Hypothesis Tests and Confidence Intervals

Jochen Ott

June 5, 2013

About this talk

- My background: I'm a physicist, not a statistician; my knowledge of statistics is mainly self-taught from problems in HEP (mainly LHC).
- This talk will cover a lot of material – but there are only a few central concepts, which I will repeat often.
- Please ask questions any time.

Part I

Hypothesis Tests

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 $s = 0$:

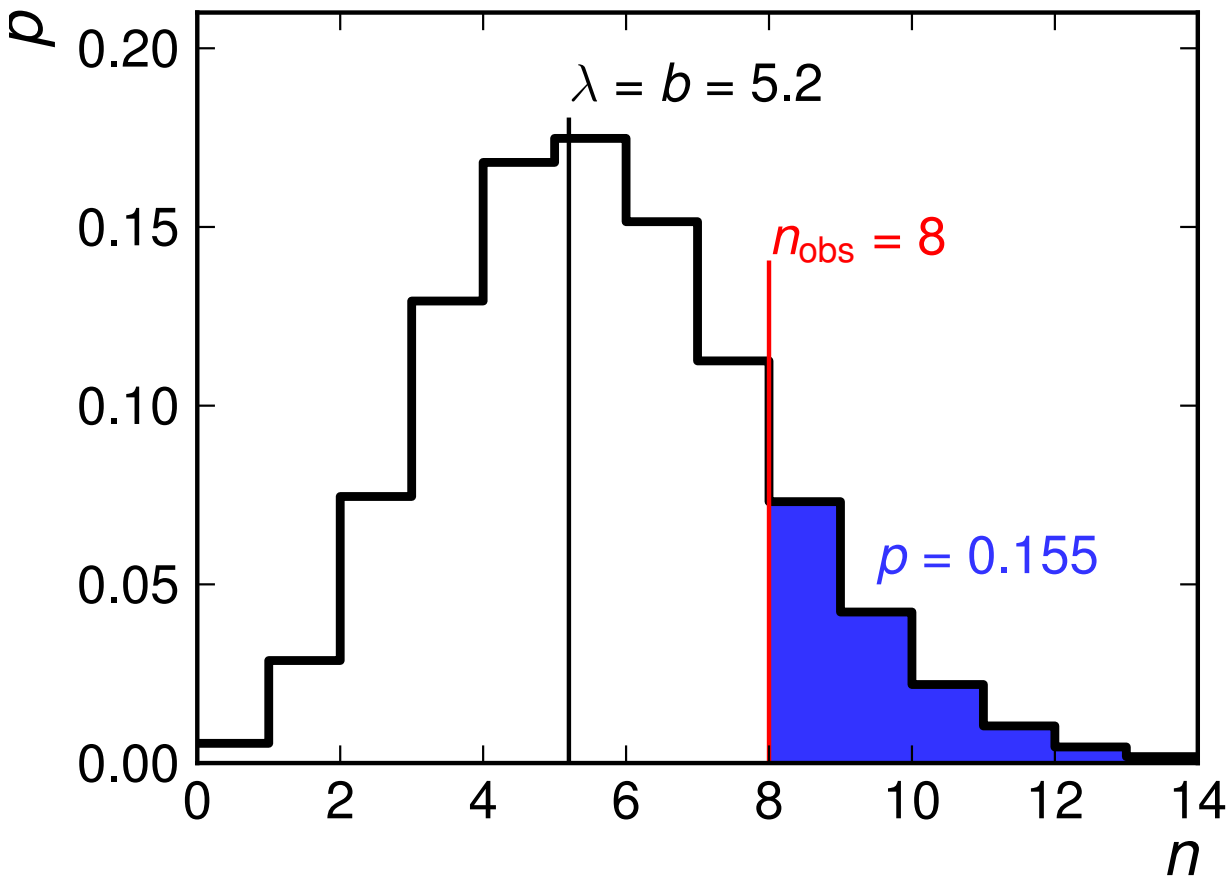
$$p(n_{\text{obs}}) = \sum_{n=n_{\text{obs}}}^{\infty} \text{Poisson}(n|\lambda = b = 5.2)$$

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 (cf. exercise 1).



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.

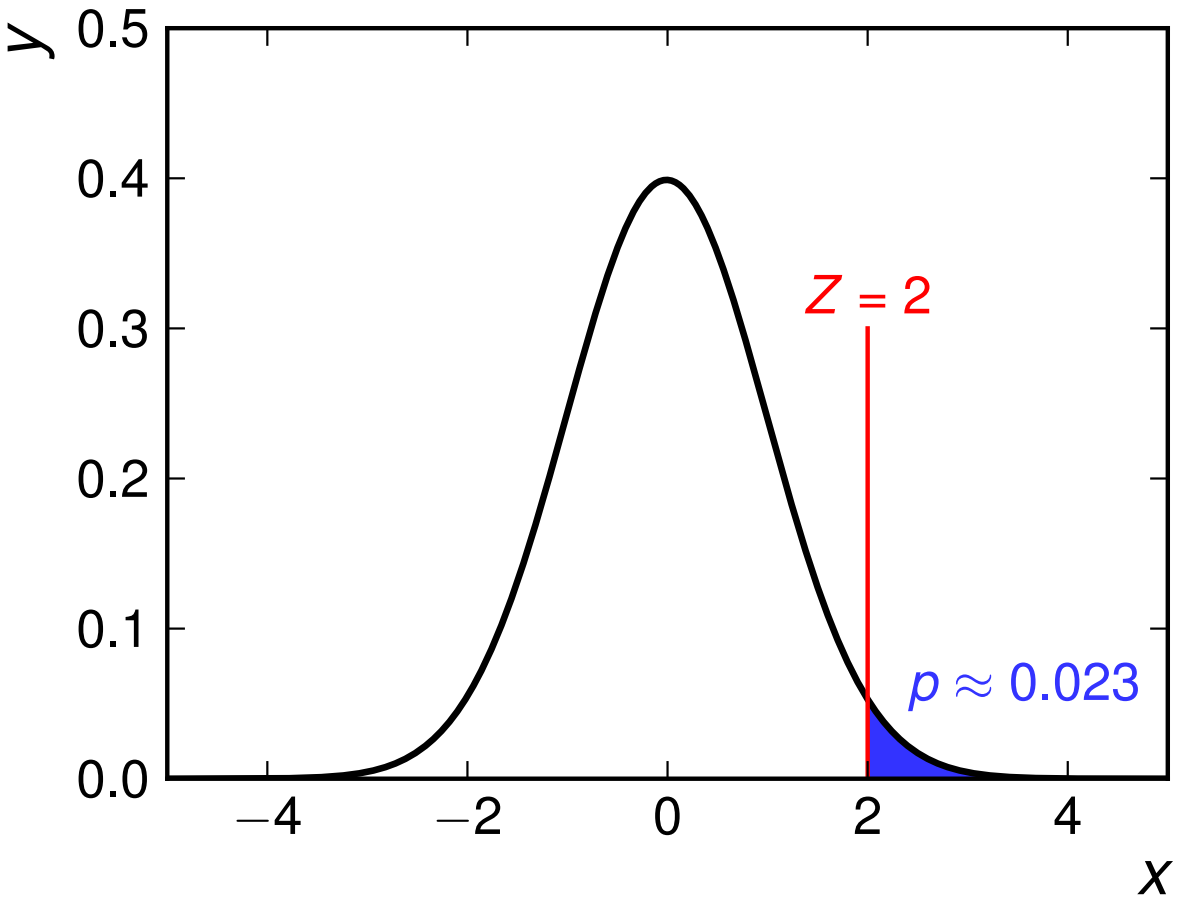
The type-II error is denoted β.
The probability to (correctly) reject H_0 if the alternative H_1 is true is the power $(1 - \beta)$.

Remarks

- For a given fixed type-I error rate α, one would prefer the test with high power $(1 - \beta)$; this will be used as a criterion later.
- 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, as explained later).
- 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.

Z-value

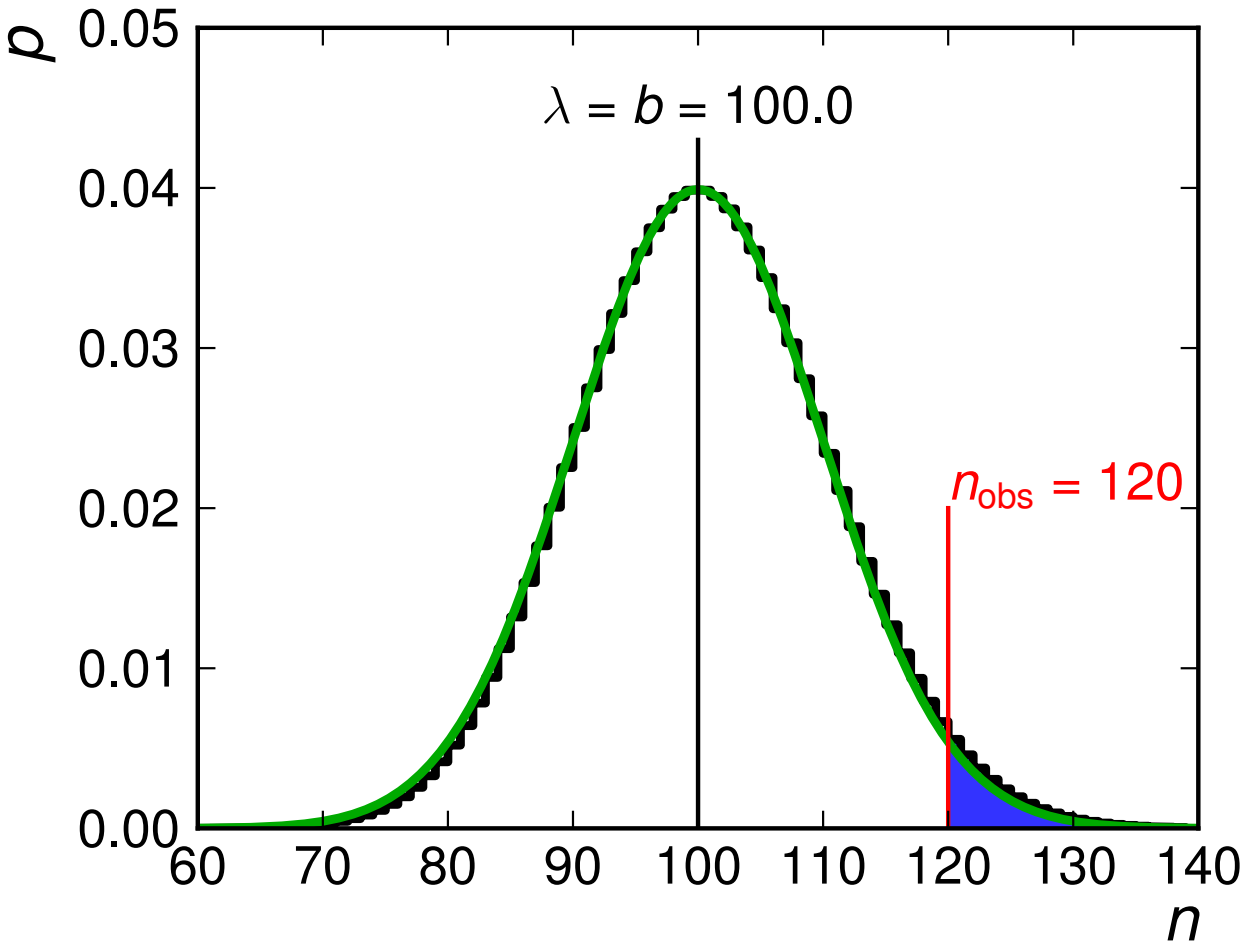
To avoid handling small p-values, p is often expressed as Z-value ("number of sigma"), defined as the lower integration bound for a standard normal distribution such that the integral reproduces the p-value:



Z	p
-1	0.84
0	0.5
1	0.16
2	0.023
3	1.3×10^{-3}
4	3.2×10^{-5}
5	2.9×10^{-7}

Normal Approximation I

For large b, the Poisson distribution is approximately normal with mean b and standard deviation $\sqrt{b}$.



$\rightsquigarrow$ The Z-value is approximated by

$$Z_\mathrm{a} = \frac{s}{\sqrt{b}} = \frac{n-b}{\sqrt{b}}.$$

Here: $Z_\mathrm{a} = 2.00$ (true Poisson $Z = 1.91$).

Normal Approximation II

Numerical example for the normal approximation

$$Z_{\mathrm{a}} = \frac{s}{\sqrt{b}} = \frac{n-b}{\sqrt{b}}$$

For $b = 5.2$:

n	p	Z	Z_{a}
6	0.419	0.20	0.35
8	0.155	1.01	1.23
10	0.040	1.75	2.10
12	0.0073	2.44	2.98

For $b = 100$:

n	Z	Z_{a}
110	0.95	1.00
120	1.91	2.00
130	2.83	3.00
140	3.74	4.00

$\rightsquigarrow$ Z_{a} overestimates actual Z, but reasonable approximation for many applications; agreement becomes better for larger n.
(see also exercise 1. for more examples)

Expected Significance

The *expected significance* is defined as the "typical" (median) result for an ensemble of data distributed according to a specific *expected* hypothesis H_e.

It is a useful concept to

- measure the (future) sensitivity (e.g. for more luminosity, ...),
- compare the performance of two different analyses/experiments.

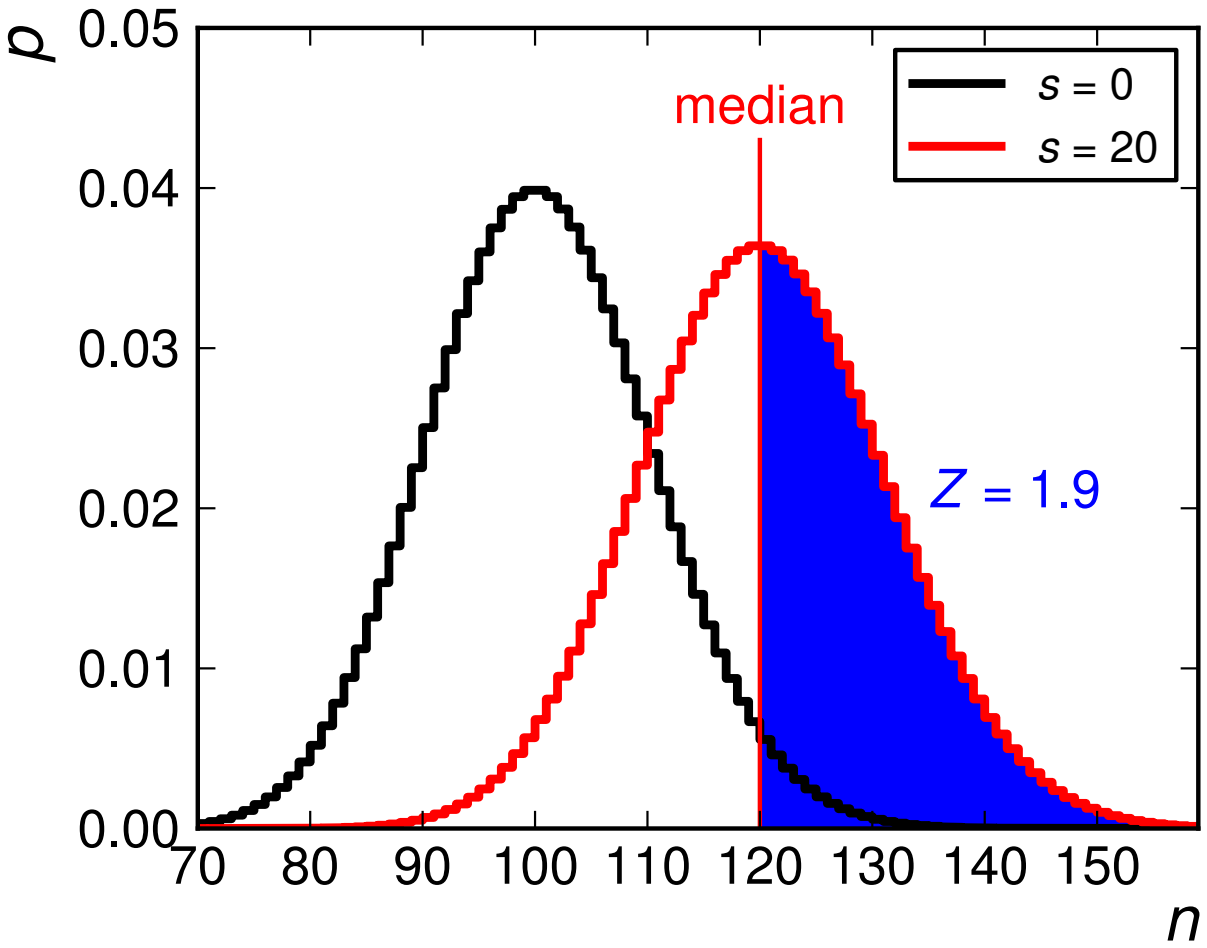
Note that the *observed* significance is a random variable and not a good measure for performance comparison (see Backup).

Expected Significance: Notes

- Can use Monte-Carlo method: generate toy data according to H_e, calculate Z-value for each toy $\rightsquigarrow$ distribution of Z-values. The median of this Z-value distribution is the expected Z-value.
- The *expected hypothesis* H_e has to be specified (fixing model parameter values) to make "expected" statement well-defined.
- Can also look not only at median but also at "typical spread" (central $1\sigma = 68\%$ / $2\sigma = 95\%$) $\rightsquigarrow$ "bands" of expected result.
- "Expected result" can be generalized to other statistical methods (e.g. limits $\rightsquigarrow$ "Brazil band" plots).

Expected Significance: Example

For $b = 100$, and an *expected* signal $s = 20$, the expected number of n is given by a Poisson distribution with mean $\lambda = 120$.

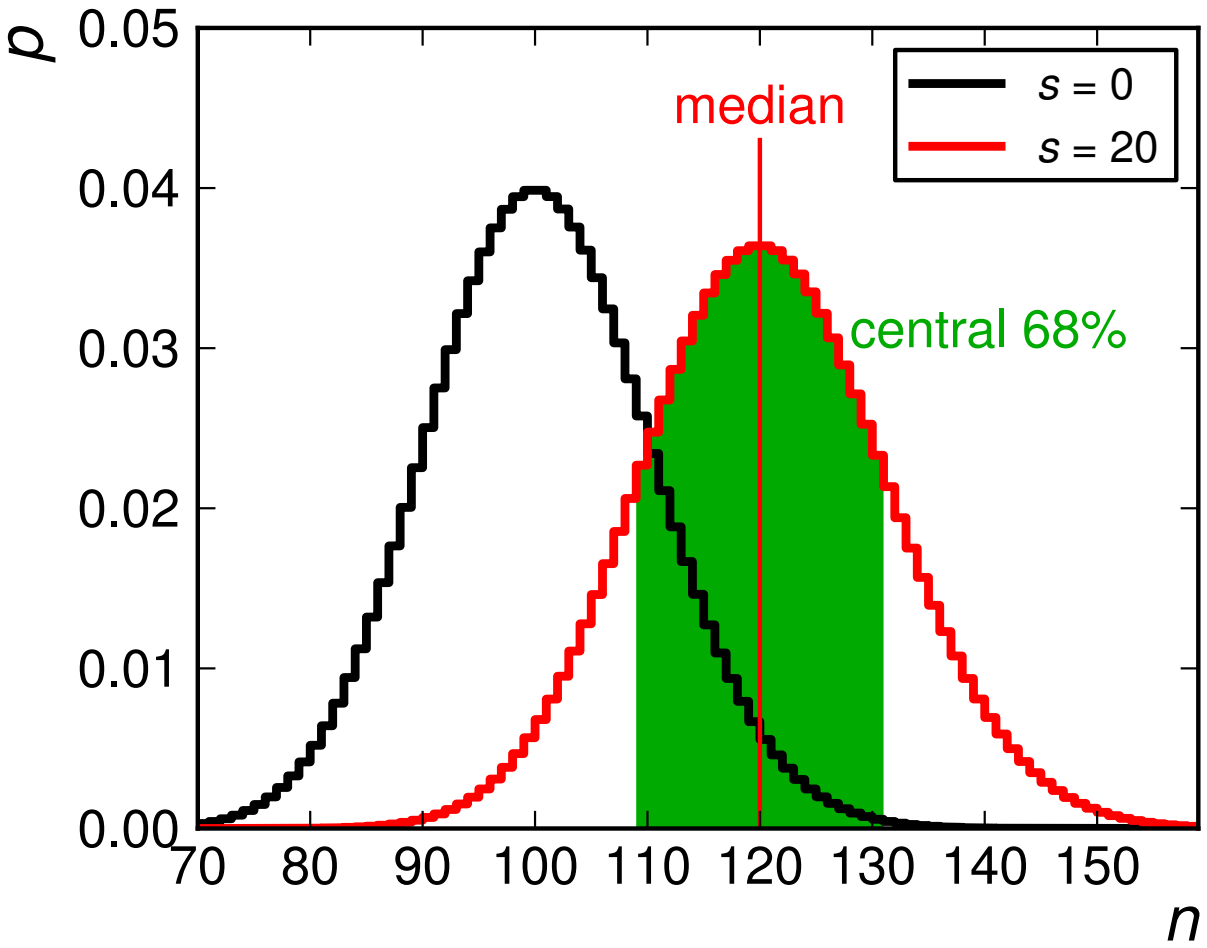


Median expected $n = 120 \rightsquigarrow$ expected significance $Z = 1.91$.

Central 68% of $s = 20$ distribution (109–131) $\rightsquigarrow 1\sigma$ expected significance range: 0.85–2.93.

Expected Significance: Example

For $b = 100$, and an *expected* signal $s = 20$, the expected number of n is given by a Poisson distribution with mean $\lambda = 120$.



Median expected $n = 120 \rightsquigarrow$ expected significance $Z = 1.91$. Central 68% of $s = 20$ distribution (109–131) $\rightsquigarrow 1\sigma$ expected significance range: 0.85–2.93.

Expected Significance in the Normal Approximation

For $b = 100$, $s = 20$, median n is 120. $\rightsquigarrow$ expected $Z_a = 2.0$.

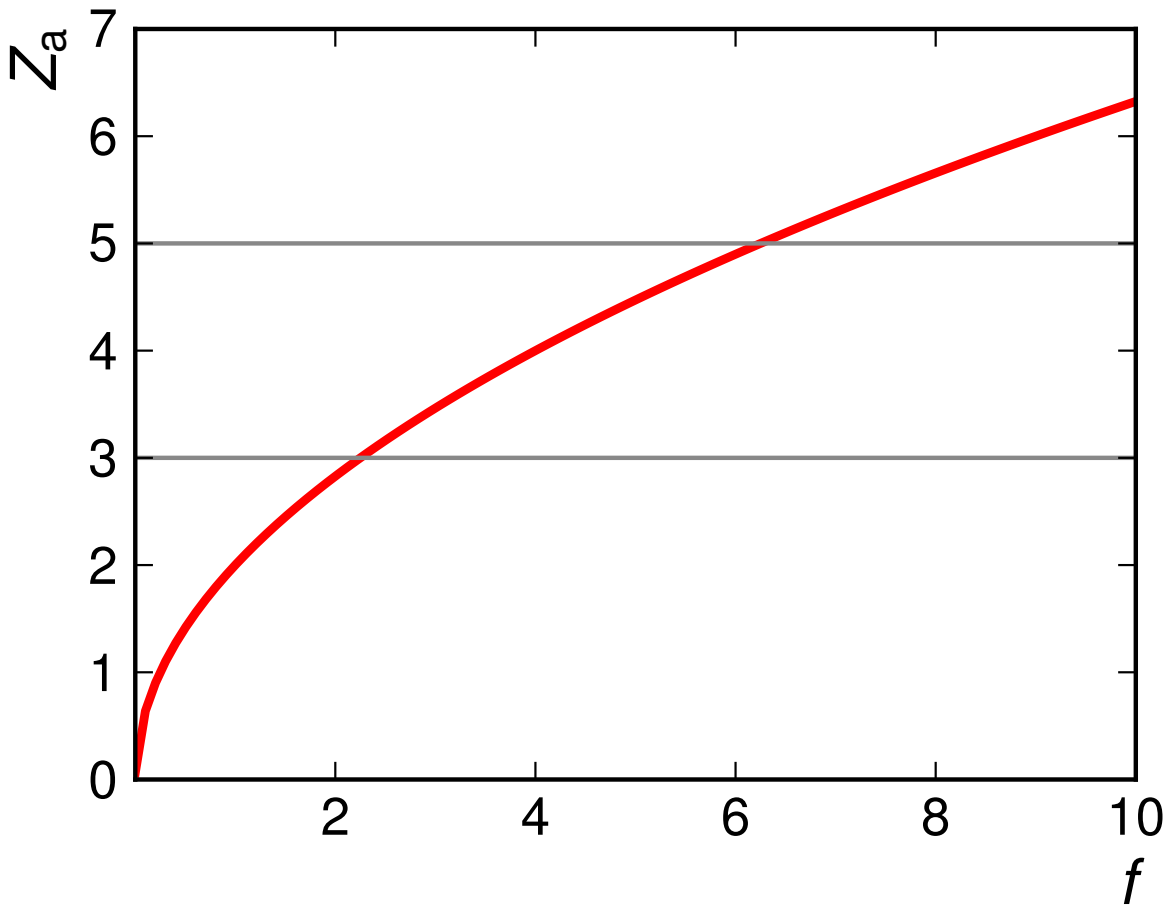
Approximation useful for luminosity projections: Increasing luminosity by a factor f leads to a "$\sqrt{f}$" behavior:

$\rightsquigarrow$ to increase Z-value by a factor of 2, need 4 times the data.

But: with systematic uncertainties, picture will change dramatically [see later]

Expected Significance in the Normal Approximation

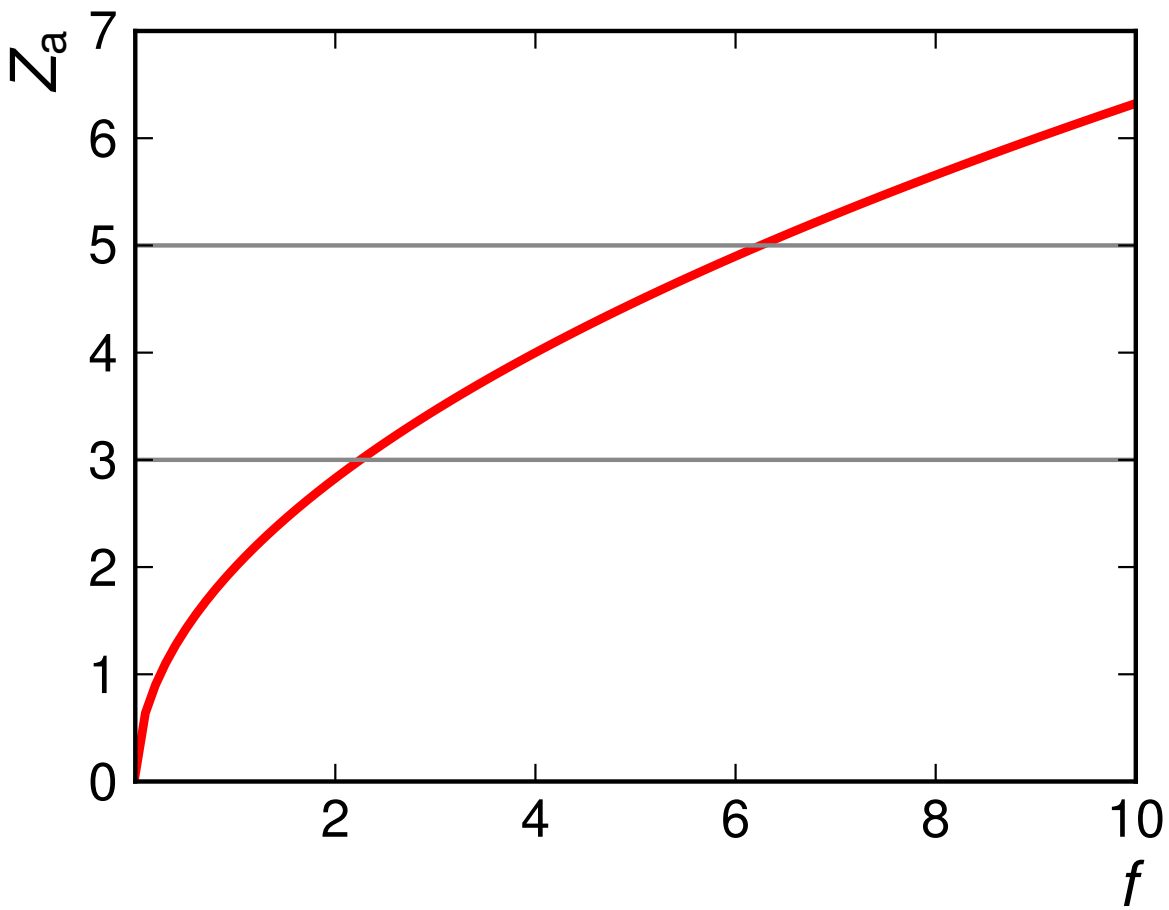
For $b = 100$, $s = 20$, median n is 120. $\rightsquigarrow$ expected $Z_a = 2.0$.
Approximation useful for luminosity projections: Increasing luminosity by a factor f leads to a "$\sqrt{f}$" behavior:



$\rightsquigarrow$ to increase Z-value by a factor of 2, need 4 times the data.
But: with systematic uncertainties, picture will change dramatically [see later]

Expected Significance in the Normal Approximation

For $b = 100$, $s = 20$, median n is 120. $\rightsquigarrow$ expected $Z_a = 2.0$.
Approximation useful for luminosity projections: Increasing luminosity by a factor f leads to a "$\sqrt{f}$" behavior:



$\rightsquigarrow$ to increase Z-value by a factor of 2, need 4 times the data.
But: with systematic uncertainties, picture will change dramatically [see later]

2 Generalization; Shape Model

- Introduction
- Test Statistic, MC method

Introduction

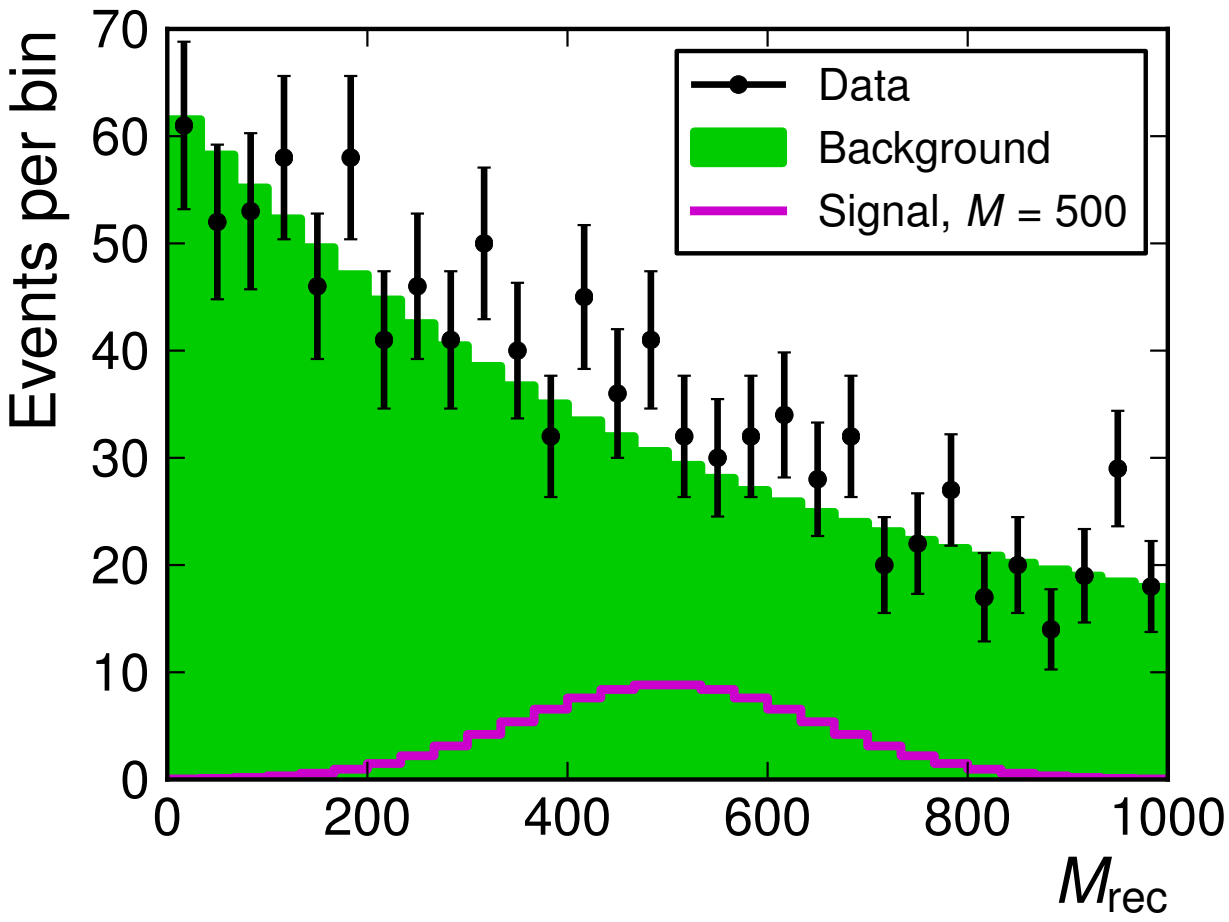
So far: simplistic Poisson model with only one event count ("counting experiment").

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:



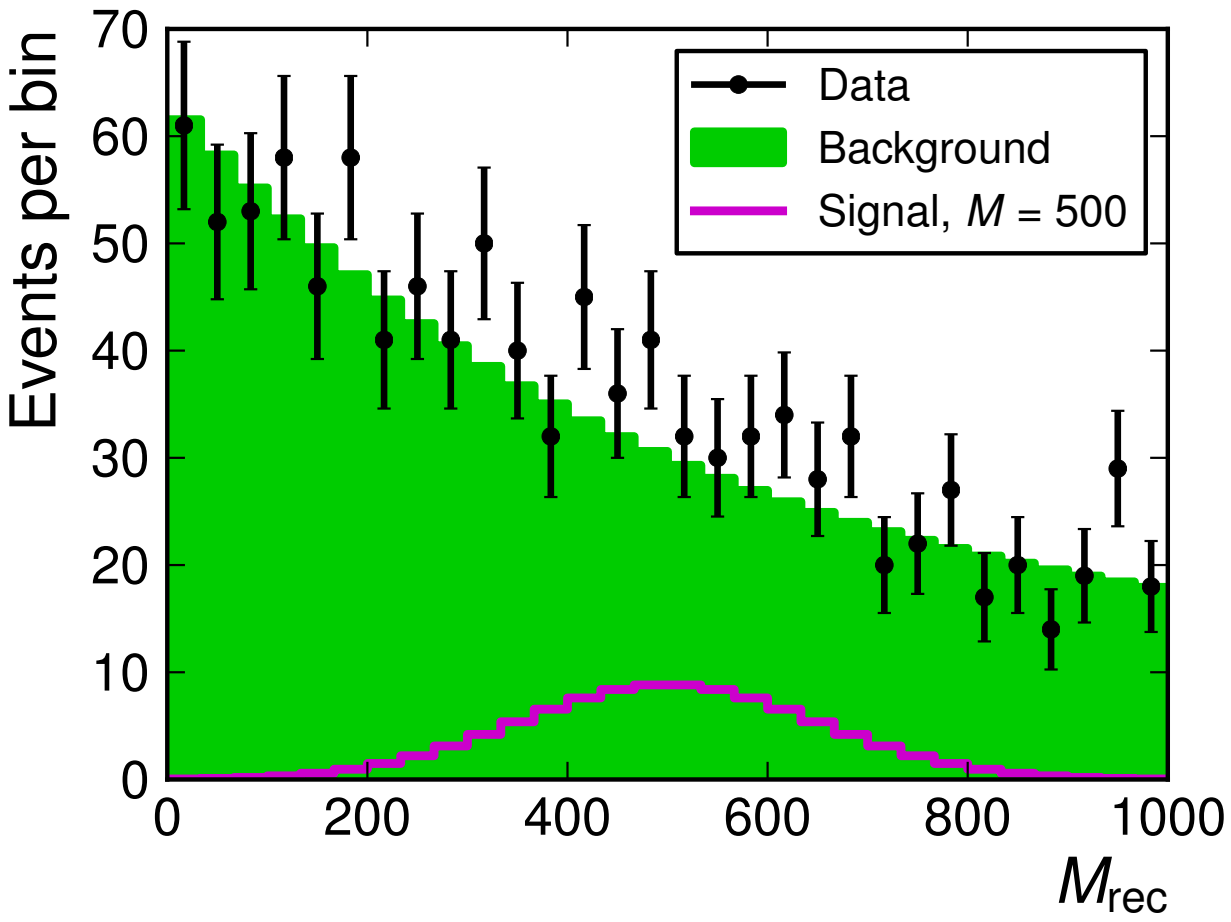
Can this data be seen as evidence against the background-only model?

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:



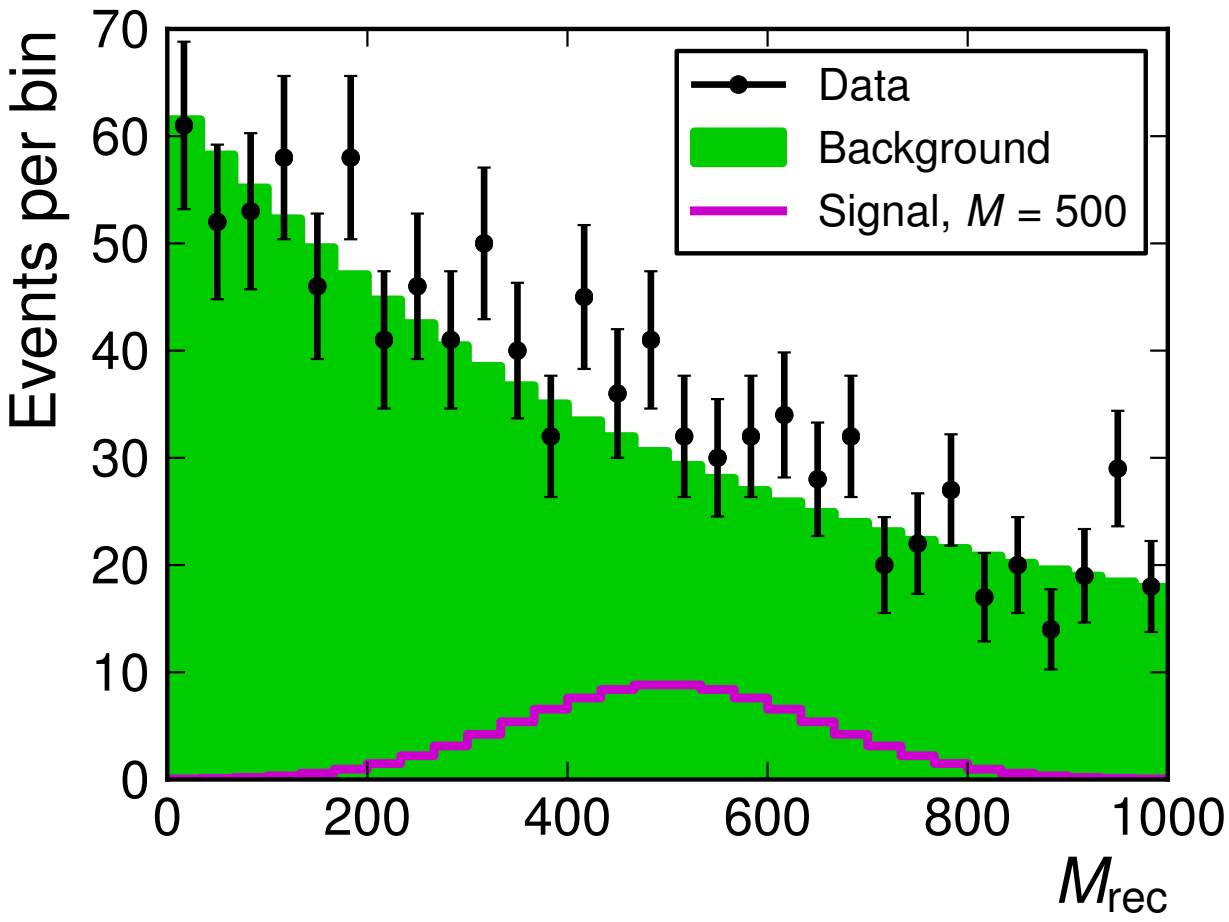
Can this data be seen as evidence against the background-only model?

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:



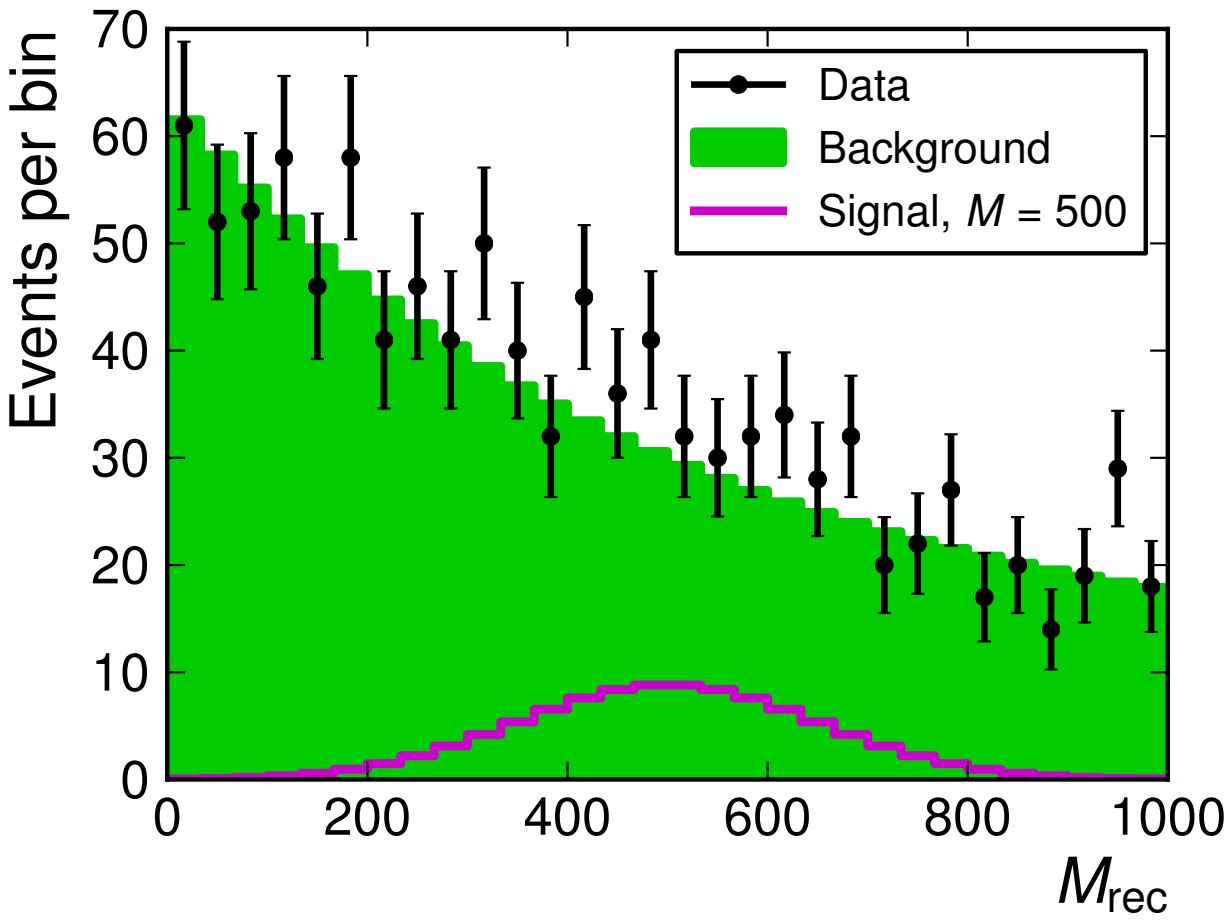
Can this data be seen as evidence against the background-only model?

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.

2 Generalization; Shape Model

- Introduction
- Test Statistic, MC method

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 a single 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 a single 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 a single 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:

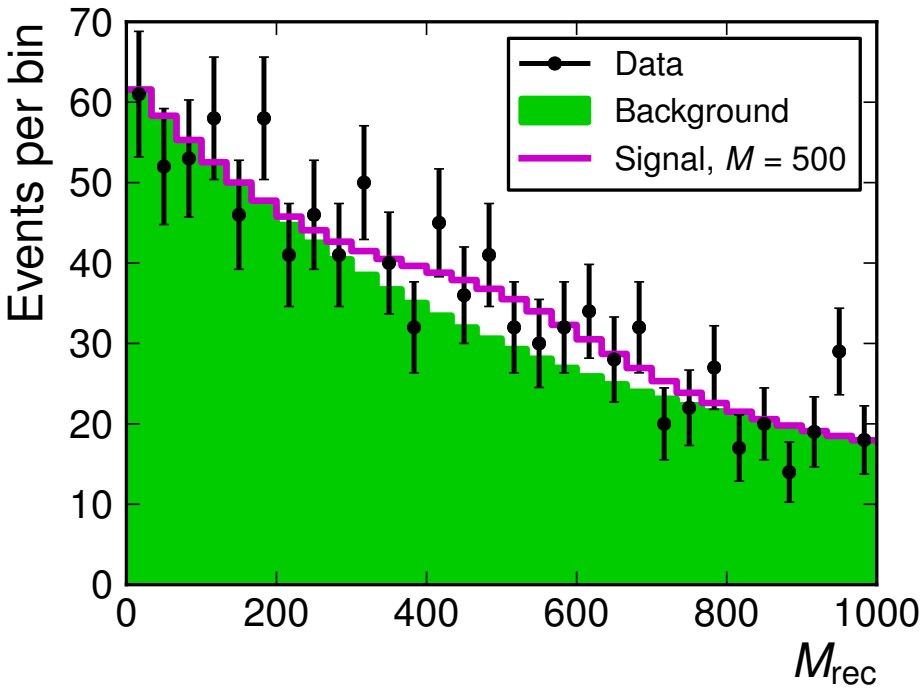
$$p = \Pr(t \geq t_{\text{obs}} | H_0).$$

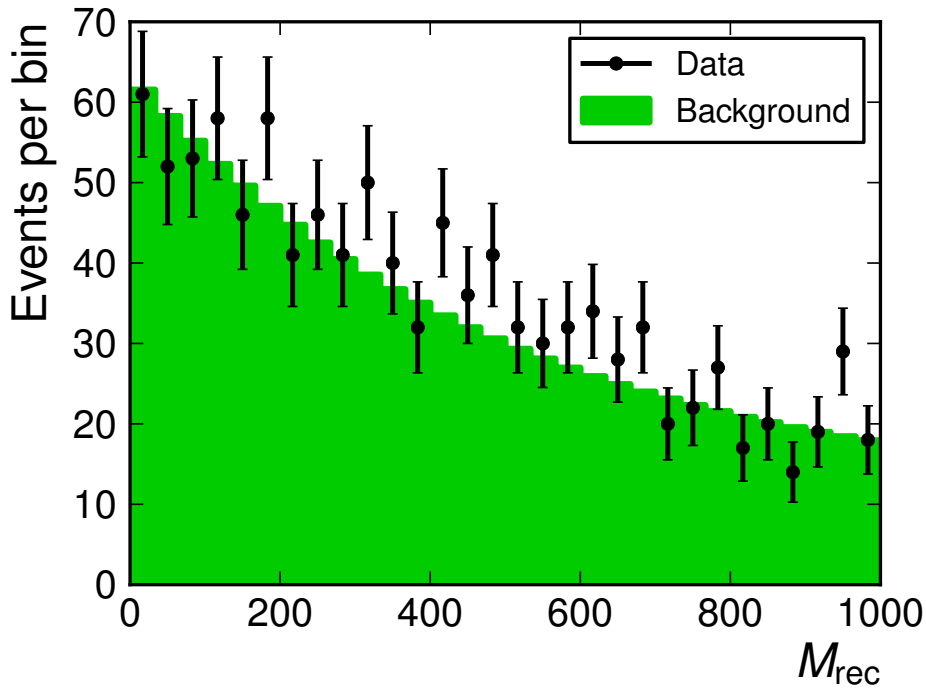
This suggests using a Monte-Carlo method for calculating the p-value:

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 I

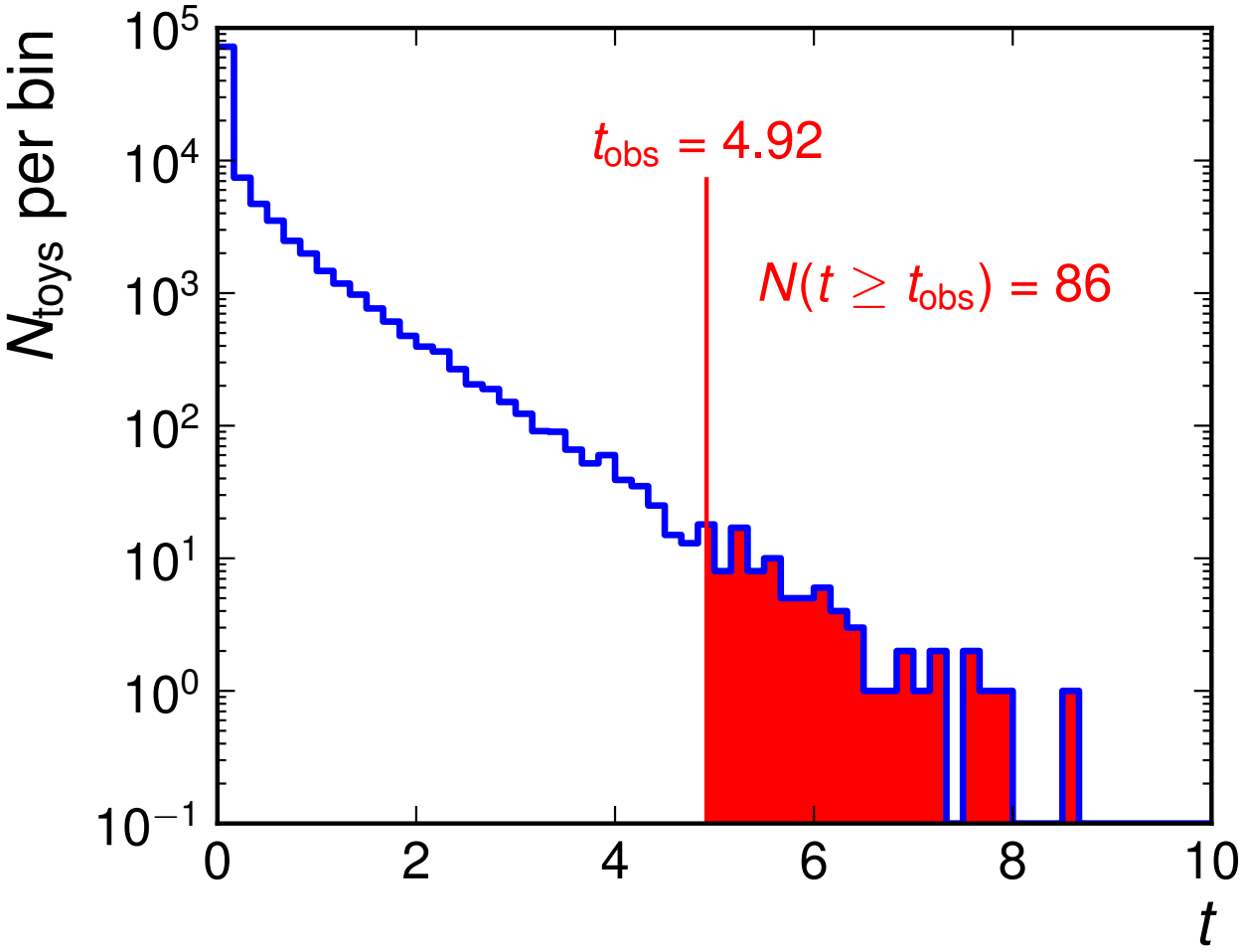




- Calculate t for data by making two likelihood fits (left: $\hat{\mu} = 0.72$; right: $\mu = 0 \rightsquigarrow t_{\text{obs}} = 4.92$).
- Generate 100,000 toy datasets for the null hypothesis $\mu = 0$ by generating Poisson random number in each bin according to the b histogram
- For each toy dataset, calculate t

MC method for p: Example II

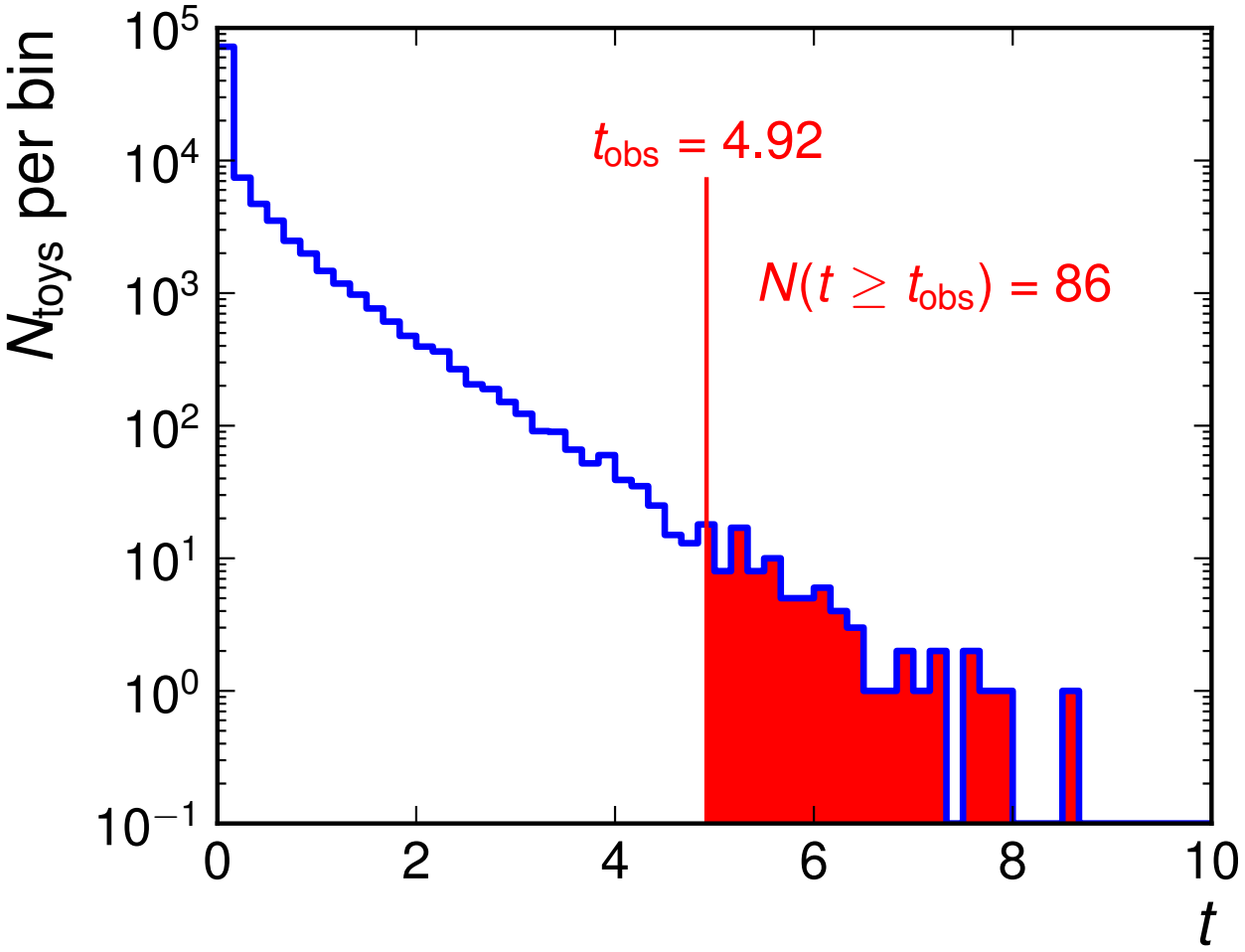
Result of 100,000 toys: $\hat{p} = 86/10^5 \rightsquigarrow \hat{Z} = 3.13$



Limited number of toys should be considered as uncertainty on $\hat{p}$.
Normal approximation for Binomial error:
$\Delta p = \sqrt{\hat{p}(1-\hat{p})/N} \rightsquigarrow$
$\hat{p} = (8.6 \pm 0.9) \times 10^{-4}$
$\rightsquigarrow \hat{Z} = 3.13 \pm 0.03.$

MC method for p: Example II

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



Limited number of toys should be considered as uncertainty on $\hat{p}$.
Normal approximation for Binomial error:
$\Delta p = \sqrt{\hat{p}(1-\hat{p})/N} \rightsquigarrow$
$\hat{p} = (8.6 \pm 0.9) \times 10^{-4}$
$\rightsquigarrow \hat{Z} = 3.13 \pm 0.03.$

Test Statistic Choice: Power

There are many reasonable choices for the test statistic definition (e.g. $t = \hat{\mu}$, the fitted signal cross section).

A criterion for the test statistic choice is the power $1 - \beta$, the probability to reject the null hypothesis H_0, if H_1 is true. This is only well defined is H_1 is a simple hypothesis, i.e. it has no free parameters.

For simple H_1 the most powerful test statistic is the likelihood ratio,

$$r = \frac{L(\theta_1|d)}{L(\theta_0|d)}$$

where θ_0 and θ_1 are the parameter values for H_0 and H_1, resp. (Neyman-Pearson Lemma).

The profile likelihood ratio is a generalization of this test statistic for non-simple hypotheses.

Test Statistic Choice: Power

There are many reasonable choices for the test statistic definition (e.g. $t = \hat{\mu}$, the fitted signal cross section).

A criterion for the test statistic choice is the power $1 - \beta$, the probability to reject the null hypothesis H_0, if H_1 is true. This is only well defined is H_1 is a simple hypothesis, i.e. it has no free parameters.

For simple H_1 the most powerful test statistic is the likelihood ratio,

$$r = \frac{L(\theta_1|d)}{L(\theta_0|d)}$$

where θ_0 and θ_1 are the parameter values for H_0 and H_1, resp. (Neyman-Pearson Lemma).

The profile likelihood ratio is a generalization of this test statistic for non-simple hypotheses.

Test Statistic Choice: Equivalence

Two apparently different test statistic definitions will lead to the same p-value if they only differ by monotonic transformation, as statements about quantiles are invariant under monotonic transformation.

- One could also define t with switched sign and take the convention that smaller values of the test statistic mean greater incompatibility with H_0; this would not change the result.
- In a counting experiment with a search for a positive signal of unknown magnitude, any reasonable test statistic is a monotonic function in n – including the profile likelihood ratio t –; and will lead to the exact same result.

Test Statistic Choice: Equivalence

Two apparently different test statistic definitions will lead to the same p-value if they only differ by monotonic transformation, as statements about quantiles are invariant under monotonic transformation.

- One could also define t with switched sign and take the convention that smaller values of the test statistic mean greater incompatibility with H_0; this would not change the result.
- In a counting experiment with a search for a positive signal of unknown magnitude, any reasonable test statistic is a monotonic function in n – including the profile likelihood ratio t –; and will lead to the exact same result.

Test Statistic Choice: Equivalence

Two apparently different test statistic definitions will lead to the same p-value if they only differ by monotonic transformation, as statements about quantiles are invariant under monotonic transformation.

- One could also define t with switched sign and take the convention that smaller values of the test statistic mean greater incompatibility with H_0; this would not change the result.
- In a counting experiment with a search for a positive signal of unknown magnitude, any reasonable test statistic is a monotonic function in n – including the profile likelihood ratio t –; and will lead to the exact same result.

Section Summary

The test statistic summarizes the data in a single number to quantify the incompatibility with the null hypothesis.

We also saw:

- How the test statistic is used in toy Monte-Carlo to define the p-value.
- The power as a criterion to choose a test statistic.
- That monotonic test statistic transformations do not change the result.

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.).

3 Handling of Systematic Uncertainties

- Bayesian/Hybrid Method; Counting Experiment
- Test Statistic Definition
- Shape Model Uncertainties

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 $t \geq t_{\text{obs}}$.

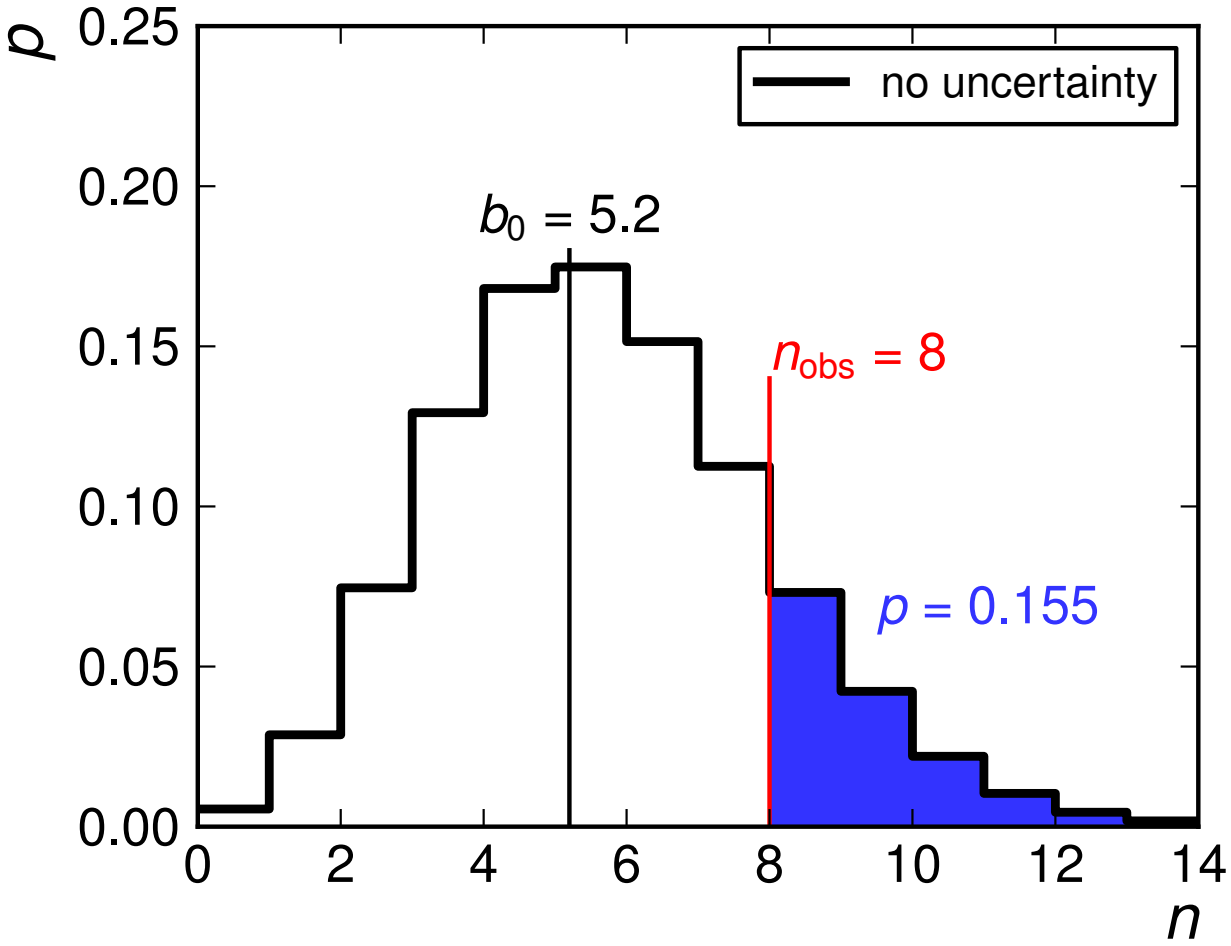
Formally, the resulting p-value is:

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

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

Example I: Counting Experiment

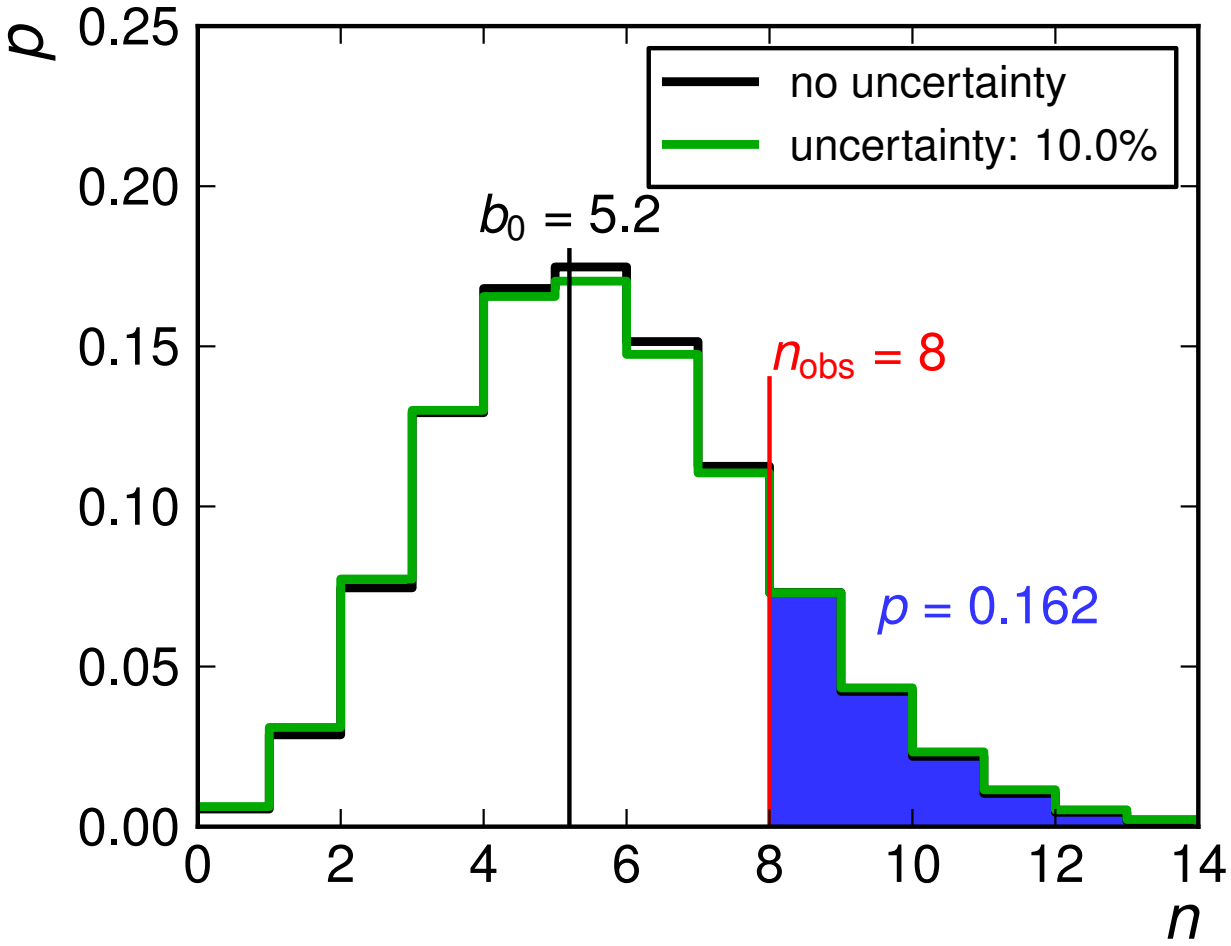
For a normal prior on b with mean $b_0 = 5.2$ distribution for n changes:



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

Example I: Counting Experiment

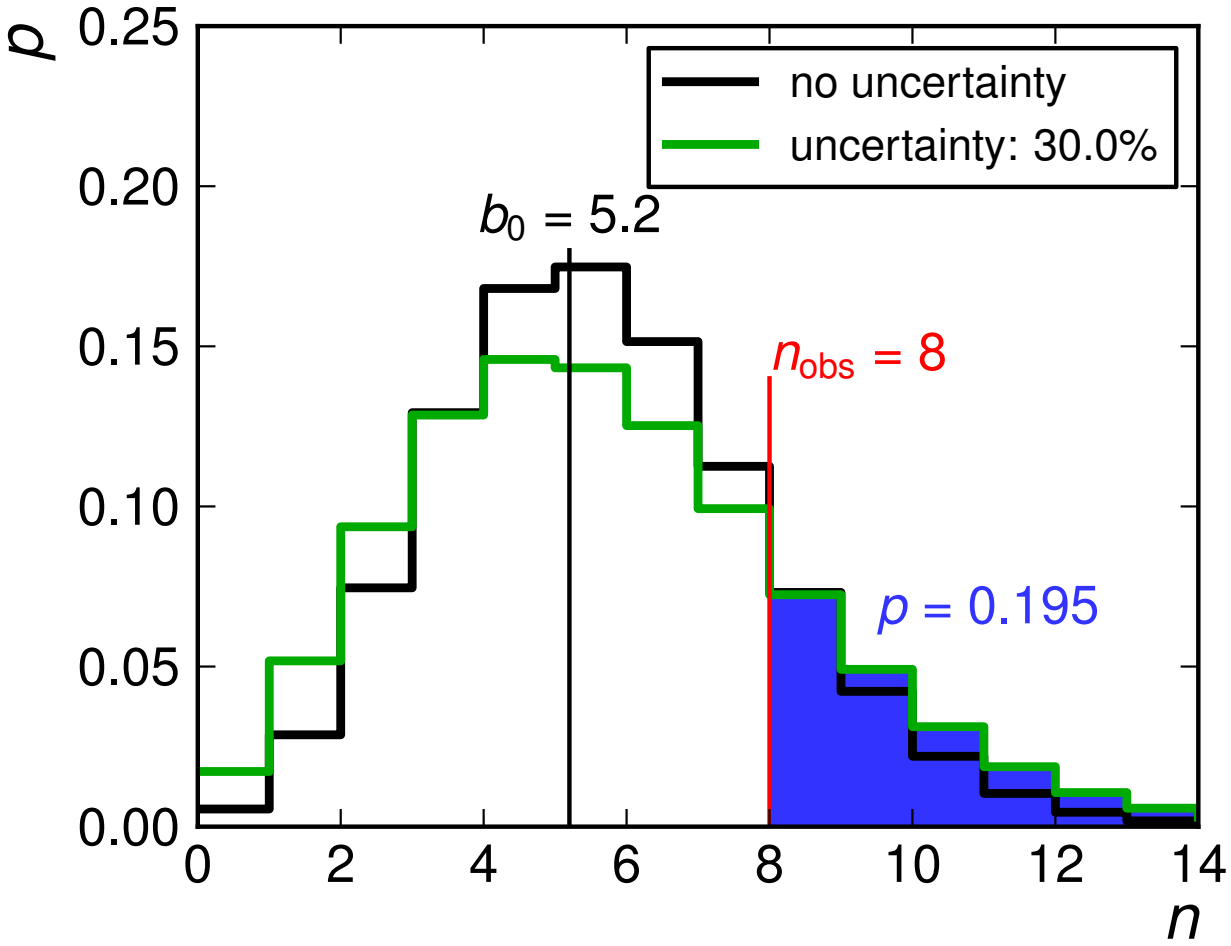
For a normal prior on b with mean $b_0 = 5.2$ distribution for n changes:



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

Example I: Counting Experiment

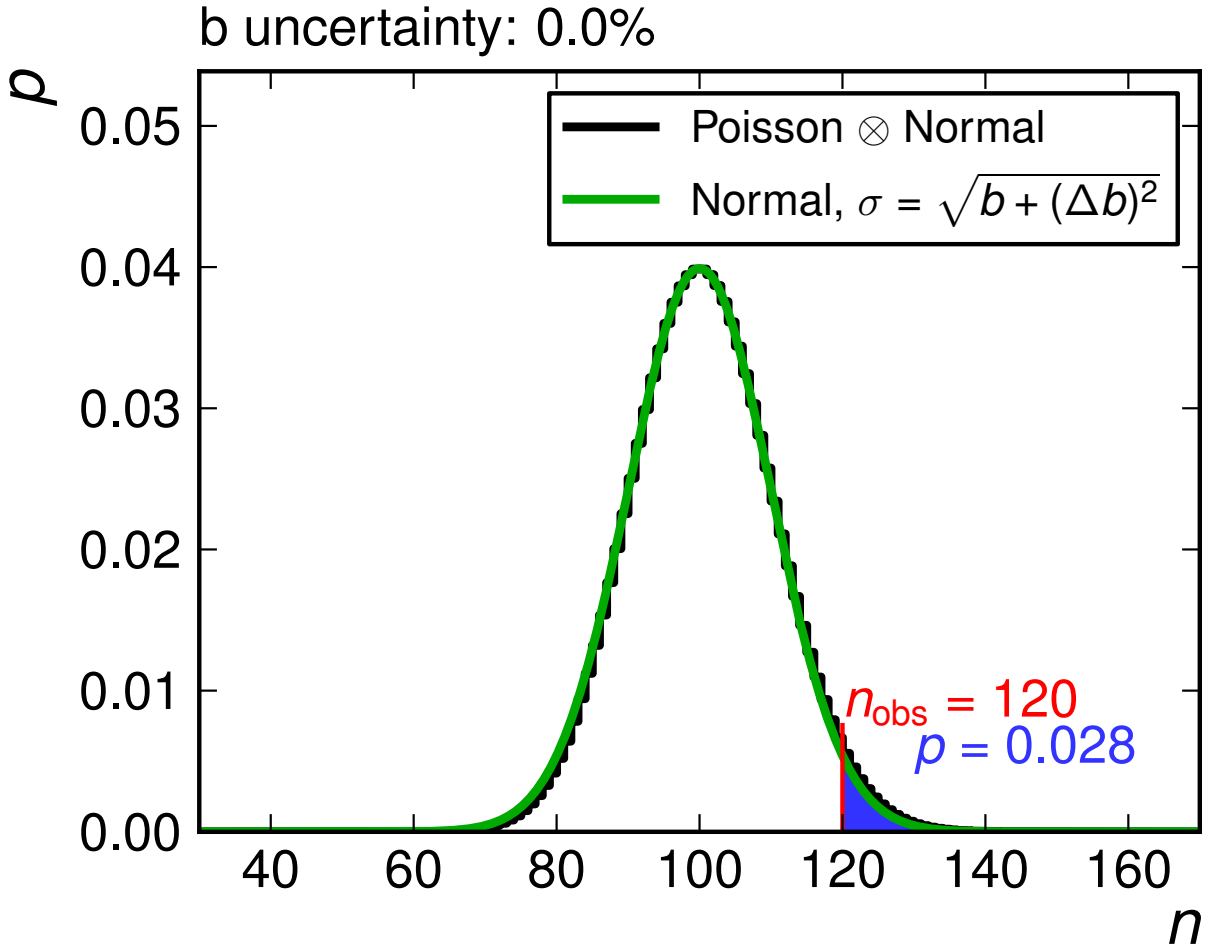
For a normal prior on b with mean $b_0 = 5.2$ distribution for n changes:



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

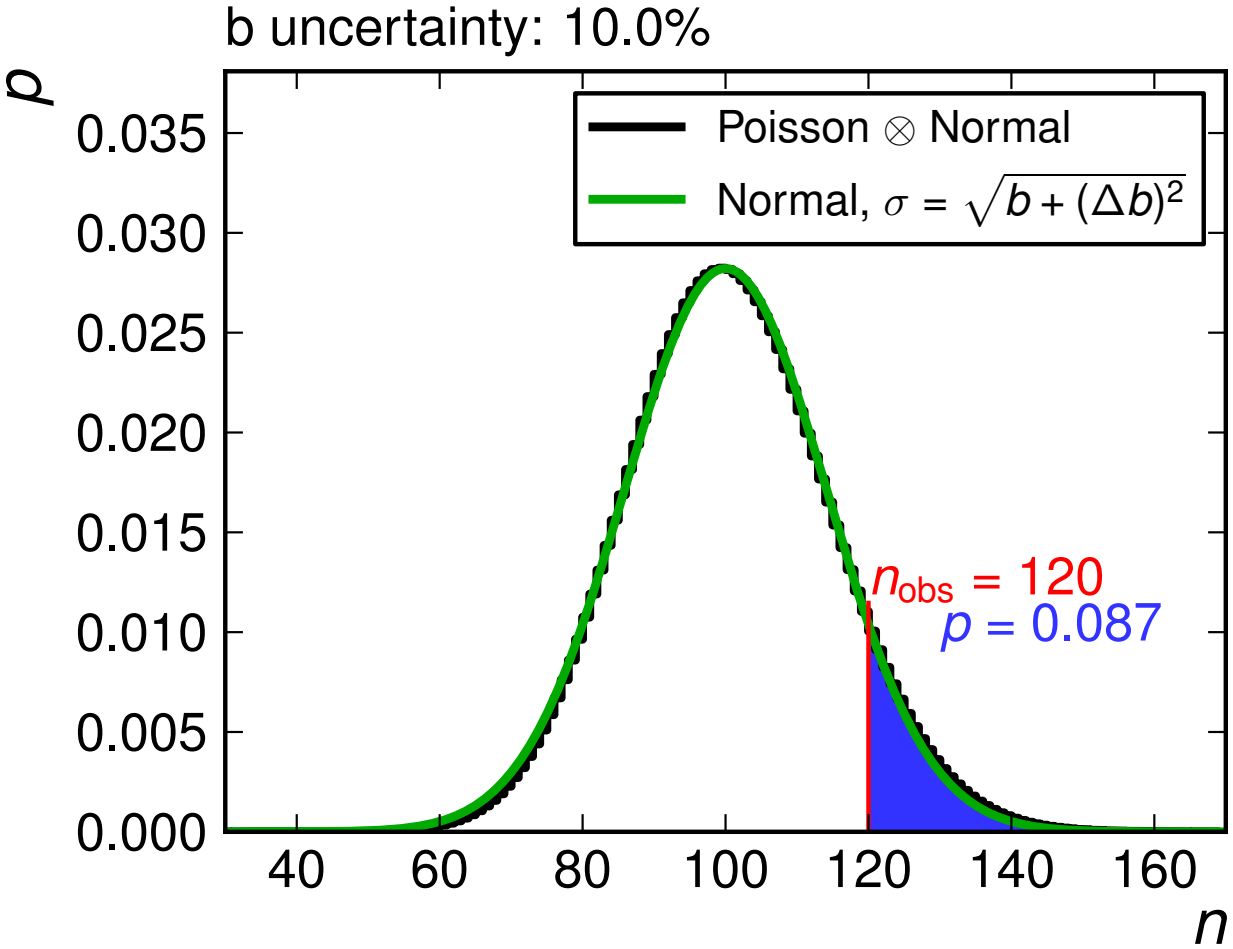
Example II: Normal Approximation

For large b, Poisson is approximately normal. Distribution for n approximately convolution of two normals with $\sigma_1 = \sqrt{b}$ and $\sigma_2 = \Delta b$ $\rightsquigarrow$ normal with std. dev. $\sigma = \sqrt{b + (\Delta b)^2}$.



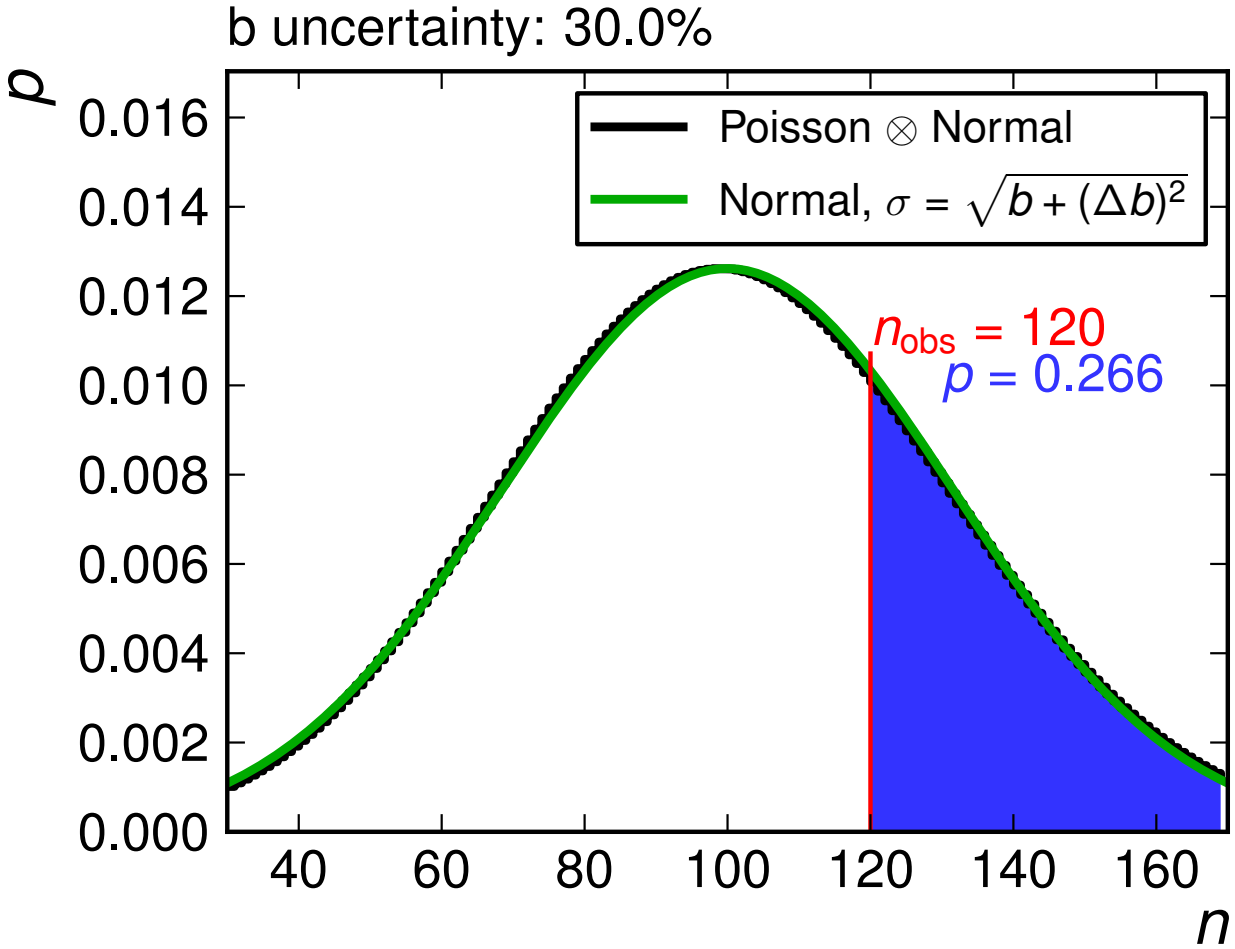
Example II: Normal Approximation

For large b, Poisson is approximately normal. Distribution for n approximately convolution of two normals with $\sigma_1 = \sqrt{b}$ and $\sigma_2 = \Delta b$
$\rightsquigarrow$ normal with std. dev. $\sigma = \sqrt{b + (\Delta b)^2}$.



Example II: Normal Approximation

For large b, Poisson is approximately normal. Distribution for n approximately convolution of two normals with $\sigma_1 = \sqrt{b}$ and $\sigma_2 = \Delta b$
$\rightsquigarrow$ normal with std. dev. $\sigma = \sqrt{b + (\Delta b)^2}$.



Normal Approximation: Formula

For large b, n follows a normal distribution with standard deviation

$$\sigma = \sqrt{b + (\Delta b)^2}$$

and the approximate Z-value is thus given by

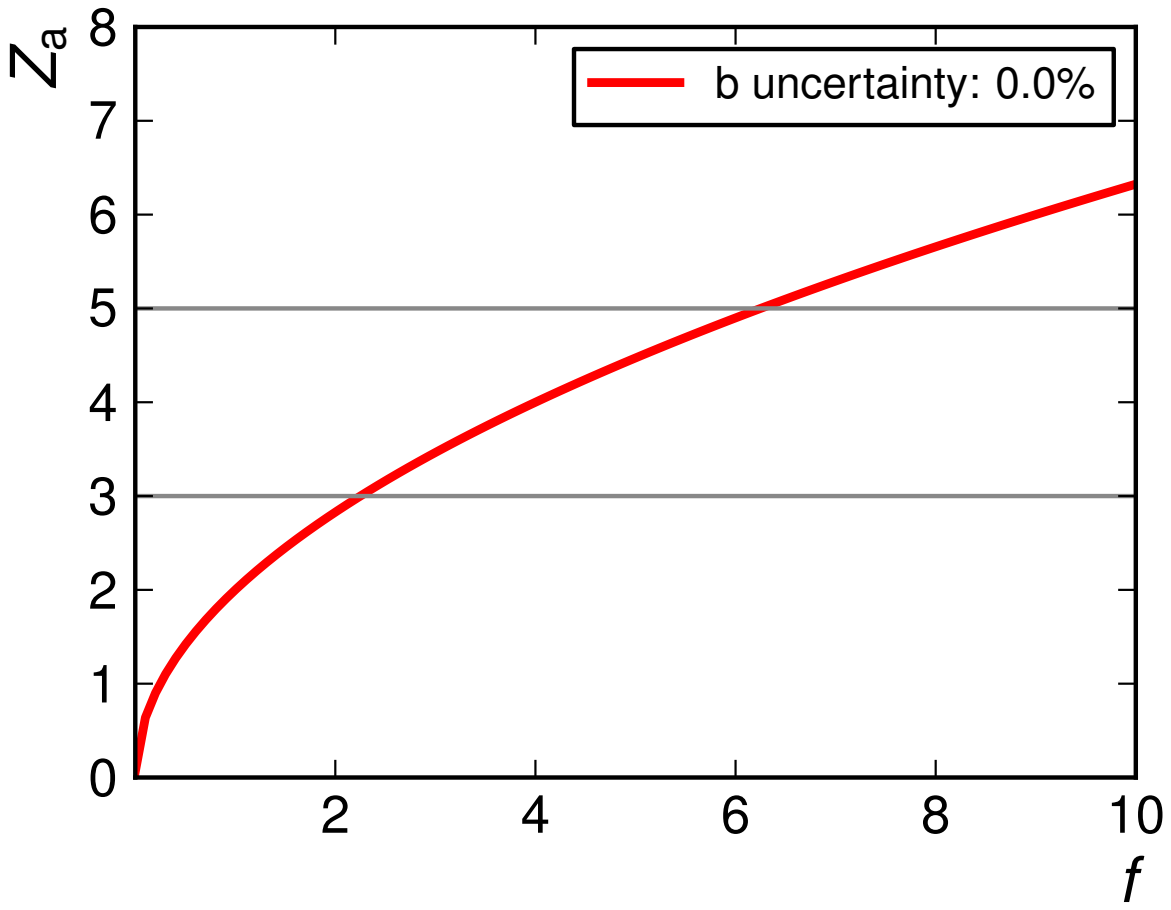
$$Z_{\mathrm{a}} = \frac{s}{\sqrt{b + (\Delta b)^2}} = \frac{n - b}{\sqrt{b + (\Delta b)^2}},$$

which is a generalization of the formula given earlier for $\Delta b = 0$.

This very simple formula exhibits some interesting features that help to understand some general properties of HEP analyses.

Normal Approximation: Luminosity Projection

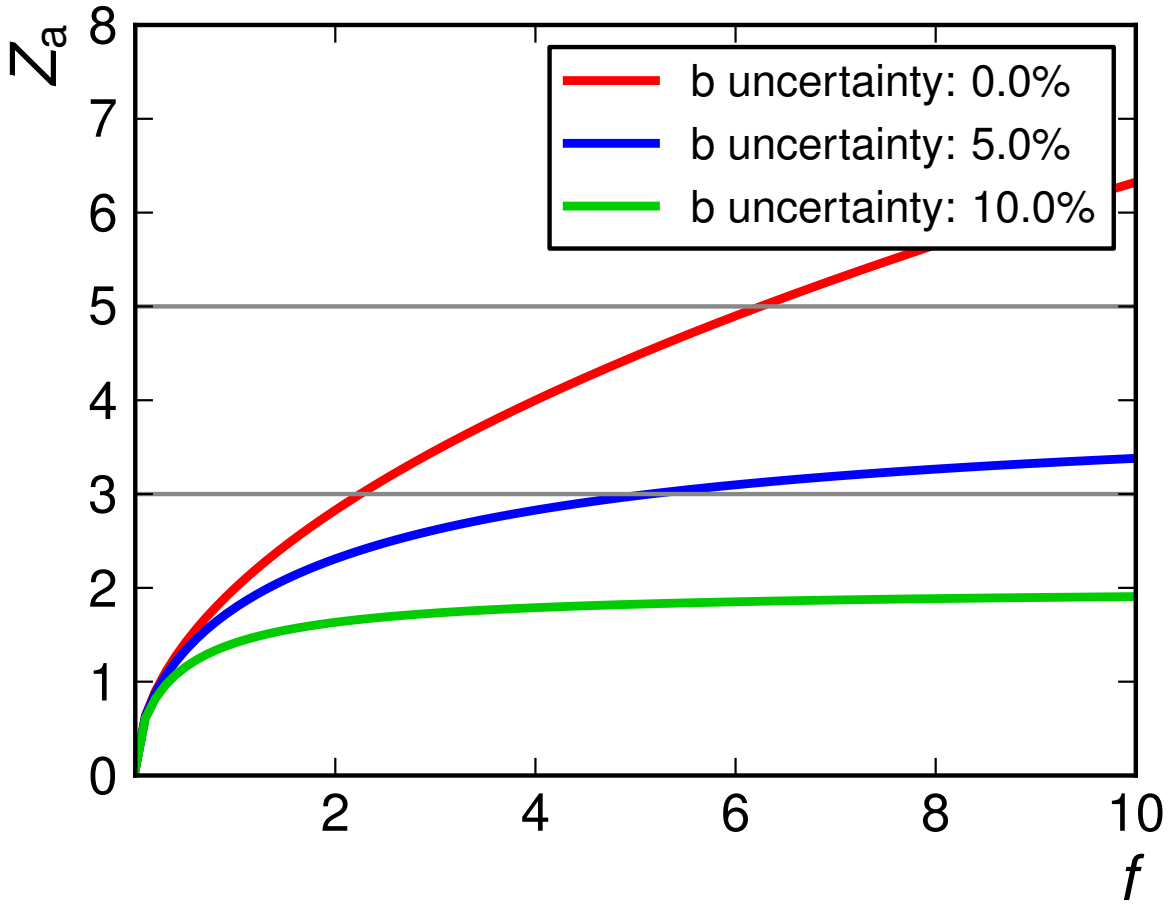
For $b = 100$, $s = 20$, the expected significance as a function of the luminosity scaling factor f:



Without uncertainty, increase like $\sqrt{f}$. With systematic uncertainty: expected significance limited by $s/\Delta b$.
$\rightsquigarrow s/\sqrt{b}$ is a good "figure of merit" if statistical uncertainties dominate; s/b better if background systematics dominate.

Normal Approximation: Luminosity Projection

For $b = 100$, $s = 20$, the expected significance as a function of the luminosity scaling factor f:



Without uncertainty, increase like $\sqrt{f}$. With systematic uncertainty: expected significance limited by $s/\Delta b$.
$\rightsquigarrow$ $s/\sqrt{b}$ is a good "figure of merit" if statistical uncertainties dominate; s/b better if background systematics dominate.

Section Summary

The Bayesian method assigns a prior to the nuisance parameter; the p-value is defined as the prior-averaged probability to observe such an H_0-incompatible test statistic value.

We also saw:

- How this can be applied to a counting experiment with a background rate uncertainty
- Asymptotic expected significance for the counting experiment and how this can motivate the widely-used "figures of merit" $s/\sqrt{b}$ and s/b

3 Handling of Systematic Uncertainties

- Bayesian/Hybrid Method; Counting Experiment
- Test Statistic Definition
- Shape Model Uncertainties

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.

3 Handling of Systematic Uncertainties

- Bayesian/Hybrid Method; Counting Experiment
- Test Statistic Definition
- Shape Model Uncertainties

Introduction

In general: Each uncertainty $\rightsquigarrow$ one (additional) nuisance parameter in the statistical model.

Next slides give examples for this principle for typical uncertainties of binned shape analyses

- Rate uncertainties from theory prediction, sideband estimation, ...
- Shape uncertainties from energy calibration/efficiency uncertainties, MC parameters, ...
- MC Statistic Uncertainties from limited size of MC sample (not covered here)

Formal Shape Model

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

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

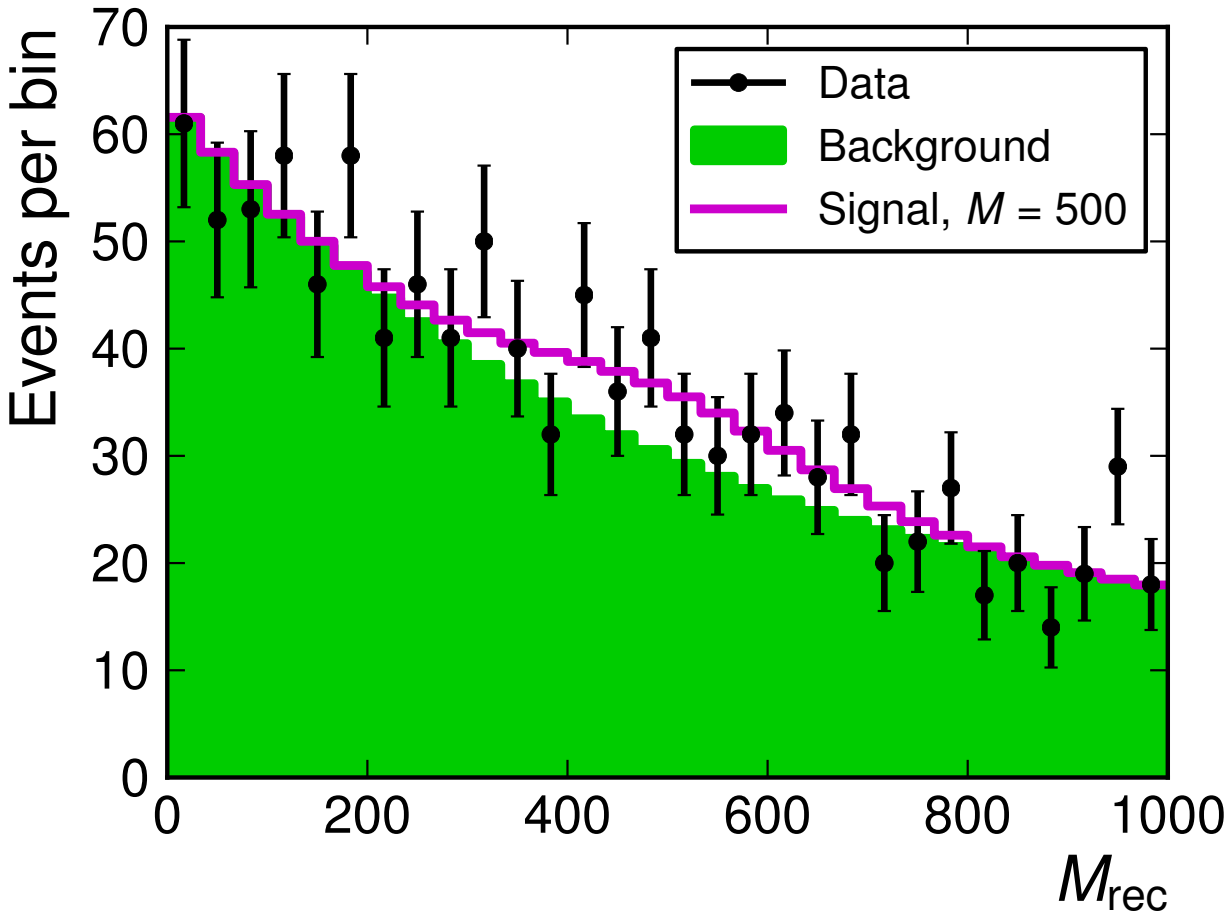
where i is the bin index and the expected number of events in bin i, λ_i, is given by the sum of (scaled) signal and background histograms:

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

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:

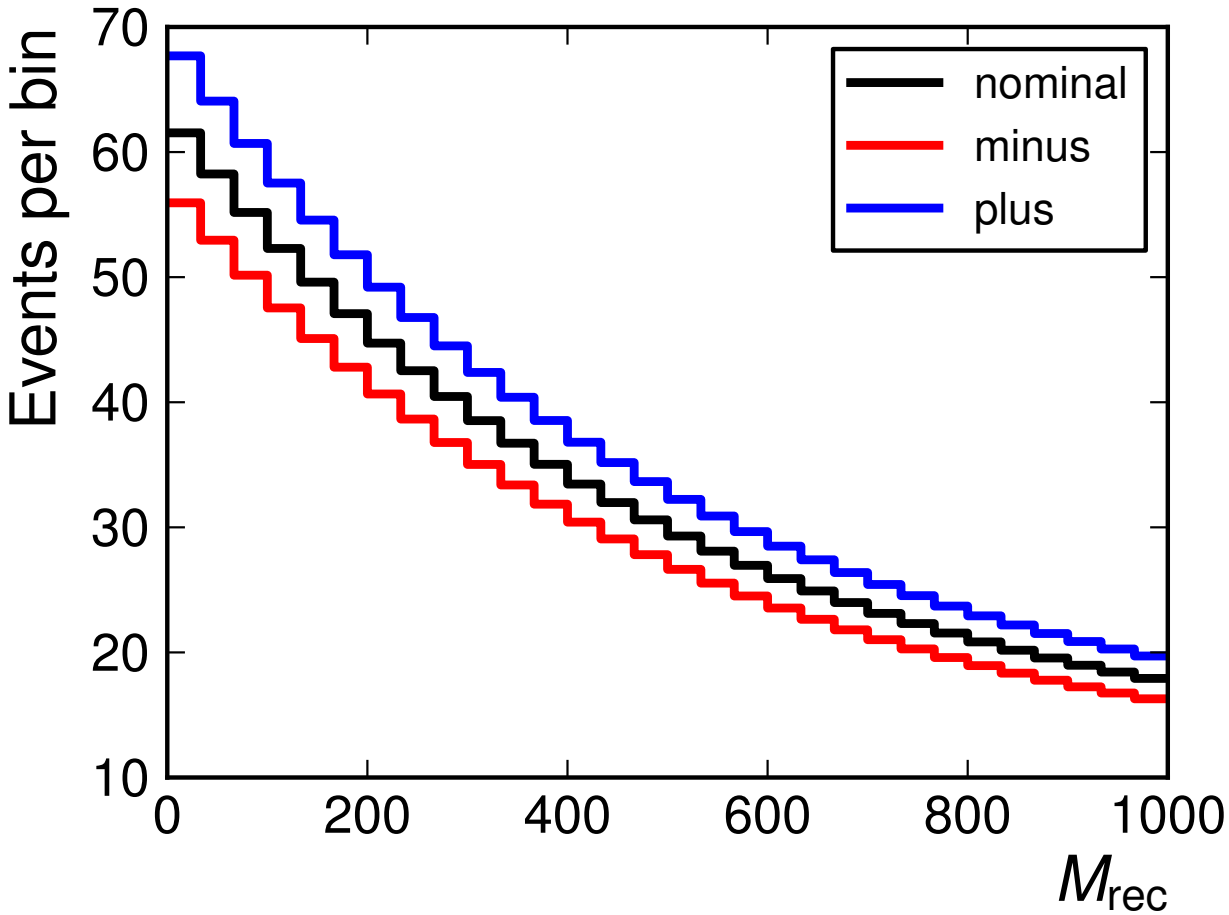


$\rightsquigarrow$ Have "plus" and "minus" templates by scaling the "nominal template up and down by 10%.

$\rightsquigarrow$ introduce nuisance parameter which scales the template accordingly.

Rate Uncertainties: Example

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

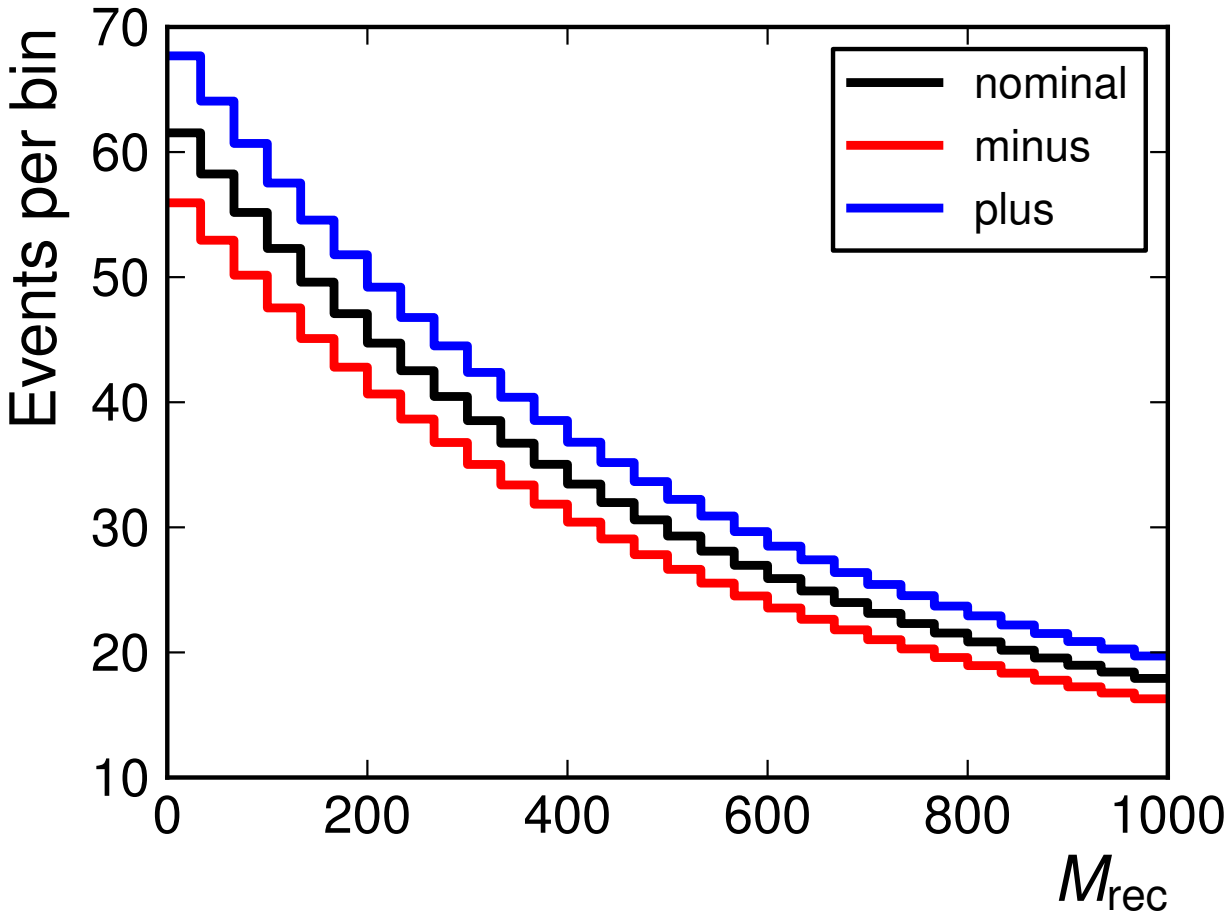


$\rightsquigarrow$ Have "plus" and "minus" templates by scaling the "nominal template up and down by 10%.

$\rightsquigarrow$ introduce nuisance parameter which scales the template accordingly.

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.

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

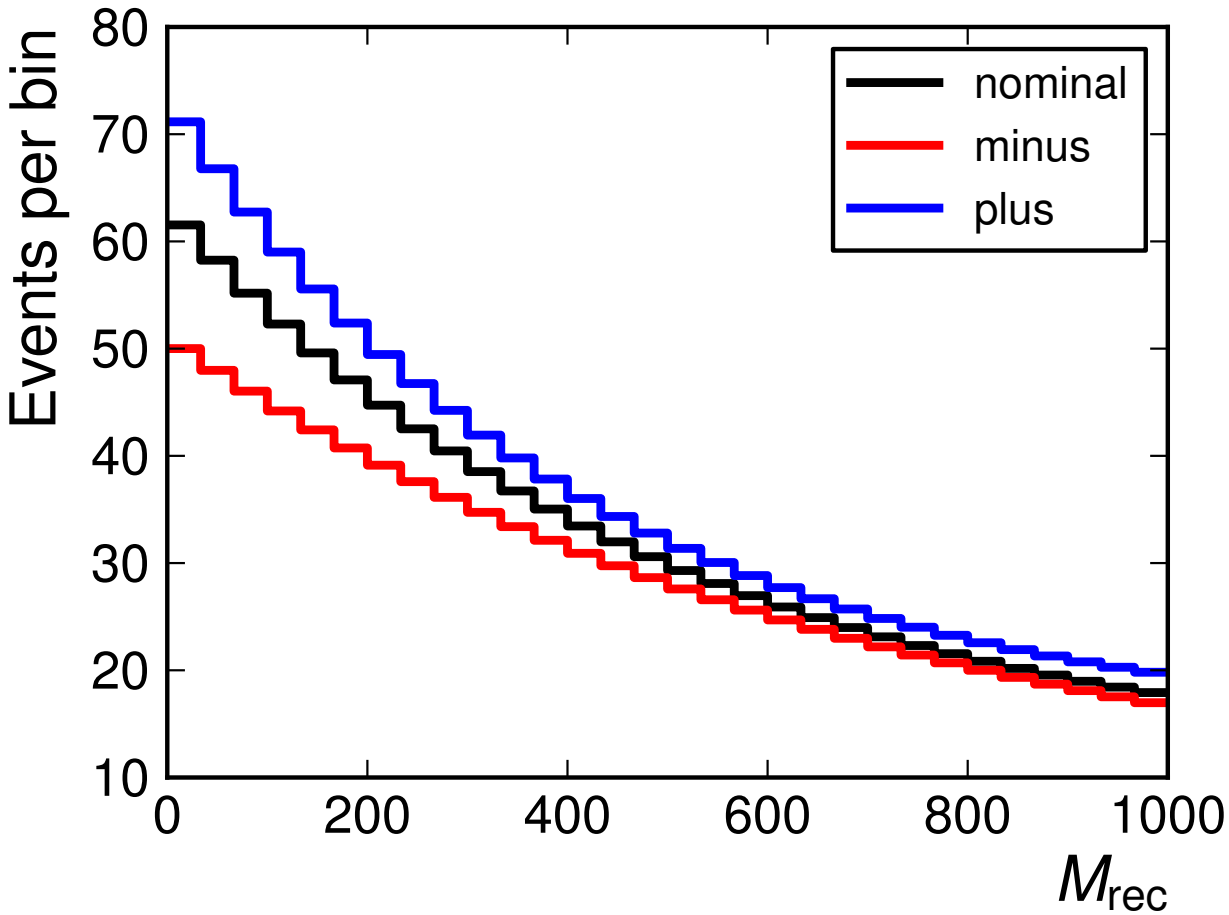
We just saw two different, but equivalent methods to introduce a log-normal scale factor.

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.

Shape Uncertainty: Introductory Example

In the shape model, assume you have three different methods to get the background shape, which look 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.

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

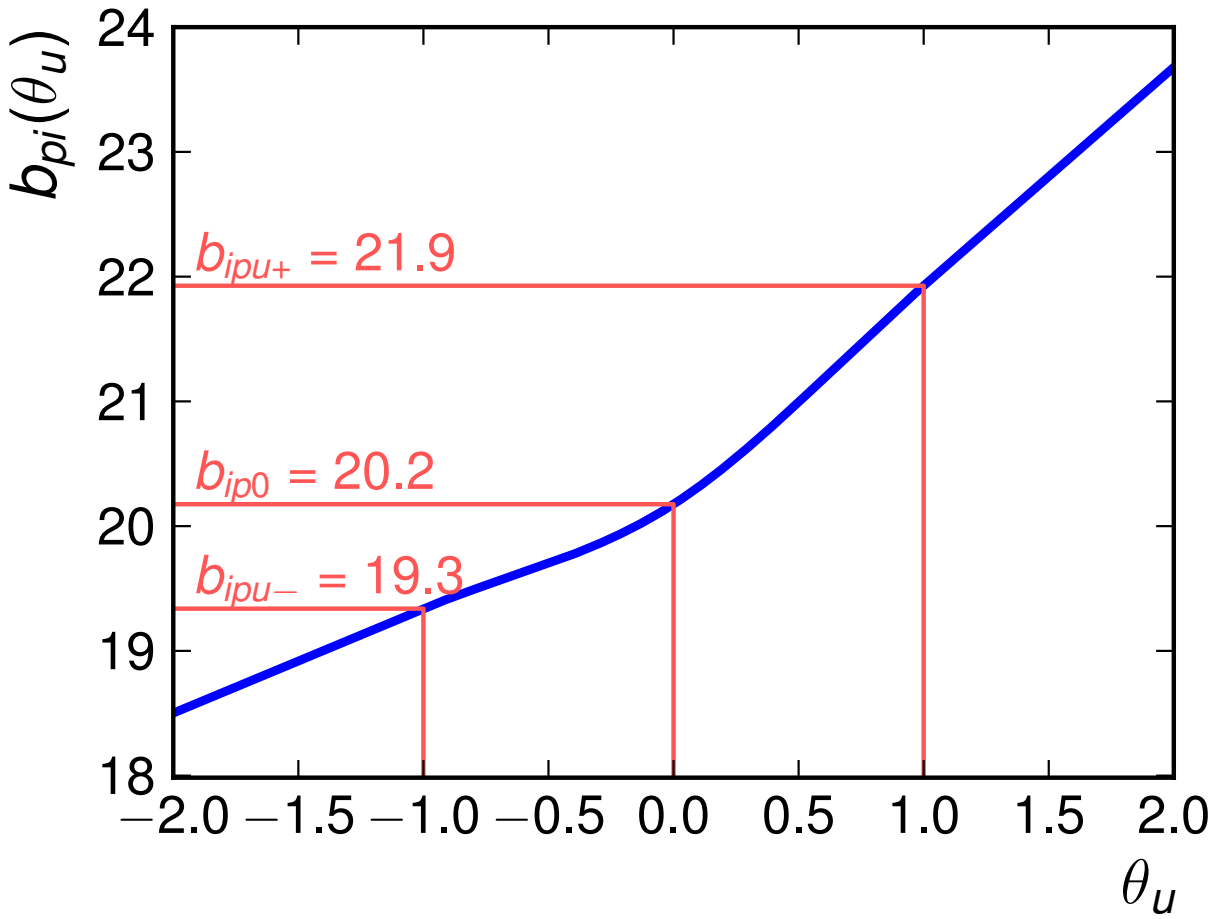
Follow general recipe: introduce nuisance parameter θ_u with standard normal prior, and write the model prediction for the Poisson mean λ_i as a function of the new parameter.

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

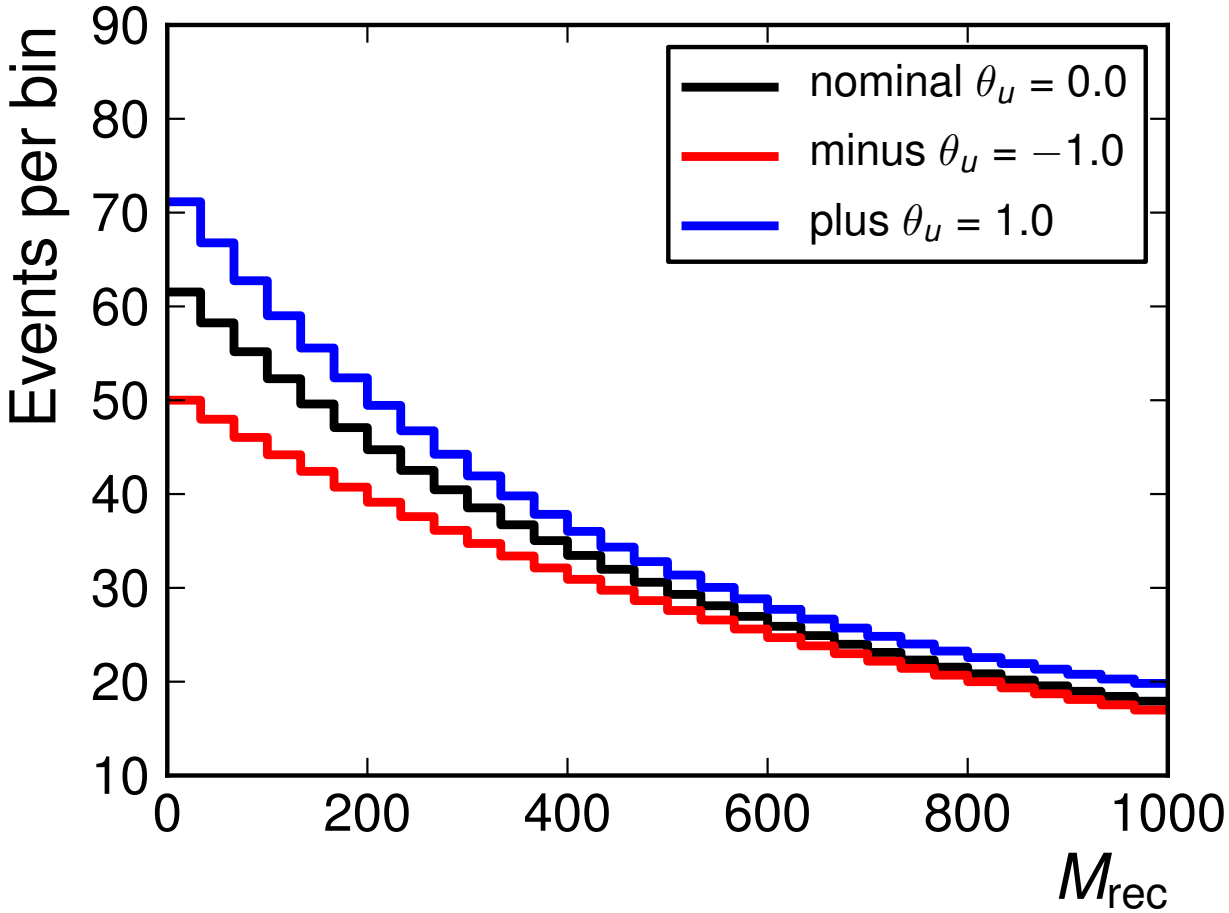
Example interpolation for the bin around $M_{\text{rec}} = 835$:



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

Shape Uncertainties: Shape Behavior

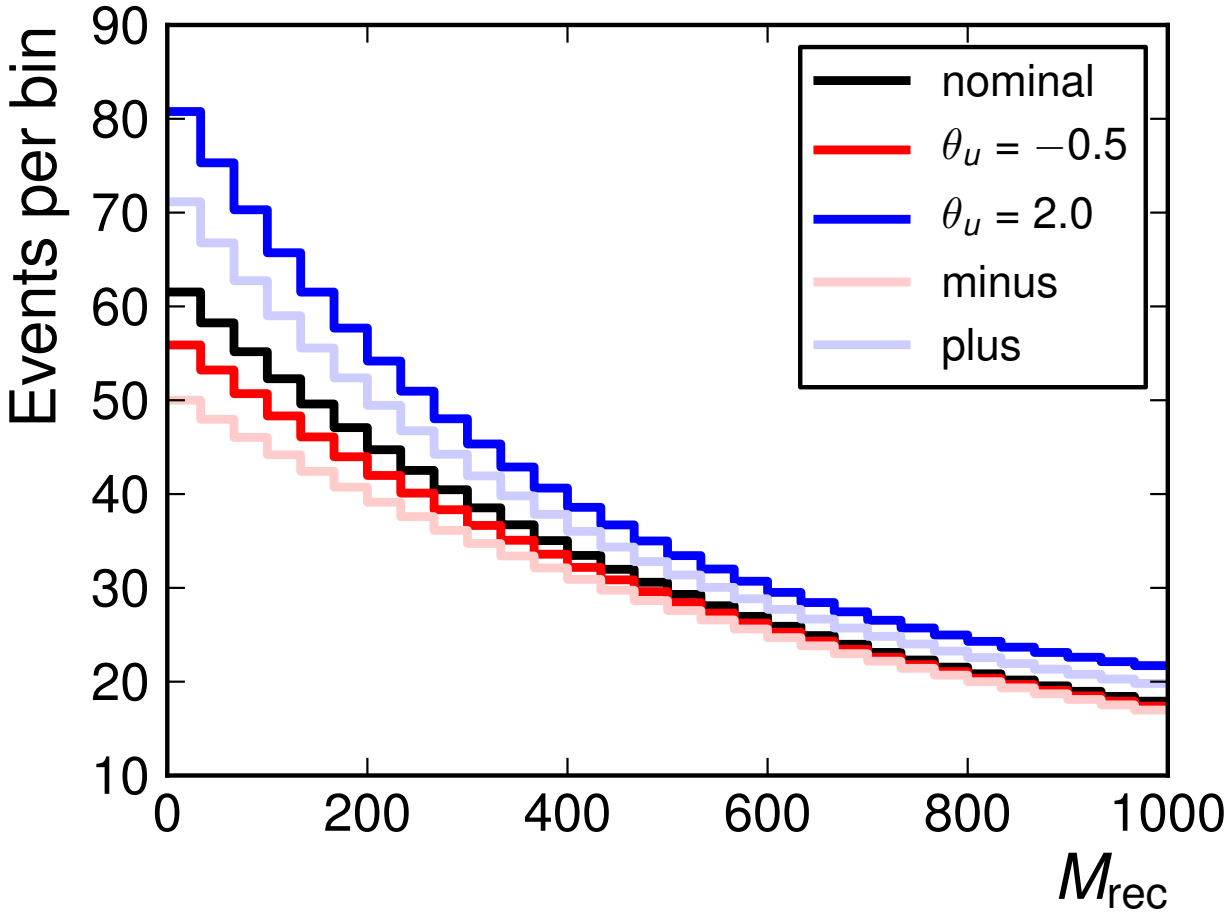
Applying template morphing for certain values θ_u, the background template looks like this:



Interpolation agrees with intuitive expectation.

Shape Uncertainties: Shape Behavior

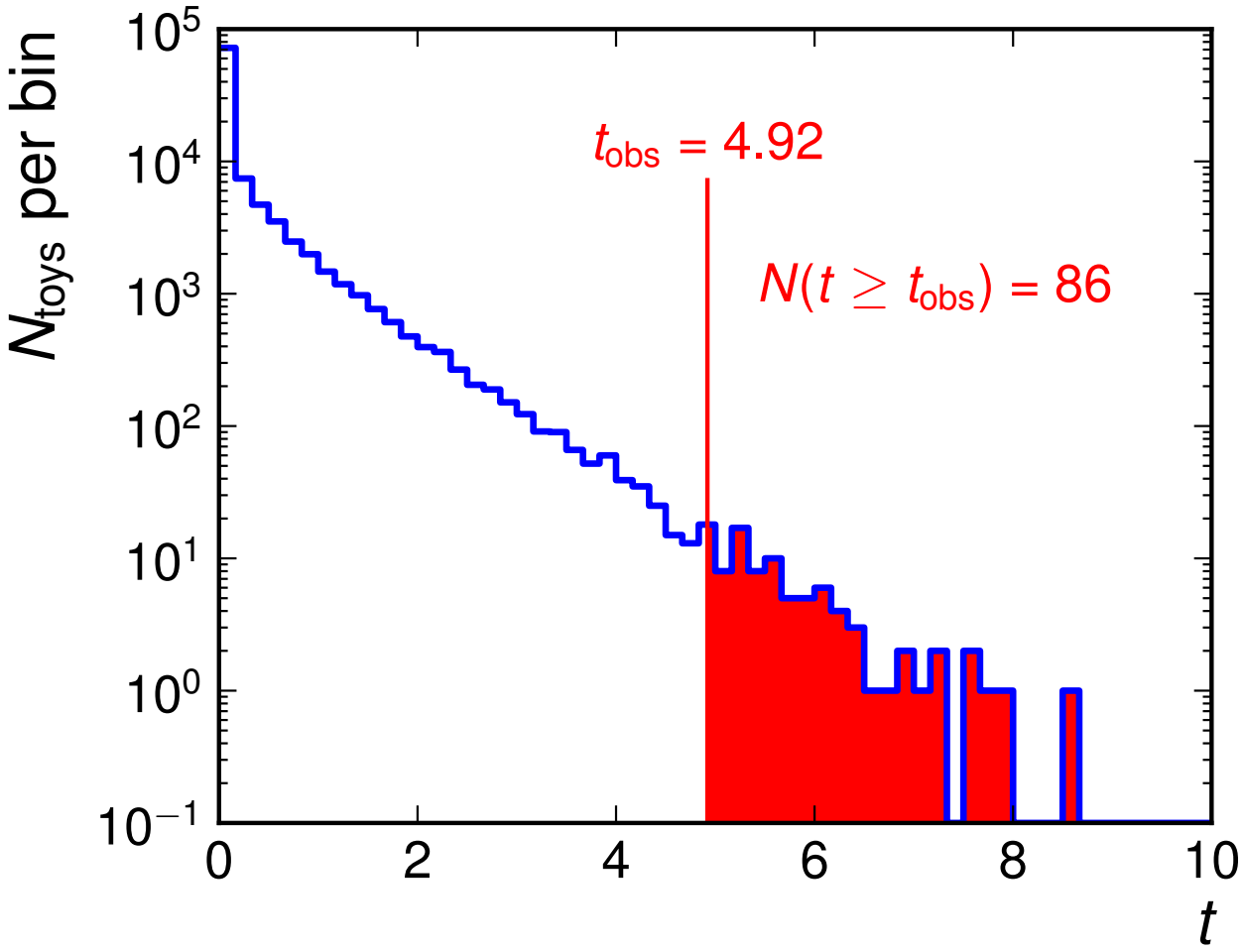
Applying template morphing for certain values θ_u, the background template looks like this:



Interpolation agrees with intuitive expectation.

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:



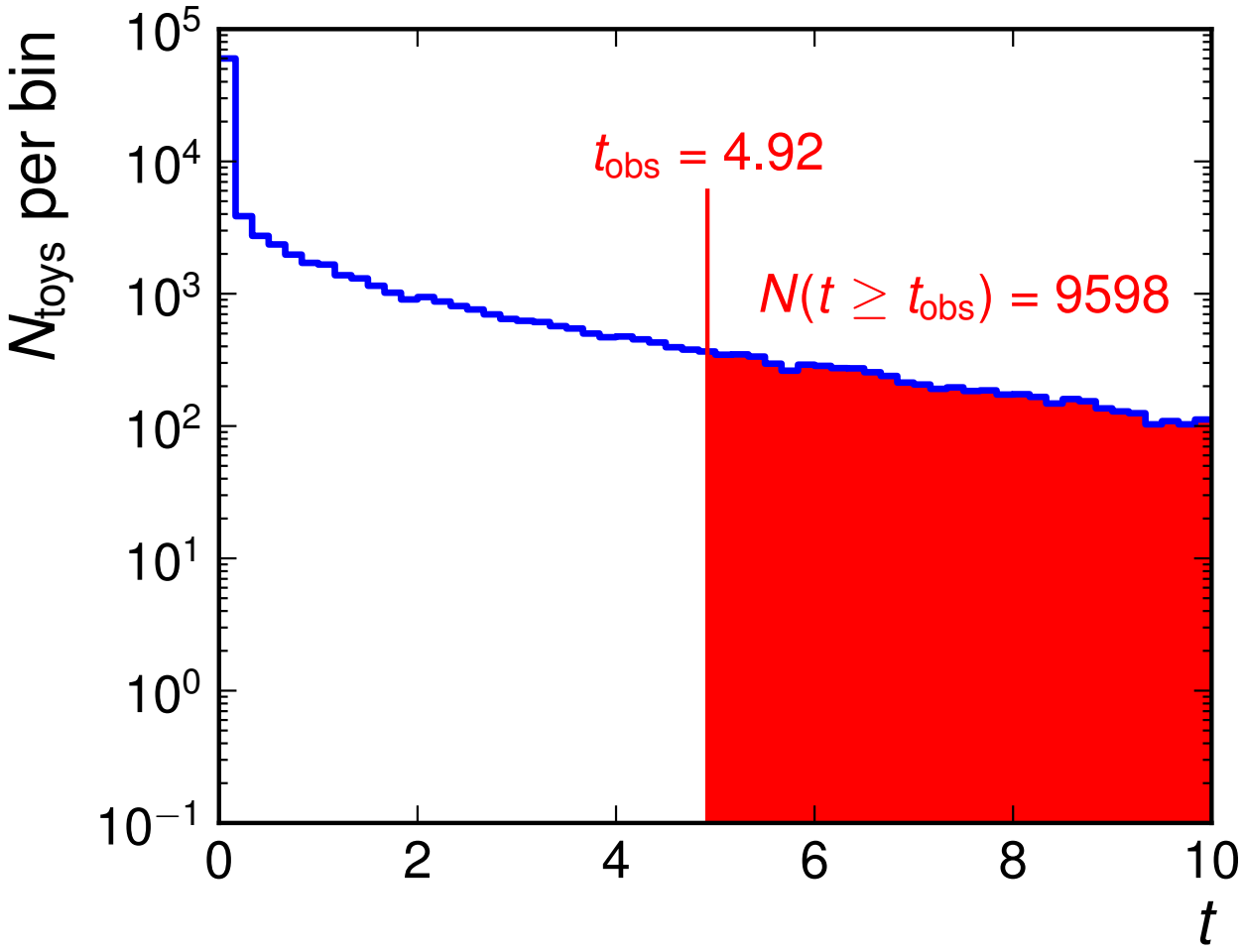
No uncertainties:
$\hat{p} = 0.00086$, $\hat{Z} = 3.1$.

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

With shape uncertainty: $\hat{p} = 0.052$, $\hat{Z} = 1.6$.

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:



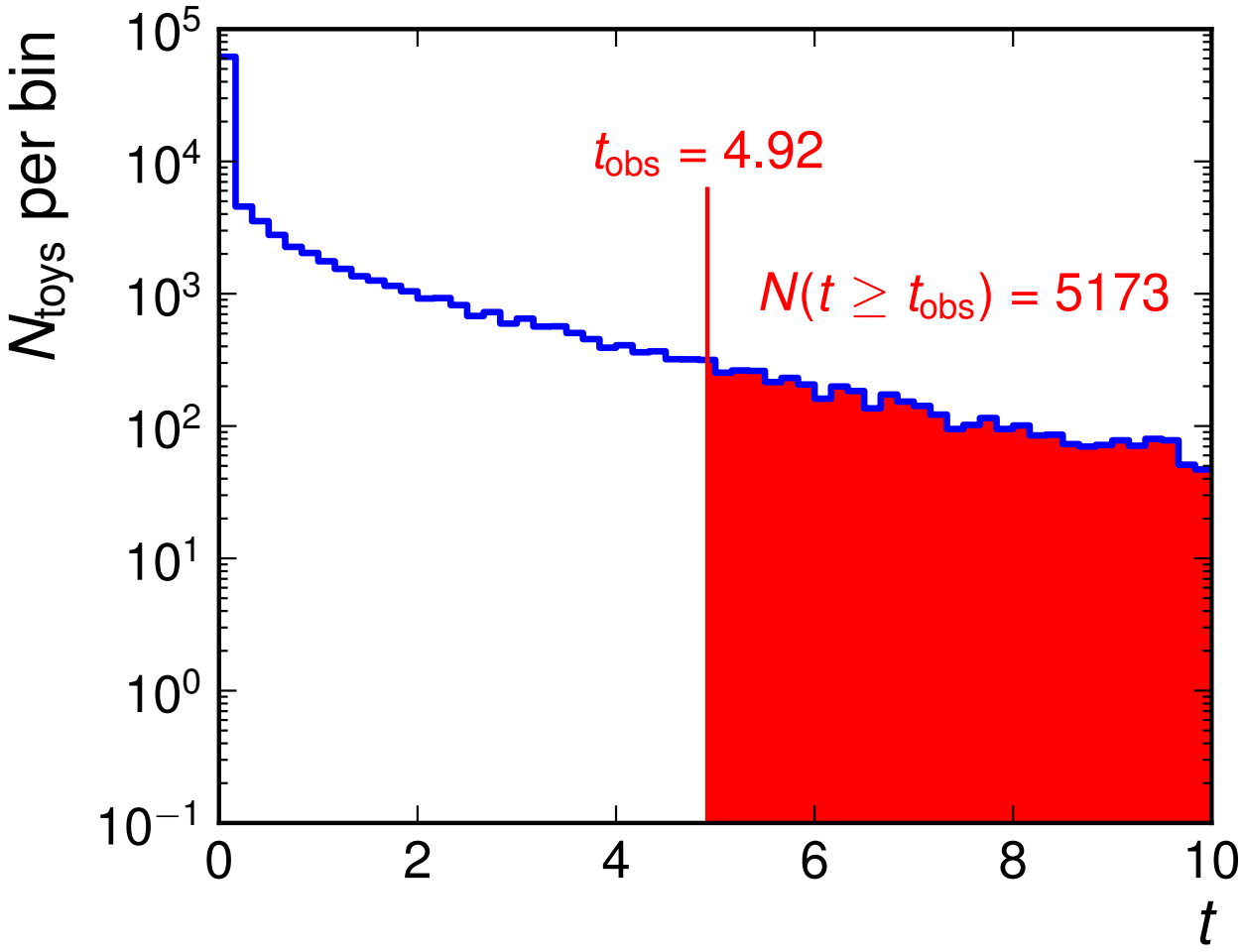
No uncertainties:
$\hat{p} = 0.00086$, $\hat{Z} = 3.1$.

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

With shape uncertainty: $\hat{p} = 0.052$, $\hat{Z} = 1.6$.

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: $\hat{p} = 0.096$, $\hat{Z} = 1.3$.

With shape uncertainty:
$\hat{p} = 0.052$, $\hat{Z} = 1.6$.

Comments

- Most methods are somewhat ad-hoc and use the simplest functional dependence for $\lambda(\theta)$. $\rightsquigarrow$ if result sensitive to this choice, something less arbitrary should be used.
- The shape model with template morphing are advanced techniques.
- Only covered binned analyses, but general principle also applies to models describing unbinned data: introduce nuisance parameters and parametrize model response.

Asymptotics: Motivation

For the p-value, need to know test statistic distribution for H_0. In general, this is not known analytically $\rightsquigarrow$ use toy Monte-Carlo.

Practical problem: large significance requires large number of toys ($\approx 10^8$ or more for $Z \geq 5$) $\rightsquigarrow$ long running time.

For counting experiment with background systematics, had simple formula for significance,

$$Z_{\mathrm{a}} = \frac{n - b}{\sqrt{b + (\Delta b)^2}}$$

which is valid for large n.

This also allowed easy computation for expected significance.

Now: Generalization of for more complicated models.

Wilks' Theorem

Suppose the null hypothesis H_0 corresponds to fixing certain model parameters in the statistical model, and the alternative hypothesis H_1 is the complement.

For $n \to \infty$, the likelihood-ratio test statistic

$$\Lambda = 2 \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)} = 2t$$

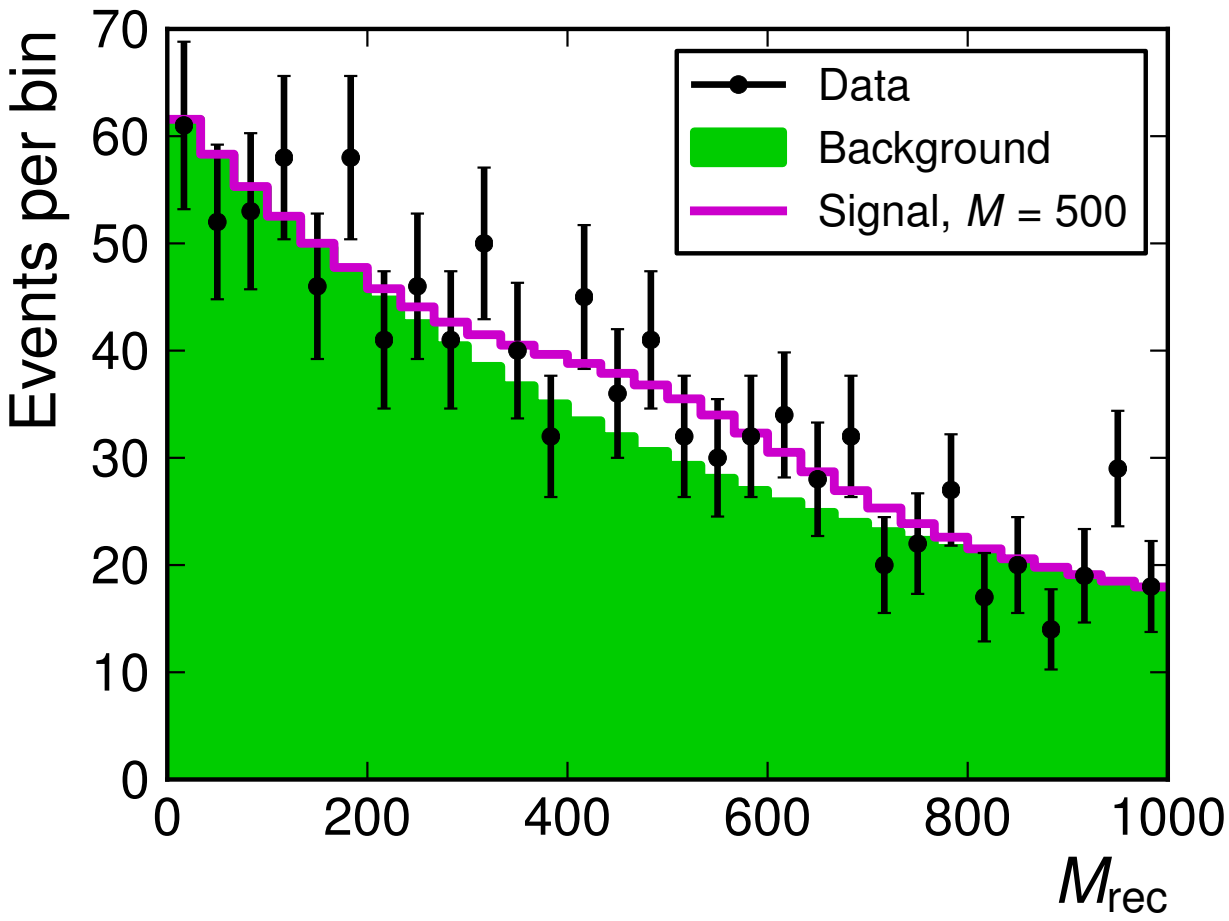
follows a χ^2_k-distribution where the number of degrees of freedom k is the number of fixed parameters in H_0.

For the case of searching a signal, $H_0 : \mu = 0$, $H_1 : \mu > 0$, there is 1 degree of freedom and the asymptotic Z-value is given by

$$Z_\mathrm{a} = \sqrt{2t}.$$

Asymptotics: Shape Example I

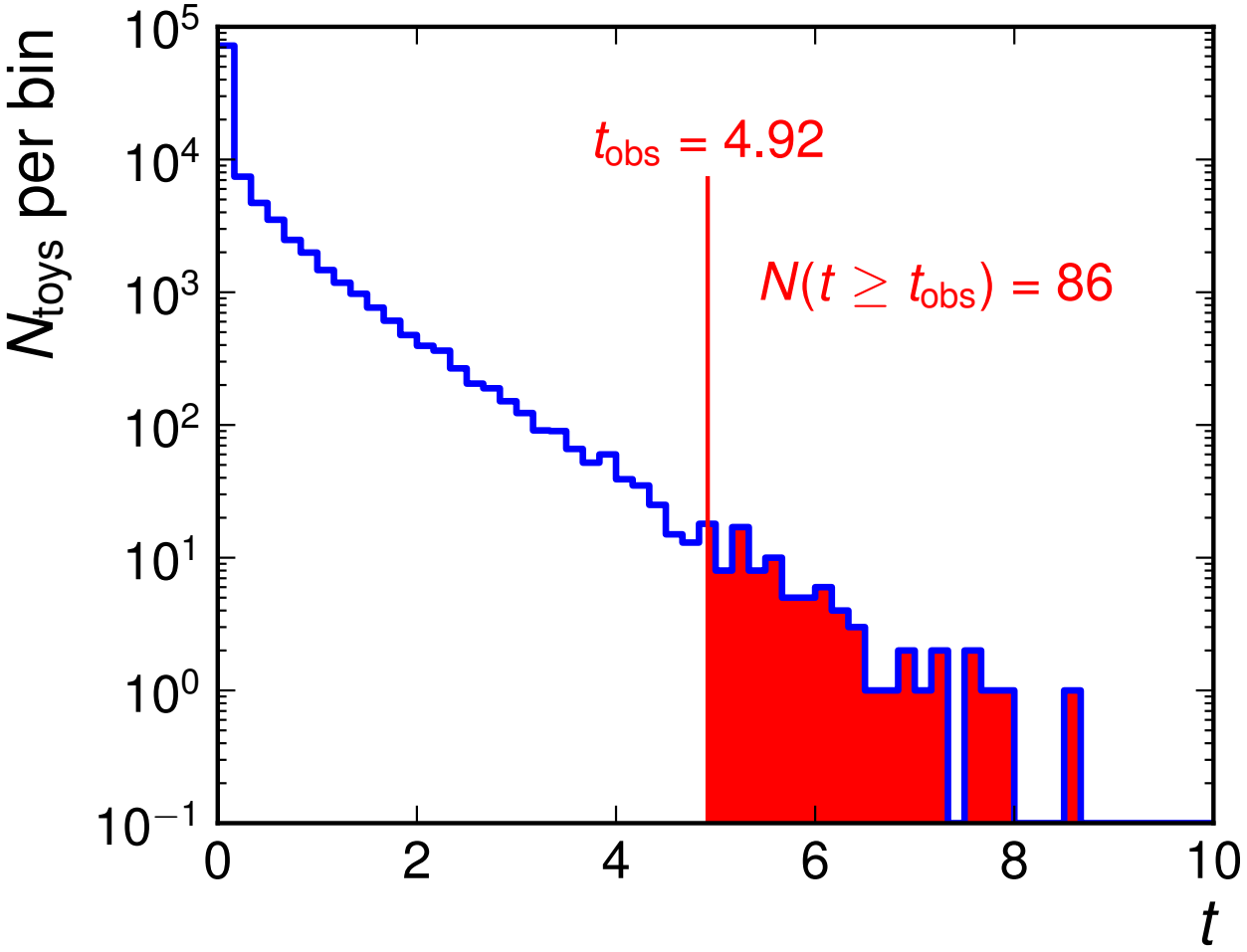
For the shape model, determines the Z-value via the tail distribution of the test statistic t.



Toy-based Z-value was $Z_t = 3.13 \pm 0.03$.
Asymptotic Z-value is $Z_a = \sqrt{2t_{\text{obs}}} = 3.14$.

Asymptotics: Shape Example I

For the shape model, determines the Z-value via the tail distribution of the test statistic t.



Toy-based Z-value was $Z_{\text{t}} = 3.13 \pm 0.03$.
Asymptotic Z-value is $Z_{\text{a}} = \sqrt{2t_{\text{obs}}} = 3.14$.

Comments

Only applicable in the "asymptotic regime" in case n is "large":

- In the asymptotic regime, can estimate all model parameters from data with errors shrinking as $\frac{1}{\sqrt{n}}$
- Often, can apply formulae even for smaller n, e.g. to the binned shape model if it approximately normal (n large in all bins; prior on θ is normal) and linear ($\lambda(\theta)$ is a linear function in θ)

If in doubt, check with Monte-Carlo.

Can generalize formula for asymptotic test statistic distribution for $\mu > 0$.
This is useful for expected significance and CLs limit construction.
This is used in Higgs searches; see Cowan, Cranmer, Gross, Vitells: Eur.Phys.J.C71:1554,2011; arXiv:1007.1727

Comments

Only applicable in the "asymptotic regime" in case n is "large":

- In the asymptotic regime, can estimate all model parameters from data with errors shrinking as $\frac{1}{\sqrt{n}}$
- Often, can apply formulae even for smaller n, e.g. to the binned shape model if it approximately normal (n large in all bins; prior on θ is normal) and linear ($\lambda(\theta)$ is a linear function in θ)

If in doubt, check with Monte-Carlo.

Can generalize formula for asymptotic test statistic distribution for $\mu > 0$.
This is useful for expected significance and CLs limit construction.
This is used in Higgs searches; see Cowan, Cranmer, Gross, Vitells: Eur.Phys.J.C71:1554,2011; arXiv:1007.1727

Comments

Only applicable in the "asymptotic regime" in case n is "large":

- In the asymptotic regime, can estimate all model parameters from data with errors shrinking as $\frac{1}{\sqrt{n}}$
- Often, can apply formulae even for smaller n, e.g. to the binned shape model if it approximately normal (n large in all bins; prior on θ is normal) and linear ($\lambda(\theta)$ is a linear function in θ)

If in doubt, check with Monte-Carlo.

Can generalize formula for asymptotic test statistic distribution for $\mu > 0$.
This is useful for expected significance and CLs limit construction.
This is used in Higgs searches; see Cowan, Cranmer, Gross, Vitells: Eur.Phys.J.C71:1554,2011; arXiv:1007.1727

Problem formulation

The threshold α is the type-I error rate: if p-value $< \alpha$, the null hypothesis is rejected.

If repeating this procedure many times using the same null hypothesis (but e.g. many different alternatives), the probability that some test rejects H_0 becomes much larger than α.

This is the look-elsewhere effect, and it has to be considered in the statistical procedure.

Simple Case

For n independent alternatives, the probability to reject H_0 is actually:

$$1 - (1 - \alpha)^n \approx n\alpha$$

(where the approximation assumes $n\alpha \ll 1$).

Example: If making 10 independent tests with $H_0 =$ Standard Model, the chance to see a "3σ" effect ($\alpha = 0.0013$) in one of them is 1.3%; if looking at 100 channels it is 12%.

The "independence" assumption made here means that the p-values follow independent uniform distributions on $[0, 1]$ under H_0.

In such a case, can correct the "local" p-value to a "global" p-value by dividing it by the "trial factor" n.

Simple Case

For n independent alternatives, the probability to reject H_0 is actually:

$$1 - (1 - \alpha)^n \approx n\alpha$$

(where the approximation assumes $n\alpha \ll 1$).

Example: If making 10 independent tests with $H_0 =$ Standard Model, the chance to see a "3σ" effect ($\alpha = 0.0013$) in one of them is 1.3%; if looking at 100 channels it is 12%.

The "independence" assumption made here means that the p-values follow independent uniform distributions on $[0, 1]$ under H_0.

In such a case, can correct the "local" p-value to a "global" p-value by dividing it by the "trial factor" n.

Simple Case

For n independent alternatives, the probability to reject H_0 is actually:

$$1 - (1 - \alpha)^n \approx n\alpha$$

(where the approximation assumes $n\alpha \ll 1$).

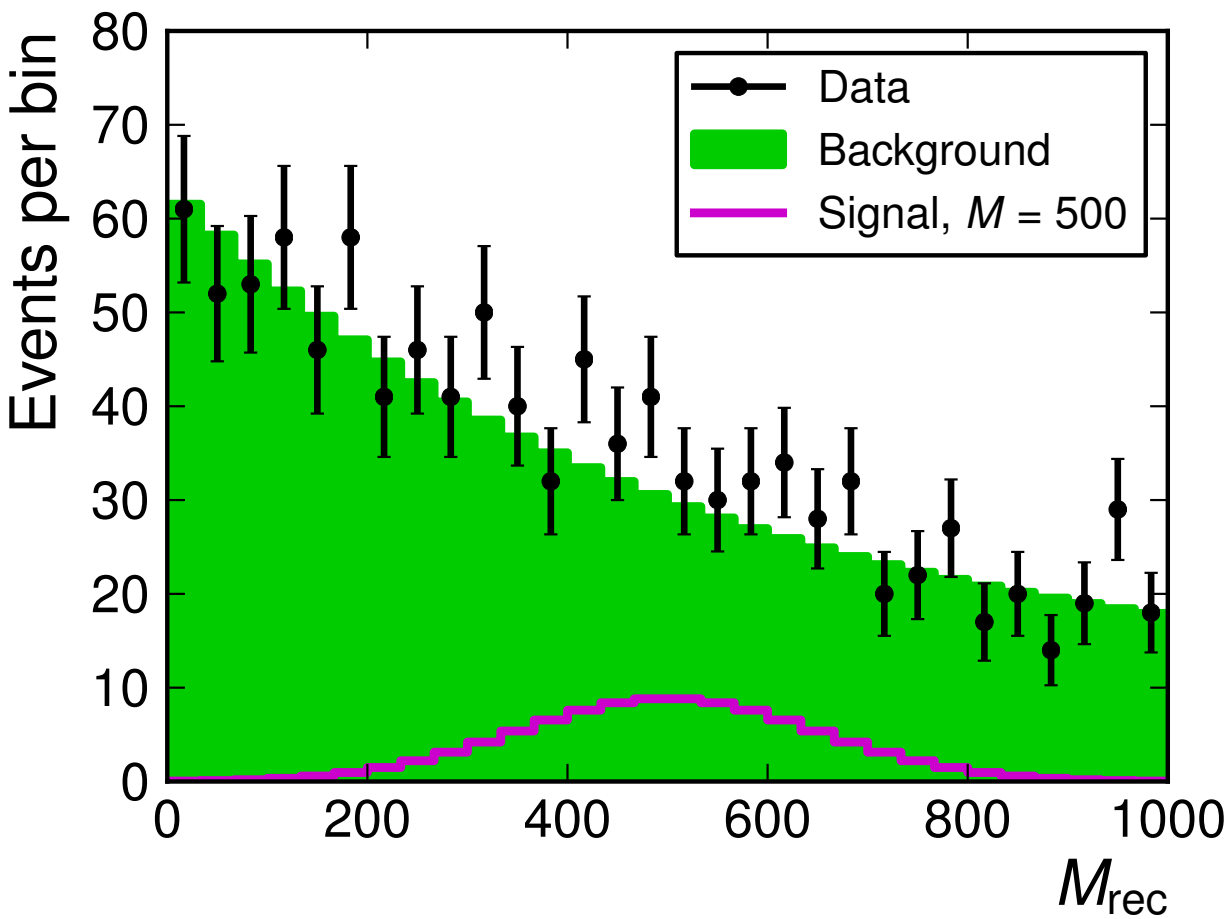
Example: If making 10 independent tests with $H_0 =$ Standard Model, the chance to see a "3σ" effect ($\alpha = 0.0013$) in one of them is 1.3%; if looking at 100 channels it is 12%.

The "independence" assumption made here means that the p-values follow independent uniform distributions on $[0, 1]$ under H_0.

In such a case, can correct the "local" p-value to a "global" p-value by dividing it by the "trial factor" n.

Realistic Example

Search for a particle with unknown mass, so instead of 1 possible signal at $M_{\text{rec}} = 500$, consider 3 potential signals.

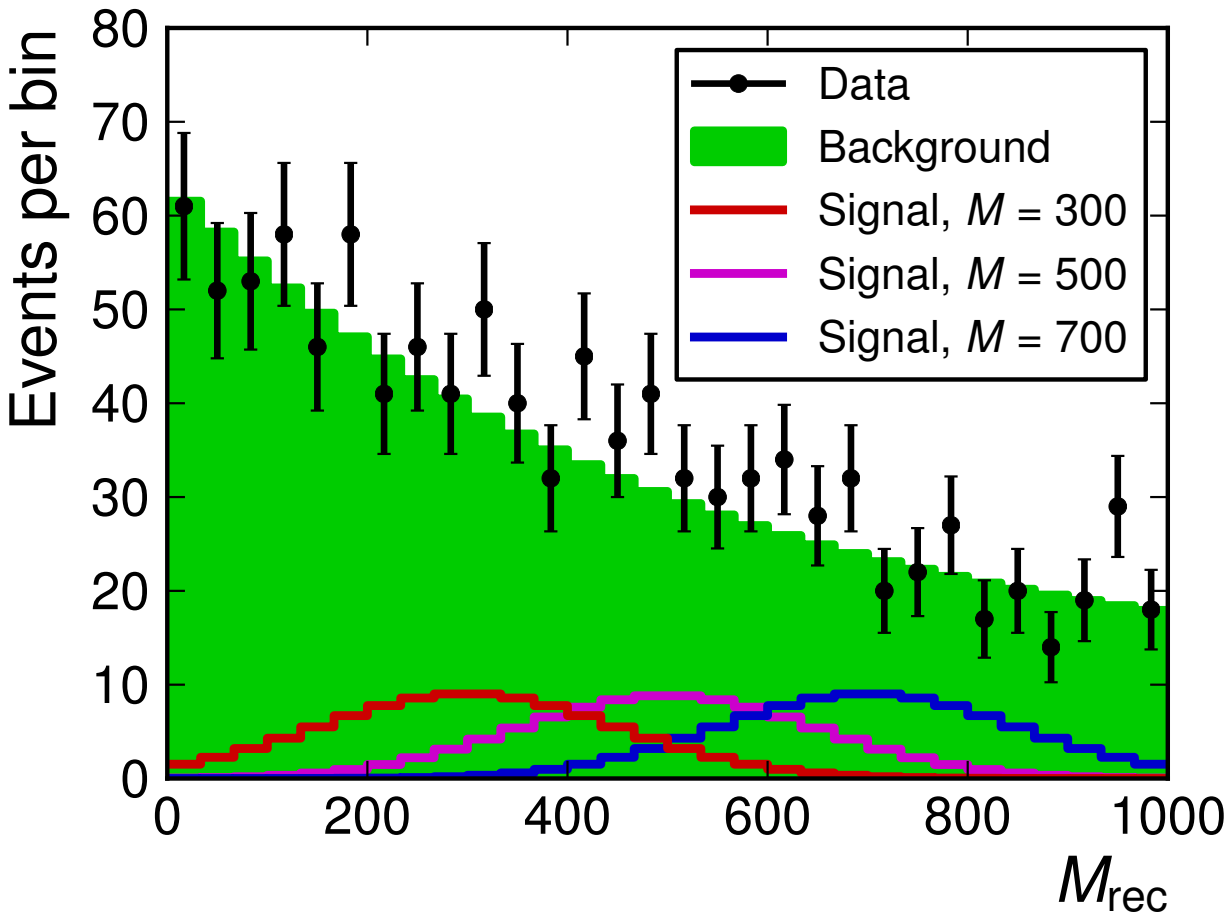


We had: $p = 8.6 \cdot 10^{-4}$; $Z = 3.1$ for a signal $m = 500$. But: if we searched on the whole spectrum, the probability to observe such a deviation anywhere is larger!

How likely is it to observe such a low p-value for any of the three masses?

Realistic Example

Search for a particle with unknown mass, so instead of 1 possible signal at $M_{\text{rec}} = 500$, consider 3 potential signals.



We had: $p = 8.6 \cdot 10^{-4}$; $Z = 3.1$ for a signal $m = 500$. But: if we searched on the whole spectrum, the probability to observe such a deviation anywhere is larger! How likely is it to observe such a low p-value for any of the three masses?

Example: Solution

In this case, can use a Monte-Carlo method:

1. Generate toy data according to background-only model H_0
2. For each toy, calculate the p-values for all three signal hypothesis "as usual" (using tail of TS distribution).
3. For each toy, calculate the minimum p-value over the three mass hypotheses. $\rightsquigarrow$ distribution for p_{min} under H_0
4. The "global", look-elsewhere-corrected p-value is given by the probability to observe $p_{\text{min}} \leq p_{\text{min}}^{\text{obs}}$.

Here: $p_{\text{min}}^{\text{obs}} = 8.6 \cdot 10^{-4}$. Probability to observe such low p_{min} is $2.2 \cdot 10^{-3}$.

Note that this is just the usual MC method for a hypothesis test with p_{min} as test statistic.

Example: Solution

In this case, can use a Monte-Carlo method:

1. Generate toy data according to background-only model H_0
2. For each toy, calculate the p-values for all three signal hypothesis "as usual" (using tail of TS distribution).
3. For each toy, calculate the minimum p-value over the three mass hypotheses. $\rightsquigarrow$ distribution for p_{min} under H_0
4. The "global", look-elsewhere-corrected p-value is given by the probability to observe $p_{\text{min}} \leq p_{\text{min}}^{\text{obs}}$.

Here: $p_{\text{min}}^{\text{obs}} = 8.6 \cdot 10^{-4}$. Probability to observe such low p_{min} is $2.2 \cdot 10^{-3}$.

Note that this is just the usual MC method for a hypothesis test with p_{min} as test statistic.

Example: Solution

In this case, can use a Monte-Carlo method:

1. Generate toy data according to background-only model H_0
2. For each toy, calculate the p-values for all three signal hypothesis "as usual" (using tail of TS distribution).
3. For each toy, calculate the minimum p-value over the three mass hypotheses. $\rightsquigarrow$ distribution for p_{min} under H_0
4. The "global", look-elsewhere-corrected p-value is given by the probability to observe $p_{\text{min}} \leq p_{\text{min}}^{\text{obs}}$.

Here: $p_{\text{min}}^{\text{obs}} = 8.6 \cdot 10^{-4}$. Probability to observe such low p_{min} is $2.2 \cdot 10^{-3}$.

Note that this is just the usual MC method for a hypothesis test with p_{min} as test statistic.

Example: Trial Factor and Resolution

- If testing many signals in small spacing: signal shapes overlap due to finite resolution.
- Data with a small p-value for e.g. $m = 500$ will also have small p-value if testing for $m = 501$.
- Joint distribution of p-values for different hypothesis no longer independent.
- Factor between global p-value and local p-value is smaller than number of tested signals.

In the example, had a trial factor of 2.6, somewhat smaller than the number of channels 3. For many tested masses, the number of "independent" channels is of the order

$$\frac{\text{tested mass range}}{\text{mass resolution}}$$

and is independent of the number of tested masses.

Example: Trial Factor and Resolution

- If testing many signals in small spacing: signal shapes overlap due to finite resolution.
- Data with a small p-value for e.g. $m = 500$ will also have small p-value if testing for $m = 501$.
- Joint distribution of p-values for different hypothesis no longer independent.
- Factor between global p-value and local p-value is smaller than number of tested signals.

In the example, had a trial factor of 2.6, somewhat smaller than the number of channels 3. For many tested masses, the number of "independent" channels is of the order

$$\frac{\text{tested mass range}}{\text{mass resolution}}$$

and is independent of the number of tested masses.

Example: Trial Factor and Resolution

- If testing many signals in small spacing: signal shapes overlap due to finite resolution.
- Data with a small p-value for e.g. $m = 500$ will also have small p-value if testing for $m = 501$.
- Joint distribution of p-values for different hypothesis no longer independent.
- Factor between global p-value and local p-value is smaller than number of tested signals.

In the example, had a trial factor of 2.6, somewhat smaller than the number of channels 3. For many tested masses, the number of "independent" channels is of the order

$$\frac{\text{tested mass range}}{\text{mass resolution}}$$

and is independent of the number of tested masses.

Example: Trial Factor and Resolution

- If testing many signals in small spacing: signal shapes overlap due to finite resolution.
- Data with a small p-value for e.g. $m = 500$ will also have small p-value if testing for $m = 501$.
- Joint distribution of p-values for different hypothesis no longer independent.
- Factor between global p-value and local p-value is smaller than number of tested signals.

In the example, had a trial factor of 2.6, somewhat smaller than the number of channels 3. For many tested masses, the number of "independent" channels is of the order

$$\frac{\text{tested mass range}}{\text{mass resolution}}$$

and is independent of the number of tested masses.

General Handling of Look-Elsewhere Effect

To cite a "global" p-value, one can:

- Can use trial factor n directly. But: Only possible if channels are (approximately) independent, or if "effective" number of independent channels known.
- Use Monte-Carlo to generate toy data according to background-only as in the example. But: assumes that we can generate background-only toys suitable for all signal hypotheses. (If we e.g. use a different event selection for each signal, that's very hard and might be unfeasible.)
- Look at the fluctuations of the observed limit w.r.t. the expected limits in data $\rightsquigarrow$ estimate for effective number of independent channels, which can be used as trial factor

  (see Gross, Vittels: Eur.Phys.J.C70:525-530,2010; arXiv:1005.1891; this is the method used in LHC Higgs searches)

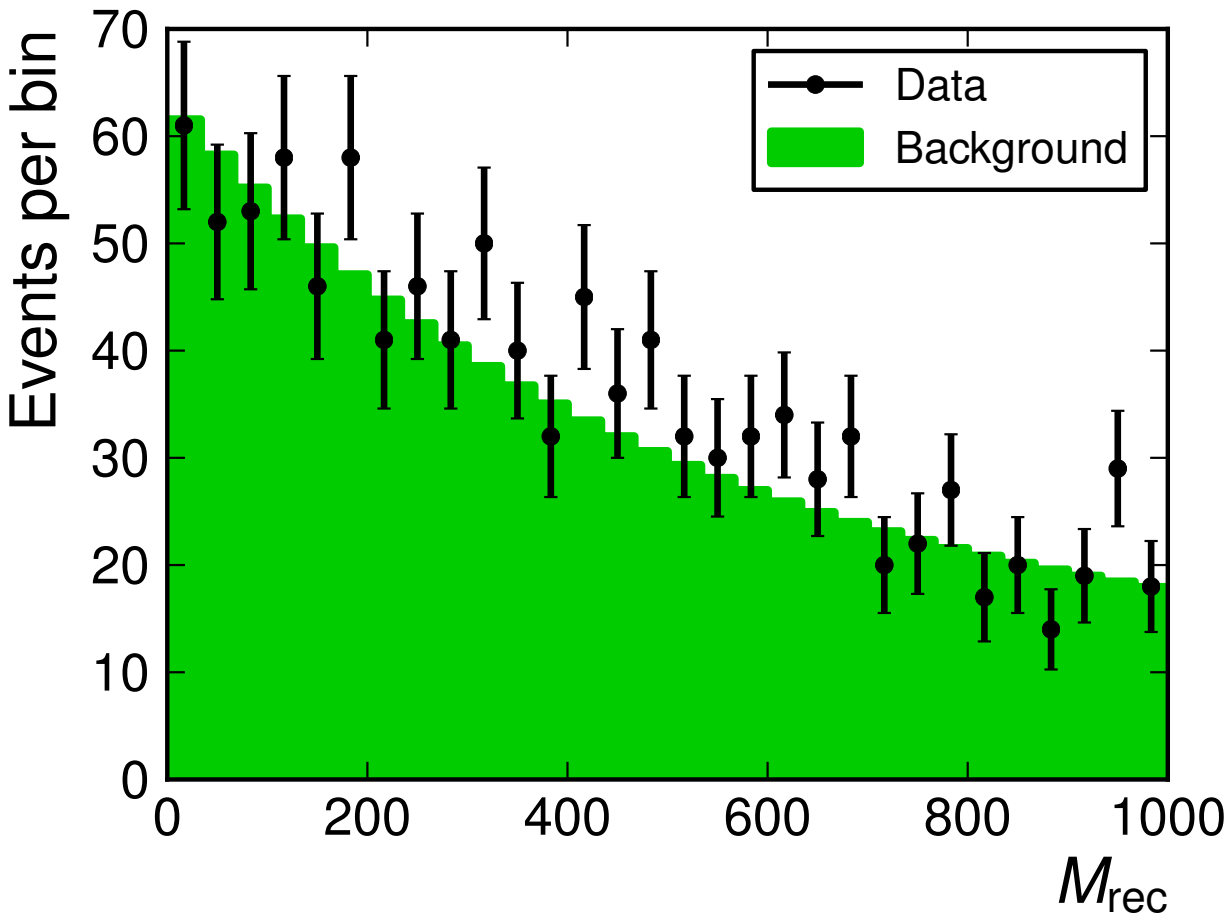
Other Incarnations of LEE

Not only in searches for particles with unknown mass, but whenever making many hypothesis tests, one might be susceptible to the LEE and should correct for it, e.g. when

- repeating tests for different dataset sizes n (cf. to Luc's example at the end of first set of slides),
- using different test statistics for the same data–model comparisons.
- making a large number of data–model or data–data comparisons (e.g. making many monitoring histograms)

Definition

A goodness of fit test is a hypothesis test where the null hypothesis H_0 is: the data follows the distribution of the statistical model.



Are the data compatible with the background-only hypothesis?

Note: no explicit alternative.

χ^2 Test

χ^2 test statistic is defined as the sum of k independent normally distributed quantities X_i,

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - \mu_i)^2}{\sigma_i^2}.$$

For making data-model comparisons, have to use binned data (k =number of bins):

- μ_i is the Poisson mean for this bin as predicted by the model
- X_i is the number of events in bin i
- $\sigma_i = \sqrt{\mu_i}$ is the standard deviation for the normal approximation of a Poisson distribution

χ^2 follows the (well-known) χ^2 distribution with k degrees of freedom.

χ^2 Test

χ^2 test statistic is defined as the sum of k independent normally distributed quantities X_i,

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - \mu_i)^2}{\sigma_i^2}.$$

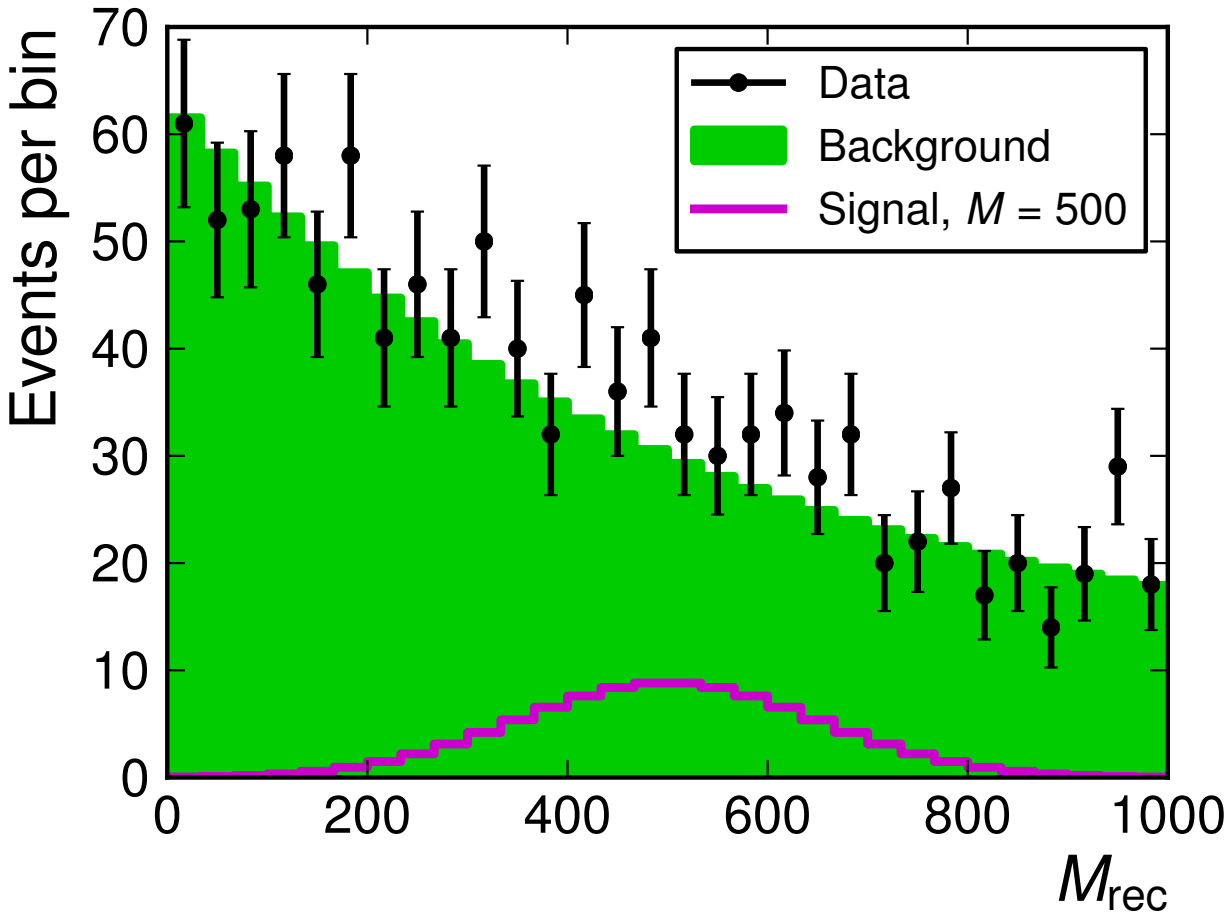
For making data-model comparisons, have to use binned data (k =number of bins):

- μ_i is the Poisson mean for this bin as predicted by the model
- X_i is the number of events in bin i
- $\sigma_i = \sqrt{\mu_i}$ is the standard deviation for the normal approximation of a Poisson distribution

χ^2 follows the (well-known) χ^2 distribution with k degrees of freedom.

χ^2 Test: Example

For the shape model, compare data to background-only with χ^2 test (remember: had $Z = 3.1$ for hypothesis test sensitive to $M = 500$):

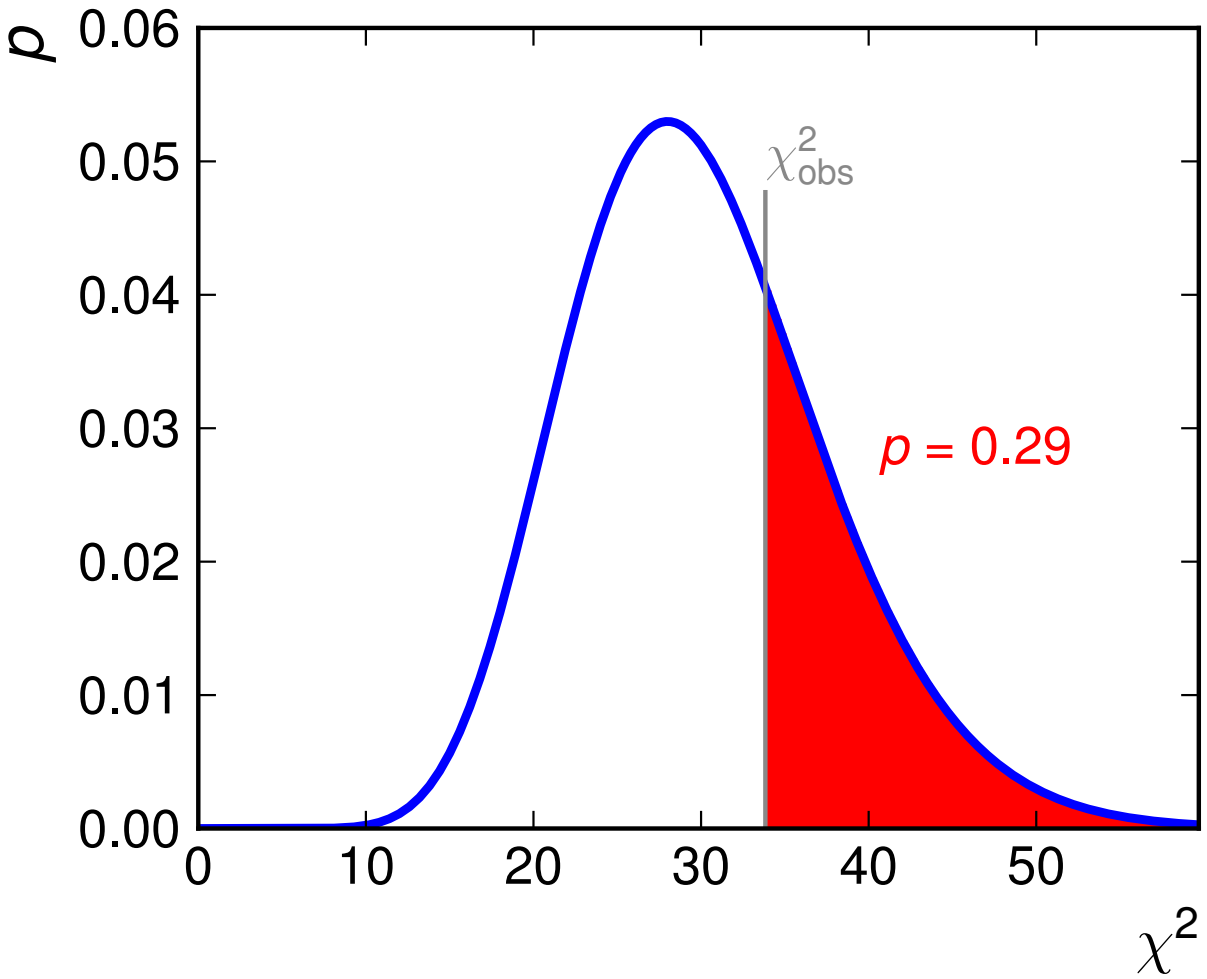


$\chi^2_{\text{obs}} = 33.8$

p-value: probability to observe even more extreme (larger) χ^2 value: $p_{\chi^2} = 0.29$, $Z_{\chi^2} = 0.56$.

χ^2 Test: Example

For the shape model, compare data to background-only with χ^2 test (remember: had $Z = 3.1$ for hypothesis test sensitive to $M = 500$):



$\chi^2_{obs} = 33.8$
p-value: probability to observe even more extreme (larger) χ^2 value: $p_{\chi^2} = 0.29$, $Z_{\chi^2} = 0.56$.

χ^2 Test: Comments

- χ^2 test works requires binned data
- μ_i should be large enough to justify the normal approximation of the Poisson.
- In general, fitting the model to the data will reduce data–model difference $\rightsquigarrow$ degrees of freedom for χ^2 distribution is reduced by number of fit parameters.
- χ^2 test sensitive to "overall deviation": In current example, χ^2 test was not sensitive to "local" 3.1σ effect of $M = 500$ peak.

KS Test

As test statistic, use the maximum of the difference in the cumulative distributions of model and data:

$$d_{\max} = \max_x |F_d(x) - F_m(x)|$$

where $F_d(x)$ is the cumulative distribution for the data ("empirical distribution function"), and $F_m(x)$ for the tested model.

KS test does not require binning of the data; binning decreases $d_{\max} \rightsquigarrow$ data look more compatible.

For large n: $p \approx \frac{1}{2} e^{-2nd_{\max}}$.

KS Test

As test statistic, use the maximum of the difference in the cumulative distributions of model and data:

$$d_{\text{max}} = \max_x |F_d(x) - F_m(x)|$$

where $F_d(x)$ is the cumulative distribution for the data ("empirical distribution function"), and $F_m(x)$ for the tested model.

KS test does not require binning of the data; binning decreases $d_{\text{max}} \rightsquigarrow$ data look more compatible.

For large n: $p \approx \frac{1}{2} e^{-2nd_{\text{max}}}$.

KS Test

As test statistic, use the maximum of the difference in the cumulative distributions of model and data:

$$d_{\text{max}} = \max_x |F_d(x) - F_m(x)|$$

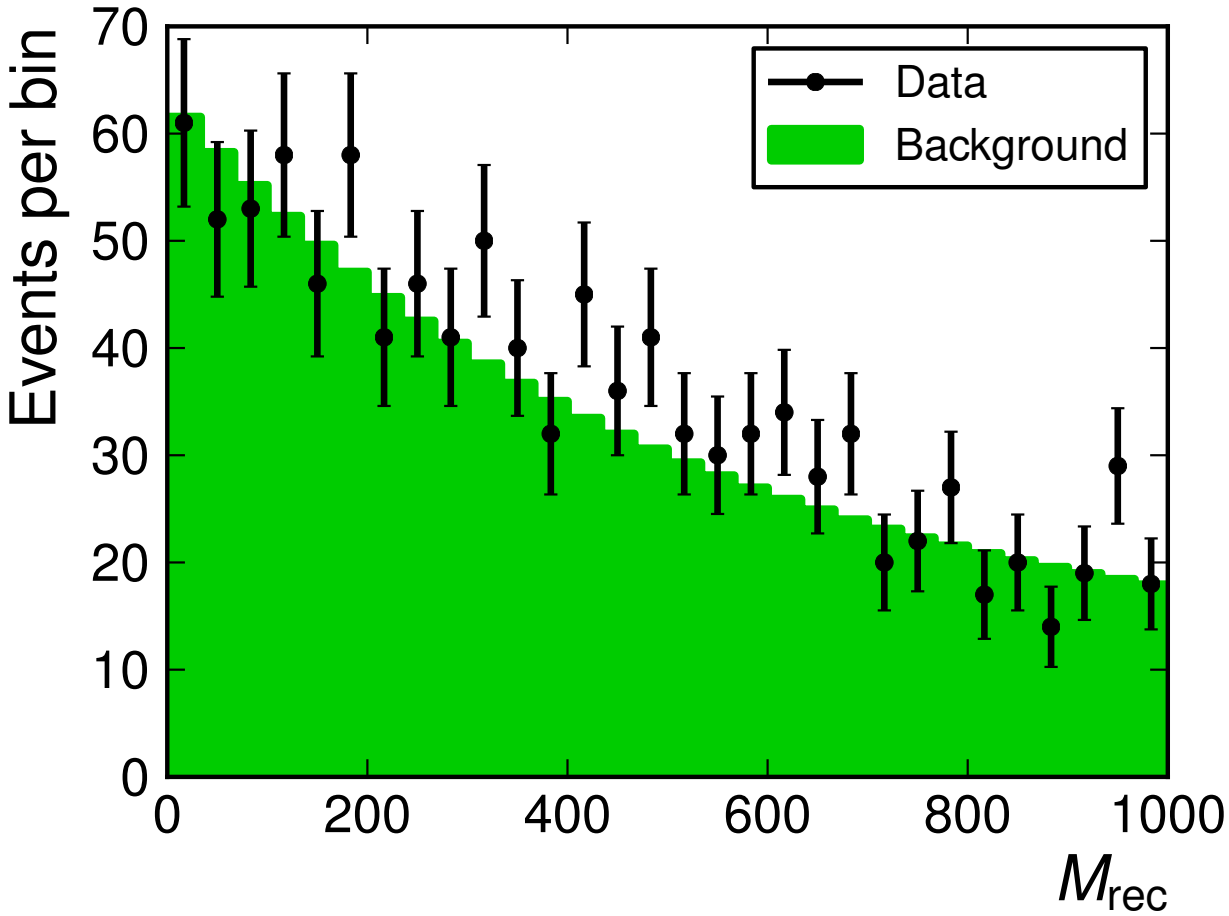
where $F_d(x)$ is the cumulative distribution for the data ("empirical distribution function"), and $F_m(x)$ for the tested model.

KS test does not require binning of the data; binning decreases $d_{\text{max}} \rightsquigarrow$ data look more compatible.

For large n: $p \approx \frac{1}{2} e^{-2nd_{\text{max}}}$.

KS Test: Example

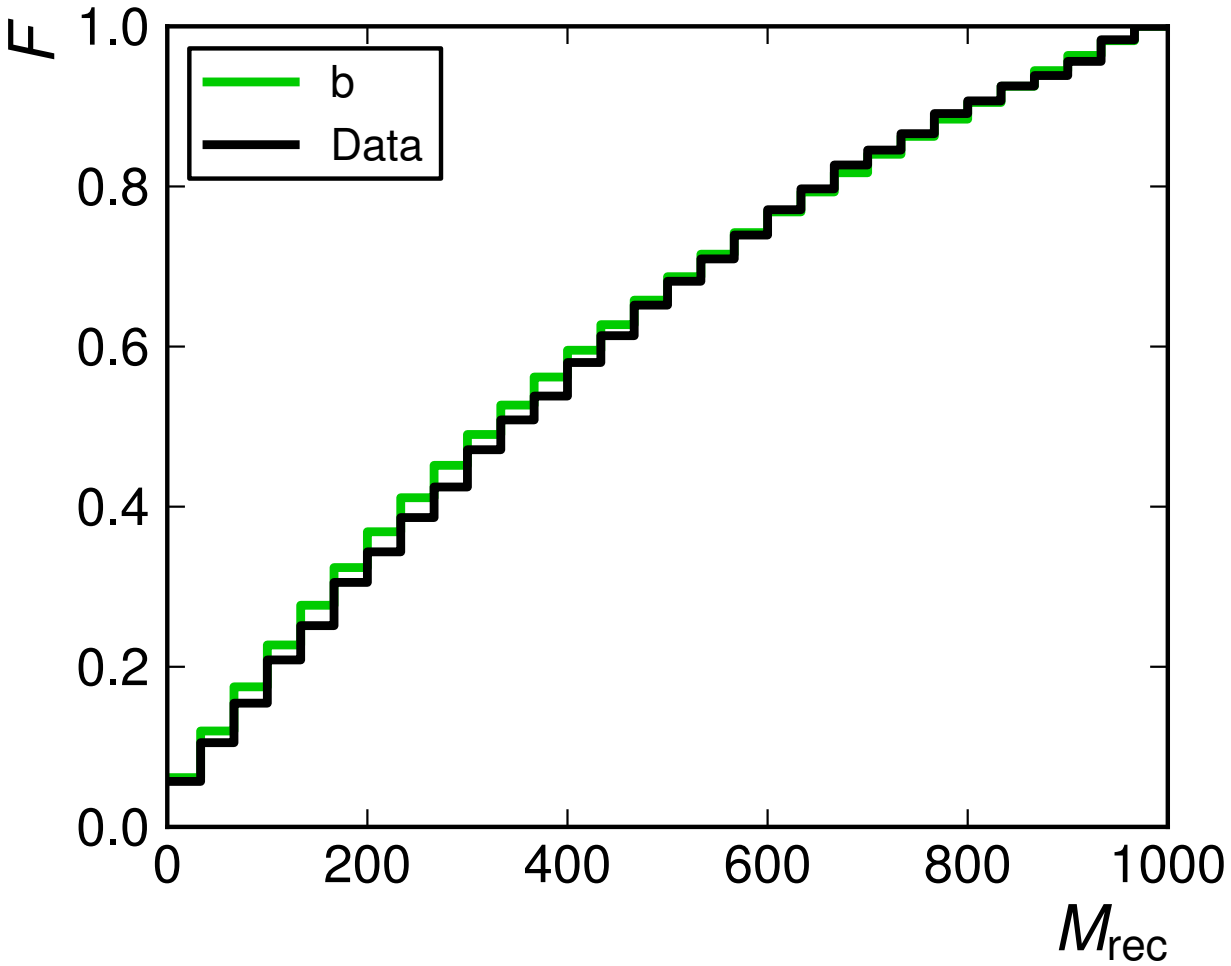
For the shape model, can determine cumulative distributions F, determine $d_{\max}$.



Using asymptotic formula: $p_{\text{KS}} = 0.10$

KS Test: Example

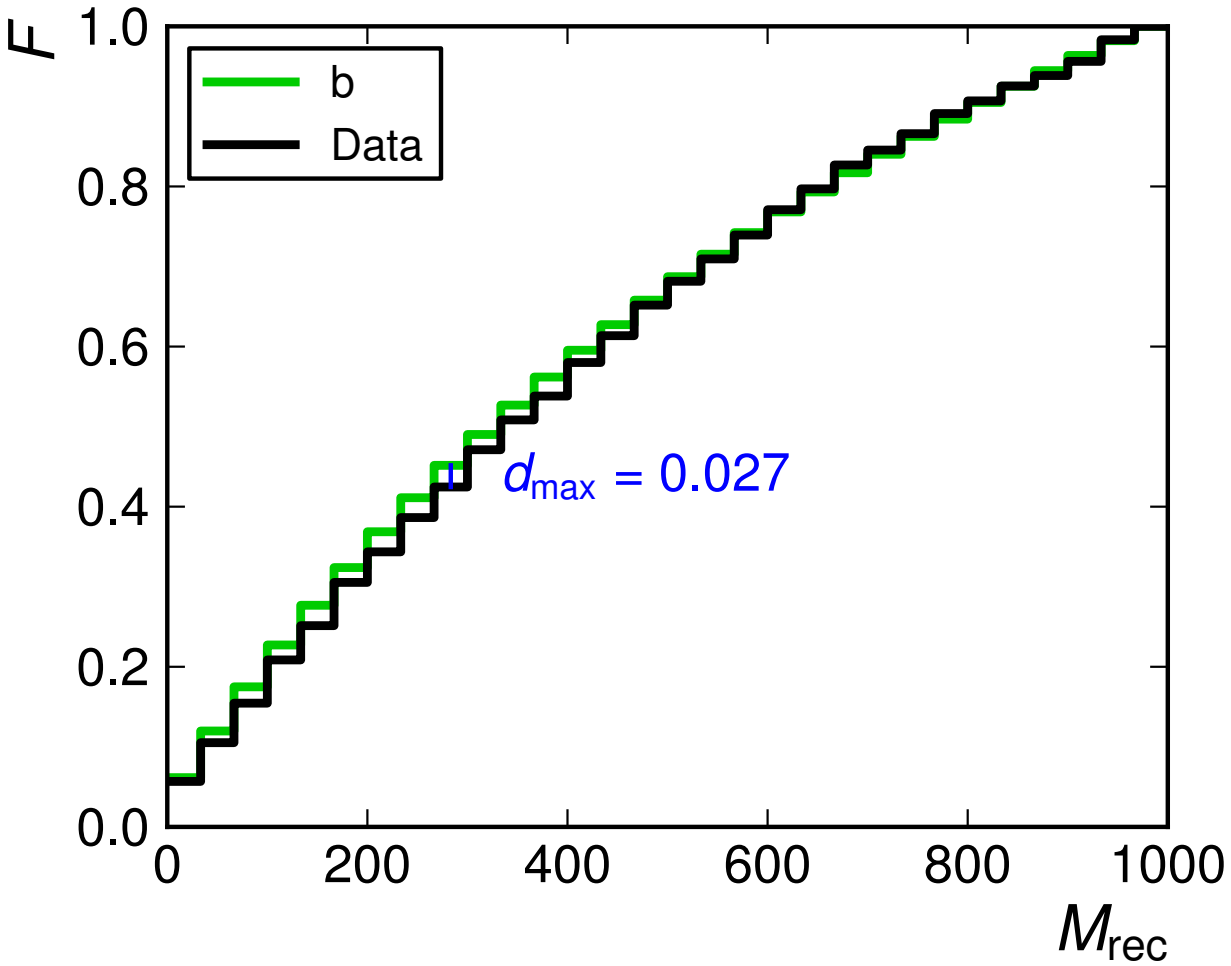
For the shape model, can determine cumulative distributions F, determine $d_{\max}$.



Using asymptotic formula: $p_{\text{KS}} = 0.10$

KS Test: Example

For the shape model, can determine cumulative distributions F, determine $d_{\max}$.



Using asymptotic formula: $p_{\text{KS}} = 0.10$

Comments

- Many more test statistic definitions possible (sensitive to different kinds of deviation): Run test, Cramér-von-Mises test, ...
- If test statistic distribution for H_0 not known, e.g.
  - not in asymptotic limit or
  - if fitting parameters to data before calculating test statistic, or
  - using bins for KS

  can (should!) determine test statistic distribution for H_0 with toy Monte-Carlo.
- Here: compared data to model. But: can use tests also to make data–data comparisons where H_0 is that both datasets follow same (unknown) parent distribution.

Beyond present Examples

I did not cover the more general case of unbinned or non-Poisson data. The main steps remain the same: Write down statistical model, introduce systematics via nuisances, construct likelihood function and test statistic, and sample toy data to derive the p-value (or use asymptotic likelihood properties).

Summary

We have seen:

- Definition of the p-value as the probability to observe such a large ("signal-like") deviation for the null hypothesis H_0; small p-values are regarded as evidence against H_0, preferring the alternative
- The role of the test statistic as measure of "signal-like"
- Monte-Carlo approach for p-value computation
- How nuisance parameters are introduced in the statistical model, how test statistic is modified, and the Bayesian ("hybrid") method to treat prior information
- Approximate formula for the Z-value for a simple counting experiment, which qualitatively shows the same properties as more complicated models

Part II

Confidence Intervals

Contents

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.

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.

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.

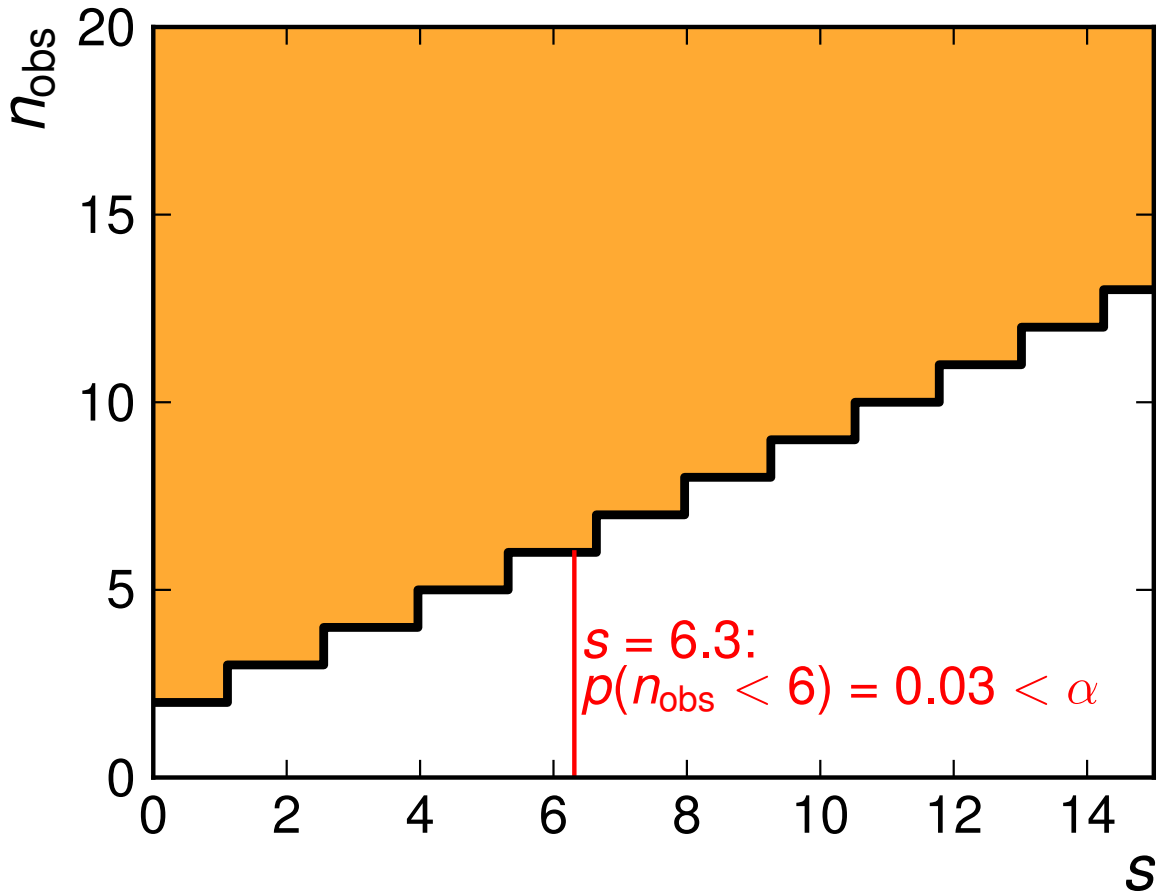
Comment

This close relation to hypothesis testing means many aspects from HT also apply here:

- 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)
- ...

Neyman Construction: Belt

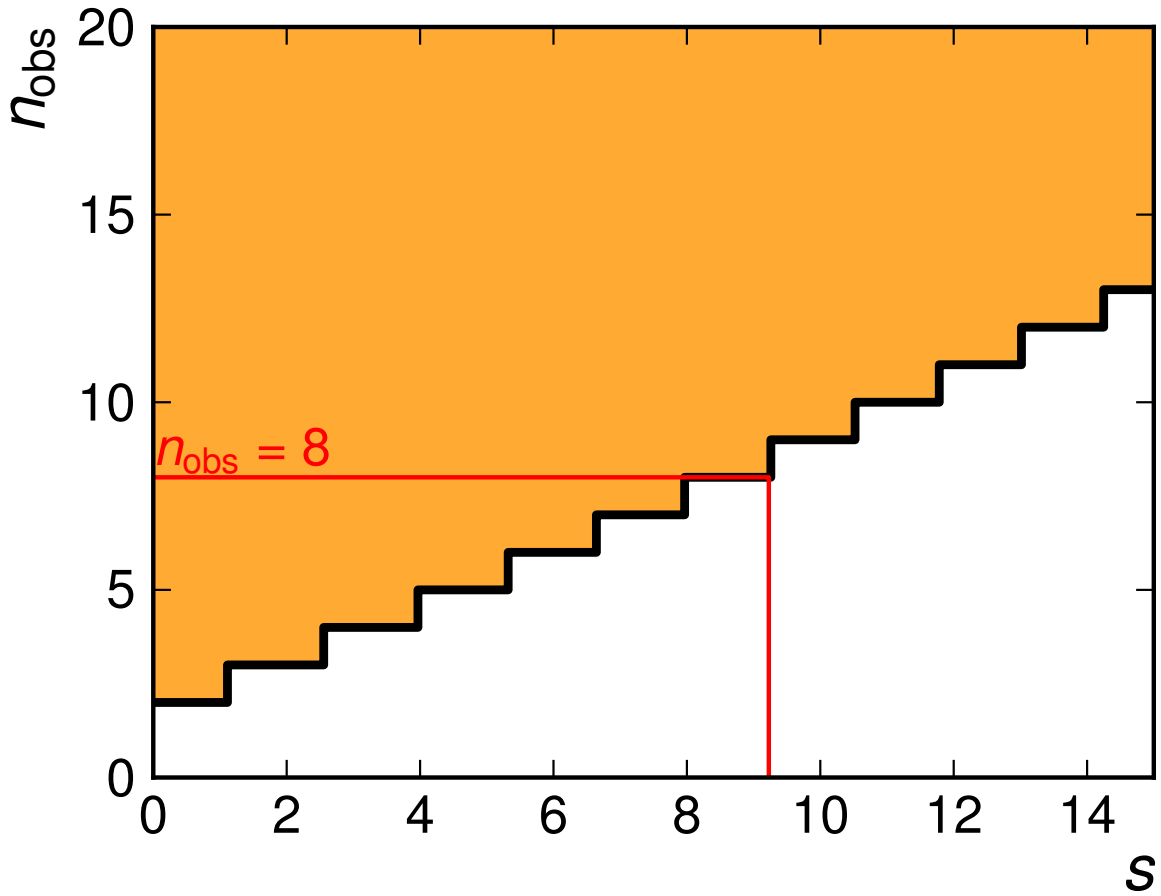
Example: Counting experiment with $b = 5.2$. As a function of s, determine n_0 for which $p(n_{\text{obs}} < n_0|s) \leq \alpha$:



Example: $n_{\text{obs}} = 8$, the 95% C.L. upper limit for s is 9.2.

Neyman Construction: Belt

Example: Counting experiment with $b = 5.2$. As a function of s, determine n_0 for which $p(n_{\text{obs}} < n_0|s) \leq \alpha$:



Example: $n_{\text{obs}} = 8$, the 95% C.L. upper limit for s is 9.2.

Neyman Construction: Coverage

Coverage for $\mu = \mu_0$:

- The probability to observe a number of events included in the belt is $[\geq](1 - \alpha) = 0.95$, by construction of the belt.
- In exactly those cases, the resulting interval will include μ_0.

$\rightsquigarrow$ The probability that the interval includes μ_0 (coverage) is $(1 - \alpha)$, as desired.

Exact coverage is not always possible due to discreteness. In this case, one usually chooses to be conservative, i.e. to over-cover.

The coverage can be determined with toy Monte-Carlo for any method by counting the fraction of toys for which the interval contains the true value.

Neyman Construction: Coverage

Coverage for $\mu = \mu_0$:

- The probability to observe a number of events included in the belt is $[\geq](1 - \alpha) = 0.95$, by construction of the belt.
- In exactly those cases, the resulting interval will include μ_0.

$\rightsquigarrow$ The probability that the interval includes μ_0 (coverage) is $(1 - \alpha)$, as desired.

Exact coverage is not always possible due to discreteness. In this case, one usually chooses to be conservative, i.e. to over-cover.

The coverage can be determined with toy Monte-Carlo for any method by counting the fraction of toys for which the interval contains the true value.

Neyman Construction: Coverage

Coverage for $\mu = \mu_0$:

- The probability to observe a number of events included in the belt is $[\geq](1 - \alpha) = 0.95$, by construction of the belt.
- In exactly those cases, the resulting interval will include μ_0.

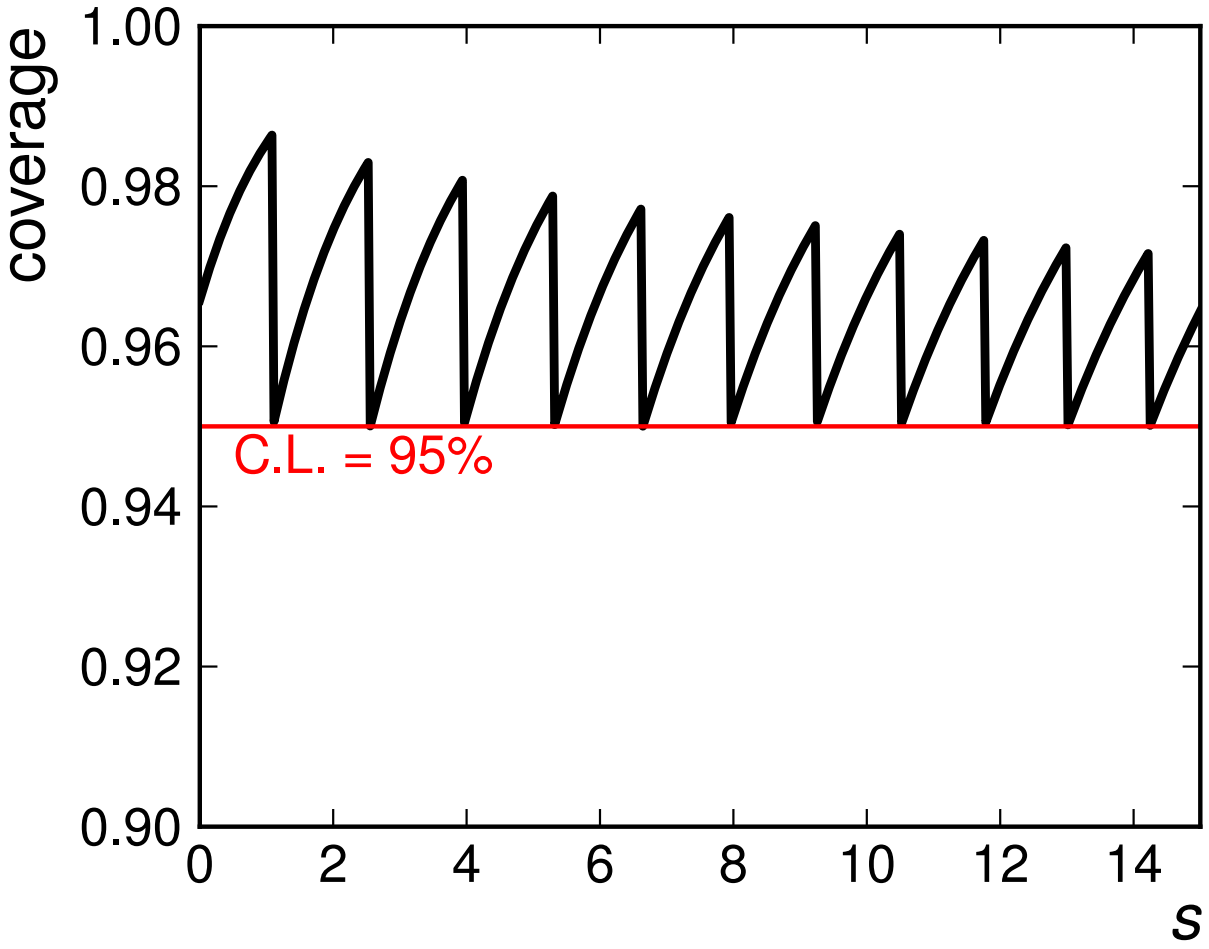
$\rightsquigarrow$ The probability that the interval includes μ_0 (coverage) is $(1 - \alpha)$, as desired.

Exact coverage is not always possible due to discreteness. In this case, one usually chooses to be conservative, i.e. to over-cover.

The coverage can be determined with toy Monte-Carlo for any method by counting the fraction of toys for which the interval contains the true value.

Coverage Example

For the counting experiment example, the coverage as a function of s is:



This is the "belt size" as a function of s.

Now: Generalize construction to shape models.

Test Statistic

As for HT, introduce test statistic as a measure for incompatibility with null hypothesis that signal strength $\mu = \mu_0$ versus alternative $\mu < \mu_0$:

$$t_{\mu_0} = \log \frac{\max_{\theta \in H_1} L(\theta|d)}{\max_{\theta \in H_0} L(\theta|d)}.$$

As for t, large values incompatibility with H_0 (which is now different for each μ_0!).

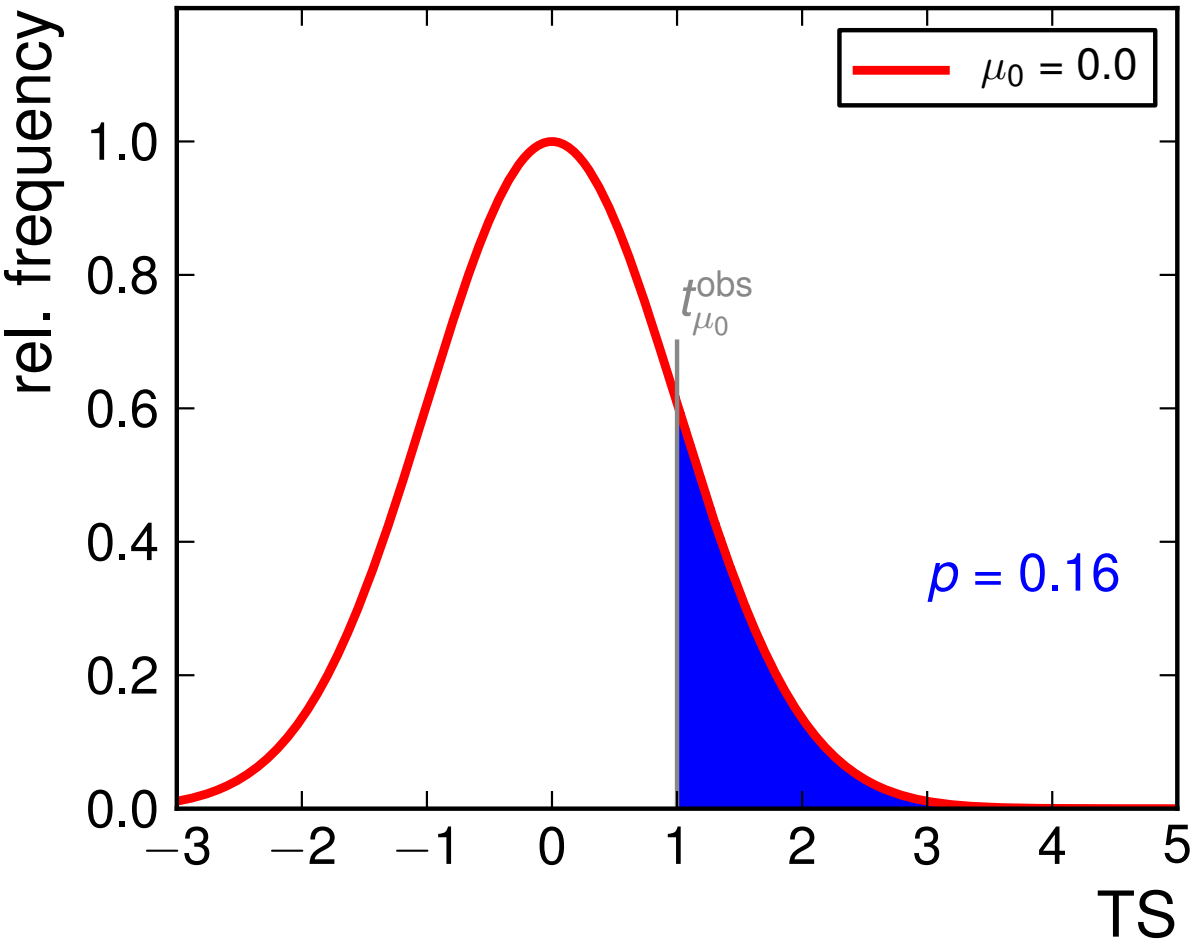
$\rightsquigarrow$ MC description for Neyman construction: For a fixed μ_0

1. generate toy data and calculate test statistic $t_{\mu_0} \rightsquigarrow$ test statistic distribution
2. calculate test statistic value $t_{\mu_0}^{\text{obs}}$ for data; the p-value is the fraction of toys with $t_{\mu_0} \geq t_{\mu_0}^{\text{obs}}$

Repeat for many μ_0; interval is given by μ_0 for which $p > \alpha$.

Neyman Construction: MC method

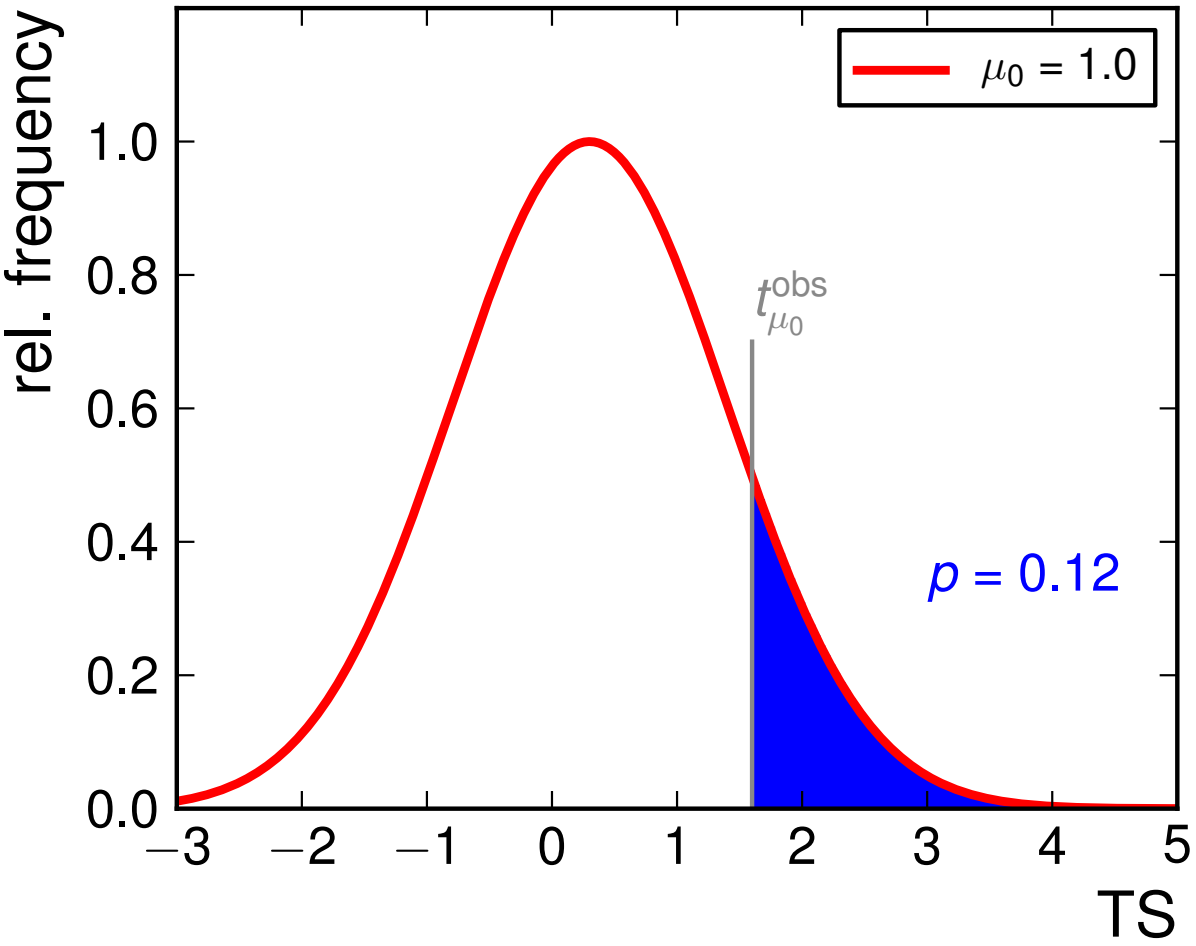
Sketch of test statistic distributions for different values for μ_0:



Frequentist 95% C.L. upper limit: $\mu = 2$.

Neyman Construction: MC method

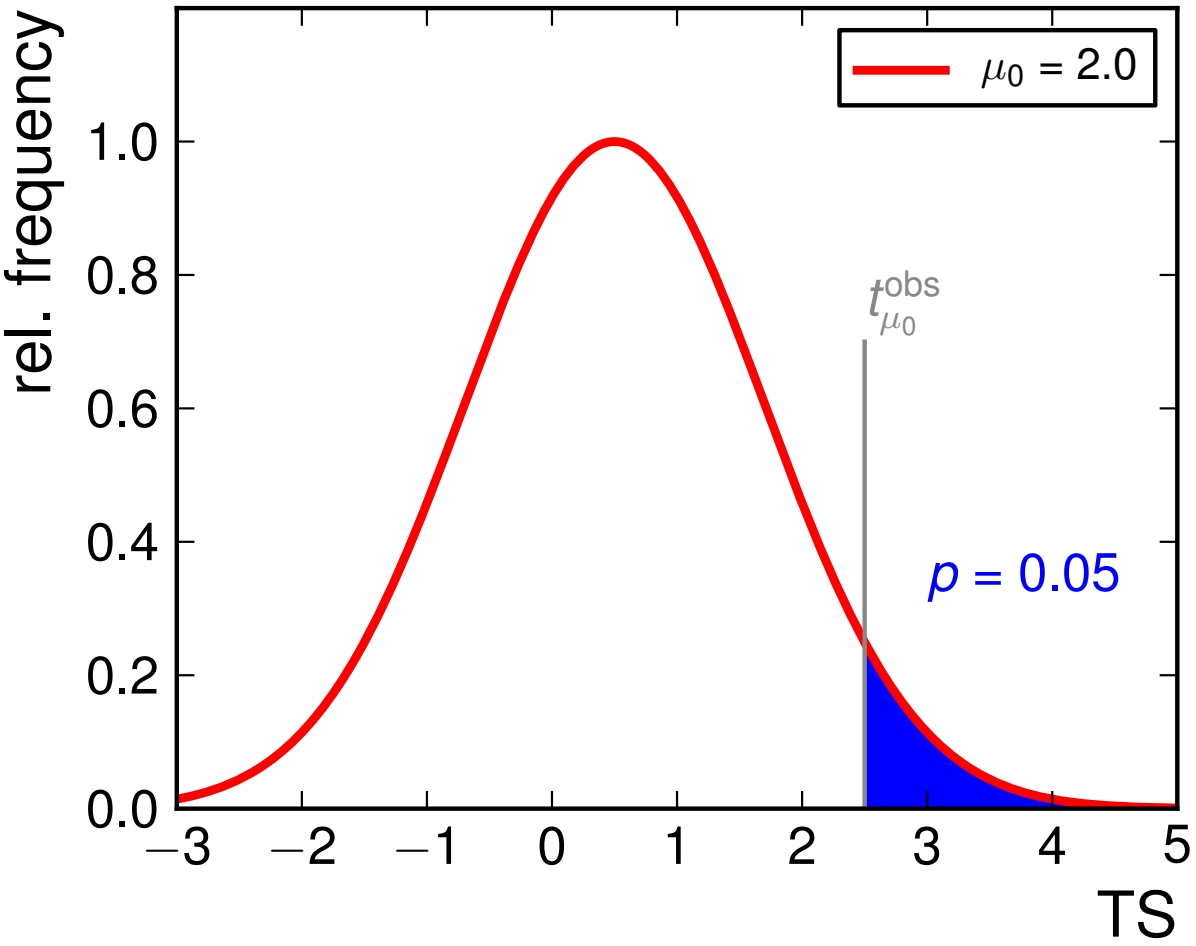
Sketch of test statistic distributions for different values for μ_0:



Frequentist 95% C.L. upper limit: $\mu = 2$.

Neyman Construction: MC method

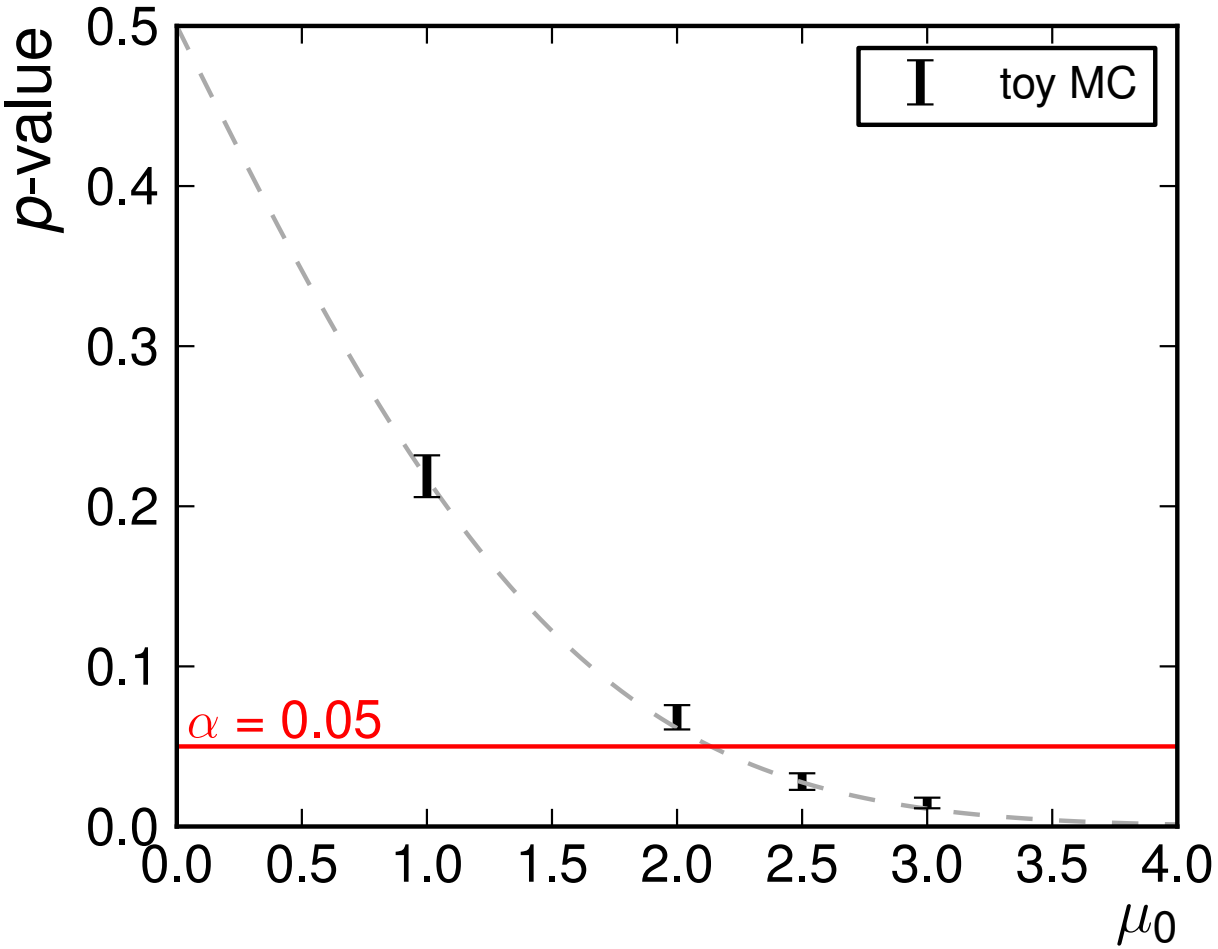
Sketch of test statistic distributions for different values for μ_0:



Frequentist 95% C.L. upper limit: $\mu = 2$.

Neyman Construction: p vs. μ_0

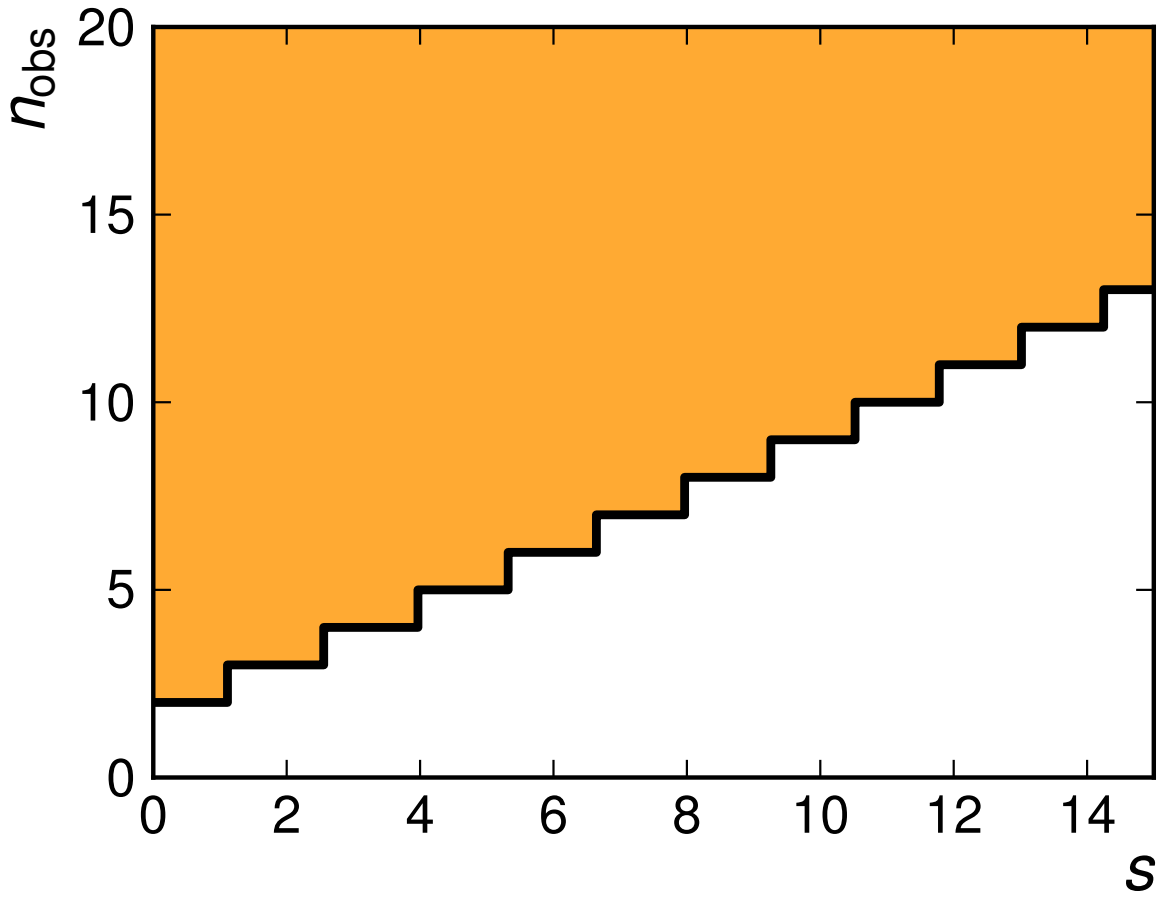
Plot the p-value vs. μ_0 for the hypothesis test $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu < \mu_0$:



The 95% C.L. upper limit is the value μ_0 at which $p = 0.05$.
For MC method, one would "scan through" μ_0, make toys at each point to get the value of μ_0 for which $p = 0.05$.
(See exercise 2).

Empty Intervals: Example

In the Neyman construction, it can happen that one cites empty intervals, e.g. for $n_{\text{obs}} = 1$, one would state $s < 0$ at 95% C.L.:



For a correct-coverage method and true $\mu = 0$, this happens in 5% of the cases, if n_{obs} happens to be small.

Empty Intervals: Discussion

Empty (or very small) intervals are unsatisfactory:

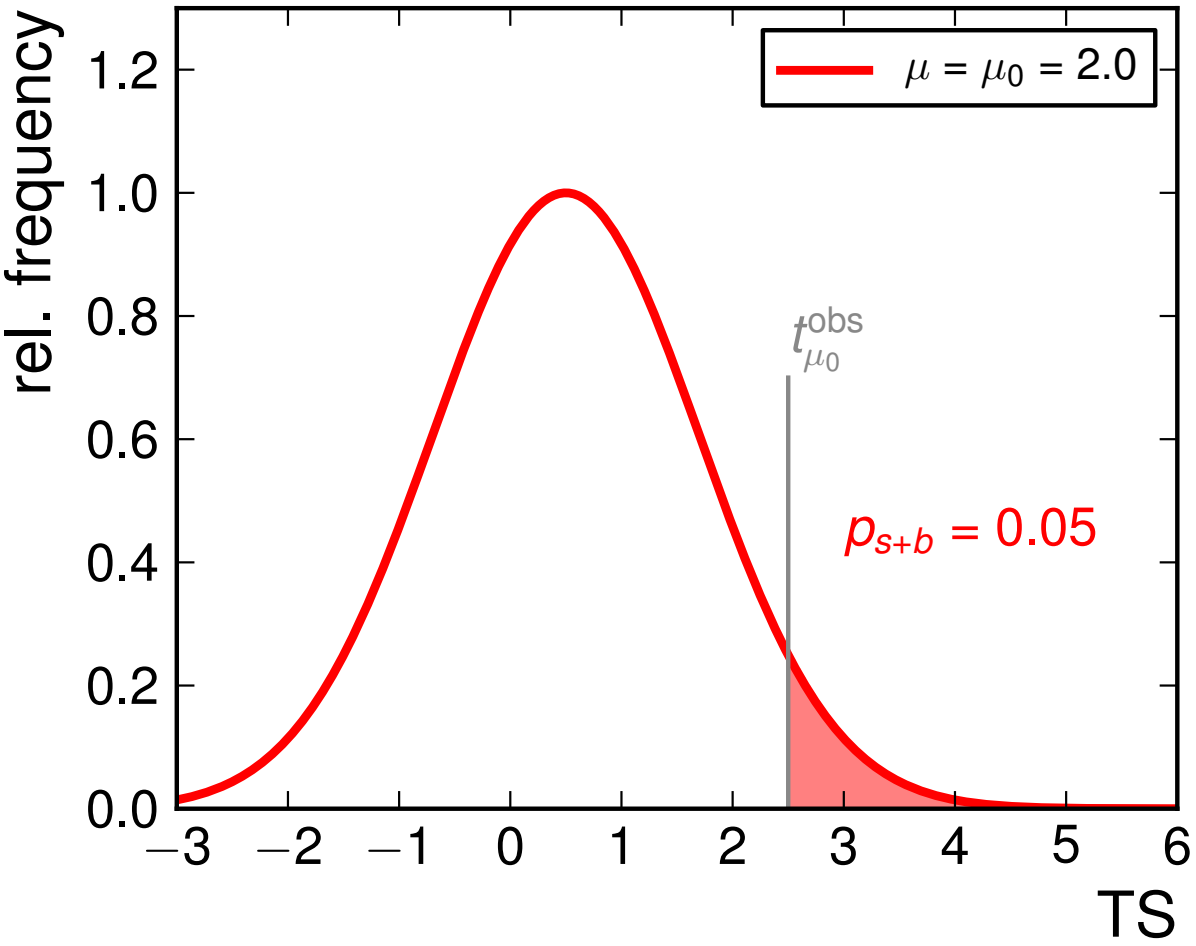
- We know we are in the "5% type I error" case.
- We would cite a very strong limit although there is no experimental sensitivity for such small values.

To avoid citing such intervals, one can modify the frequentist construction
$\rightsquigarrow$ "modified frequentist intervals" also known as the "CLs method".

A closer look

Small/empty intervals happen in case of incompatibility with background-only model (e.g. very few events even for background-only).

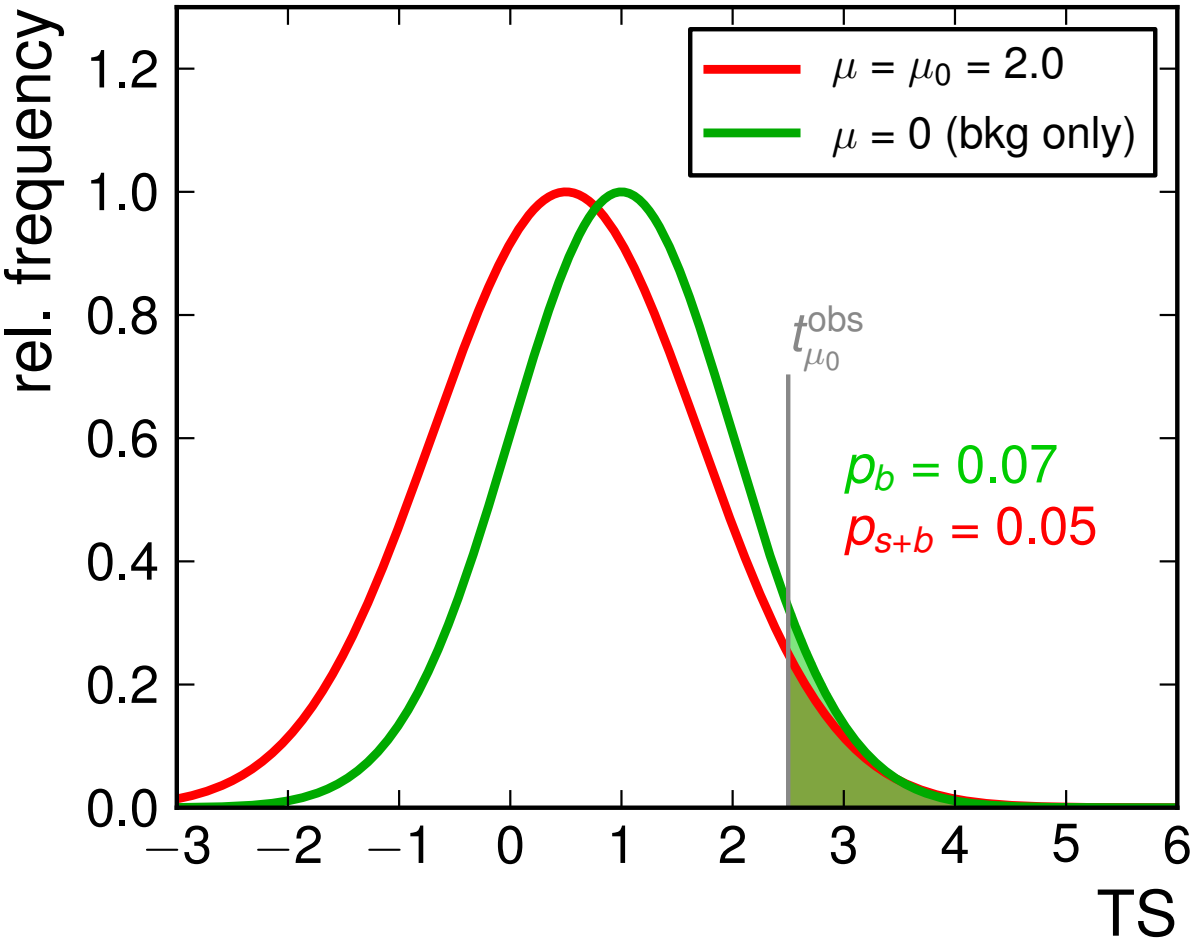
$\rightsquigarrow$ Also look at test statistic distribution for background-only model $\mu = 0$.



Idea: increase limit if data is incompatible with background-only hypothesis $\mu = 0$. $\rightsquigarrow$ increase interval in case of small values for p_b.

A closer look

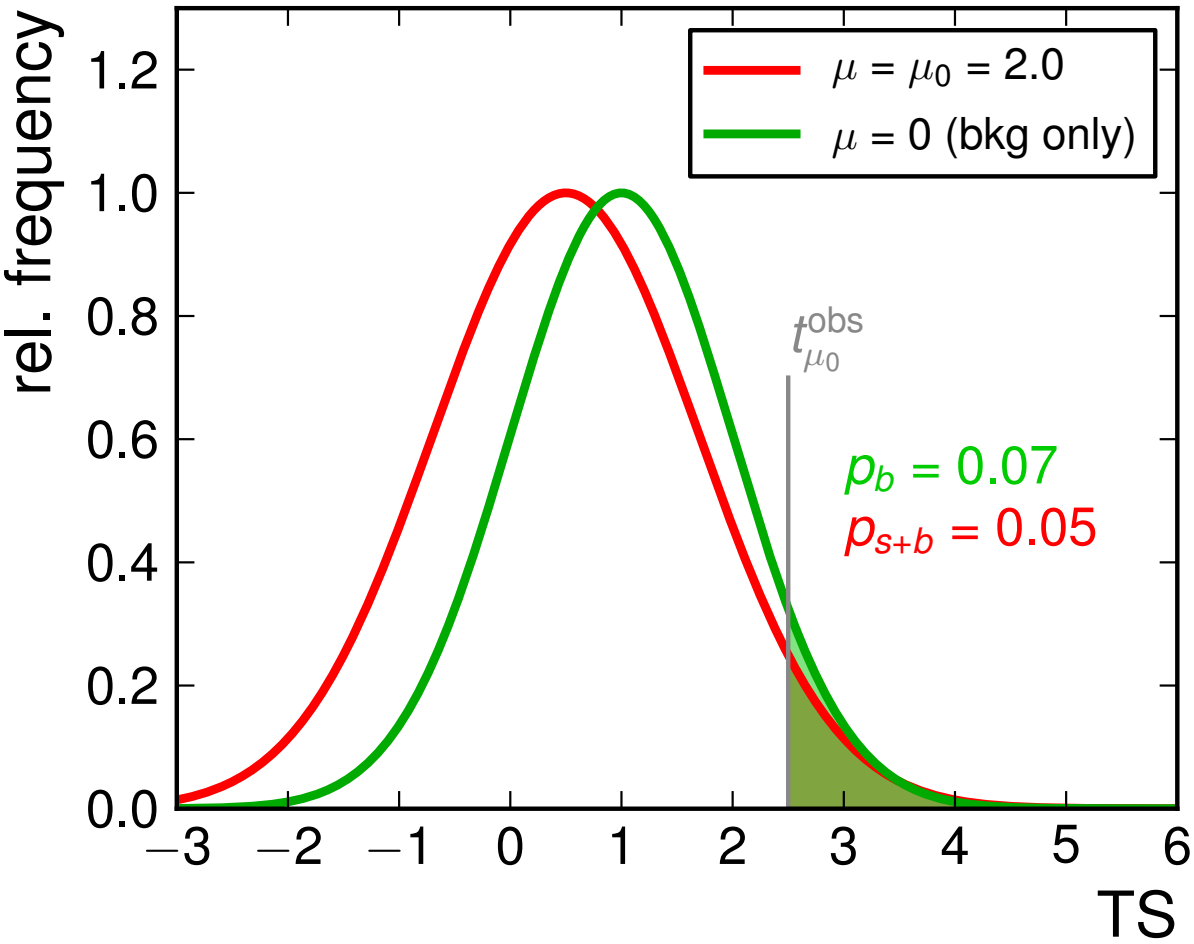
Small/empty intervals happen in case of incompatibility with background-only model (e.g. very few events even for background-only).
$\rightsquigarrow$ Also look at test statistic distribution for background-only model $\mu = 0$.



Idea: increase limit if data is incompatible with background-only hypothesis $\mu = 0$. $\rightsquigarrow$ increase interval in case of small values for p_b.

A closer look

Small/empty intervals happen in case of incompatibility with background-only model (e.g. very few events even for background-only).
$\rightsquigarrow$ Also look at test statistic distribution for background-only model $\mu = 0$.



Idea: increase limit if data is incompatible with background-only hypothesis $\mu = 0$. $\rightsquigarrow$ increase interval in case of small values for p_b.

CLs definition

The CL_s-value is a modified p-value such that it is large for small p_b

$$CL_s := \frac{p_{s+b}}{p_b} \left(= \frac{CL_{s+b}}{CL_b} \right).$$

[Note that in the literature, p_b is often defined differently and the denominator then is $(1 - p_b)$.]

In the limit construction, use CL_s in place of p_{s+b} as before: Find limit is μ for which $CL_s = \alpha$.

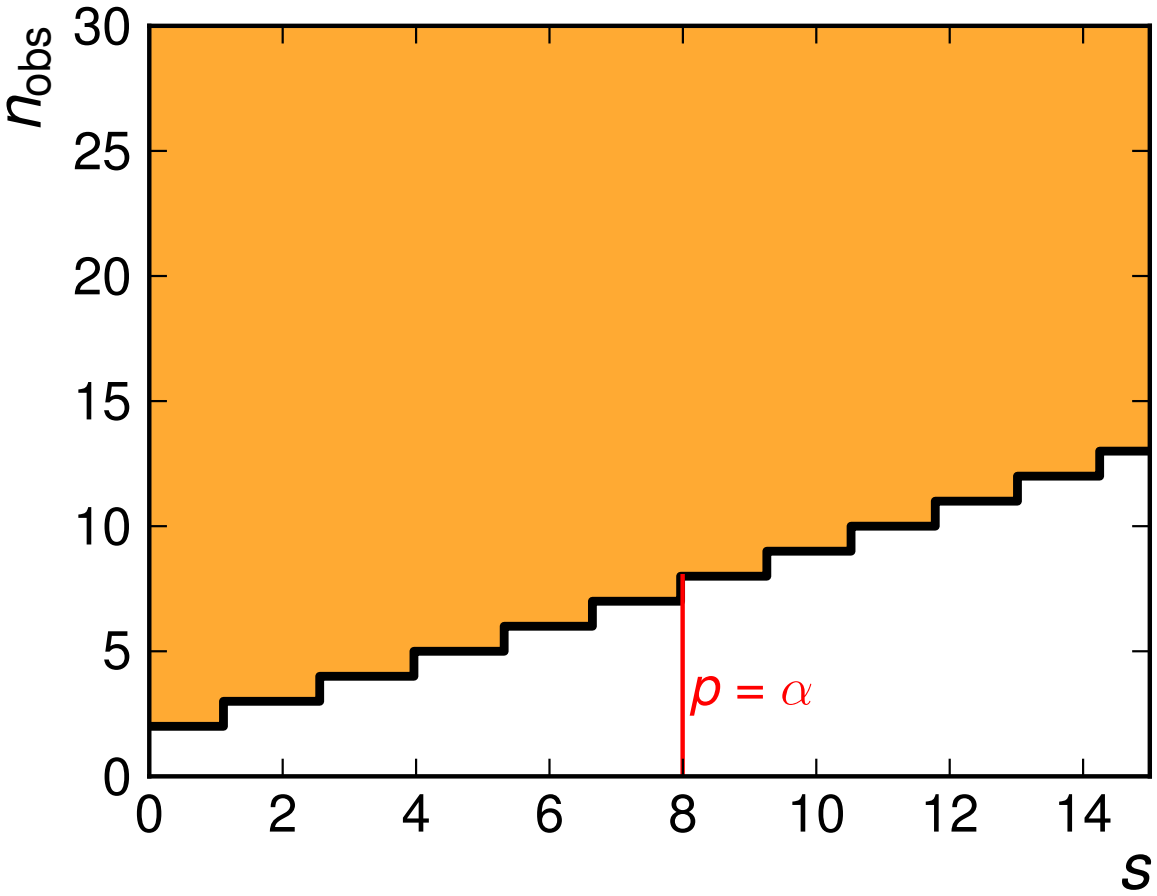
Notes:

- $CL_s \geq p_{s+b}$ by construction $\rightsquigarrow$ CL_s limits are always more conservative than "purely frequentist" (Neyman) limits
- The CL_s method prevents citing limits with no experimental sensitivity

Neyman Construction

So far: For each s, include upper $(1 - \alpha) = 95\%$ of n in belt $\rightsquigarrow$ upper limits.

Now: include central 95% $\rightsquigarrow$ "central" intervals.

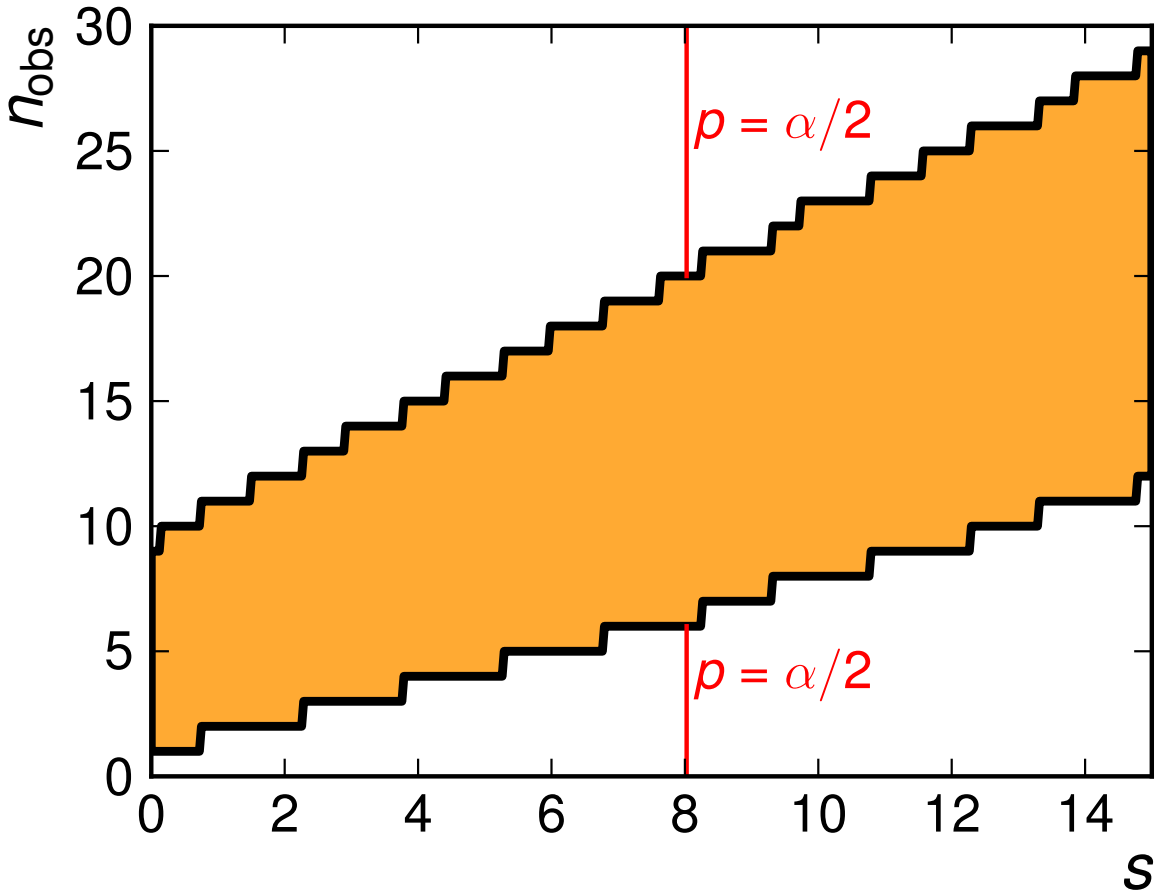


But: Ordering is somewhat arbitrary; central intervals only one choice.

Neyman Construction

So far: For each s, include upper $(1 - \alpha) = 95\%$ of n in belt $\rightsquigarrow$ upper limits.

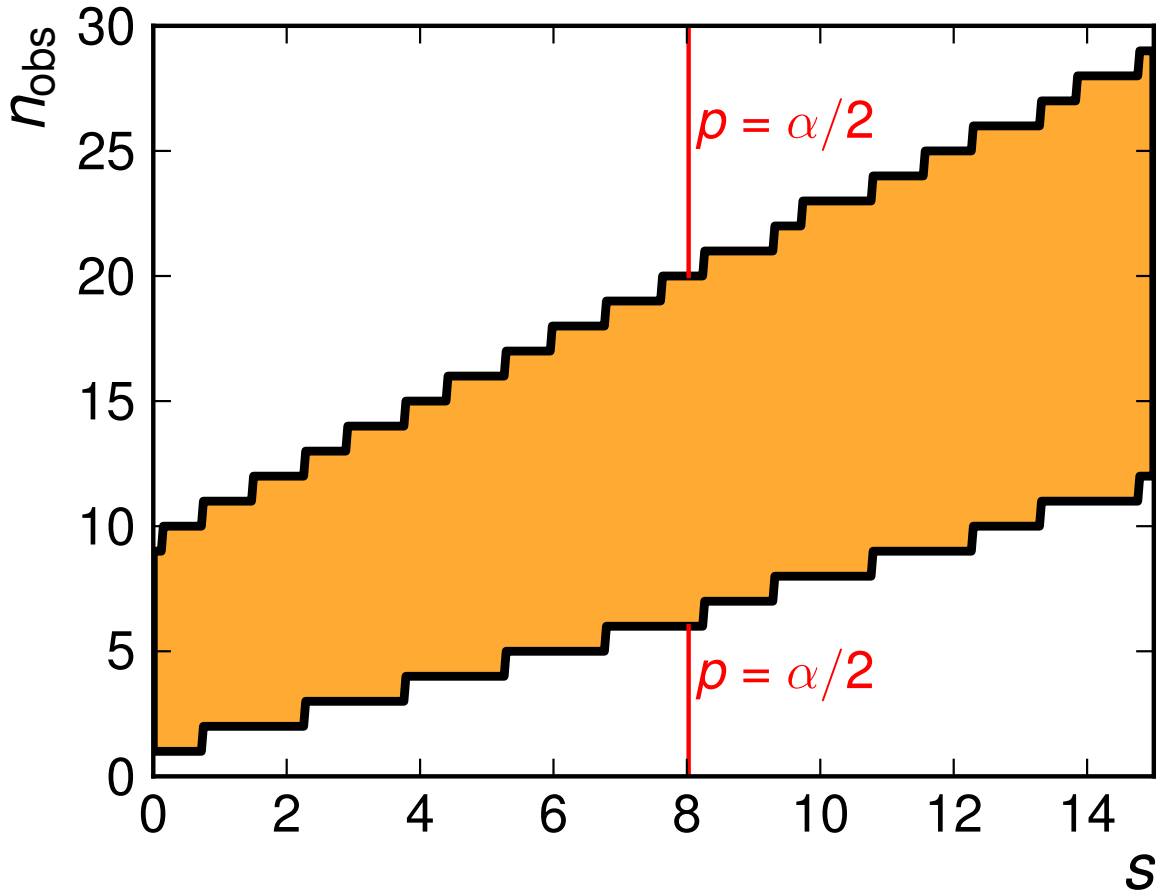
Now: include central 95% $\rightsquigarrow$ "central" intervals.



But: Ordering is somewhat arbitrary; central intervals only one choice.

Neyman Construction

So far: For each s, include upper $(1 - \alpha) = 95\%$ of n in belt $\rightsquigarrow$ upper limits.
Now: include central 95% $\rightsquigarrow$ "central" intervals.



But: Ordering is somewhat arbitrary; central intervals only one choice.

Ordering Rule; Feldman-Cousins

Previous slide: central Neyman intervals.

But: Have to decide beforehand whether we want intervals or limits.

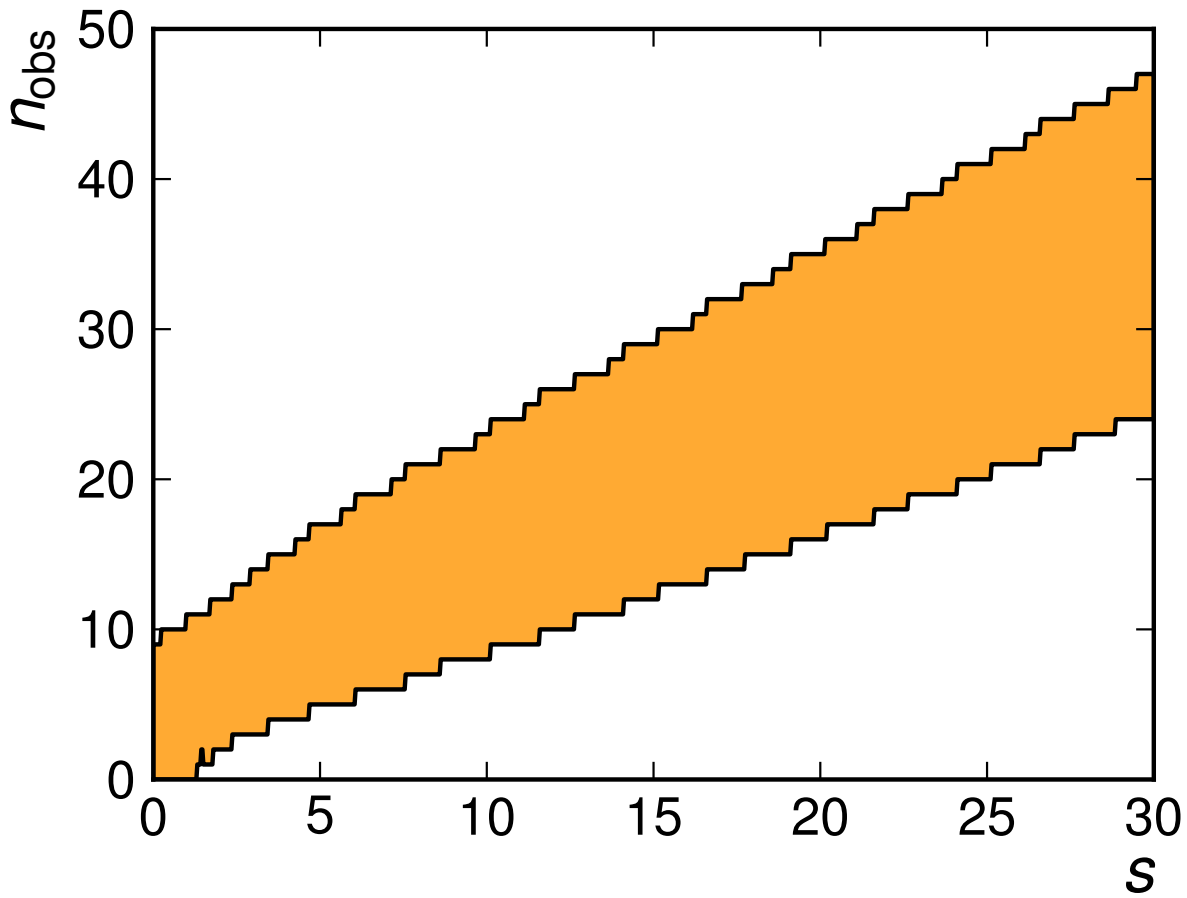
Feldman and Cousins proposed to include those values of n in the belt in decreasing order of the likelihood ratio test statistic:

$$S = \frac{L(s_0|n)}{\max_s L(s|n)},$$

i.e. include those points first where $s = s_0$ is a "good fit" to data, compared to other values for s.
As usual, include points in the belt until reaching a probability of $(1 - \alpha)$ for each value of s_0.

Feldman-Cousins: Example

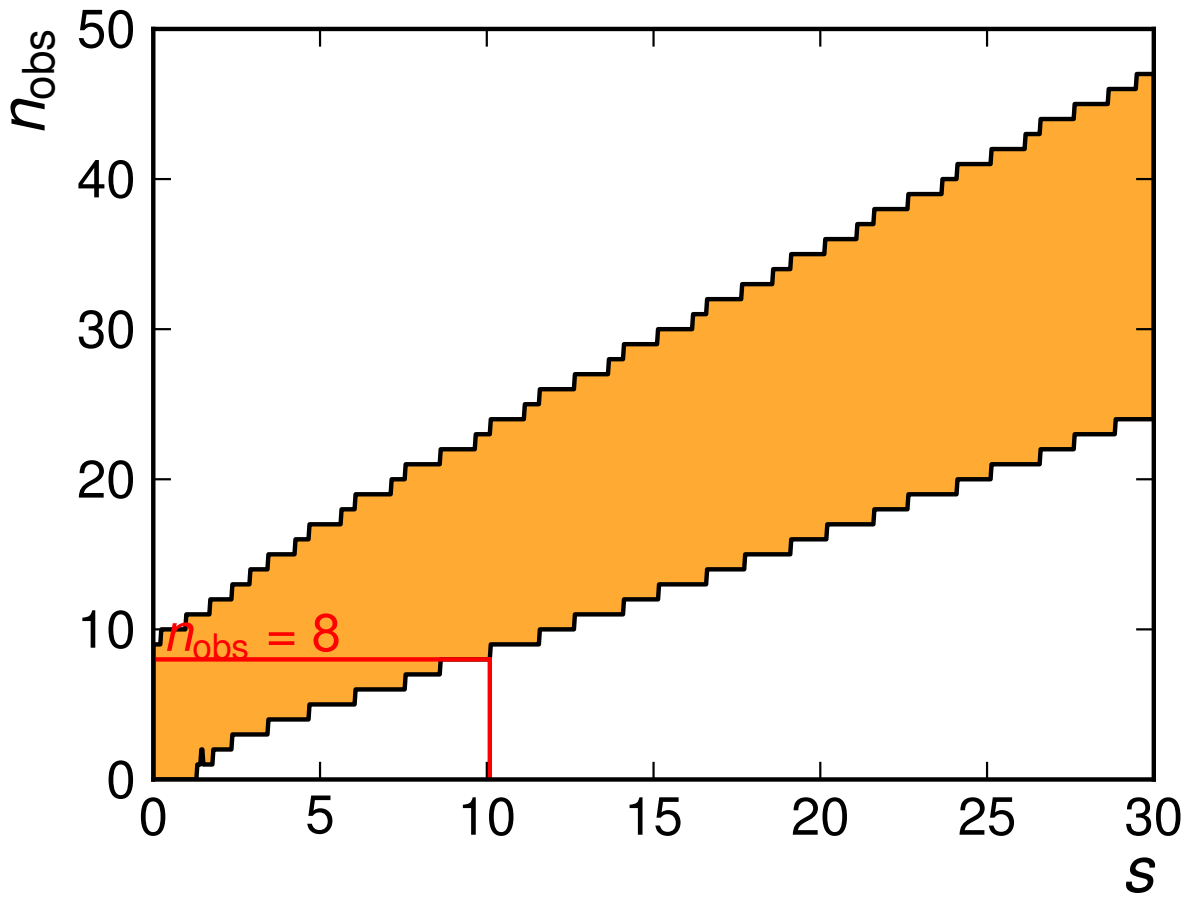
For counting experiment with $b = 5.2$, the Feldman-Cousins band looks like this:



Automatic transition from "limit-like" intervals to two-sided intervals.

Feldman-Cousins: Example

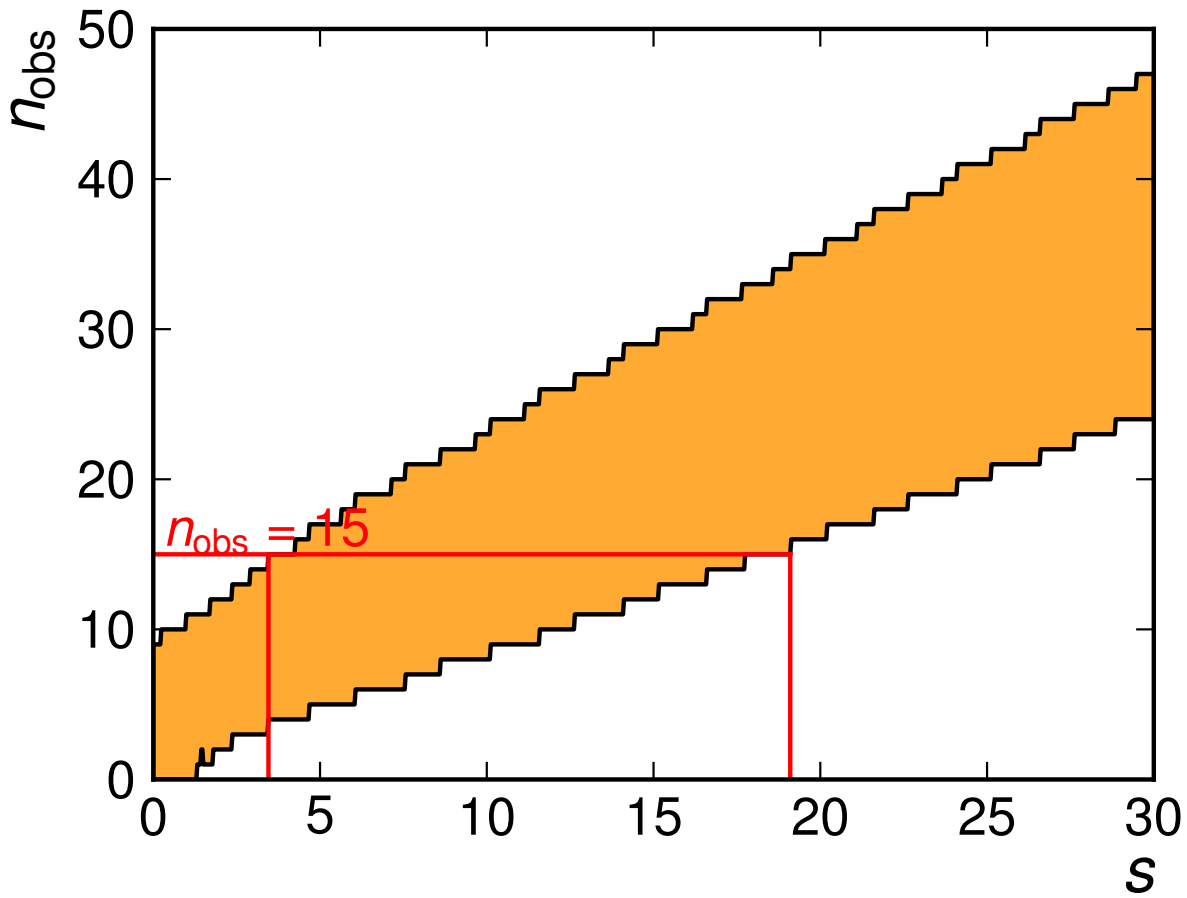
For counting experiment with $b = 5.2$, the Feldman-Cousins band looks like this:



Automatic transition from "limit-like" intervals to two-sided intervals.

Feldman-Cousins: Example

For counting experiment with $b = 5.2$, the Feldman-Cousins band looks like this:



Automatic transition from "limit-like" intervals to two-sided intervals.

Feldman-Cousins: Comments

- Suitable for calculating intervals for parameters with physical limits (e.g. cross sections, ratios, masses ...).
- Does not require to decide beforehand whether to cite a limit or a two-sided interval ("unified" construction).
- Empty/small intervals are also avoided to some degree $\rightsquigarrow$ alternative to CLs
- Generalization: use profile likelihood ratio test statistic $\tilde{t}_{\mu_0}$ with $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ as ordering criterion.

Reference: Feldman, Cousins Phys.Rev.D57:3873-3889,1998; arXiv: 9711021.

Introduction

In contrast to frequentist statistics, Bayesian statistics allows probabilistic statements about model parameters.

Bayes' Theorem gives a formula for the posterior of the parameters θ, given data d:

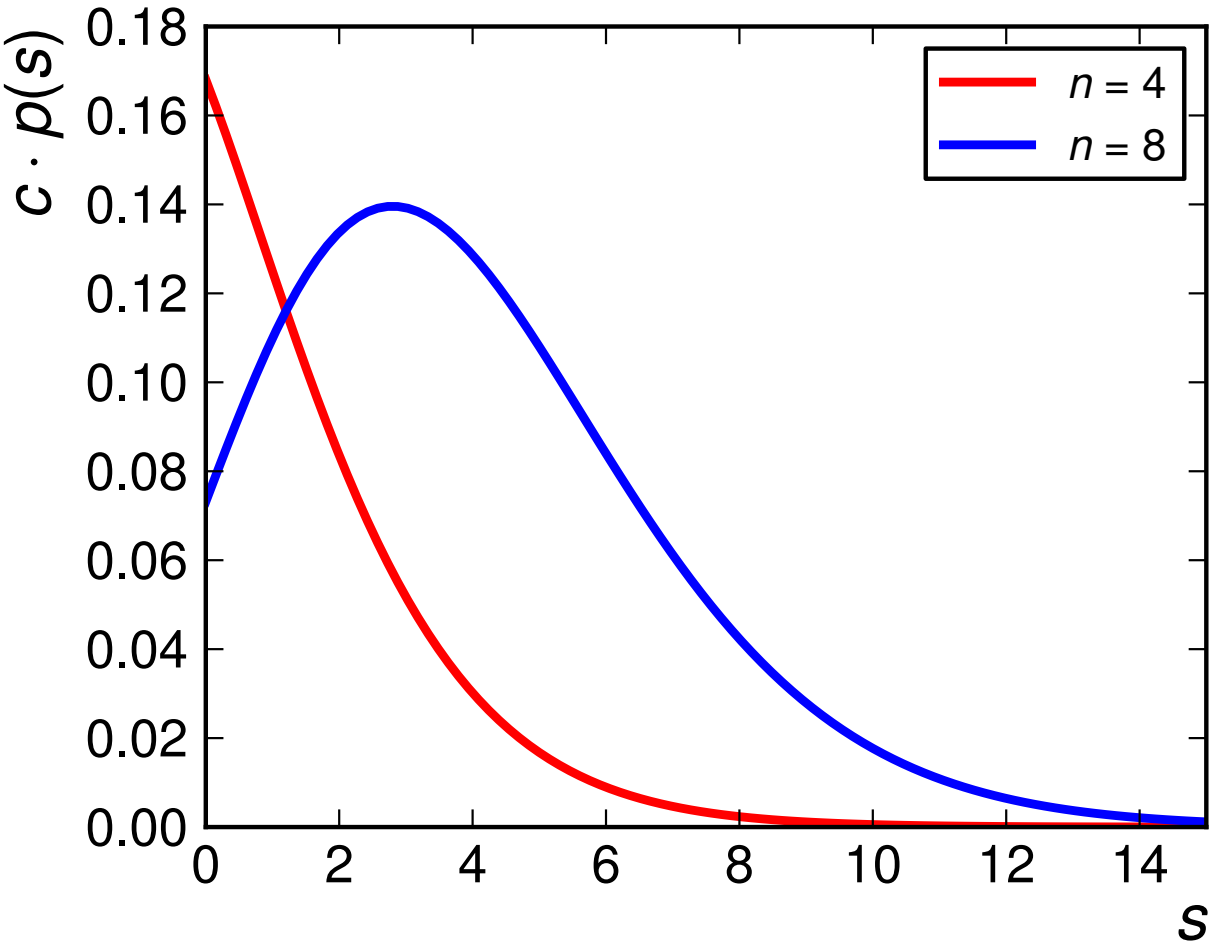
$$p(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{\pi(d)}$$

where

- $\pi(\theta)$ is the prior for the parameters θ
- $p(d|\theta)$ is the probability to observed data d, given θ – this is the statistical model
- $\pi(d)$ is the prior probability to observe data d (for fixed d, this is just a number making the left side a properly normalized probability; we usually do not need it.)

Example

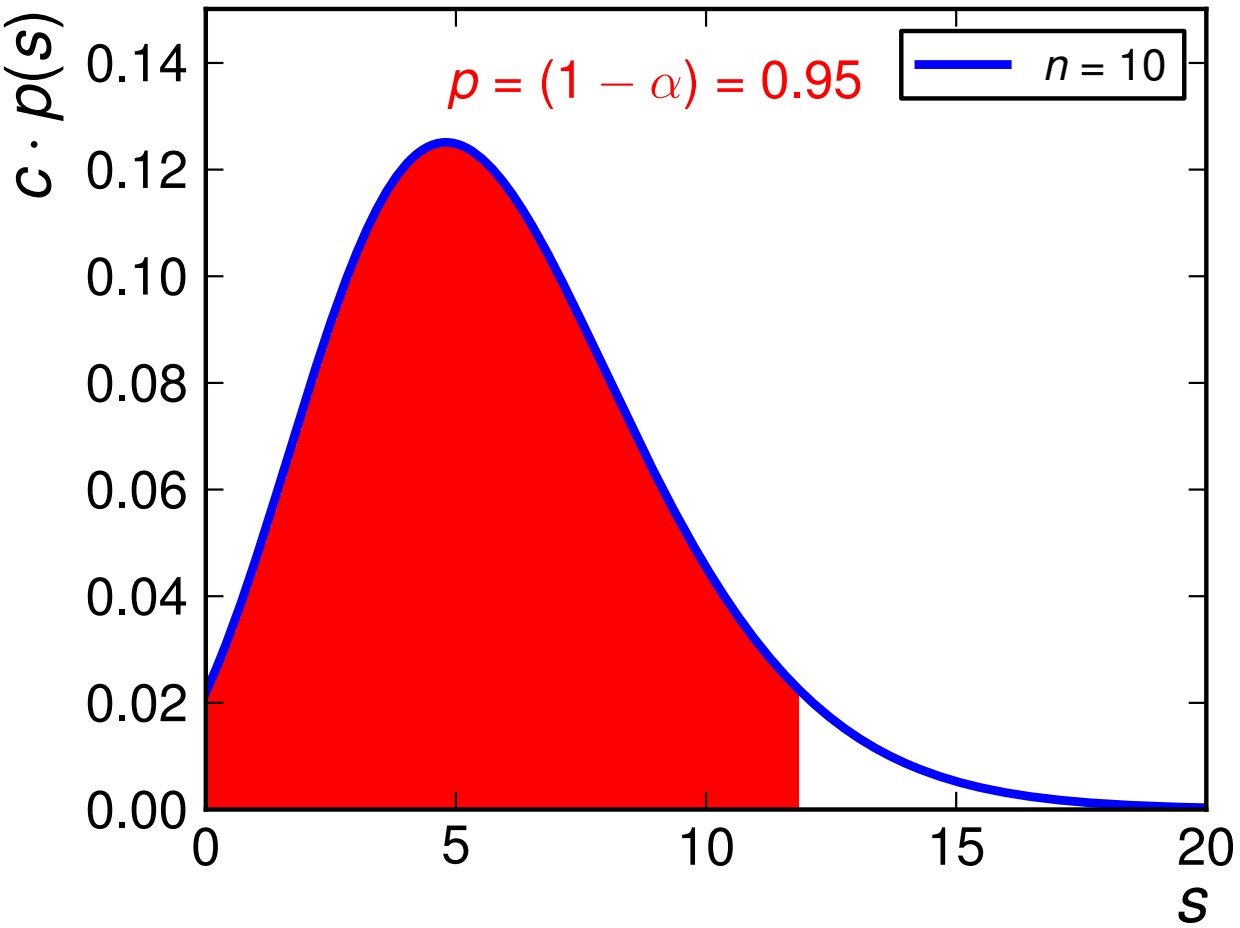
For the counting experiment with known $b = 5.2$ and a flat prior for the signal s, the posterior is (apart from normalization) Poisson$(n|s+b)$, but read as a probability in s, not in n(!)



As might be expected intuitively, maximum posterior at $s = 0$ for $n < b$ and at $n - b$ for $n > b$.

Intervals

To get confidence intervals from posterior, choose range of s s.t. posterior probability coincides with confidence level $(1 - \alpha)$ [cf. Luc's talk].



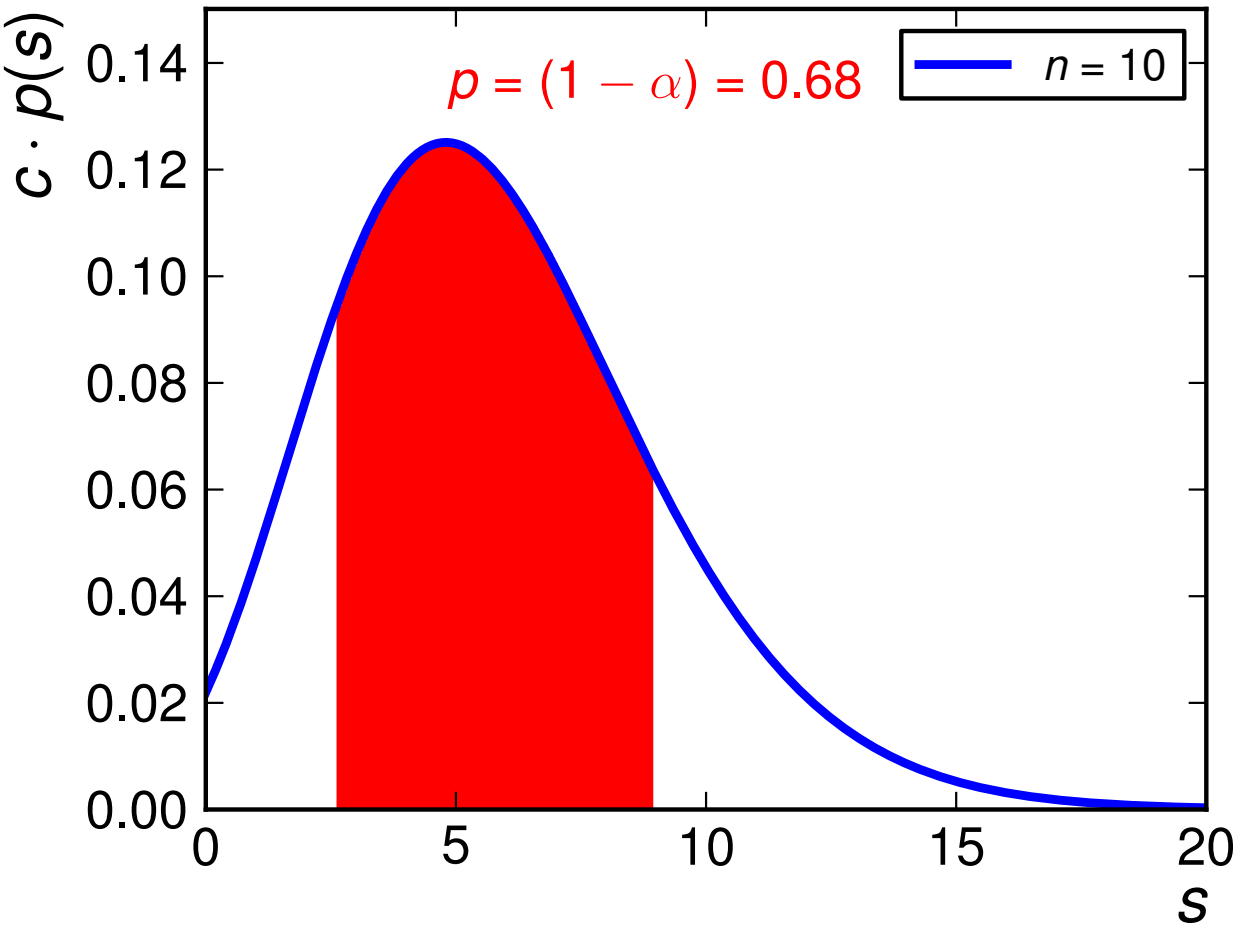
Include lower 95% $\rightsquigarrow$ 95% C.L. upper limit.

Central 68%

Maximum posterior $\rightsquigarrow$ shortest interval

Intervals

To get confidence intervals from posterior, choose range of s s.t. posterior probability coincides with confidence level $(1 - \alpha)$ [cf. Luc's talk].



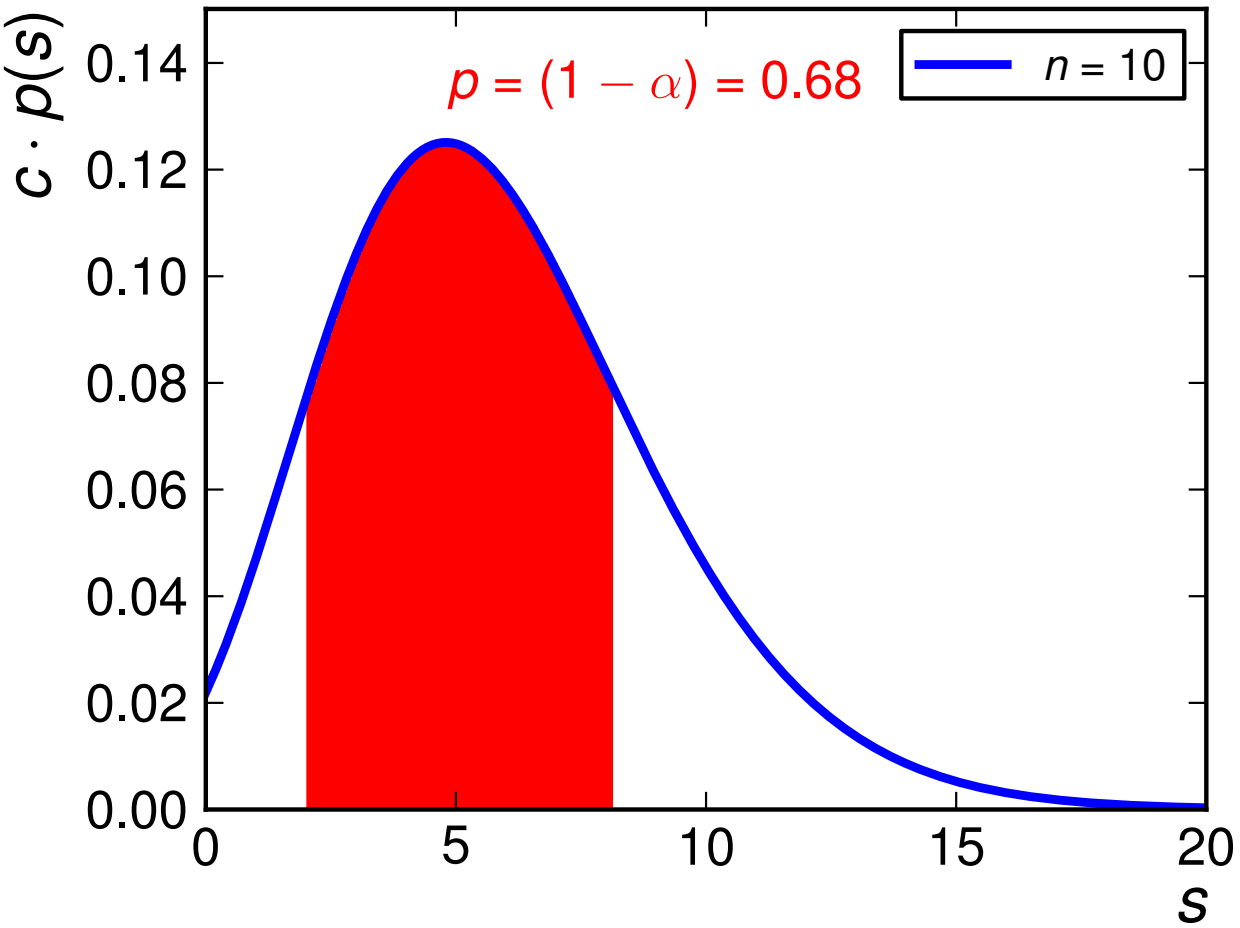
Include lower 95% $\rightsquigarrow$ 95% C.L. upper limit.

Central 68%

Maximum posterior $\rightsquigarrow$ shortest interval

Intervals

To get confidence intervals from posterior, choose range of s s.t. posterior probability coincides with confidence level $(1 - \alpha)$ [cf. Luc's talk].



Include lower 95% $\rightsquigarrow$ 95% C.L. upper limit.
Central 68%
Maximum posterior $\rightsquigarrow$ shortest interval

Marginal Posterior

In general, the model parameters θ include nuisance parameters θ_n.

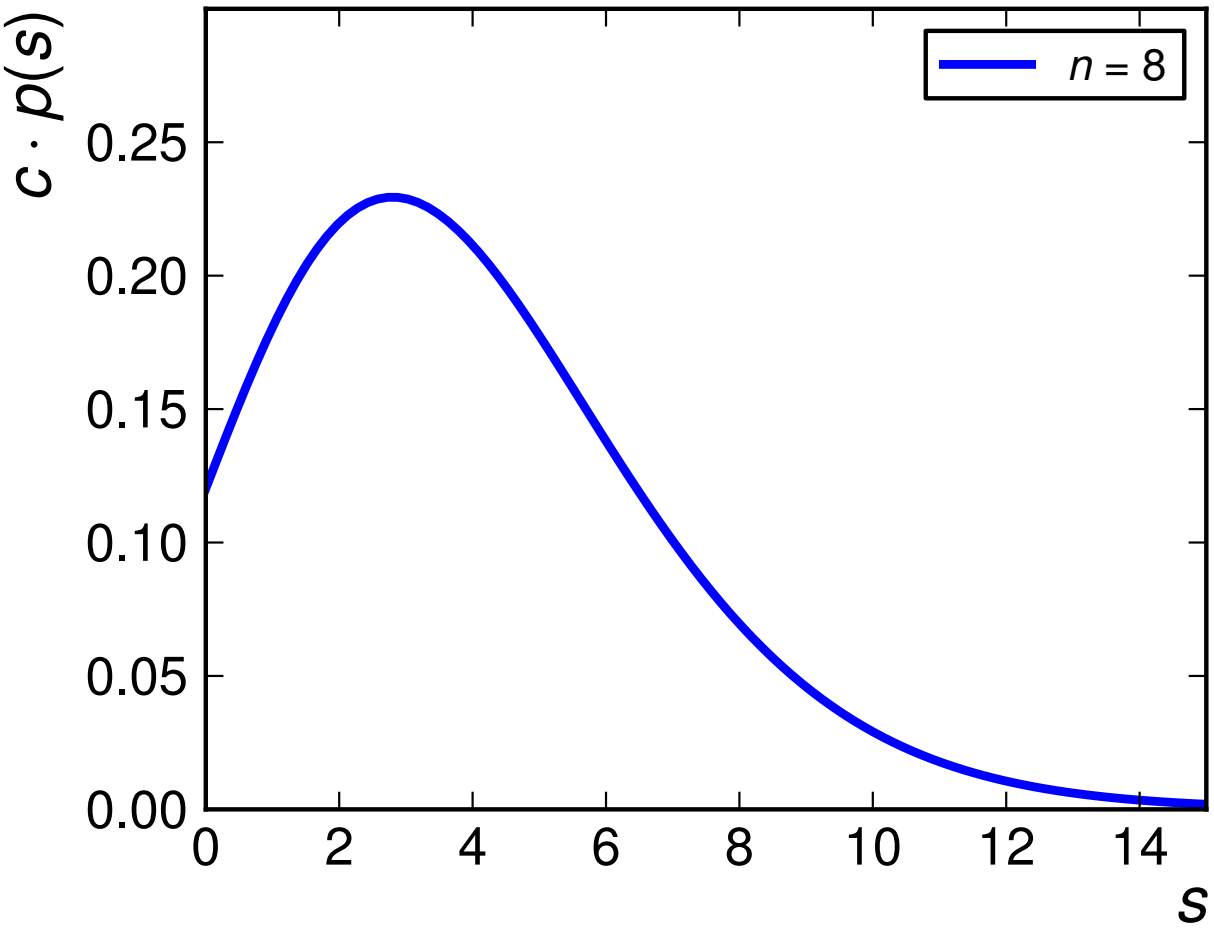
For inferences about μ, use the marginal posterior in μ, in which the nuisance parameters have been "integrating out":

$$p(\mu|d) = \int_{\theta_n} p(\theta|d) d\theta_n = c \int_{\theta_n} p(d|\theta)\pi(\theta) d\theta_n$$

where c is a normalization constant.

Example

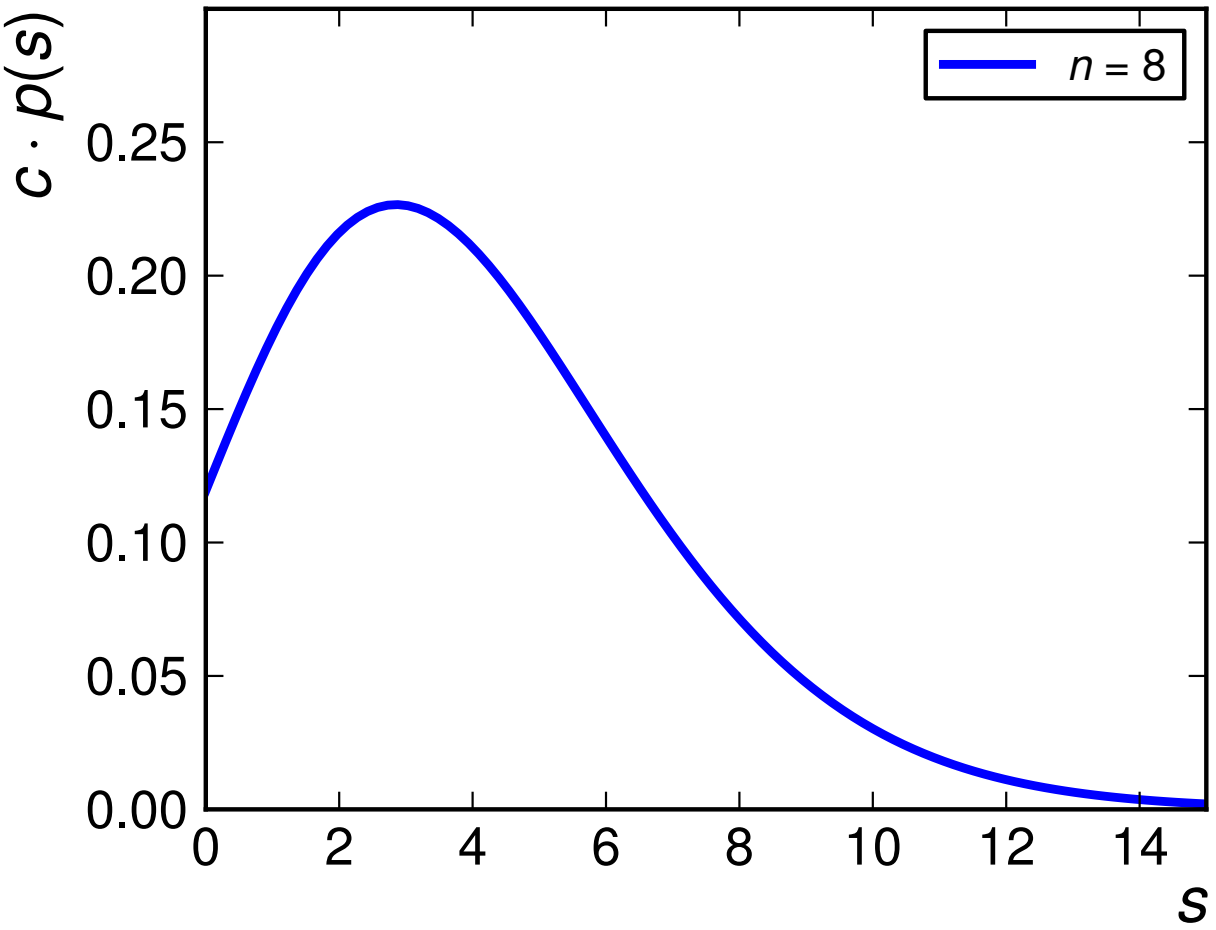
For a normal prior on b with mean $b_0 = 5.2$, look at the marginal posterior for $n_{\text{obs}} = 8$ for $\Delta b = 0\%$



As expected intuitively, posterior is broadened $\rightsquigarrow$ intervals and limits derived from marginal posterior will be larger. (See also exercise 3.)

Example

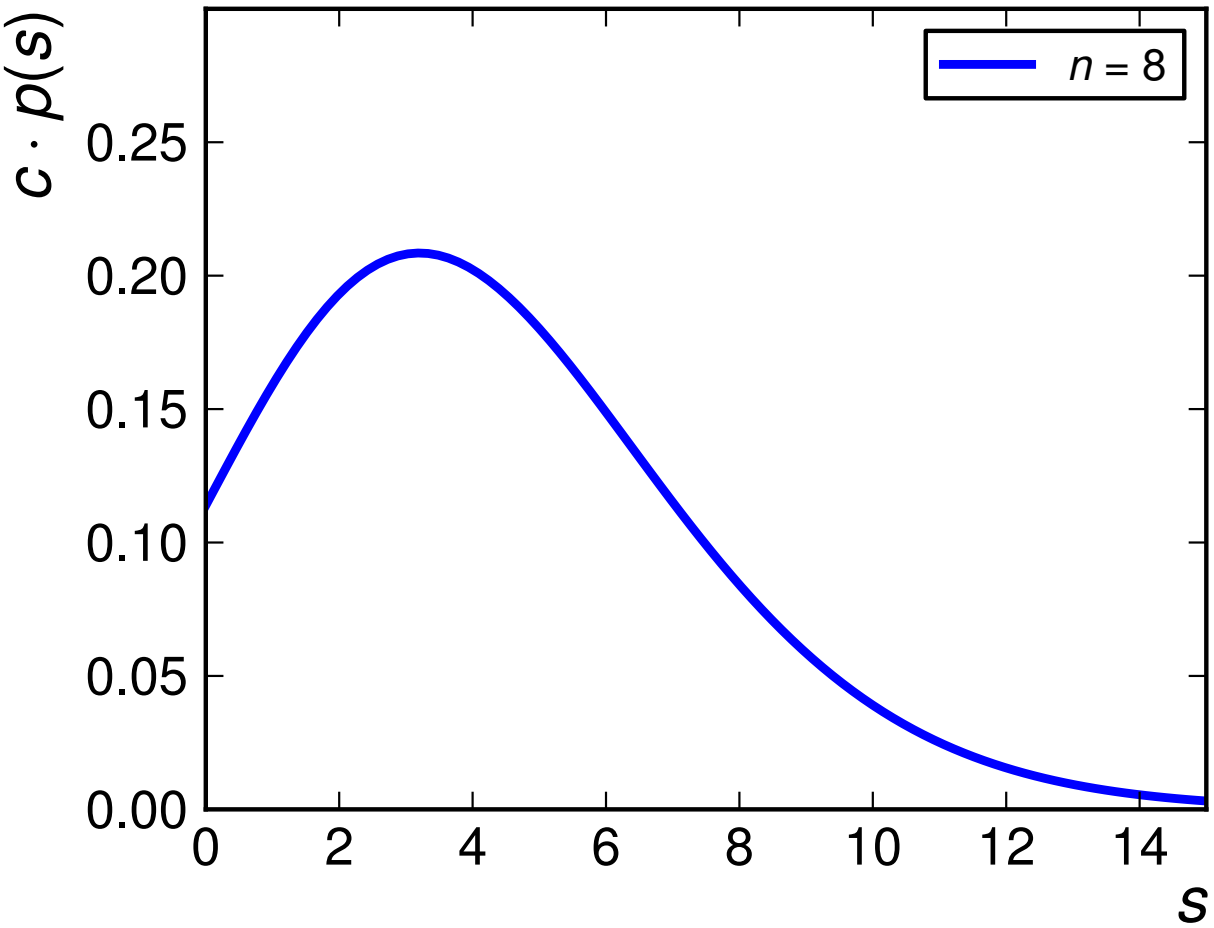
For a normal prior on b with mean $b_0 = 5.2$, look at the marginal posterior for $n_{\text{obs}} = 8$ for $\Delta b = 10\%$



As expected intuitively, posterior is broadened $\rightsquigarrow$ intervals and limits derived from marginal posterior will be larger. (See also exercise 3.)

Example

For a normal prior on b with mean $b_0 = 5.2$, look at the marginal posterior for $n_{\text{obs}} = 8$ for $\Delta b = 30\%$



As expected intuitively, posterior is broadened $\rightsquigarrow$ intervals and limits derived from marginal posterior will be larger. (See also exercise 3.)

Properties of Bayesian Intervals

- Bayesian intervals do not have guaranteed coverage; this is not a central concept in Bayesian statistic. But: can still determine coverage. In practice, coverage for limits in case of flat prior for signal is often fulfilled approximately.
- Posterior depends on prior; this is a source of fundamental criticism of the Bayesian method; even using a "flat" prior is ambiguous: transformation of variables can make a "flat" prior non-flat! Can try to make this subjectiveness objective by formal prior selection rules (cf. Luc's talk).
  Current HEP practice:
  - Check (frequentist) coverage properties
  - Check whether result is sensitive to prior by checking different "reasonable" priors

Interval Summary

- Relation between hypothesis tests and (frequentist) confidence intervals $\rightsquigarrow$ inherit concepts and methods from hypothesis tests (test statistic, systematics handling, ...)
- Additional degree of freedom: Ordering rule in Neyman construction $\rightsquigarrow$ central intervals, Feldman-Cousins
- CLs method to "fix" very small / zero exclusions
- Purely Bayesian method: Derive (marginal) posterior and choose interval from there.

Part III

Backup

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.

12 Frequentist Interpretation; Bootstrapping

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.

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:

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

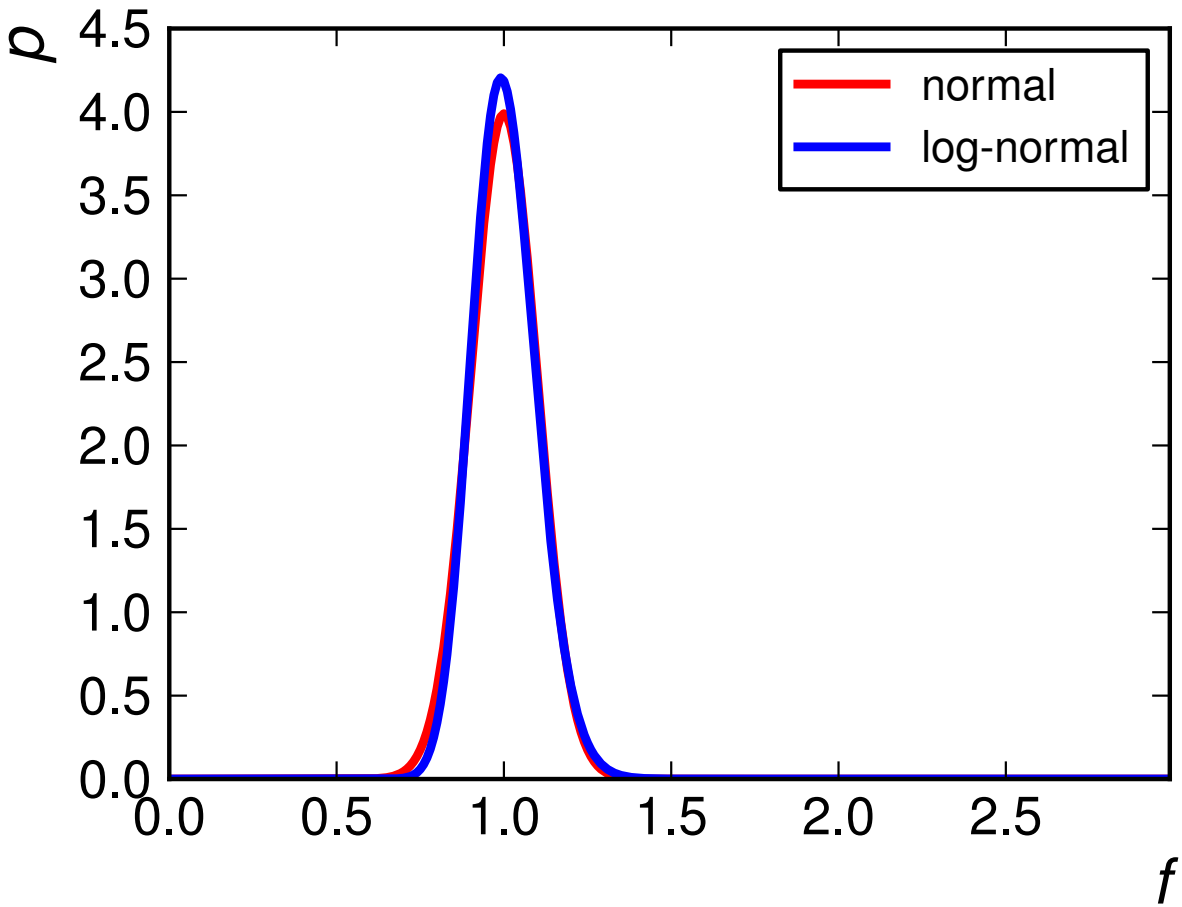
where f has a log-normal prior, i.e., $\log f$ has a normal distribution.
An equivalent formulation is

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

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

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

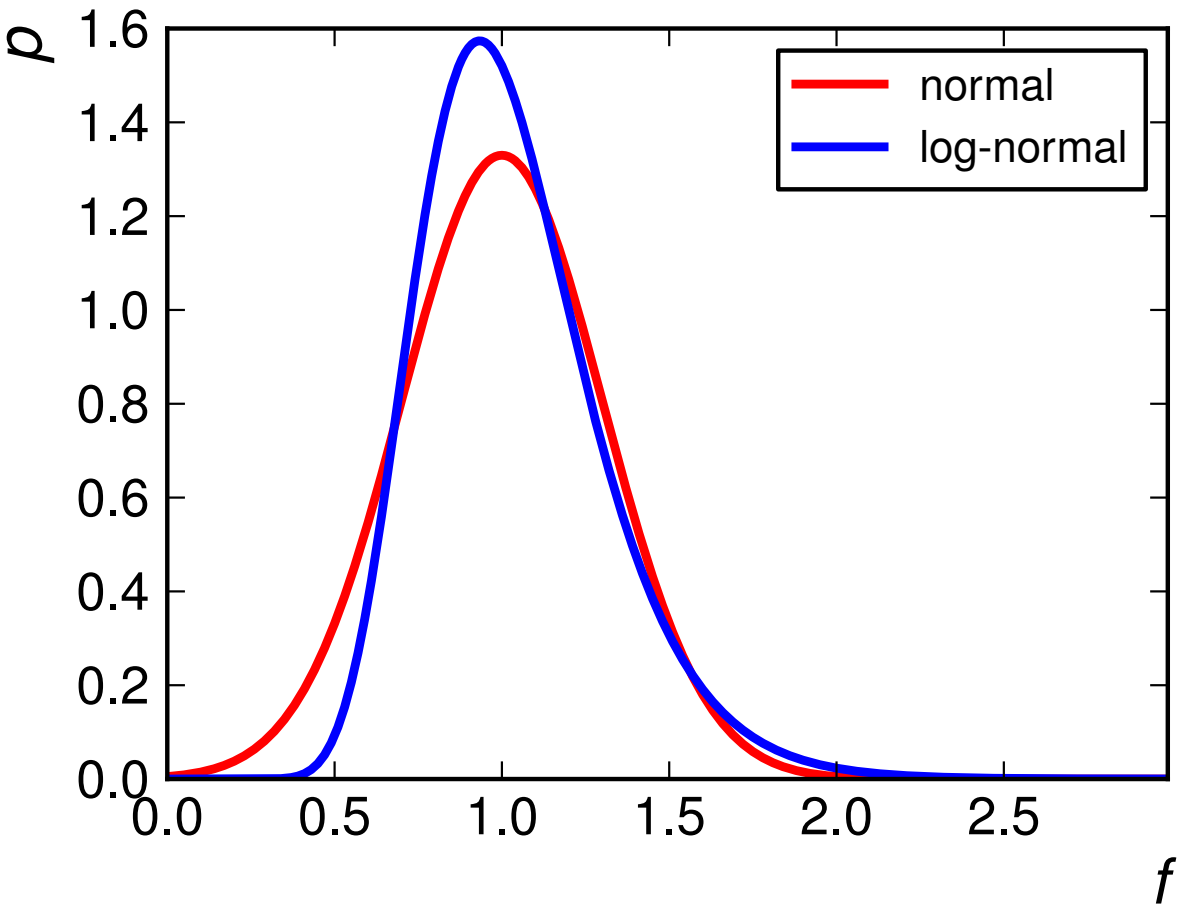
Comparing the prior for the scale factor between normal and log-normal: $\Delta b = 0.1$:



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

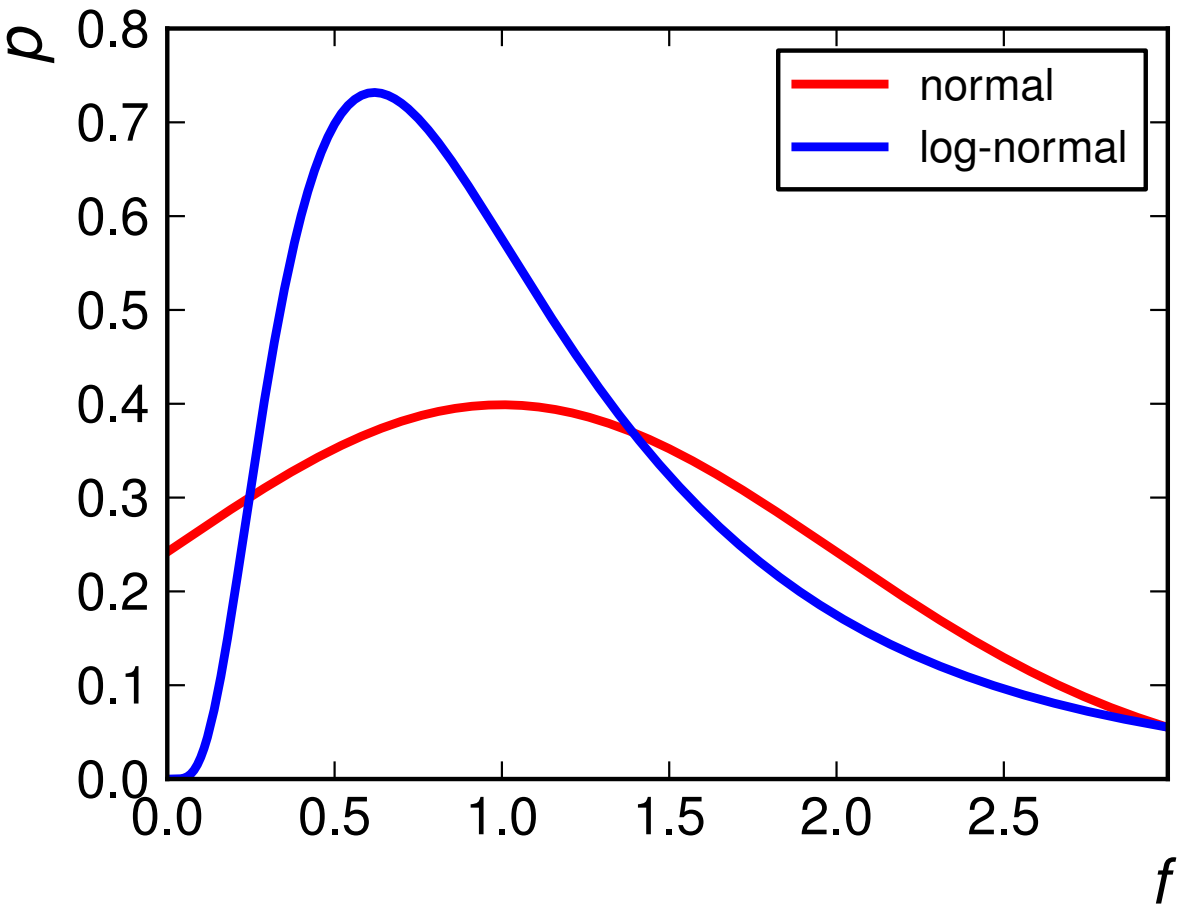
Comparing the prior for the scale factor between normal and log-normal: $\Delta b = 0.3$:



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.