Interpretability of Bayes Factors in Diffuse Analyses

Hans Niederhausen

Technical University of Munich

diffuse workshop, 09/14/19, Tokyo, Japan

Reminder Classical Testing

formulate **hypotheses H0 (“null”)** and **H1 (“alternative”)** about how data x is generated

choose **test-statistic T(x)** - some function of data x with different pdfs depending hypothesis
(choice should be guided by considerations of **statistical power** = probability of test to reject H0)

de-facto standard choice: **likelihood ratio test statistic**

asymptotics well understood (Wilks' theorem)

$$T(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\boldsymbol{\theta} \mid \mathbf{x})}{\sup_{\Theta} L(\boldsymbol{\theta} \mid \mathbf{x})}$$

Reminder Classical Testing

formulate **hypotheses H0 (“null”)** and **H1 (“alternative”)** about how data x is generated

choose **test-statistic T(x)** - some function of data x with different pdfs depending hypothesis
(choice should be guided by considerations of **statistical power** = probability of test to reject H0)

de-facto standard choice: **likelihood ratio test statistic**

asymptotics well understood (Wilks' theorem)

$$T(\mathbf{x}) = \frac{\sup_{\boldsymbol{\Theta_0}} L\left(\boldsymbol{\theta} \,|\, \mathbf{x}\right)}{\sup_{\boldsymbol{\Theta}} L\left(\boldsymbol{\theta} \,|\, \mathbf{x}\right)}$$

one (out of many) alternative choices: **Bayes Factor** = ratio of marginal likelihoods
(in all plots I use -2 log BF to maintain scale)

$$T(\mathbf{x}) = \frac{\int_{\boldsymbol{\Theta_0}} d\boldsymbol{\theta_0}\, L\left(\boldsymbol{\theta_0} \,|\, \mathbf{x}\right)\, q_0\left(\boldsymbol{\theta_0}\right)}{\int_{\boldsymbol{\Theta_1}} d\boldsymbol{\theta_1}\, L\left(\boldsymbol{\theta_1} \,|\, \mathbf{x}\right)\, q_1\left(\boldsymbol{\theta_1}\right)}$$

select **threshold for discovery** (typically 5 sigma) by fixing Type I error (false discovery)
observe data and **perform test**

Bayes Factor = ratio of marginal likelihoods

$$T(\mathbf{x}) = \frac{\int_{\Theta_0} d\boldsymbol{\theta_0}\, L\left(\boldsymbol{\theta_0} \,|\, \mathbf{x}\right)\, q_0\left(\boldsymbol{\theta_0}\right)}{\int_{\Theta_1} d\boldsymbol{\theta_1}\, L\left(\boldsymbol{\theta_1} \,|\, \mathbf{x}\right)\, q_1\left(\boldsymbol{\theta_1}\right)}$$

marginal likelihood

average value of likelihood function throughout entire parameter space w.r.t prior pdf

$$x \sim f(x)$$

$$E(g(x)) = \int dx\, g(x)\, f\,(x)$$

Bayes Factor = ratio of marginal likelihoods

$$T(\mathbf{x}) = \frac{\int_{\Theta_0} d\boldsymbol{\theta_0}\, L\left(\boldsymbol{\theta_0} \,|\, \mathbf{x}\right)\, q_0\left(\boldsymbol{\theta_0}\right)}{\int_{\Theta_1} d\boldsymbol{\theta_1}\, L\left(\boldsymbol{\theta_1} \,|\, \mathbf{x}\right)\, q_1\left(\boldsymbol{\theta_1}\right)}$$

marginal likelihood

average value of likelihood function throughout entire parameter space w.r.t prior pdf

$$x \sim f(x)$$

$$E(g(x)) = \int dx\, g(x)\, f\,(x)$$

$$m_0 = \int_{\Theta_0} d\boldsymbol{\theta_0}\, L\left(\boldsymbol{\theta_0} \,|\, \mathbf{x}\right)\, q_0\left(\boldsymbol{\theta_0}\right)$$

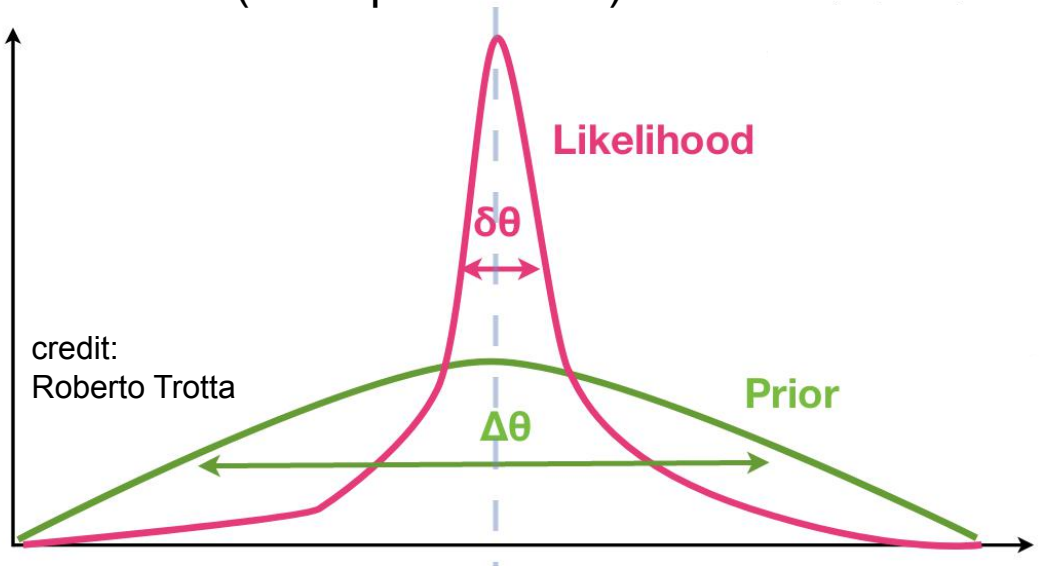
average value

likelihood = some function in theta (not a pdf in theta)

prior = pdf in theta

$$\approx \frac{\delta\theta}{\Delta\theta} L(\hat{\theta})$$

Occam's factor

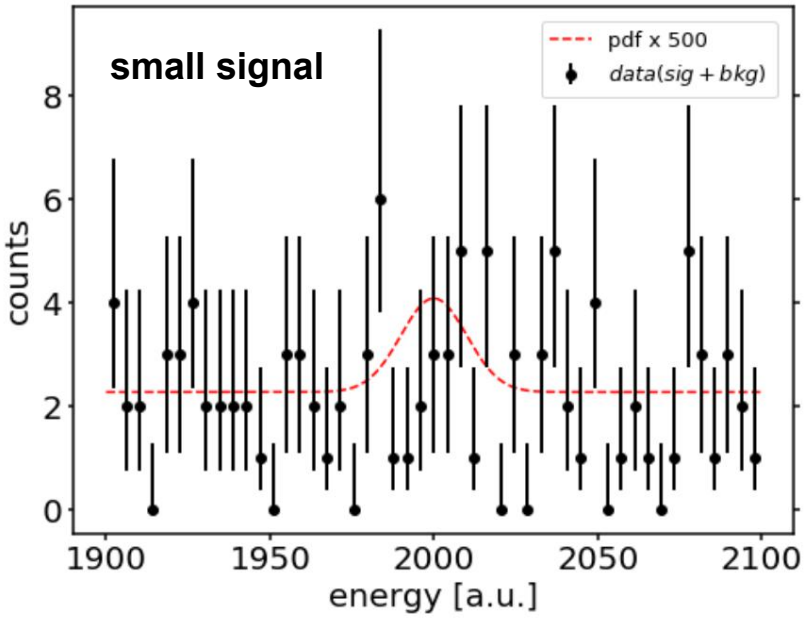


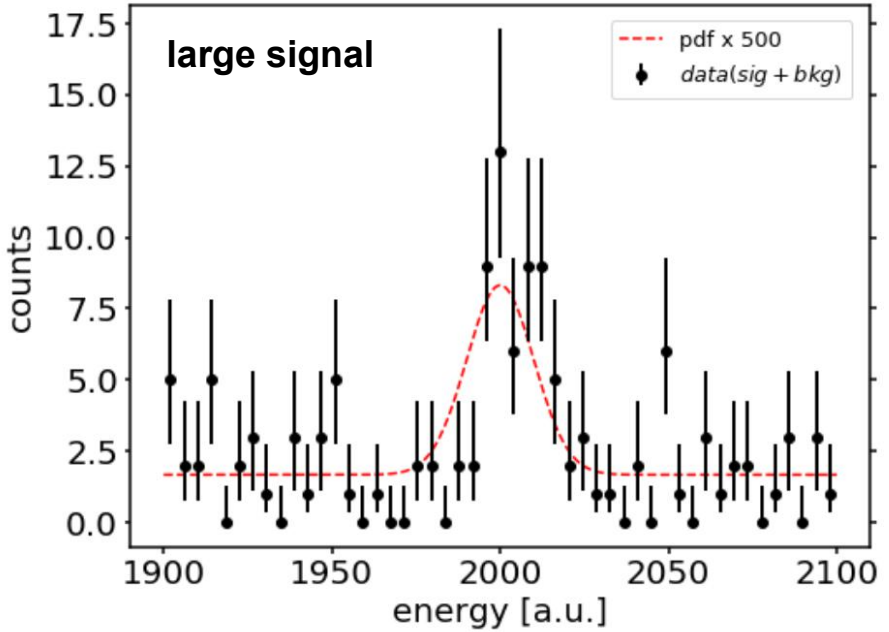
credit:
Roberto Trotta

marginal likelihood very sensitive to prior width (unlike posterior intervals)

a toy example

uniform background with possible gaussian signal (known location) - marked poisson process





possible parameterizations

1) **fraction of signal events** (parameter of interest) and **total rate** (nuisance parameter)

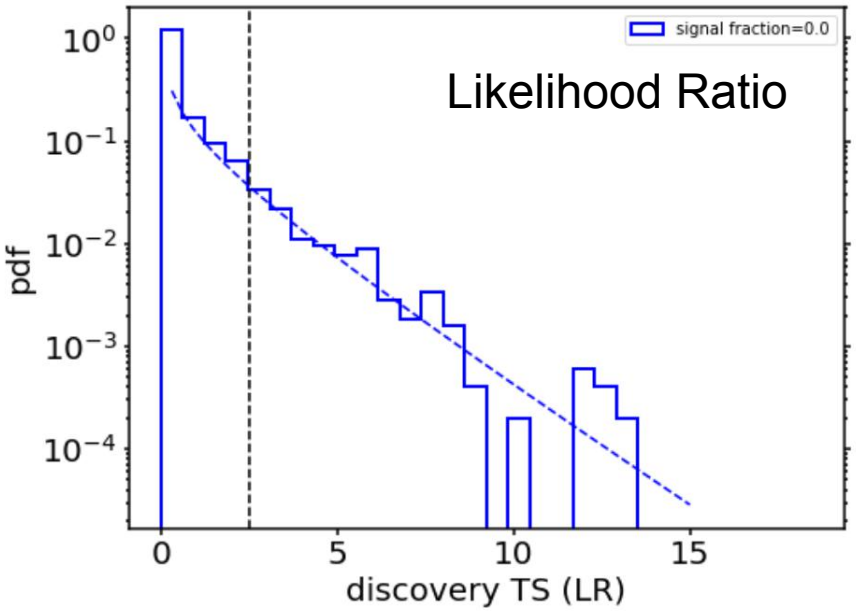
2) **total signal rate** (parameter of interest) and **total background rate** (nuisance parameter)

Example: **fraction of signal events** (parameter of interest) and **total rate** (nuisance parameter)

calculate **distributions for BF** - prior on signal fraction: uniform(0,1) and **LR as function of true signal fraction**

H0: p_s = 0.0 against H1: p_s > 0.0

calculate power of corresponding statistical tests for Type I error rate of alpha=0.05

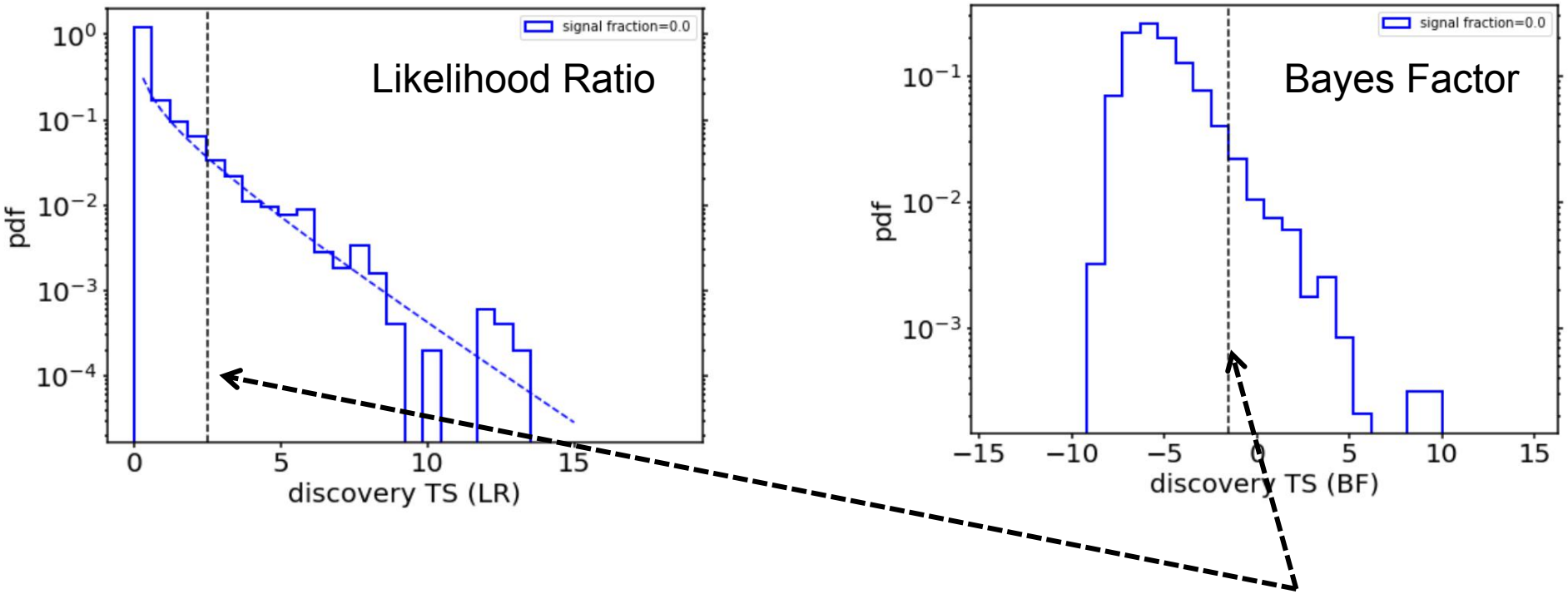


Example: **fraction of signal events** (parameter of interest) and **total rate** (nuisance parameter)

calculate **distributions for BF** - prior on signal fraction: uniform(0,1) and **LR as function of true signal fraction**

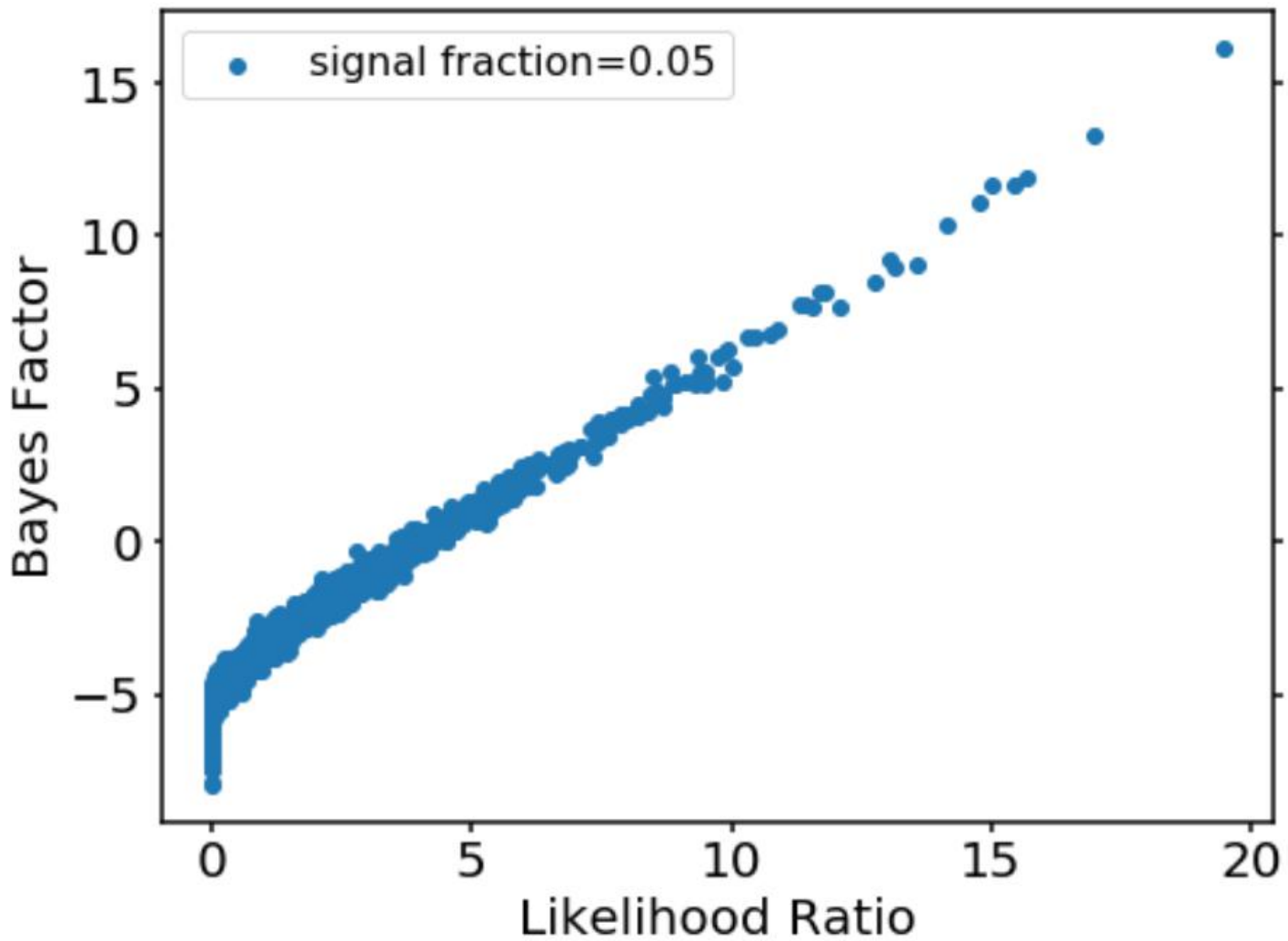
H0: p_s = 0.0 against H1: p_s > 0.0

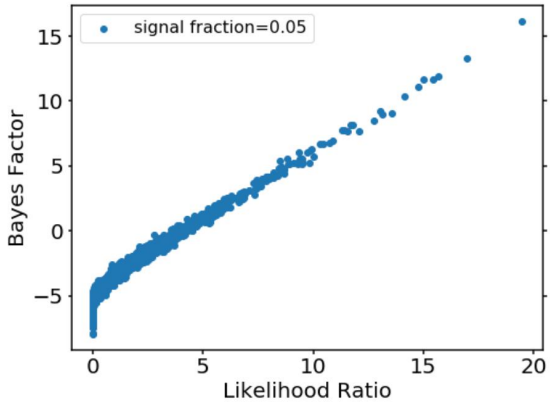
calculate power of corresponding statistical tests for Type I error rate of alpha=0.05

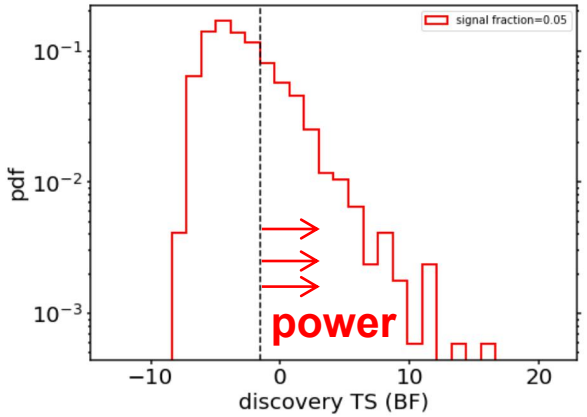


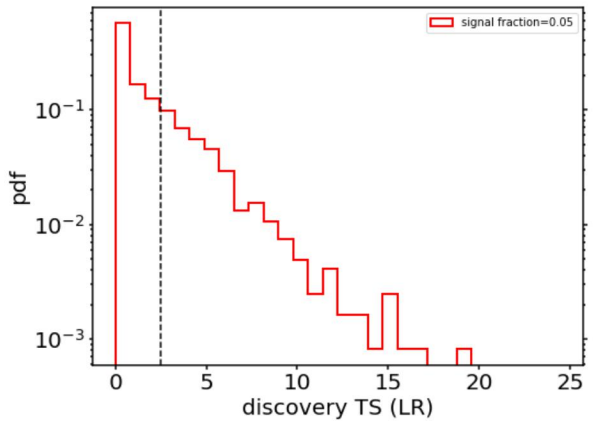
TS distributions different under H0: p_s = 0.0 -> **critical values are different** (dashed lines)

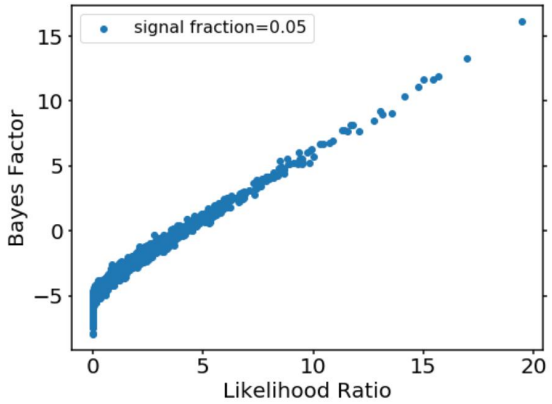
But are the tests different?

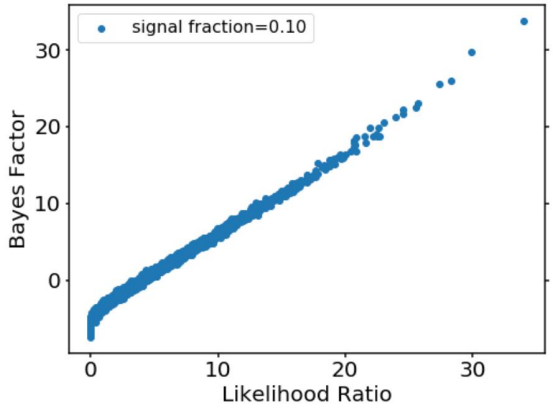

signal fraction=0.05
Bayes Factor
Likelihood Ratio
15
10
5
0
−5
0
5
10
15
20

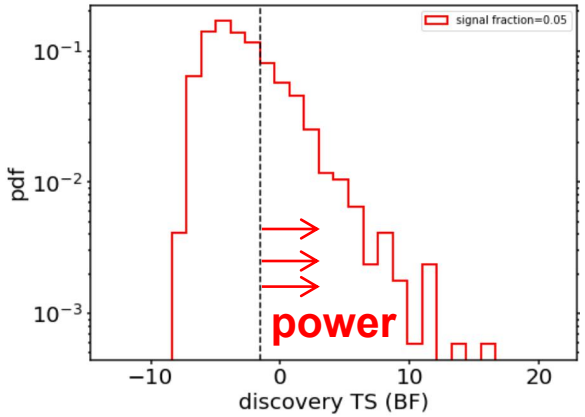

signal fraction=0.05
Bayes Factor
Likelihood Ratio
15
10
5
0
−5
0
5
10
15
20

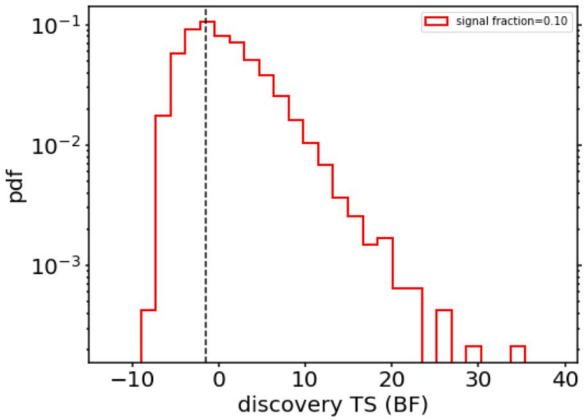

signal fraction=0.05
pdf
discovery TS (BF)
power
10^{-1}
10^{-2}
10^{-3}
−10
0
10
20

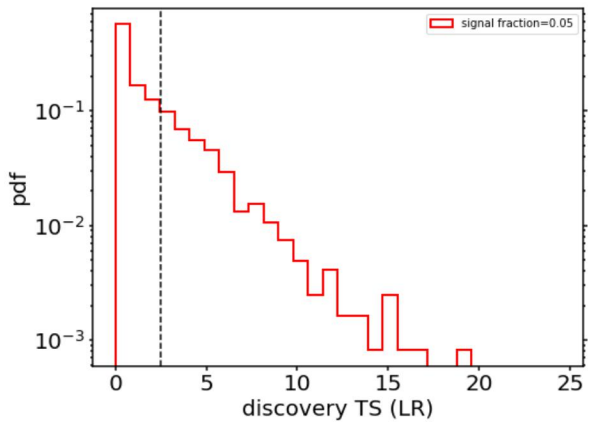

signal fraction=0.05
pdf
discovery TS (LR)
10^{-1}
10^{-2}
10^{-3}
0
5
10
15
20
25

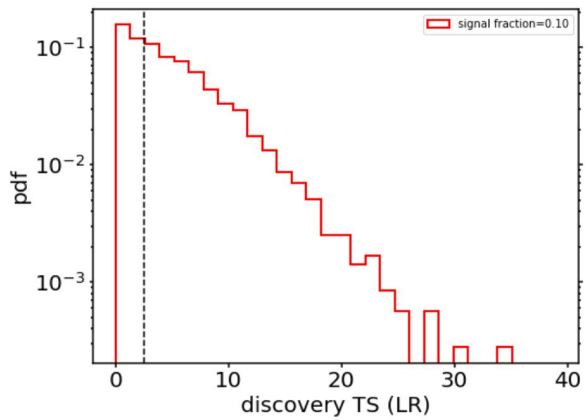

signal fraction=0.05
Bayes Factor
Likelihood Ratio

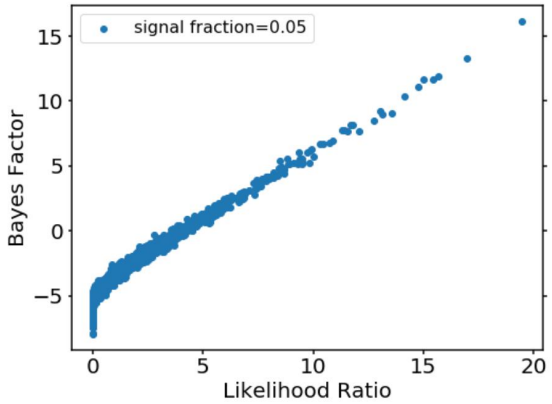

signal fraction=0.10
Bayes Factor
Likelihood Ratio

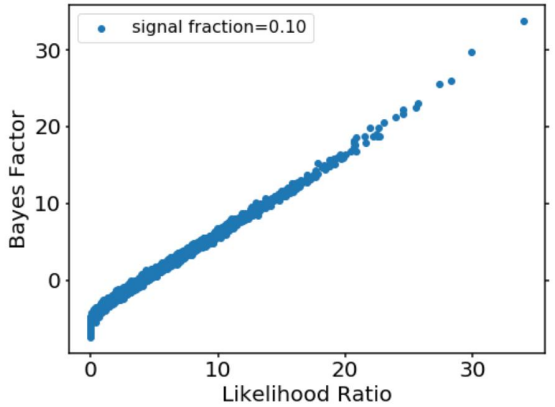

signal fraction=0.05
pdf
power
discovery TS (BF)

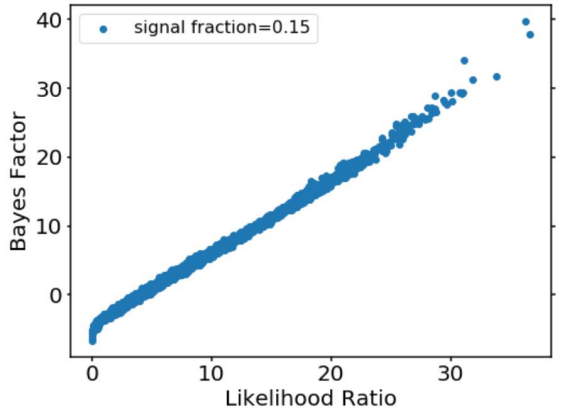

signal fraction=0.10
pdf
discovery TS (BF)

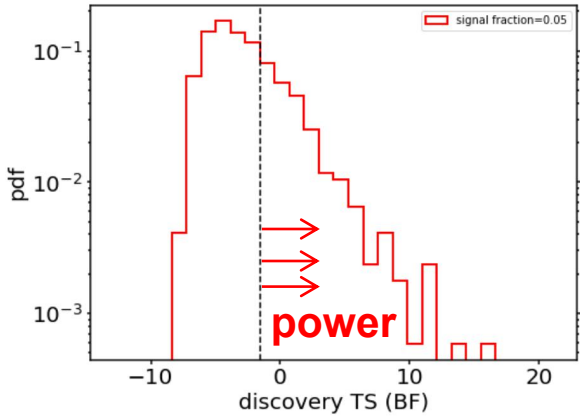

signal fraction=0.05
pdf
discovery TS (LR)

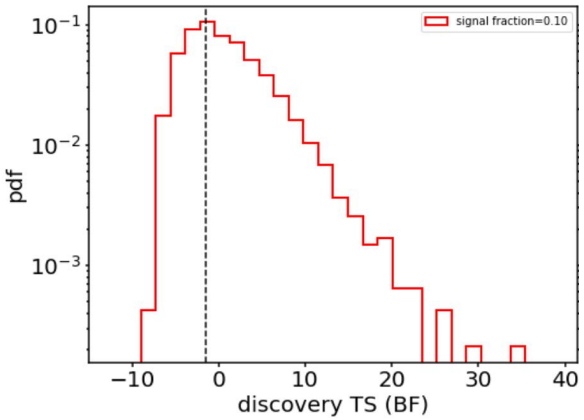

signal fraction=0.10
pdf
discovery TS (LR)

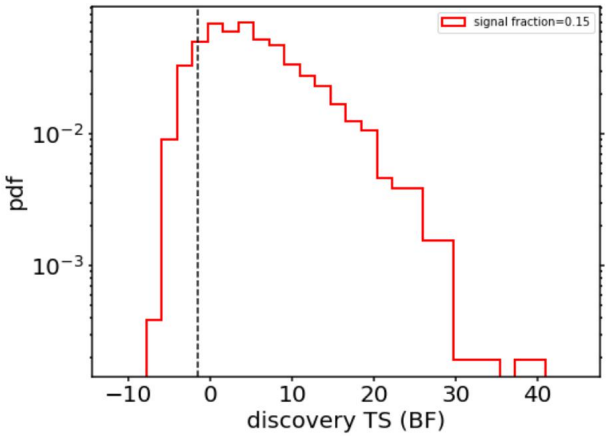

signal fraction=0.05
Bayes Factor
Likelihood Ratio

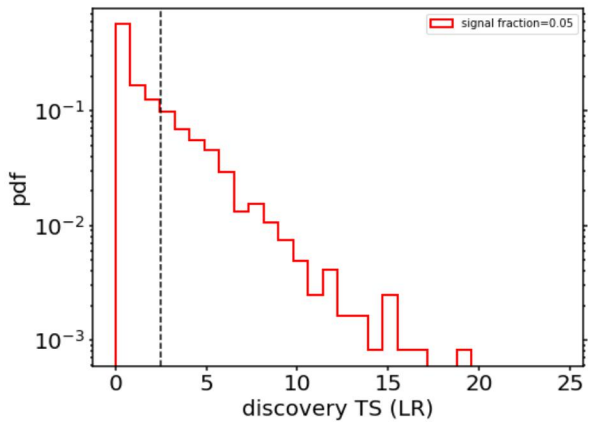

signal fraction=0.10
Bayes Factor
Likelihood Ratio

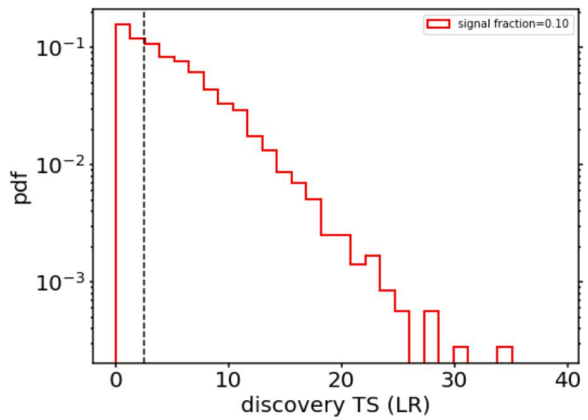

signal fraction=0.15
Bayes Factor
Likelihood Ratio

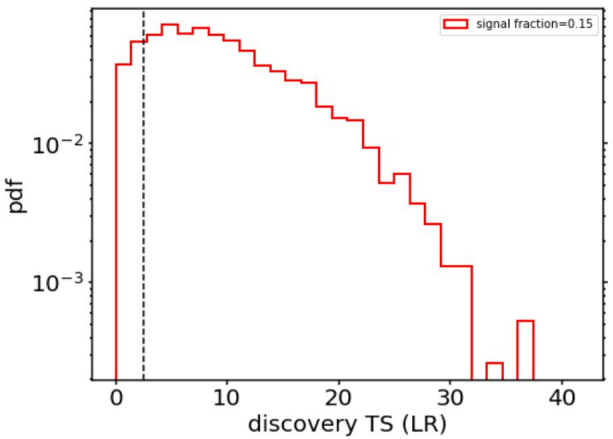

signal fraction=0.05
pdf
power
discovery TS (BF)

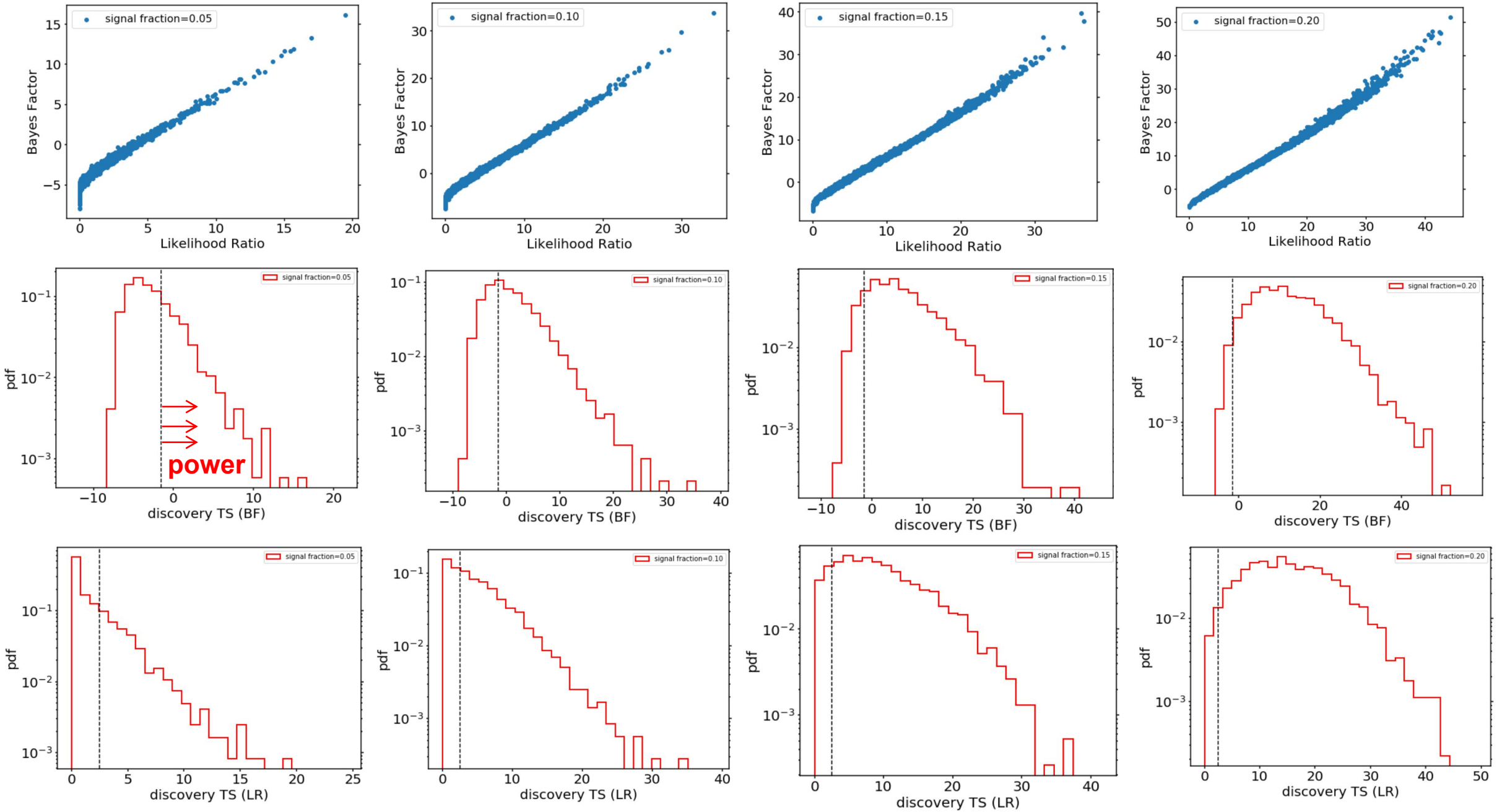

signal fraction=0.10
pdf
discovery TS (BF)

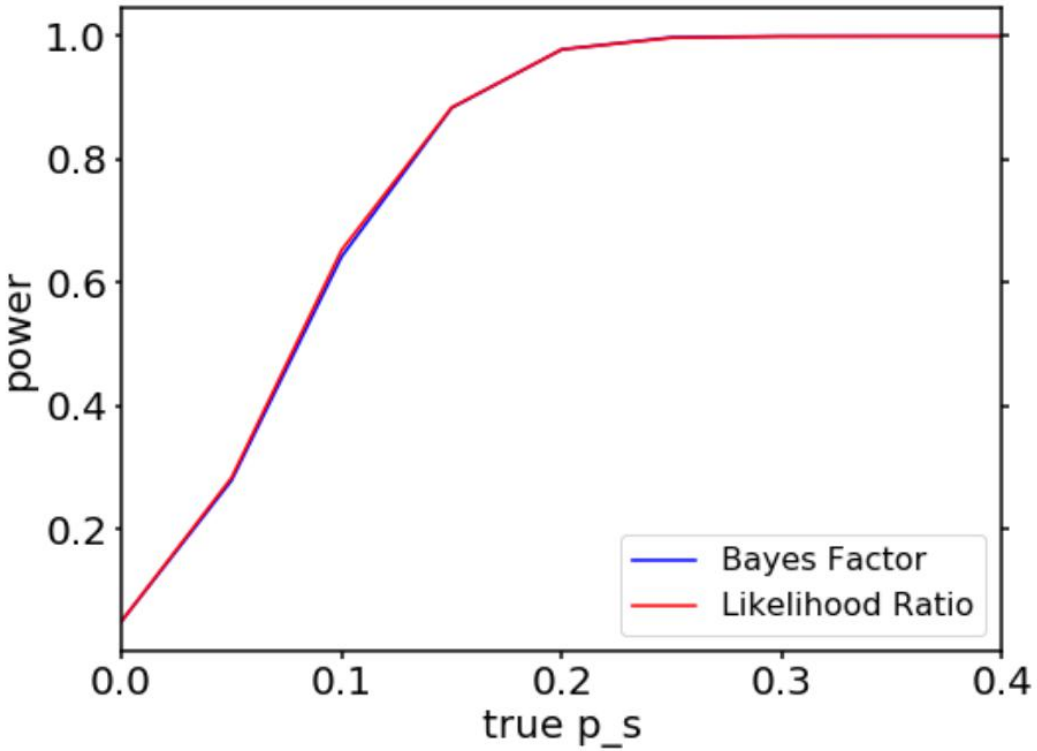

signal fraction=0.15
pdf
discovery TS (BF)


signal fraction=0.05
pdf
discovery TS (LR)


signal fraction=0.10
pdf
discovery TS (LR)

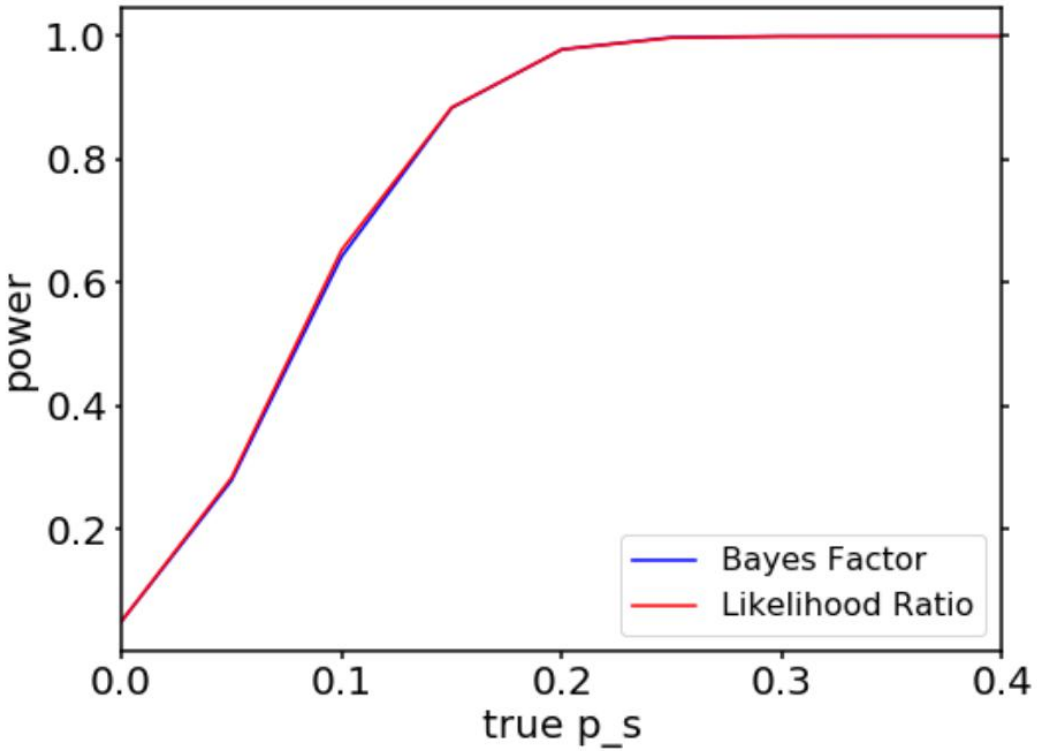

signal fraction=0.15
pdf
discovery TS (LR)


signal fraction=0.05
signal fraction=0.10
signal fraction=0.15
signal fraction=0.20
Bayes Factor
Likelihood Ratio
pdf
discovery TS (BF)
discovery TS (LR)
power



to within statistical precision (nested sampling to compute BF) both tests deliver **same power**.

since TS values correlate well - one would draw **identical conclusions** from both frequentist analyses!



to within statistical precision (nested sampling to compute BF) both tests deliver **same power**.

since TS values correlate well - one would draw **identical conclusions** from both frequentist analyses!

This agreement is NOT a general result.
Relative performance will depend on the problem.

You will not know which one is the better test, unless you compute the power curve!
(in our problems uniformly most powerful tests very likely don't exist)

Example from IceCube

(not a full bayes factor - average only computed across time dimension of parameter space)

Braun et al (2010)

We correct this behavior by marginalizing with respect to the burst time in a Bayesian manner, using a uniform prior. For a search window bounded in time by T_{min} and T_{max}, the likelihood is

average likelihood across central time

$$\mathcal{L}(n_s, \gamma, \sigma_{\mathrm{T}}) = \int_{\mathrm{T}_{min}}^{\mathrm{T}_{max}} \prod_{i=1}^{N} \left[\frac{n_s}{N} \mathcal{S}_i + (1 - \frac{n_s}{N}) \mathcal{B}_i \right] P(\mathrm{T}_\circ) d\mathrm{T}_\circ, \qquad (10)$$

where $P(\mathrm{T}_\circ) = \frac{1}{\mathrm{T}_{max} - \mathrm{T}_{min}}$ is a constant prior for the time window, and the PDFs $\mathcal{S}_i$ and $\mathcal{B}_i$ are given by Eqs. 7 and 8, respectively. If the data contains a significant burst, only times within $\sim\hat{\sigma}_{\mathrm{T}}$ of the burst contribute to the integral, shown in Fig. 1. The integrand has Gaussian dependence with respect to $\mathrm{T}_\circ$, and the maximum is the maximized likelihood $\mathcal{L}(\hat{n}_s, \hat{\gamma}, \hat{\sigma}_{\mathrm{T}}, \hat{\mathrm{T}}_\circ)$. The marginal likelihood in Eq. 10 can therefore be approximated by

$$\mathcal{L}(\hat{n}_s, \hat{\gamma}, \hat{\sigma}_{\mathrm{T}}) \sim \frac{\sqrt{2\pi}\hat{\sigma}_{\mathrm{T}}}{\mathrm{T}_{max} - \mathrm{T}_{min}} \mathcal{L}(\hat{n}_s, \hat{\gamma}, \hat{\sigma}_{\mathrm{T}}, \hat{\mathrm{T}}_\circ), \qquad (11)$$

occam's factor

proposed analysis fully frequentist but not a standard likelihood ratio test

Bayesian Use of the Bayes Factor

$$\mathbf{BF} = T(\mathbf{x}) = \frac{\int_{\Theta_0} d\boldsymbol{\theta_0}\, L\left(\boldsymbol{\theta_0} \,|\, \mathbf{x}\right)\, q_0\left(\boldsymbol{\theta_0}\right)}{\int_{\Theta_1} d\boldsymbol{\theta_1}\, L\left(\boldsymbol{\theta_1} \,|\, \mathbf{x}\right)\, q_1\left(\boldsymbol{\theta_1}\right)}$$

in a full Bayesian analysis there is **no need to look at sampling distributions**.
BF contains all relevant information (assuming your priors “make sense”).
see Kass & Raftery (1995) ,13k citations ...

Bayesian Use of the Bayes Factor

$$\mathbf{BF} = T(\mathbf{x}) = \frac{\int_{\Theta_0} d\boldsymbol{\theta_0}\, L\left(\boldsymbol{\theta_0} \mid \mathbf{x}\right)\, q_0\left(\boldsymbol{\theta_0}\right)}{\int_{\Theta_1} d\boldsymbol{\theta_1}\, L\left(\boldsymbol{\theta_1} \mid \mathbf{x}\right)\, q_1\left(\boldsymbol{\theta_1}\right)}$$

in a full Bayesian analysis there is **no need to look at sampling distributions**.
BF contains all relevant information (assuming your priors “make sense”).
see Kass & Raftery (1995) ,13k citations ...

calculate BF and derive conclusion from Jeffreys or Kass/Rafterys scales

jeffreys scale

$\log_{10}(B_{10})$	B_{10}	Evidence against H_0
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

Kass/Raftery scale

$2\log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

similar scale as LRT TS

assuming your priors “make sense”

1) theory / previous experiments provide priors -> good.

2) no prior information

BF is not defined for improper priors
BF strongly depends on any cutoff used to make prior proper
(example: upper end of uniform prior)

innocent prior?

signal_rate ~ uniform(0, maximum rate)

assuming your priors “make sense”

1) theory / previous experiments provide priors -> good.

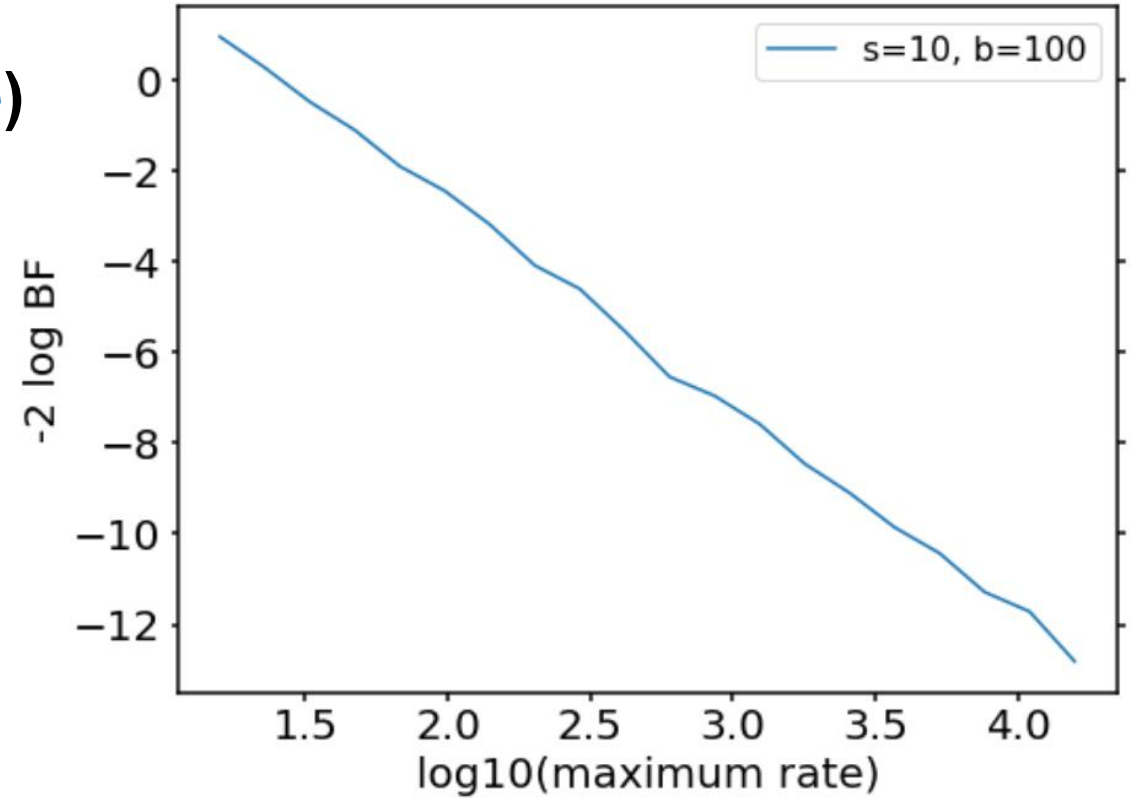
2) no prior information

BF is not defined for improper priors
BF strongly depends on any cutoff used to make prior proper
(example: upper end of uniform prior)

innocent prior?

signal_rate ~ uniform(0, maximum rate)

example **large maximum rate**
-> “artifically” **large evidence for H0**



2 options:

A) use some cutoff value and perform frequentist analysis with BF as test-statistic

B) read Jim Berger's papers on objective Bayesian analysis. e.g. Berger & Pericchi (2004)

Summary

- Bayes Factors are important tool for Bayesian model selection
- can also be a useful test-statistic for frequentist analysis

- BUT pure Bayesian use of BF hard/tricky in situations without useful prior information

  in these cases suggest the following

  1) frequentist calibration

  or

  2) additional steps (training samples, see Berger et al. refs) to generate objective but proper priors for BF computation and subsequent Bayesian testing