Astronomical Statistics: Tutorial Solutions 1

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

1. What we know is that $p(+|C) = 0.9$ and $p(-|N) = 0.9$, where $+$ and $-$ denote positive and negative tests, C denotes cancer and N denotes no cancer. As a simple consequence, $p(-|C) = 0.1$ and $p(+|N) = 0.1$. What we want is $p(C|+)$, which we get from Bayes' Theorem:
$$p(C|+) = \frac{p(+|C)p(C)}{p(+)} = \frac{p(+|C)p(C)}{p(+|C)p(C) + p(+|N)p(N)},$$
where the last step uses the theorem of total probability. Since we are told that $p(C) = 0.01$ (implying $p(N) = 0.99$), the number we want is
$$p(C|+) = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} = \frac{0.009}{0.009 + 0.099} = 1/12.$$
So the odds are reasonably in favour of you being healthy, even though it might naively sound like your probability of having the disease is the reliability of the test – i.e. $p = 0.9$.

   This result can also be seen in a frequentist way: 1% of the population have cancer, of whom 90% will generate a positive test. The other way of getting a positive test is to be in the 99% who are healthy, of whom 10% will get a false alarm. The ratio of probabilities for these two routes is 0.009/0.099, or 1:11, consistent with the Bayesian argument.

2. Frequentist view: The car will be behind one of the three doors with equal probability $p = 1/3$; whichever door you pick, this is your chance of success. But whether you pick a door with a car or a toy, it is always possible to show you another door containing a toy. Therefore this adds no information about whether your initial choice was a good one, and so your chance of success is still 1/3. But now the only other possibility is that the car may lurk behind the 3rd door – the one that you didn't choose, and which is still unopened. Since probabilities sum to unity, the probability of the car being behind that door must be $1 - 1/3 = 2/3$, and you double your chance of success by switching.

   In more detail, label the door you choose 1. Then the possibilities for door 1,2,3 are (a) CTT, (b) TCT, (c) TTC – all equally probable. For (a), you lose by switching; for (b) and (c) you win by switching. Therefore the chance of winning by switching is 2/3.

   Bayesian view: there are three relevant numbers: S (the door you pick); C (the door with the car); H (the door the host opens). Label the door you pick number 1, and the one the host opens number 3 (we can always choose these labels). So we want

$p(C = 2|S = 1, H = 3)$. By Bayes' Theorem, this is

$$p(C = 2|S = 1, H = 3) = \frac{p(H = 3|C = 2, S = 1)p(C = 2|S = 1)}{\sum_{i=1}^{3} p(H = 3|C = i, S = 1)p(C = i|S = 1)}.$$

We know all the components of this:

$$p(H = 3|C = 1, S = 1) = 1/2;\ p(H = 3|C = 2, S = 1) = 1;\ p(H = 3|C = 3, S = 1) = 0,$$

so the required probability is

$$p(C = 2|S = 1, H = 3) = \frac{1 \times (1/3)}{(1/2) \times (1/3) + 1 \times (1/3) + 0 \times (1/3)} = 2/3.$$

3. If N_T is the number of photons from the source field, and N_B the number of background counts, then our best estimate of the source counts is evidently $\hat{N}_S = N_T - N_B$. The variance of this is the sum of the variances of the rhs terms, $\sigma_S^2 = \sigma_T^2 + \sigma_B^2 = N_T + N_B$, since the photons obey Poisson statistics. Thus the signal to noise is

$$\frac{S}{N} = \frac{N_T - N_B}{N_T + N_B} = \frac{300}{\sqrt{2500}} = 6.$$

With a second background field, detecting N_{B2} photons, the estimate of the background count is $(N_B + N_{B2})/2$, so the estimate of the source counts changes to

$$\hat{N}_S = N_T - (N_B + N_{B2})/2 = 275.$$

The variance is (remembering that the variance of $N_B/2$ is $\sigma_B^2/4$) given by $\sigma_S^2 = N_T + N_B/4 + N_{B2}/4$. Putting in the numbers gives $S/N = 6.21$, so the significance goes up from 6σ to 6.2σ.

4. First imagine numbering the trials from 1 to N; each trial has an outcome, which is a number between 1 and m. If we get n_1 instances of outcome 1, n_2 instances of outcome 2, etc., then this gives a a set of n_1 trial numbers that gave 1, etc. The probability of exactly those trials giving exactly those outcomes is $p_1^{n_1} p_2^{n_2} p_3^{n_3} \cdots$. But we only care about the total number of outcomes, not which trial gave them, so there are many other possible lists of trial numbers, each of which have the same probability; we just need to count how many ways there are of getting this set of n_i outcomes. Start by picking the trial numbers that result in 1: we are choosing n_1 out of N, so there are $N(N-1)(N-2)\cdots(N-n_1+1) = N!/(N-n_1)$ different sequences. But we want to have these listed in order of trial number, since doing trial 2 before trial 1 makes no sense. Thus we divide by $n_1!$ and get the usual $C_{n_1}^N$ expression as the number of ways we could choose our n_1. Now we choose the trials that yield n_2: there are $N - n_1$ left, so the number of ways is $C_{n_2}^{N-n_1}$. Continuing this arguments, the total number of ways of getting our n_1 etc. is

$$C_{n_1}^N C_{n_2}^{N-n_1} C_{n_3}^{N-n_1-n_2} \cdots = \frac{N!}{(N-n_1)!n_1!} \frac{(N-n_1)!}{(N-n_1-n_2)!n_2!} \frac{(N-n_1-n_2)!}{(N-n_1-n_3-n_3)!n_3!} \cdots$$

It should be clear that factorials cancel in the top and bottom of successive terms, so that only the $N!/n_1!n_2!n_3!\cdots$ factor survives.

A more direct route to the same expression is to imagine making a permutation of all the N trial numbers, which we lay out in a line. Now chop the line into pieces of length n_1, n_2 etc., which picks the trials that succeeded in each case. Overall, there are $N!$ permutations; but for each sublist there are $n_i!$ permutations, only one of which has the trials ordered. Therefore the number of distinct ways of partitioning the trials, but having them ordered is $N!$ divided by all the $n_i!$ factors.

5. It is convenient to introduce a variable y, which is a zero-mean Gaussian of unit variance: $\langle y\rangle = 0$; $\langle y^2\rangle = 1$; $\langle y^3\rangle = 0$; $\langle y^4\rangle = 3$. The latter two results are obtained (a) by symmetry and (b) by integration by parts. Now properties of the variable x can be obtained by writing $x = \mu + \sigma y$.

   (i) $\langle x\rangle = \mu + \langle y\rangle = \mu$, using linearity.

   (ii) $\langle x^2\rangle = \langle(\mu^2 + \sigma^2 y^2 + 2\mu\sigma y\rangle = \mu^2 + \sigma^2\langle y^2\rangle + 2\mu\sigma\langle y\rangle = \mu^2 + \sigma^2$.

   (iii) $\langle x^3\rangle = \langle(\mu^3 + 3\mu^2\sigma y + 3\mu\sigma^2 y^2 + \sigma^3 y^3\rangle = \mu^3 + 3\mu\sigma^2$

   (iv) $\langle x^4\rangle = \langle(\mu^4 + 4\mu^3\sigma y + 6\mu^2\sigma^2 y^2 + 4\mu\sigma^3 y^3 + \sigma^4 y^4\rangle = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$.

   The sample variance is $s^2 = (1/N)\sum(x_i - m)^2$, where m, the sample mean, is $m = (1/N)\sum x_i$ – so $s^2 = (1/N)\sum x_i^2 - m^2$, as usual. Thus $\langle s^2\rangle = \langle x^2\rangle - \langle m^2\rangle$. The first term is $\langle x^2\rangle = \mu^2 + \sigma^2$. For the second, write $m^2 = (1/N^2)\sum_i x_i^2 + (1/N^2)\sum_{i\neq j} x_i x_j$, of which the expectation is $(1/N)(\mu^2 + \sigma^2) + (1/N^2)(N^2 - N)\mu^2$. Overall, then,

$$\langle s^2\rangle = \mu^2 + \sigma^2 - (1/N)(\mu^2 + \sigma^2) - (1/N^2)(N^2 - N)\mu^2 = \sigma^2(1 - 1/N).$$

   The same exercise for the sample skewness, $\langle(1/N)\sum(x_i - m)^3\rangle$ looks like it will be more complicated, but fortunately we can make a symmetry argument. Consider inverting all deviations from the mean, i.e. replacing all $x_i - \mu$ by $-x_i + \mu$, which has to be equally likely to the uninverted data, by the symmetry of the pdf for the x_i. But this inverts the sign of $x_i - m$, giving two equal and opposite contributions to $\langle(x_i - m)^3\rangle$, which therefore vanishes by symmetry. Thus both the sample and population skewness are zero for Gaussian data.

6. (i) The distribution $f(x)$ will only be a valid pdf if it is normalized to unity:

$$\begin{aligned}
\int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{+\infty}\phi(x)dx + \alpha\int_{-\infty}^{+\infty}\phi(x)(x^3 - 3x)dx + \beta\int_{-\infty}^{+\infty}\phi(x)(x^4 - 6x^2 + 3)dx \\
&= 1 + \alpha\times 0 + \beta\left((2\times 2 - 1)!!\langle x^2\rangle_\phi^2 - 6\langle x^2\rangle_\phi + 3\right) \\
&= 1 + \beta(3 - 6 + 3) \\
&= 1 \qquad (1)
\end{aligned}$$

   where we used $\phi(x)(x^3 - 3x)$ is an odd function integrated over an even space, hence its integral over that space is zero and that we are dealing with a unit-variance Gaussian, with zero mean, i.e. $\langle x\rangle_\phi = 0$ and $\langle x^2\rangle_\phi = 1$. The required condition therefore holds for any values of the two constants.

   However, the constants α and β must be such that $\alpha(x^3 - 3x) + \beta(x^4 - 6x^2 + 3) \geq -1$ for all x, in order to avoid unphysical negative values of $f(x)$.

   (ii) The mean is given by:

$$\begin{aligned}
\langle x\rangle_f &= \int_{-\infty}^{+\infty} xf(x)dx \\
&= \int_{-\infty}^{+\infty} x\phi(x)dx + \alpha\int_{-\infty}^{+\infty} x\phi(x)(x^3 - 3x)dx + \beta\int_{-\infty}^{+\infty} x\phi(x)(x^4 - 6x^2 + 3)dx \\
&= 0 + \alpha\int_{-\infty}^{+\infty}\phi(x)(x^4 - 3x^2)dx + \beta\times 0 \\
&= \alpha\left((2\times 2 - 1)!!\langle x^2\rangle_\phi^2 - 3\langle x^2\rangle_\phi\right) = 0
\end{aligned}$$

and the variance is given by:

$$\begin{aligned}
s^2 &= \langle (x - \langle x\rangle_f)^2\rangle_f = \langle x^2\rangle_f \\
&= \int_{-\infty}^{+\infty} x^2 f(x)dx \\
&= \int_{-\infty}^{+\infty} x^2\phi(x)dx + \alpha\int_{-\infty}^{+\infty} x^2\phi(x)(x^3-3x)dx + \beta\int_{-\infty}^{+\infty} x^2\phi(x)(x^4-6x^2+3)dx \\
&= 1 + \alpha\times 0 + \beta\int_{-\infty}^{+\infty}\phi(x)(x^6-6x^4+3x^2)dx \\
&= 1 + \beta\left((3\times 2-1)!!\langle x^2\rangle_\phi^3 - 6(2\times 2-1)!!\langle x^2\rangle_\phi^2 + 3\langle x^2\rangle_\phi\right) \\
&= 1 + \beta(15-18+3) = 1
\end{aligned}$$

(iii) The skewness is given by:

$$\begin{aligned}
\langle x^3\rangle_f &= \int_{-\infty}^{+\infty} x^3 f(x)dx \\
&= \int_{-\infty}^{+\infty} x^3\phi(x)dx + \alpha\int_{-\infty}^{+\infty} x^3\phi(x)(x^3-3x)dx + \beta\int_{-\infty}^{+\infty} x^3\phi(x)(x^4-6x^2+3)dx \\
&= 0 + \alpha\int_{-\infty}^{+\infty}\phi(x)(x^6-3x^4)dx + \beta * 0 \\
&= 0 + \alpha\left((3\times 2-1)!!\langle x^2\rangle_\phi^3 - 3(2\times 2-1)!!\langle x^2\rangle_\phi^2\right) = 6\alpha
\end{aligned}$$

and the kurtosis is estimated in a similar way:

$$\begin{aligned}
\langle x^4\rangle_f &= \int_{-\infty}^{+\infty} x^4 f(x)dx \\
&= \int_{-\infty}^{+\infty} x^4\phi(x)dx + \alpha\int_{-\infty}^{+\infty} x^4\phi(x)(x^3-3x)dx + \beta\int_{-\infty}^{+\infty} x^4\phi(x)(x^4-6x^2+3)dx \\
&= (2\times 2-1)!!\langle x^2\rangle_\phi^2 + \alpha\times 0 + \beta\int_{-\infty}^{+\infty}\phi(x)(x^8-6x^6+3x^4)dx \\
&= 3 + \beta\left((4\times 2-1)!!\langle x^2\rangle_\phi^4 - 6(3\times 2-1)!!\langle x^2\rangle_\phi^3 + 3(2\times 2-1)!!\langle x^2\rangle_\phi^2\right) \\
&= 3 + \beta(105-90+9) = 3+24\beta
\end{aligned}$$

Note that this question would have been more complicated if the mean m ($= \langle x\rangle_f$) was non-zero. In general, the variance is $\langle (x-m)^2\rangle = \langle x^2\rangle - m^2$; the skewness is $\langle (x-m)^3\rangle = \langle x^3\rangle + 2m^2 - 3m\langle x^2\rangle$; and the kurtosis is $\langle (x-m)^4\rangle = \langle x^4\rangle - 4m\langle x^3\rangle + 6m^2\langle x^2\rangle - 3m^4$.

7. Although this is a 2D pdf, expectation values of 1D variables like x are well-defined, but we need to integrate over both x and y. Start with $\langle x\rangle$:

$$\langle x\rangle = \iint x\; A\exp[-(ax^2+by^2+cxy)/2]\;dx\,dy,$$

where A is a normalization constant. As is common with Gaussian problems, we have to complete the square in the exponential: $by^2+cxy = b(y+cx/2b)^2 - c^2x^2/4b = bz^2 - c^2x^2/4b$, where $z \equiv y + cx/2a$. The Jacobian between (x,z) and (x,y) is a constant, so we can immediately integrate over z, leaving

$$\langle x\rangle \propto \int x\;\exp[-(ax^2 - c^2x^2/4b)/2]\;dx,$$

which vanishes through symmetry (this wasn't obvious in the initial 2D expression). Similarly, $\langle y \rangle = 0$.

For the variances and covariance, we need $\langle x^2 \rangle$, $\langle y^2 \rangle$, and $\langle xy \rangle$. The first of these follows the same line as above:

$$\langle x^2 \rangle = \frac{\int x^2 \ \exp[-(ax^2 - c^2x^2/4b)/2] \, dx}{\int \exp[-(ax^2 - c^2x^2/4b)/2] \, dx}.$$

The integral on the top line is done by parts, yielding

$$\langle x^2 \rangle = (a - c^2/4b)^{-1},$$

from which it is obvious that

$$\langle y^2 \rangle = (b - c^2/4a)^{-1}.$$

To get $\langle xy \rangle$, it may like we have more work to do, but we can save this by considering $\langle xz \rangle$: this must vanish because the z dependence is odd. But from the definition of z, $\langle xz \rangle = \langle xy \rangle + c\langle x^2 \rangle / 2a$. Hence $\langle xy \rangle = -c\langle x^2 \rangle / 2a$. So finally the correlation coefficient is

$$r \equiv \frac{\langle xy \rangle}{(\langle x^2 \rangle \langle y^2 \rangle)^{1/2}} = -\frac{c}{2(ab)^{1/2}}.$$

All these relations can be obtained much more simply by the general expression for a zero-mean Gaussian, in which the pdf is proportional to

$$\exp[-(x, y) \cdot \mathbf{C^{-1}} \cdot (x, y)/2] = \exp[-(x^2/\sigma_x^2 + y^2/\sigma_y^2 - 2rxy/\sigma_x\sigma_y)/2(1 - r^2)],$$

where $\mathbf{C}$ is the covariance matrix.

To deal with the rotation to new coordinates, a useful shorthand is to write the rotation using $C \equiv \cos\theta$ and $S \equiv \sin\theta$. Then we have

$$ax^2 + by^2 + cxy = a(C^2X^2 + S^2Y^2 - 2SCXY) + b(C^2Y^2 + S^2X^2 + 2SCXY) + c(C^2XY - S^2XY).$$

We can eliminate the cross term in XY if

$$c(C^2 - S^2) + 2SC(b - a) = 0 \Rightarrow \tan 2\theta = c/(b - a),$$

so now

$$ax^2 + by^2 + cxy = X^2(aC^2 + bS^2) + Y^2(aS^2 + bC^2) \equiv X^2/\sigma_X^2 + Y^2/\sigma_Y^2$$

and the pdf factorises into independent Gaussians in X and Y, whose variances can be read from the formula. In making this statement, we need to be clear that we are using the fact that the Jacobian between (x, y) and (X, Y) is unity.