Practice session: Statistics for Astronomy

Kapteyn Learning Community Sirius A
Give motivations and/or derivations for your answers.

1. **Basic Statistics**

   (a) What is the difference between deductive logic and plausible reasoning (top-down vs. bottom-up logic respectively)?

   (b) Formulate Cox's sum rule and product rule. Given that $\text{prob}(a \mid I) = 1$, what is $\text{prob}(\text{not } a \mid I)$?

   (c) Describe Ockham's razor.

   (d) What are two main differences between Bayesian statistics and frequentist statistics?

   (e) Use the product rule to formulate Bayes' theorem.

   (f) Give the name and the formula of the technique used to obtain $\text{prob}(X \mid I)$ from $\text{prob}(X, Y \mid I)$.

Solution:

(a) The difference between deductive logic and plausible reasoning is, that deductive logic starts from the theory and applies this to the results, while plausible reasoning starts with the results and implies from these the theory.

(b)

$$\text{prob}(X \mid I) + \text{prob}(\overline{X} \mid I) = 1$$

$$\text{prob}(X, Y \mid I) = \text{prob}(X \mid Y, I) \times \text{prob}(Y \mid I)$$

$$\text{prob}(a \mid I) = 1,\ \text{prob}(a \mid I) + \text{prob}(\overline{a} \mid I) = 1 \implies \text{prob}(\overline{a} \mid I) = 0$$

(c) *"Accept the simplest explanation that fits the data"*.

(d) Bayesian statistics uses a prior function and allows data to be added later.

(e)

$$\begin{aligned} \text{prob}(X, Y \mid I) &= \text{prob}(Y, X \mid I) \\ &\downarrow \text{ product rule} \\ \text{prob}(X \mid Y, I) \times \text{prob}(Y \mid I) &= \text{prob}(Y \mid X, I) \times \text{prob}(X \mid I) \\ \text{prob}(X \mid Y, I) &= \frac{\text{prob}(Y \mid X, I) \times \text{prob}(X \mid I)}{\text{prob}(Y \mid I)} \end{aligned}$$

if we now replace X with Hypothesis and Y with data you get Bayes theorem.

(f) You use marginalization

$$\text{prob}(X \mid I) = \int_{-\infty}^{\infty} \text{prob}(X, Y \mid I) dY$$

2. **Distributions**

   (a) Formulate the principle of indifference.

   (b) For what type of data are the following distributions useful: binomial, Poisson and Gaussian.

(c) Give the general formula for finding an expectation value $\langle x \rangle$ from a general continuous distribution $f(x)$.

(d) Formulate the central limit theorem.

(e) Which distribution type does not tend to a Gaussian distribution, and why?

Solution:

(a) Suppose that there are $n > 1$ mutually exclusive and collectively exhaustive possibilities. The principle of indifference states that if the n possibilities are indistinguishable except for their names, then each possibility should be assigned a probability equal to $\frac{1}{n}$.

(b) Binomial distribution is used when the outcome of the data is either a succes or a failure.
The Poisson distribution is used when the mean amount of independent events is known and we want to know the probability of N events happening.
Gaussian can be used for almost all data that is scattered around a single mean value.

(c)
$$\langle x \rangle = \int_{-\infty}^{\infty} x f(x) dx$$

(d) When the amount of data points goes towards infinity all distributions with a defined mean and variance tend towards a Gaussian distribution.

(e) The Cauchy distribution does not tend towards a Gaussian since it has an undefined mean and an infinite variance.

3. **Double Parameter Estimation and Tests**

(a) Can you marginalize out parameters in the 2D probability distribution? (Example: a signal on a background.)

(b) What additional variance term shows up when you have two parameters and what does it say about the relation between the two parameters?

(c) Draw a correlated, anti-correlated and uncorrelated probability distribution.

Solution:

(a) Yes.

(b) The covariance σ_{xy}^2 pops up and holds information about the correlation between x and y

(c) Correlated should look like an ellipse with the axis along $x = y$, anti-correlated should be an ellipse with the axis along $x = -y$. uncorrelated should have it's axis parallel to the x or y axis

4. **Your very own Spaceship**
The total change in velocity for a rocket-like vehicle is described by Tsiolkovsky's rocket equation:

$$\Delta v = v_e \ln\left(\frac{m_i}{m_f}\right) \tag{1}$$

Where v_e is the effective exhaust velocity and m_i and m_f are the total initial mass (including propellant) and total final mass (without propellant), respectively.

To your best knowledge, you have determined the following data concerning the fuel consumption of your very own rocket:

v_e (km/s)	σ_{v_e}	m_i (kg)	σ_{m_i}	m_f (kg)	σ_{m_f}
5.1450	0.0522	5.085×10^5	7250	1.018×10^5	2545

Assume that v_e, m_i and m_f are independent of each other. Let's say we want to calculate the uncertainty in Δv.

(a) Express $\sigma^2_{\frac{m_i}{m_f}}$ in quantities from the table, and then plug in the numbers from the table.

(b) Express $\sigma^2_{\Delta v}$ using the quantities from the table. What is the uncertainty in Δv in km/s?

Solution:

(a) The covariance $\sigma^2_{m_i m_f}$ is 0, hence:

$$\sigma^2_{\frac{m_i}{m_f}} = \left(\frac{m_i}{m_f}\right)^2 \cdot \left(\left(\frac{\sigma_{m_i}}{m_i}\right)^2 + \left(\frac{\sigma_{m_f}}{m_f}\right)^2\right) = 2.067 \times 10^{-2}$$

(b)

$$\begin{aligned}\sigma^2_{\Delta v} &= \left(\frac{\partial(\Delta v)}{\partial v_e}\right)^2 \sigma^2_{v_e} + \left(\frac{\partial(\Delta v)}{\partial m_i}\right)^2 \sigma^2_{m_i} + \left(\frac{\partial(\Delta v)}{\partial m_f}\right)^2 \sigma^2_{m_f} \\ &= \left(\ln\left[\frac{m_i}{m_f}\right]\sigma_{v_e}\right)^2 + \left(\frac{v_e \sigma_{m_i}}{m_i}\right)^2 + \left(\frac{v_e \sigma_{m_f}}{m_f}\right)^2\end{aligned}$$

Then,

$\sigma^2_{\Delta v} = 2.90 \times 10^{-2}\,\mathrm{km}^2/\mathrm{s}^2$

5. **Parameter Estimation using Bayesian Statistics**

In astronomy, it is a standard practice to calculate the flux of a star. The first task in making this calculation is to measure the number of photons observed per second of that star. A single observation would give a good first estimate, but we ideally would want to take multiple exposures and combine them to get a better estimate for the average number of observed photons (we call this λ). In this problem we will find the best estimate for this number and the associated uncertainty.

Since we are essentially counting photons, the probability of observing k photons (in a one second exposure) for some value of λ (also per second) is given by:

$$\mathrm{prob}(k \mid \lambda, I) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{2}$$

Note, that we ignore the background signal and all uncertainties associated with the instrumentation.

(a) What is this distribution called?

Suppose now, we perform N observations, resulting in the set of observed number of photons $\mathbf{k} = \{k_1, \ldots, k_N\}$.

(b) Assuming independent observations, write down the joint likelihood of observing $\mathbf{k}$

$$\mathrm{prob}(\mathbf{k} = \{k_1, \ldots, k_N\} \mid \lambda, I) \tag{3}$$

(c) Write down Bayes' Theorem. What are each of the terms called? Which terms can usually be ignored in the context of parameter estimation?

(d) Use Bayes' Theorem to write down the posterior for λ given our observations $\mathbf{k}$:

$$\text{prob}(\lambda \mid \mathbf{k} = \{k_1, \ldots, k_N\}, I) \tag{4}$$

What prior should we assume here?

We now want to find the best estimate for the average number of photons per second given our observations. In order to do this we will find the value of λ_0 for which the posterior is maximal.

(e) First, write down the logarithm of the posterior L:

$$L = \ln\left[\text{prob}(\lambda \mid \mathbf{k}, I)\right] \tag{5}$$

and then find λ_0 for which L is maximal.

Finally, we want to find the uncertainty σ_{λ_0} in our estimate. For the Gaussian assumption the uncertainty is given by:

$$\sigma_{\lambda_0} = \left[-\frac{\partial^2 L}{\partial \lambda^2}\bigg|_{\lambda_0}\right]^{-1/2} \tag{6}$$

(f) Find σ_{λ_0} in terms of our best estimate λ_0 and the number of exposures N.

(g) Using the previously obtained result(s), reason what would happen to λ_0 when we take additional observations (say, twice as many as we did before).

Solution:

(a) The Poisson distribution

(b)

$$\text{prob}(\mathbf{k} = \{k_1, \ldots, k_N\} \mid \lambda, I) = \prod_{i=1}^{N} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!}$$

You can rewrite this even more by using the fact that $\Pi_i \lambda^{k_i} = \lambda^{\sum k_i}$ and $\Pi_i e^{-\lambda} = e^{-\sum \lambda} = e^{-N\lambda}$ to get:

$$\prod_{i=1}^{N} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} = \lambda^{\sum_i k_i} e^{-N\lambda} \prod_{i=1}^{N} \frac{1}{k_i!}$$

(c) Bayes' theorem for some hypothesis A and data B:

$$\text{prob}(A \mid B, I) = \frac{\text{prob}(B \mid A, I)\text{prob}(A, I)}{\text{prob}(B, I)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

The evidence is a normalization factor (i.e. a sum over possible hypotheses) and so it does not depend on the hypothesis / parameter in the posterior. We can therefore ignore it since we will take the derivative of the posterior (w.r.t. the parameter) and all constant factors will be zero. We therefore usually write Bayes' theorem with a proportionality sign:

$$\text{prob}(A \mid B, I) \propto \text{prob}(B \mid A, I)\text{prob}(A, I)$$

If we also assume a flat/uniform prior, this will be a constant factor as well. Bayes' theorem then reduces to:

$$\text{prob}(A \mid B, I) \propto \text{prob}(B \mid A, I)$$

(d) We assume a uniform / flat prior here so:

$$\begin{aligned}\text{prob}(\lambda \mid \mathbf{k}, I) &\propto \text{prob}(\mathbf{k} \mid \lambda, , I) \\ &\propto \lambda^{\sum_i k_i} e^{-N\lambda} \frac{1}{\prod_{i=1}^{N} k_i!} \\ &\propto \lambda^{\sum_i k_i} e^{-N\lambda}\end{aligned}$$

In the last step, we ignore the term that does not involve λ

(e) We first find L, with C being some proportionality constant:

$$L = \ln\left(C \times \lambda^{\sum_i k_i} e^{-N\lambda}\right) = \ln C + \sum_i k_i \ln\lambda - N\lambda$$

Then, the maximum value of L is given by the following:

$$\left.\frac{\partial L}{\partial \lambda}\right|_{\lambda_0} = 0$$

we compute this derivative and then we solve for λ_0:

$$\begin{aligned}\left.\frac{\partial L}{\partial \lambda}\right|_{\lambda_0} &= \frac{\sum_i k_i}{\lambda_0} - N = 0 \\ \implies \lambda_0 &= \frac{\sum_i k_i}{N}\end{aligned}$$

(f) We first compute the second derivative, and use the definition of $\sum_i k_i = N\lambda_0$ (what we just found).

$$\begin{aligned}\left.\frac{\partial^2 L}{\partial \lambda^2}\right|_{\lambda_0} &= -\frac{\sum_i k_i}{\lambda_0^2} \\ &= -\frac{N\lambda_0}{\lambda_0^2} \\ &= -\frac{N}{\lambda_0}\end{aligned}$$

The uncertainty is then given by:

$$\sigma_{\lambda_0} = \left[-\left.\frac{\partial^2 L}{\partial \lambda^2}\right|_{\lambda_0}\right]^{-1/2} = \left[\frac{N}{\lambda_0}\right]^{-1/2} = \sqrt{\frac{\lambda_0}{N}}$$

(g) Since the standard deviation scales with $1/\sqrt{N}$, doubling the amount of observations would reduce the standard deviation with a factor $\sqrt{2}$. Therefore, the estimate λ_0 will become more precise with additional observations.