641 PS1

Mohan Agrawal *Sep 21, 2022*

Q 1

Let's start with the standard Poisson probability distribution

$$P(k;\lambda) = e^{-\lambda}\lambda^k/k! \tag{1}$$

and express k in terms of the distance from the mean (λ) denoted by x, such that $x = k-\lambda$. This gives

$$P(x) = e^{-\lambda}\lambda^{(\lambda+x)}/(\lambda+x)! \tag{2}$$

or

$$log[P(x)] = -\lambda + (\lambda+x)log[\lambda] - log[(\lambda+x)!] \tag{3}$$

Since we let $\lambda \longrightarrow \infty$, we can apply Stirling's approximation, which says —

$$log(n!) = nlog(n) - n + \sqrt{2\pi n} \tag{4}$$

for very large n, to the right-most term in (3). Rearranging the terms, we get

$$\begin{aligned}
log[P(x)] &= x - log[\sqrt{2\pi(\lambda+x)}] - (\lambda+x)log\Big[\frac{\lambda+x}{\lambda}\Big] && (5)\\
&= x - log[\sqrt{2\pi(\lambda+x)}] - (\lambda+x)log\Big[1+\frac{x}{\lambda}\Big] && (6)\\
&= x - log[\sqrt{2\pi(\lambda+x)}] - (\lambda+x)\Big(\frac{x}{\lambda} - \frac{x^2}{2\lambda^2}\Big) && (7)\\
&= -log[\sqrt{2\pi(\lambda+x)}] - \frac{x^2}{\lambda} + \frac{x^2}{2\lambda} + \frac{x^3}{2\lambda^2} && (8)\\
&\approx -log[\sqrt{2\pi\lambda}] - \frac{x^2}{2\lambda} && (9)
\end{aligned}$$

Since we're probing the vicinity of the mean [1], $x < \lambda$, and $\lambda >> 1$. Therefore, right-most term in (8) has been dropped, and $\lambda + x \approx \lambda$.

Thus, we get, finally

$$P(x) = \frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{x^2}{2\lambda}} = \frac{1}{\sqrt{2\pi\lambda}}e^{-\frac{(k-\lambda)^2}{2\lambda}} \tag{10}$$

which has the same form as a Gaussian probability distribution function with $\sigma = \sqrt{\lambda}$.

[1] NB - Whether or not this approximation is valid at large distances from the mean is the whole point of question 2, where it is analyzed in detail.

Q 2

NB - the variable called k here and in the code is what Jon calls n.

If Gaussian approximation is good enough, then the probability obtained from (10) should match that obtained from (2) to within a factor of 2 for any x. We test this at two values of x, 3σ and 5σ.

But we know from Q 1 that $\sigma = \sqrt{\lambda}$, and $x = k - \lambda$. Therefore, we can write

$$k = \lambda + s\sqrt{\lambda} \tag{11}$$

where s is either 3 or 5.

Since we have a relation between k and λ, we can either get a Poisson probability distribution as function of k or λ. If we choose to express k in terms of λ, i.e. $P(k(\lambda), \lambda|s)$, our probability at $x = s\sigma$ will fluctuate since k is constrained to be an integer. We'll have to take a floor/ceil of (11). Therefore, it is better to instead express λ in terms of k, since λ can be continuous.

Let $\lambda = d^2$, then the above equation becomes a quadratic in d.

$$k = d^2 + sd \tag{12}$$

The equation will have two solutions for d, but d has to be positive since k is constrained to be positive. Taking the positive solution we get

$$d = \frac{\sqrt{s^2 + 4k} - s}{2}. \tag{13}$$

Therefore,

$$\lambda(k) = \left(\frac{\sqrt{s^2 + 4k} - s}{2}\right)^2. \tag{14}$$

Furthermore, Gaussian distribution (10) at $x = s\sigma$ becomes

$$P = \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{s^2}{2}} \tag{15}$$

We can now compare the probability given by two distributions, see how close they are, and obtain the critical value of k where the ratio hits 2. This has been done in the Jupyter notebook.

We need $k \geq 16$ for the Gaussian to line up within 2x of Poisson at 3σ and $k \geq 694$ for 5σ.

Q 3

The fundamental equation for linear least-squares, maximum likelihood estimate of parameters can be written as:

$$A^T N^{-1} A m = A^T N^{-1} d \tag{16}$$

where A is the model, m the parameters, N the underlying noise-covariance matrix, and d the observed data.

Estimating maximum-likelihood mean of data is equivalent to fitting a constant, such that m is a scalar, and A is a (n,1) vector of ones, where n is the number of data points. Since the noise in the observed data is known to be σ, N is a (n.n) diagonal matrix of σ^2. In such a case then (16) reduces trivially to:

$$m = \frac{\sum_i d_i}{n} \tag{17}$$

which is simply the mean of the data. In a slightly different case, where the noise in each observation was different, we would have instead had the noise-weighted mean

$$m = \frac{\sum_i d_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \tag{18}$$

Above equation will be useful in Q 4.

We also know that the covariance matrix of the error in the fit parameters is given as:

$$< \delta m \delta m^T >= (A^T N^{-1} A)^{-1} \tag{19}$$

Once again, in the present case of estimating the mean, and where noise in each observation is the same, the uncertainty in the best-fit mean is simply

$$< \delta m^2 >= \frac{\sigma^2}{n} \tag{20}$$

or the RMSE being the familiar formula

$$\sigma_m = \frac{\sigma}{\sqrt{n}} \tag{21}$$

Let us write down (20) again with different noise for each data-point.

$$< \delta m^2 >= \left(\sum_i \frac{1}{\sigma_i^2} \right)^{-1} \tag{22}$$

If we got errors on half (n/2) of our data-points wrong (e.g. overestimated) by a factor of $\sqrt{2}$, we'd have:

$$< \delta m^2 >= \left(\sum_{i=1}^{n/2} \frac{1}{\sigma^2} + \sum_{i=n/2}^{n} \frac{1}{2\sigma^2} \right)^{-1} = \frac{\sigma^2}{3/4n} \tag{23}$$

or $\sigma_m = \sigma/\sqrt{0.75n}$.

Similarly, had we underestimated by $\sqrt{2}$, we'd have gotten

$$< \delta m^2 >= \left(\sum_{i=1}^{n/2} \frac{1}{\sigma^2} + \sum_{i=n/2}^{n} \frac{2}{\sigma^2} \right)^{-1} = \frac{\sigma^2}{3/2n} \tag{24}$$

or $\sigma_m = \sigma/\sqrt{1.5n}$. So we see that underestimation leads to a decrease in error-bars, and overestimation leads to an increase, as expected. (Further on this below.)

If we overweight 1% of our data by 100, it implies, weights in (18) are $100/\sigma_i^2$, which is numerically equivalent underestimation of the noise in data-points by 10 times. In this case, the error on the fit parameter will be

$$< \delta m^2 >= \left(\sum_{i=1}^{0.01n} \frac{100}{\sigma^2} + \sum_{i=0.01n}^{n} \frac{1}{\sigma^2} \right)^{-1} = \frac{\sigma^2}{1.99n} \tag{25}$$

or $\sigma_m \approx \sigma/\sqrt{2n}$.

Similarly, if we underweight 1% of our data (overestimation of noise in data)

$$< \delta m^2 >= \left(\sum_{i=1}^{0.01n} \frac{1}{100\sigma^2} + \sum_{i=0.01n}^{n} \frac{1}{\sigma^2} \right)^{-1} = \frac{\sigma^2}{0.9901n} \tag{26}$$

or $\sigma_m \approx \sigma/\sqrt{n}$.

Thus, we see that we must be careful with overweighting parts of our data since it will produce artificially low error-bars on our fit parameters. In real life, it manifests itself as underestimation of noise in data (which is same as overweighting as shown above). This can be deadly since we might be lead to believe there is a signal in our data when there's none — we just didn't characterize our noise properly. Underweighting data-points is practically neglecting 1% of data from our analysis. This is a common practice since we regularly have bad/corrupt chunks in the data being analyzed. The slightly higher error-bars indicate the incompleteness of our knowledge about the data being analyzed (since chunks of it have been removed), and thus are perfectly acceptable.

4

Please refer to the Jupyter notebook.

5

Let us write down chi-square for un-correlated data, where N is a diagnonal matrix.

$$\chi^2 = (d - A(m))^T N^{-1} (d - A(m)) \tag{27}$$

We can introduce an orthogonal coupling matrix V (any invertible matrix will work. Choosing orthogonal for ease of notation), such that $V^T V = V V^T = I$, in the above expression as

$$\chi^2 = (d - A(m))^T V^T V N^{-1} V^T V (d - A(m)) \tag{28}$$

This can be re-written as

$$\chi^2 = (\tilde{d} - \tilde{A}(m))^T \tilde{N}^{-1} (\tilde{d} - \tilde{A}(m)) \tag{29}$$

where $\tilde{N} = V N V^T$. This is straightforward to see. $\tilde{N}^{-1} = (V^T)^{-1} N^{-1} V^{-1}$, but $V^T = V^{-1}$, so $\tilde{N}^{-1} = V N^{-1} V^T$ as written in (28). Data and model's prediction have both been rotated into this new, correlated space, $\tilde{d} = V d$.

Now, consider the case where we have only uncorrelated noise as the data, and no model, and denote the realization of the noise as n. (Analogously, n is the residual in chi-square equation, e.g. (27)). Noise realization in correlated space will be $\tilde{n} = V n$. Covariance of the new, correlated noise can be written as:

$$< \tilde{n}\tilde{n}^T > = < V n (V n)^T > = < V n n^T V^T > \tag{30}$$

Since V is not random, we can take it outside the ensemble average.

$$< \tilde{n}\tilde{n}^T > = V < n n^T > V^T = V N V^T = \tilde{N} \tag{31}$$

QED.

Although I have proved this with vector/matrix notation, but it can just as easily be done in element notation by noting that

$$\begin{aligned} < \tilde{n}_i \tilde{n}_j > &= \sum_k V_{ik} n_k \sum_m V_{jm} n_m \\ &= \sum_m \sum_k V_{ik} < n_k n_m > V_{jm} \\ &= \sum_k V_{ik} \sigma_k^2 V_{jk} \end{aligned}$$

where $< n_k n_m >$ is non-zero only along the diagonal. Similarly,

$$\begin{aligned}\tilde{N}_{ij} &= \sum_m \sum_k V_{ik} N_{km} V_{jm} \\ &= \sum_k V_{ik} \sigma_k^2 V_{jk}\end{aligned}$$

Both expressions match.