LIGO VIRGO KAGRA NANOGrav

Introduction to frequentist statistics and Bayesian inference

Joe Romano, Texas Tech University
Wednesday, 20 July 2022
(HUST GW Summer School 2022, Lecture 1)

- thanks to Yan for inviting me to give these lectures
- please ask questions!! this is a summer school not a presentation at a conference

References

- Romano and Cornish, Living Reviews in Relativity article, 2017 (section 3)
- Rover, Messenger, Prix, "Bayesian versus frequentist upper limits," PHYSTAT 2011 workshop
- Gregory, "Bayesian Logical data analysis", 2005
- Howson and Urbach, "Scientific reasoning: the Bayesian approach", 2006
- Helstrom, "Statistical theory of signal detection", 1968
- Wainstein and Zubakov, "Extraction of signals from noise," 1971

- list of references
- will draw from LRR for all my lectures
- references to both bayesian and frequentist literature

Outline

1. Probabilistic inference (broadly defined)
2. Frequentist statistics
3. Bayesian inference
4. Exercises - worked examples

- brief outline

Frequentist vs Bayesian "pre-test"

- An astronomer measures the mass of a NS in a binary pulsar system to be $M = (1.39 \pm 0.02) M_\odot$ with 90% confidence. How do you interpret the quoted result?
- Answer 1: You are 90% confident that the true mass of the NS lies in the interval $[1.37 M_\odot, 1.41 M_\odot]$
- Answer 2: You interpret 90% as the long-term relative frequency with which the true mass of the NS lies in the set of intervals $\{[\hat{M} - 0.02 M_\odot, \hat{M} + 0.02 M_\odot]\}$ where $\{\hat{M}\}$ is the set of measured masses.

- start with a pretest to judge the way that you think about probabilistic statements

Frequentist vs Bayesian "affiliation"

- If you chose answer 1, then you are a Bayesian
- If you chose answer 2, then you are a frequentist

- A1-> bayesian, A2 -> frequentist

Goal of science is to infer nature's state from observations

- Observations are:
  - **incomplete** (problem of induction)
  - **imprecise** (measurement noise, quantum mechanics, …)

$\Longrightarrow$ conclusions are uncertain!!

- **Probabilistic inference** (aka "plausible inference", "statistical inference") is a way of **dealing with uncertainty**
- **Different from** mathematical deduction

- big picture overview
- observation are both incomplete and imprecise -> conclusions are necessarily uncertain
- statistical / plausible / probabilistic inference is the framework that we have for dealing with uncertainty
- emphasize "infer" not "deduce" -> not mathematical induction

I. Probabilistic inference

Definitions of probability

- Frequentist definition: **Long-run relative frequency** of occurrence of an event in a set of repeatable identical experiments
- Bayesian definition: **Degree of belief** (or confidence, plausibility) in any proposition

NOTE: For the frequentist definition, probabilities can only be assigned to propositions about outcomes of repeatable identical experiments (i.e., **random variables**), not to hypotheses or parameters describing the state of nature, which have fixed but unknown values

- need to define what we mean by probability
- frequentist: probability equals long-run relative frequency of occurrence of an event in a set of repeatable
- bayesian: degree of belief, confidence, or plausibility in a proposition (more subjective, but more general)
- for frequentists, assign probabilities only to random variables not to hypotheses or parameters describing the state of nature
- frequentists have a way of making statements about parameters and hypothesis, but in a somewhat *indirect* manner

Algebra of probability

- Possible values:

$$P(X = \text{true}) = 1$$
$$P(X = \text{false}) = 0$$
$$0 < P(X = \text{not sure}) < 1$$

- Sum rule:

$$P(X) + P(\bar{X}) = 1$$

- Product rule:

$$P(X\,|\,Y)P(Y) = P(X, Y)$$

- NOTE: $P(X\,|\,Y)$ is the probability of X conditioned on Y (assuming Y is true)
- $P(X\,|\,Y) \neq P(Y\,|\,X)$ in general. Example X="person is pregnant", Y="person is female"

- the algebra of probabilities is extremely simple
- values between 0 and 1 inclusing
- sum rule (Xbar is complement of X)
- product rule relates joint probabilities and conditional probabilities
- P(X|Y) \ne P(Y|X)

Bayes' theorem (a simple consequence of the product rule!!)

posterior, likelihood, prior, evidence

$$\underbrace{P(H|D)}_{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\,\overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}$$

where $P(D) = P(D|H)P(H) + P(D|\bar{H})P(\bar{H})$

"Learning from experience": the probability of H being true (in light of new data) increases by the ratio of the probability of obtaining the new data D when H is true to the probability of obtaining D in any case

- Bayes' theorem is a simple consequence of the product rule P(X.Y) = P(Y|X)
- also valid for frequentist statistic if H,D = X,Y are random variables
- Bayes' theorem incorporates learning from experience: updating degree of belief in H in light of new data D

Bayes' theorem (for parameters associated with a given hypothesis or model)

$$p(a|d,H) = \frac{p(d|a,H)p(a|H)}{p(d|H)}$$

where $p(d|H) = \int \mathrm{d}a\, p(d|a,H)p(a|H)$

"marginalization" over a

- Bayes' theorem for parameters a associated with hypothesis/model H.
- H conditions all the probabilities, goes along for the ride
- the evidence is calculated by integrating over (marginalizing over) a

---

Comparing frequentist & Bayesian inference

Frequentist statistics	Bayesian inference
Probabilities are long-run relative occurrences of outcomes of repeatable expts —> can't be assigned to hypotheses	Probabilities are degree of belief —> can be assigned to hypotheses
Usually start with a likelihood function p(d|H)	Same as frequentist
Construct a statistic (some function of the data d) for parameter estimation or hypothesis testing	Need to specify priors for parameters and hypotheses
Calculate sampling distribution of the statistics (e.g., using time slide)	Use Bayes' theorem to update degree of belief in a parameter or hypothesis
Calculates confidence intervals (for parameter estimation) and p-values (for hypothesis testing)	Construct posteriors (for parameter estimation) and odds ratios (Bayes factors) (for hypothesis testing)

- single slide summarizing the remainder of the talk

II. Frequentist statistics

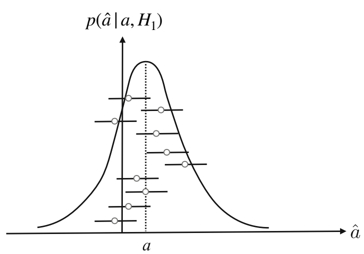
- spend the next several slides on frequentist statistics

Frequentist parameter estimation

- Construct a statistic (**estimator**) $\hat{a}$ for the parameter you are interested in
- Calculate the **sampling distribution** $p(\hat{a}\,|\,a, H_1)$ where $H_1 = \cup_{a>0} H_a$
- Statements like Prob$(a - \Delta < \hat{a} < a + \Delta)$ make sense since $\hat{a}$ is a random variable
- Statements like $a = \hat{a} \pm \Delta$ with 90% confidence must be interpreted as statements about the **randomness of the intervals**—i.e., 90% is the long-term relative frequency with which the true value of the parameter lies in the set of intervals $\{[\hat{a} - \Delta, \hat{a} + \Delta]\}$ where $\{\hat{a}\}$ is the set of measured parameter estimates

- for parameter estimation need to construct and estimator (function of the data) for a particular parameter
- then need to know the sampling distribution (probability distribution) for that estimator conditioned on the relevant hypothesis for that parameter
- ahat is random variable so probabilistic statements about ahat make sense
- can’t make probabilistic statements about a
- instead a = ahat +\- delta should be interpreted as statements about the randomness of the intervals

Frequentist parameter estimation



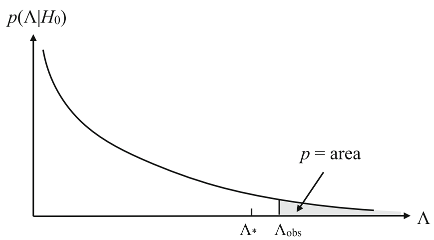
- intervals are random, 90% contain the true value a

Frequentist hypothesis testing

- Suppose you want to test a hypothesis H_1 that a GW signal with some fixed but unknown amplitude $a > 0$ is present in the data ($H_1 \equiv \cup_{a>0} H_a$)
- Since you can't assign probabilities to hypotheses as a frequentist, you introduce the null hypothesis $H_0 = \bar{H}_1$ (for this example, $a = 0$), and then **argue for H_1 by arguing against H_0** (like proof by contradiction)
- So you construct a **test statistic** Λ and calculate its sampling distributions $p(\Lambda \,|\, H_0)$ and $p(\Lambda \,|\, a, H_1)$ conditioned on H_0 and H_1
- If the observed value of Λ lies far out in the tail for the null distribution, $p(\Lambda \,|\, H_0)$, you reject H_0 (accept H_1) at the $p \times 100\,\%$ level where $p = \text{Prob}(\Lambda > \Lambda_{\text{obs}} \,|\, H_0)$ is the so-called **p-value**

- frequentist hypothesis testing is like proof for contradiction
- you argue for H1 by arguing against its complement H0
- construct a test statistic and calculate its sampling distribution for the different hypothesis
- if the observed value of Lambda lies far out in the tail for the null distribution, you reject H0 and accept H1 at the p% level, where p=prob(Lambda>Lambda_obs | H0)

Frequentist p-value

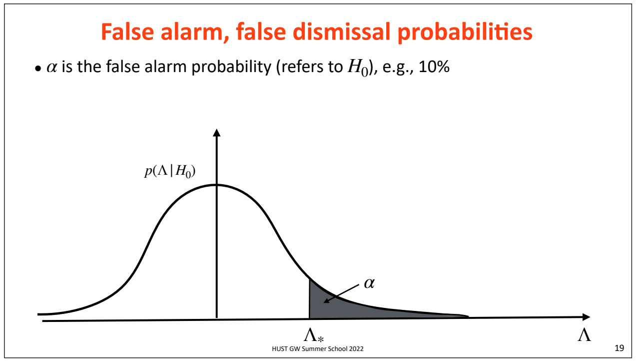


- graphical representation of the p-value; Lambda_* will be described next

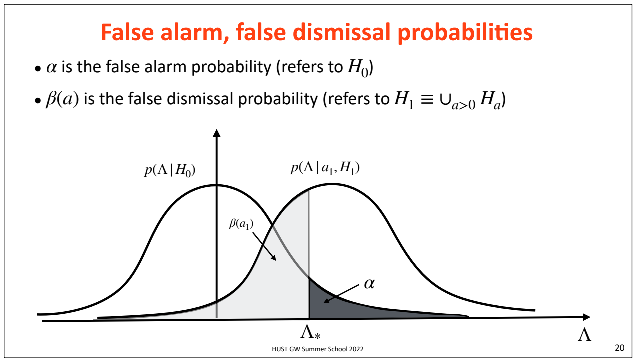
False alarm, false dismissal probabilities

- The p value needed to reject the null hypothesis defines a **threshold** Λ_*
- There are **two types of errors** when using the test statistic Λ:
  - False alarm: Reject the null hypothesis ($\Lambda_{\text{obs}} > \Lambda_*$) when it is true
  - False dismissal: Accept the null hypothesis ($\Lambda_{\text{obs}} \leq \Lambda_*$) when it is false
- Different test statistics are **judged according to their false alarm and false dismissal probabilities**
- In GW data analysis, one typically sets the false alarm probability to some acceptably low level (e.g., 1 in 1000), then finds the test statistic that minimizes the false dismissal probability for fixed false alarm probability (called the **Neyman-Pearson criterion**)

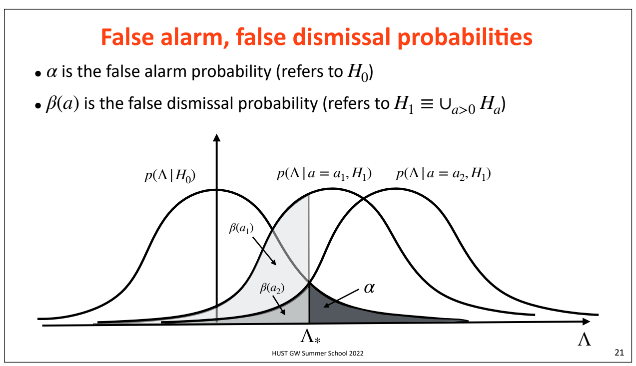
- the choice of p-value needed to reject the null hypothesis defines the threshold Lambda_*
- two types of errors associated with frequentist hypothesis testing:
- false alarm: reject null hypothesis when it is true
- false dismissal: accept the null hypothesis when it is false
- test statistics judged by false alarm / false dismissal probabilities
- Neyman-Pearson: fix false alarm probability, then choose test statistic that minimizes the false dismissal probability for fixed false alarm



- false alarm probability is the probability that Lambda lies to the right of Lambda_* conditioned on H0



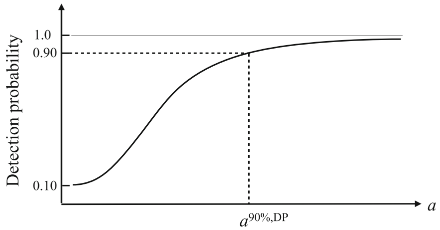
- false dismissal probability is the area to the left of Lambda_* conditioned on the hypothesis that a signal is present in the data ('a' dependent)



- illustration for a_2 > a_1

Detection probability

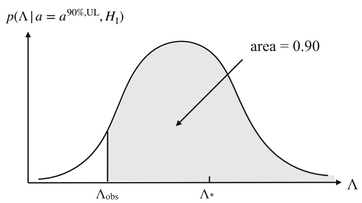
- $\gamma(a) \equiv 1 - \beta(a)$ is the fraction of the time that the test statistic Λ **correctly identifies the presence of a signal** with amplitude a



- detection probability = 1 - false dismissal probability
- fraction of the time that the test statistic correctly identifies the presence of a signal with amplitude ‘a’

Frequentist upper limits

- If $\Lambda_{\mathrm{obs}} < \Lambda_*$ one often sets an UL on the amplitude a of the signal
- $a^{90\%,\mathrm{UL}}$ is the value of a for which $\mathrm{Prob}\,(\Lambda \geq \Lambda_{\mathrm{obs}}\,|\,a = a^{90\%,\mathrm{UL}}, H_1) = 0.90$



- frequentist UL is defined as the value of ‘a’ for which the probability that Lambda>Lambda_obs conditioned on that a is 90%.
- somewhat counter-intuitive to me

III. Bayesian inference

Bayesian parameter estimation

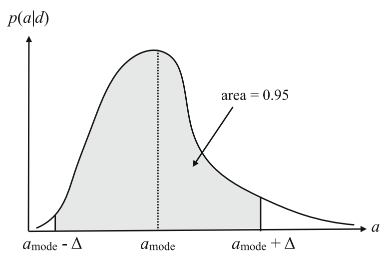
- Bayesian parameter estimation is via the **posterior** distribution $p(a \mid d, H)$
- The **posterior distributions contains all the information** about the parameter, but you can reduce it to a few numbers (e.g., mode, mean, stddev, …)
- If the posterior distribution depends on several parameters, you can obtain the posterior for one parameter by **marginalizing** over the others,

$$p(a \mid d, H) = \int \mathrm{d}b \, p(a, b \mid d, H) = \int \mathrm{d}b \, p(a \mid b, d, H) p(b \mid H)$$

- A Bayesian **credible interval** or **upper limit** defined in terms of the area under the posterior distribution

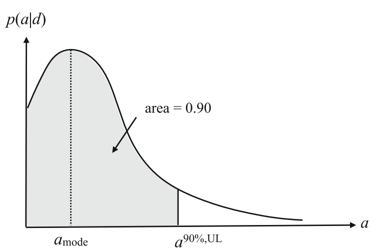
- Bayesian parameter estimation is via the posterior
- posterior distribution contains all info, but you can reduce it to a few numbers
- might need to marginalize
- Bayesian credible interval or UL defined in terms of area under the posterior

Bayesian credible interval



- 95% Bayesian credible interval

Bayesian credible upper limit



- Bayesian 90% credible UL

Bayesian hypothesis testing / model selection

- Compare two hypotheses H_1 and H_0 by taking their posterior **odds ratio**:

$$\frac{p(H_1|d)}{p(H_0|d)} = \frac{p(d|H_1)}{p(d|H_0)}\frac{p(H_1)}{p(H_0)}$$

posterior odds

Bayes factor $\mathscr{B}_{10}(d)$
(ratio of marginalized likelihoods or "evidences")

prior odds

- Bayesian model selection is via posterior odds ratio
- Posterior odds = prior odds * ratio of marginalized likelihoods

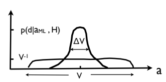
Relating Bayes factors and maximum-likelihood ratios

- Calculation of the evidence (=likelihood of an hypothesis) usually involves **marginalization over the parameters** associated with the hypothesis/model:

$$p(d|H) = \int \mathrm{d}a\, p(d|a,H)p(a|H)$$

- When the **data are informative**:

$$p(d|H) \simeq p(d|a_{\rm ML},H)p(a_{\rm ML}|H)\Delta a = \mathscr{L}_{\rm ML}(d|H)\Delta V/V$$



- Bayes factor:

$$\mathscr{B}_{10}(d) \equiv \frac{p(d|H_1)}{p(d|H_0)} = \frac{\int \mathrm{d}a_1\, p(d|a_1,H_1)p(a_1|H_1)}{\int \mathrm{d}a_0\, p(d|a_0,H_0)p(a_0|H_0)} \simeq \Lambda_{\rm ML}(d)\frac{\Delta V_1/V_1}{\Delta V_0/V_0}$$

- The $\Delta V/V$ factors penalize hypotheses that uses more parameter space volume V than necessary to fit the data ΔV (**Occam's penalty factor**)

- evidence calculation usually requires marginalization over parameer values
- when the data are informative can write evidence in terms of ML value of the likelihood
- BF \simeq ratio of maxima of likelihood functions time penalty factors related to how much parameter space volume is needed to fit the data
- using more parameters or parameter space volume than necessary is penalized in the Bayesian approach (occam's factor)

Significance of Bayes factor values

approximately equal to the squared SNR of the data

$\mathcal{B}_{\alpha\beta}(d)$	$2\ln\mathcal{B}_{\alpha\beta}(d)$	Evidence for model $\mathcal{M}_\alpha$ relative to $\mathcal{M}_\beta$
<1	<0	Negative (supports model $\mathcal{M}_\beta$)
1–3	0–2	Not worth more than a bare mention
3–20	2–6	Positive
20–150	6–10	Strong
>150	>10	Very strong

Adapted from Kass and Raftery (1995)

- evidence for one model relative to another and its connection to different values of BFs and ln BFs.
- note that 2 ln BF \simeq the squared SNR of the data

IV. Exercises / worked examples

- some exercises / worked examples that you can work on either now or later
- solutions are available in romano_notes1.pdf and romano_code1.ipynb

---

1. Practical application of Bayes' theorem

- Suppose on your last visit to the doctor's office you took a test for some rare disease. This type of disease occurs in only 1 out of 10,000 people, as determined by a random sample of the population. The test that you took is rather effective in that it can correctly identify the presence of the disease 95% of the time, but it gives false positives 1% of the time.
- Suppose the test came up positive. What is the probability that you have the disease?

- BT example in terms of a rare disease

Solution to Bayes' theorem problem

- H = have the disease; + = test positive
- Information:

$$P(H) = 0.0001 \qquad P(\bar{H}) = 0.9999$$
$$P(+\,|\,H) = 0.95 \qquad P(+\,|\,\bar{H}) = 0.01$$

- Calculate:

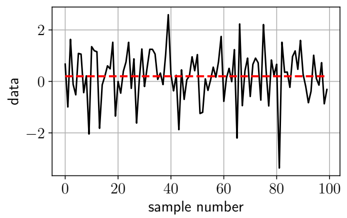
$$P(H\,|\,+) = \frac{P(+\,|\,H)P(H)}{P(+)} \qquad \begin{aligned} P(+) &= P(+\,|\,H)P(H) + P(+\,|\,\bar{H})P(\bar{H}) \\ &= 0.95 \times 0.0001 + 0.01 \times 0.9999 \\ &\approx 0.01 \end{aligned}$$

- Final result:

$$P(H\,|\,+) \approx 0.0095 \approx 0.01$$

- even though you tested positive, the chance for a false positive (0.01) is larger than your chance probability of having the disease (0.0001).
- the probability that you have the disease given that you tested positive is 100 times greater than your chance of having the disease without doing the test

2. Comparing frequentist and Bayesian analyses for a constant amplitude signal in white noise



- constant amplitude signal in white noise

Key formulae

Likelihoods functions:

$$p(d\,|\,\mathscr{M}_0) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N d_i^2\right]$$

$$p(d\,|\,a, \mathscr{M}_1) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N (d_i - a)^2\right]$$

Prior:

$$p(a\,|\,\mathscr{M}_1) = \frac{1}{a_{\max}}$$

Parameter choices:

$N = 100\,,\quad \sigma = 1\,,\quad 0 \le a \le a_{\max}\,,\quad a_0 =$ true value

- key formulae for this expression (should derive analytically then implement using code)

Key formulae

Maximum-likelihood estimator:

$$\hat{a} \equiv a_{\rm ML}(d) = \frac{1}{N}\sum_{i=1}^N d_i \equiv \bar{d} \qquad \sigma_{\hat{a}}^2 = \frac{\sigma^2}{N}$$

Useful identity:

$$\sum_{i=1}^N (d_i - a)^2 = \sum_i d_i^2 - N\hat{a}^2 + N(a-\hat{a})^2 = N\left(\mathrm{Var}[d] + (a-\hat{a})^2\right)$$

Likelihood function (in terms of ML estimator):

$$p(d\,|\,a, \mathscr{M}_1) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{\mathrm{Var}[d]}{2\sigma_{\hat{a}}^2}\right] \exp\left[-\frac{(a-\hat{a})^2}{2\sigma_{\hat{a}}^2}\right]$$

Evidence:

$$p(d\,|\,\mathscr{M}_1) = \frac{\exp\left[-\frac{\mathrm{Var}[d]}{2\sigma_{\hat{a}}^2}\right]\left[\mathrm{erf}\left(\frac{a_{\max}-\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right) + \mathrm{erf}\left(\frac{\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)\right]}{2a_{\max}\left(\sqrt{2\pi}\sigma\right)^{N-1}\sqrt{N}}$$

Posterior distribution:

$$p(a\,|\,d, \mathscr{M}_1) = \frac{1}{\sqrt{2\pi}\sigma_{\hat{a}}}\exp\left[-\frac{(a-\hat{a})^2}{2\sigma_{\hat{a}}^2}\right] 2\left[\mathrm{erf}\left(\frac{a_{\max}-\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right) + \mathrm{erf}\left(\frac{\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)\right]^{-1}$$

Key formulae

Bayes factor:

$$\mathscr{B}_{10}(d) = \exp\left[\frac{\hat{a}^2}{2\sigma_{\hat{a}}^2}\right]\left(\frac{\sqrt{2\pi}\sigma_{\hat{a}}}{a_{\rm max}}\right)\frac{1}{2}\left[\mathrm{erf}\left(\frac{a_{\rm max}-\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)+\mathrm{erf}\left(\frac{\hat{a}}{\sqrt{2}\sigma_{\hat{a}}}\right)\right] \simeq \exp\left[\frac{\hat{a}^2}{2\sigma_{\hat{a}}^2}\right]\left(\frac{\sqrt{2\pi}\sigma_{\hat{a}}}{a_{\rm max}}\right)$$

Maximum likelihood ratio statistic:

$$\Lambda_{\rm ML}(d) = \exp\left(\frac{\hat{a}^2}{2\sigma_{\hat{a}}^2}\right)$$

Frequentist test statistic:

$$\Lambda(d) \equiv 2\ln\Lambda_{\rm ML}(d) = \frac{\hat{a}^2}{\sigma_{\hat{a}}^2} = \left(\frac{\sqrt{N}\bar{d}}{\sigma}\right)^2 \equiv \rho^2$$

Sampling distributions of the test statistic:

$$p(\Lambda\,|\,\mathscr{M}_0) = \frac{1}{\sqrt{2\pi\Lambda}}e^{-\Lambda/2}$$

$$p(\Lambda\,|\,a,\mathscr{M}_1) = \frac{1}{\sqrt{2\pi\Lambda}}\frac{1}{2}\left[e^{-\frac{1}{2}(\sqrt{\Lambda}-\sqrt{\lambda})^2}+e^{-\frac{1}{2}(\sqrt{\Lambda}+\sqrt{\lambda})^2}\right] \qquad \lambda = \langle\rho\rangle^2 = \frac{Na^2}{\sigma^2}$$

See romano_notes1.pdf and romano_code1.ipynb for solutions