# Introduction to Bayesian Statistics

Daniel Greenwald

Technische Universität München

SOS 2016, Autrans

# from Jim Berger, Duke University

"Bayesian Multiplicity Control: i.e., Bayesian methods of controlling for the 'look elsewhere effect"'



## The need for multiplicity control:

In a recent talk about the drug discovery process, the following numbers were given in illustration.

- 10,000 relevant compounds were screened for biological activity.
- 500 passed the initial screen and were studied in vitro.
- 25 passed this screening and were studied in Phase I animal trials.
- 1 passed this screening and was studied in a Phase II human trial.

# from Jim Berger, Duke University

"Bayesian Multiplicity Control: i.e., Bayesian methods of controlling for the 'look elsewhere effect"'



## The need for multiplicity control:

In a recent talk about the drug discovery process, the following numbers were given in illustration.

- 10,000 relevant compounds were screened for biological activity.
- 500 passed the initial screen and were studied in vitro.
- 25 passed this screening and were studied in Phase I animal trials.
- 1 passed this screening and was studied in a Phase II human trial.

This could be nothing but noise, if screening was done based on 'significance at the 0.05 level.'

If no compound had any effect,

- about $10,000 \times 0.05 = 500$ would initially be significant at the 0.05 level;
- about $500 \times 0.05 = 25$ of those would next be significant at the 0.05 level;
- about $25 \times 0.05 = 1.25$ of those would next be significant at the 0.05 level
- the 1 that went to Phase II would fail with probability 0.95.

# Scientific knowledge

**Scientific knowledge = justified belief.**

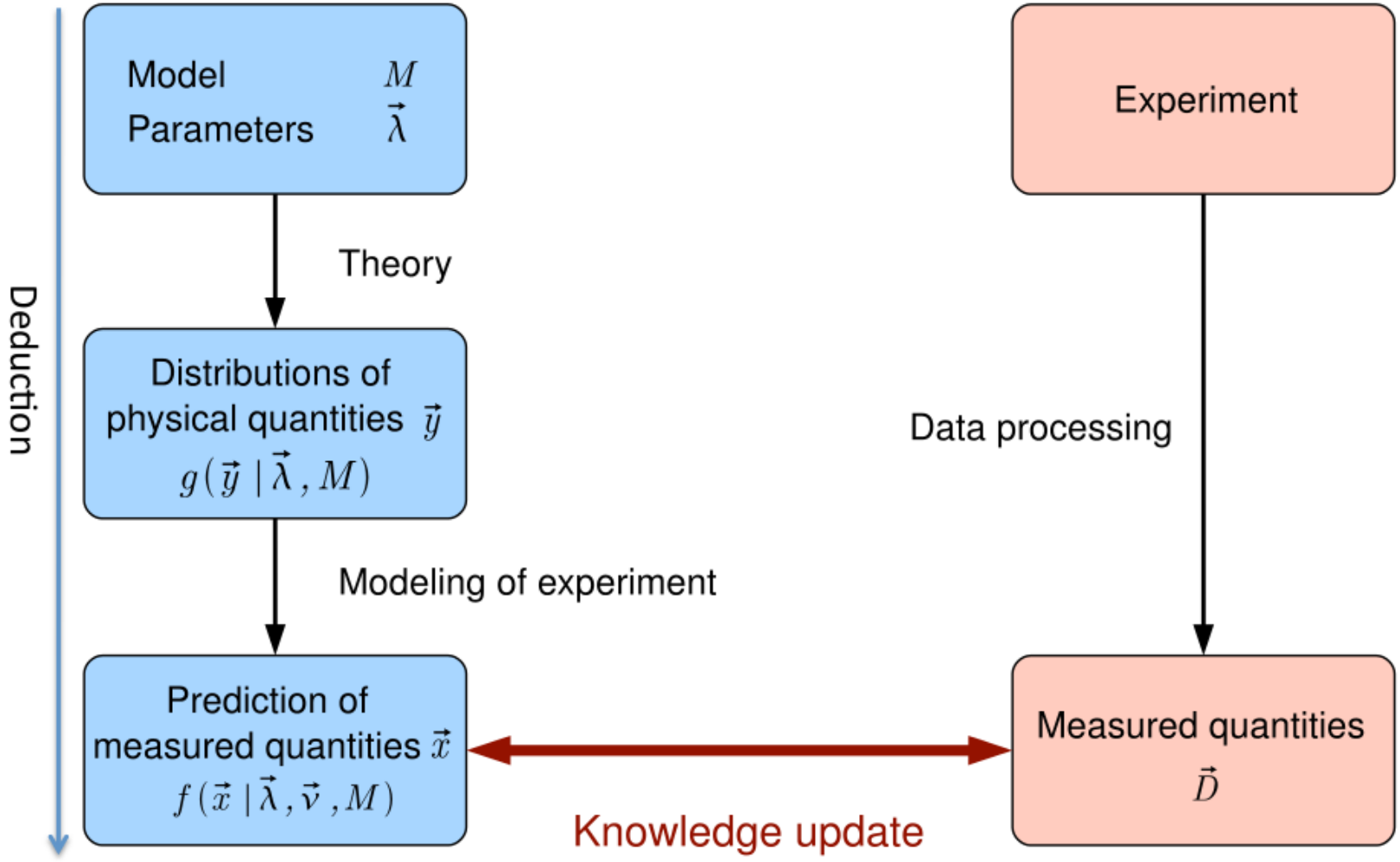
- The standard definition from philosophy adds the adjective "true." But as scientists we understand truth is unproveable without ruling out all other possible models—something that is feasibly impossible.
- "Belief" is self-explanatory;
- so what is "justification"?

  It is embodied in a **model** that

  - is logically consistent and explains known facts,
  - and makes testable predictions.

  Given several models that fulfil the above, we often apply "Occam's razor," selecting the simplest model as the best. But in the absence of mathematical differences in their predictive powers, this is only an aesthetic preference.

# Learning


Model $M$
Parameters $\vec{\lambda}$
Theory
Distributions of physical quantities $\vec{y}$
$g(\vec{y}\,|\,\vec{\lambda}, M)$
Modeling of experiment
Prediction of measured quantities $\vec{x}$
$f(\vec{x}\,|\,\vec{\lambda}, \vec{\nu}, M)$
Deduction
Experiment
Data processing
Measured quantities
$\vec{D}$
Knowledge update

# Learning

We factorize the model into two categories:

- The **scientific model** ("theory" in the diagram)
  - predict distributions of physical quantities from parameters
- The **statistical model** ("modeling of experiment" in the diagram)
  - predict measured quantities from physical quantities

These correspond to knowledge:

- Our **scientific knowledge** consists in knowing
  - how to calculate physical quantities from parameters
  - the likely values of the parameters (or constraints on them).
- Our **stastical knowledge** consists in knowing
  - how to propagate information (probability distributions) from one step to the next

By comparing predicted quantities with actual measured quantities in the appropriate way, we learn about the model and parameters.

# Probability

We will focus on the statistical models, which requires we define probability.

We define probability according to the axioms of **Kolmogorov**:

1. Define:
   - $S$, the set of all possible states.
   - $\mathbb{S}$, the space of subsets of $S$
2. Probability is a mapping fulfilling
   - 2.1 $P : \mathbb{S} \to \mathbb{R}_{\geq 0}$, That is:
   $$P(A) \geq 0, \quad \forall A \subset S.$$
   - 2.2 The probability of $S$ is unity:
   $$P(S) = 1.$$
   - 2.3 The join probability of **disjoint** sets is the sum of their individual probabilities:
   $$P(A \cup B) = P(A) + P(B), \quad \forall A, B \mid A \cap B = \varnothing$$

# Probability

It is important to understand subtle differences in probabilities.

We illustrate with an easy problem:

Flip a (fair) coin 10 times, resulting in:

$$A: \quad \text{T H T H H T T H T H}$$

Repeat, resulting in:

$$B: \quad \text{T T T T T T T T T T}$$

Which one is more likely? What do I mean by that?

# Probability

$A$: T H T H H T T H T H

$B$: T T T T T T T T T T

1. The probability of the sequence:

$$P(A) =$$

$$P(B) =$$

2. The probability of the set (regardless of sequence):

$$P(A) =$$

$$P(B) =$$

# Probability

$A$: T H T H H T T H T H

$B$: T T T T T T T T T T

1. The probability of the sequence:

$$P(A) = \left(\frac{1}{2}\right)^{10}$$

$$P(B) = \left(\frac{1}{2}\right)^{10}$$

2. The probability of the set (regardless of sequence):

$$P(A) =$$

$$P(B) =$$

# Probability

$A$: T H T H H T T H T H

$B$: T T T T T T T T T T

1. The probability of the sequence:

$$P(A) = \left(\frac{1}{2}\right)^{10}$$

$$P(B) = \left(\frac{1}{2}\right)^{10}$$

2. The probability of the set (regardless of sequence):

$$P(A) = \binom{10}{5}\left(\frac{1}{2}\right)^{5}\left(\frac{1}{2}\right)^{5} = 252/1024$$

$$P(B) = \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0} = 1/1024$$

# Bayes' Theorem

1. We can relate the **joint probability** to the **conditional probability** two ways
   1.1 Kolmogorov definition of conditional probability:
   $$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$
   1.2 Axiom of conditional probability:
   $$P(A \cap B) = P(A|B)P(B)$$
2. Note that joint probability is **commutative**:
   $$P(A \cap B) = P(B \cap A)$$
3. From which Bayes' theorem practically falls out:

# Bayes' Theorem

1. We can relate the **joint probability** to the **conditional probability** two ways
   1.1 Kolmogorov definition of conditional probability:
   $$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$
   1.2 Axiom of conditional probability:
   $$P(A \cap B) = P(A|B)P(B)$$
2. Note that joint probability is **commutative**:
   $$P(A \cap B) = P(B \cap A)$$
3. From which Bayes' theorem practically falls out:
   $$P(A|B)P(B) = P(B|A)P(A)$$

# Bayes' Theorem

1. We can relate the **joint probability** to the **conditional probability** two ways
   1.1 Kolmogorov definition of conditional probability:
   $$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$
   1.2 Axiom of conditional probability:
   $$P(A \cap B) = P(A|B)P(B)$$
2. Note that joint probability is **commutative**:
   $$P(A \cap B) = P(B \cap A)$$
3. From which Bayes' theorem practically falls out:
   $$P(A|B)P(B) = P(B|A)P(A)$$
   $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Bayes' Theorem

We can expand the theorem using the **law of total probability**, which states

$$P(X) = \sum_i P(X|Y_i)P(Y_i),$$

if

$$\sum_i P(Y_i) = 1 \text{ and } Y_i \cap Y_j = \varnothing, \ \forall i, j$$

That first requirement is the same as requiring $\cup_i Y_i$ spans all possible states.

So:

$$P(A|B) = \frac{P(B|A)P(B)}{\sum_i P(B|A_i)P(A_i)}$$

In the context of data ($D$) and a model ($M$) we write this:

$$\underbrace{P(M|D)}_{\text{posterior}} = \frac{P(D|M)\overbrace{P_0(M)}^{\text{prior}}}{\underbrace{\sum_m P(D|m)P_0(m)}_{\text{evidence } (Z)}}$$

# Simple example of Bayes' Theorem

Let us explore the theorem in a simple example:

- We have a sensor that detects only samples of type $A$ or type $B$.

  It is efficient at detecting $A$'s and inefficient at detecting $B$'s

  $$P(\text{signal}|A) = 95\% \qquad P(\text{signal}|B) = 2\% \qquad P(\text{signal}|\overline{A \text{ or } B}) = 0$$

- We use the sensor on an unknown sample and it produces a signal ($S$).

  What do we know?

  What is $P(A|S)$?

  What is $P(B|S)$?

# Simple example of Bayes' Theorem

$$P(\text{signal}|A) = 95\% \qquad P(\text{signal}|B) = 2\% \qquad P(\text{signal}|\overline{A \text{ or } B}) = 0$$

Apply Bayes' theorem:

$$P(X|S) = \frac{P(S|X)P_0(X)}{P(S|A)P_0(A) + P(S|B)P_0(B) + \underbrace{P(S|\overline{A \text{ or } B})}_{0} P_0(\overline{A \text{ or } B})}$$

$X = A$ or $B$

Obviously we need to know the priors!

# Simple example of Bayes' theorem

$$P(\text{signal}|A) = 95\% \qquad P(\text{signal}|B) = 2\%$$

$$P(X|S) = \frac{P(S|X)P_0(X)}{P(S|A)P_0(A) + P(S|B)P_0(B)}$$

Let us try a couple of scenarios:

1. Given no information, we reasonably choose

$$P_0(A) = P_0(B)$$

# Simple example of Bayes' theorem

$$P(\text{signal}|A) = 95\% \qquad P(\text{signal}|B) = 2\%$$

$$P(X|S) = \frac{P(S|X)P_0(X)}{P(S|A)P_0(A) + P(S|B)P_0(B)}$$

Let us try a couple of scenarios:

1. Given no information, we reasonably choose

$$P_0(A) = P_0(B)$$

So the priors cancel and:

$$P(A|S) = \frac{95\%}{95\% + 2\%} = 97.9\% \qquad P(B|S) = 2.1\%$$

# Simple example of Bayes' theorem

$$P(\text{signal}|A) = 95\% \qquad P(\text{signal}|B) = 2\%$$

$$P(X|S) = \frac{P(S|X)P_0(X)}{P(S|A)P_0(A) + P(S|B)P_0(B)}$$

Let us try a couple of scenarios:

1. Given no information, we reasonably choose

$$P_0(A) = P_0(B)$$

2. We know that $B$'s are far more likely than $A$'s:

$$P_0(B)/P_0(A) = 1000$$

# Simple example of Bayes' theorem

$$P(\text{signal}|A) = 95\% \qquad P(\text{signal}|B) = 2\%$$

$$P(X|S) = \frac{P(S|X)P_0(X)}{P(S|A)P_0(A) + P(S|B)P_0(B)}$$

Let us try a couple of scenarios:

1. Given no information, we reasonably choose

$$P_0(A) = P_0(B)$$

2. We know that $B$'s are far more likely than $A$'s:

$$P_0(B)/P_0(A) = 1000$$

So:

$$P(A|S) = \frac{95\%}{95\% + 2\% \cdot 100} = 4.5\%, \qquad P(B|S) = 95.5\%$$

# Efficiencies

Efficiency calculations illustate where the naive calculation breaks down:

Suppose we have $N$ trials of which $H$ are successes.

Naively:

$$\epsilon = \frac{H}{N} \pm \frac{\sqrt{H}}{N}$$

Suppose out of 100 trials, 98 are successes:

$$\epsilon = 98\% \pm 9.9\%$$

What does it mean for an efficiency to be above 100%?

# Efficiencies

Efficiency calculations illustate where the naive calculation breaks down:

Suppose we have $N$ trials of which $H$ are successes.

Naively:

$$\epsilon = \frac{H}{N} \pm \frac{\sqrt{H}}{N}$$

Suppose out of 100 trials, 98 are successes:

$$\epsilon = 98\% \pm 9.9\%$$

What does it mean for an efficiency to be above 100%?

The Bayesian formulation is a little more complicated.

We expand our formulation of the theorem for **parameters** in an assumed model:

$$P(\vec{\lambda}|M, D) = \frac{P(D|M, \vec{\lambda})P_0(\lambda)}{\int P(D|M, \vec{\lambda}')P_0(\vec{\lambda}')\mathrm{d}\vec{\lambda}'}$$

And let us first discuss the **binomial distribution**

# Binomial distribution

$$P(n|N,r) = \binom{N}{n} p^n (1-p)^{N-n}$$

The **binomial distribution** is used when there are

- two possible outcomes, with fixed probabilities: $p$ and $(1-p)$;
- and a fixed number of trials ($N$)

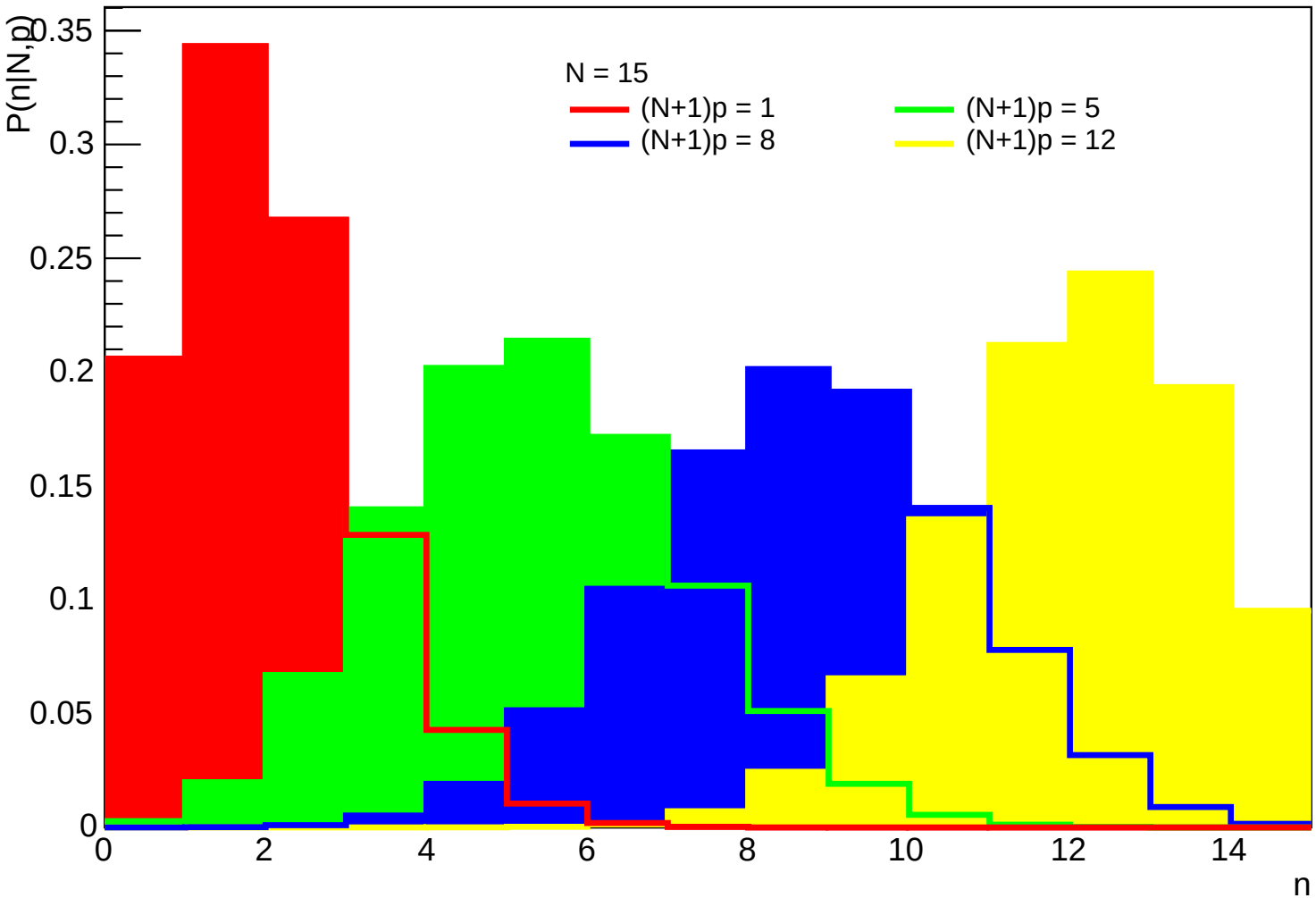
(perfect to describe efficiencies)

There are obvious constraints:

1. $0 \leq p \leq 1$
2. $0 \leq N \leq \infty$
3. $0 \leq n \leq N$, $n$ the number of outcomes occuring with prob. $p$

# Binomial distribution

$$P(n|N,r) = \binom{N}{n} p^n (1-p)^{N-n}$$



# Binomial distribution

Certain properties useful to know:

1. The binomial distribution is normalized:

$$\sum_{n=0}^{N} P(n|N,p) = 1$$

2. The expectation value for $n$ is

$$E[n] = \sum_{n=0}^{N} nP(n|N,p) = Np$$

3. The variance of $n$ is

$$V[n] = \sum_{n=0}^{N} n^2 P(n|N,p) - E[n]^2 = Np(1-p)$$

4. The mode of $n$, $n*$, depends on $N$ and $p$:

   4.1 $n^* = 0$ for $p = 0$;

   4.2 $n^* = N$ for $p = 1$;

   4.3 $n^* = \lfloor (N+1)p \rfloor$; for $0 \leq p \leq 1$ and $(N+1)p \notin \mathbb{Z}$;

   4.4 $n^* = \lfloor (N+1)p \rfloor$ and $\lfloor (N+1)p \rfloor - 1$; for $0 \leq p \leq 1$ and $(N+1)p \in \mathbb{Z}$;

# Bayesian Efficiency

$$P(\epsilon|N,H) = \frac{\binom{N}{H}\epsilon^H(1-\epsilon)^{N-H}\cdot P_0(\epsilon)}{\int_0^1 \binom{N}{H}\epsilon'^H(1-\epsilon')^{N-H}\cdot P_0(\epsilon')\mathrm{d}\epsilon'}$$

Again, we require information about the prior.

Let's try a flat **flat prior**

$$P_0(\epsilon) = 1$$

So

$$P(\epsilon|N,H) = \frac{\epsilon^H(1-\epsilon)^{N-H}}{\underbrace{\int_0^1 \epsilon'^H(1-\epsilon')^{N-H}\mathrm{d}\epsilon'}}$$

# Bayesian Efficiency

$$P(\epsilon|N,H) = \frac{\binom{N}{H}\epsilon^H(1-\epsilon)^{N-H}\cdot P_0(\epsilon)}{\int_0^1 \binom{N}{H}\epsilon'^H(1-\epsilon')^{N-H}\cdot P_0(\epsilon')\mathrm{d}\epsilon'}$$

Again, we require information about the prior.

Let's try a flat **flat prior**

$$P_0(\epsilon) = 1$$

So

$$P(\epsilon|N,H) = \frac{\epsilon^H(1-\epsilon)^{N-H}}{\underbrace{\int_0^1 \epsilon'^H(1-\epsilon')^{N-H}\mathrm{d}\epsilon'}_{\beta(H+1,N-H+1)=\frac{H!(N-H)!}{(N+1)!}}}$$

# Bayesian Efficiency

$$P(\epsilon|N,H) = \frac{\binom{N}{H}\epsilon^H(1-\epsilon)^{N-H}\cdot P_0(\epsilon)}{\int_0^1 \binom{N}{H}\epsilon'^H(1-\epsilon')^{N-H}\cdot P_0(\epsilon')\mathrm{d}\epsilon'}$$

Again, we require information about the prior.

Let's try a flat **flat prior**

$$P_0(\epsilon) = 1$$

So

$$P(\epsilon|N,H) = \frac{\epsilon^H(1-\epsilon)^{N-H}}{\underbrace{\int_0^1 \epsilon'^H(1-\epsilon')^{N-H}\mathrm{d}\epsilon'}_{\beta(H+1,N-H+1)=\frac{H!(N-H)!}{(N+1)!}}} = \frac{(N+1)!}{H!(N-H)!}\epsilon^H(1-\epsilon)^{N-H}$$

# Bayesian Efficiency

$$P(\epsilon|N,H) = \frac{\binom{N}{H}\epsilon^H(1-\epsilon)^{N-H}\cdot P_0(\epsilon)}{\int_0^1 \binom{N}{H}\epsilon'^H(1-\epsilon')^{N-H}\cdot P_0(\epsilon')\mathrm{d}\epsilon'}$$

Again, we require information about the prior.

Let's try a flat **flat prior**

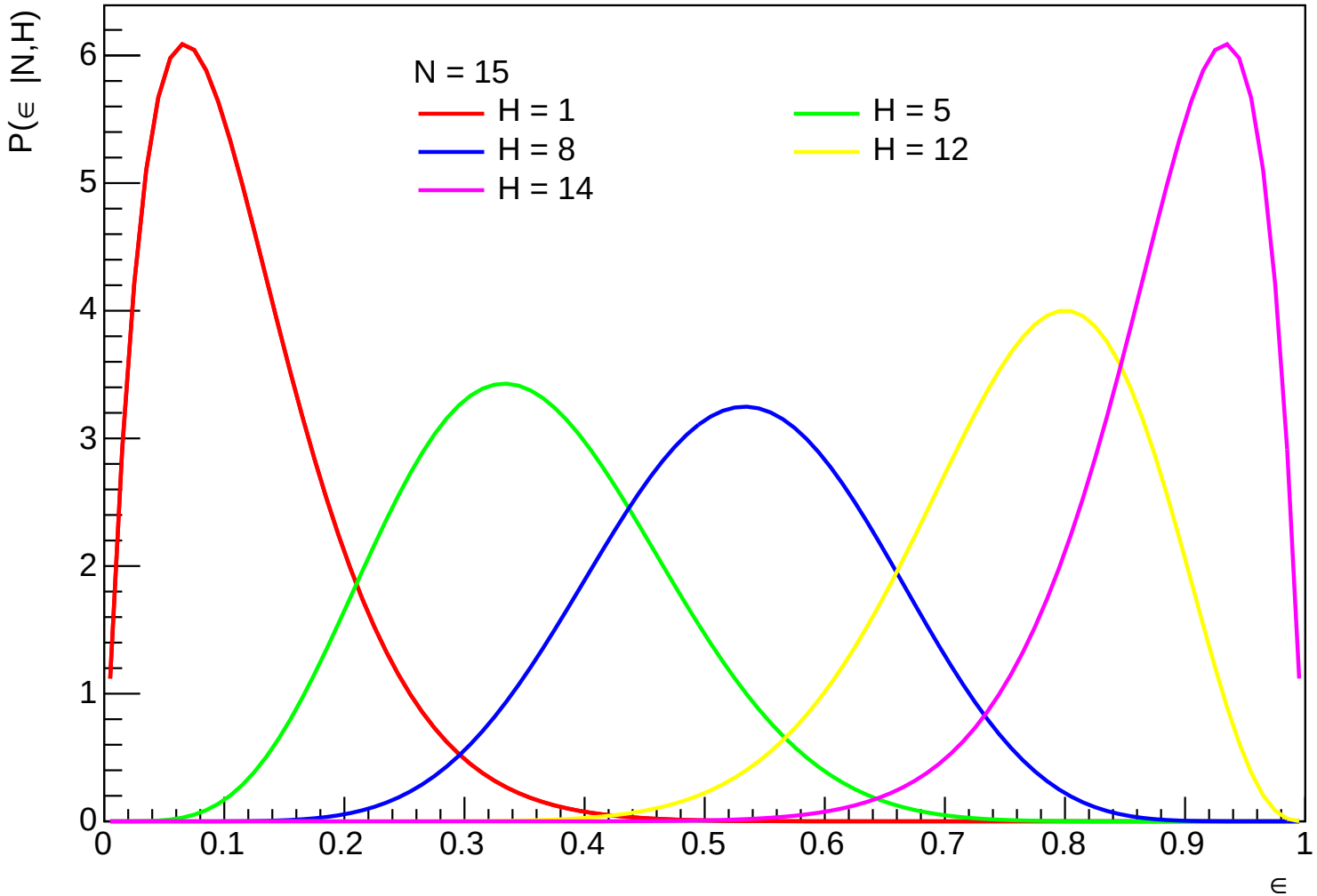
$$P_0(\epsilon) = 1$$

So

$$P(\epsilon|N,H) = \frac{\epsilon^H(1-\epsilon)^{N-H}}{\underbrace{\int_0^1 \epsilon'^H(1-\epsilon')^{N-H}\mathrm{d}\epsilon'}_{\beta(H+1,N-H+1)=\frac{H!(N-H)!}{(N+1)!}}} = \frac{(N+1)!}{H!(N-H)!}\epsilon^H(1-\epsilon)^{N-H} = (N+1)\cdot P(H|N,\epsilon)$$

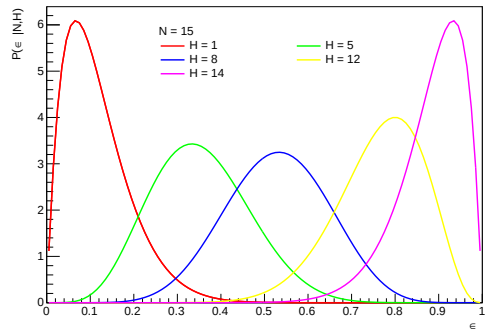
# Bayesian Efficiency

$$P(\epsilon|N,H) = (N+1)\cdot P(H|N,\epsilon)$$



# Bayesian Efficiency

$$P(\epsilon|N,H) = (N+1)\cdot P(H|N,\epsilon)$$




The mode of this distribution is

$$\epsilon^* = \frac{H}{N}$$

The mean is

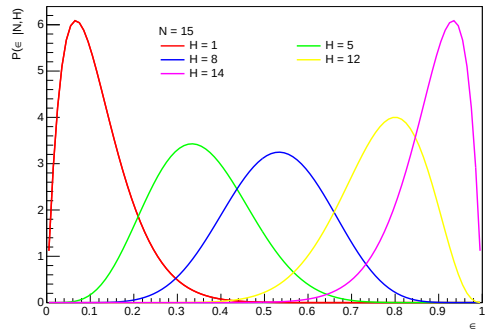
$$E[\epsilon] = \frac{H+1}{N+2}$$

The variance is

$$V[\epsilon] = \frac{E[\epsilon](1-E[\epsilon])}{N+3}$$

# Bayesian Efficiency

$$P(\epsilon|N,H) = (N+1)\cdot P(H|N,\epsilon)$$



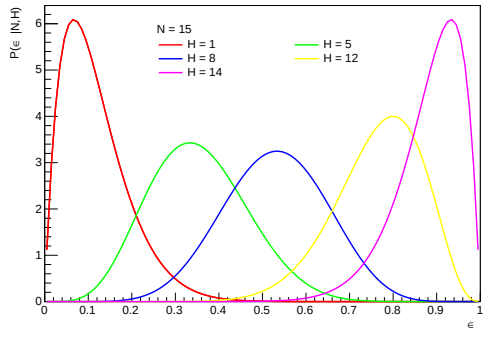

| $N$ | $H$ | $\frac{H}{N} \pm \frac{\sqrt{H}}{N}$ | $E[\epsilon] \pm (V[\epsilon])^{\frac{1}{2}}$ |
|---|---|---|---|
| 15 | 1 | 6.7% ± 6.7% | 11.8% ± 7.6% |
| 15 | 5 | 33.3% ± 14.9% | 35.3% ± 11.3% |
| 15 | 8 | 53.3% ± 18.9% | 52.9% ± 11.8% |
| 15 | 12 | 80.0% ± 23.1% | 76.5% ± 10.0% |
| 15 | 14 | 93.3% ± 24.9% | 88.2% ± 7.6% |

Notice that the uncertainty grows with $H$ in the naive calculation!

But it's symmetric for the Bayesian calculation.

# Bayesian Efficiency

$$P(\epsilon|N,H) = (N+1)\cdot P(H|N,\epsilon)$$




| $N$ | $H$ | $\frac{H}{N} \pm \frac{\sqrt{H}}{N}$ | $E[\epsilon] \pm (V[\epsilon])^{\frac{1}{2}}$ |
|---|---|---|---|
| 15 | 14 | 93.3% ± 24.9% | 88.2% ± 7.6% |
| 30 | 28 | 93.3% ± 17.6% | 90.6% ± 5.1% |
| 45 | 42 | 93.3% ± 14.4% | 91.5% ± 4.0% |
| 60 | 56 | 93.3% ± 12.5% | 91.9% ± 3.4% |
| 75 | 70 | 93.3% ± 11.2% | 92.2% ± 3.0% |

Keeping the mode, $\epsilon^* = H/N$, fixed

# Poisson distribution

The **Poisson distribution** can be obtained from the **binomial distribution** via

$$N \to \infty \qquad p \to 0$$

while maintaining

$$Np = \text{finite}$$

# Poisson distribution

The **Poisson distribution** can be obtained from the **binomial distribution** via

$$N \to \infty \qquad p \to 0$$

while maintaining

$$Np = \text{finite}$$

We introduce

$$\nu \equiv Np$$

so:

$$P(n|N,\nu) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

# Poisson distribution

The **Poisson distribution** can be obtained from the **binomial distribution** via

$$N \to \infty \qquad p \to 0$$

while maintaining

$$Np = \text{finite}$$

We introduce

$$\nu \equiv Np$$

so:

$$P(n|N,\nu) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

We have (if $n$ remains finite)

$$\lim_{N\to\infty}\frac{N!}{(N-n)!} \to N^n$$

and

$$\lim_{N\to\infty}\left(1-\frac{\nu}{N}\right)^{N-n} \to e^{-\nu}$$

# Poisson distribution

The **Poisson distribution** can be obtained from the **binomial distribution** via

$$N \to \infty \qquad p \to 0$$

while maintaining

$$Np = \text{finite}$$

We introduce

$$\nu \equiv Np$$

so:

$$P(n|N,\nu) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

We have (if $n$ remains finite)

$$\lim_{N\to\infty}\frac{N!}{(N-n)!} \to N^n$$

and

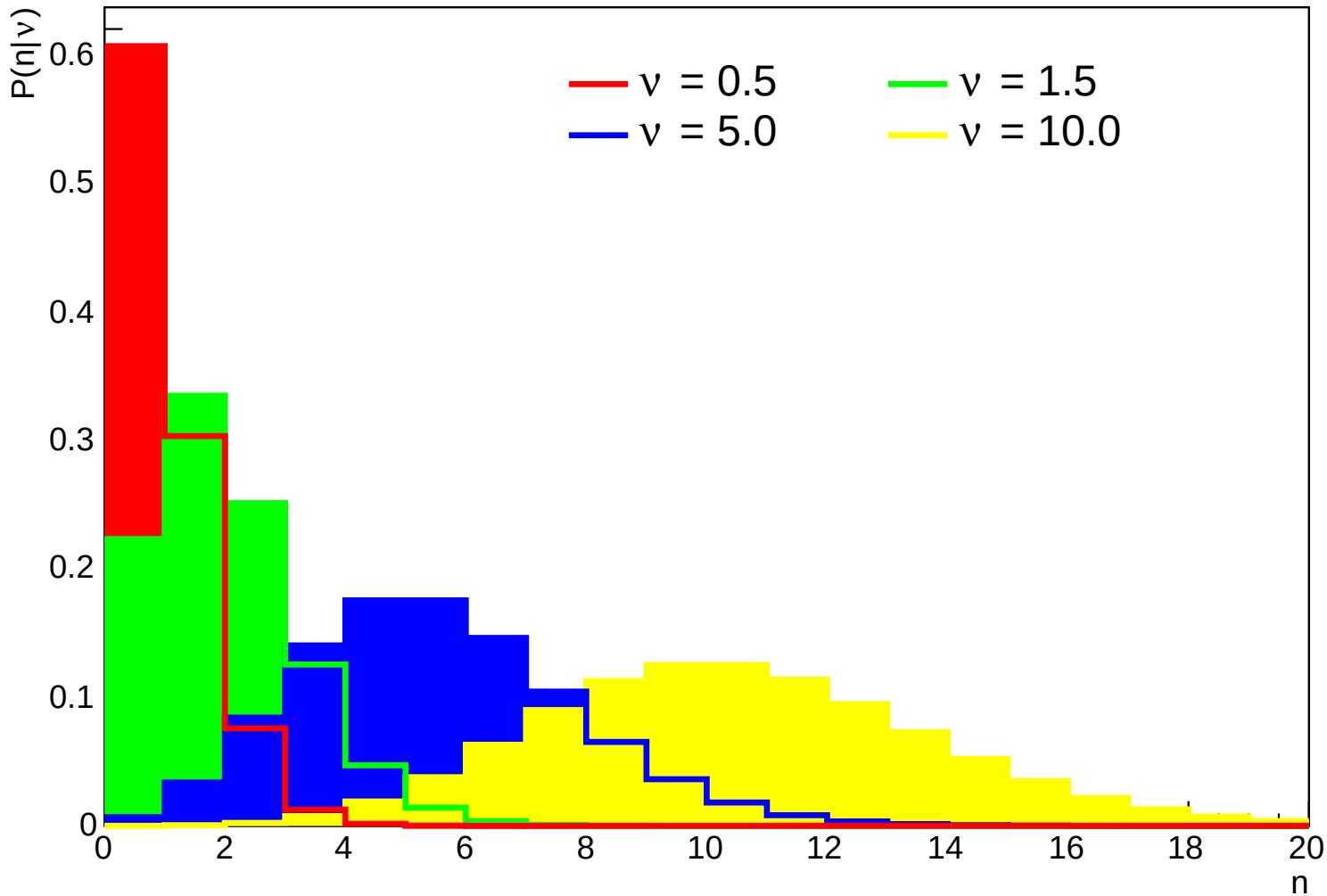
$$\lim_{N\to\infty}\left(1-\frac{\nu}{N}\right)^{N-n} \to e^{-\nu}$$

So

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}$$

# Poisson distribution

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}, \qquad \nu \in \mathbb{R}_{\geq 0}, \qquad n \in \mathbb{N}_0$$



# Poisson distribution

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}, \qquad \nu \in \mathbb{R}_{\geq 0}, \qquad n \in \mathbb{N}_0$$

The expectation value of $n$ is

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n e^{-\nu}}{n!}$$

# Poisson distribution

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}, \qquad \nu \in \mathbb{R}_{\geq 0}, \qquad n \in \mathbb{N}_0$$

The expectation value of $n$ is

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n e^{-\nu}}{n!} = \nu e^{-\nu} \underbrace{\sum_{n=1}^{\infty}}_{n=0 \text{ negl.}} \frac{\nu^{n-1}}{(n-1)!}$$

# Poisson distribution

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}, \qquad \nu \in \mathbb{R}_{\geq 0}, \qquad n \in \mathbb{N}_0$$

The expectation value of $n$ is

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n e^{-\nu}}{n!} = \nu e^{-\nu} \underbrace{\sum_{n=1}^{\infty}}_{n=0 \text{ negl.}} \frac{\nu^{n-1}}{(n-1)!} = \nu e^{-\nu} \underbrace{\sum_{m=0}^{\infty} \frac{\nu^m}{m!}}_{e^{\nu}}$$

# Poisson distribution

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}, \qquad \nu \in \mathbb{R}_{\geq 0}, \qquad n \in \mathbb{N}_0$$

The expectation value of $n$ is

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n e^{-\nu}}{n!} = \nu e^{-\nu} \underbrace{\sum_{n=1}^{\infty}}_{n=0 \text{ negl.}} \frac{\nu^{n-1}}{(n-1)!} = \nu e^{-\nu} \underbrace{\sum_{m=0}^{\infty} \frac{\nu^m}{m!}}_{e^{\nu}} = \nu$$

# Poisson distribution

$$P(n|\nu) = \frac{\nu^n e^{-\nu}}{n!}, \qquad \nu \in \mathbb{R}_{\geq 0}, \qquad n \in \mathbb{N}_0$$

The expectation value of $n$ is

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n e^{-\nu}}{n!} = \nu e^{-\nu} \underbrace{\sum_{n=1}^{\infty}}_{n=0 \text{ negl.}} \frac{\nu^{n-1}}{(n-1)!} = \nu e^{-\nu} \underbrace{\sum_{m=0}^{\infty} \frac{\nu^m}{m!}}_{e^{\nu}} = \nu$$

other useful properties:

1. Normalization:
$$\sum_{n=0}^{\infty} = \frac{\nu^n e^{-\nu}}{n!} = e^{-\nu} \frac{\nu^n}{n!} = 1$$

2. $V[n] = \nu$ (also check as an exercise)

3. $n^* = \lfloor \nu \rfloor$, if $\nu \notin \mathbb{Z}$
   $n^* = \nu$ and $\nu - 1$, if $\nu \in \mathbb{Z}$

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$?

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$? Bayes theorem!

$$P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu')P_0(\nu')\mathrm{d}\nu'}$$

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$? Bayes theorem!

$$P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu')P_0(\nu')\mathrm{d}\nu'}$$

Again, we need info about the prior.

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$? Bayes theorem!

$$P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu')P_0(\nu')\mathrm{d}\nu'}$$

Again, we need info about the prior.

Let us take a **flat prior**:

$$P_0(\nu) = \frac{1}{\nu_\infty}, \text{ for } 0 \leq \nu \leq \nu_\infty$$

0 otherwise.

$\nu_\infty \gg n$ finite but large: finite $\nu_\infty \rightarrow$ nonzero prior!

$$P(\nu|n) = \frac{\nu^n e^{-\nu}}{\underbrace{\int_0^{\nu_\infty} \nu'^n e^{-\nu'}\mathrm{d}\nu}}$$

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$? Bayes theorem!

$$P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu')P_0(\nu')\mathrm{d}\nu'}$$

Again, we need info about the prior.

Let us take a **flat prior**:

$$P_0(\nu) = \frac{1}{\nu_\infty}, \text{ for } 0 \leq \nu \leq \nu_\infty$$

0 otherwise.

$\nu_\infty \gg n$ finite but large: finite $\nu_\infty \rightarrow$ nonzero prior!

$$P(\nu|n) = \frac{\nu^n e^{-\nu}}{\underbrace{\int_0^{\nu_\infty} \nu'^n e^{-\nu'}\mathrm{d}\nu}_{\approx n!}}$$

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$? Bayes theorem!

$$P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu')P_0(\nu')\mathrm{d}\nu'}$$

Again, we need info about the prior.

Let us take a **flat prior**:

$$P_0(\nu) = \frac{1}{\nu_\infty}, \text{ for } 0 \leq \nu \leq \nu_\infty$$

0 otherwise.

$\nu_\infty \gg n$ finite but large: finite $\nu_\infty \rightarrow$ nonzero prior!

$$P(\nu|n) = \frac{\nu^n e^{-\nu}}{\underbrace{\int_0^{\nu_\infty} \nu'^n e^{-\nu'}\mathrm{d}\nu}_{\approx n!}} = \frac{\nu^n e^{-\nu}}{n!}$$

# Poisson distribution in Bayesian context

$P(n|\nu)$ is distribution of $n$. What about $\nu$? Bayes theorem!

$$P(\nu|n) = \frac{P(n|\nu)P_0(\nu)}{\int_0^\infty P(n|\nu')P_0(\nu')\mathrm{d}\nu'}$$

Again, we need info about the prior.

Let us take a **flat prior**:

$$P_0(\nu) = \frac{1}{\nu_\infty}, \text{ for } 0 \leq \nu \leq \nu_\infty$$

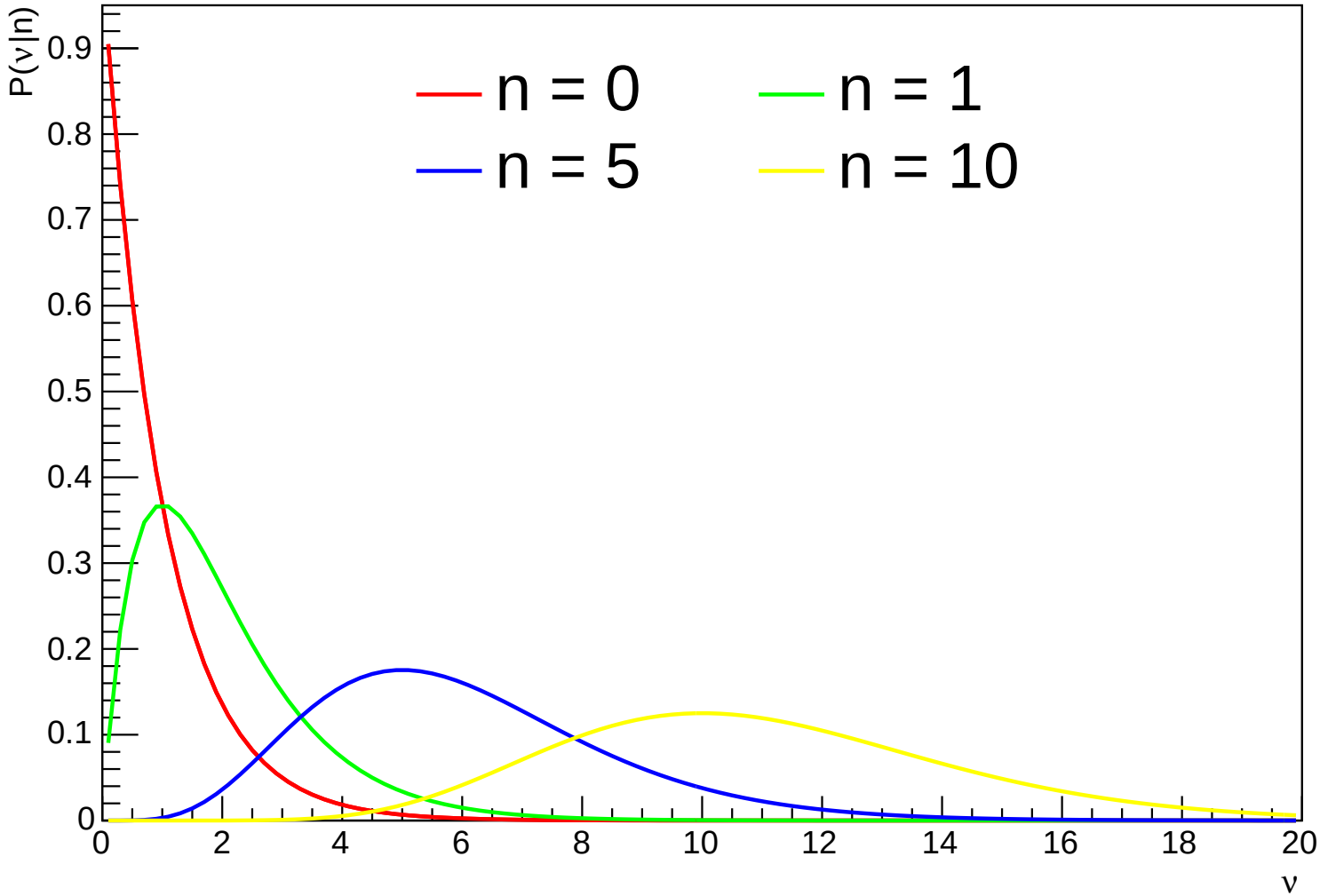
0 otherwise.

$\nu_\infty \gg n$ finite but large: finite $\nu_\infty \rightarrow$ nonzero prior!

$$P(\nu|n) = \frac{\nu^n e^{-\nu}}{\underbrace{\int_0^{\nu_\infty} \nu'^n e^{-\nu'}\mathrm{d}\nu}_{\approx n!}} = \frac{\nu^n e^{-\nu}}{n!} = P(n|\nu)$$

# Poisson distribution in Bayesian context

$$P_0(\nu) = \text{flat} \rightarrow P(\nu|n) = P(n|\nu)$$



# Poisson distribution in Bayesian context

$$P_0(\nu) = \text{flat} \rightarrow P(\nu|n) = P(n|\nu)$$

Properties:

- Normalized:
$$\int_0^\infty \frac{\nu^n e^{-\nu}}{n!} \mathrm{d}\nu = 1$$
- $\nu^* = n$
- $E[\nu] = \int_0^{\nu\infty} \nu \frac{\nu^n e^{-\nu}}{n!} \mathrm{d}\nu = \frac{(n+1)!}{n!} = n+1$
- $V[\nu] = E[\nu^2] - E[\nu]$
$$= \int_0^{\nu\infty} \nu^2 P(\nu|n) \mathrm{d}\nu - (n+1)^2$$
$$= \frac{(n+2)!}{n!} - (n+1)^2 = n+1$$

# Poisson distribution in Bayesian context

What if we measure $n = 0$?

$$\nu^* = ?, \qquad E[\nu] = ?$$

# Poisson distribution in Bayesian context

What if we measure $n = 0$?

$$\nu^* = 0, \qquad E[\nu] = 1$$

# Poisson distribution in Bayesian context

What if we measure $n = 0$?

$$\nu^* = 0, \qquad E[\nu] = 1$$

Remember $n = 0$ is a **measurement**!

$$P(\nu|0) = e^{-\nu}$$

# Poisson distribution in Bayesian context

What if we measure $n = 0$?

$$\nu^* = 0, \qquad E[\nu] = 1$$

Remember $n = 0$ is a **measurement**!

$$P(\nu|0) = e^{-\nu}$$

And remember this is still a big change from the prior:

$$E_0[\nu] = \int_0^{\nu_\infty} \nu P_0(\nu) \mathrm{d}\nu = \left[\frac{\nu^2}{2\nu_\infty}\right]_0^{\nu_\infty} = \frac{\nu_\infty}{2}$$

So:

$$E_0[\nu] \to E[\nu] = \frac{\nu_\infty}{2} \to 1$$

# Poisson distribution in Bayesian context

No let's look at the cumulative distribution:

$$F(\nu|n) = \int_0^\nu \frac{\nu'^n e^{-\nu'}}{n!} \mathrm{d}\nu' = \left[ -\frac{\nu'^n e^{-\nu'}}{n!} \right]_0^\nu - \int_0^\nu \frac{\nu'^{n-1} e^{-\nu'}}{(n-1)!} \mathrm{d}\nu' = \ldots$$

repeat until denominator in integral $= -1!$, since $(-1!)^{-1} = 0$

$$F(\nu|n) = 1 - e^{-\nu} \sum_{m=0}^{n} \frac{\nu^m}{m!}$$

So if we measure $n = 0$, what is the 95%-credibility upper limit?

$$F(\nu_{95}|0) = 1 - e^{-\nu_{95}} = 0.95 \quad \rightarrow \quad \nu_{95} = -\ln 0.05 \approx 3$$

# Cautionary interlude!

If you want to use the so-called "poisson uncertainty", remember this only applies at large $n$!

The 68% probability region is $[n - \sqrt{n}, n + \sqrt{n}]$ (for large $n$)

# Poisson distribution in Bayesian context

What about a **nonflat prior**?

In general we cannot analytically solve the problem.

(Generally solve numerically—as we'll see in the exercises with BAT.)

But let us take one example:

$$P_0(\nu) = \frac{1}{\nu_0} e^{-\nu/\nu_0}$$

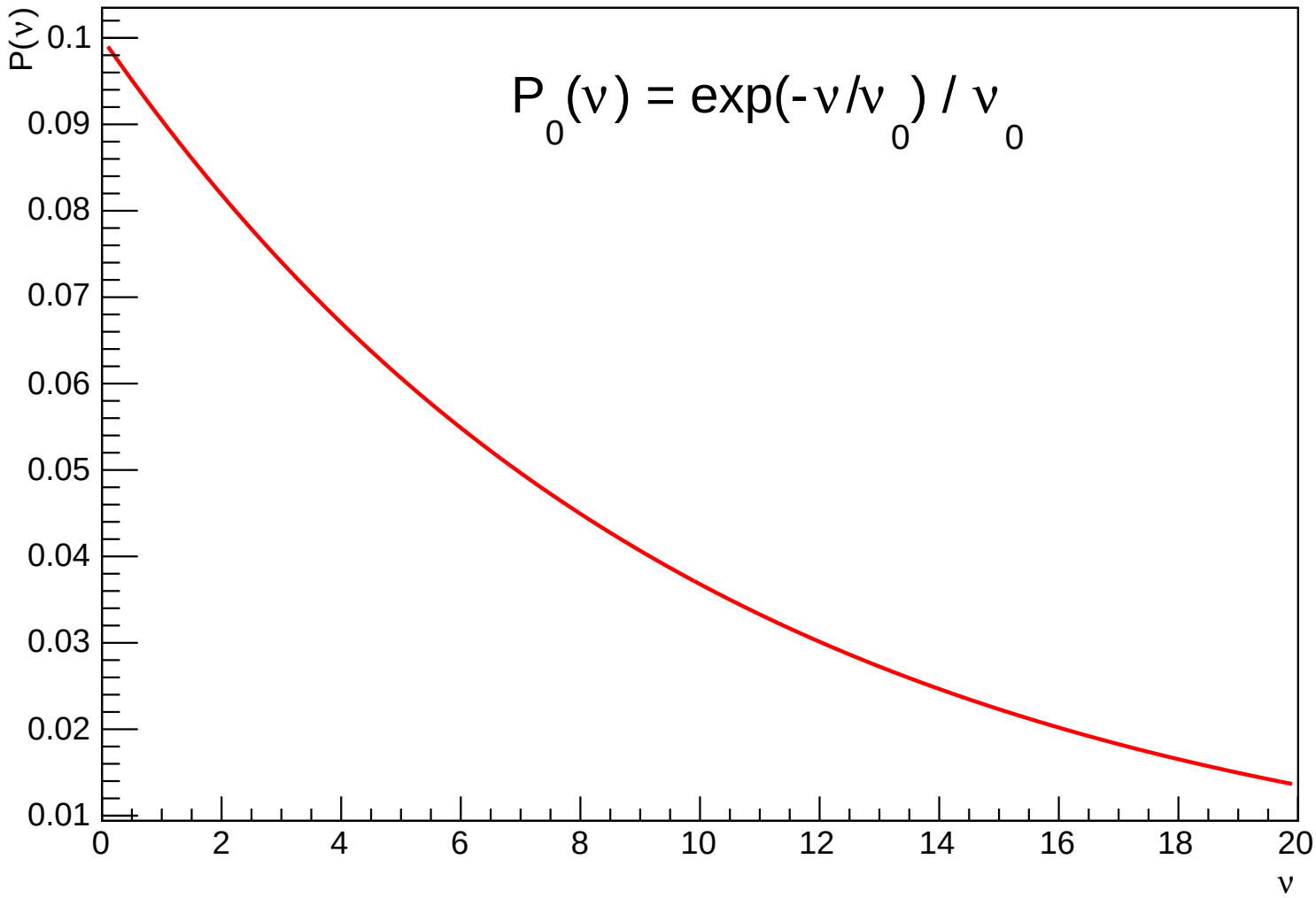
This is normalized:

$$\int_0^\infty \frac{1}{\nu_0} e^{-\nu/\nu_0} \mathrm{d}\nu = 1$$

And has expectation value:

$$E_0[\nu] = \nu_0$$

# Poisson distribution in Bayesian context



# Poisson distribution in Bayesian context

$$P_0(\nu) = \frac{1}{\nu_0} e^{-\nu/\nu_0}$$

So:

$$P(\nu|n) = \frac{\nu^n e^{-\nu \frac{\nu_0+1}{\nu_0}}}{\int_0^\infty \nu^n e^{-\nu \frac{\nu_0+1}{\nu_0}}} = \left(\frac{\nu_0+1}{\nu_0}\right)^{n+1} \frac{\nu^n e^{-\nu \frac{\nu_0+1}{\nu_0}}}{n!}$$

And

$$E[\nu] = (n+1)\left(\frac{\nu_0}{\nu_0+1}\right)$$

Again, let's look at $n = 0$:

$$P(\nu|0) = \left(\frac{\nu_0+1}{\nu_0}\right) e^{-\nu \frac{\nu_0+1}{\nu_0}}$$

and

$$F(\nu|0) = 1 - e^{-\nu \frac{\nu_0+1}{\nu_0}}$$

so

$$\nu_{95} \approx 3 \cdot \frac{\nu_0}{\nu_0+1} < \nu_{95}^{\text{flat}}$$

Why is it less than the upper limit with the flat prior?

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\mathrm{S}},\nu_{\mathrm{B}}) = \sum_{n_{\mathrm{S}}=0}^{n} P(n_{\mathrm{S}}|\nu_{\mathrm{S}})P(n-n_{\mathrm{S}}|\nu_{\mathrm{B}})$$

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\mathrm{S}},\nu_{\mathrm{B}}) = \sum_{n_{\mathrm{S}}=0}^{n} P(n_{\mathrm{S}}|\nu_{\mathrm{S}})P(n-n_{\mathrm{S}}|\nu_{\mathrm{B}})$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \sum_{n_{\mathrm{S}}=0}^{n} \frac{\nu_{\mathrm{S}}^{n_{\mathrm{S}}}\nu_{\mathrm{B}}^{n_{\mathrm{B}}}}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}$$

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\mathrm{S}},\nu_{\mathrm{B}}) = \sum_{n_{\mathrm{S}}=0}^{n} P(n_{\mathrm{S}}|\nu_{\mathrm{S}})P(n-n_{\mathrm{S}}|\nu_{\mathrm{B}})$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \sum_{n_{\mathrm{S}}=0}^{n} \frac{\nu_{\mathrm{S}}^{n_{\mathrm{S}}}\nu_{\mathrm{B}}^{n_{\mathrm{B}}}}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}$$

$$\text{pull out a factor of } \frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^{n}}{n!}$$

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\mathrm{S}},\nu_{\mathrm{B}}) = \sum_{n_{\mathrm{S}}=0}^{n} P(n_{\mathrm{S}}|\nu_{\mathrm{S}})P(n-n_{\mathrm{S}}|\nu_{\mathrm{B}})$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \sum_{n_{\mathrm{S}}=0}^{n} \frac{\nu_{\mathrm{S}}^{n_{\mathrm{S}}}\nu_{\mathrm{B}}^{n_{\mathrm{B}}}}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}$$

$$\text{pull out a factor of } \frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!}$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!} \underbrace{\sum_{n_{\mathrm{S}}=0}^{n} \frac{n!}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!} \left(\frac{\nu_{\mathrm{S}}}{\nu_{\mathrm{S}}+\nu_{\mathrm{B}}}\right)^{n_{\mathrm{S}}} \left(\frac{\nu_{\mathrm{B}}}{\nu_{\mathrm{S}}+\nu_{\mathrm{B}}}\right)^{n-n_{\mathrm{S}}}}$$

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\mathrm{S}},\nu_{\mathrm{B}}) = \sum_{n_{\mathrm{S}}=0}^{n} P(n_{\mathrm{S}}|\nu_{\mathrm{S}})P(n-n_{\mathrm{S}}|\nu_{\mathrm{B}})$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \sum_{n_{\mathrm{S}}=0}^{n} \frac{\nu_{\mathrm{S}}^{n_{\mathrm{S}}}\nu_{\mathrm{B}}^{n_{\mathrm{B}}}}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}$$

$$\text{pull out a factor of } \frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!}$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})}\frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!} \underbrace{\sum_{n_{\mathrm{S}}=0}^{n} \frac{n!}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}\left(\frac{\nu_{\mathrm{S}}}{\nu_{\mathrm{S}}+\nu_{\mathrm{B}}}\right)^{n_{\mathrm{S}}}\left(\frac{\nu_{\mathrm{B}}}{\nu_{\mathrm{S}}+\nu_{\mathrm{B}}}\right)^{n-n_{\mathrm{S}}}}_{\sum_{n_{\mathrm{S}}=0}^{n}\binom{n}{n_{\mathrm{S}}}p^{n_{\mathrm{S}}}(1-p)^{n-n_{\mathrm{S}}}=1}$$

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\rm S},\nu_{\rm B}) = \sum_{n_{\rm S}=0}^{n} P(n_{\rm S}|\nu_{\rm S})P(n-n_{\rm S}|\nu_{\rm B})$$

$$= e^{-(\nu_{\rm S}+\nu_{\rm B})} \sum_{n_{\rm S}=0}^{n} \frac{\nu_{\rm S}^{n_{\rm S}}\nu_{\rm B}^{n_{\rm B}}}{n_{\rm S}!(n-n_{\rm S})!}$$

$$\text{pull out a factor of } \frac{(\nu_{\rm S}+\nu_{\rm B})^n}{n!}$$

$$= e^{-(\nu_{\rm S}+\nu_{\rm B})} \frac{(\nu_{\rm S}+\nu_{\rm B})^n}{n!} \underbrace{\sum_{n_{\rm S}=0}^{n} \frac{n!}{n_{\rm S}!(n-n_{\rm S})!} \left(\frac{\nu_{\rm S}}{\nu_{\rm S}+\nu_{\rm B}}\right)^{n_{\rm S}} \left(\frac{\nu_{\rm B}}{\nu_{\rm S}+\nu_{\rm B}}\right)^{n-n_{\rm S}}}_{\sum_{n_{\rm S}=0}^{n} \binom{n}{n_{\rm S}} p^{n_{\rm S}}(1-p)^{n-n_{\rm S}}=1}$$

$$= e^{-(\nu_{\rm S}+\nu_{\rm B})} \frac{(\nu_{\rm S}+\nu_{\rm B})^n}{n!}$$

# Superposition of Poisson distributions

We often have several Poisson processes contributing signals.

Basic scenario: signal (S) and background (B)

Then

$$P(n|\nu_{\mathrm{S}},\nu_{\mathrm{B}}) = \sum_{n_{\mathrm{S}}=0}^{n} P(n_{\mathrm{S}}|\nu_{\mathrm{S}})P(n-n_{\mathrm{S}}|\nu_{\mathrm{B}})$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \sum_{n_{\mathrm{S}}=0}^{n} \frac{\nu_{\mathrm{S}}^{n_{\mathrm{S}}}\nu_{\mathrm{B}}^{n_{\mathrm{B}}}}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}$$

$$\text{pull out a factor of } \frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!}$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})}\frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!} \underbrace{\sum_{n_{\mathrm{S}}=0}^{n} \frac{n!}{n_{\mathrm{S}}!(n-n_{\mathrm{S}})!}\left(\frac{\nu_{\mathrm{S}}}{\nu_{\mathrm{S}}+\nu_{\mathrm{B}}}\right)^{n_{\mathrm{S}}}\left(\frac{\nu_{\mathrm{B}}}{\nu_{\mathrm{S}}+\nu_{\mathrm{B}}}\right)^{n-n_{\mathrm{S}}}}_{\sum_{n_{\mathrm{S}}=0}^{n}\binom{n}{n_{\mathrm{S}}}p^{n_{\mathrm{S}}}(1-p)^{n-n_{\mathrm{S}}}=1}$$

$$= e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})}\frac{(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})^n}{n!} = P(n|\nu_{\mathrm{S}}+\nu_{\mathrm{B}})$$

# Superposition of Poisson distributions in Bayesian context

Let's take the prior on the signal flat:

$$P(\nu_{\rm S}) = \text{constant}$$

And look at when the background is **perfectly known**:

prior for background is a delta function:

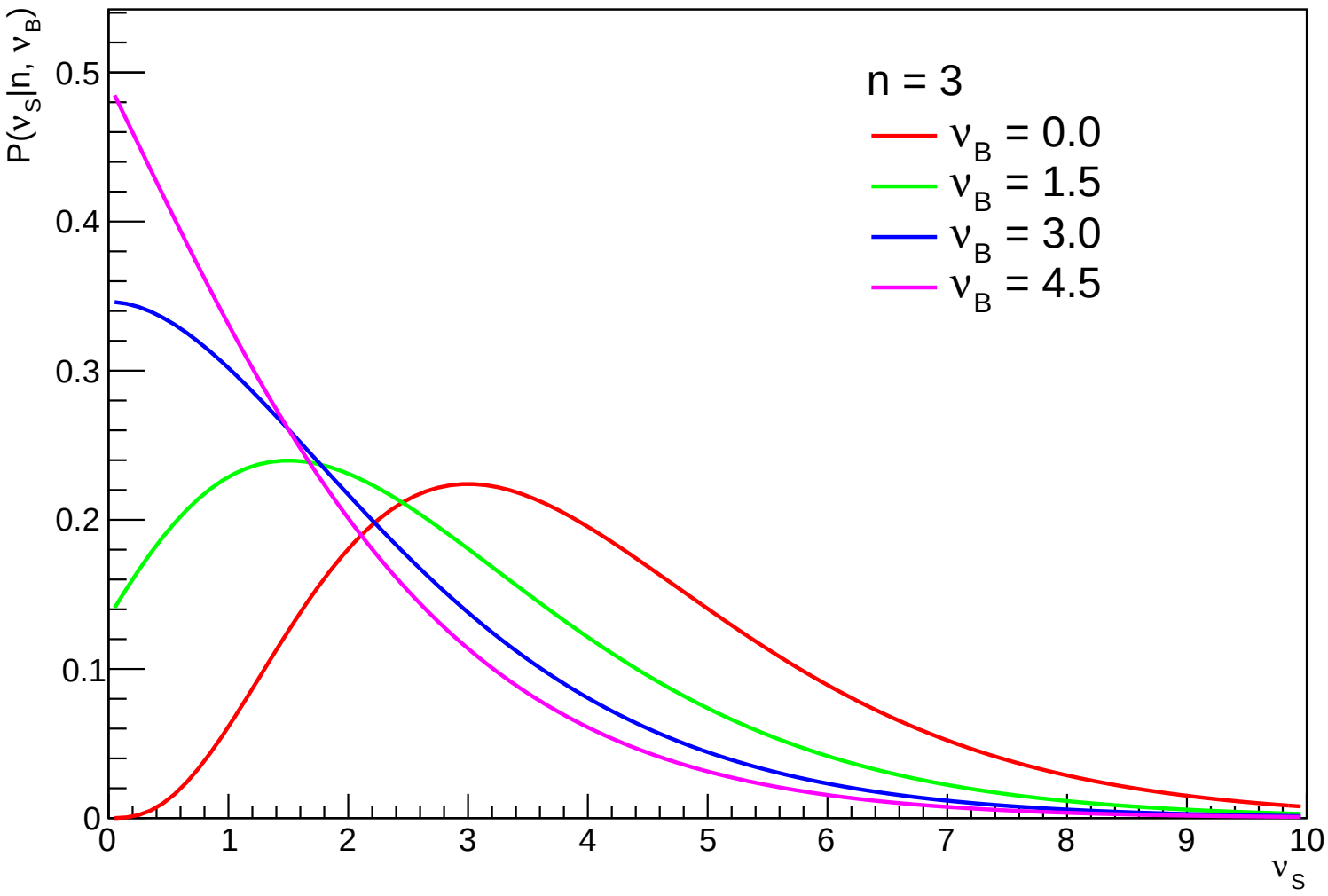
$$P_0(\nu) = \delta(\nu_{\rm B} - \nu_{\rm B}^*)$$

and so the posterior for $\nu_{\rm S}$ is

$$P(\nu_{\rm S}|n,\nu_{\rm B}^*) = \frac{e^{-(\nu_{\rm S}+\nu_{\rm B})}(\nu_{\rm S}+\nu_{\rm B}^*)^n}{\int_0^\infty e^{-(\nu'_{\rm S}+\nu_{\rm B})}(\nu'_{\rm S}+\nu_{\rm B}^*)^n {\rm d}\nu'_{\rm S}} = \frac{e^{-\nu_{\rm S}}(\nu_{\rm S}+\nu_{\rm B}^*)^n}{n!} \Big/ \sum_{m=0}^{n} \frac{\nu_{\rm B}^m}{m!}$$

# Superposition of Poisson distributions in Bayesian context

$$P(\nu_{\mathrm{S}}) = \text{constant}, \qquad P_0(\nu) = \delta(\nu_{\mathrm{B}} - \nu_{\mathrm{B}}^*)$$



# Superposition of Poisson distributions in Bayesian context

$$P(\nu_{\mathrm{S}}) = \text{constant}, \qquad P_0(\nu) = \delta(\nu_{\mathrm{B}} - \nu_{\mathrm{B}}^*)$$

$$P(\nu_{\mathrm{S}}|n, \nu_{\mathrm{B}}^*) = \frac{e^{-\nu_{\mathrm{S}}}(\nu_{\mathrm{S}} + \nu_{\mathrm{B}}^*)^n}{n!} \Bigg/ \sum_{m=0}^{n} \frac{\nu_{\mathrm{B}}^m}{m!}$$

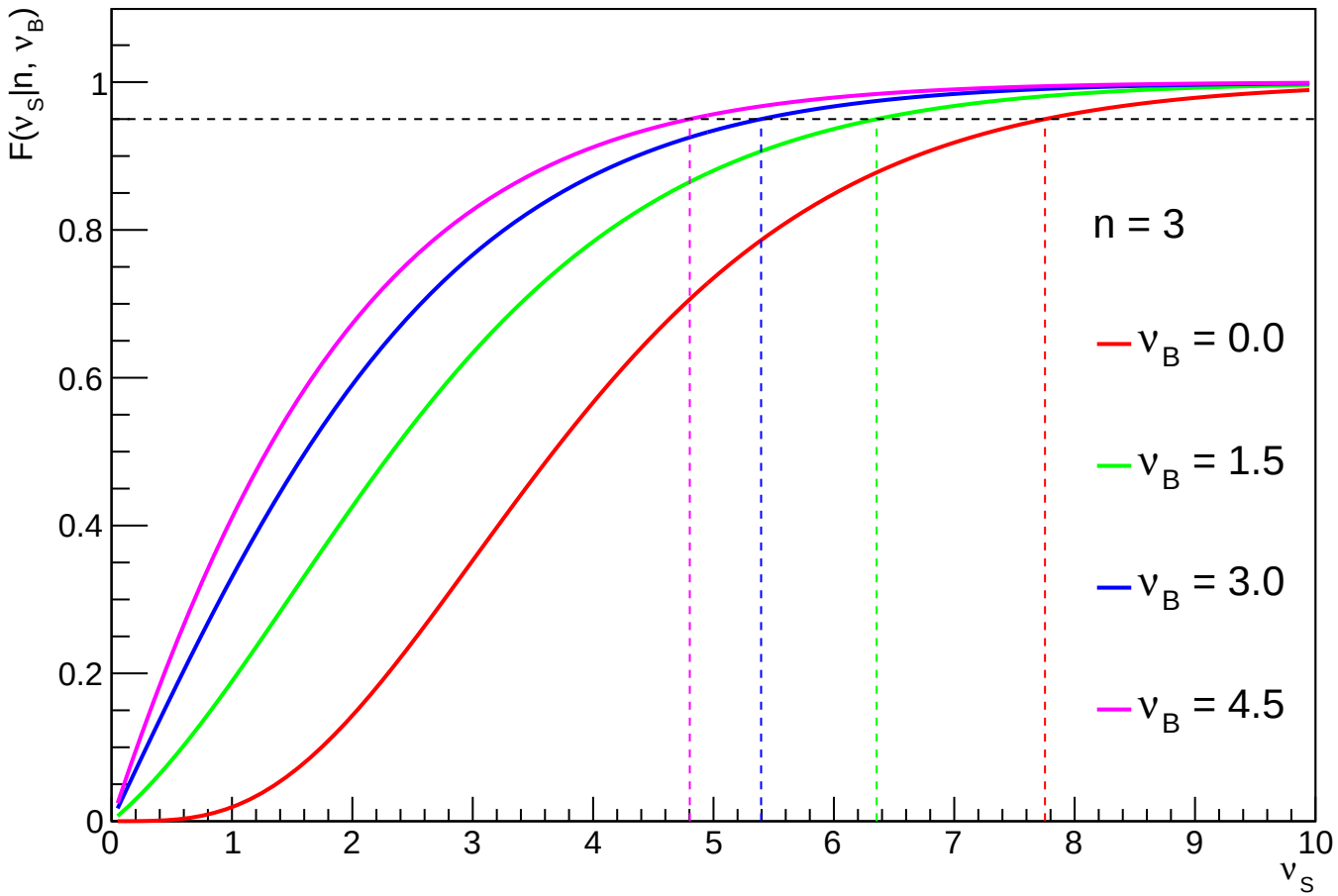
So the cumulative is:

$$F(\nu_{\mathrm{S}}|n, \nu_{\mathrm{B}}^*) = 1 - e^{-\nu_{\mathrm{S}}} \sum_{m=0}^{n} \frac{(\nu_{\mathrm{S}} + \nu_{\mathrm{B}})^m}{m!} \Bigg/ \sum_{m=0}^{n} \frac{\nu_{\mathrm{B}}^m}{m!}$$

# Superposition of Poisson distributions in Bayesian context

$$P(\nu_S) = \text{constant}, \qquad P_0(\nu) = \delta(\nu_B - \nu_B^*)$$

$$F(\nu_S | n, \nu_B^*) = 1 - e^{-\nu_S} \sum_{m=0}^{n} \frac{(\nu_S + \nu_B)^m}{m!} \Bigg/ \sum_{m=0}^{n} \frac{\nu_B^m}{m!}$$



# Superposition of Poisson distributions in Bayesian context

Now let's look at when we we have imprecise knowledge of the background:

$$P_0(\nu_{\mathrm{B}}) = \mathcal{N}(\nu_{\mathrm{B}}|\nu_{\mathrm{B}}^*, \sigma_{\mathrm{B}}^2) \cdot P_0(\nu_{\mathrm{B}}^*, \sigma_{\mathrm{B}}^2)$$

with delta priors for the parameters $\nu_{\mathrm{B}}^*$ and $\sigma_{\mathrm{B}}^2$; and again a flat prior for the signal:

$$P_0(\nu_{\mathrm{S}}) = \text{constant}$$

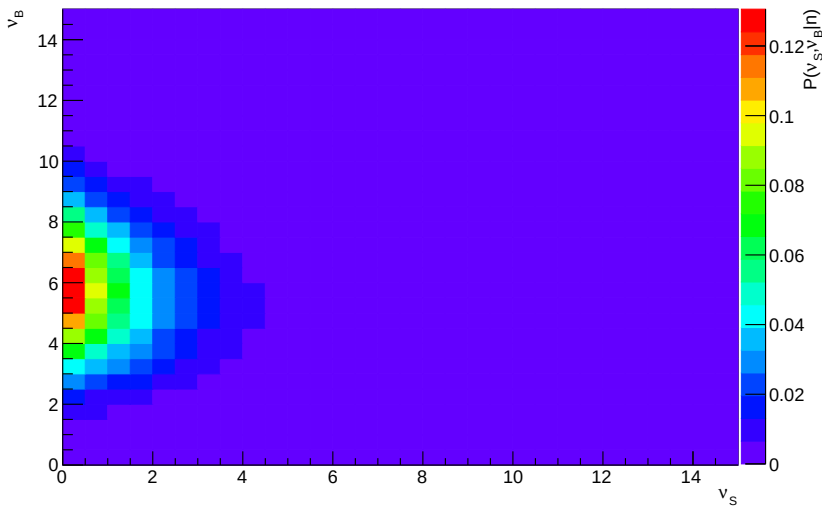
our posterior is

$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}}|n; \nu_{\mathrm{B}}^*, \sigma_{\mathrm{B}}^2) \propto P(n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) \cdot \mathcal{N}(\nu_{\mathrm{B}}|\nu_{\mathrm{B}}^*, \sigma_{\mathrm{B}}^2)$$

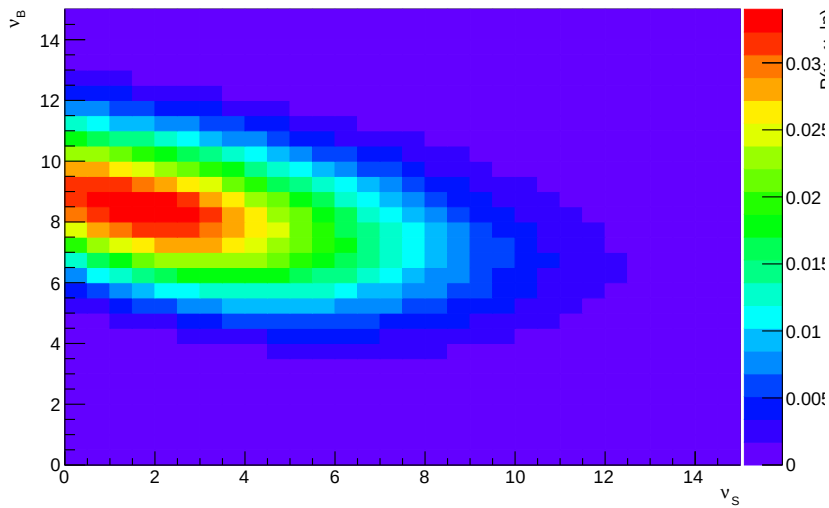
# Superposition of Poisson distributions in Bayesian context

$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}} | n; \nu_{\mathrm{B}}^{*}, \sigma_{\mathrm{B}}^{2}) \propto P(n | \nu_{\mathrm{S}}, \nu_{\mathrm{B}}) \cdot \mathcal{N}(\nu_{\mathrm{B}} | \nu_{\mathrm{B}}^{*}, \sigma_{\mathrm{B}}^{2})$$

n = 2, $\nu_B^*$ = 8.5, $\sigma_B$ = 2.0

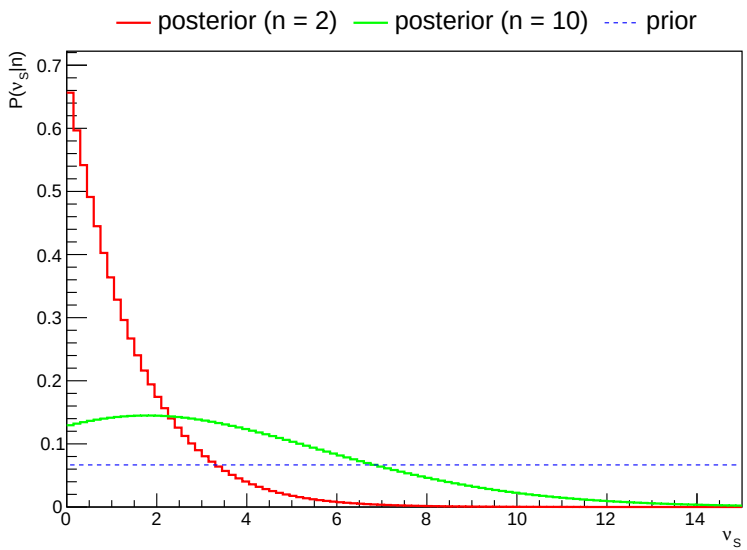


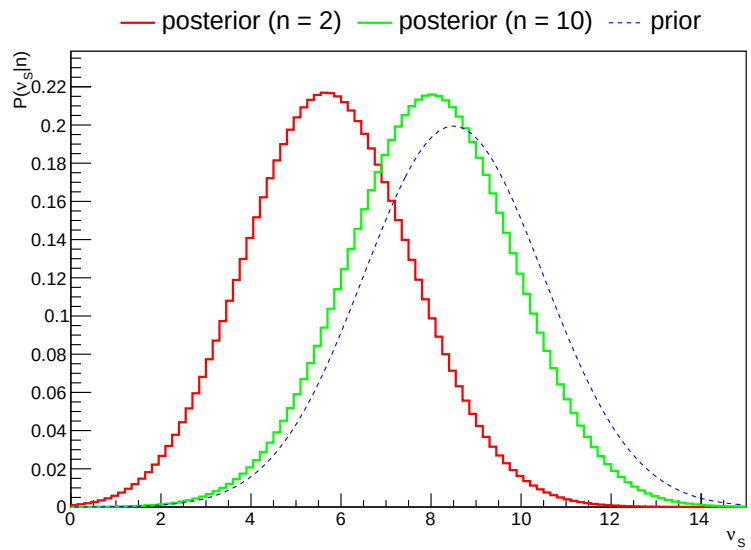

n = 10, $\nu_B^*$ = 8.5, $\sigma_B$ = 2.0



# Superposition of Poisson distributions in Bayesian context

$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}} | n; \nu_{\mathrm{B}}^{*}, \sigma_{\mathrm{B}}^{2}) \propto P(n | \nu_{\mathrm{S}}, \nu_{\mathrm{B}}) \cdot \mathcal{N}(\nu_{\mathrm{B}} | \nu_{\mathrm{B}}^{*}, \sigma_{\mathrm{B}}^{2})$$



# Superposition of Poisson distributions in Bayesian context

Now let us suppose we have two **independent** pieces of data:

1. $b$ : data containing background only
2. $n$ : data containing background and signal.

# Superposition of Poisson distributions in Bayesian context

Now let us suppose we have two **independent** pieces of data:

1. $b$ : data containing background only
2. $n$ : data containing background and signal.

And correspondingly, two **independent** probabilities:

1. $P(b|\nu_{\mathrm{B}}) =$
2. $P(n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) =$

# Superposition of Poisson distributions in Bayesian context

Now let us suppose we have two **independent** pieces of data:

1. $b$ : data containing background only
2. $n$ : data containing background and signal.

And correspondingly, two **independent** probabilities:

1. $P(b|\nu_{\mathrm{B}}) = \nu_{\mathrm{B}}^{b} e^{-\nu_{\mathrm{B}}} / b!$
2. $P(n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) =$

# Superposition of Poisson distributions in Bayesian context

Now let us suppose we have two **independent** pieces of data:

1. $b$ : data containing background only
2. $n$ : data containing background and signal.

And correspondingly, two **independent** probabilities:

1. $P(b|\nu_{\mathrm{B}}) = \nu_{\mathrm{B}}^{b} e^{-\nu_{\mathrm{B}}} \big/ b!$
2. $P(n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) = P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}}) = (\nu_{\mathrm{S}} + \nu_{\mathrm{B}})^{n} e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \Big/ n!$

# Superposition of Poisson distributions in Bayesian context

Now let us suppose we have two **independent** pieces of data:

1. $b$ : data containing background only
2. $n$ : data containing background and signal.

And correspondingly, two **independent** probabilities:

1. $P(b|\nu_{\mathrm{B}}) = \nu_{\mathrm{B}}^{b} e^{-\nu_{\mathrm{B}}} \big/ b!$
2. $P(n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) = P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}}) = (\nu_{\mathrm{S}} + \nu_{\mathrm{B}})^{n} e^{-(\nu_{\mathrm{S}}+\nu_{\mathrm{B}})} \Big/ n!$

So:

$$P(b, n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) = P(b|\nu_{\mathrm{B}}) P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}})$$

# Superposition of Poisson distributions in Bayesian context

So:

$$P(b, n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) = P(b|\nu_{\mathrm{B}})P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}})$$

So our posterior is:

$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}}|b, n) = \frac{P(b|\nu_{\mathrm{B}})P(n|\nu_{\mathrm{S}}\nu_{\mathrm{B}})P_0(\nu_{\mathrm{S}})P_0(\nu_{\mathrm{B}})}{\underbrace{\int_0^\infty \int_0^\infty P(b|\nu'_{\mathrm{B}})P(n|\nu'_{\mathrm{S}}\nu'_{\mathrm{B}})P_0(\nu'_{\mathrm{S}})P_0(\nu'_{\mathrm{B}})\mathrm{d}\nu'_{\mathrm{S}}\mathrm{d}\nu'_{\mathrm{B}}}_{\text{numerically solveable}}}$$

# Superposition of Poisson distributions in Bayesian context

So:

$$P(b, n|\nu_{\mathrm{S}}, \nu_{\mathrm{B}}) = P(b|\nu_{\mathrm{B}})P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}})$$

So our posterior is:

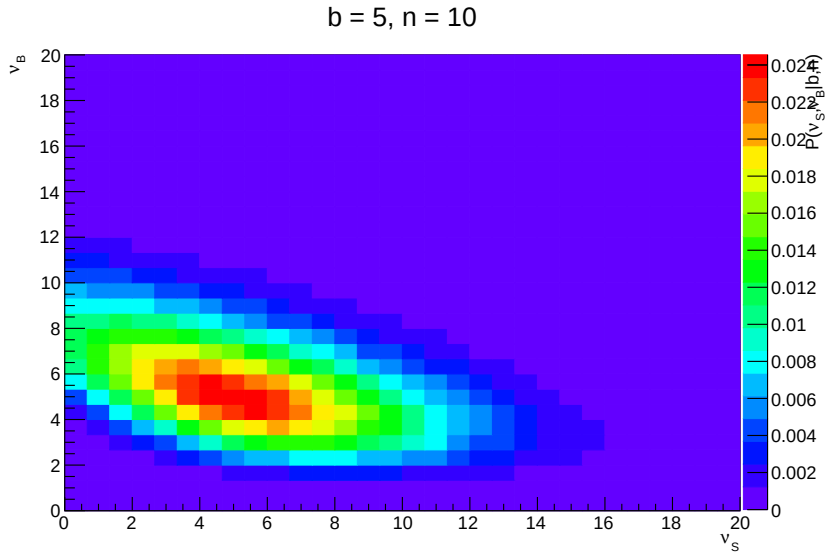
$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}}|b, n) = \frac{P(b|\nu_{\mathrm{B}})P(n|\nu_{\mathrm{S}}\nu_{\mathrm{B}})P_0(\nu_{\mathrm{S}})P_0(\nu_{\mathrm{B}})}{\underbrace{\int_0^\infty \int_0^\infty P(b|\nu'_{\mathrm{B}})P(n|\nu'_{\mathrm{S}}\nu'_{\mathrm{B}})P_0(\nu'_{\mathrm{S}})P_0(\nu'_{\mathrm{B}})\mathrm{d}\nu'_{\mathrm{S}}\mathrm{d}\nu'_{\mathrm{B}}}_{\text{numerically solveable}}}$$

But we can rely on the proportionality (with flat priors here)

$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}}|b, n) \propto \nu_{\mathrm{B}}^b(\nu_{\mathrm{S}} + \nu_{\mathrm{B}})^n e^{-(\nu_{\mathrm{S}}+2\nu_{\mathrm{B}})} \Big/ b!n!$$

# Superposition of Poisson distributions in Bayesian context

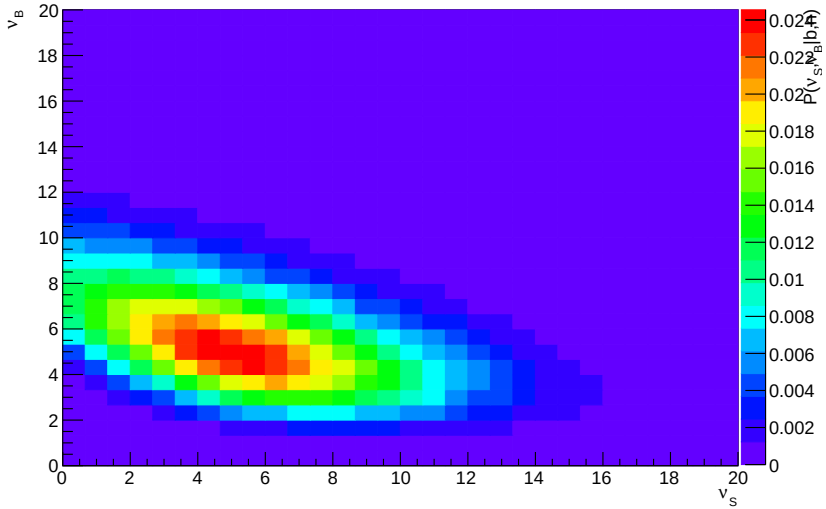
$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}}|b, n) \propto P(b|\nu_{\mathrm{B}})P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}})$$



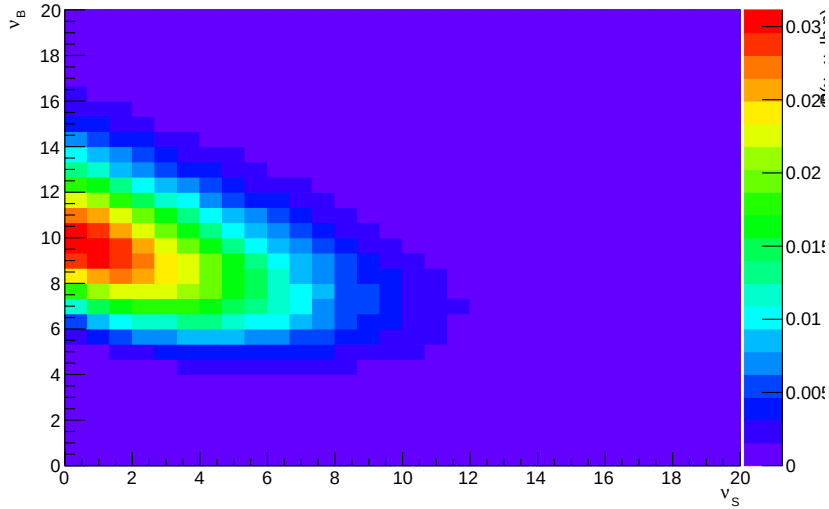
# Superposition of Poisson distributions in Bayesian context

$$P(\nu_S, \nu_B | b, n) \propto P(b|\nu_B) P(n|\nu_S + \nu_B)$$

b = 5, n = 10



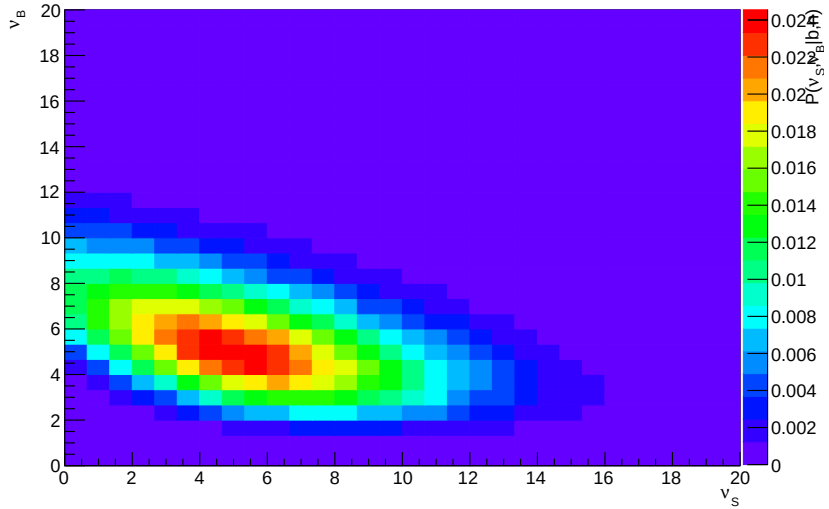

b = 10, n = 10



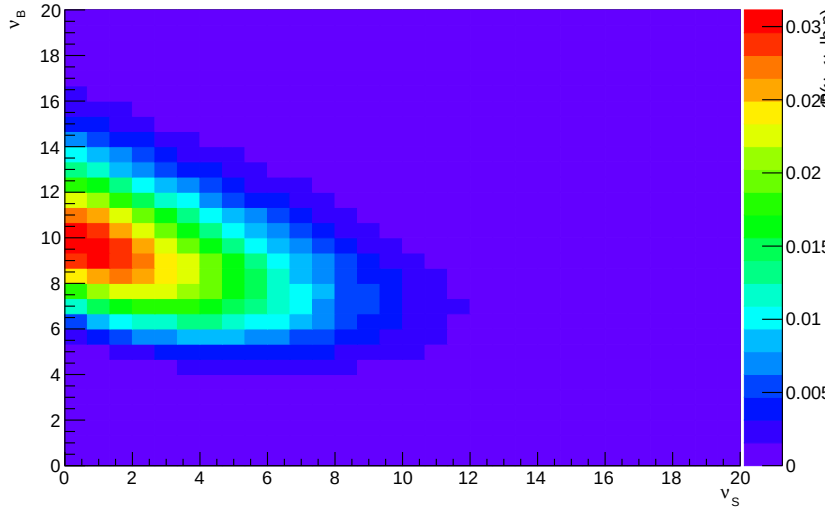
# Superposition of Poisson distributions in Bayesian context

$$P(\nu_S, \nu_B | b, n) \propto P(b|\nu_B) P(n|\nu_S + \nu_B)$$

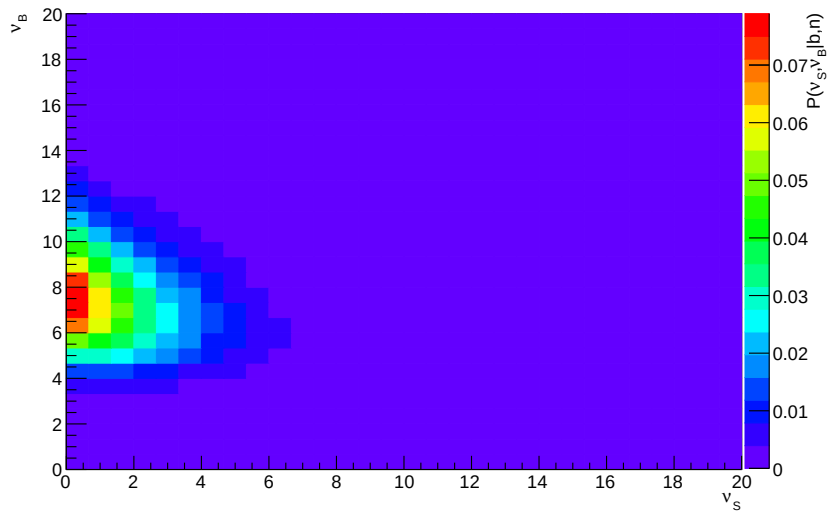
b = 5, n = 10




b = 10, n = 10




b = 10, n = 5

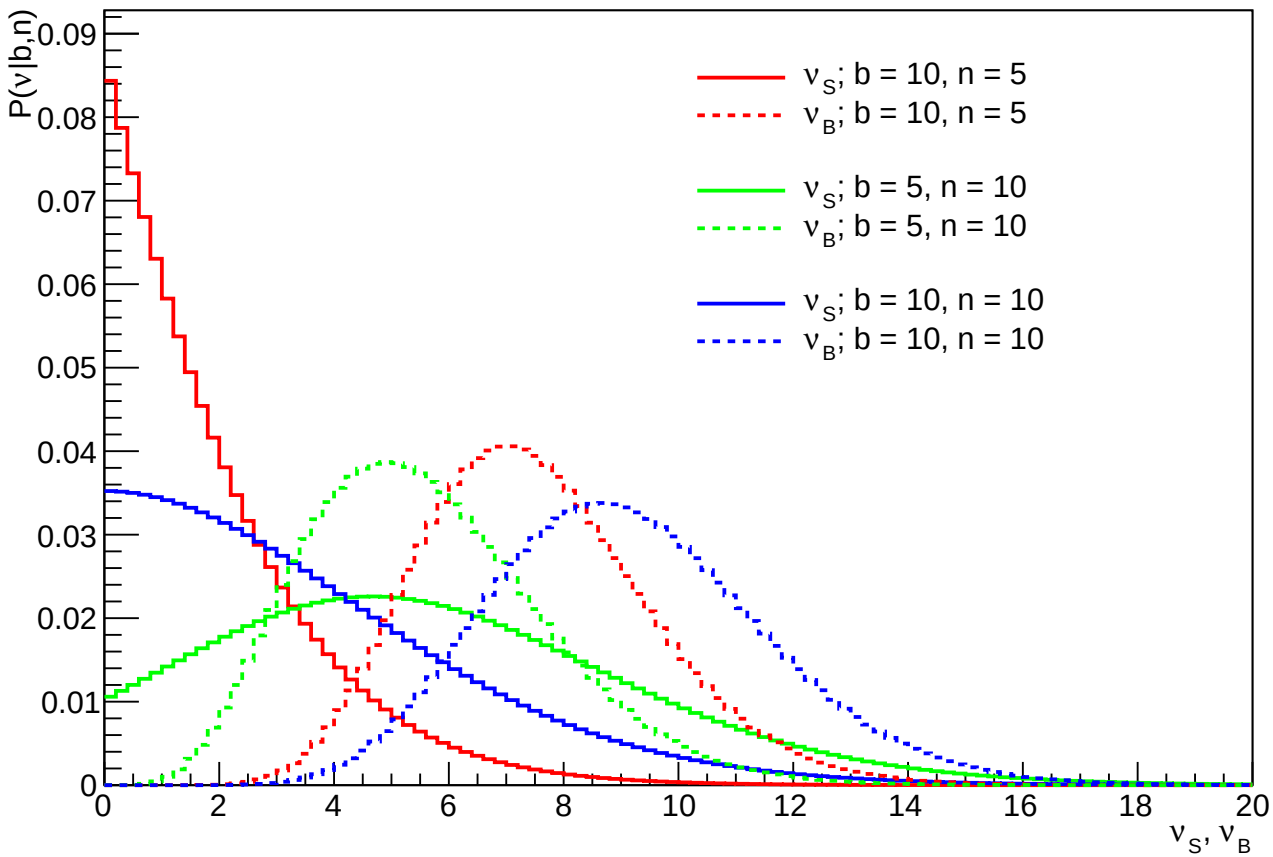


# Superposition of Poisson distributions in Bayesian context

$$P(\nu_S, \nu_B | b, n) \propto P(b|\nu_B) P(n|\nu_S + \nu_B)$$

we can **marginalize** the 2D dist. to get information about the 1D dists:

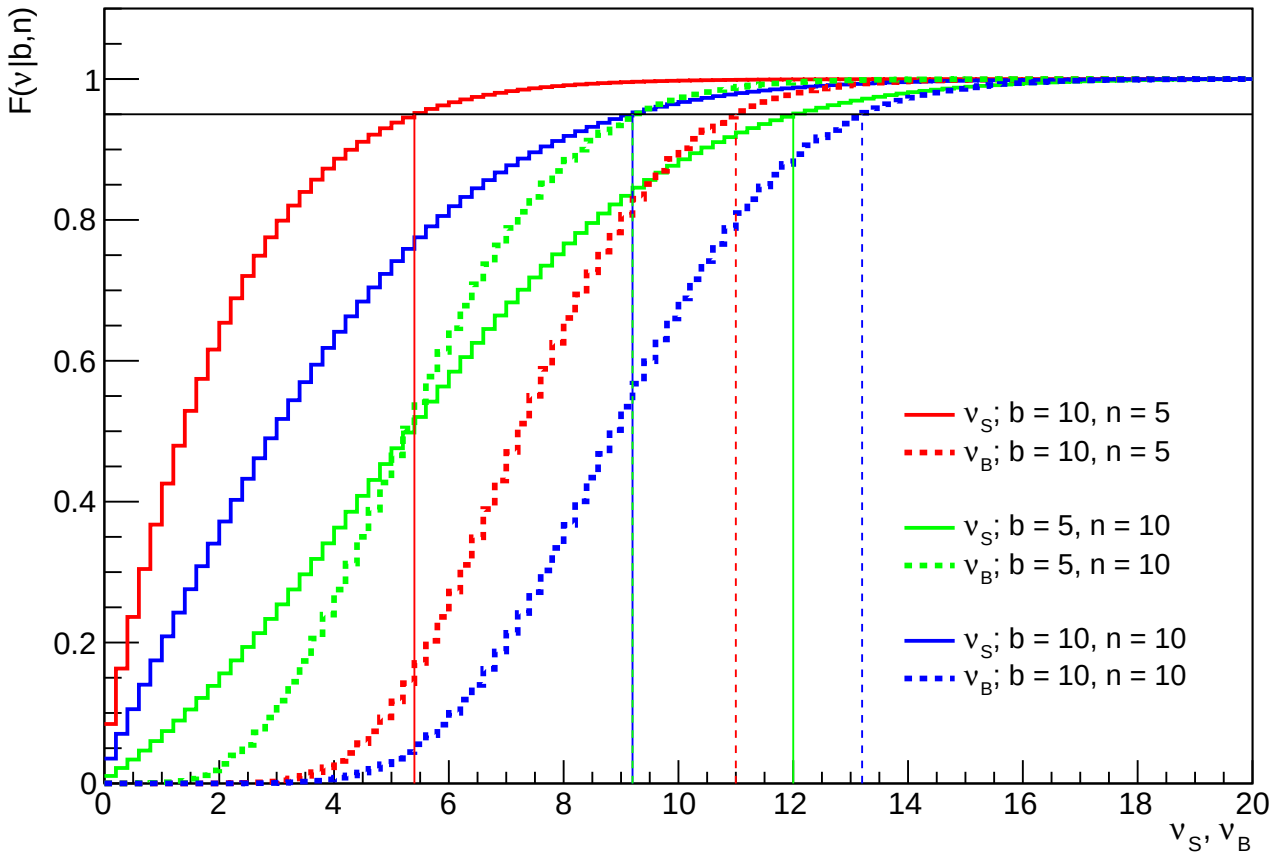
$$P(\nu_S|b,n) = \int_0^\infty P(\nu_S, \nu_B|n,b)\mathrm{d}\nu_B, \qquad P(\nu_B|b,n) = \int_0^\infty P(\nu_S, \nu_B|n,b)\mathrm{d}\nu_S$$



# Superposition of Poisson distributions in Bayesian context

$$P(\nu_{\mathrm{S}}, \nu_{\mathrm{B}}|b, n) \propto P(b|\nu_{\mathrm{B}})P(n|\nu_{\mathrm{S}} + \nu_{\mathrm{B}})$$

And can then calculate the cumulatives:



# Bayes factor

In the last example we had two measurements: $b$ and $n$

And we presumed:

1. $b$ was measured with only background contributing
2. $n$ was measured with background and signal constributing

**This is a model!**

# Bayes factor

In the last example we had two measurements: $b$ and $n$

And we presumed:

1. $b$ was measured with only background contributing
2. $n$ was measured with background and signal constributing

**This is a model!**

Another valid model is:

1. $b$ was measured with only background contributing
2. $n$ was also measured with only background contributing

How can we compare them?

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$, each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) =$$

$$Z_2 = P(D|M_2) =$$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) = \qquad P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1)$$

$$Z_2 = P(D|M_2) =$$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) = \int P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1) \mathrm{d}\vec{\lambda}_1$$

$$Z_2 = P(D|M_2) =$$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) = \int P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1) \mathrm{d}\vec{\lambda}_1$$

$$Z_2 = P(D|M_2) = \int P(D|\vec{\lambda}_2; M_2) P_0(\vec{\lambda}_2|M_2) \mathrm{d}\vec{\lambda}_2$$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) = \int P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1) \mathrm{d}\vec{\lambda}_1$$

$$Z_2 = P(D|M_2) = \int P(D|\vec{\lambda}_2; M_2) P_0(\vec{\lambda}_2|M_2) \mathrm{d}\vec{\lambda}_2$$

And (total) marginalization:

$$P(M_1|D) =$$

$$P(M_2|D) =$$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) = \int P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1) \mathrm{d}\vec{\lambda}_1$$

$$Z_2 = P(D|M_2) = \int P(D|\vec{\lambda}_2; M_2) P_0(\vec{\lambda}_2|M_2) \mathrm{d}\vec{\lambda}_2$$

And (total) marginalization:

$$P(M_1|D) = \int P(\vec{\lambda}_1; M_1|D) \mathrm{d}\vec{\lambda}_1$$

$$P(M_2|D) =$$

# Bayes Factor

Again, turn to Bayes' theorem:

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

Recall the evidence (now for each model separately):

$$Z_1 = P(D|M_1) = \int P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1) \mathrm{d}\vec{\lambda}_1$$

$$Z_2 = P(D|M_2) = \int P(D|\vec{\lambda}_2; M_2) P_0(\vec{\lambda}_2|M_2) \mathrm{d}\vec{\lambda}_2$$

And (total) marginalization:

$$P(M_1|D) = \int P(\vec{\lambda}_1; M_1|D) \mathrm{d}\vec{\lambda}_1$$

$$P(M_2|D) = \int P(\vec{\lambda}_2; M_2|D) \mathrm{d}\vec{\lambda}_2$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$, each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

The posteriors are:

$$P(M_1|D) =$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

The posteriors are:

$$P(M_1|D) = \int \frac{P(D|\vec{\lambda}_1; M_1)P_0(\vec{\lambda}_1|M_1)P_0(M_1)}{} \mathrm{d}\vec{\lambda}_1$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

The posteriors are:

$$P(M_1|D) = \int \frac{P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_1) P_0(M_1)}{Z_1 P_0(M_1) + Z_2 P_0(M_2)} \mathrm{d}\vec{\lambda}_1$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

The posteriors are:

$$P(M_1|D) = \int \frac{P(D|\vec{\lambda}_1; M_1)P_0(\vec{\lambda}_1|M_1)P_0(M_1)}{Z_1 P_0(M_1) + Z_2 P_0(M_2)} \mathrm{d}\vec{\lambda}_1$$

$$= \frac{P_0(M_1) \overbrace{\int P(D|\vec{\lambda}_1; M_1)P_0(\vec{\lambda}_1|M_1)\mathrm{d}\vec{\lambda}_1}^{Z_1}}{Z_1 P_0(M_1) + Z_2 P_0(M_2)}$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

The posteriors are:

$$P(M_1|D) = \int \frac{P(D|\vec{\lambda}_1; M_1)P_0(\vec{\lambda}_1|M_1)P_0(M_1)}{Z_1 P_0(M_1) + Z_2 P_0(M_2)} \mathrm{d}\vec{\lambda}_1$$

$$= \frac{P_0(M_1) \overbrace{\int P(D|\vec{\lambda}_1; M_1)P_0(\vec{\lambda}_1|M_1)\mathrm{d}\vec{\lambda}_1}^{Z_1}}{Z_1 P_0(M_1) + Z_2 P_0(M_2)}$$

and similarly

$$P(M_2|D) = \frac{P_0(M_2) \overbrace{\int P(D|\vec{\lambda}_2; M_2)P_0(\vec{\lambda}_2|M_2)\mathrm{d}\vec{\lambda}_2}^{Z_2}}{Z_1 P_0(M_1) + Z_2 P_0(M_2)}$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} =$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

   You must decide this!

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

   You must decide this!

2. The ratio of evidence: $Z_1/Z_2$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

   You must decide this!

2. The ratio of evidence: $Z_1/Z_2$

   This is the **Bayes factor**

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

   You must decide this!

2. The ratio of evidence: $Z_1/Z_2$

   This is the **Bayes factor**

So the Bayes factor is

$$K_{12} \equiv \frac{Z_1}{Z_2}$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

   You must decide this!

2. The ratio of evidence: $Z_1/Z_2$

   This is the **Bayes factor**

So the Bayes factor is

$$K_{12} \equiv \frac{Z_1}{Z_2} = \frac{P(D|M_1)}{P(D|M_2)}$$

# Bayes Factor

Let's say we have two models: $M_1$ and $M_2$,
each with parameters $\vec{\lambda}_1$ and $\vec{\lambda}_2$

To compare the models, we compare their posteriors:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{Z_1}{Z_2}\frac{P_0(M_1)}{P_0(M_2)}$$

the denominators, $Z = Z_1 P_0(M_1) + Z_2 P_0(M_2)$, cancel.

To quarantine the "subjective" assignment of prior beliefs in the models, we factorize this into two terms

1. The ratio of prior beliefs in the two models: $P_0(M_1)/P_0(M_2)$

   You must decide this!

2. The ratio of evidence: $Z_1/Z_2$

   This is the **Bayes factor**

So the Bayes factor is

$$K_{12} \equiv \frac{Z_1}{Z_2} = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\vec{\lambda}_1; M_1) P_0(\vec{\lambda}_1|M_2) \mathrm{d}\vec{\lambda}_1}{\int P(D|\vec{\lambda}_2; M_2) P_0(\vec{\lambda}_2|M_2) \mathrm{d}\vec{\lambda}_2}$$

# Bayes Factor Interlude

Everyon's favorite example:

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

OK . . . What is the equivalent of $6\sigma$ in a Bayes factor?

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

OK . . . What is the equivalent of $6\sigma$ in a Bayes factor?

Let us take the "Jeffreys scale":

| K | Comparative strength |
|---|---|
| $< 1$ | negative |
| 1 to $10^{\frac{1}{2}}$ | not substantial |
| $10^{\frac{1}{2}}$ to $10^{1}$ | substantial |
| $10^{1}$ to $10^{\frac{3}{2}}$ | strong |
| $10^{\frac{3}{2}}$ to $10^{2}$ | very strong |
| $> 100$ | decisive |

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

Let us say $6\sigma$ is at the threshold of "decisive":

$$K = 10^2$$

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

Let us say $6\sigma$ is at the threshold of "decisive":

$$K = 10^2$$

What do we require now to say we believe neutrinos move faster than the speed of light?

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

Let us say $6\sigma$ is at the threshold of "decisive":

$$K = 10^2$$

What do we require now to say we believe neutrinos move faster than the speed of light?

$$\frac{P(\text{super}|D)}{P(\text{sub}|D)} = K * \frac{P_0(\text{super})}{P_0(\text{sub})} > 1$$

# Bayes Factor Interlude

Everyon's favorite example: Superluminal Neutrinos!

Let us presume that a Bayesian analysis produces a Bayes factor equivalent to the original $6\sigma$ result:

$$K = \frac{\text{Neutrinos move faster than the speed of light}}{\text{neutrinos don't move faster than the speed of light}}$$
$$= \text{equivalent of } 6\sigma$$

Let us say $6\sigma$ is at the threshold of "decisive":

$$K = 10^2$$

What do we require now to say we believe neutrinos move faster than the speed of light?

$$\frac{P(\text{super}|D)}{P(\text{sub}|D)} = K * \frac{P_0(\text{super})}{P_0(\text{sub})} > 1$$

So:

$$\frac{P_0(\text{super})}{P_0(\text{sub})} > K^{-1} = 1\%$$

# Bayes Factor

Let us return to our example of two models for the two measurements:

1. "SB" : $b$ contains background only, $n$ contains background and signal
2. "B" : both $b$ and $n$ contain background only

and calculate the evidences

$$Z_B = \int_0^{\nu_\infty} \frac{\nu_\mathrm{B}^b e^{-\nu_\mathrm{B}}}{b!} \cdot \frac{\nu_\mathrm{B}^n e^{-\nu_\mathrm{B}}}{n!} \cdot \frac{1}{\nu_\infty} \mathrm{d}\nu_\mathrm{B}$$

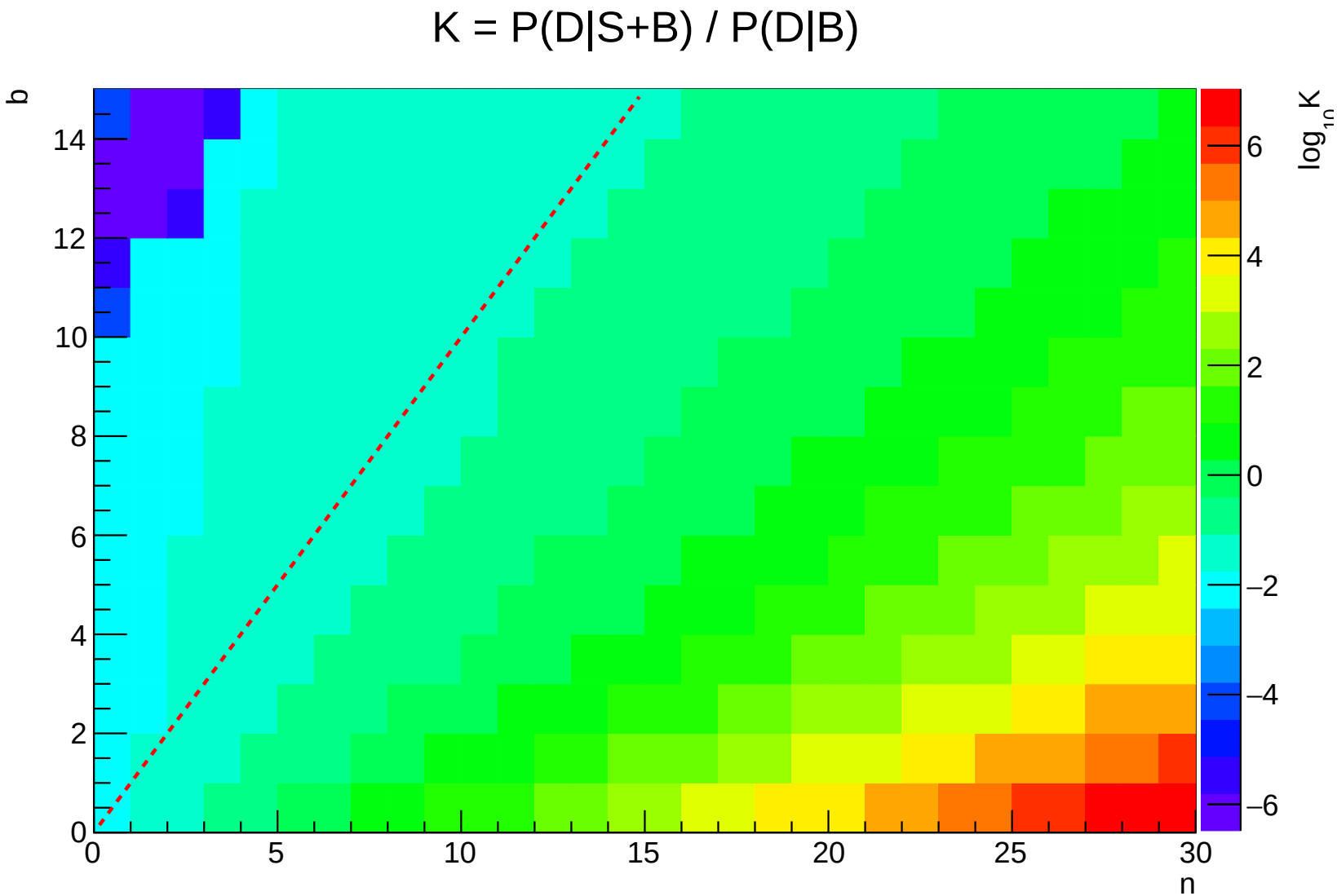
$$Z_{SB} = \int_0^{\nu_\infty} \int_0^{\nu_\infty} \frac{\nu_\mathrm{B}^b e^{-\nu_\mathrm{B}}}{b!} \cdot \frac{(\nu_\mathrm{S} + \nu_\mathrm{B})^n e^{-(\nu_\mathrm{S}+\nu_\mathrm{B})}}{n!} \cdot \frac{1}{\nu_\infty} \cdot \frac{1}{\nu_\infty} \mathrm{d}\nu_\mathrm{B} \mathrm{d}\nu_\mathrm{S}$$

and the Bayes factor:

$$K = \frac{Z_{SB}}{Z_B}$$

We have to do this numercally.

# Bayes factor



# Think carefully about your prior

As we saw in the previous examples, the prior has an impact on the posterior.

Let us look at an example of how to carefully choose the prior.

We measure an energy with a calorimeter, which has a Gaussian resolution:

$$P(E_{\text{meas}}|E_{\text{true}}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(E_{\text{meas}} - E_{\text{true}})^2}{2\sigma^2}\right)$$

We make several measurements from which we want to construct a spectrum.

We need to know $E_{\text{true}}$:

$$P(E_{\text{true}}|E_{\text{meas}}) \propto P(E_{\text{meas}}|E_{\text{true}}) \cdot P_0(E_{\text{true}})$$

We commonly make the implicit assumption of a flat prior:

$$P(E_{\text{true}}|E_{\text{meas}}) = P(E_{\text{meas}}|E_{\text{true}})$$

# Think carefully about your prior

But in many applications, the underlying spectrum is **very much not flat**. For example,

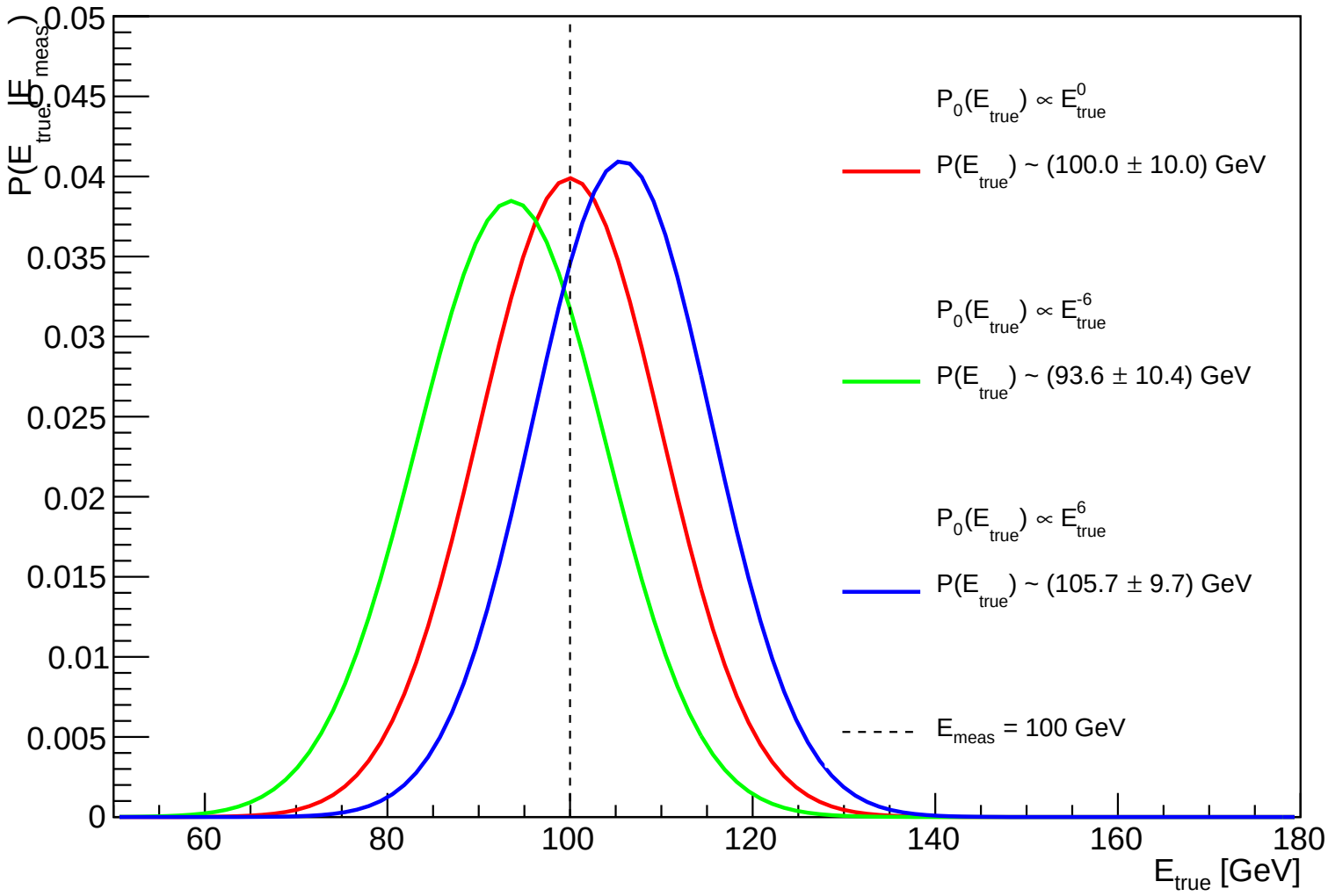
$$f(E_{\text{true}}) \propto E_{\text{true}}^{-6}$$

leading to:

$$P(E_{\text{true}}|E_{\text{meas}}) \propto P(E_{\text{meas}}|E_{\text{true}}) \cdot E_{\text{true}}^{-6}$$

This affects the posterior.

# Think carefully about your prior



# Building a larger model

Now let us extend the previous example to build a larger model:

What we'd probably like to learn about is not $E_{\rm true}$, but the power of the underlying spectrum:

$$f(E_{\rm true}) \propto E^{\lambda}$$

And our data is a series of **independent** measurements:

$$\vec{E}_{\rm meas} = \{E^{(1)}_{\rm meas}, E^{(2)}_{\rm meas}, \dots\}$$

each of which is related to a true value:

$$P(E^{(i)}_{\rm meas}|E^{(i)}_{\rm true}) = \mathcal{N}(E^{(i)}_{\rm meas}|E^{(i)}_{\rm true}, \sigma^2)$$

So now we apply Bayes' theorem:

$$P(\lambda|\vec{E}_{\rm meas}) \propto \left(\prod_i P(E^{(i)}_{\rm meas}|\lambda)\right) P_0(\lambda)$$

But what is $P(E_{\rm meas}|\lambda)$?

# Building a larger model

$$P(\lambda|\vec{E}_{\mathrm{meas}}) \propto \left(\prod_i P(E^{(i)}_{\mathrm{meas}}|\lambda)\right) P_0(\lambda)$$

But what is $P(E_{\mathrm{meas}}|\lambda)$?

We can apply the law of total probability to get it:

$$P(E^{(i)}_{\mathrm{meas}}|\lambda) = \int P(E^{(i)}_{\mathrm{meas}}|E_{\mathrm{true}})P(E_{\mathrm{true}}|\lambda)\mathrm{d}E_{\mathrm{true}}$$

And our full posterior is

$$P(\lambda|\vec{E}_{\mathrm{meas}}) \propto \left(\prod_i \int \mathcal{N}(E^{(i)}_{\mathrm{meas}}|E_{\mathrm{true}}, \sigma^2)E^{\lambda}_{\mathrm{true}}\mathrm{d}E_{\mathrm{true}}\right) P_0(\lambda)$$

Simplest approach: numerically integrate the term in the product during parameter scan.

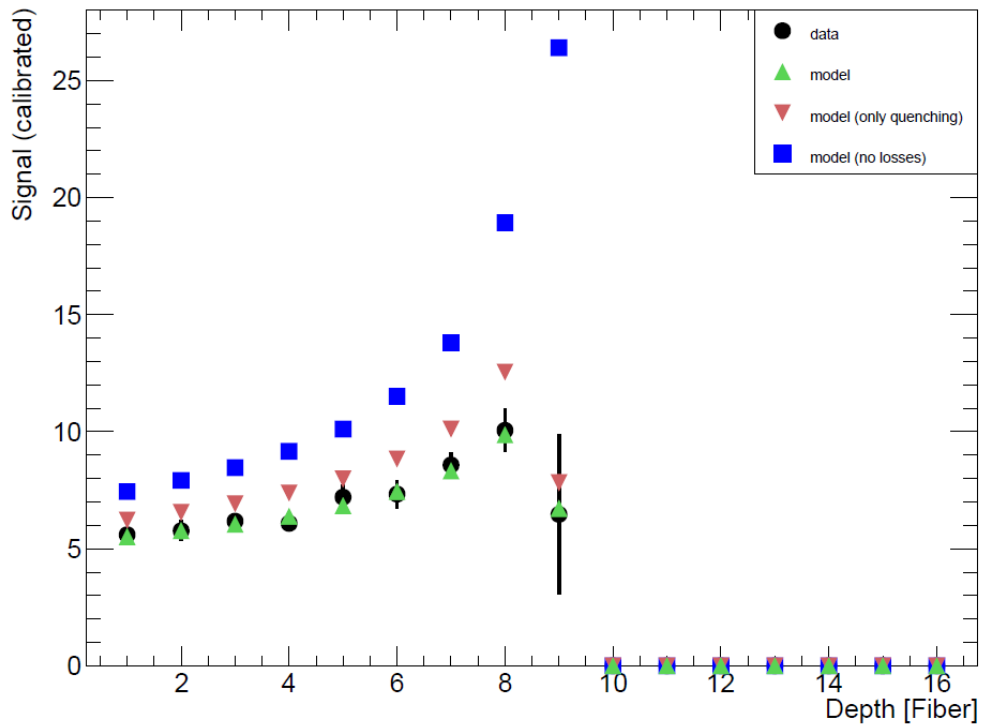
(This can be expensive if the data set is very large.)

# Models

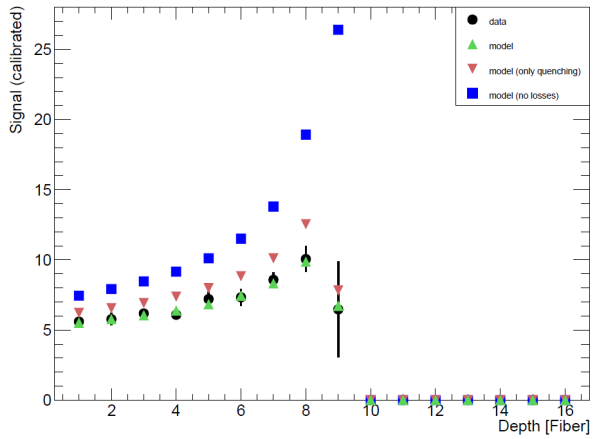
Your model need not just calculate numbers of events and apply such matching of model to data.

Let us look at a different problem:

Here is a data set where we have a particle pass through segmented layers of scintilating material slowing down, producing scintilation light:



# Models

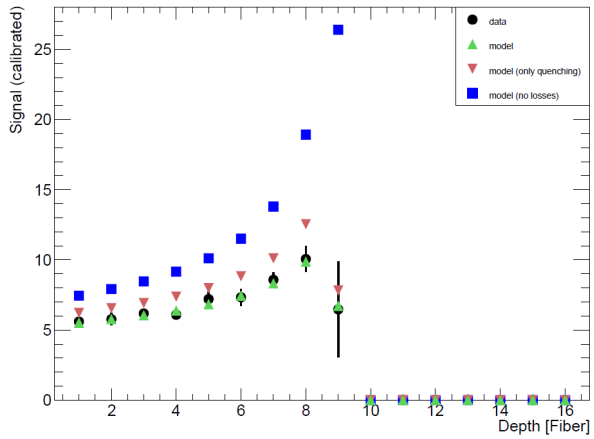



We have a multi-step model:

1. The particle slows down, according to the stopping power, producing a charecteristic "Bragg peak" (blue squares)

   This is dependent on the particle species, and its incoming energy

# Models




We have a multi-step model:

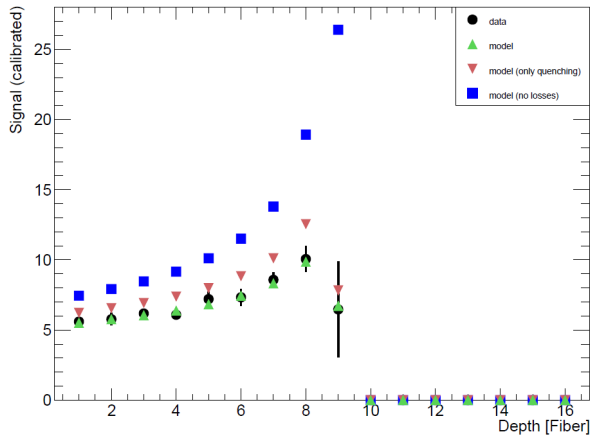
1. The particle slows down, according to the stopping power, producing a charecteristic "Bragg peak" (blue squares)

   This is dependent on the particle species, and its incoming energy

2. The scintilator produces photons, with non-linear dependence on the energy (red triangles)

   This is dependent on "Birks' coefficient"—a property of the scintilator

# Models




We have a multi-step model:

1. The particle slows down, according to the stopping power, producing a charecteristic "Bragg peak" (blue squares)

   This is dependent on the particle species, and its incoming energy

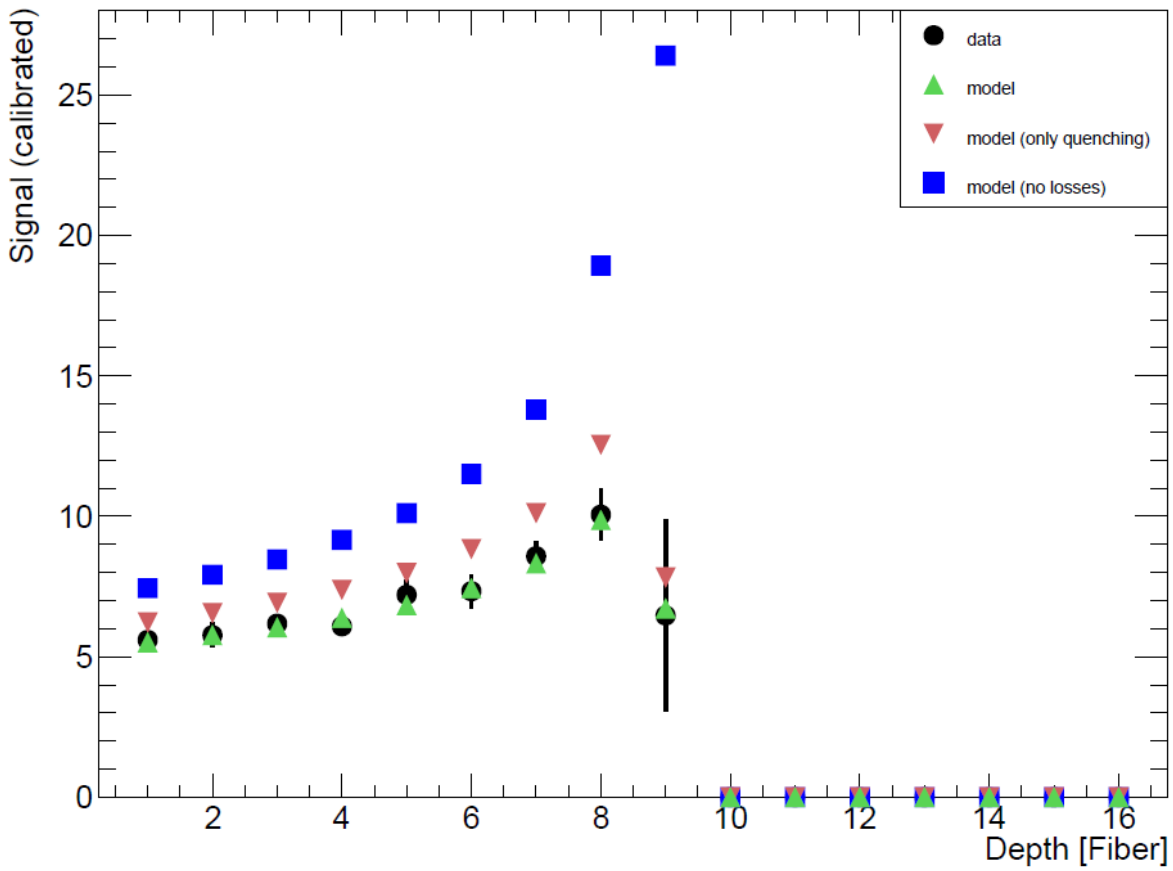
2. The scintilator produces photons, with non-linear dependence on the energy (red triangles)

   This is dependent on "Birks' coefficient"—a property of the scintilator

3. Finally the light sensor detects the photons with non-linear dependence on the number of photons (green triangles)

   This is dependent on a "photo-detection efficiency"—a property of the detector.

# Models

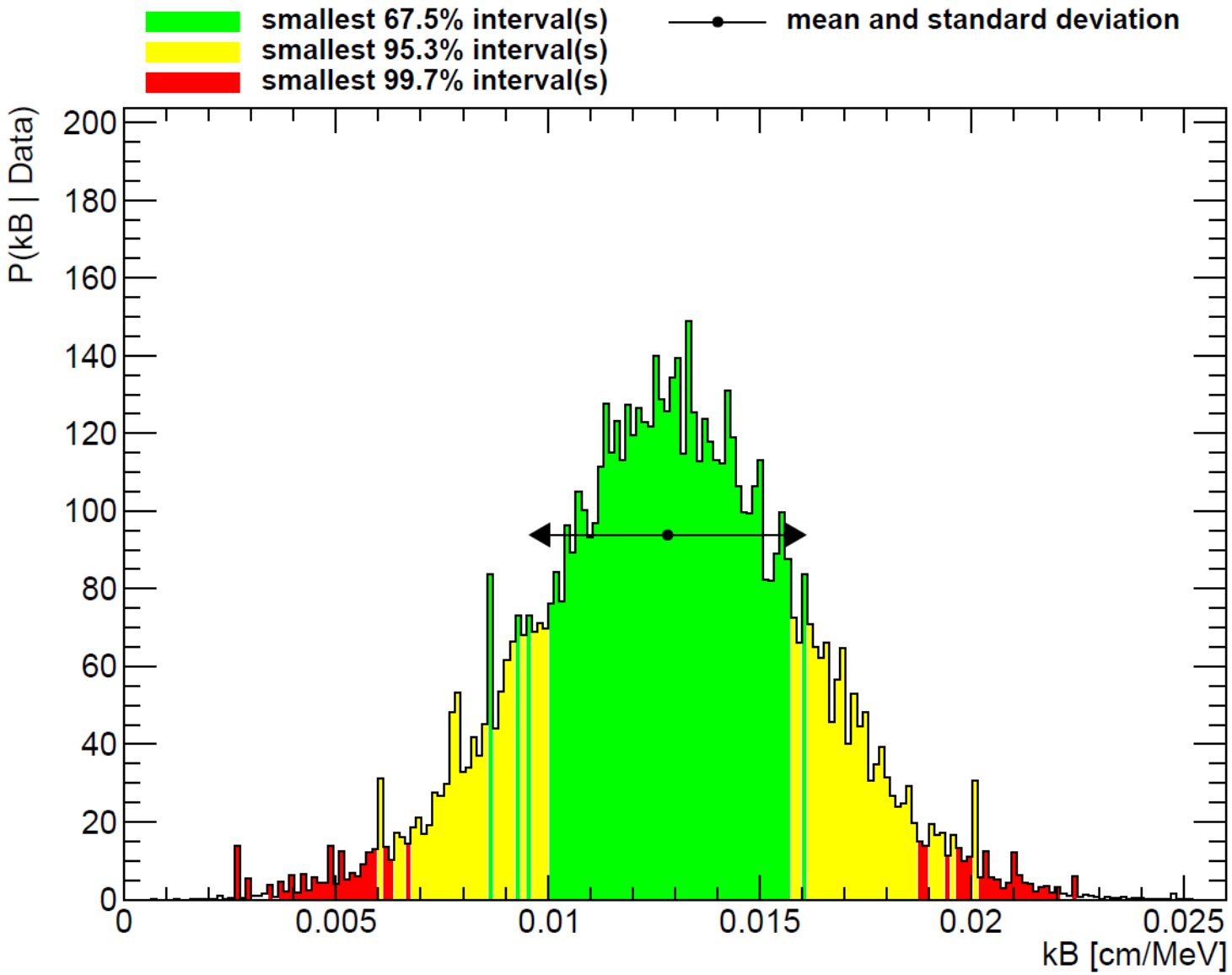



We match the predictions ($p_i$) against the measurements ($m_i$), using the measurement uncertainties ($\sigma_i^2$):

$$P(\vec{\lambda}|D) \propto \left( \prod_i \mathcal{N}\left( p_i(\vec{\lambda}) | m_i, \sigma_i^2 \right) \right) P_0(\vec{\lambda})$$

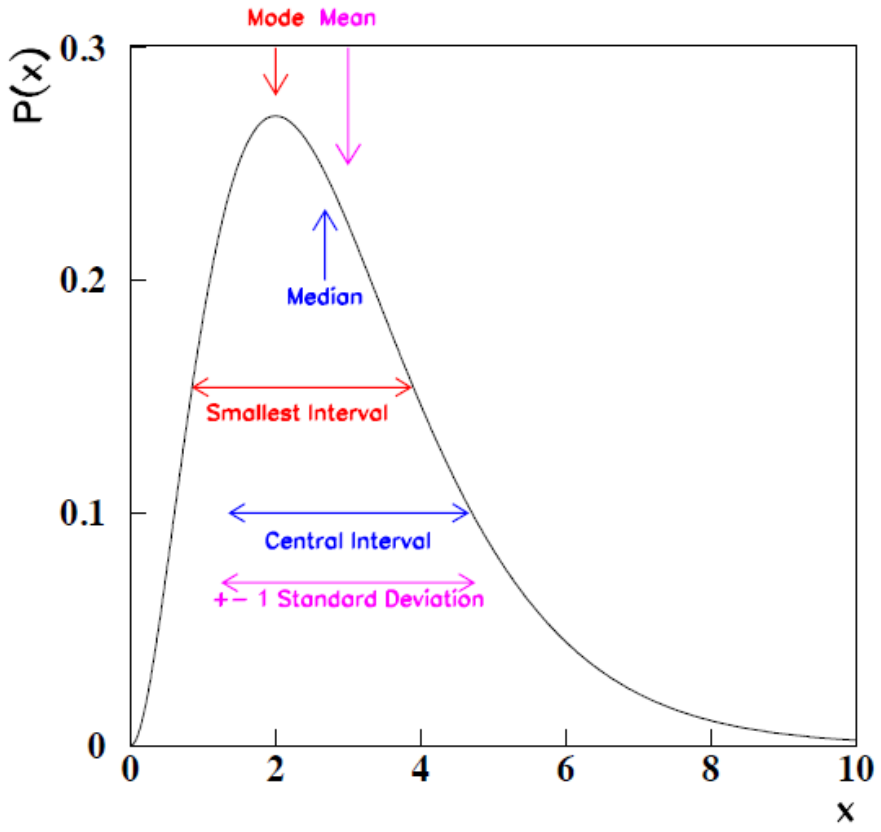
# Models

We sample from the posterior using BAT, and marginalize to measure Birk's coefficient:



# Credibility Intervals

Let us take a look at the possible intervals we could quote to summarize a posterior:




BUT: **The actual distribution is still more important than the summarizing interval!**

# Markov Chain Monte Carlo

We have seen that only the simplest of Bayesian analyses can be solved analytically.

Nearly all problems we encounter in physics require a numerical solution.

We need to

1. **Sample** parameter points from our posterior distribution to understand what it looks like
2. **Marginalize** over (integrate out) some parameters to project onto subspaces . . .

   . . . typically 1D and 2D subspaces.

**Markov Chain Monte Carlo** algorithms are well suited to both these tasks.

# Markov Chain Monte Carlo

**Monte Carlo** = stochastic process

**Markov Chain** = a sequence of points in which the conditional probability for a point given all those before it, depends only on the point directly preceding it:

$$P(\vec{x}_{n+1}|\vec{x}_1, \ldots, \vec{x}_n) = P(\vec{x}_{n+1}|\vec{x}_n)$$

As well, $P(\vec{x}_{n+1}|\vec{x}_n)$ need not depend on $\vec{x}_n$:

This is typical of the most naive algorithms. For example: Hit or Miss.

We will focus on the **Metropolis-Hastings algorithm**

# Metropolis-Hastings algorithm

We want to sample according to a function: $f(\vec{x}) : \mathbb{X} \to \mathbb{R}_{\geq 0}$
(for simplicity assume $\mathbb{X} \subset \mathbb{R}^n$)

1. from a current point $\vec{x}_i$ propose a new point $\vec{y}$

   Denote the probability of selecting $\vec{y}$ given $\vec{x}_i$ by $T(\vec{y}|\vec{x}_i)$

# Metropolis-Hastings algorithm

We want to sample according to a function: $f(\vec{x}) : \mathbb{X} \to \mathbb{R}_{\geq 0}$
(for simplicity assume $\mathbb{X} \subset \mathbb{R}^n$)

1. from a current point $\vec{x}_i$ propose a new point $\vec{y}$

   Denote the probability of selecting $\vec{y}$ given $\vec{x}_i$ by $T(\vec{y}|\vec{x}_i)$

2. calculate the **Hastings ratio**:

$$r(\vec{x}_i, \vec{y}) = \frac{f(\vec{y})}{T(\vec{y}|\vec{x}_i)} \Big/ \frac{f(\vec{x}_i)}{T(\vec{x}_i|\vec{y})} = \frac{f(\vec{y}) \cdot T(\vec{x}_i|\vec{y})}{f(\vec{x}_i) \cdot T(\vec{y}|\vec{x}_i)}$$

# Metropolis-Hastings algorithm

We want to sample according to a function: $f(\vec{x}) : \mathbb{X} \to \mathbb{R}_{\geq 0}$
(for simplicity assume $\mathbb{X} \subset \mathbb{R}^n$)

1. from a current point $\vec{x}_i$ propose a new point $\vec{y}$

   Denote the probability of selecting $\vec{y}$ given $\vec{x}_i$ by $T(\vec{y}|\vec{x}_i)$

2. calculate the **Hastings ratio**:

$$r(\vec{x}_i, \vec{y}) = \frac{f(\vec{y})}{T(\vec{y}|\vec{x}_i)} \Big/ \frac{f(\vec{x}_i)}{T(\vec{x}_i|\vec{y})} = \frac{f(\vec{y}) \cdot T(\vec{x}_i|\vec{y})}{f(\vec{x}_i) \cdot T(\vec{y}|\vec{x}_i)}$$

3. accept/reject $\vec{y}$ with probability $\min(1, r)$

# Metropolis-Hastings algorithm

We want to sample according to a function: $f(\vec{x}) : \mathbb{X} \to \mathbb{R}_{\geq 0}$
(for simplicity assume $\mathbb{X} \subset \mathbb{R}^n$)

1. from a current point $\vec{x}_i$ propose a new point $\vec{y}$

   Denote the probability of selecting $\vec{y}$ given $\vec{x}_i$ by $T(\vec{y}|\vec{x}_i)$

2. calculate the **Hastings ratio**:

$$r(\vec{x}_i, \vec{y}) = \frac{f(\vec{y})}{T(\vec{y}|\vec{x}_i)} \Big/ \frac{f(\vec{x}_i)}{T(\vec{x}_i|\vec{y})} = \frac{f(\vec{y}) \cdot T(\vec{x}_i|\vec{y})}{f(\vec{x}_i) \cdot T(\vec{y}|\vec{x}_i)}$$

3. accept/reject $\vec{y}$ with probability $\min(1, r)$

   3.1 if $r \geq 1$, $\vec{x}_{i+1} = \vec{y}$

# Metropolis-Hastings algorithm

We want to sample according to a function: $f(\vec{x}) : \mathbb{X} \to \mathbb{R}_{\geq 0}$
(for simplicity assume $\mathbb{X} \subset \mathbb{R}^n$)

1. from a current point $\vec{x}_i$ propose a new point $\vec{y}$

   Denote the probability of selecting $\vec{y}$ given $\vec{x}_i$ by $T(\vec{y}|\vec{x}_i)$

2. calculate the **Hastings ratio**:

$$r(\vec{x}_i, \vec{y}) = \frac{f(\vec{y})}{T(\vec{y}|\vec{x}_i)} \Big/ \frac{f(\vec{x}_i)}{T(\vec{x}_i|\vec{y})} = \frac{f(\vec{y}) \cdot T(\vec{x}_i|\vec{y})}{f(\vec{x}_i) \cdot T(\vec{y}|\vec{x}_i)}$$

3. accept/reject $\vec{y}$ with probability $\min(1, r)$

   3.1 if $r \geq 1$, $\vec{x}_{i+1} = \vec{y}$

   3.2 else throw uniform random number in unit interval:

$$a \sim U(0, 1)$$

   if $a \leq r$, $\vec{x}_{i+1} = \vec{y}$;

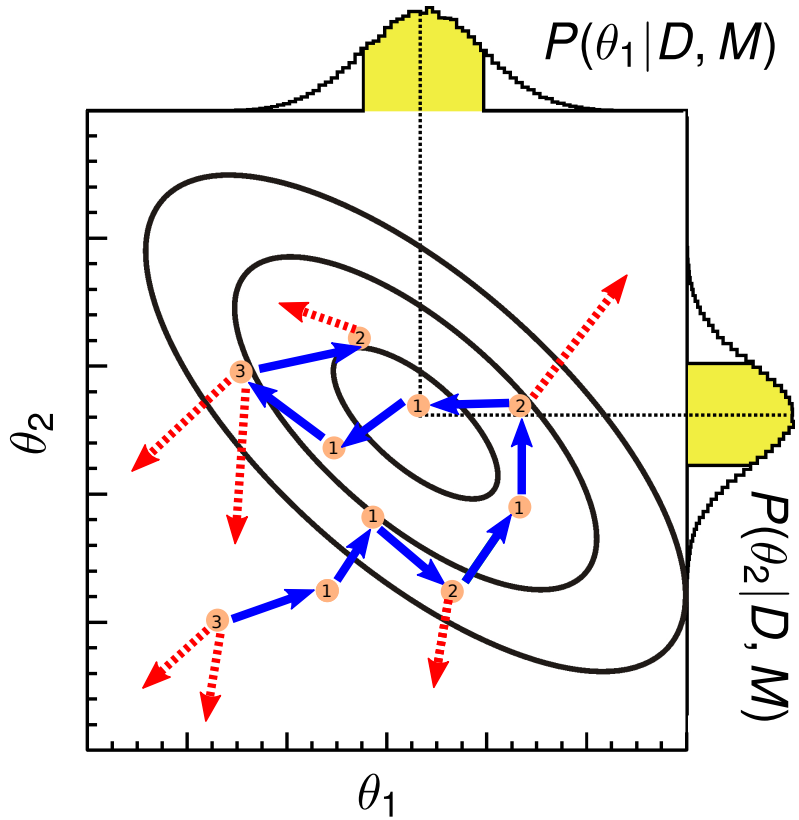
   else $\vec{x}_{i+1} = \vec{x}_i$

# Metropolis-Hastings algorithm

In this way we scan our parameter space, spending our time wisely:

- concetrating on areas of interest
- but not completely avoiding areas of less interest

by storing the parameters in histograms of the sub-spaces we **marginalize**



# Metropolis-Hastings algorithm

Some things to note:

1. in the accept/reject state we **always produce a new point**!

   if we reject $\vec{y}$, we accept $\vec{x}_i$ as $\vec{x}_{i+1}$

   This of course introduces auto-correlations.

   This does not impact marginalizations with large numbers of samples.

   But if you need to avoid auto-correlation (say for MC), **you cannot simply change the accept/reject step**

   It is vital for the functioning of the algorithm.

   Instead you apply a lag: take every n'th sample

# Metropolis-Hastings algorithm

Some things to note:

2. the algorithm can propose a $\vec{y}_0$ for which $f(\vec{y}_0) = 0$

   but it will never go to it, since:

$$r(\vec{x}_i, \vec{y}_0) \propto f(\vec{y}_0) = 0$$

   So from a point for which $f(\cdot) \neq 0$ you can never reach a point $f(\cdot) = 0$, **regardless of the proposal function**

# Metropolis-Hastings algorithm

Some things to note:

3. $r$ is **undefined** if

$$T(\vec{x}_i|\vec{y}) = 0, \text{for some } \vec{x}_i, \vec{y}|\, T(\vec{y}|\vec{x}_i) \neq 0$$

since:

$$r(\vec{x}_i, \vec{y}) \propto \frac{T(\vec{y}|\vec{x}_i)}{T(\vec{x}_i|\vec{y})}$$

SO: **the proposal function must be reversible!**
If the transition $\vec{x}_i \rightarrow \vec{y}$ can occur,
then the transition $\vec{y} \rightarrow \vec{x}_i$ must be able to occur

In fact, the proposal function is usually chosen symmetric:

$$B_{\vec{r}}(\vec{x}_i), \qquad \vec{x}_i + \mathcal{N}(0, \Sigma), \qquad \text{hypercube}(\vec{x}_i)$$

This is the original Metropolis algorithm

# Metropolis-Hastings algorithm

Some things to note:

4. finally, $r(\vec{x}_i, \vec{y})$ is **undefined** if $f(\vec{x}_i) = 0$

   but since we can never accept a new point for which $f(\cdot) = 0$, we need only insure:

$$f(\vec{x}_0) \neq 0$$

# MCMC in practice

1. we will need to tune the proposal function so that the algorithm is efficient.

   most common:

   - Gaussian,
   - Cauchy,
   - Student's $t$

   All require radius:

   - If radius is too large: too often select unlikely points $\rightarrow$ chain becomes ineffecient
   - If radius is too small: though efficient, we take small steps, move too slowly, see only part of parameter space

   So we monitor efficiency:

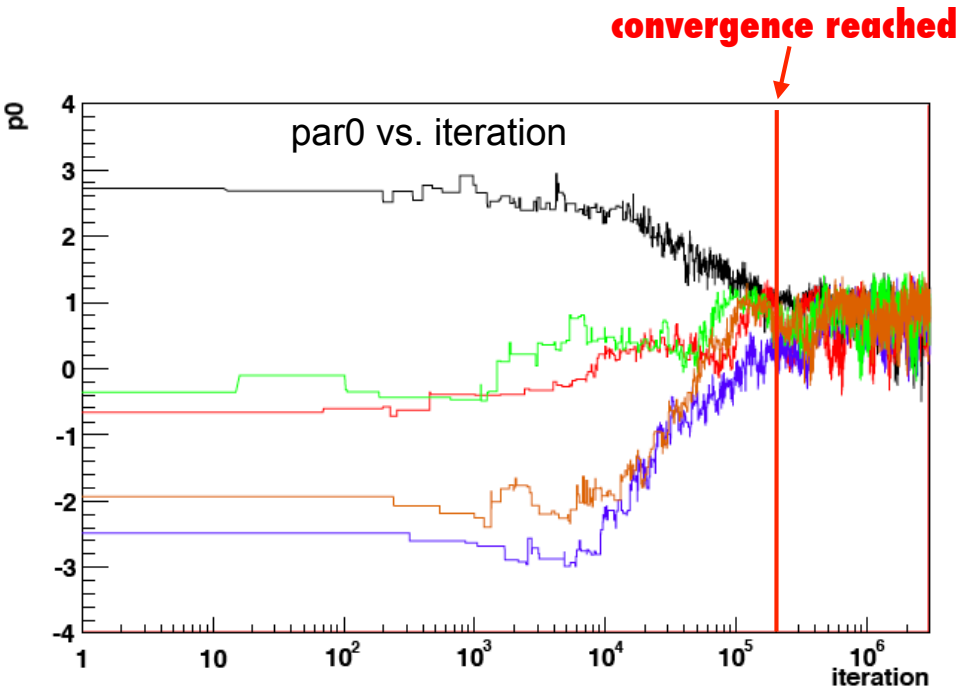
   - if efficiency is too low, decrease radius
   - if efficiency is too large, increase radius

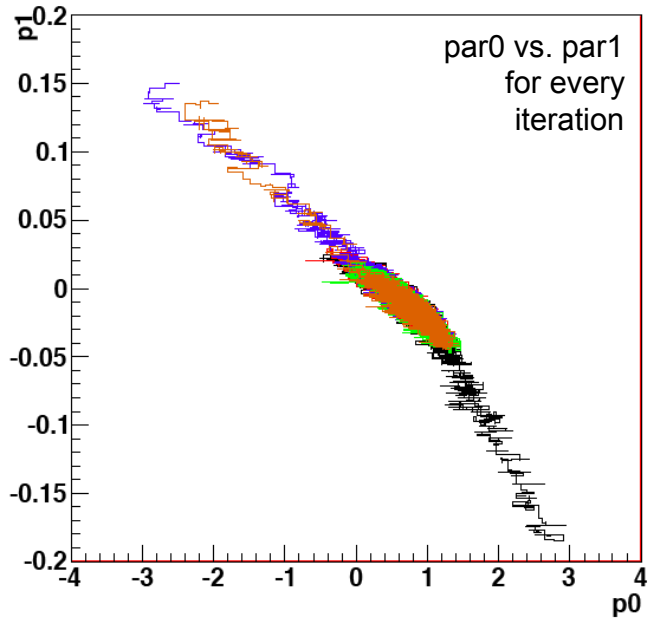
# MCMC in practice

2. It takes some initial number of iterations before the Markov Chain **converges** to its equilbrium distribution

   There are many methods for judging whether a chain has converged.

   Many judge graphically.



# MCMC in practice

2. It takes some initial number of iterations before the Markov Chain **converges** to its equilbrium distribution

   There are many methods for judging whether a chain has converged.

   Many judge graphically.

   In BAT we judge by modified $R$ values (Brooks  Gelman, 1998)

   1. run several chains for many iterations
   2. for each parameter, calculate

$$R'(\lambda) \propto R(\lambda) = \frac{\text{variance over all samples in all chains for } \lambda}{\text{mean of chains' individual variances for } \lambda}$$

   3. chains have converged when $R'$ is below threshold for all parameters

   **What is the threshold? Wait—what even is $R'$?**

   $R'$ is a quasi-measurement of the changing of our variance with respect to the true variance:

   $R' - 1 \approx$ the shrinking of our variance towards the true variance if we run for more iterations

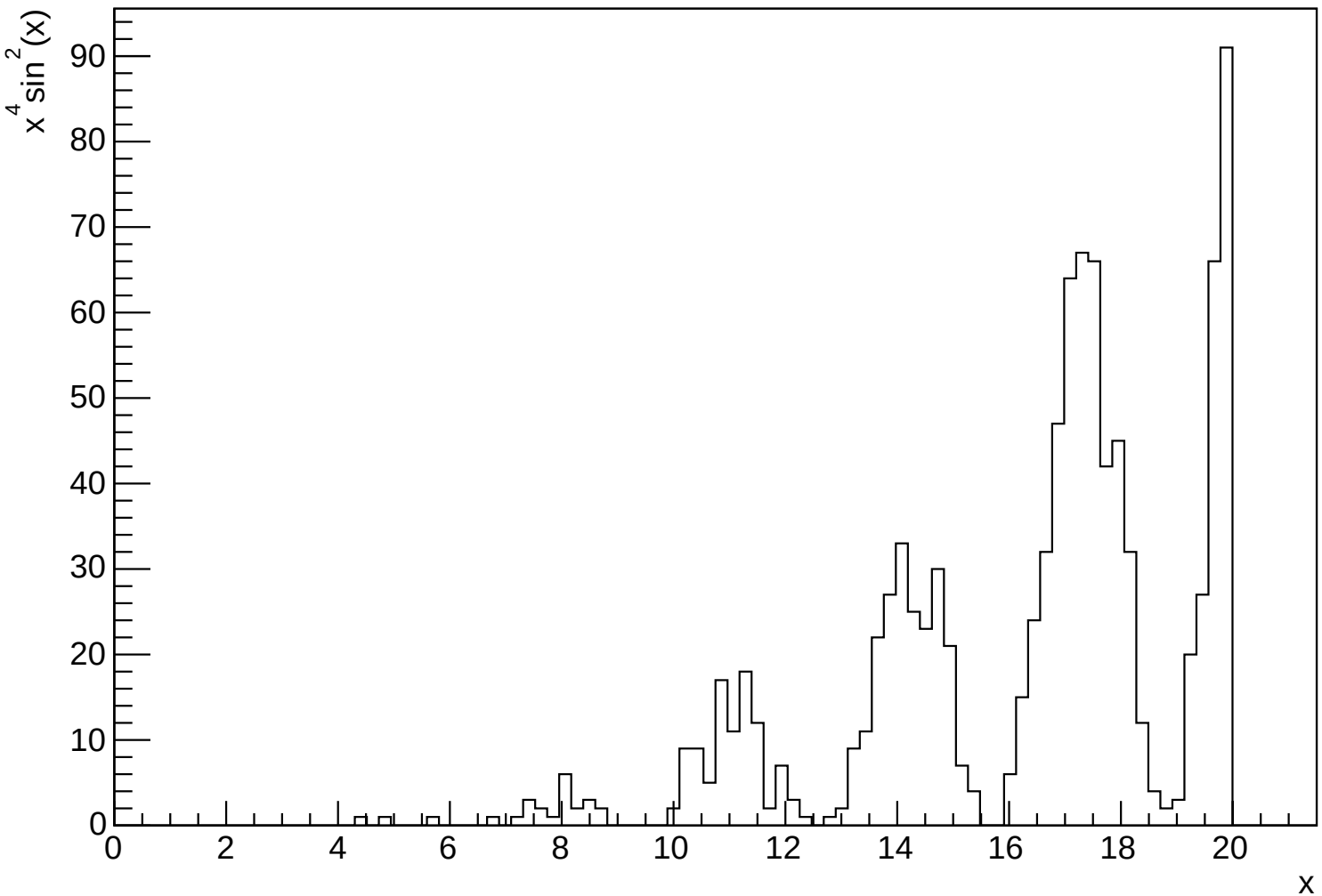
   in BAT, we default to requiring $R' \leq 1.1$

# MH MCMC demonstration

Let us look at how well the algorithm attacks equations with multiple peaking structures and dead spaces:

$$f(x) = x^4 \sin^2(x)$$

1000 iterations

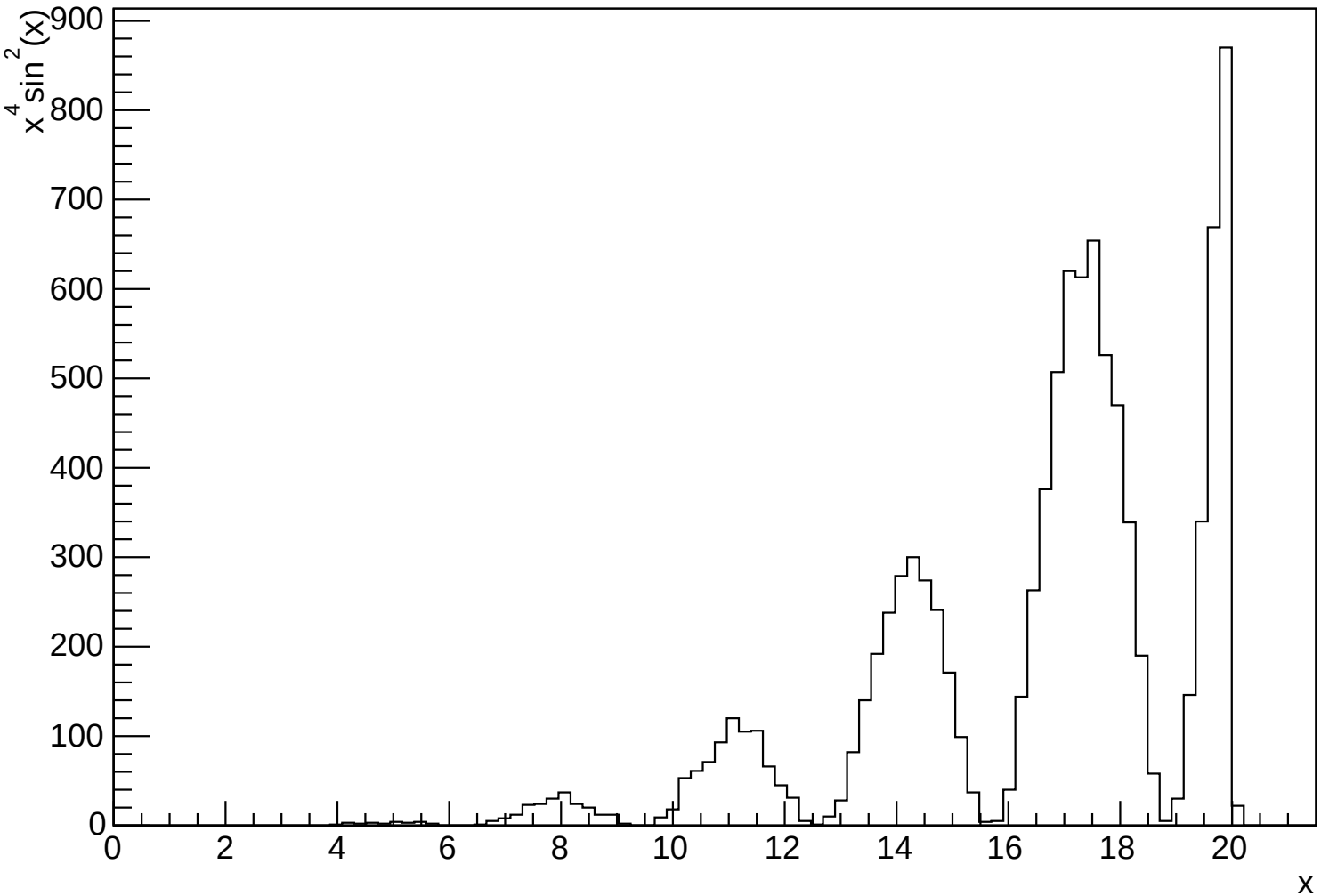


# MH MCMC demonstration

Let us look at how well the algorithm attacks equations with multiple peaking structures and dead spaces:

$$f(x) = x^4 \sin^2(x)$$

10,000 iterations

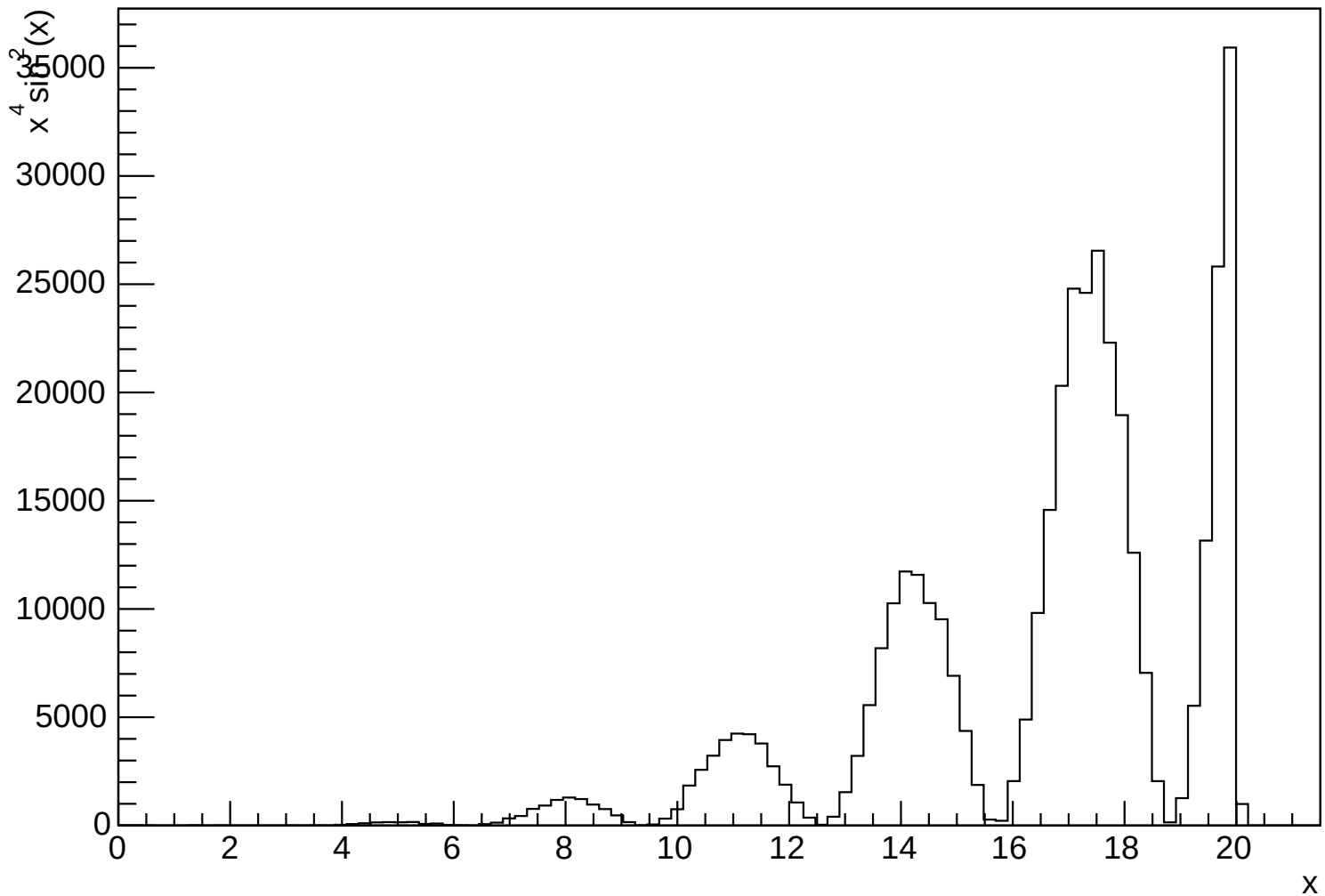


# MH MCMC demonstration

Let us look at how well the algorithm attacks equations with multiple peaking structures and dead spaces:

$$f(x) = x^4 \sin^2(x)$$

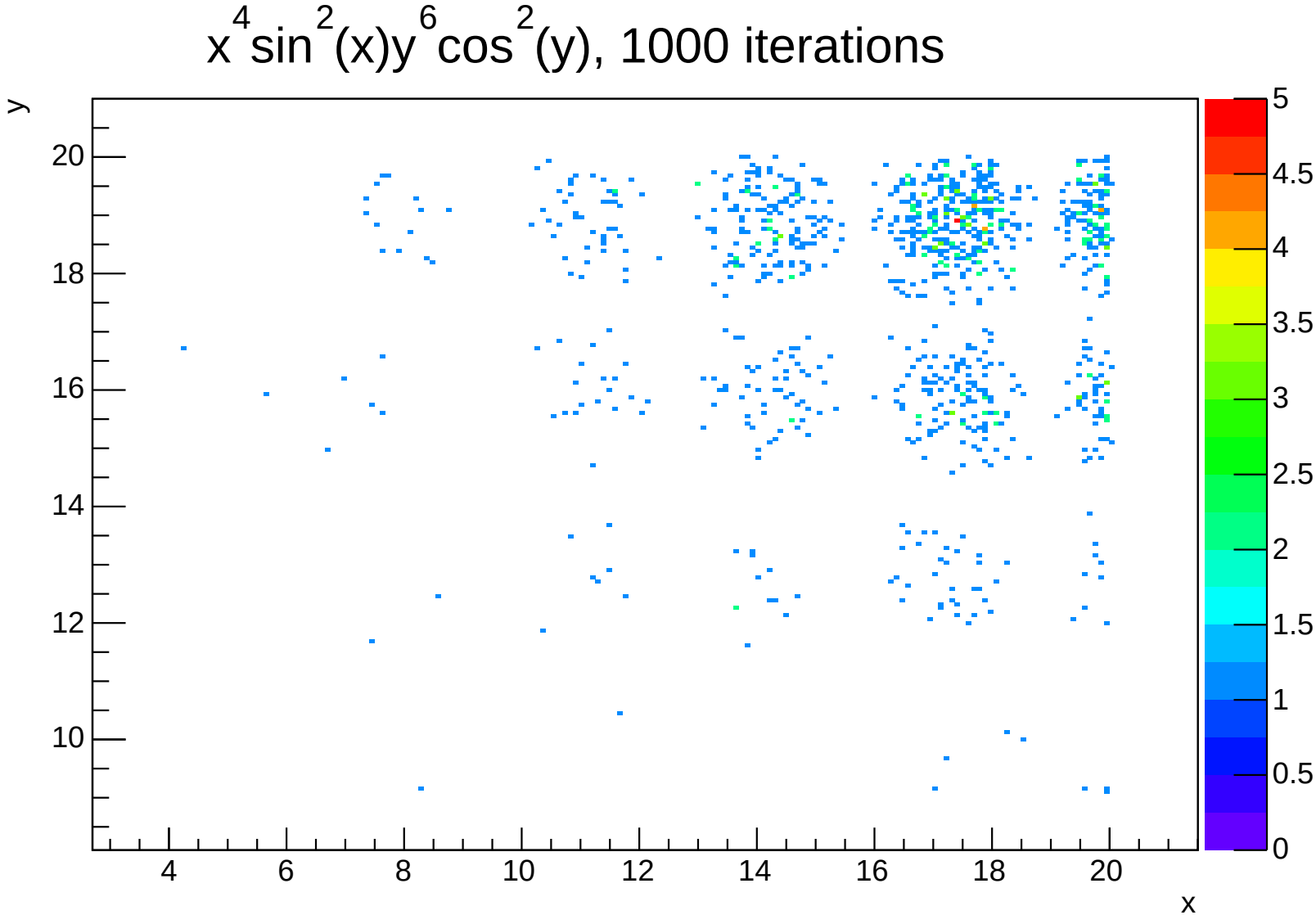
400,000 iterations



# MH MCMC demonstration

Let us look at how well the algorithm attacks equations with multiple peaking structures and dead spaces:

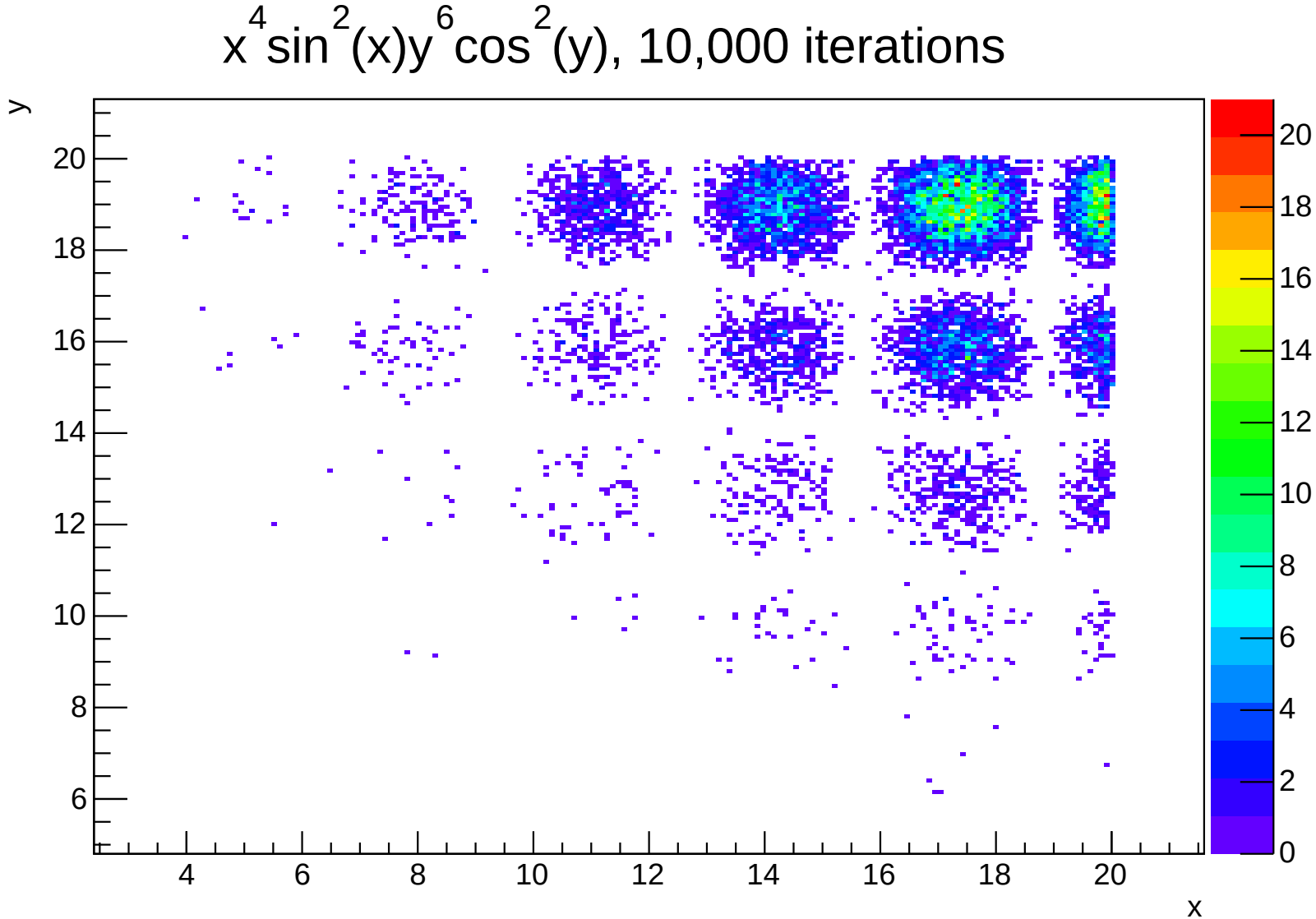
$$f(x, y) = x^4 \sin^2(x) y^6 \cos^2(y)$$



# MH MCMC demonstration

Let us look at how well the algorithm attacks equations with multiple peaking structures and dead spaces:

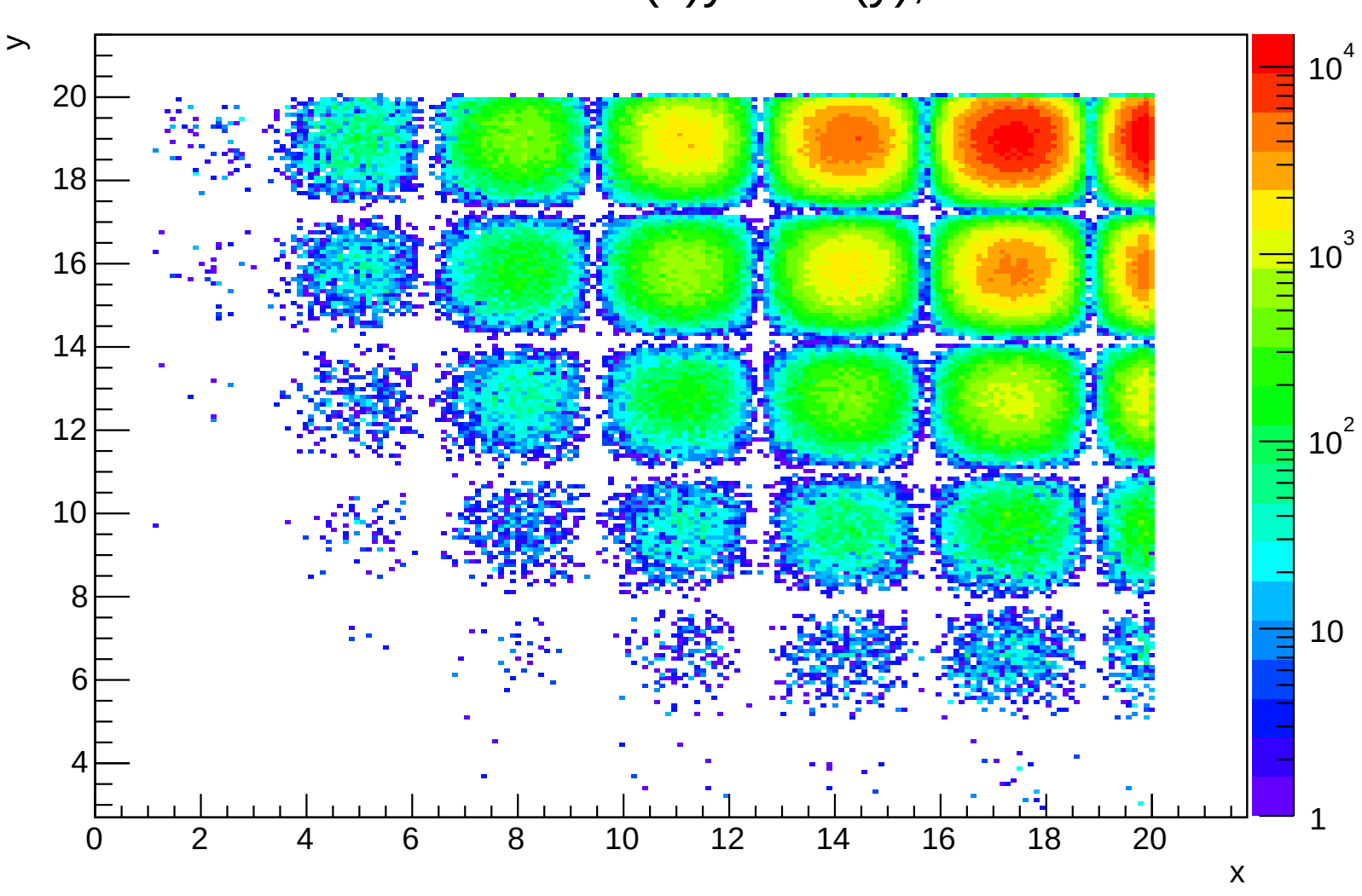
$$f(x, y) = x^4 \sin^2(x) y^6 \cos^2(y)$$



# MH MCMC demonstration

Let us look at how well the algorithm attacks equations with multiple peaking structures and dead spaces:

$$f(x, y) = x^4 \sin^2(x) y^6 \cos^2(y)$$



# More practical experience

And in the tutorial, we'll explore using the

**Bayesian Analysis Toolkit (BAT)**

to run MCMC to sample from posteriors.