Astronomical Statistics: Tutorial Questions 1

John Peacock

C20, Royal Observatory; jap@roe.ac.uk

October 2012

1. A test for cancer is known to be 90% accurate either in detecting cancer if present or in giving an all-clear if cancer is absent. The prevalence of cancer in the population is 1% – how worried should you be if you test positive? (hint: apply Bayes' Theorem).

2. A game-show host invites you to choose one of three doors for a chance to win a car (behind one door) or a cuddly toy (behind the other two). You choose a door, but the host opens a different door to reveal a toy – and asks you if you want to switch your choice to the other unopened door. Does switching improve your chance of winning the car? Try to justify your answer using both frequentist and Bayesian reasoning. (Hint: from the frequentist point of view, the three permutations CTT, TCT, TTC are equally likely).

3. A detector receives 1400 photons from region of sky containing a possible source. 1100 photons are received from a blank piece of sky of the same area, in the same exposure time. What is the significance of the detection of the source? (state your assumptions, and list possible complicating effects that cannot be addressed with this information). Another observer then detects 1150 photons from another blank piece of sky of the same area in the same time. Does the significance of the detection go up or down?

4. For a binomial distribution, the probability of n successes out of N trials with probability of success p is
$$P_n = C_n^N p^n (1-p)^{N-n}, \quad (0 \le n \le N), \tag{1}$$
where
$$C_n^N = \frac{N!}{n!(N-n)!}. \tag{2}$$
This can be generalized to the multinomial distribution, in which there are possible outcomes 1 to m, with probabilities p_1 to p_m. The probability of obtaining n_1 cases of outcome 1 through to n_m cases of outcome m is
$$p(n_1 \dots n_m) = \frac{N!}{n_1! \, n_2! \, \dots \, n_m!} \, p_1^{n_1} p_2^{n_2} \dots p_m^{n_m}. \tag{3}$$
Give two separate arguments to justify the form of this result.

[PTO]

5. For a Gaussian-distributed variable, x, with mean μ and rms σ, evaluate (i) $\langle x \rangle$; (ii) $\langle x^2 \rangle$; (iii) $\langle x^3 \rangle$; (iv) $\langle x^4 \rangle$. (hint: use the properties of a Gaussian centred at zero).

   Taking N independent samples, x_i, verify that the sample variance of the data has an expectation value of $\sigma^2(1 - 1/N)$. What is the expectation of $\sum_i (x_i - m)^3/N$, where m is the sample mean?

6. Consider the probability density function

$$f(x) = \phi(x)(1 + \alpha(x^3 - 3x) + \beta(x^4 - 6x^2 + 3)), \qquad (4)$$

   where ϕ is the unit-variance Gaussian, $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.

   (i) What conditions must the constants α and β satisfy in order for this to be a valid pdf?

   (ii) Calculate the mean, m, and variance, s^2, of this distribution.

   (iii) Calculate the skewness, $\langle (x-m)^3 \rangle$, and kurtosis, $\langle (x-m)^4 \rangle$, of this distribution.

   Hints:
   a. You may use the result that for a Gaussian variable:
   $\langle x^{2n} \rangle = (2n-1)!!\langle x^2 \rangle^n$
   where
   $(2n-1)!! = (2n-1).(2n-3).(2n-5)\cdots 3.1.$
   b. If hint (a) is not a known fact, it is worth showing it by recurrence.
   c. Exploit symmetries in the integrals.

7. Consider the 2-dimensional Gaussian pdf

$$p(x,y) \propto \exp[-(ax^2 + by^2 + cxy)/2]. \qquad (5)$$

   Evaluate the variances in x and y, together with the correlation coefficient between x and y.

   New variables are defined as linear combinations corresponding to a rotation in the $x - y$ plane: $x = (\cos\theta)X - (\sin\theta)Y$; $y = (\cos\theta)Y + (\sin\theta)X$. Show that, for a suitable choice of angle, the variables X and Y are independent. What are the variances in X and Y?