Statistics for Astronomers
Solutions to Midterm Exam

Prof. Sundar Srinivasan

November 11, 2020

Note: Solutions to Questions 5, 6, 7, and 8 use modules in the script here.

1. (a) The marginal pdf for x is

$$p_X(x) = \int_x^\infty p_{X,Y}(x,y)dy = \int_x^\infty e^{-y}dy = e^{-x}$$

The marginal pdf for y is

$$p_Y(y) = \int_0^y p_{X,Y}(x,y)dx = e^{-y}\int_0^y dx = ye^{-y}.$$

(b) The conditional probabilities are

$$p_X(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y}.$$
$$p_Y(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{x-y}.$$

(c) Since $p_X(x|y) \neq p_X(x)$, x and y are not independent.

2. (a) If $t = u^\alpha$, $\frac{dt}{du} = \alpha u^{\alpha-1} = \alpha t/u \implies \frac{du}{dt} = u/(t\alpha) = \frac{t^{-1+1/\alpha}}{\alpha}$. Therefore,

$$p_T(t) = p_U(u)\frac{du}{dt} = \frac{t^{-1+1/\alpha}}{\alpha}$$

Comparing the above equation with the desired distribution, we can draw from a uniform distribution and then set $t = u^3$.

(b) Let the normalisation constant be C, so that

$$C\int_{-\infty}^{\infty} \frac{dx}{\frac{x^2}{\nu}+1} = 1 \tag{1}$$

Substituting $x = \sqrt{\nu}\ \tan\theta$, so that $\theta \in (-\pi/2, \pi/2)$, Equation (1) becomes

$$C \int_{-\pi/2}^{\pi/2} \frac{d\theta\ \sqrt{\nu}\ \sec^2\theta}{\sec^2\theta} = C\ \sqrt{\nu} \int_{-\pi/2}^{\pi/2} d\theta = \pi\ \sqrt{\nu}\ C = 1 \Longrightarrow C = \frac{1}{\sqrt{\nu}\ \pi}$$

The CDF for X is given by

$$F_x(x) = \frac{1}{\pi\sqrt{\nu}} \int_{-\infty}^{x} \frac{dy}{\frac{y^2}{\nu} + 1} = \frac{1}{\pi}\Big[\tan^{-1}\Big(\frac{x}{\sqrt{\nu}}\Big) + \frac{\pi}{2}\Big] = \frac{1}{\pi}\tan^{-1}\Big(\frac{x}{\sqrt{\nu}}\Big) + \frac{1}{2}$$

If we set $Y \equiv F_x(x)$, then the Probability Integral Transform ensures that Y is a uniform random variable. We can invert this to write X in terms of Y:

$$y = F_x(x) = \frac{1}{\pi}\tan^{-1}\Big(\frac{x}{\sqrt{\nu}}\Big) + \frac{1}{2} \Longrightarrow x = \sqrt{\nu}\ \tan\Big[\pi\Big(y - \frac{1}{2}\Big)\Big] \tag{2}$$

Thus, we can draw uniform random variates Y from $U(0,1)$ and then relate X to Y using Equation (2) to ensure that X will have the desired distribution.

3. The studentised version of the sample mean is $T = \dfrac{m-\mu}{s/\sqrt{N}}$, where N is the number of observations. T has the Student's t distribution with $\nu = N-1$ degrees of freedom (and not N, because the sample standard deviation s is computed using the sample mean as an estimator for the unknown population mean).

   Therefore, T has mean 0 and variance $\dfrac{\nu}{\nu-2} = \dfrac{N-1}{N-3}$.

   The problem asks about the quantity $\dfrac{m-\mu}{s}$, which is $\sqrt{N}$ times T. It has the same distribution as T, with the same mean (zero), but its variance is N times Var$[T]$, which is $N\ \dfrac{N-1}{N-3} = 27.27$.

4. The likelihood for N data points $x_i,\ i = 1, \cdots, N$ is

$$\mathscr{L}(\mu) \propto \prod_{i=1}^{N} \exp\left(-|x_i - \mu|\right) \Longrightarrow \ln\mathscr{L}(\mu) = \text{constant} - \sum_{i=1}^{N} |x_i - \mu|$$

   The term inside the summation is $x_i - \mu$ for $x_i > \mu$ and $\mu - x_i$ for $x_i < \mu$. The derivative of each term is, accordingly, minus or plus 1. Therefore,

$$\frac{\partial}{\partial\mu}\ln\mathscr{L}(\mu) = (\#\text{ of points with values} > \mu) - (\#\text{ of points with values} < \mu)$$

   That is, the derivative equals the **net** number of points with values above μ. Since the MLE $\hat{\mu}$ is the value of μ at which the derivative vanishes, it is also the value of μ that has an equal number of data points on either side of it. **This is the definition of the median of the data set**. Therefore, $\hat{\mu} = \text{Median}(\{x_i\})$. **The sample median minimises the absolute deviation**, just as the sample mean minimises the (signed) deviation.

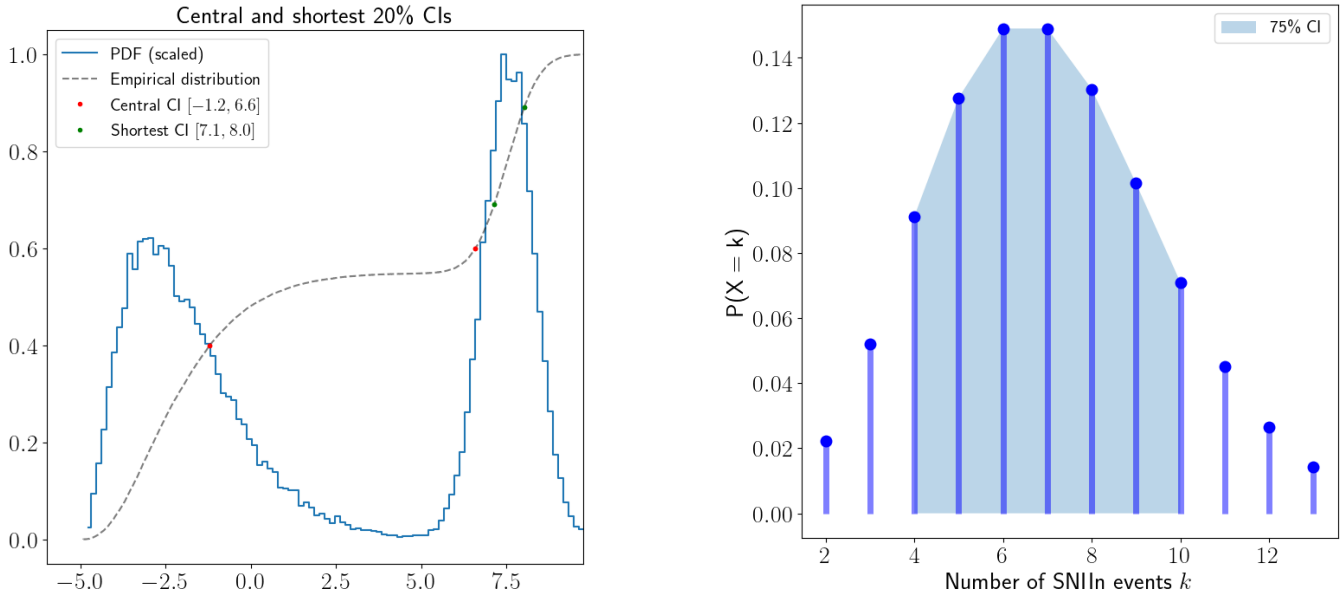

Figure 1: *Left*: the lower and upper limits of the central and shortest 20% CIs plotted onto the PDF and EDF for Problem 5. *Right*: the 75% central CI plotted onto the PMF for Problem 6.

5. The PDF and EDF are shown in Figure 1. The distribution is bimodal; although the global maximum is located at $x = 7.5$, there is also a local maximum at $x \approx -3$. These peaks are separated by a region of low probability ($1 \lesssim x \lesssim 6$). From the empirical distribution, it is clear that the region around the global maximum has at least 40% of the total probability mass. Therefore, the 20% shortest CI must be inside the range $[5, 8]$.

   The module `prob5` in `midterm.py` first computes the empirical distribution function for the sample. It then computes the central and shortest CIs:
   ```
   The central 20.0% CI is [-1.2, 6.6].
   The shortest 20.0% CI is [7.1, 8.0].
   ```
   The lower and upper limits of each interval are shown on the EDF in the figure (filled circles), and it is easy to verify that the enclosed probability in each case is 20%.

6. As the SN events are independent of each other, this is a Poisson problem. The rate parameter is $7\ \mathrm{yr}^{-1} \times 1\ \mathrm{yr} = 7$. For a Poisson problem, the expectation value equals the rate parameter; our symmetric CI is then centred at 7. The module `prob6` in `midterm.py` stretches outward in both directions from this centre and selects the symmetric interval enclosing a probability as close to 75% as possible. This turns out to be the interval $[4, 10]$. Figure 1 shows this CI overlaid onto the PMF for the problem.

7. The module `prob7` in `midterm.py` is used for this problem. It is a Poisson problem, since we are counting photons. Therefore, the variance associated with N counts is also N. For this problem, we assume that the count in each bin is independent of the counts in the other bins.

   (a) To compute the mean continuum counts, we sum over the eight bins (four on either side of the line region) and divide by eight. We can use the Bienaymé Identity to compute the variance on

this mean, which turns out to be the mean divided by eight.

$$C_{\text{cont.mean}} = \frac{1}{8}\Big(C_1 + C_2 + C_3 + C_4 + C_{12} + C_{13} + C_{14} + C_{15}\Big)$$
$$\text{Var}[C_{\text{cont.mean}}] = \frac{1}{8^2}\Big(\text{Var}[C_1] + \text{Var}[C_2] + \text{Var}[C_3] + \text{Var}[C_4] + \text{Var}[C_{12}] + \text{Var}[C_{13}] + \text{Var}[C_{14}] + \text{Var}[C_{15}]\Big)$$
$$= \frac{1}{8^2}\Big(C_1 + C_2 + C_3 + C_4 + C_{12} + C_{13} + C_{14} + C_{15}\Big) = \frac{1}{8}\, C_{\text{cont.mean}} \qquad (3)$$

The standard error is then the square root of this variance. The script outputs
`The mean continuum count is 1415 ± 13 photons.`

(b) There are a total of seven bins in the line region. The total continuum contribution is therefore seven times the mean continuum count. The variance on this total is 7^2 times the variance on the mean continuum count.

$$C_{\text{cont.underline}} = 7\; C_{\text{mean}} \Longrightarrow \text{Var}[C_{\text{cont.underline}}] = 7^2\; \text{Var}[C_{\text{mean}}] = \frac{7^2}{8}\; C_{\text{cont.mean}} \qquad (4)$$

The standard error is then the square root of this variance. The script outputs
`The total continuum count under the line is 9903 ± 93 photons.`

(c) The total count in the line region is the sum of the counts in each of the seven bins. The variance on this total is the sum of the variances of the counts in each bin. When the continuum contribution is subtracted from the total count, the variance on this difference is nothing but the sum of the variance on the total count and the variance on the continuum contribution.

$$C_{\text{line,total}} = \sum_{i=5}^{11} C_i - C_{\text{cont.underline}} = \sum_{i=5}^{11} C_i - 7\; C_{\text{cont.mean}}$$
$$\text{Var}[C_{\text{line,total}}] = \sum_{i=5}^{11} \text{Var}[C_i] + \text{Var}[C_{\text{cont.underline}}] = \sum_{i=5}^{11} C_i + \frac{7^2}{8}\; C_{\text{cont.mean}} \qquad (5)$$

The standard error is then the square root of this variance. The script outputs
`The integrated line flux is 7177 ± 160 photons.`

(d) The advantage of reporting the integrated flux in the line as opposed to the flux at line centre is the increased signal-to-noise ratio in the result. Due to the Bienaymé Identity, the standard errors adding in quadrature means that the relative uncertainty on the total is lower than that of just the central value. The script outputs:
`The integrated line flux is 7177 ± 160 photons.`
`The line flux at the line centre is 1861 ± 68 photons.`
The signal-to-noise in the integrated flux is ≈ 45, whereas the signal-to-noise in the central flux is ≈ 27.

8. (a) There are four bins (three nearest-neighbour pairs) on either side of the line region, leading to a total of six covariance terms added to the computation of the mean continuum level. The following must be added to the RHS of Equation 3:

$$\text{Cov}[C_1, C_2] + \text{Cov}[C_2, C_3] + \text{Cov}[C_3, C_4] + \text{Cov}[C_{12}, C_{13}] + \text{Cov}[C_{13}, C_{14}] + \text{Cov}[C_{14}, C_{15}].$$

In general, for each of these terms, $\mathrm{Cov}[C_i, C_{i+1}] = 0.5\sqrt{\mathrm{Var}[C_i]\mathrm{Var}[C_{i+1}]} = 0.5\sqrt{C_i C_{i+1}}$. Since the standard error adds this term in quadrature to the other terms in Equation 3, the net effect is not very large. In the absence of any covariance, the covariance matrix for the first four wavelength bins is a diagonal matrix, with the diagonal elements equal to $\mathrm{Var}[C_i]$ $(i = 1, 2, 3, 4)$. With covariance among nearest-neighbour bins, the off-diagonal elements closest to the diagonal are now populated. That is,

$$\Sigma = \begin{bmatrix} \mathrm{Var}[C_1] & \mathrm{Cov}[C_1, C_2] & 0 & 0 \\ \mathrm{Cov}[C_1, C_2] & \mathrm{Var}[C_2] & \mathrm{Cov}[C_2, C_3] & 0 \\ 0 & \mathrm{Cov}[C_2, C_3] & \mathrm{Var}[C_3] & \mathrm{Cov}[C_2, C_4] \\ 0 & 0 & \mathrm{Cov}[C_3, C_4] & \mathrm{Var}[C_4] \end{bmatrix}$$

$$= \begin{bmatrix} C_1 & 0.5\sqrt{C_1 C_2} & 0 & 0 \\ 0.5\sqrt{C_1 C_2} & C_2 & 0.5\sqrt{C_2 C_3} & 0 \\ 0 & 0.5\sqrt{C_2 C_3} & C_3 & 0.5\sqrt{C_3 C_4} \\ 0 & 0 & 0.5\sqrt{C_3 C_4} & C_4 \end{bmatrix} \tag{6}$$

The `prob8` module outputs:

```
The covariance matrix for the first four wavelength bins:
[[1337  676    0    0]
 [ 676 1365  705    0]
 [   0  705 1454  688]
 [   0    0  688 1299]]
```

(b) The `prob7` module, when run with the argument `rho = 0.5`, outputs:

```
The mean continuum count is 1415 ± 16 photons.
```

There is a modest increase in the standard error associated with the mean continuum count (13 photons to 16 photons), which propagates into the standard error on the line flux.

The variance associated with the line flux also increases, and has not been accounted for in the above as the problem only required us to compute the change in variance in the continuum.