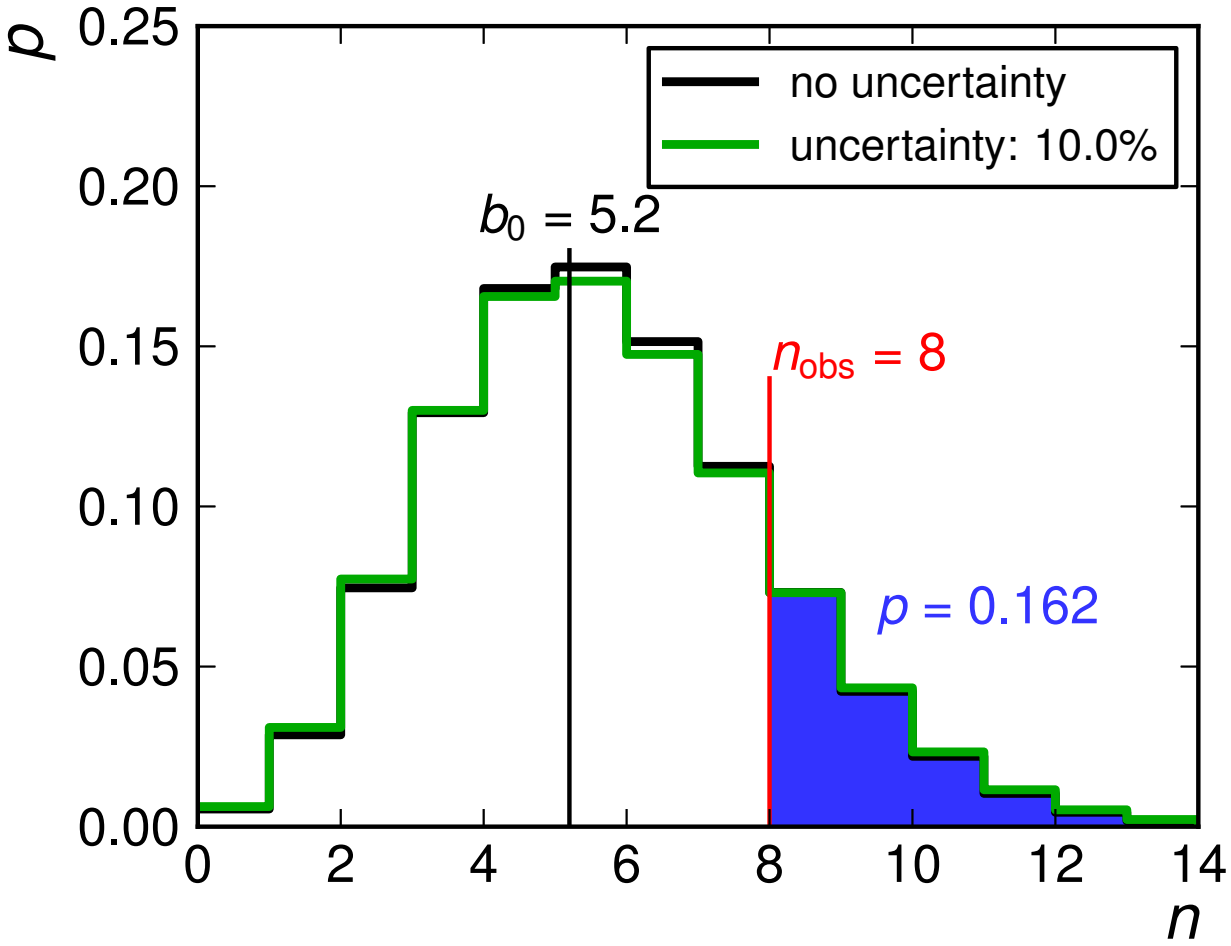
For a normal prior on b with mean $b_0 = 5.2$ distribution for n changes:



In general: adding systematic uncertainties "broadens" the test statistic distribution, thus enlarging the p-value, reducing the Z-value.

Example I: Counting Experiment

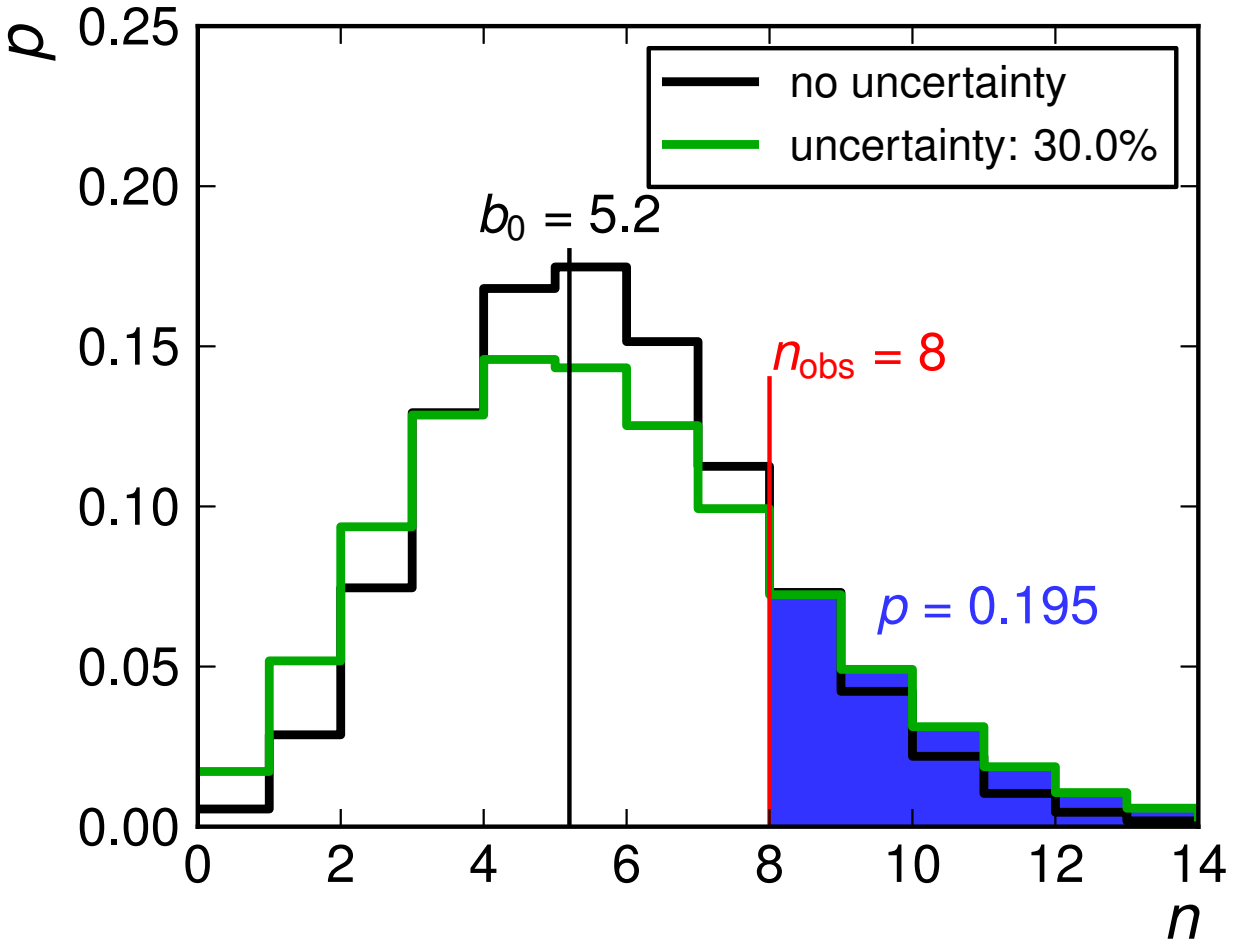
For a normal prior on b with mean $b_0 = 5.2$ distribution for n changes:



In general: adding systematic uncertainties "broadens" the test statistic distribution, thus enlarging the p-value, reducing the Z-value.

Example I: Counting Experiment

For a normal prior on b with mean $b_0 = 5.2$ distribution for n changes:



In general: adding systematic uncertainties "broadens" the test statistic distribution, thus enlarging the p-value, reducing the Z-value.

Contents

Test Statistic

Test Statistic t defined as ratio of profile likelihoods for null and alternative:

$$t = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)} = \log \frac{\max_{\mu \geq 0} L(\mu|d)}{L(\mu = 0|d)}.$$

Now, model parameters θ include the parameter of interest (signal strength μ) and nuisance parameters θ_n: $\theta = (\mu, \theta_n)$. Change t:

1. Fix nuisance parameters to most probable value $\theta_{n,0}$ in maximization, i.e. only vary μ:

$$t' = \log \frac{\max_{\mu \geq 0} L(\mu, \theta_n = 0|d)}{L(\mu = 0, \theta_n = 0|d)}$$

2. Replace L with the posterior, i.e. multiply by the nuisance parameter prior π:

$$\tilde{t} = \log \frac{\max_{\mu \geq 0, \theta_n} L(\mu, \theta_n|d) \times \pi(\theta_n)}{\max_{\theta_n} L(\mu = 0, \theta_n|d) \times \pi(\theta_n)}$$

Test Statistic

Test Statistic t defined as ratio of profile likelihoods for null and alternative:

$$t = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)} = \log \frac{\max_{\mu \geq 0} L(\mu|d)}{L(\mu = 0|d)}.$$

Now, model parameters θ include the parameter of interest (signal strength μ) and nuisance parameters θ_n: $\theta = (\mu, \theta_n)$. Change t:

1. Fix nuisance parameters to most probable value $\theta_{n,0}$ in maximization, i.e. only vary μ:

$$t' = \log \frac{\max_{\mu \geq 0} L(\mu, \theta_n = 0|d)}{L(\mu = 0, \theta_n = 0|d)}$$

2. Replace L with the posterior, i.e. multiply by the nuisance parameter prior π:

$$\tilde{t} = \log \frac{\max_{\mu \geq 0, \theta_n} L(\mu, \theta_n|d) \times \pi(\theta_n)}{\max_{\theta_n} L(\mu = 0, \theta_n|d) \times \pi(\theta_n)}$$

Test Statistic

Test Statistic t defined as ratio of profile likelihoods for null and alternative:

$$t = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)} = \log \frac{\max_{\mu \geq 0} L(\mu|d)}{L(\mu = 0|d)}.$$

Now, model parameters θ include the parameter of interest (signal strength μ) and nuisance parameters θ_n: $\theta = (\mu, \theta_n)$. Change t:

1. Fix nuisance parameters to most probable value $\theta_{n,0}$ in maximization, i.e. only vary μ:

$$t' = \log \frac{\max_{\mu \geq 0} L(\mu, \theta_n = 0|d)}{L(\mu = 0, \theta_n = 0|d)}$$

2. Replace L with the posterior, i.e. multiply by the nuisance parameter prior π:

$$\tilde{t} = \log \frac{\max_{\mu \geq 0, \theta_n} L(\mu, \theta_n|d) \times \pi(\theta_n)}{\max_{\theta_n} L(\mu = 0, \theta_n|d) \times \pi(\theta_n)}$$

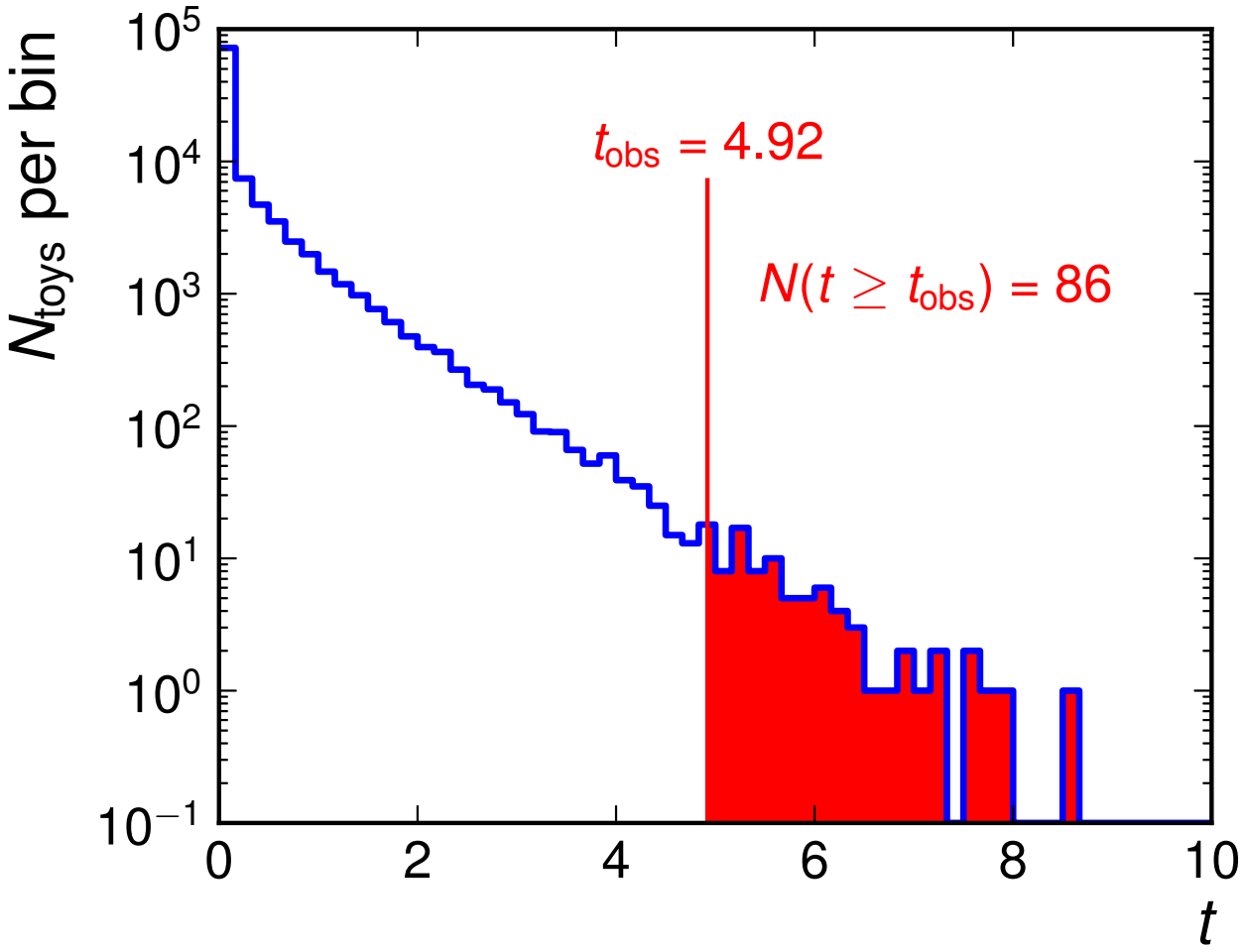
Test Statistic and p-value

Both t' and $\tilde{t}$ have been used in HEP analyses.

The definition of the test statistic is orthogonal to the definition of the ensemble H_0 used to define the p-value:
For a MC method, this means it is crucial to vary the nuisance parameters in the toy data generation, while it's not necessary to vary them in the definition of t.

Applying uncertainties

Using the prior averaging method means that toy data is drawn for random values for $\theta_u \rightsquigarrow$ test statistic distribution is broadened:

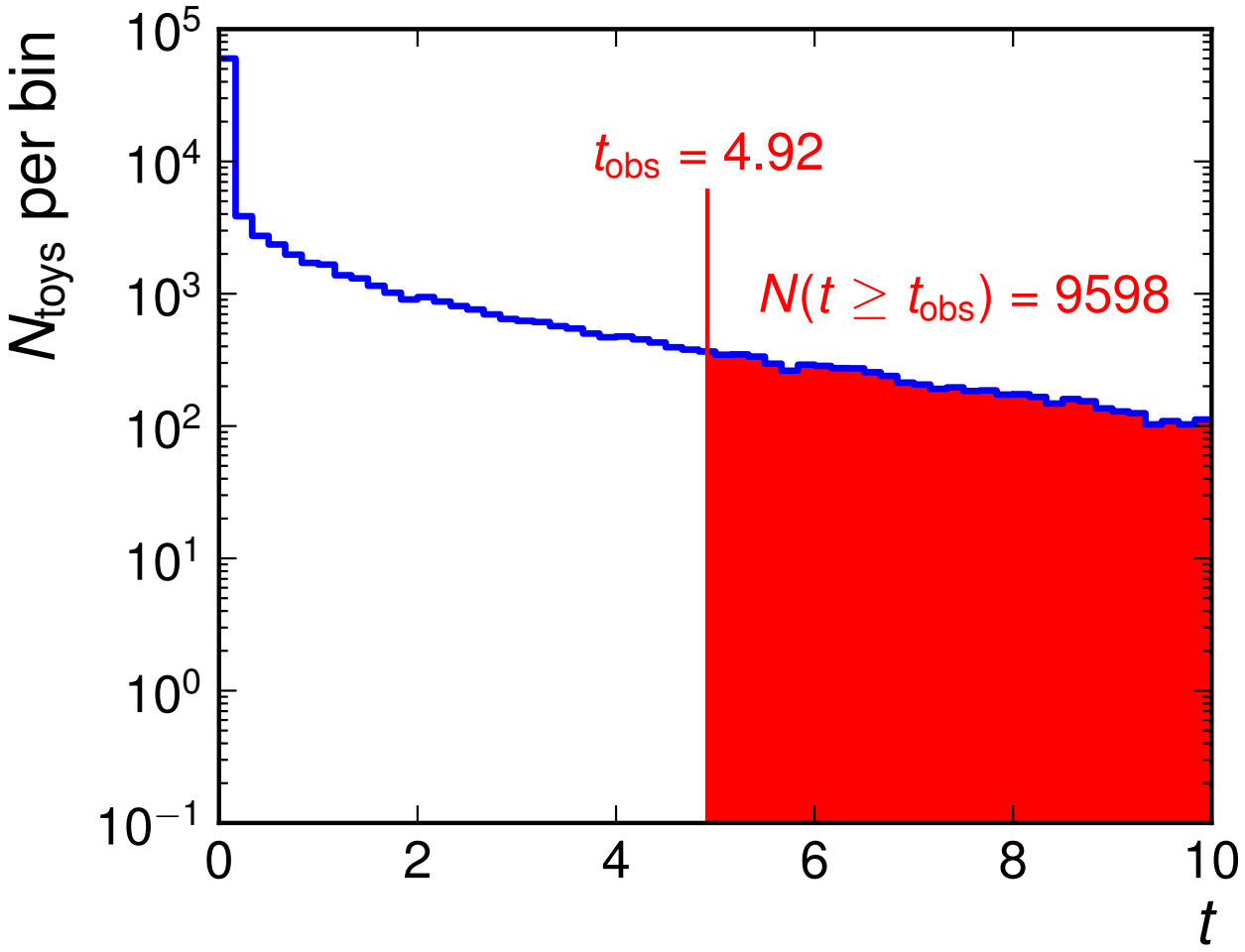


No uncertainties:
$\hat{p} = 0.00086$, $\hat{Z} = 3.1$.

With 10% rate uncertainty: $\hat{p} = 0.096$, $\hat{Z} = 1.3$.

Applying uncertainties

Using the prior averaging method means that toy data is drawn for random values for $\theta_u \rightsquigarrow$ test statistic distribution is broadened:

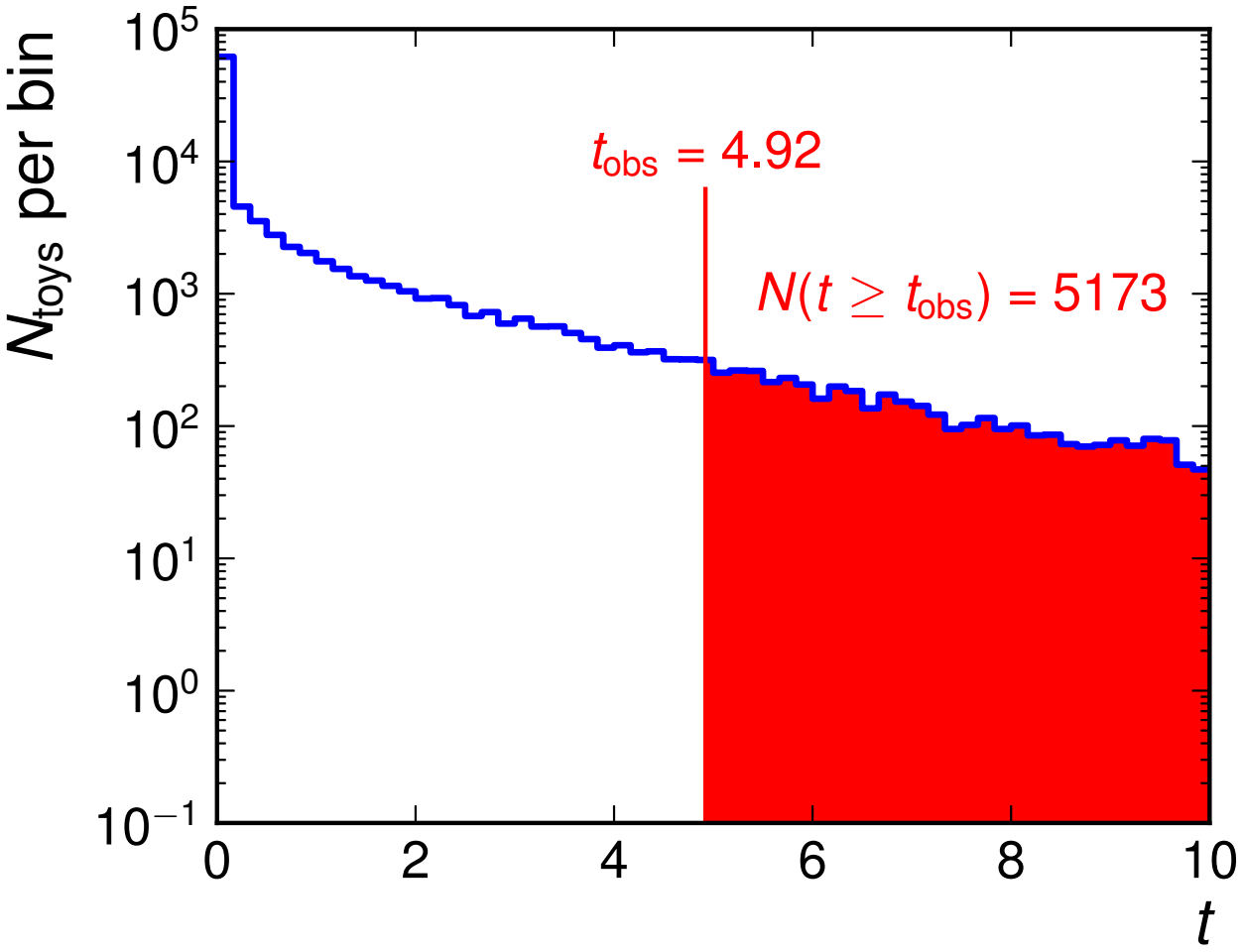


No uncertainties:
$\hat{p} = 0.00086$, $\hat{Z} = 3.1$.

With 10% rate uncertainty: $\hat{p} = 0.096$, $\hat{Z} = 1.3$.

Applying uncertainties

Using the prior averaging method means that toy data is drawn for random values for $\theta_u \rightsquigarrow$ test statistic distribution is broadened:



No uncertainties: $\hat{p} = 0.00086$, $\hat{Z} = 3.1$.

With 10% rate uncertainty: $\hat{p} = 0.096$, $\hat{Z} = 1.3$.

Part II

Confidence Intervals

Introduction and Definitions

Confidence intervals are probabilistic statement about the value of parameters of a statistical model. They are calculated at a given confidence level which specifies the (claimed/desired) coverage.

The coverage is the probability that the interval contains the true value. In general, the coverage is a function of the true parameter values. (Note that this does not assign probabilities to the true value but only to the interval construction!)

A method is said to over-cover (and conservative) if the coverage is above the confidence level; the opposite is under-coverage. Sometimes, exact coverage cannot be reached (e.g. due to discrete data); in this case one usually chooses to construct the method to over-cover.

Introduction and Definitions

Confidence intervals are probabilistic statement about the value of parameters of a statistical model. They are calculated at a given confidence level which specifies the (claimed/desired) coverage.

The coverage is the probability that the interval contains the true value. In general, the coverage is a function of the true parameter values. (Note that this does not assign probabilities to the true value but only to the interval construction!)

A method is said to over-cover (and conservative) if the coverage is above the confidence level; the opposite is under-coverage. Sometimes, exact coverage cannot be reached (e.g. due to discrete data); in this case one usually chooses to construct the method to over-cover.

Counting Experiment

Consider the simple counting model with $b = 5.2$ and unknown $s \geq 0$ with

$$p(n|s) = \text{Poisson}(n|s + b).$$

For a given observation (e.g. $n_{\text{obs}} = 5$), what statement can be made about s?
For example, we would expect to rule out large values for s, e.g. $s = 100$.

This is a question for a hypothesis test.

Counting Experiment

Consider the simple counting model with $b = 5.2$ and unknown $s \geq 0$ with

$$p(n|s) = \text{Poisson}(n|s + b).$$

For a given observation (e.g. $n_{\text{obs}} = 5$), what statement can be made about s?
For example, we would expect to rule out large values for s, e.g. $s = 100$.

This is a question for a hypothesis test.

Intervals and Hypothesis Tests

Upper limit construction by hypothesis test "inversion":

1. For a given $s = s_0$, make a hypothesis test with the null hypothesis $s = s_0$ and the alternative $s < s_0$ with type-I error α (e.g., $\alpha = 0.05$).
2. Repeat step 1 for different values of s_0.
3. The confidence interval for s comprises exactly those values s_0 for which the hypothesis test could not reject the null hypothesis $s = s_0$.

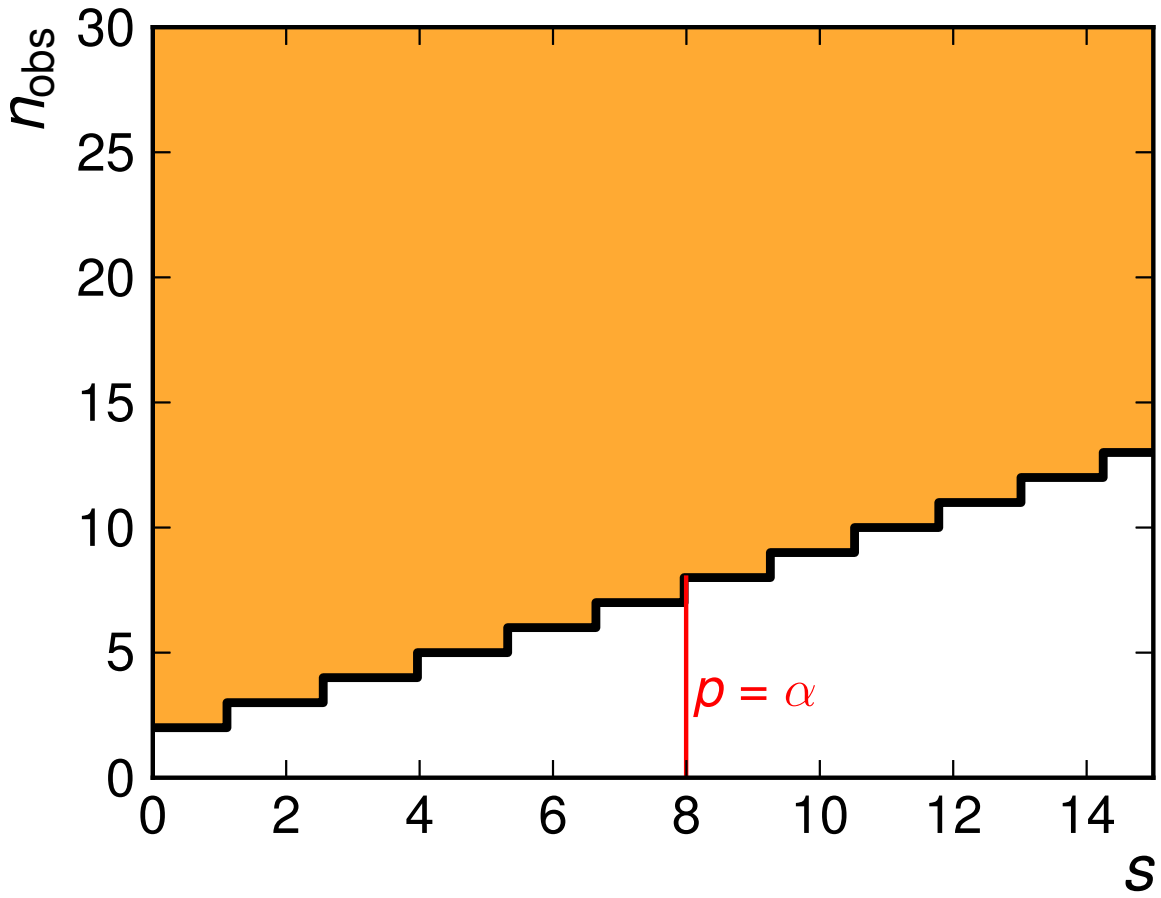
For this formulation of the hypothesis test ($s = s_0$ vs. $s < s_0$), we get an upper limit.

The confidence level is $(1 - \alpha)$ (here: 95%).

This is known as the *Neyman Construction*. It can be visualized as "belt construction" on the (n–s) plane.

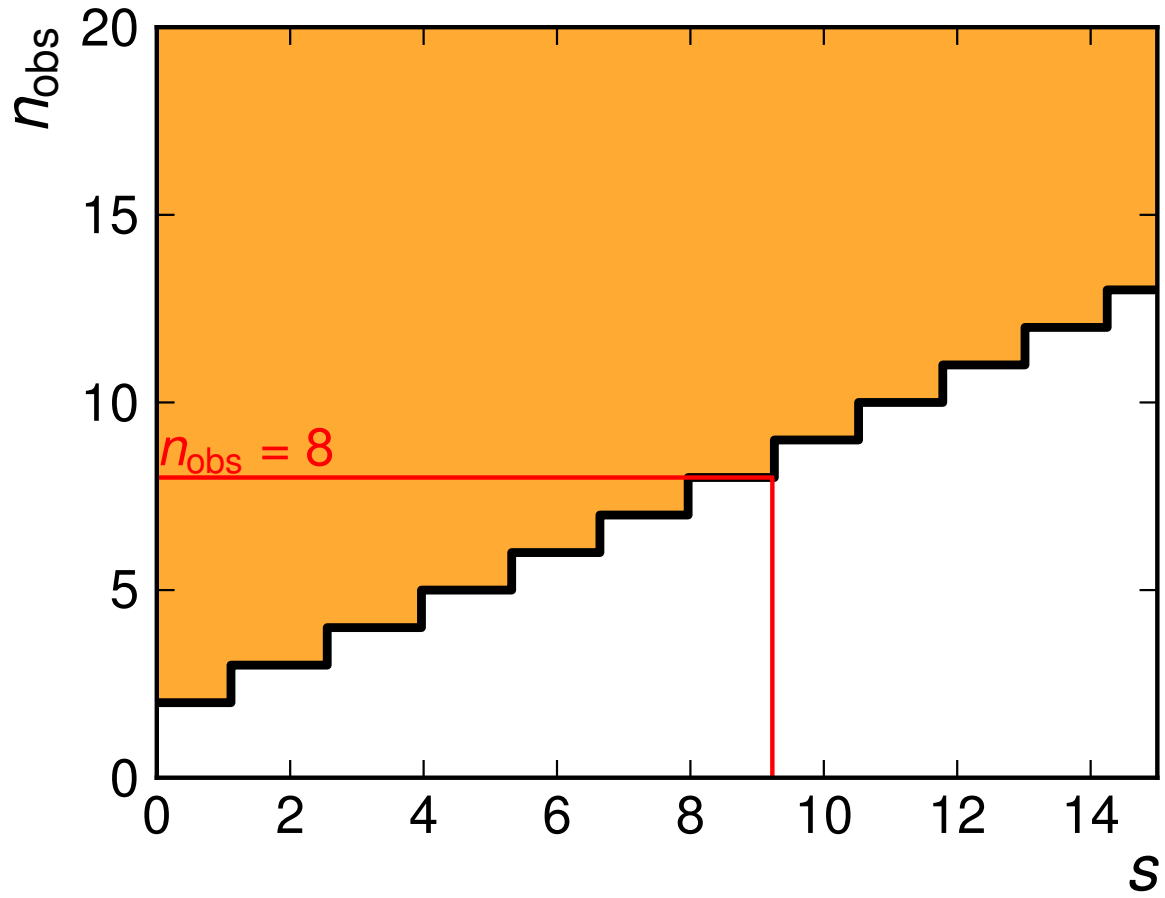
Belt

The Neyman construction as "belt" in the s–n_{obs} plane: For each s, the "belt" in n_{obs} has probability $\geq 1 - \alpha$. For given n_{obs}, the interval is given by intersecting the horizontal line with the belt.



Belt

The Neyman construction as "belt" in the s–n_{obs} plane: For each s, the "belt" in n_{obs} has probability $\geq 1 - \alpha$. For given n_{obs}, the interval is given by intersecting the horizontal line with the belt.



Comment

The close relationship with hypothesis testing means many aspects from HT also apply to limits:

- explicitly introduce test statistic for shape models; again: Test statistic choice not unique, can use profile likelihood ratio t
- use toy Monte-Carlo to get test statistic distribution for H_0
- handling of systematic uncertainties via Bayesian/hybrid method
- use of asymptotic methods
- concept of "expected limit" by running the method (imaginatively or with MC) on an ensemble representing the expected data (usually background-only)
- ...

Extensions

So far, discussed Neyman construction for limits. Modifications:

- Modify "belt" construction: for each s, include the "central" set of n_{obs} into belt $\rightsquigarrow$ "central intervals", which are two-sided
- Use likelihood ratio as ordering criterion to decide what to include in the belt $\rightsquigarrow$ Feldman-Cousins "unified intervals" with smooth transition between limit-like and central-like intervals
- Purely frequentist methods might lead to empty intervals $\rightsquigarrow$ modify frequentist p-values by a "penalty factor" which is high if experimental sensitivity is low $\rightsquigarrow$ modified frequentist CLs limits

Part III

theta

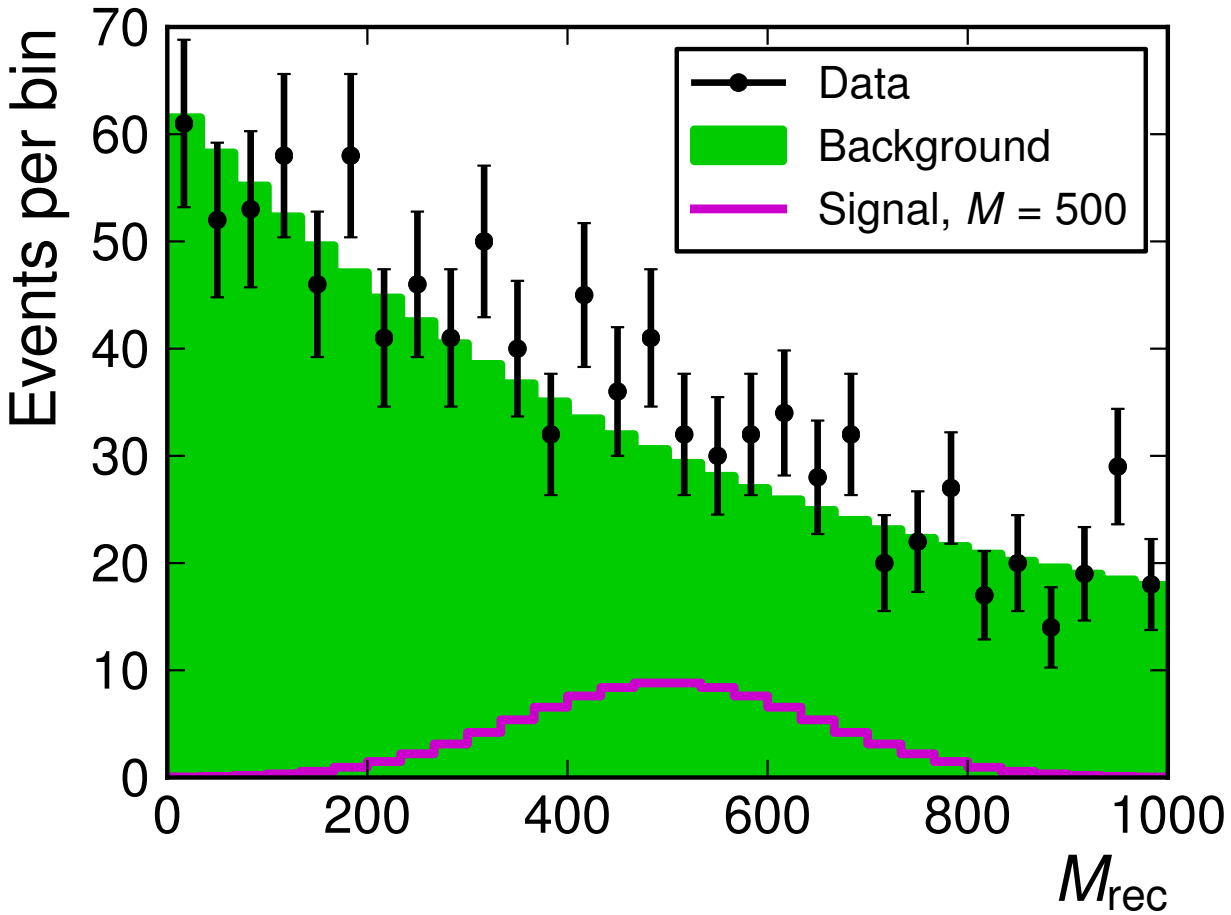
6 Statistical Model

7 Using theta

Contents

Overview

theta is a program for statistical inference. It only handles binned models that use histograms for the prediction such as the example:



Important limits: one bin = counting experiment; many narrow bins $\rightsquigarrow$ almost equivalent to unbinned case

Formal Shape Model

The statistical model is the product of Poisson in each bin:

$$p(n|\theta) = \prod_i \text{Poisson}(n_i|\lambda_i(\theta))$$

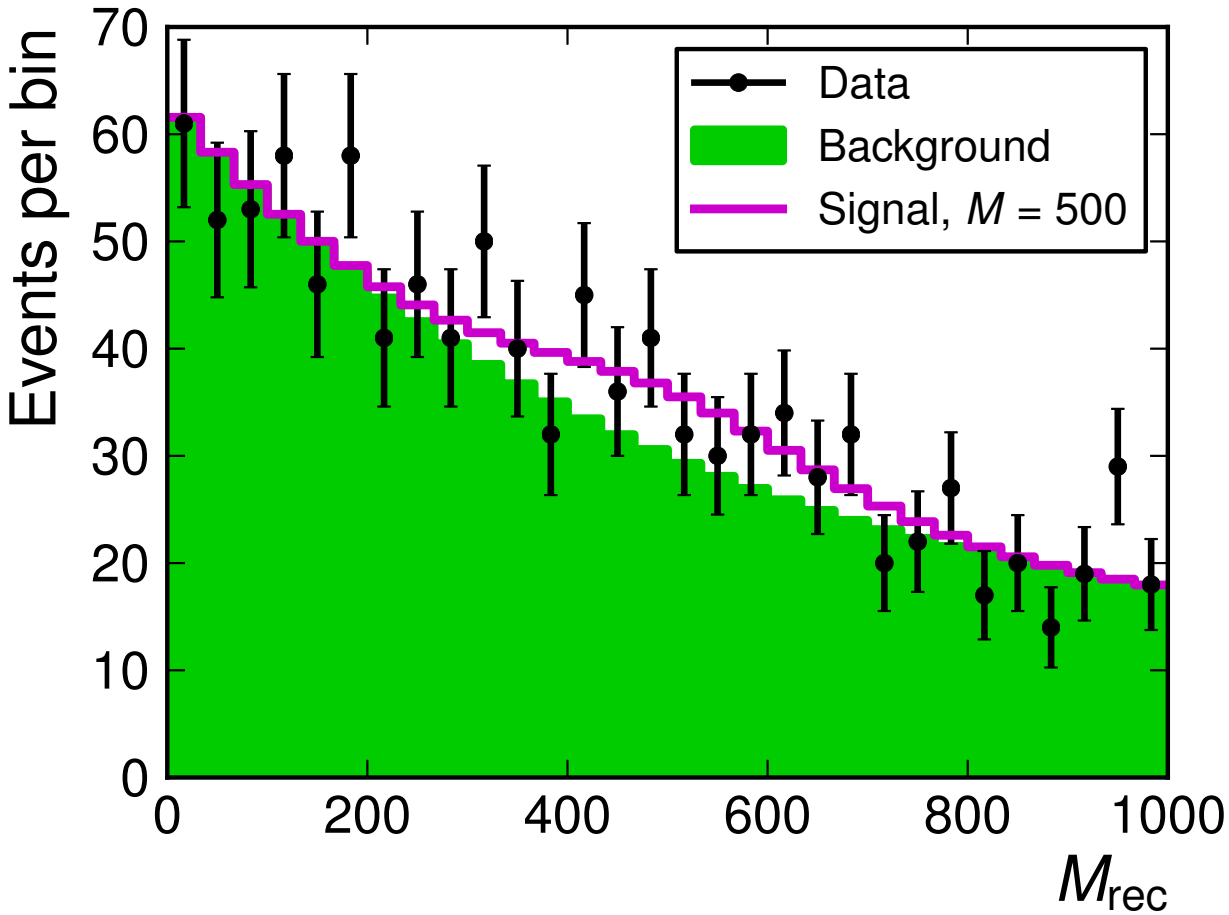
where i is the bin index and the expected number of events in bin i, λ_i, is given by the sum of (scaled) signal and background histograms:

$$\lambda_i(\theta) = \mu s_i + \sum_p c_p(\theta_n) b_{pi}(\theta_n).$$

p denotes the different background processes which are expected to contribute. The bin-independent coefficient $c_p(\theta_n)$ encodes (process-specific) rate uncertainties, while $b_{pi}(\theta_n)$ is the most general dependence, a shape uncertainty.

Rate Uncertainties: Example

In the shape example model, we might estimate a 10% uncertainty on the overall rate of the background:

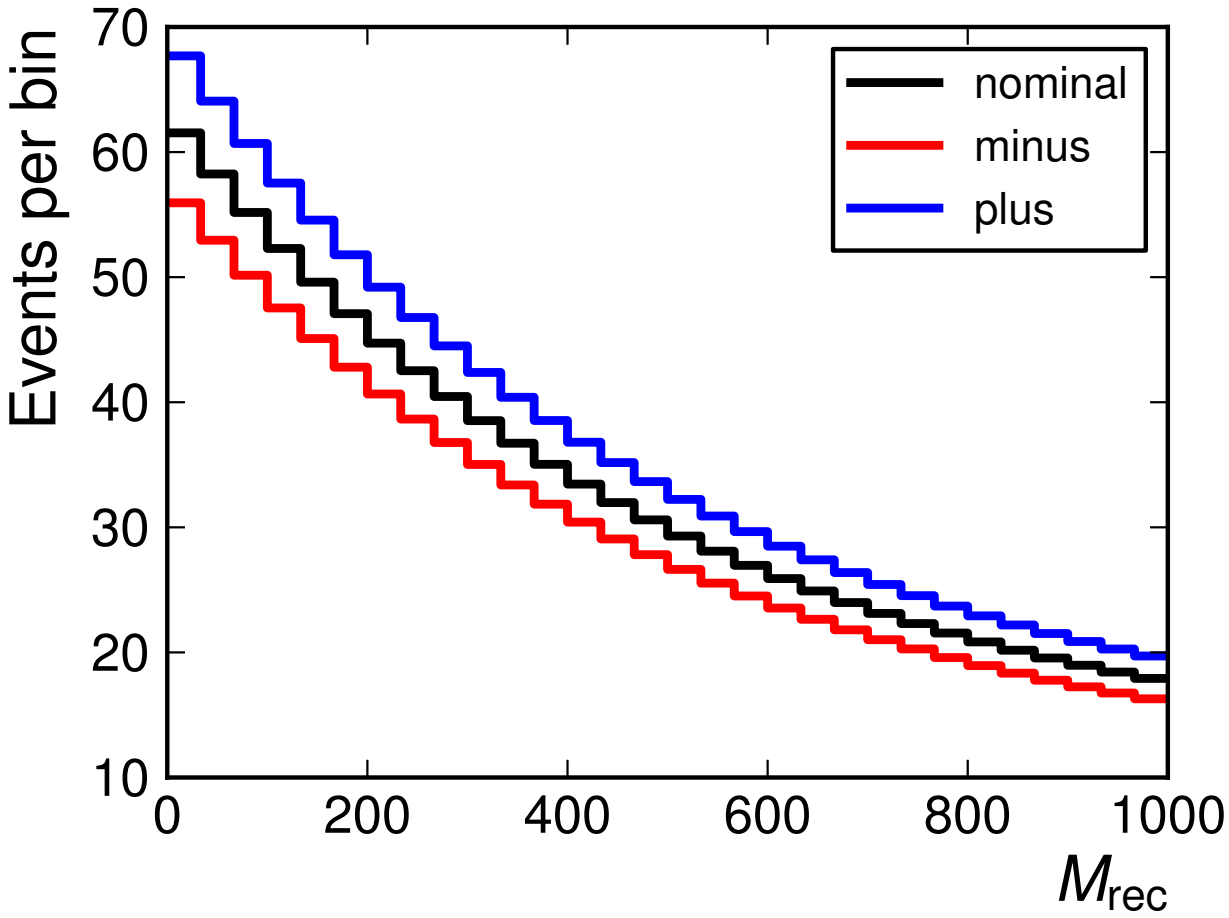


$\rightsquigarrow$ Have "plus" and "minus" templates by scaling the "nominal template up and down by 10%.

$\rightsquigarrow$ introduce nuisance parameter which scales the template accordingly (see backup for implementation).

Rate Uncertainties: Example

In the shape example model, we might estimate a 10% uncertainty on the overall rate of the background:

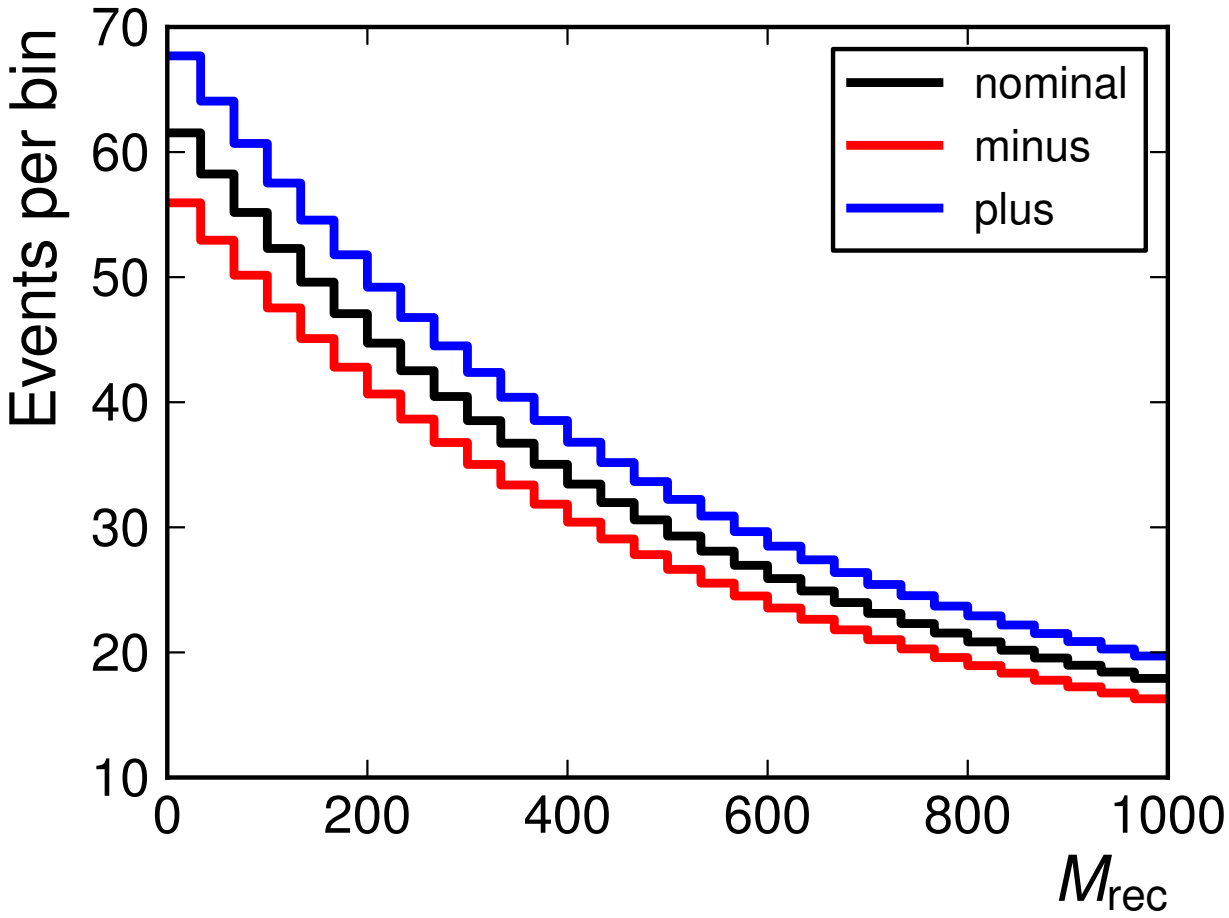


$\rightsquigarrow$ Have "plus" and "minus" templates by scaling the "nominal template up and down by 10%.

$\rightsquigarrow$ introduce nuisance parameter which scales the template accordingly (see backup for implementation).

Rate Uncertainties: Example

In the shape example model, we might estimate a 10% uncertainty on the overall rate of the background:

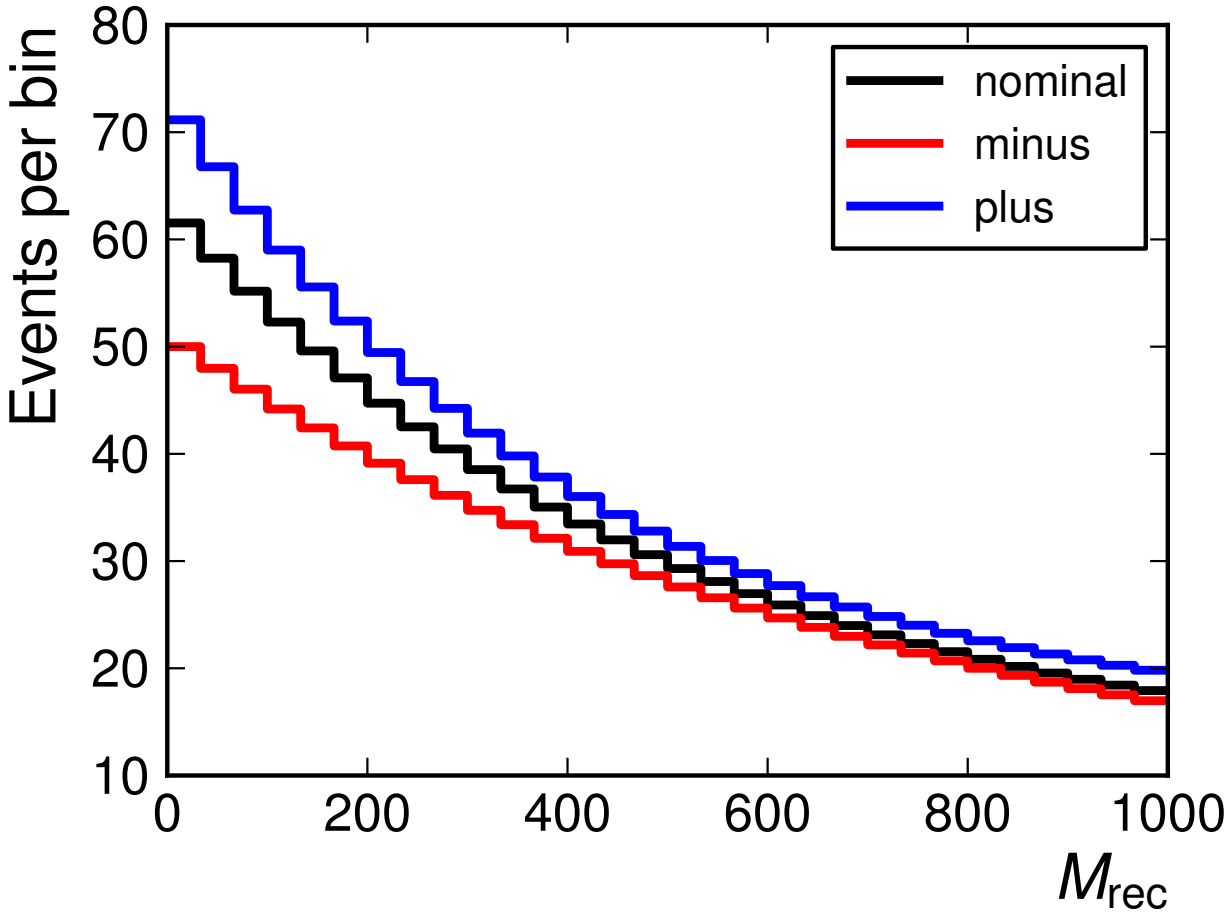


$\rightsquigarrow$ Have "plus" and "minus" templates by scaling the "nominal template up and down by 10%.

$\rightsquigarrow$ introduce nuisance parameter which scales the template accordingly (see backup for implementation).

Shape Uncertainty: Example

In the shape model, assume you have an uncertainty affecting the background shape like this:



$\rightsquigarrow$ Can introduce the nuisance parameter and use it in the model to interpolate smoothly between the three templates, if assuming the "plus" and "minus" correspond to a $\pm 1\sigma$ uncertainty (see backup for details)

Contents

theta Overview

theta is a software package for making statistical inference using *template*-based models.

- The "core" of theta is a (heavily optimized) C++ program, controlled by a text-based cfg-file, writing the result to a sqlite database or a root tree.
- A python interface "theta-auto" can be used for many cases which makes usage easier than writing the cfg-file manually.

Here: concentrate on the python interface.
Usually, the script has two separate steps: Building the statistical model (as class `Model`), and then applying different statistical methods.

Building Models I

Formally, the general statistical model in theta is

$$p(n|\theta) = \prod_i \text{Poisson}(n_i|\lambda_i(\theta))$$

where i is the bin index and the mean expected number of events in bin i, λ_i, is given by the sum of (scaled) signal and background histograms:

$$\lambda_i(\theta) = \mu s_i(\theta) + \sum_p c_p(\theta_n) b_{pi}(\theta_n).$$

where p runs over the background processes.

The python interface supports log-normal coefficients for c_p and the template morphing for $b_{pi}(\theta_n)$, as discussed earlier. theta also supports handling MC statistical uncertainties with the "Barlow-Beeston light" method.

Building Models II

In theta-auto, models can be built e.g. from a text-based "datacard" (as used for "combine"), which specifies the model and the observed data:

```
imax 1  number of channels
jmax 3  number of backgrounds
kmax 2  number of nuisance parameters
bin  1
observation  0
bin                    1      1      1      1
process              ggH  qqWW  ggWW others
process                0      1      2      3
rate                1.47   0.63   0.06   0.22
lumi      lnN       1.11      -   1.11      -
xs_ggWW lnN            -      -   1.50      -
```

Note that it is possible to include template morphing by refering to histograms in a root file. See SWGuideHiggsAnalysisCombinedLimit for details (CMS-internal).

Method Overview

Once the statistical model is built, one can use it to

- perform a maximum likelihood fit to get estimates for all model parameters
- compute profile likelihood intervals
- compute toy-based or asymptotic CLs limits (observed and expected)
- evaluate the posterior using Markov-Chain Monte-Carlo (Bayesian method for intervals and limits)
- ...

Example: Maximum Likelihood Estimate

Most methods are implemented as a single python function, taking the model, the "input" dataset and the number of evaluations as input. To make the fit on data, use

```
res = mle(model, 'data', 1)
print res
```

To make the fit on 1000 toys with $\mu = 1.2$, use:

```
res = mle(model, 'toys:1.2', 1000)
# do some calculations with res
```

where the "calculations" could e.g. study the bias and pull of the maximum likelihood estimate.

More information on theta: `http://theta-framework.org/`.

Comments on Maximum Likelihood

The Maximum likelihood estimate has some nice asymptotic properties:

- The bias goes to zero
- Among the unbiased estimators, it has the smallest variance
- Can use curvature of negative log-likelihood to estimate variance

$\rightsquigarrow$ Maximum Likelihood is the "default choice" for an estimator in HEP.

Summary

Introduced concepts in one sentence:

- Hypothesis tests: The p-value is the probability to observe at least as "extreme" data.
- Test Statistic is the measure of "extreme".
- Neyman Interval construction as an "inversion" of the hypothesis test: Include those points μ_0 in the interval for which we can not reject the null hypothesis $\mu = \mu_0$.
- Handle rate (shape) uncertainties in the statistical model by introducing a nuisance parameter scaling (interpolating) the template

Part IV

Backup

8 Misc

9 Frequentist Interpretation; Bootstrapping

10 Rate Uncertainty Implementation

11 Shape Uncertainties

12 BB light method

Contents

p-value Distribution for Discrete Data

The p-value is defined to observe at least as extreme data for H_0.

If the data in the statistical model is discrete, the p-value can't follow a proper uniform distribution on $[0, 1]$.

In general (also for discrete data):

$$\Pr(p \leq p_0 | H_0) \leq p_0 \qquad \text{for } 0 \leq p_0 \leq 1.$$

In words: The probability to observe a p-value below some threshold p_0 is at most p_0 (and if equality holds, p-value is indeed uniform on $[0, 1]$).

"Expected" vs. "Observed" Result

Consider two counting experiments A, B searching for the same signal, both expecting background $b = 100$, and signals $s_A = 20$ and $s_B = 15$, clearly indicating a better performance for A.

By chance, $n_A = 120$, $s_B = 120$, giving $Z_A \approx 1$ and $Z_B \approx 1.3$, so using the "observed" significance, experiment B is "better", which is of course nonsense.

Related issue in the statement: "Experiment A sees 3σ effect, experiment B sees 3.5σ effect, so in summary we have a 3.5σ effect" (or similar statement for limits).
However, this is wrong from the statistical point of view, as minimum of two p-values is not a proper p-value (refer to look-elsewhere effect).
$\rightsquigarrow$ solution: Always use *expected significance* and decide which analysis/experiment to use without using the (random) data result.

"Expected" vs. "Observed" Result

Consider two counting experiments A, B searching for the same signal, both expecting background $b = 100$, and signals $s_A = 20$ and $s_B = 15$, clearly indicating a better performance for A.

By chance, $n_A = 120$, $s_B = 120$, giving $Z_A \approx 1$ and $Z_B \approx 1.3$, so using the "observed" significance, experiment B is "better", which is of course nonsense.

Related issue in the statement: "Experiment A sees 3σ effect, experiment B sees 3.5σ effect, so in summary we have a 3.5σ effect" (or similar statement for limits).
However, this is wrong from the statistical point of view, as minimum of two p-values is not a proper p-value (refer to look-elsewhere effect).
$\rightsquigarrow$ solution: Always use *expected significance* and decide which analysis/experiment to use without using the (random) data result.

Contents

Model Reminder

The observed data can be summarized as the number of observed events n. The probability to observe n events is given by a Poisson probability:

$$p(n|\theta) = \text{Poisson}(n|\lambda(\theta)).$$

The Poisson mean $\lambda(\theta)$ is given by the sum of (scaled) signal and background yields,

$$\lambda_i(\theta) = \mu s + b(\theta_n),$$

where the model parameters θ comprise the signal strength parameter μ and the nuisance parameters θ_n: $\theta = (\mu, \theta_n)$.

External knowledge about the nuisance parameters is encoded in the prior $\pi(\theta_n)$.

Frequentist Interpretation

The Bayesian posterior f is given by the likelihood times the prior (assumed to be normal here):

$$f(\theta) = L(\theta|d) \times \mathcal{N}(\theta_n),$$

(apart from an unimportant normalization). $f(\theta)$ can be used in place of plain L at many places (e.g. parameter estimation, definition of t).

Frequentist re-interpretation:

$$f(\theta) \propto L(\theta|d) \times \prod_u e^{-\frac{(\theta_u - \mu_u)^2}{2\delta_u^2}}$$

where $\mu_u = 0$ and $\delta_u = 1$, u runs over all nuisance parameters.
This can be interpreted as the likelihood function of a slightly different model by swapping θ_u and μ_u. The μ_u now are random variables, part of the data. The data comprise the number of observed events and the values for μ_u (with $\mu_u = 0$ for the observed data).

Frequentist Interpretation

The Bayesian posterior f is given by the likelihood times the prior (assumed to be normal here):

$$f(\theta) = L(\theta|d) \times \mathcal{N}(\theta_n),$$

(apart from an unimportant normalization). $f(\theta)$ can be used in place of plain L at many places (e.g. parameter estimation, definition of t).

Frequentist re-interpretation:

$$f(\theta) \propto L(\theta|d) \times \prod_u e^{-\frac{(\theta_u - \mu_u)^2}{2\delta_u^2}}$$

where $\mu_u = 0$ and $\delta_u = 1$, u runs over all nuisance parameters.
This can be interpreted as the likelihood function of a slightly different model by swapping θ_u and μ_u. The μ_u now are random variables, part of the data. The data comprise the number of observed events and the values for μ_u (with $\mu_u = 0$ for the observed data).

Frequentist Interpretation: Comments

- No longer need (Bayesian) concept of prior for model parameters θ_u; instead, have extended the data by μ_u.
- Allows to use purely frequentist concepts for defining ensembles of toys data; but: requires to choose parameter values.
- Choose parameter values by fitting to data: "(parametric) bootstrapping".
- If want to keep structure for f, have to use conjugate distribution for μ_u in the frequentist model. Normal distribution is self-conjugate $\rightsquigarrow$ use normal model for distribution of μ_u.

Updated Monte-Carlo Method for p-value

For the p-value calculation with Monte-Carlo, the steps are modified:

1. Make a maximum likelihood fit to data (with null hypothesis H_0) to get estimates for nuisance parameters θ_u, $\tilde{\theta}_u$.
2. Generate toy data by sampling from the model at the fitted values for θ_u; in particular, draw a Gaussian for μ_u around $\tilde{\theta}_u$ with width 1.

For each toy data, calculate the test statistic value, e.g. using the t' or $\tilde{t}$ definitions. The fraction of toys for which $t \geq t_{\text{obs}}$ is the p-value.

Summary; Comments

The expression for the posterior can be interpreted purely frequentist way of a slightly different statistical model with an extended dataset. For that model, can apply parametric bootstrapping and proceed with a purely frequentist framework.

Notes:

- The frequentist approach allows the application of asymptotic formulas
- This is the method used in the LHC Higgs combination.

Contents

Rate Uncertainties: Implementation

In the statistical model, had expression for Poisson mean

$$\lambda_i(\theta) = \mu s_i + \sum_p c_p(\theta_n) b_{pi}(\theta_n).$$

Rate uncertainties can be included in the coefficient c_p, e.g. by using

$$c_p(\theta_u) = \theta_u$$

where θ_u has a log-normal prior around 1 (see Backup.). Equivalently, we can also use:

$$c_p(\theta_u) = e^{\theta_u \Delta b}$$

where the nuisance parameter θ_u has a normal prior around 0 with standard deviation 1, which corresponds to a scale factor with a log-normal prior.

Rate Uncertainties: Implementation

In the statistical model, had expression for Poisson mean

$$\lambda_i(\theta) = \mu s_i + \sum_p c_p(\theta_n) b_{pi}(\theta_n).$$

Rate uncertainties can be included in the coefficient c_p, e.g. by using

$$c_p(\theta_u) = \theta_u$$

where θ_u has a log-normal prior around 1 (see Backup.). Equivalently, we can also use:

$$c_p(\theta_u) = e^{\theta_u \Delta b}$$

where the nuisance parameter θ_u has a normal prior around 0 with standard deviation 1, which corresponds to a scale factor with a log-normal prior.

Equivalent Parametrizations

We just saw two different, but equivalent methods to introduce a log-normal scale factor.

This is an example of a more general principle:
The statistical model can be re-parametrized, which changes both the "model response" to the nuisance parameter and the parameter prior.

Using this freedom, one can use independent standard normal priors for the nuisance parameters, which will be assumed from now on.

Rate Uncertainties: Normal vs. log-normal I/II

The uncertainty on b was implemented by using the stat. model

$$p(n|s,b) = \text{Poisson}(\lambda = s + b)$$

with a normal prior for b around known b_0 with known width Δb.

But: λ can become negative with non-zero probability, for which a Poisson is not defined.

Instead, use a *log-normal* prior for a scale factor for b_0:

$$\lambda(s,f) = s + f \cdot b_0$$

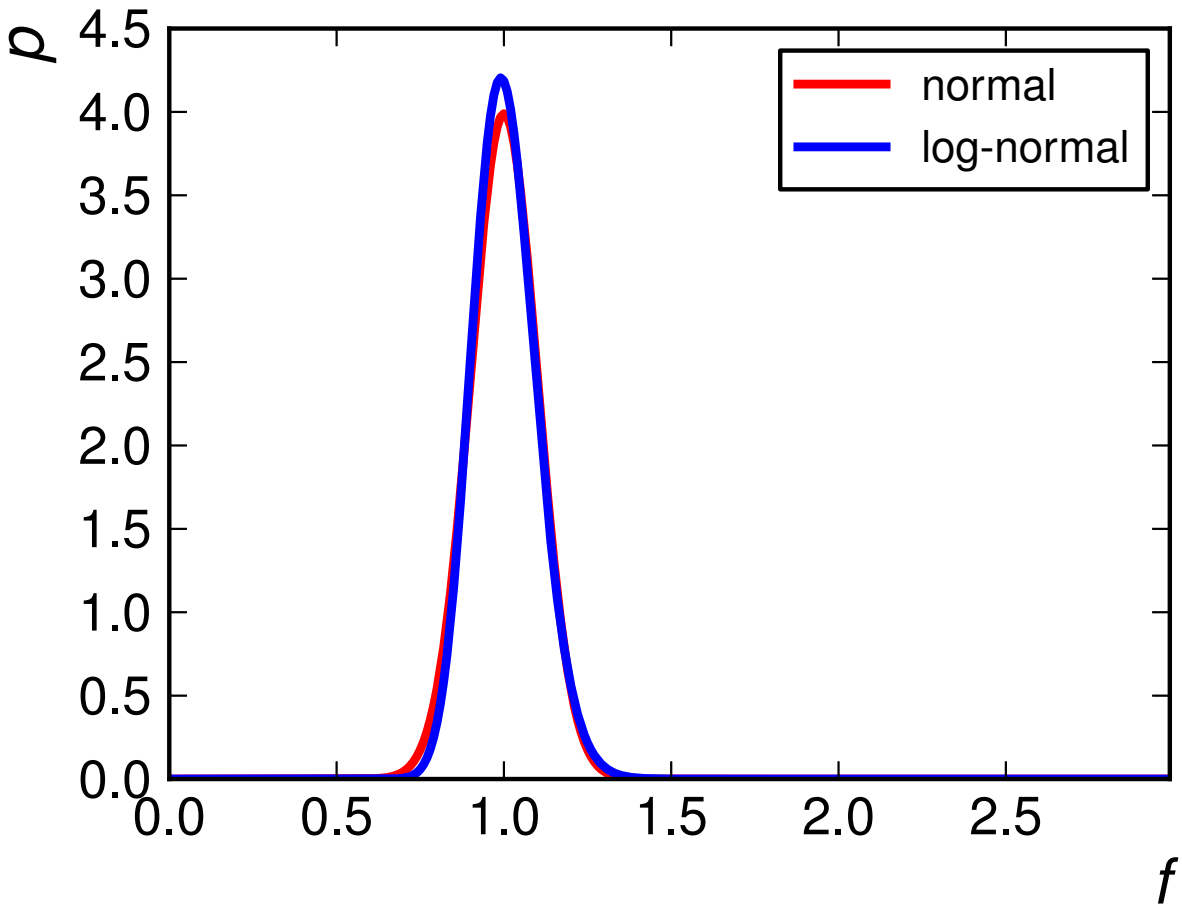
where f has a log-normal prior, i.e., $\log f$ has a normal distribution.
An equivalent formulation is

$$\lambda(s,\theta) = s + e^{\theta \log(1+\Delta b)} b_0$$

where the n.p. θ has a normal prior with mean 0 and standard deviation 1.

Rate Uncertainties: Normal vs. log-normal II/II

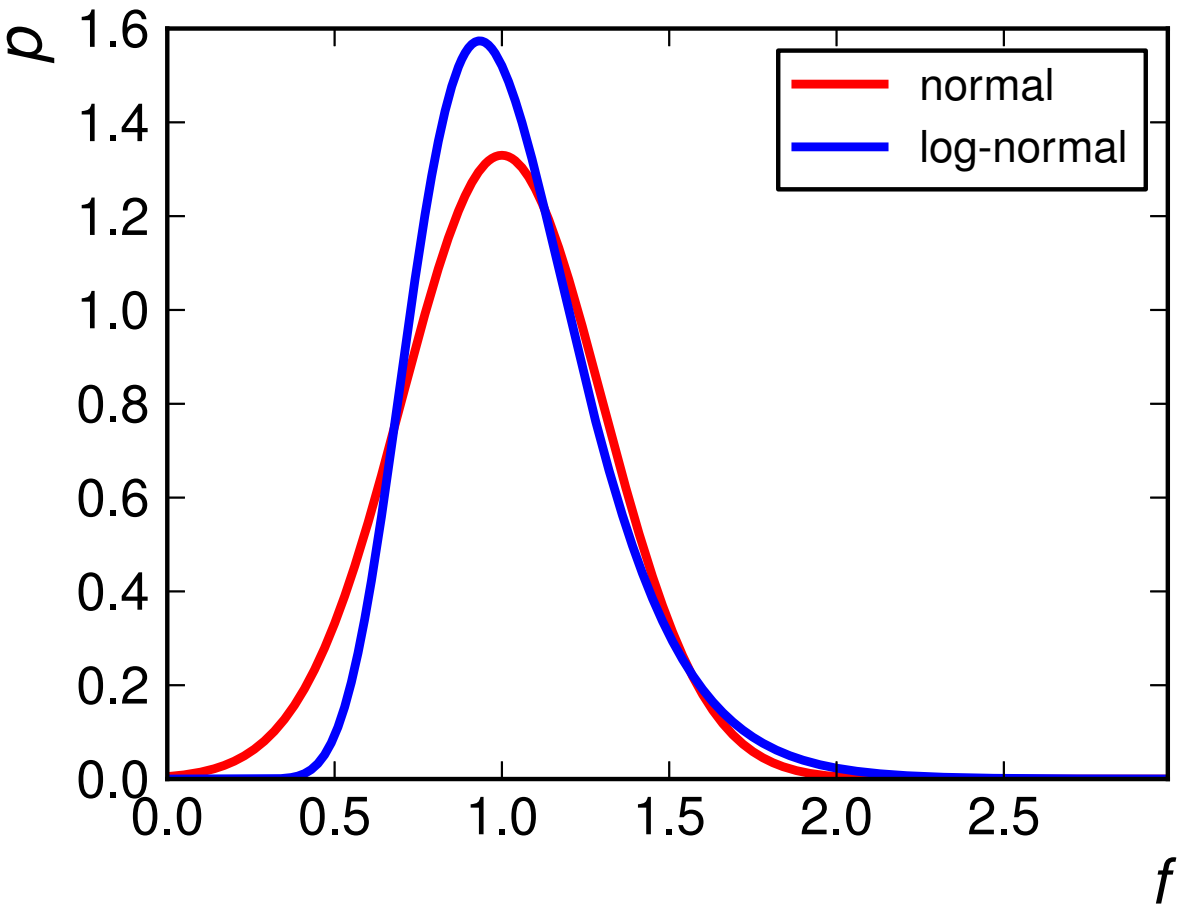
Comparing the prior for the scale factor between normal and log-normal: $\Delta b = 0.1$:



Log-normal avoids unphysical jump / truncation at $f = 0$, while normal prior requires that for large Δb.

Rate Uncertainties: Normal vs. log-normal II/II

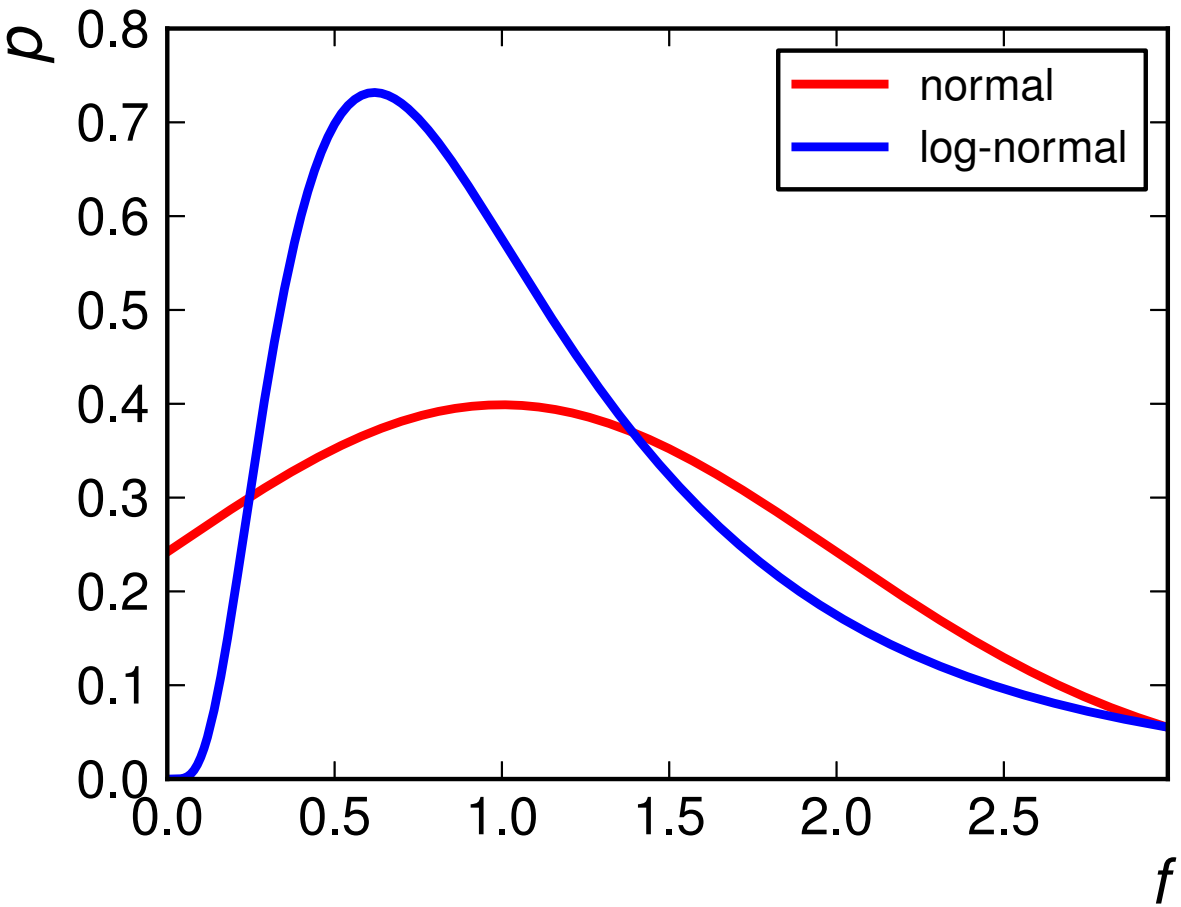
Comparing the prior for the scale factor between normal and log-normal: $\Delta b = 0.3$:



Log-normal avoids unphysical jump / truncation at $f = 0$, while normal prior requires that for large Δb.

Rate Uncertainties: Normal vs. log-normal II/II

Comparing the prior for the scale factor between normal and log-normal: $\Delta b = 1.0$:



Log-normal avoids unphysical jump / truncation at $f = 0$, while normal prior requires that for large Δb.

Contents

Shape Uncertainties: Introduction

Can have uncertainties also affecting shape in a general way (e.g. by energy calibration, . . .).

Typically, in an analysis, one would

- Use MC sample (or sideband) to get a shape for a process "nominal template"
- Modify the MC (or sideband) to get "$\pm 1\sigma$ effects" of some uncertainty, e.g. by re-weighting events, modifying the energy calibration, using a different sideband, . . .
  $\rightsquigarrow$ "plus" / "minus" template

Shape Uncertainties: Statistical Model

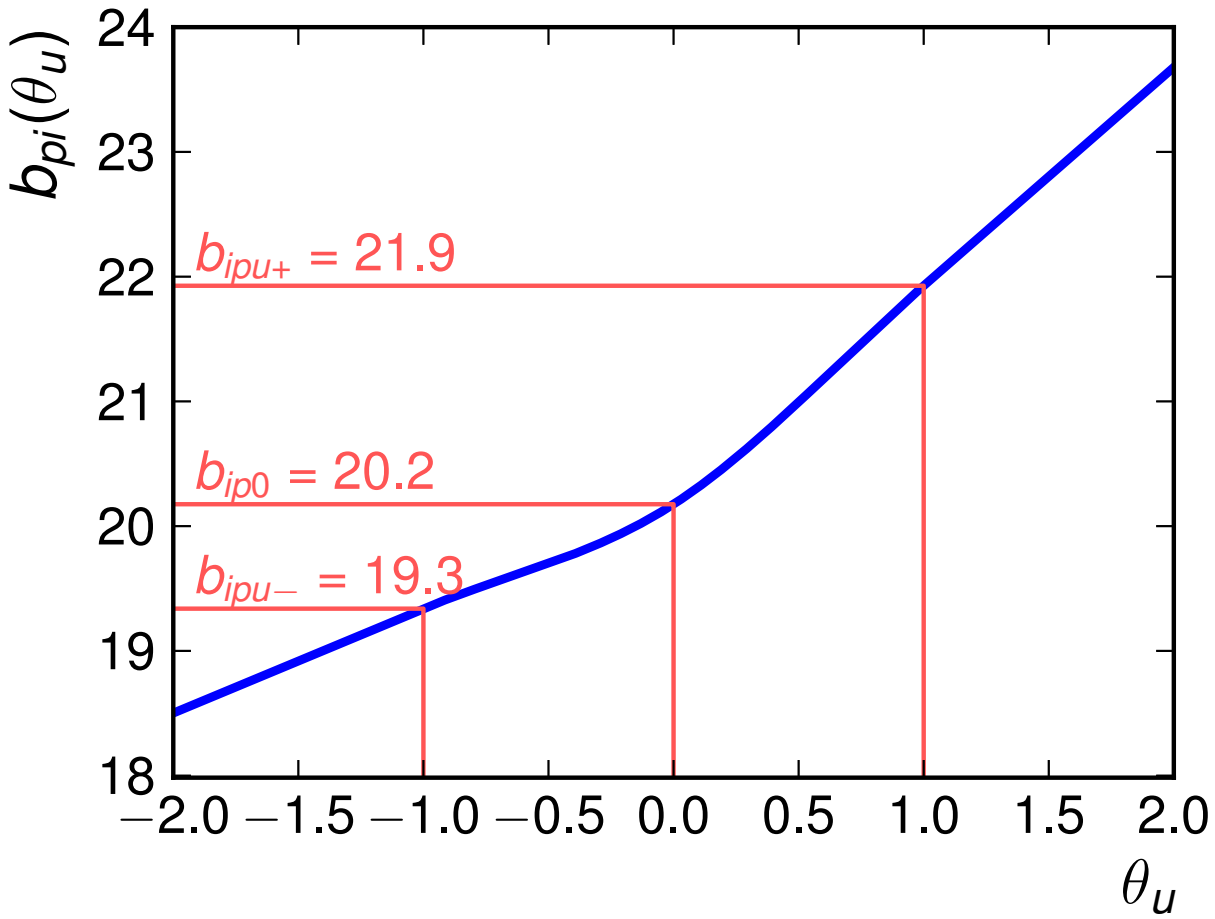
Follow general recipe: introduce nuisance parameter θ_u with standard normal prior, and write the model prediction for the Poisson mean λ_i as a function of the new parameter.

It should interpolate smoothly between the "minus" template for $\theta_u = -1$, the "nominal" template at $\theta_u = 0$ and the "plus" template at $\theta_u = +1$.

There are many possibilities to achieve this; here: Use cubic interpolation for $|\theta_u| < 1$ and linear extrapolation for $|\theta_u| > 1$.

Shape Uncertainties: Single Bin Behavior

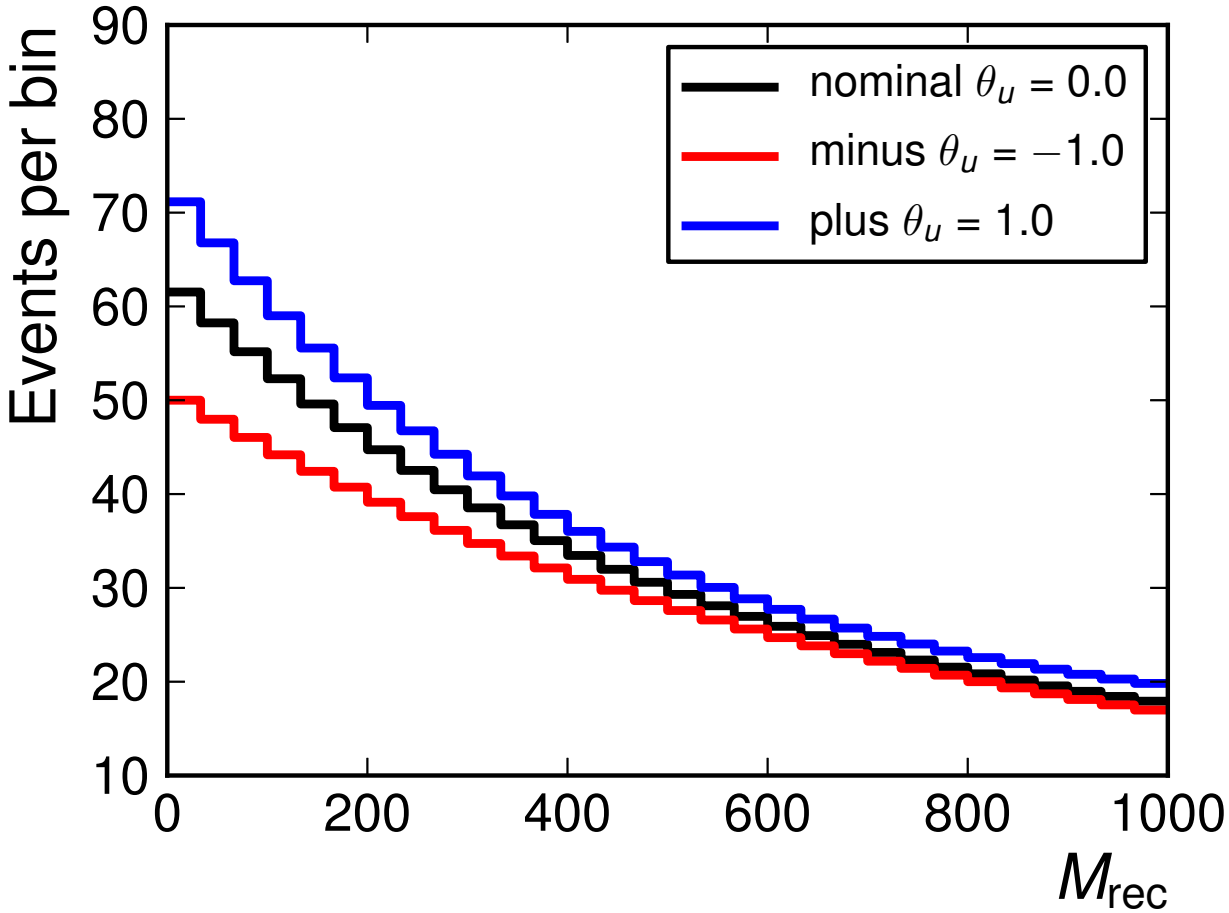
Example interpolation for the bin around $M_{\text{rec}} = 835$:



$b_{ipu\pm}$ are the bin contents for the "plus" and "minus" templates; b_{ip0} for the "nominal".

Shape Uncertainties: Shape Behavior

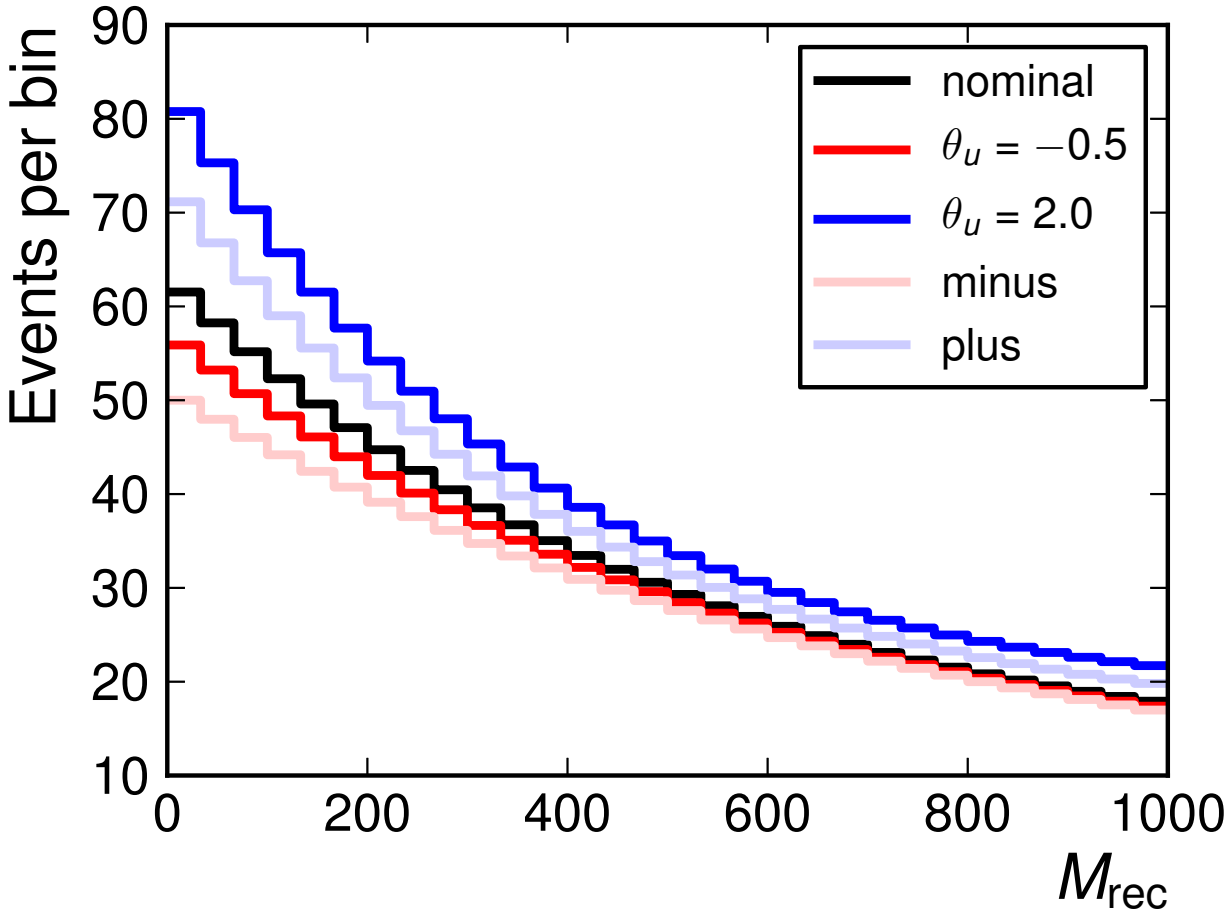
Applying template morphing for certain values θ_u, the background template looks like this:



Interpolation agrees with intuitive expectation.

Shape Uncertainties: Shape Behavior

Applying template morphing for certain values θ_u, the background template looks like this:



Interpolation agrees with intuitive expectation.

Contents

Introduction

If searching for a small signal using Monte-Carlo (or the data in a sideband) to define the statistical model, the limited size of the Monte-Carlo sample might introduce an additional uncertainty on the predicted mean λ_{ci}.

Example: counting experiment in which only $N_{\mathrm{MC}} = 1$ background event survives, weighted to $b = 0.2$, and $n = 3$ events are observed. If $b = 0.2$ was known exactly, this would correspond to a 3σ effect, but it is possible that the true mean corresponds to $N_{\mathrm{MC}} = 2$ ($b = 0.4$), which corresponds to a much smaller significance ...

Barlow-Beeston Method

Idea of Barlow and Beeston (Comp. Phys. Comm. 77 (1993)):

- Introduce the true mean in bin i as nuisance parameter ϵ_i in the statistical model
- Extend the statistical model to model the joint probability to observe n_i data events and $N_{\mathrm{MC},i}$ Monte-Carlo events, given ϵ_i (both are basically Poisson)
- For maximum likelihood methods, note that maximizing w.r.t. ϵ_i leads to de-coupled equations that can be solved numerically

This is implemented in ROOT's TFractionFitter.

Disadvantage: Need to numerically solve equations at each step in the minimization.

Modification: Barlow-Beeston light

Same idea: add one nuisance parameter δ_{ci} per bin to model MC stat. uncertainty:

$$\mu_{ci}(\theta, \delta) = \mu_{ci}(\theta) + \delta_{ci}$$

where δ_{ci} here has a normal prior (with mean 0).

This can be implemented using template morphing: Add N_{bins} nuisance parameters δ_{ci}, each shifting exactly one bin w.r.t. the nominal template. But: can lead to hundreds of nuisance parameters $\rightsquigarrow$ numerical minimization slow, unstable.

Analytical Solution

To maximize the Poisson likelihood function $L(\theta, \delta|n)$, set all derivatives to zero and solve for the parameters $\delta_{ci} \rightsquigarrow$ analytical solution $\delta_m(\theta)$!

Then, the actual maximization algorithm is run on the likelihood function L_m in which the δ parameters have been "maximized away":

$$L_m(\theta|d) := L(\theta, \delta_m(\theta)|d)$$

$\rightsquigarrow$ Allows to handle MC stat. uncertainties without introducing an additional "visible" nuisance parameter.

Comments

BB light uses normal approximation $\rightsquigarrow$ need enough MC events in each bin.

Possible work-arounds in case of low MC statistics include:

- re-bin to reach a minimum number of MC events
- use a non-parametric smoothing procedure (neighbor bins)
- fit a function ("parametric smoothing")

Often, first option is the easiest one, but: could loose sensitivity w.r.t. others.