# Nested sampling for frequentist computation

Andrew Fowlie, Sebastian Hoof, and Will Handley (May 2021). In: arXiv: 2105.13923 [physics.data-an]

---

Andrew Fowlie

9 July 2021

Nanjing Normal University

# Overview

1. What is a $p$-value?
2. How to compute small $p$?
3. Nested sampling
4. Results

# What is a *p*-value?

# *P*-values

We can say a lot about *p*-values. Some good (Cousins 2018; Lakens 2021), some bad (Fowlie 2021; Wagenmakers 2007). Here I give only some facts.

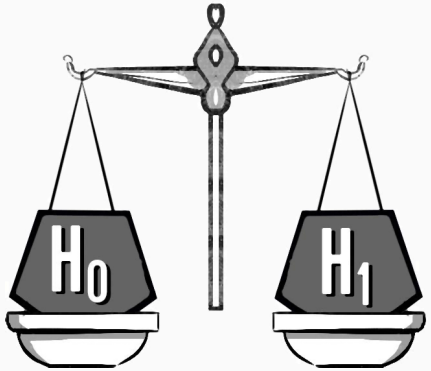
**_P_-value**

The *p*-value, $p$, is the probability of observing data as or more extreme than that observed, given the null hypothesis, $H_0$, i.e.,

$$p = P(\lambda \geq \lambda_{\text{Observed}} \mid H_0)$$

where $\lambda$ is a test-statistic that summarises the data and defines extremeness.

# *P*-values in high-energy physics

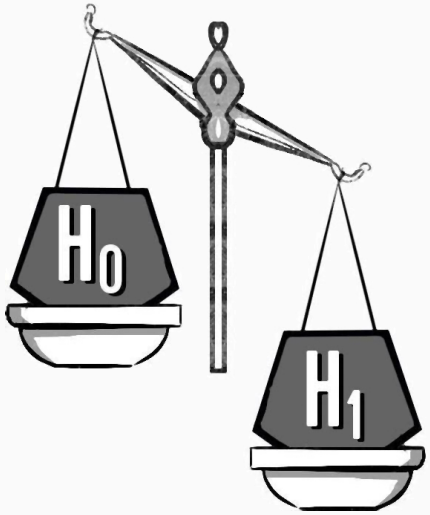



In high-energy physics, we want to discover new phenomena and new particles. Perform null hypothesis test:

- $H_0$ — Standard Model (SM) backgrounds only
- $H_1$ — SM + new physics, e.g. Higgs boson or supersymmetric particles

Original artwork Viktor Beekman and concepts Eric-Jan Wagenmakers

# *P*-values in high-energy physics




For a discovery we conventionally require a tiny global $p$-value of

$$p \lesssim 10^{-7} \, (5\sigma)$$

i.e., $\alpha \simeq 10^{-7}$ (Lyons 2013). In high-energy physics, we need to compute tiny $p$-values.

Threshold in evidence — extraordinary claims require extraordinary evidence — and imposes a $10^{-7}$ type-1 error rate.

Original artwork Viktor Beekman and concepts Eric-Jan Wagenmakers

# Choice of test statistic

Conventionally based on profiled likelihood ratio

$$\lambda = \frac{p\left(\mathbf{x} \mid \hat{\boldsymbol{\Theta}}_1, H_1\right)}{p\left(\mathbf{x} \mid \hat{\boldsymbol{\Theta}}_0, H_0\right)}$$

where $\hat{\boldsymbol{\Theta}}_0$ are the best-fit parameters under $H_0$ etc and $\mathbf{x}$ are the data.

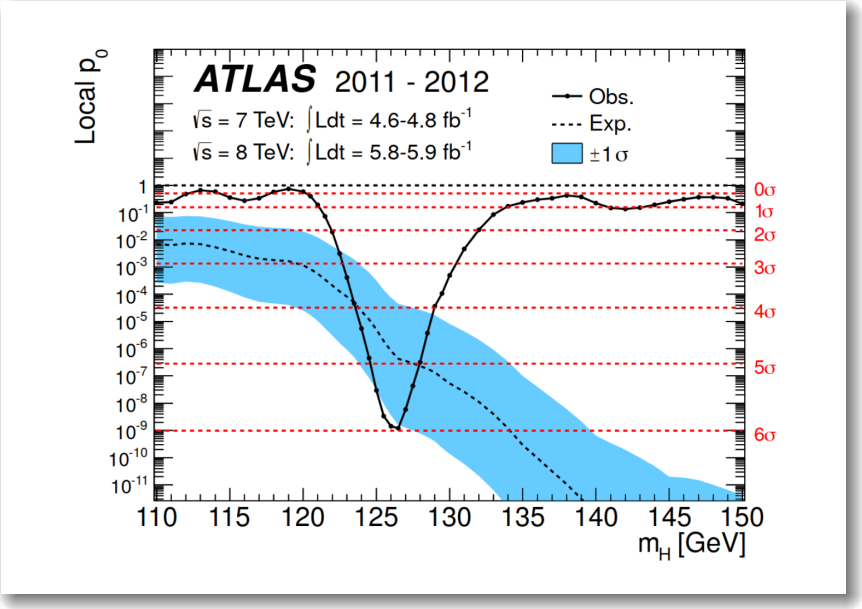
Optimal in simple cases (Neyman-Pearson lemma) and some slightly non-simple cases (Karlin-Rubin theorem).

*Won't dwell on choice of test-statistic in this talk or how to compute it from a given dataset, which could involve multi-dimensional optimisation.*

*Take it as a given.*

# Higgs discovery

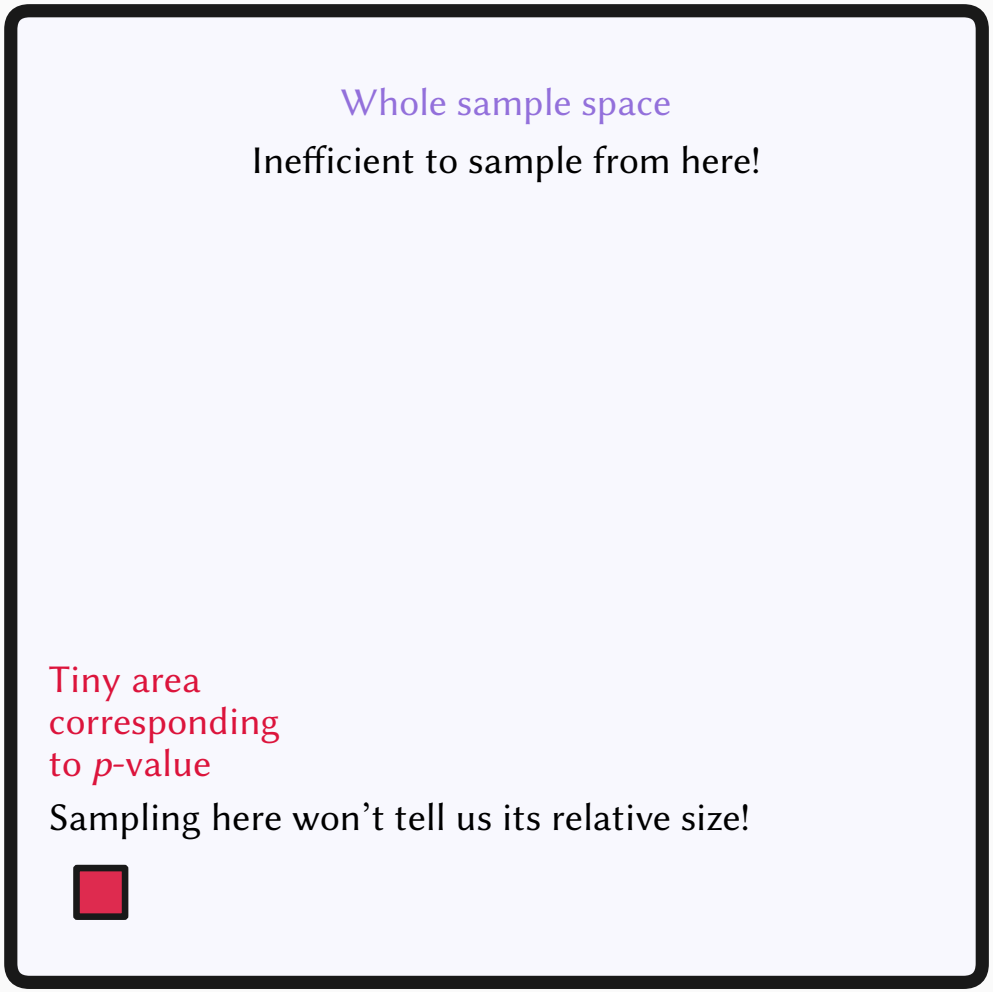
Classic example. Higgs discovery in 2012 (Aad et al. 2012).




Wait until reach $5\sigma$ global. We need to compute tiny $p$-values.

# How to compute small $p$?

Illustrate problem with two-dimensional data, $\mathbf{x}$.



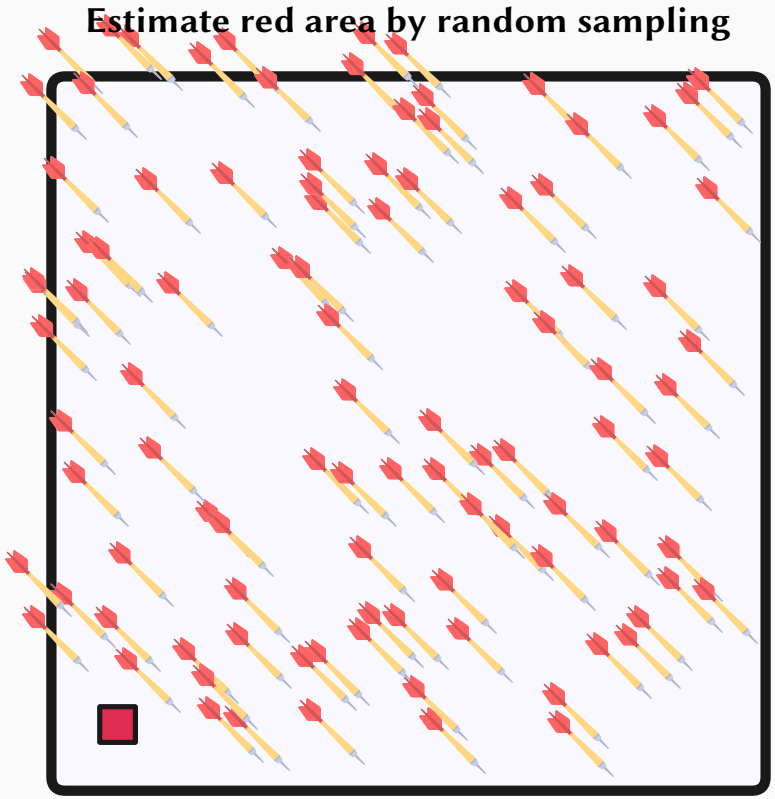

In reality, red region exponentially tiny. Illustrate with squares but assume nothing about geometry/topology in problem or solution.

# Sample from whole sample space — Monte Carlo

Draw $n$ samples from whole sampling distribution. Estimate $p$ by fraction of them that fall in red region

$$\hat{p} = \frac{m}{n}$$

# Sample from whole sample space — Monte Carlo




We really need at least one sample to fall in red region.

# Sample from whole sample space — Monte Carlo

Error of order Wald (Brown, Cai, and DasGupta 2001)

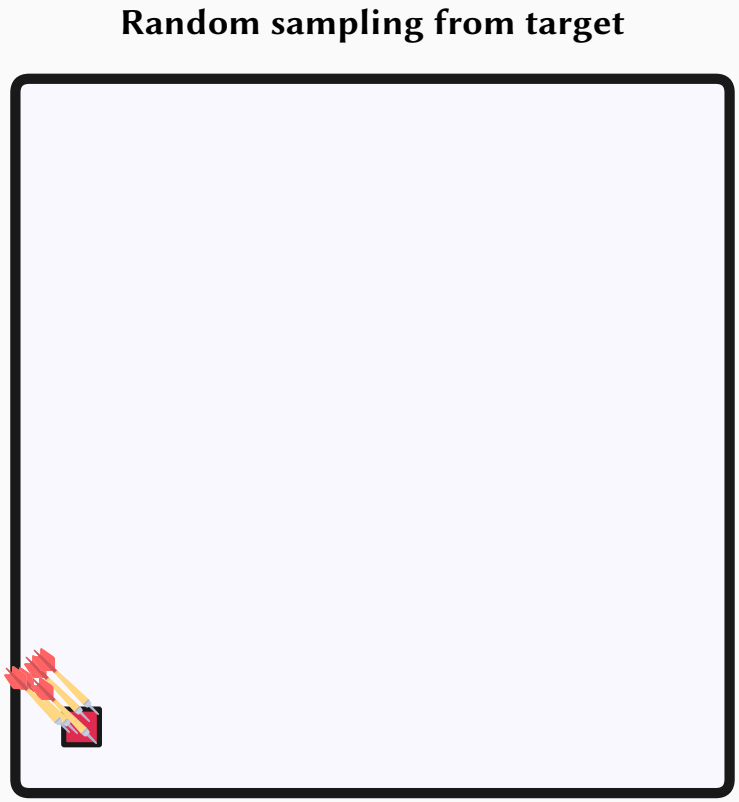
$$\frac{\Delta p}{p} = \sqrt{\frac{1/p}{n}}$$

Usual $1/\sqrt{n}$ statistical error scaling. For fixed fractional uncertainty, number of samples scales as $1/p$

*Need $n \gtrsim 1/p$ at very least for reasonable estimate.*

# Sample from target region

Try to sample from region in which $\lambda \geq \lambda_{\text{Observed}}$.




Even if we could, that won't tell us $p$!

# Asymptotics

If we have some more information, we can compute $p$ analytically.

E.g., I know the red area is a box of side $0.05$. $p = 0.05^2$.

# Asymptotics

If we have some more information, we can compute $p$ analytically.

Sometimes, we know (or hope!) our problem satisfies certain regularity conditions and large sample limit. We can apply asymptotics (Cowan et al. 2011).

E.g., $\lambda \sim \chi^2$. If you're lucky, finding $p$ a matter of calling the right survival function from `scipy.stats` or `root`.

It might still be slightly involved, e.g., simulating to find unknown constants in the Gross-Vitells (Vitells and Gross 2011) method (though at a favourable threshold).

# Asymptotics

If we have some more information, we can compute $p$ analytically.

*Wonderful. Right? But the conditions aren't always satisfied.*

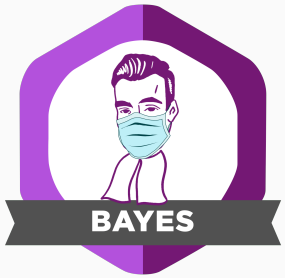
*We want generality. Generality is power to tackle any problem we want.*

# State of play for computing small $p$-values

- Sampling from the entire sampling space inefficient.
- The region correspond to $p$ is important but sampling from it won't tell us $p$!
- Asymptotics needn't apply. When they do, the extra assumptions yield quick answers.

# Compression connection to Bayesian computation

# Connection to Bayesian computation




As for *p*-values, we can say a lot about Bayes. Some good, some bad. As before, only facts.

Bayes' theorem decomposes

$$\text{Likelihood} \times \text{Prior} = \text{Evidence} \times \text{Posterior}$$

Posterior tackled by e.g., Markov Chain Monte Carlo. Method works for ill-normalised distributions. Never need to know evidence.

Original artwork Viktor Beekman and concepts Eric-Jan Wagenmakers

# Bayesian evidence

The Bayesian evidence often ignored, though enables Bayesian model comparison (Kass and Raftery 1995), and so some effort made to compute it efficiently (Martin, Frazier, and Robert 2020).

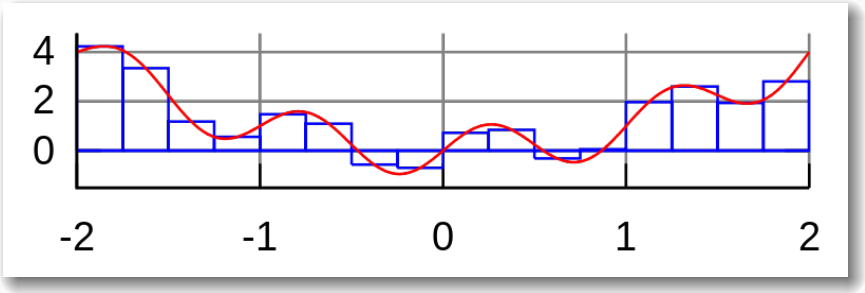
Integral of the likelihood over the model's parameter space

$$\text{Evidence} = \mathcal{Z} = \int \mathcal{L}(\mathbf{x})\pi(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

*High-dimensional integral. Not easy to compute.*

# Don’t even think about it

Quadrature won’t work in high-dimensions. Curse of dimensionality. We would need $\mathcal{O}(e^d)$ cells.



# Sample form the prior

Just try sampling from the prior!

$$\hat{\mathcal{Z}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathbf{x}_i)$$

for $\mathbf{x}_i \sim$ Prior. Doesn't care about dimensionality!

But prior could be very different from posterior! The problem is compression. Most draws contribute nothing. Just like for $p$-values.

Usual $1/\sqrt{n}$ scaling, but potentially terrible variance.

# Sample from the target

This won't work either. Just like for $p$-values.

**Heuristic (Neal 2008)**

The posterior depends only weakly on prior. Evidence depends strongly on it.

Thus hard to deduce evidence from posterior.

E.g., likelihood concentrated around $\theta \simeq 0$ and zero outside $|\theta| > 1$. Extending prior range from $|\theta| \leq 1$ to $|\theta| \leq 1000$ makes no difference to posterior. It changes evidence by factor 1000.

# Asymptotics

Well, the posterior might be asymptotically normal (Bernstein-von Mises theorem). Approximate the integral by a Gaussian! This is a Laplace approximation.

- Fast — we know how to integrate Gaussians, just need to fit it to the posterior
- But not reliable in so many problems of interest
- You'll get a fast answer, maybe not a correct answer.

# State of play for computing evidence

- Sampling from the prior inefficient — enormous compression.
- Posterior important, but sampling from it won't easily tell us the evidence.
- Asymptotics (e.g., Laplace approximation) won't always hold – we want generality.

# Sound familiar?

- Sampling from the ~~prior~~ whole sampling space inefficient — enormous compression.
- ~~Posterior~~ Region corresponding to $p$ important, but sampling from it won't easily tell us the ~~evidence~~ $p$-value
- Asymptotics (e.g., ~~Laplace~~ Wilks' approximation) won't always hold – we want generality.

# Path sampling

Path sampling could be the answer (Gelman and Meng 1998).

- Prior and posterior miles apart — enormous compression.
- Build a path — a sequence of distributions — between them.
- Evolve a collection of particles along that path.

You may be familiar with annealing — cool from the prior to the posterior.

# Path sampling

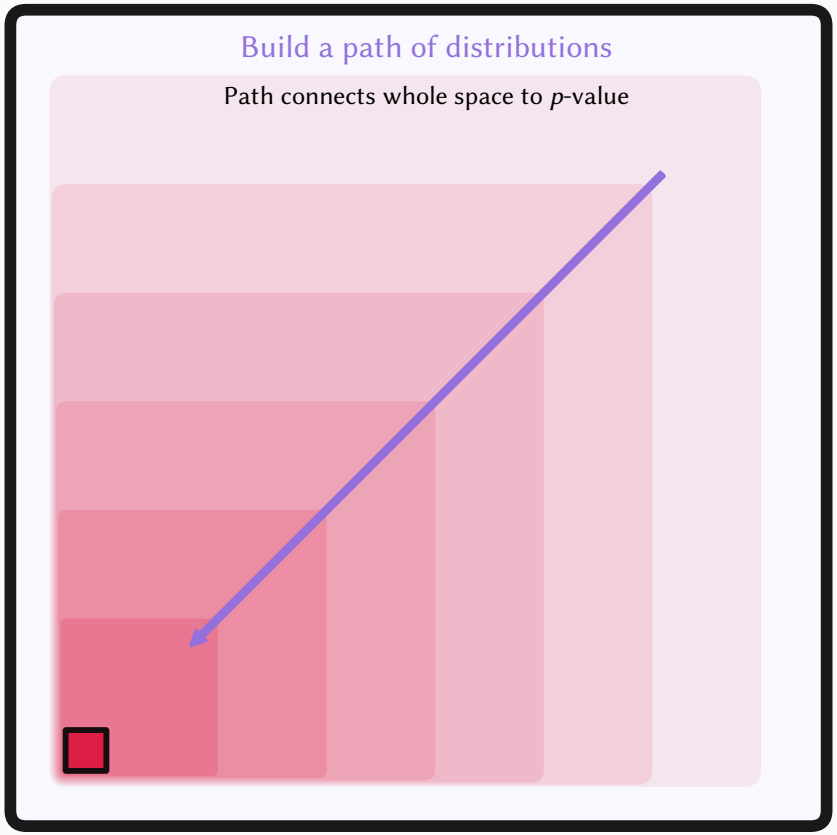
Path sampling could be the answer.

- ~~Prior~~ Sampling space and ~~posterior~~ region corresponding to $p$ miles apart — enormous compression.
- Build a path — a sequence of distributions — between them.
- Evolve a collection of particles along that path.

You may be familiar with annealing — cool from the prior to the posterior.

# Path sampling

**Solution — path sampling**




Sequence of distributions from whole sampling space to the $p$-value.

# Which path to the *p*-value?


Which path though? Don't claim optimality. But what follows is simple and it gets us there.

# Which path to the *p*-value?

Natural that the path of intermediate distributions are based on contours of test-statistic

$$\pi^\star(\mathbf{x}) \propto \begin{cases} \pi(\mathbf{x}) & \lambda(\mathbf{x}) \geq \lambda^\star \\ 0 & \text{otherwise} \end{cases}$$

for some threshold $\lambda^\star$. This is the constrained sampling distribution.

We start at $\lambda^\star = -\infty$ with the whole sampling distribution. The threshold monotonically increases along the path, until we reach the region corresponding to the *p*-value at $\lambda^\star = \lambda_{\text{Observed}}$.

# Which path to the *p*-value?

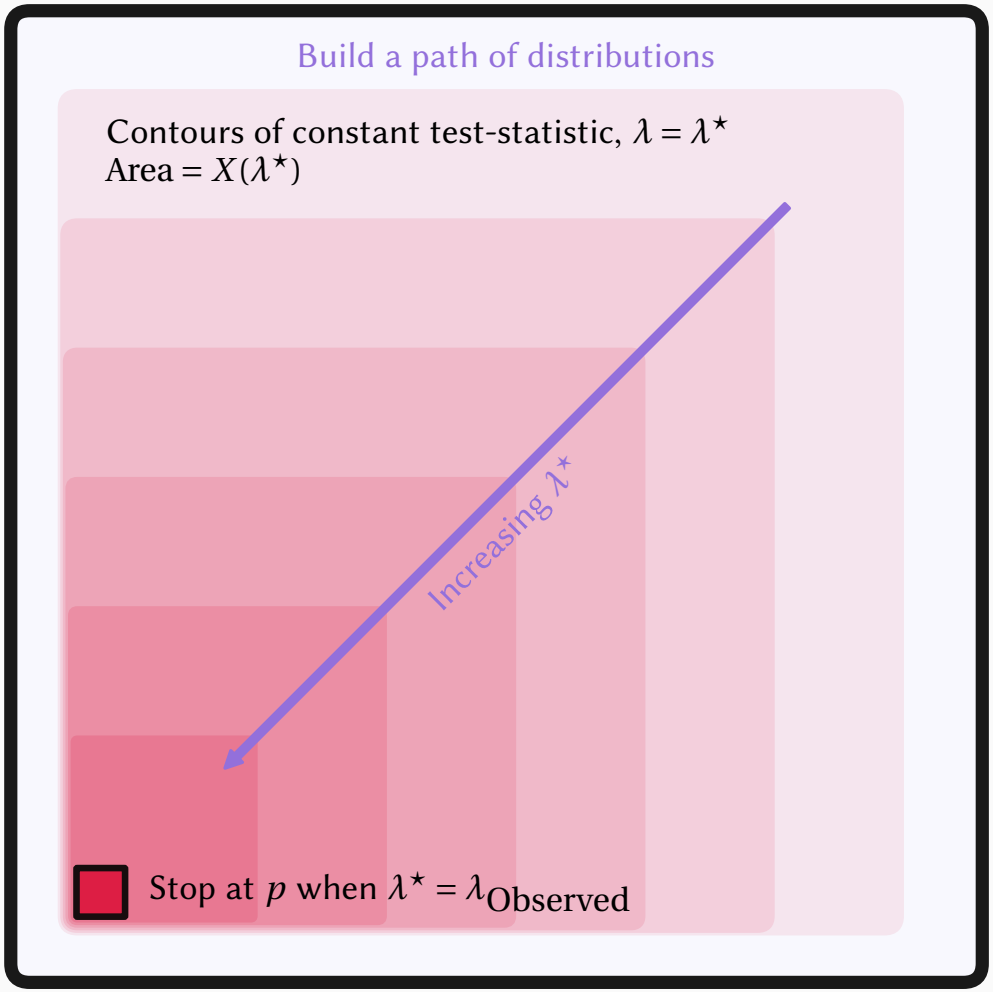
The survival function

$$X(\lambda^\star) \equiv \int_{\lambda(\mathbf{x}) \geq \lambda^\star} \pi(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

tells us the compression required to reach the threshold $\lambda^\star$.

Along the path, the survival function monotonically decreases from $X = 1$ to the $p$-value at $p = X(\lambda_{\text{Observed}})$.

# Path sampling



# How fast to travel on the path to the $p$-value?

We have the path. How fast should we go? How about controlling the speed of compression through the survival function?

**Take steps $\Delta X = \text{const.}$?**

- Start at $X_0 = 1$. $X_1 = 1 - \Delta X \ldots$ Stop at $X_i = 1 - i\Delta X \leq p$.
- Need $\Delta X \approx p$ for any reasonable precision.
- Too slow. Would require $1/p$ iterations to reach $p$ and we're back to $1/p$ scaling!

# How fast to travel on the path to the $p$-value?

We have the path. How fast should we go? How about controlling the speed of compression through the survival function?

**Take steps $\Delta \log X = \text{const.}$**

- Go faster. Constant exponential compression.
- Avoid $1/p$.
- Let's try that.

# How to find the path?

The sampling space could be complicated. It won't be nice squares like my pictures.

*How on Earth can we evolve a collection of particles along path whilst compressing at a constant exponential rate?*

*And how can we estimate the survival function along the way?*

# Nested sampling


Original artwork Viktor Beekman and concepts Eric-Jan Wagenmakers

# Nested sampling

Nested sampling (Skilling 2004; Skilling 2006) originally algorithm for Bayesian computation — computes evidence and posterior simultaneously.

- Evolves collection of $n_{\text{live}}$ live points along path, to greater and greater ~~likelihoods~~ test-statistics
- Evolution controlled by single user parameter — $n_{\text{live}}$
- Approximately constant exponential compression
- Statistical estimates of survival function along the way
- All about compression — no need to even talk about evidence or posterior here

# General idea

- Sample $n$ points from the sampling distribution
- Rank them by test-statistic
- Delete the least extreme half

*You just compressed by factor* $1/2$*! Repeat it* $i$ *times and you'll achieve exponential compression* $1/2^i$ *at a constant rate!*

# Nested sampling

0. Draw $n_{\text{live}}$ samples from the whole sampling distribution — the live points

So we evolve a set of $n_{\text{live}}$ live points to more and more extreme test-statistics, replacing one live point at a time.

# Nested sampling

0. Draw $n_{\text{live}}$ samples from the whole sampling distribution — the live points
1. Denote the smallest test-statistic among the live points by $\lambda^\star$

So we evolve a set of $n_{\text{live}}$ live points to more and more extreme test-statistics, replacing one live point at a time.

# Nested sampling

0. Draw $n_{\text{live}}$ samples from the whole sampling distribution — the live points
1. Denote the smallest test-statistic among the live points by $\lambda^\star$
2. Replace that live point by one drawn from the constrained sampling distribution

$$\pi^\star(\mathbf{x}) \propto \begin{cases} \pi(\mathbf{x}) & \lambda(\mathbf{x}) \geq \lambda^\star \\ 0 & \text{otherwise} \end{cases}$$

So we evolve a set of $n_{\text{live}}$ live points to more and more extreme test-statistics, replacing one live point at a time.

# Nested sampling

0. Draw $n_{\text{live}}$ samples from the whole sampling distribution — the live points
1. Denote the smallest test-statistic among the live points by $\lambda^\star$
2. Replace that live point by one drawn from the constrained sampling distribution

$$\pi^\star(\mathbf{x}) \propto \begin{cases} \pi(\mathbf{x}) & \lambda(\mathbf{x}) \geq \lambda^\star \\ 0 & \text{otherwise} \end{cases}$$

3. Make a statistical estimate of $X(\lambda^\star)$ from this procedure

So we evolve a set of $n_{\text{live}}$ live points to more and more extreme test-statistics, replacing one live point at a time.

# Nested sampling

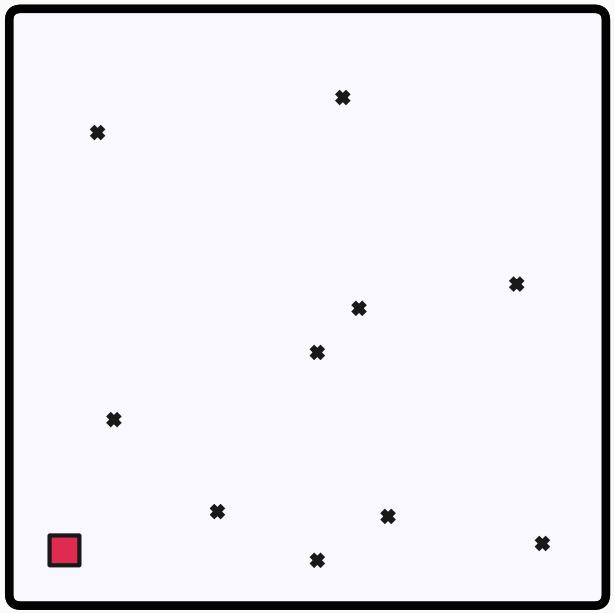
0. Draw $n_{\text{live}}$ samples from the whole sampling distribution — the live points
1. Denote the smallest test-statistic among the live points by $\lambda^\star$
2. Replace that live point by one drawn from the constrained sampling distribution

$$\pi^\star(\mathbf{x}) \propto \begin{cases} \pi(\mathbf{x}) & \lambda(\mathbf{x}) \geq \lambda^\star \\ 0 & \text{otherwise} \end{cases}$$

3. Make a statistical estimate of $X(\lambda^\star)$ from this procedure
4. If $\lambda^\star \geq \lambda_{\text{Observed}}$, we reached $p$. Stop. Else go to 1.

So we evolve a set of $n_{\text{live}}$ live points to more and more extreme test-statistics, replacing one live point at a time.

## Nested sampling



Live Dead Replacement

1. Uniformly distributed live points
2. Identify least extreme point
3. Update threshold
4. Draw replacement

# Compression

*Make a statistical estimate of $X(\lambda^\star)$ from this procedure*

How? We know that $X_0 = 1$. How much do we expect $X$ to contract when we replace the least extreme test-statistic?

# Estimating the compression

Drawing from the constrained sampling distribution means live points are distributed uniformly in $X$ from 0 to $X(\lambda^\star)$.

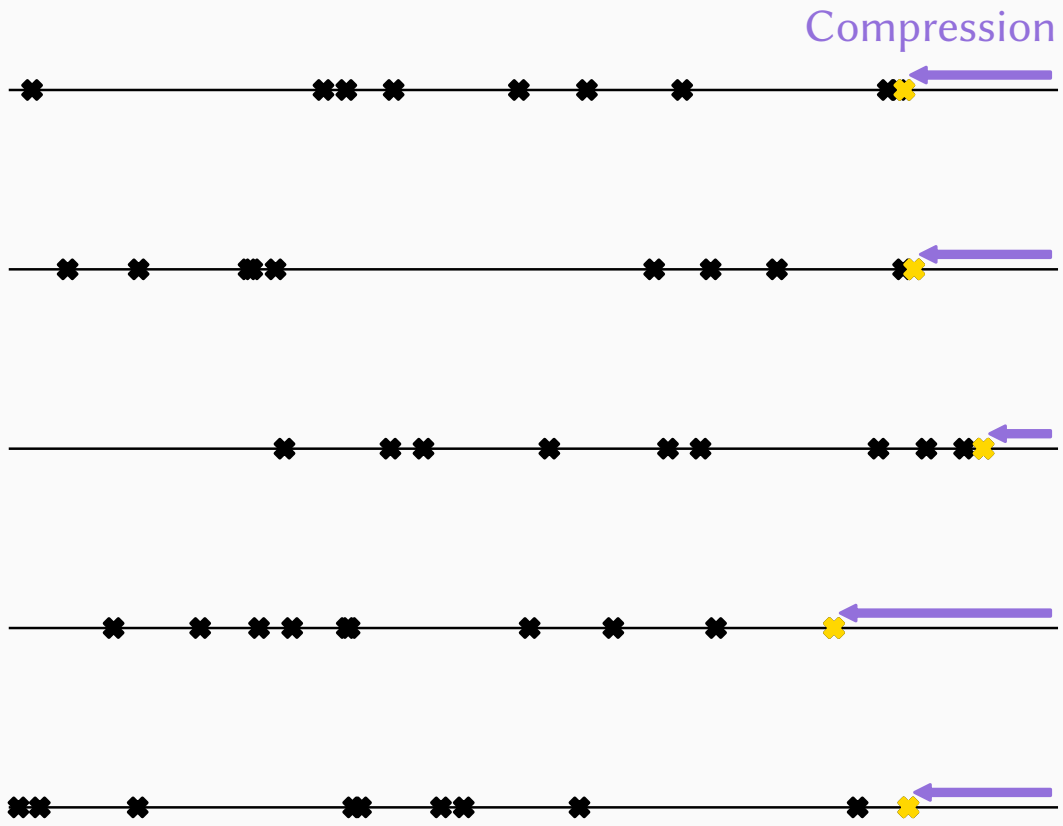
In other words,

$$y_i = \frac{X(\lambda_i)}{X(\lambda^\star)}$$

for $i = 1$ to $n_{\text{live}}$ are uniformly distributed from 0 to 1.

# Estimating the compression

The largest one, $t \equiv \max y_i$, gives us the compression.



# Estimating the compression

What can we say about $t$? We can write the pdf!

$$p(t) = \binom{n_{\text{live}}}{1} \cdot t^{n_{\text{live}}-1} \cdot 1 = n_{\text{live}} t^{n_{\text{live}}-1}$$

where the factors are combinatorial, the probability of $n_{\text{live}} - 1$ samples less than $t$, and lastly the probability density of a point at $t$.

This is a $\beta(n_{\text{live}}, 1)$ distribution.

# Estimating the compression

We can find the expected compression:

$$\mathrm{E}[\log t] = n_{\text{live}} \int_0^1 t^{n_{\text{live}}-1} \log t \, \mathrm{d}t = -\frac{1}{n_{\text{live}}}$$

Look at $\log t$ because logarithms add and expectation is linear.

*Thus we may estimate that at iteration $i$*

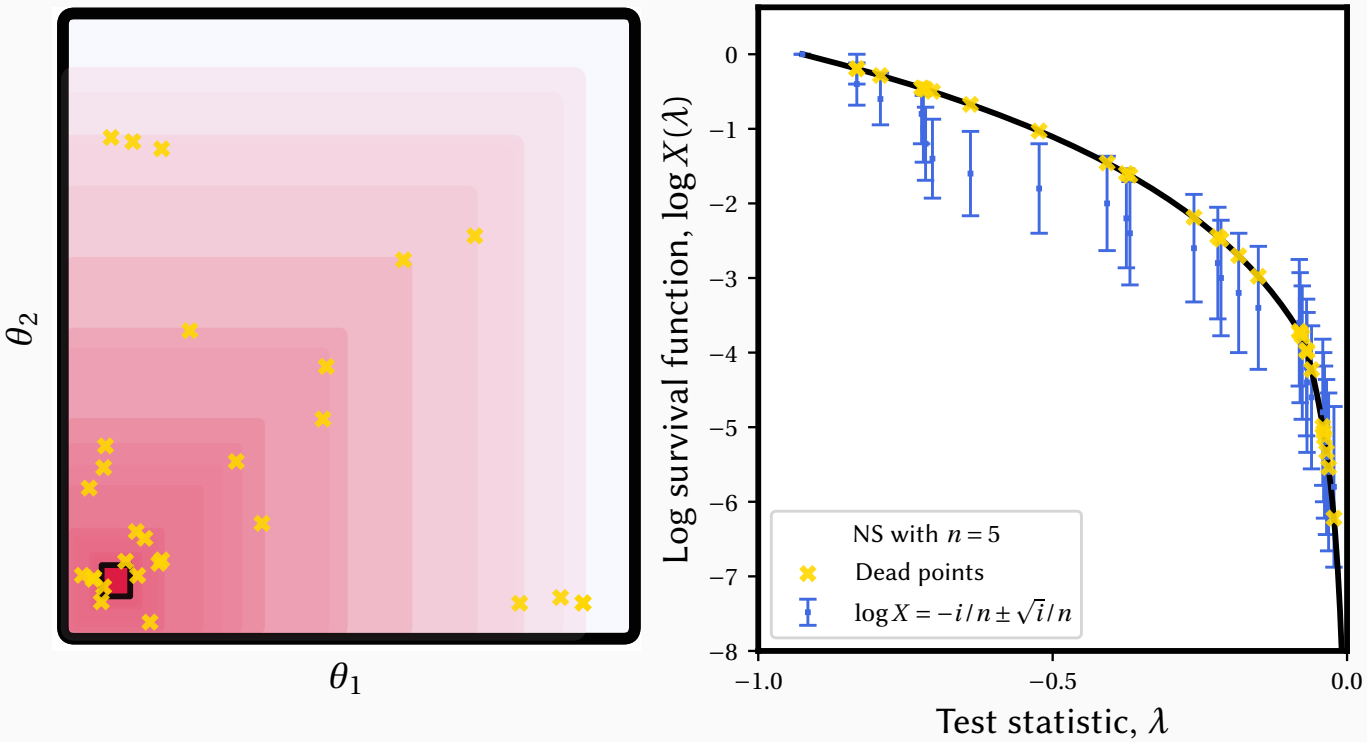
$$X(\lambda_i^\star) = \prod_{j=1}^{i} t_j \approx \prod_{j=1}^{i} e^{-1/n_{live}} = e^{-i/n_{live}}$$

*This is the desired constant exponential compression! and a way to estimate it!*

The user parameter $n_{\text{live}}$ controls the speed and consequently the precision.

# Estimating the compression

We don't know $X$ but make statistical estimates of it along the path



# Stopping

*NS builds a path straight to the p-value in roughly equal steps of $\Delta \log p \simeq 1/n_{live}$. We just stop once we're there.*

Stop as soon as $\lambda^\star \geq \lambda_{\text{Observed}}$ at iteration $n_{\text{iter}}$. Estimate $p$ from compression

$$p = X(\lambda_{\text{Observed}}) \simeq \prod_{i=1}^{n_{\text{iter}}} e^{-1/n_{\text{live}}} = e^{-n_{\text{iter}}/n_{\text{live}}}$$

Thus NS breaks tiny $p$ into a product of moderate factors.

# Assumptions

*Basically no assumptions about problem, no asymptotics, completely general*

- Assumes we can evaluate the test-statistic
- Assumes we can draw from the sampling distribution
- Plateaus in the test-statistic (regions of sample space where it is constant) cause subtelty but easily overcome (Fowlie, Handley, and Su 2020b).

For it to work efficiently, further assumes we can efficiently and correctly draw from the constrained sampling distribution

# Uncertainties

Let's look at the errors from the $\beta$ distributions. As $\log t \sim \beta(n_{\text{live}}, 1)$,

$$\text{Var}[\log t] = \frac{1}{n_{\text{live}}^2}$$

As $\log p$ is estimated as the sum of $n_{\text{iter}}$ independent $\log t$ draws,

$$\text{Var}[\log p] = n_{\text{iter}}\text{Var}[\log t] = \frac{n_{\text{iter}}}{n_{\text{live}}^2} = \frac{\log 1/p}{n_{\text{live}}}$$

where we plugged in the estimate of $\log p$. Finally,

$$\frac{\Delta p}{p} = \Delta \log p = \sqrt{\frac{\log 1/p}{n_{\text{live}}}}$$

Note the usual $1/\sqrt{n}$ scaling.

# Uncertainties — log $p$ not $p$

Exponential compression. Estimating log $p$ not $p$.

It is log $p$ that has approximately symmetric, Gaussian uncertainty.

*Just fine. When p is small, it is magnitude* log $p$ *that matters.*

# What did we gain? — Theoretical speedup

For nested sampling,

$$\frac{\Delta p}{p} = \sqrt{\frac{\log 1/p}{n_{\text{live}}}}$$

For Monte Carlo,

$$\frac{\Delta p}{p} = \sqrt{\frac{1/p}{n}}$$

Looking good. Ripped an exponential factor out of the problem.

# Closer look at theoretical speedup

We need to write the behaviour in terms of test-statistic evaluations rather than $n_{\text{live}}$,

$$n = \text{average calls per iteration} \cdot n_{\text{iter}} = \frac{n_{\text{iter}}}{\epsilon}$$

We expect though that $n_{\text{iter}} = n_{\text{live}} \log 1/p$ from the exponential compression,

$$n = \frac{n_{\text{live}} \log 1/p}{\epsilon}$$

Thus finally,

$$\frac{\Delta p}{p} = \sqrt{\frac{\log^2 1/p}{\epsilon n}}$$

# Closer look at theoretical speedup

For fixed fractional uncertainty on $p$, we expect to obtain a speed-up versus MC

$$\frac{\text{Evaluations for NS}}{\text{Evaluations for MC}} = \frac{(\log^2 1/p)/\epsilon}{1/p}$$

*Massive gains for small p! Provided that the efficiency factor $\epsilon$ doesn't spoil things.*

# Exploration

# Exploration

So far nothing depended on dimension! Let alone geometry! I glossed over a detail though.

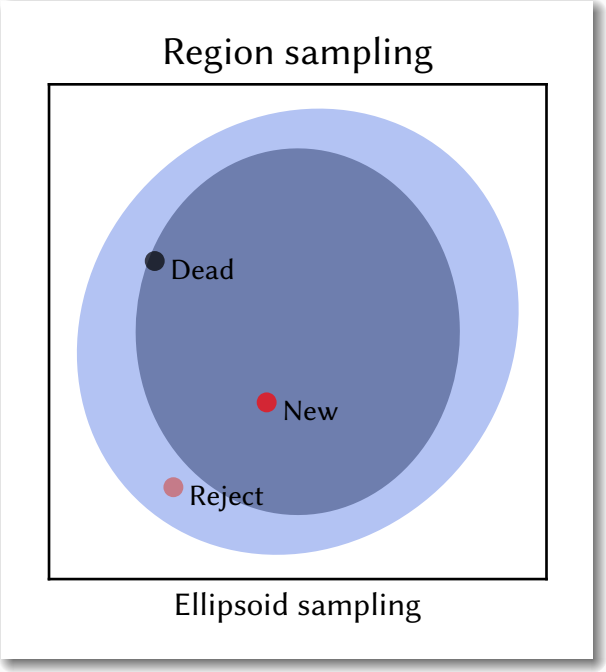
- How to draw a replacement live point from the constrained sampling distribution, $\lambda > \lambda^\star$?
- Involves trial and error, and some inefficiency (the factor $\epsilon$)
- Re-introduces dependence on dimensionality of the sampling space (though it needn't be exponential)

# Exploration

- This requires an exploration strategy. Fortunately, the current set of live points can guide the exploration.
- There are well-established implementations of nested sampling that do this, developed and used for Bayesian inference in other scientific settings
- There are potentially optimisations for this setting
- Correctness can be checked (Fowlie, Handley, and Su 2020a)

# Rejection sampling

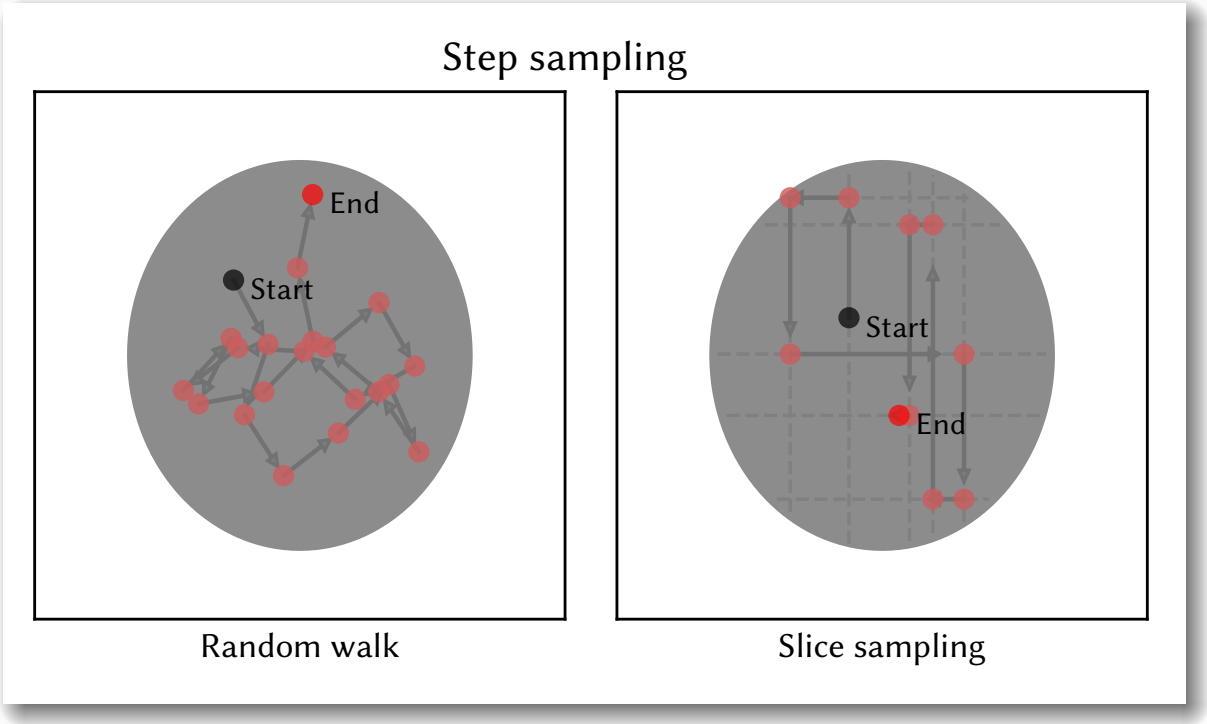
`MultiNest` (Feroz and Hobson 2008; Feroz, Hobson, and Bridges 2009; Feroz et al. 2013) — bound live points by ellipsoids. Use them to approximate the $\lambda^\star$ contour. Sample from the ellipsoids.




Rejection sampler — efficient at small $d$, but curse of dimensionality ultimately strikes.

# Step samplers

Take walk starting from a randomly chosen existing live point. E.g., `PolyChord` (Handley, Hobson, and Lasenby 2015a; Handley, Hobson, and Lasenby 2015b) — slice sampling walk.




Efficient and good scaling, $\epsilon \propto 1/d$.

# Results

# Does it work?

That's all for theory. We need some numerical investigation.

*Does it work as I say it does?*

*Do we benefit from the* $\log^2 1/p$ *scaling? or does the* $\epsilon$ *efficiency factor spoil things?*

Here are our results, you should try it too!

# A simple problem

The $p$-value associated with $d$ independent Gaussian measurements

- $d$ dimensional sampling space, $\mathbf{x}_i \sim \mathcal{N}(\mu, \sigma^2)$.
- Test-statistic

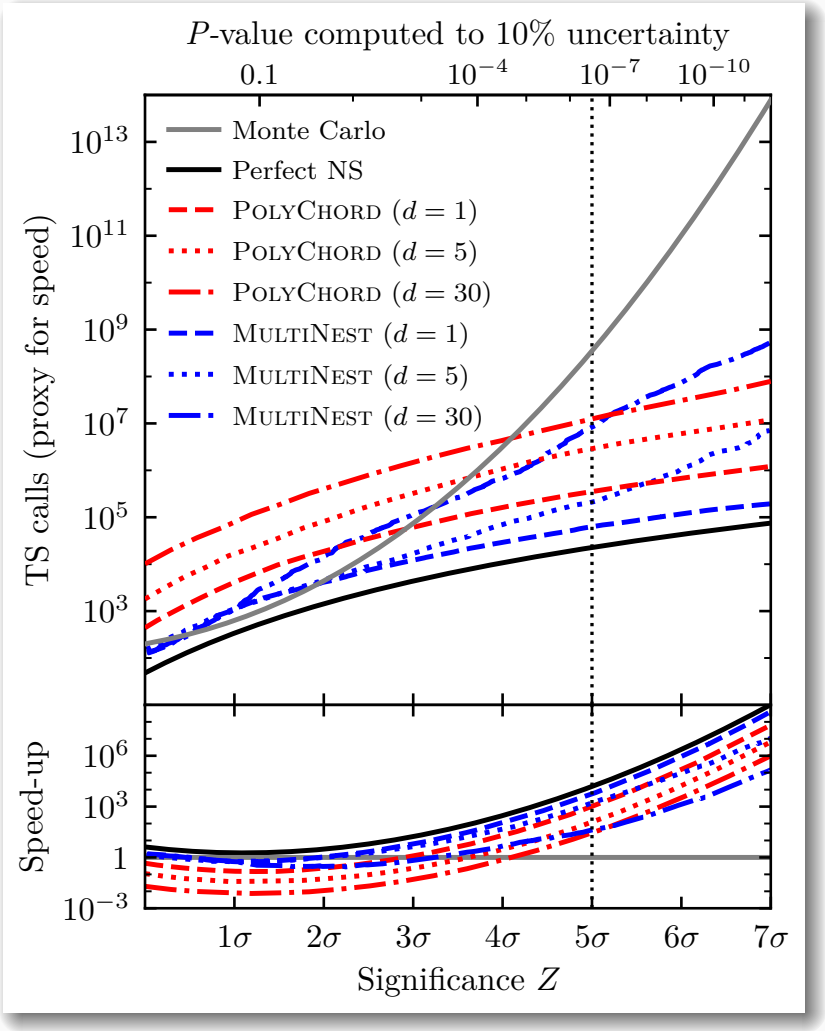
$$\lambda = \sum_{i=1}^{d} \left( \frac{\mathbf{x}_i - \mu}{\sigma} \right)^2$$

- We know analytically

$$\lambda \sim \chi^2_d \quad \text{such that} \quad p = 1 - F_{\chi^2_d}(\lambda_{\text{Observed}})$$

- Toy example that allows us to easily control dimension, size of $p$ and check correctness

# Number of evaluations for fixed fractional uncertainty




Perfect NS means if 100% efficiency, $\epsilon = 1$, was possible

# Performance on simple problem

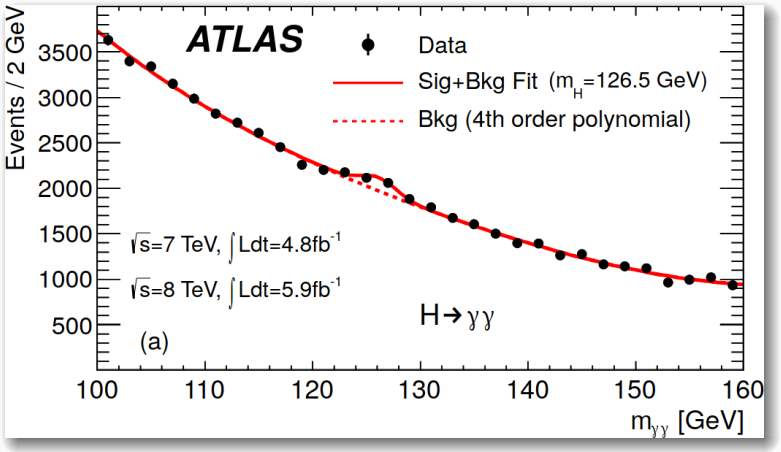
- For large $p$, nested sampling could only narrowly beat MC in the best-case scenario, and typically worse. MC just fine when $p$ is moderate
- For small $p \lesssim 4\sigma$, the scaling kicks in. Nested sampling wins by orders of magnitude
- Nested sampling performance depends on dimensionality but even for $30d$ sampling space, winning by $10^6$ at $7\sigma$

# Higgs-like example

That was a warm-up. How about a resonance search?

- Simplified version of the original Higgs discovery by ATLAS (Aad et al. 2012) in the diphoton channel
- $H_0$ — Standard Model (SM) background-only hypothesis, with a known shape and an unknown total number of background events
- $H_1$ — SM + a Higgs boson with a Gaussian signal, with a known width but an unknown mass and an unknown positive signal strength

# Higgs-like example

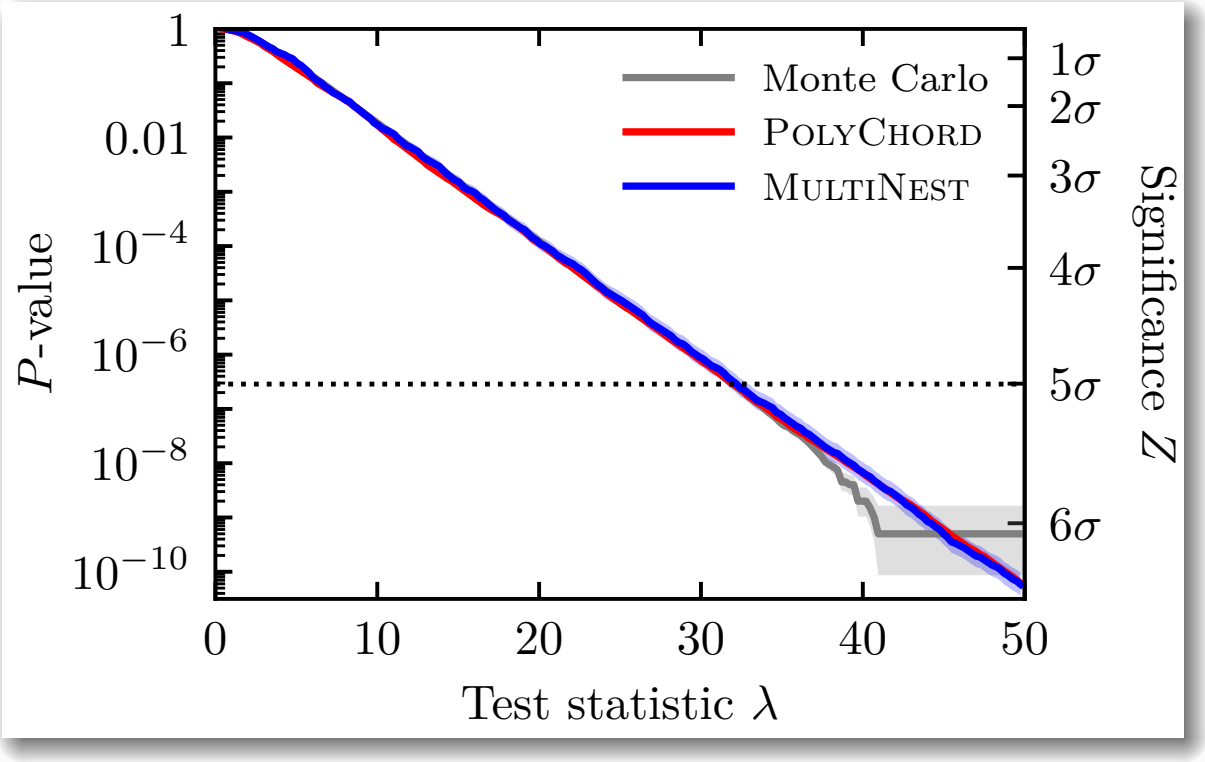



- 30 dimensional sampling space, $\mathbf{x}$ — the Poisson-distributed counts in the 30 bins
- Log-likelihood ratio test-statistic with unknown parameters profiled

$$\lambda(\mathbf{x}) = 2 \log \left( \frac{\max P(\mathbf{x} \mid m_h, \mu, b)}{\max P(\mathbf{x} \mid \mu = 0, b)} \right)$$

- Likelihood just a product of 30 Poissons.

# Result on Higgs-like example

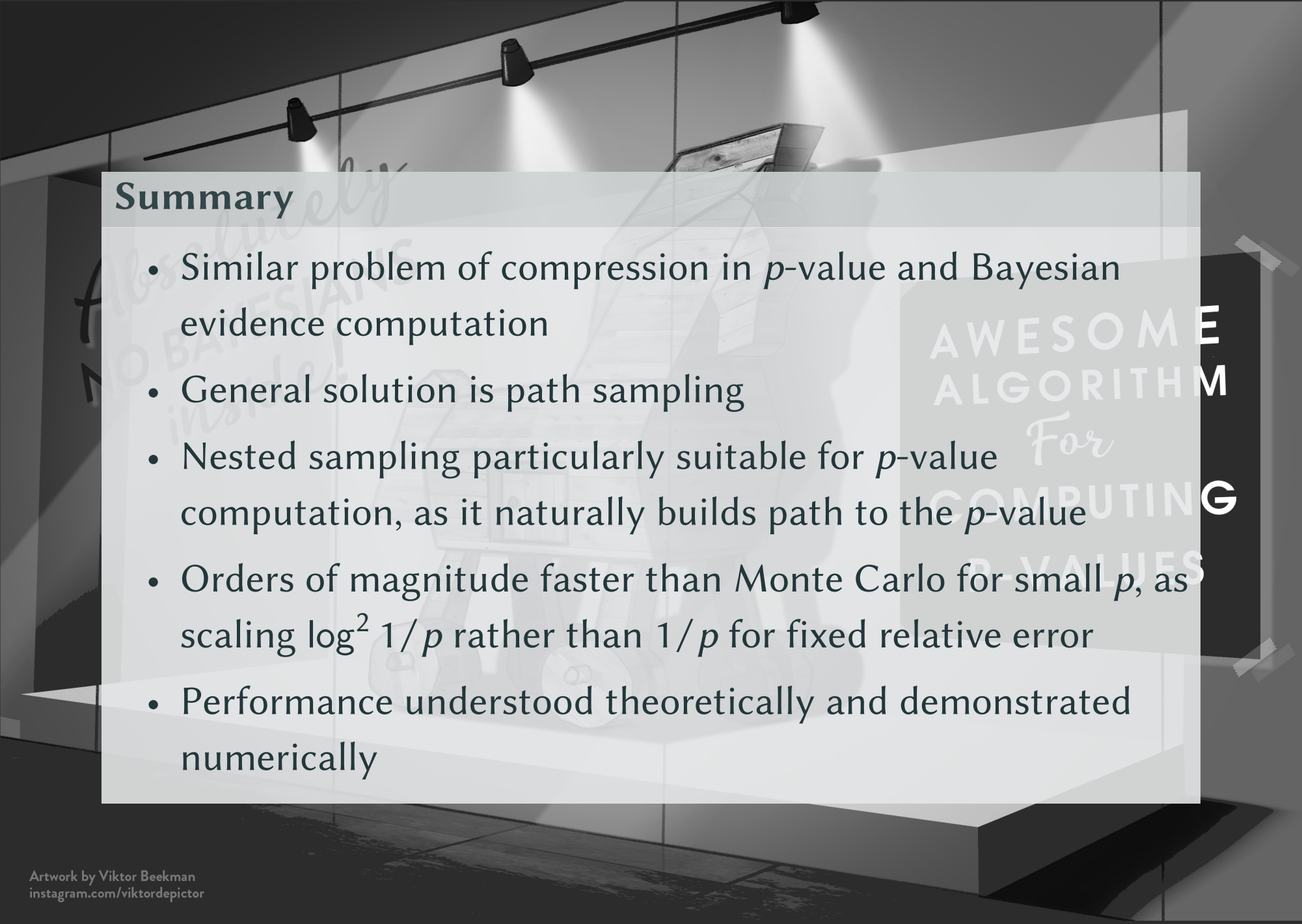



Compute $p$ for increasing test-statistic, $\lambda$

# Result on Higgs-like example

- Results consistent between MC, NS and Gross-Vitells predicted slope to $6\sigma$ and beyond
- To reach $6.5\sigma$ with uncertainty $\Delta \log_{10} p \approx 0.2$, `PolyChord` needed $3 \times 10^6$ calls and `MultiNest` needed $4 \times 10^7$
- To reach a similar uncertainty, MC would require $10^{11}$ MC simulations.
- Nested sampling winning by about $10^5$ for `PolyChord`


Absolutely
NO BAYESIANS
inside!
Summary
AWESOME
ALGORITHM
For
COMPUTING
P-VALUES
Artwork by Viktor Beekman
instagram.com/viktordepictor




## Summary

- Similar problem of compression in $p$-value and Bayesian evidence computation
- General solution is path sampling
- Nested sampling particularly suitable for $p$-value computation, as it naturally builds path to the $p$-value
- Orders of magnitude faster than Monte Carlo for small $p$, as scaling $\log^2 1/p$ rather than $1/p$ for fixed relative error
- Performance understood theoretically and demonstrated numerically

Artwork by Viktor Beekman
instagram.com/viktordepictor

# References

Aad, Georges et al. (2012). "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." In: Phys. Lett. B 716, pp. 1–29. arXiv: 1207.7214 [hep-ex].

Brown, Lawrence D., T. Tony Cai, and Anirban DasGupta (2001). "Interval Estimation for a Binomial Proportion." In: Statistical Science 16.2, pp. 101–117. URL: http://www.jstor.org/stable/2676784.

Cousins, Robert D. (July 2018). "Lectures on Statistics in Theory: Prelude to Statistics in Practice." In: arXiv e-prints, arXiv:1807.05996, arXiv:1807.05996. arXiv: 1807.05996 [physics.data-an].

# References ii


Cowan, Glen et al. (2011). "Asymptotic formulae for likelihood-based tests of new physics." In: Eur. Phys. J. C 71. [Erratum: Eur.Phys.J.C 73, 2501 (2013)], p. 1554. arXiv: 1007.1727 [physics.data-an].

Feroz, F., M. P. Hobson, and M. Bridges (2009). "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics." In: Mon. Not. Roy. Astron. Soc. 398, pp. 1601–1614. arXiv: 0809.3437 [astro-ph].

Feroz, F. et al. (2013). "Importance Nested Sampling and the MultiNest Algorithm." In: The Open Journal of Astrophysics. arXiv: 1306.2144 [astro-ph.IM].

# References iii


Feroz, Farhan and M. P. Hobson (2008). "Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis." In: Mon. Not. Roy. Astron. Soc. 384, p. 449. arXiv: 0704.3704 [astro-ph].

Fisher, R. A. (1925). Statistical Methods for Research Workers. Oliver and Boyd.

Fowlie, Andrew (Apr. 30, 2021). "Comment on "Reproducibility and Replication of Experimental Particle Physics Results"." In: Harvard Data Science Review. arXiv: 2105.03082 [physics.data-an].

Fowlie, Andrew, Will Handley, and Liangliang Su (2020a). "Nested sampling cross-checks using order statistics." In: Mon. Not. Roy. Astron. Soc. 497.4, pp. 5256–5263. arXiv: 2006.03371 [stat.CO].

# References iv


Fowlie, Andrew, Will Handley, and Liangliang Su (Oct. 2020b). "Nested sampling with plateaus." In: arXiv: 2010.13884 [stat.CO].

Fowlie, Andrew, Sebastian Hoof, and Will Handley (May 2021). "Nested sampling for frequentist computation: fast estimation of small $p$-values." In: arXiv: 2105.13923 [physics.data-an].

Gelman, Andrew and Xiao-Li Meng (1998). "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling." In: Statistical Science 13.2, pp. 163–185. URL: https://doi.org/10.1214/ss/1028905934.

Handley, W. J., M. P. Hobson, and A. N. Lasenby (2015a). "PolyChord: nested sampling for cosmology." In: Mon. Not. Roy. Astron. Soc. 450.1, pp. L61–L65. arXiv: 1502.01856 [astro-ph.CO].

# References v


Handley, W. J., M. P. Hobson, and A. N. Lasenby (Nov. 2015b). "PolyChord: next-generation nested sampling." In: Mon. Not. Roy. Astron. Soc. 453.4, pp. 4384–4398. arXiv: 1506.00171 [astro-ph.IM].

Hubbard, Raymond and M. J Bayarri (2003). "Confusion Over Measures of Evidence ($p$'s) Versus Errors ($\alpha$'s) in Classical Statistical Testing." In: Am. Stat. 57.3, pp. 171–178.

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors." In: J. Am. Stat. Assoc. 90.430, pp. 773–795.

Lakens, Daniël (2021). "The Practical Alternative to the p Value Is the Correctly Used p Value." In: Perspectives on Psychological Science.

# References vi


Lyons, Louis (Oct. 2013). "Discovering the Significance of 5 sigma." In: arXiv e-prints, arXiv:1310.1284, arXiv:1310.1284. arXiv: 1310.1284 [physics.data-an].

Martin, Gael M., David T. Frazier, and Christian P. Robert (2020). Computing Bayes: Bayesian Computation from 1763 to the 21st Cent arXiv: 2004.06425 [stat.CO].

Neal, Radford (2008). The Harmonic Mean of the Likelihood: Worst Monte Carlo Method E https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/.

Neyman, J. and E. S. Pearson (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses." In: Philos. Trans. Roy. Soc. London Ser. A 231, pp. 289–337. ISSN: 02643952. URL: http://www.jstor.org/stable/91247.

# References vii


Skilling, John (Nov. 2004). "Nested Sampling." In: American Institute of Physics Conference Series. Ed. by Rainer Fischer, Roland Preuss, and Udo Von Toussaint. Vol. 735. American Institute of Physics Conference Series, pp. 395–405.

– (2006). "Nested sampling for general Bayesian computation." In: Bayesian Analysis 1.4, pp. 833–859.

Vitells, Ofer and Eilam Gross (2011). "Estimating the significance of a signal in a multi-dimensional search." In: Astropart. Phys. 35, pp. 230–234. arXiv: 1105.4355 [astro-ph.IM].

Wagenmakers, Eric-Jan (2007). "A practical solution to the pervasive problems of *p* values." In: Psychon. Bull. Rev. 14, pp. 779–804.

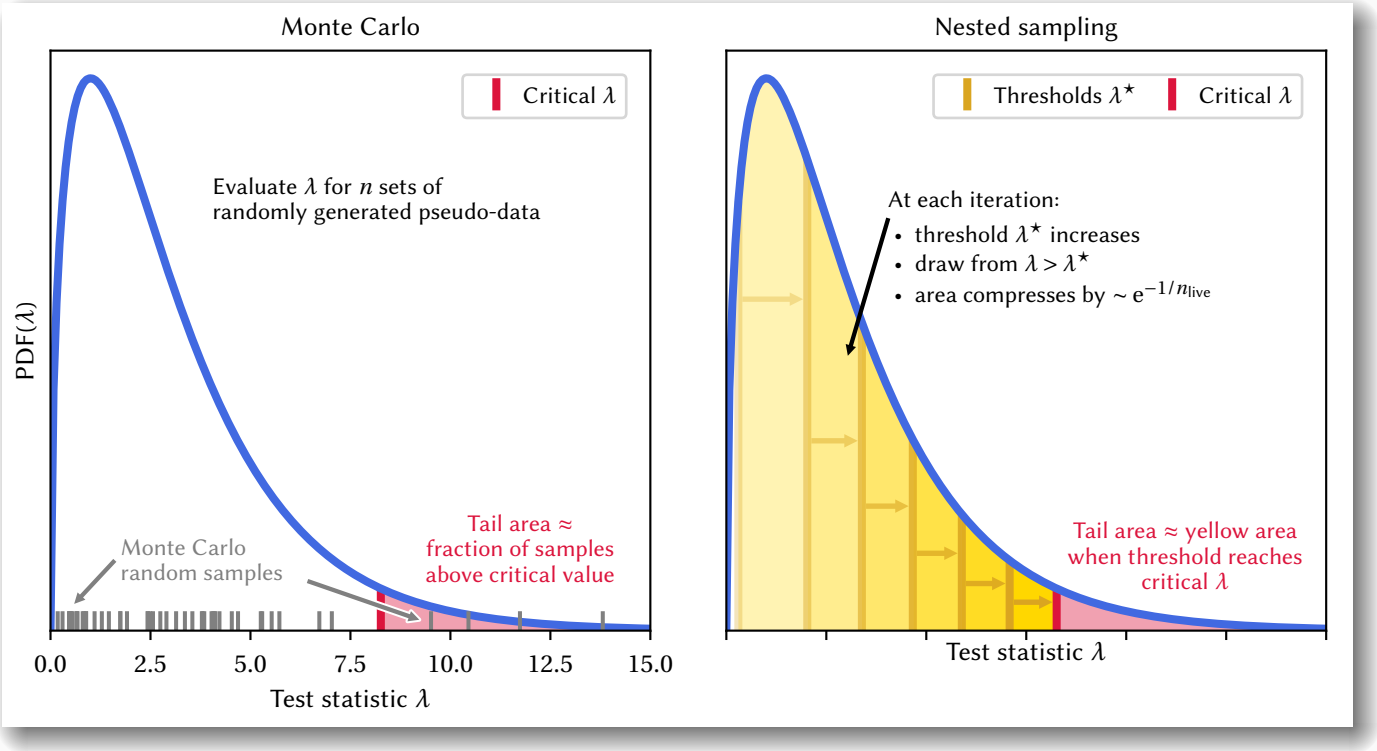
# Backup

# Translation — NS frequentist ⇔ NS Bayesian

- Sampling space ⇔ Parameter space
- Sampling distribution ⇔ Prior distribution
- Test-statistic ⇔ Likelihood function
- Region corresponding to $p$-value ⇔ Posterior distribution ⇔ Target
- Survival function ⇔ Volume variable
- Expected $p$-value under sampling distribution ⇔ Evidence (expected likelihood under prior distribution)

# Summary

Ripped an exponential factor out of the problem.



# *P*-values

*P*-values appear in two statistical frameworks (Hubbard and Bayarri 2003):

- Fisher 1925: $p$ is continuous measures of evidence against $H_0$
- Neyman and Pearson 1933: $p < \alpha$ allows us to control the type-1 error rate at $\alpha$

**Type-1 error rate**

Rate at which would be reject the null, $H_0$, when it was true.

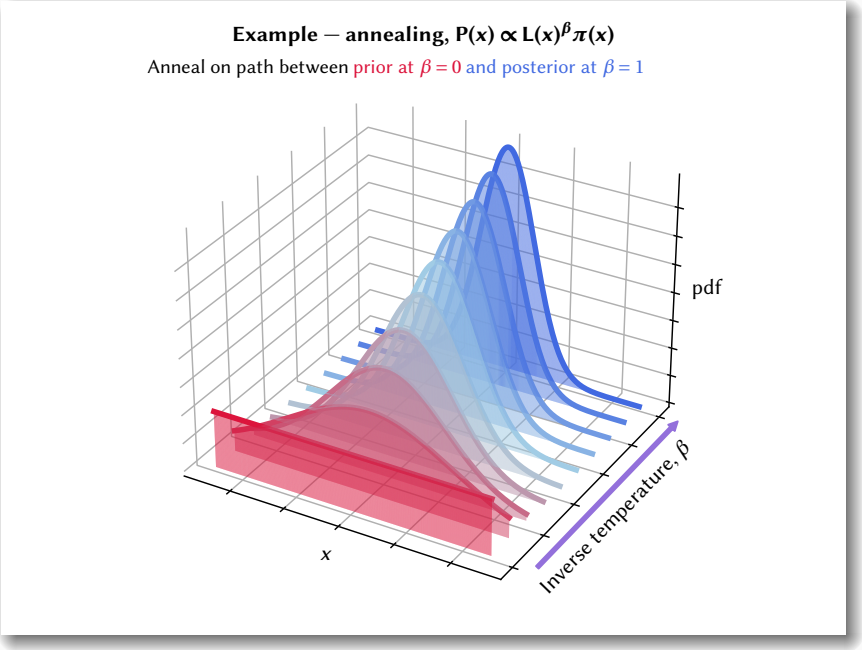
# $5\sigma$ and beyond

In some cases, we may desire even more than $5\sigma$. Table from Lyons 2013

| Search | Degree of surprise | Impact | LEE | Systematics | Number of $\sigma$ |
|---|---|---|---|---|---|
| Higgs search | Medium | Very high | Mass | Medium | 5 |
| Single top | No | Low | No | No | 3 |
| SUSY | Yes | Very high | Very large | Yes | 7 |
| $B_s$ oscillations | Medium/low | Medium | $\Delta m$ | No | 4 |
| Neutrino oscillations | Medium | High | $sin^2(2\theta), \Delta m^2$ | No | 4 |
| $B_s \to \mu\mu$ | No | Low/Medium | No | Medium | 3 |
| Pentaquark | Yes | High/very high | M, decay mode | Medium | 7 |
| $(g-2)_\mu$ anomaly | Yes | High | No | Yes | 4 |
| H spin $\neq 0$ | Yes | High | No | Medium | 5 |
| $4^{th}$ generation $q, l, \nu$ | Yes | High | M, mode | No | 6 |
| $v_\nu > c$ | Enormous | Enormous | No | Yes | >8 |
| Dark matter (direct) | Medium | High | Medium | Yes | 5 |
| Dark energy | Yes | Very high | Strength | Yes | 5 |
| Grav waves | No | High | Enormous | Yes | 7 |

*In high-energy physics, we need to compute tiny p-values.*

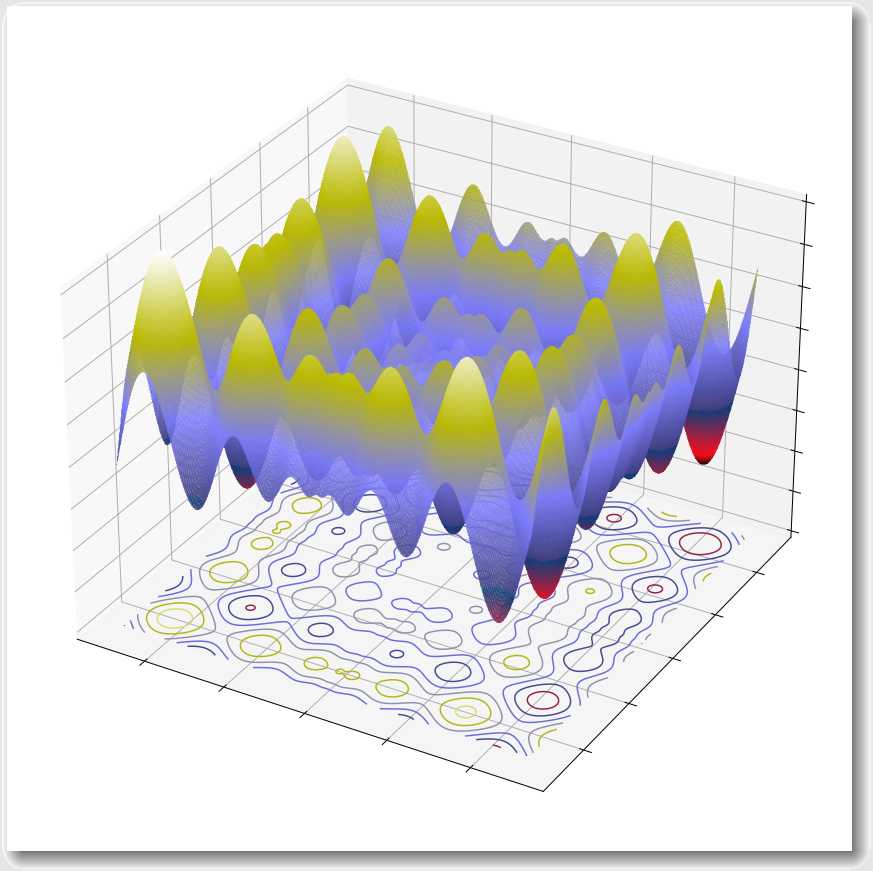
# Familiar example




Thermodynamic integration, Annealed Importance Sampling, Sequential Monte Carlo and Nested Sampling can be thought of as path samplers.

## Multi-modal

The test-statistic could contain several distinct modes — problematic?



# Sample from whole sample space — Monte Carlo

Looks wonderful. Very (computationally) expensive.

# Sample from the target

There are techniques — e.g., inverse-harmonic mean, where evidence written as posterior mean,

$$\frac{1}{\mathcal{Z}} = \left\langle \frac{1}{\mathcal{L}} \right\rangle \simeq \frac{1}{n} \sum \frac{1}{\mathcal{L}(\mathbf{x}_i)}$$

for $\mathbf{x}_i \sim$ Posterior. Compute this using draws from the posterior using MCMC!?

*No thank you. Terrible properties because of above reasoning. Radford Neal calls it the Worst MC Method Ever.*