What have we measured?

Daniel Greenwald
Technische Univerität München

December 1, 2022

Let's look at an an analysis of

$$\tau^- \to XY$$

in which we

- identify *all* sources of events that mimic the signal (backgrounds) and
- determine criteria that
  - select for signal events and
  - reject background ones.

Explicitly, for each source i, we **observe** that

- for M_i simulated events,
- m_i events pass the selection criteria.

We then look at the real data and **observe** that

- for data of a known size,
- N events pass the same selection criteria.

What's the goal?

To measure the branching fraction

$$\mathcal{B}_s \equiv \frac{\Gamma(\tau^- \to XY)}{\Gamma(\tau^- \to \text{anything})}$$

What have we observed?

- $M_i \to m_i$ in simulation
- N in data (of known size)

How do we connect the two?

A model . . . a **physics** model.

Let's think like physicists first and then like statisticians.

From collision to observation:

For signal,

$e^+e^- \to \tau^+\tau^-$

$\hookrightarrow \tau^\pm \to$ 1 prong occurs $\mathcal{B}_t$ fraction of the time (tag)

$\hookrightarrow \tau^\mp \to XY$ occurs $\mathcal{B}_s$ fraction of the time (signal)

- $2\mathcal{B}_s\,\mathcal{B}_t$ of $\tau\tau$ events are the type of event we reconstruct.
- And we have an efficiency ϵ_s to detect them.

So we expect to observe $2\epsilon_s\,\mathcal{B}_s\,\mathcal{B}_t$ of $\tau\tau$ events.

How many $\tau\tau$ events are there?

$$\sigma_{\tau\tau}\,L$$

- $\sigma_{\tau\tau} \equiv$ cross section for $e^+e^- \to \tau^+\tau^-$
- $L \equiv$ integrated luminosity of the data set

So we expect to measure $2\epsilon_s\,\mathcal{B}_s\,\mathcal{B}_t\,\sigma_{\tau\tau}\,L$ signal events

From collision to observation:

For each background,

$e^+e^- \to$ (background source)$_i$

- with cross section σ_i
- and efficiency ϵ_i

So we expect to measure $\sum_i \epsilon_i \, \sigma_i \, L$ background events

In total, we expect

$$\left[2\epsilon_s \, \mathcal{B}_s \, \mathcal{B}_t \, \sigma_{\tau\tau} + \sum_i \epsilon_i \, \sigma_i \right] L \text{ events}$$

From simulation to observation

We simulated M_i events, and observe m_i events after selection criteria.

This informs us about the efficiency.

Naively:

$$\epsilon_i = \frac{m_i}{M_i} \pm \frac{m_i}{M_i}\sqrt{\frac{1}{m_i} + \frac{1}{M_i}}$$

Naive . . . and wrong . . . $X \pm Y$ implies a normal distribution,

$$P(\epsilon_i | M_i, m_i) = \mathcal{N}\left(\frac{m_i}{M_i}, \frac{m_i}{M_i}\sqrt{\frac{1}{m_i} + \frac{1}{M_i}}\right)$$

which is not the case.

And what happens when $m_i = 0$:

$$\epsilon_i(m_i = 0) = 0 \text{ exactly.}$$

This states we're 100% confident the efficiency is 0 . . . regardless of M_i's value.

This is also not a definition of an **efficiency**.
It's just a ratio of two numbers.

We need a statistical model.

Let's try again using Bayesian statistics.

An ultra quick refresher:

$$P(A \cap B) = P(B \cap A)$$

$$P(A|B)\,P(B) = P(B|A)\,P(A)$$

$$\boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}}$$

We relate the conditional probability of A given B (the left side)

to the conditional probablity of B given A (the right side)

$$\boxed{P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}}$$

Let's look at this more concretely in terms of data and parameters:

We want to know the probability of parameters, $\vec{\lambda}$, given measured data

$$P(\vec{\lambda}|\text{data}) = \frac{P(\text{data}|\vec{\lambda})\,P(\vec{\lambda})}{P(\text{data})}$$

So we use Bayes' theorem.

These four parts have common names (and symbols):

$$\begin{array}{lll}
P(\vec{\lambda}|\text{data}) & \equiv & \textit{posterior} \text{ prob. distribution} \\
P(\text{data}|\vec{\lambda}) \equiv \mathcal{L}(\text{data}|\vec{\lambda}) & \equiv & \text{likelihood of the data} \\
P(\vec{\lambda}) \equiv P_0(\vec{\lambda}) & \equiv & \textit{a priori} \text{ prob. distribution} \\
P(\text{data}) \equiv Z(\text{data}) = \int \mathcal{L}(\text{data}|\vec{\lambda})\,P_0(\vec{\lambda})\,\mathrm{d}\vec{\lambda} & \equiv & \text{evidence}
\end{array}$$

$$P(\vec{\lambda}|\text{data}) \propto \mathcal{L}(\text{data}|\vec{\lambda})\, P_0(\vec{\lambda})$$

Since we can normalize the posterior from the above proportionality, we don't need to

- concern ourselves with the evidence,
- normalize the likelihood, or
- normalize the prior

$$P(\vec{\lambda}|\text{data}) \propto \mathcal{L}(\text{data}|\vec{\lambda})\, P_0(\vec{\lambda})$$

Let's look now at our efficiency, ϵ_i:

$$P(\epsilon_i|M_i, m_i) \propto \mathcal{L}(m_i|M_i, \epsilon_i)\, P_0(\epsilon_i)$$

What is the likelihood for selecting m_i events from a total of M_i when the efficiency for selecting each event is ϵ_i?

$$\mathcal{L}(m_i|M_i, \epsilon_i) \propto {\epsilon_i}^{m_i} (1-\epsilon_i)^{M_i - m_i}$$

$\Rightarrow$ the Binomial distribution:

What about $P_0(\epsilon_i)$? A general approach that parameterizes two common choices:

$$P_0(\epsilon_i|\delta_i) = \begin{cases} (\epsilon_i)^{-\delta_i} (1-\epsilon_i)^{-\delta_i}, & \text{if } \epsilon_i \in [0,1], \\ 0, & \text{else} \end{cases}$$

$\delta_i = 0 \Rightarrow P_0(\epsilon_i) =$ constant prior $\qquad$ $\delta_i = 1 \Rightarrow P_0(\epsilon_i) =$ double-inverse prior

Putting it all together

$$P(\epsilon_i|M_i, m_i) \propto {\epsilon_i}^{m_i-\delta_i}(1-\epsilon_i)^{M_i-m_i-\delta_i}$$

if $\epsilon_i \in [0, 1]$, else it's zero.

In terms of ϵ_i, this is the Beta distribution . . . with properties:

mean $\quad \bar{\epsilon}_i = \frac{m_i+1-\delta_i}{M_i+2-2\delta_i}$

mode $\quad \epsilon_i^* = \frac{m_i-\delta_i}{M_i-2\delta_i}$

variance $\quad \mathsf{Var}[\epsilon_i] = \bar{\epsilon}_i^2 \frac{M_i-m_i+1-\delta_i}{(m_i+1-\delta_i)(M_i+3-2\delta_i)}$

$$\frac{\mathsf{Var}[\epsilon_i]}{\bar{\epsilon}_i^2} = \frac{1}{m_i+1-\delta_i} - \frac{1}{M_i+3-2\delta_i} - \frac{1}{(m_i+1-\delta_i)(M_i+3-2\delta_i)}$$

skew $\quad = \ldots$

kurtosis $\quad = \ldots$

Let's make the full model now:

$$P(\mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L | N, \vec{M}, \vec{m}) \propto \mathcal{L}(N, \vec{M}, \vec{m} | \mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L)\, P_0(\mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L)$$

Let's focus on the likelihood first. It factorizes:

$$\mathcal{L}(N, \vec{M}, \vec{m} | \mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L) = \mathcal{L}(N | \mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L) \prod_i \mathcal{L}(m_i | M_i, \epsilon_i)$$

And we already know $\mathcal{L}(m_i | M_i, \epsilon_i)$.

The likelihood to observe N events, given an expectation is

$$\mathcal{L}(N | \mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L) \propto \nu^N e^{-\nu}$$

$\Rightarrow$ the Poisson distribution ... with

$$\nu \equiv \left[2\epsilon_\mathsf{s}\, \mathcal{B}_\mathsf{s}\, \mathcal{B}_\mathsf{t}\, \sigma_{\tau\tau} + \sum_i \epsilon_b\, \sigma_b \right] L$$

What about the prior?

It factorizes:

$$P_0(\mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L) = P_0(\mathcal{B}_\mathsf{s})\, P_0(\mathcal{B}_\mathsf{t})\, P_0(L) \prod_i P_0(\epsilon_i)\, P_0(\sigma_i)$$

We've already discussed $P_0(\epsilon_i)$.

What about the others?

- $P_0(\mathcal{B}_\mathsf{t}) = \mathcal{N}(\mathcal{B}_\mathsf{t}|\text{mean}, \text{variance})$ [PDG...]
- $P_0(L) = \mathcal{N}(L|\text{mean}, \text{variance})$ [Belle II's measured value]
- $P_0(\sigma_i) = \mathcal{N}(\sigma_i|\text{mean}, \text{variance})$
  - measured value @ $\sqrt{s} = 10.58\,\text{GeV}$
  - calculated value @ $\sqrt{s} = 10.58\,\text{GeV}$

  If necessary, $\sqrt{s}$ could itself be a parameter.

And what about $P_0(\mathcal{B}_\mathsf{s})$?

This is our parameter of interest ... we have several choices.

$P_0(\mathcal{B}_s)$ Choices

- $P_0(\mathcal{B}_s) =$ flat in [0,1] $\rightarrow$ flat in the *value* of $\mathcal{B}_s$
- $P_0(\mathcal{B}_s) \propto 1/\mathcal{B}_s \rightarrow$ flat in the *scale* of $\mathcal{B}_s$
- $P_0(\mathcal{B}_s) =$ previous measurement result $\rightarrow$ most informative

  This may lead to the lowest upper bound,
  but requires work to figure out the proper likelihood.

  Ordinarily we don't do this, so that our measurement can be used with others—cf. the PDG—but since experiments generally don't report *useful* results on upper limits, the PDG doesn't average them anyway!

The results should not vary greatly (at least between the first two choices).

Such a check is easy—and should be done.

Putting this aaaaall together ...

$$P(\mathcal{B}_s, \mathcal{B}_t, \vec{\epsilon}, \vec{\sigma}, L | N, \vec{M}, \vec{m}) \propto \mathcal{L}(N, \vec{M}, \vec{m} | \mathcal{B}_s, \mathcal{B}_t, \vec{\epsilon}, \vec{\sigma}, L)\, P_0(\mathcal{B}_s, \mathcal{B}_t, \vec{\epsilon}, \vec{\sigma}, L)$$

$$\begin{aligned}
\mathcal{L}(N, \vec{M}, \vec{m} | \mathcal{B}_s, \mathcal{B}_t, \vec{\epsilon}, \vec{\sigma}, L) &= \mathcal{L}(N | \mathcal{B}_s, \mathcal{B}_t, \vec{\epsilon}, \vec{\sigma}, L) \prod_i \mathcal{L}(m_i | M_i, \epsilon_i) \\
\mathcal{L}(N | \mathcal{B}_s, \mathcal{B}_t, \epsilon_s, \sigma_{\tau\tau}, L) &= \text{Poisson}(N | 2\epsilon_s\, \mathcal{B}_s\, \mathcal{B}_t\, L + \textstyle\sum_b \epsilon_b\, \sigma_b\, L) \\
\mathcal{L}(m_i | M_i, \epsilon_i) &= \text{Binom}(m_i | M_i, \epsilon_i)
\end{aligned}$$

$$P_0(\mathcal{B}_s, \mathcal{B}_t, \vec{\epsilon}, \vec{\sigma}, L) = P_0(\mathcal{B}_s)\, P_0(\mathcal{B}_t)\, P_0(L) \prod_i P_0(\epsilon_i)\, P_0(\sigma_i)$$

$$\begin{aligned}
P_0(\mathcal{B}_t) &= \text{world average} \\
P_0(L) &= \text{Belle II measurement from other channels} \\
P_0(\epsilon_i) &= {\epsilon_i}^{-\delta_i} (1 - \epsilon_i)^{-\delta_i} \\
P_0(\sigma_i) &= \text{measurement or calculation} \\
P_0(\mathcal{B}_s) &= {\mathcal{B}_s}^{-\delta_{\mathcal{B}}}
\end{aligned}$$

a semantic matter:

$$\mathcal{L}(m_i|M_i, \epsilon_i) \times P_0(\epsilon_i) = \text{Binom}(m_i|M_i, \epsilon_i) \times P_0(\epsilon_i)$$

can instead be taken as a prior on ϵ_i,
with no likelihood component for it then:

$$P_0(\epsilon_i) = \text{Beta}(\epsilon_i|m_i + 1 - \delta_i, M_i - m_i + 1 - \delta_i)$$

So we have a model:

$$P(\mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L | N, \vec{M}, \vec{m}) = \ldots$$

But we don't care about $\mathcal{B}_\mathsf{t}$, $\vec{\epsilon}$, $\vec{\sigma}$, L ... they're *nuissance* parameters.

So we integrate them out:

$$P(\mathcal{B}_\mathsf{s} | N, \vec{M}, \vec{m}) = \int P(\mathcal{B}_\mathsf{s}, \mathcal{B}_\mathsf{t}, \vec{\epsilon}, \vec{\sigma}, L | N, \vec{M}, \vec{m}) \, \mathrm{d}\mathcal{B}_\mathsf{t} \, \mathrm{d}\vec{\epsilon} \, \mathrm{d}\vec{\sigma} \, \mathrm{d}L$$

This is called the "marginalization" of $\mathcal{B}_\mathsf{s}$.

The integral is difficult (perhaps impossible?) to analytically calculate.

But it's very easy to calculate on the computer via sampling:

- Sample from the full posterior.
- Histogram of $\mathcal{B}_\mathsf{s}$ = marginalization of $\mathcal{B}_\mathsf{s}$

Then we have the probability distribution for $\mathcal{B}_\mathsf{s}$ and upper limits are easy:

$$\int_0^{\mathcal{B}_\mathsf{s}^{\alpha\mathsf{UL}}} P(\mathcal{B}_\mathsf{s} | N, \vec{M}, \vec{m}) = \alpha$$

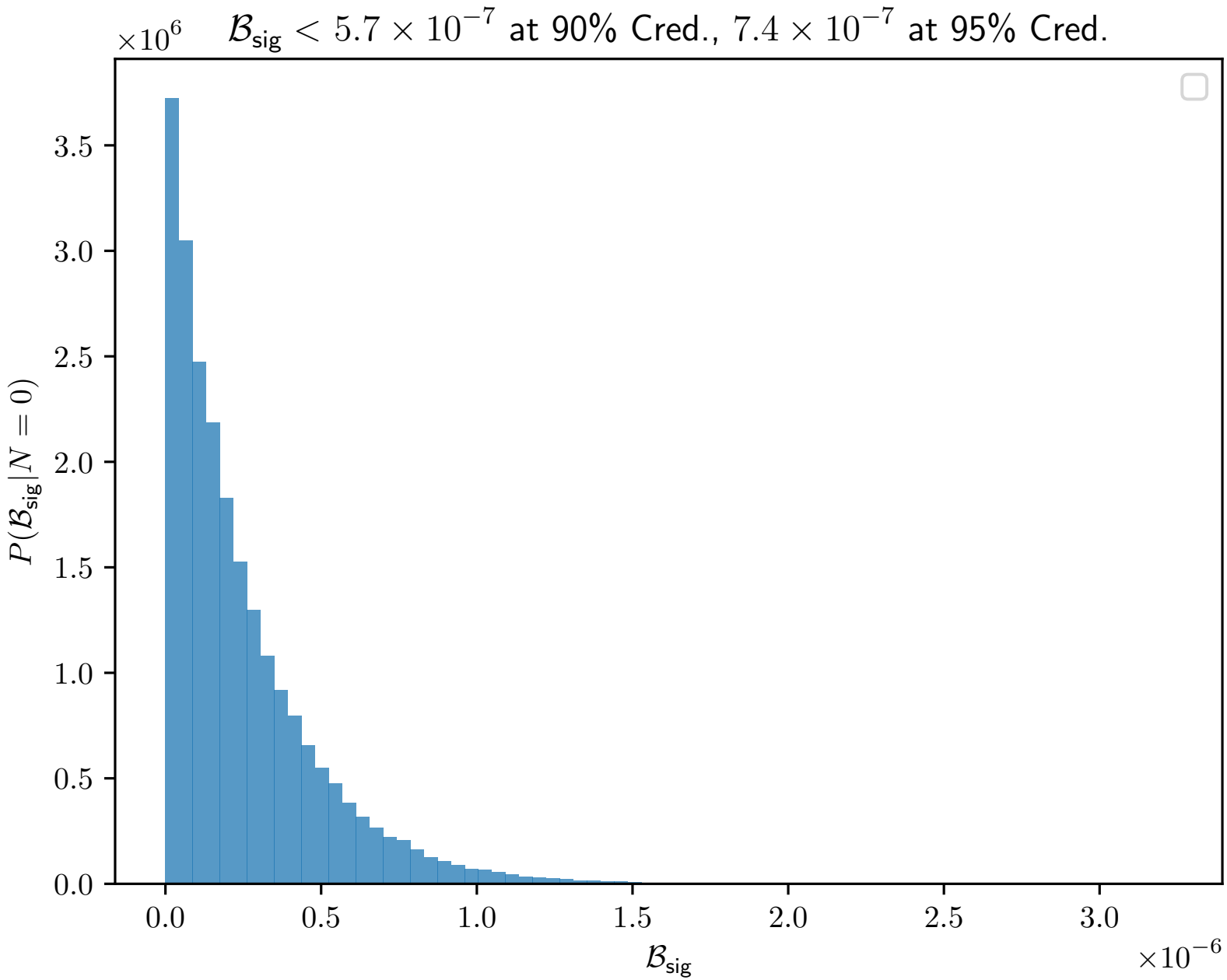
so let's do it ...

Let's say we have the following information

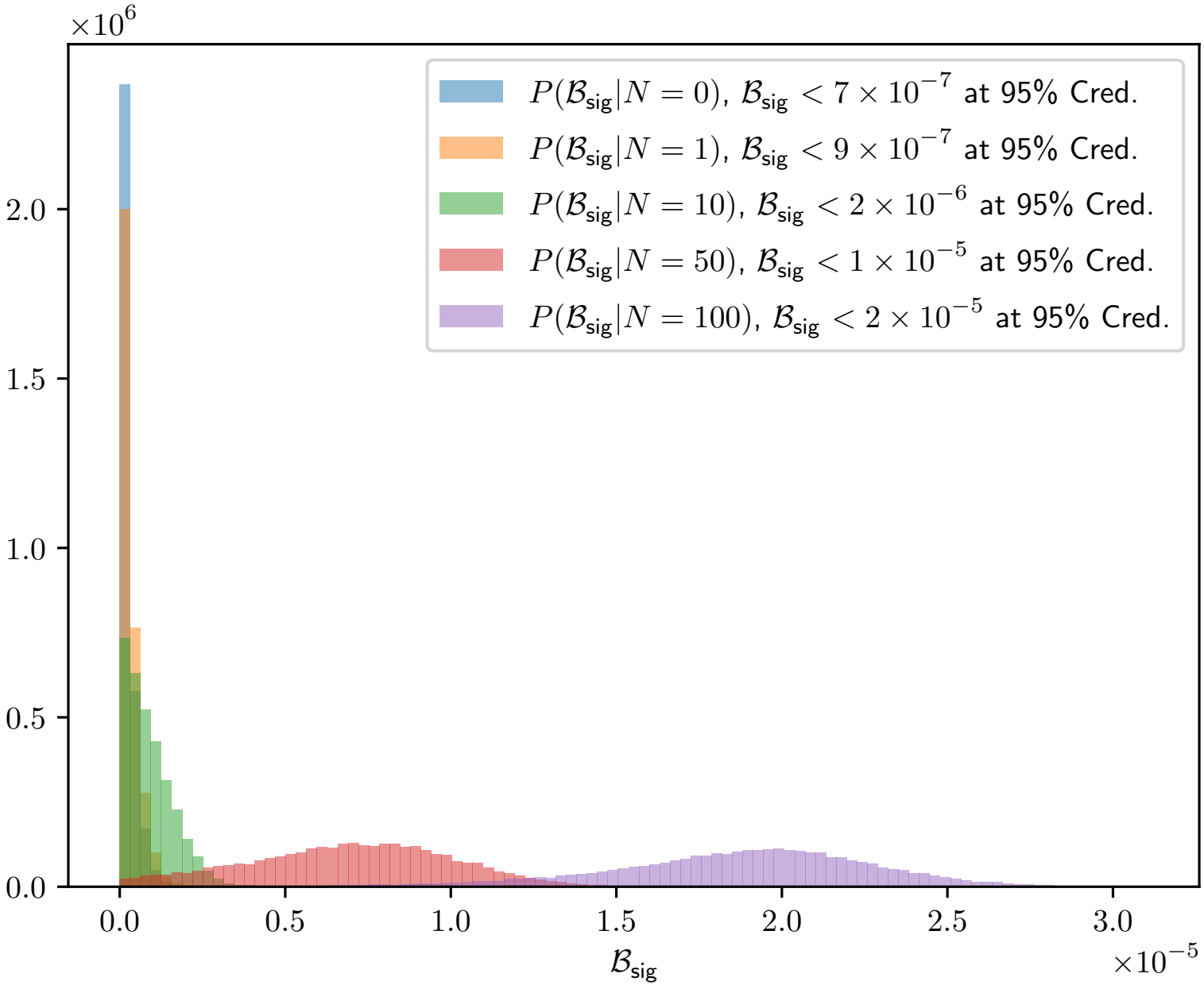
BG Source	M [10^6]	m	σ [nb]
$q\bar{q}$	372	0	$3.720\,00 \pm 0.003\,72$
$B\overline{B}$	105	0	1.100 ± 0.011
eeγ	2940	0	294 ± 2
$\mu\mu\gamma$	575	0	1.148 ± 0.005
ee$\mu\mu$	1890	0	18.97 ± 0.02
eeee	397	0	39.74 ± 0.02
$\tau\tau$ (bg)	91.9	0	0.919 ± 0.003
$\tau^- \to XY$ (sig)	1	53 489	

$\mathcal{B}_t = (84.71 \pm 0.06) \times 10^{-2}$ $\quad$ $L = (100.0 \pm 1.4)\,\text{fb}^{-1}$

$N = 0$, all $\delta = 0$ (flat priors)



Let's look at different measurement possibilities:



an interesting aside:

Let's look at a Poisson distribution with two sources:

$$P(\nu, \beta|N) \propto (\nu + \beta)^N e^{-(\nu+\beta)} P_0(\nu) P_0(\beta)$$

The marginalized distribution for ν is

$$P(\nu|N) \propto e^{-\nu} P_0(\nu) \int (\nu + \beta)^N e^{-\beta} P_0(\beta) \mathrm{d}\beta$$

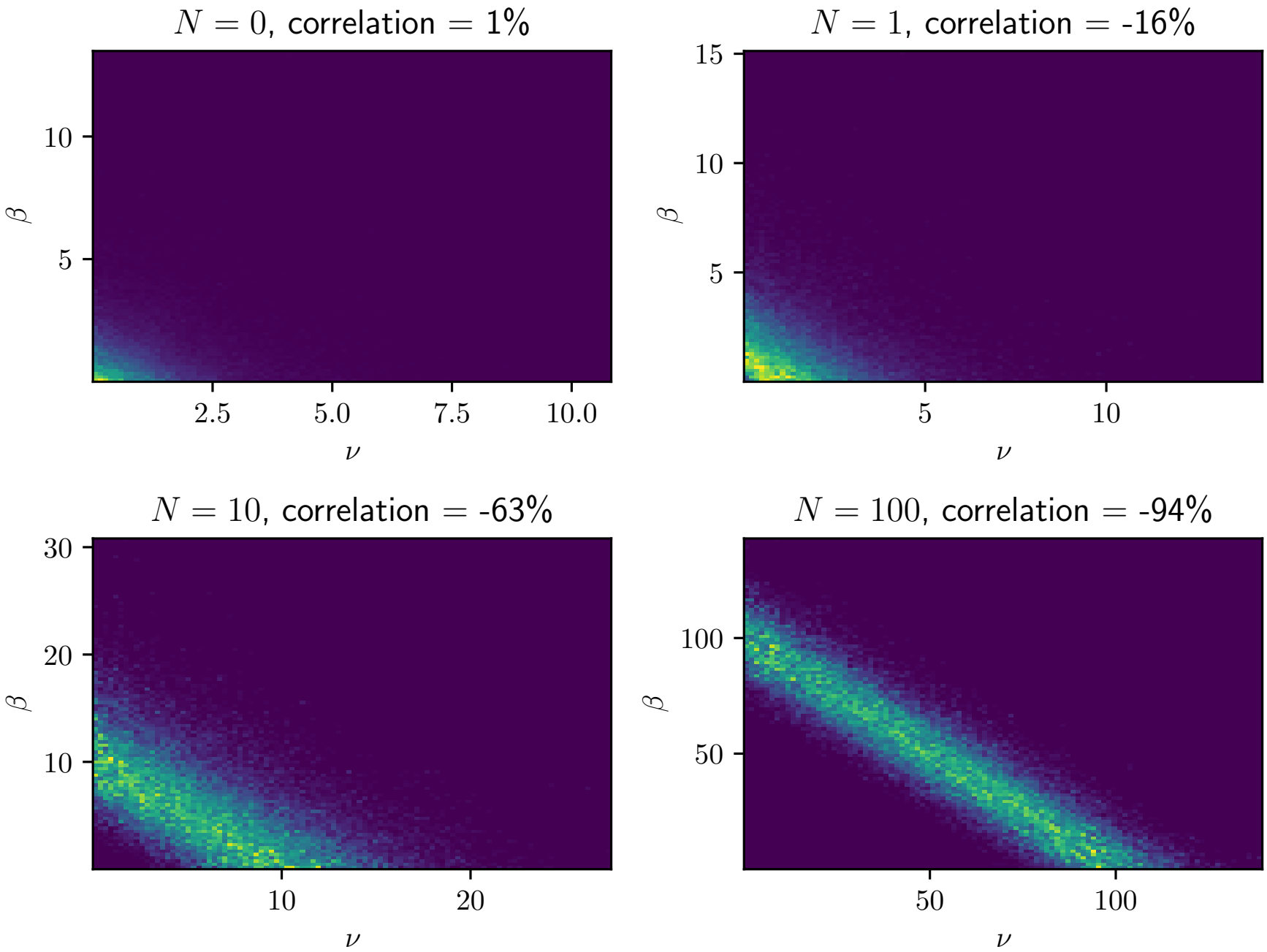
And when $N = 0$:

$$P(\nu|N = 0) \propto e^{-\nu} P_0(\nu) \int e^{-\beta} P_0(\beta) \mathrm{d}\beta$$

That is,

$$P(\nu|N = 0) \propto e^{-\nu} P_0(\nu)$$

$\Rightarrow P(\nu|0)$ is independent of prior knowledge on β!


$N = 0$, correlation = 1%
$N = 1$, correlation = -16%
$N = 10$, correlation = -63%
$N = 100$, correlation = -94%
β
ν
2.5
5.0
7.5
10.0
5
10
15
20
30
50
100