Astronomical Statistics: Tutorial Solutions 2

John Peacock
C20 Royal Observatory; jap@roe.ac.uk

1. The number of parameters is $N_p = 10$, the number of data points is $N_D = 50$ so the number of degrees of freedom is $\nu = 50 - 10 = 40$. The fit gives a $\chi^2 = 53.3$ which means a reduced $\chi^2_R = 53.3/40 = 1.333$ with an error $\sqrt{2\nu}/\nu = 8.9/40 = 0.22$ so that $\chi^2_R = 1.33 \pm 0.22$. For five parameters the calculation is similar giving $\chi^2_R = 1.20 \pm 0.21$; from this we conclude that the 10 parameters achieve a worse fit than the 5-parameter fit.

2. The optimal way to average data is by reciprocal variance weighting. Thus our best estimate of Ω_m is $[0.24/0.015^2 + 0.30/0.025^2]/[[1/0.015^2 + 1/0.025^2] = 0.256$. With weights w_i, the variance is $\sum_i w_i^2\sigma_i^2/(\sum_i w_i)^2$, and for the optimal weighting, this is $1/(\sum_i w_i)$. Thus the rms error on our estimate is 0.013.

   Following the same reasoning, with errors reduced by a factor two for both experiments, we will obtain the same combined figure, and the rms error would halve: 0.256 ± 0.006. But before accepting this answer, we should do a goodness-of-fit sanity check: is the combined figure consistent with the original data? We can compute χ^2 (on 1 d.f., since the combined figure is in effect a minimum-χ^2 value). For the original data, this is

   $$\chi^2 = (0.24 - 0.256)^2/0.015^2 + (0.30 - 0.256)^2/0.025^2 = 4.2,$$

   which is not strongly significant: a 2.1σ result. But with the reduced errors, this becomes $\chi^2 = 16.9$, which is a 4.1σ result. Thus the input data are inconsistent with each other on the stated errors, and the combined figure should not be trusted.

   There is no certain way of proceeding from this point, but one common approach is to assert that the two experiments must have neglected some additional source of error, which we assume afflicts each experiment equally. Thus, we should replace the quoted errors by $(\sigma_i^2 + \epsilon^2)$, where ϵ^2 is the variance of the additional 'systematic' error. The only information we have on how big ϵ might be is from the difference in the two measurements (call them x and y): $\langle (x-y)^2 \rangle = \sigma_1^2 + \sigma_2^2 + 2\epsilon^2$. Since $x - y = 0.06$, this yields an estimate of $\epsilon = 0.0412$, so that the realistic errors on the two measurements would be 0.0419 and 0.0431. So now the weighted result is

0.269 ± 0.030 This is *less* accurate even though the experiments now claim to be more accurate. The reason is that at first we had no proof that systematic errors were present, but this became apparent as the experiments 'improved' their results.

3. $\exp[G]$ is a reasonable candidate for a density, since it is always positive, becoming zero in the limit that $G \to -\infty$. But in order to be correctly normalized as a fluctuation around the mean density, we need $\langle\delta\rangle = 0$, i.e. $\langle\exp[G + c]\rangle = 1$. The required integral is

$$\langle\exp[G]\rangle = \frac{1}{(2\pi)^{1/2}\sigma} \int \exp[G]\ \exp[-G^2/2\sigma^2]\ dG.$$

As usual, we need to complete the square, so that $G - G^2/2\sigma^2 = -(G - \sigma^2)^2/2\sigma^2 + \sigma^2/2$, yielding $\langle\exp[G]\rangle = \exp[\sigma^2/2]$ and hence $c = -\sigma^2/2$.

The variance is obtained similarly:

$$\langle\delta^2\rangle = \langle\exp[2G + 2c] - 2\exp[G + c] + 1\rangle = \langle\exp[2G + 2c]\rangle - 1.$$

Again, we need to complete the square: $2G - G^2/2\sigma^2 = -(G - 2\sigma^2)^2/2\sigma^2 + 2\sigma^2$, so that $\langle\delta^2\rangle = \exp[\sigma^2] - 1$. Thus the variances in δ and in G are equal when they are both small, but the variance in density becomes larger than that of the 'generating field' G when σ is large.

4. The integral probability for $z \equiv x + y$ is

$$P(< z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} p(x)p(y)\,dx\,dy,$$

and differentiating this to get $p(z)$ gives a convolution integral. For a product, $z \equiv xy$ and we might think to write the same expression with z/x in the second integral. The only subtlety is that, when x is negative, increasing z makes y more negative (draw a diagram to see this). Thus the integral probability is in two parts:

$$P(< z) = \int_{0}^{\infty}\int_{-\infty}^{z/x} p(x)p(y)\,dx\,dy, + \int_{-\infty}^{0}\int_{z/x}^{\infty} p(x)p(y)\,dx\,dy.$$

Now, differentiating with respect to z changes the sign depending on whether x is positive or negative; this can be combined into

$$p(z) = \int_{-\infty}^{\infty} p(x)p(z/x)\,dx/|x|.$$

Apply this to the case where $p(x)$ is 1 when $0 < x < 1$ and zero otherwise. We see immediately that z must also lie between 0 and 1. When $x < z$, $p(z/x) = 0$, and otherwise the product of the p's is unity. Thus

$$p(z) = \int_{z}^{1} dx/x = -\ln z = \ln(1/z).$$

It is easy to check via integration by parts that this is a correctly normalized pdf over $0 < z < 1$.

5. The Central Limit Theorem states that the sum of a set of N quantities drawn independently from a pdf of finite variance will tend to have a Gaussian distribution as $N \to \infty$.

   We can allow the two pdfs to have different properties, so that

$$p(z) = \int_{-\infty}^{\infty} p_1(x)p_2(z-x)\,dx = \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty} \exp[-(x-\mu_1)^2/2\sigma_1^2]\,\exp[-(z-x-\mu_2)^2/2\sigma_2^2]\,dx.$$

   Completing the square of the term in the exponential to group the x-dependent terms into the form $A(x+B)^2$ is a standard exercise, if a bit messy in this general case. But as usual, we can then integrate over x, and the remaining term is as we would expect: $-(z-\mu_1-\mu_2)^2/2(\sigma_1^2+\sigma_2^2)$, so we get a Gaussian in z. This result is simpler to prove in Fourier space, using the fact that the Fourier transform of a Gaussian is also a Gaussian.

   The Lorentzian (Cauchy distribution) clearly has a divergent variance, which is why it violates the Central Limit Theorem. It must be normalized, so that

$$1 = \int_{-\infty}^{\infty} \frac{A}{1+x^2/\sigma^2}\,dx = A\sigma\int_{-\infty}^{\infty}\frac{1}{1+y^2}\,dy = \pi A\sigma$$

   (where the last step needs a recognition of the standard $\tan^{-1}$ integral). Hence $A = 1/(\pi\sigma)$. If the characteristic function is $\phi(k) = \exp(-|k|\sigma)$, then the characteristic function for the sum is $\exp(-2|k|\sigma)$. This is the characteristic function of a Lorentzian of twice the 'width' (normally we would expect the rms to increase by $\sqrt{2}$ on adding two independent variables).

   For the 95% confidence range, repeating the above integral over a finite range gives

$$P(x_{\rm min} < x < x_{\rm max}) = \frac{1}{\pi}[\tan^{-1}(x_{\rm max}/\sigma) - \tan^{-1}(x_{\rm min}/\sigma)] = \frac{2}{\pi}\tan^{-1}(x_{\rm max}/\sigma).$$

   (where the last expression applies for a symmetric range). So for a 95% confidence range, we want $x_{\rm max}/\sigma = \tan(0.475\pi) = 12.7$, hence the required range in x is $\pm 12.7\sigma$. Since the effective width doubles on adding two such variables, this would become $\pm 25.4\sigma$.

6. If $P(>z) = \exp(-az^b)$, then $p(z) = abz^{b-1}\exp(-az^b)$. If we only have $z = 1$ measured, then

$$\mathcal{L} = p(1) = ab\exp(-a).$$

   Normally, we differentiate to maximize, but this won't give sensible results for b, since clearly $\mathcal{L}$ is maximized as $b \to \infty$ for any given a. Differentiating wrt a gives $b\exp(-a) - ab\exp(-a)$, which vanishes for $a = 1$. However, this derivation is more than a little dubious if b is infinite. Indeed, looking at the expression for $P(>z)$, we see that a divergent b makes this a step from $P = 1$ to $P = 0$ at $z = 1$ for any value of a (a reasonable outcome).

   For a more sensible answer, we need more data. If our second measurement is $z = e$, then $\mathcal{L}$ for this is $ab\exp(b-1)\exp(-ae^b)$, so the overall likelihood from both redshifts is

$$\mathcal{L} = p(1)p(e) = a^2b^2\exp\left[b-1-a-ae^b\right].$$

Differentiating wrt a and b gives

$$2a - a^2(1 + e^b) = 0 \Rightarrow a = 2/(1 + e^b);$$

$$2b + b^2(1 - ae^b) = 0 \Rightarrow b = 2\frac{e^b + 1}{e^b - 1}.$$

The second equation requires some numerical experimentation to solve, yielding $b \simeq 2.3994$ and hence $a = 0.1664$.

To get the Hessian matrix, we need to differentiate $\ln \mathcal{L}$ twice. The first derivatives are $\partial\mathcal{L}/\partial a$ divided by $\mathcal{L}$ etc.:

$$\partial \ln \mathcal{L}/\partial a = (2/a) - (1 + e^b);$$

$$\partial \ln \mathcal{L}/\partial b = (2/b) - (ae^b - 1).$$

The second derivatives are now easy:

$$\mathbf{H} = -\begin{pmatrix} 2/a^2 & e^b \\ e^b & 2/b^2 + ae^b \end{pmatrix}.$$

Putting in the numbers for a and b, we get

$$\mathbf{H} = -\begin{pmatrix} 72.2310 & 11.0166 \\ 11.0166 & 2.1806 \end{pmatrix}.$$

The covariance matrix is the inverse of $-\mathbf{H}$:

$$\mathbf{C} = \begin{pmatrix} 0.0603 & -0.3048 \\ -0.3048 & 1.9986 \end{pmatrix}.$$

Thus $\sigma_a = 0.2456$ and $\sigma_b = 1.4137$, and the correlation coefficient is -0.8779. The conditional errors would be from the reciprocal square root of the diagonal components of the negative Hessian: 0.1177 and 0.6772, and these are too small by over a factor 2: the usual problem with strongly correlated errors.