DATA, MODELS & UNCERTAINTY IN THE NATURAL SCIENCES

Problem Set 1

1. We have introduced $p(x)$ and $P(x)$, the *probability density* and *distribution functions* of a *random variable* x, and discussed their properties. With these we have defined the *expected value* of a function of this variable, $g(x)$, to be

$$\langle g(x) \rangle = \int_{-\infty}^{+\infty} g(x)\, p(x)\, dx. \tag{1}$$

Prove, disprove and/or complete the following properties of the *expectation operator* $\langle \cdot \rangle$:

(a) $\langle a + g_1(x) + g_2(x) \rangle = a + \langle g_1(x) \rangle + \langle g_2(x) \rangle$

(b) $\langle a g(x) \rangle = a \langle g(x) \rangle$

(c) $\langle x - \langle x \rangle \rangle = \, ?$

(d) $\langle ax + b \rangle = \, ?$

for arbitrary constants a, b. Motivate your answers summarily.

2. We have introduced $p(x, y)$ and $P(x, y)$, the *joint probability density* and *distribution functions* of two *random variables* x and y, and their properties. With this we have defined the *expected value* of a function of these variables, $g(x, y)$, to be

$$\langle g(x, y) \rangle = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y)\, p(x, y)\, dx\, dy. \tag{2}$$

(a) Write a brief review of the properties of $p(x, y)$ and $P(x, y)$, and discuss the special case where x and y are *statistically independent*.

(b) Show or refute that $\langle xy \rangle = \langle x \rangle \langle y \rangle$ for independent random variables x and y.

(c) Under which conditions is $\langle a + g_1(x) + g_2(y) \rangle = a + \langle g_1(x) \rangle + \langle g_2(y) \rangle$?

As before, the subscripted g's are some functions of x and a is some constant.

3. We have introduced the *conditional probability density function* $p(x|y)$ as the probability density function of x *given* y, and defined it from the joint probability density function $p(x, y)$ of x *and* y, and the (marginal) probability density functions $p(x)$ and $p(y)$ of x *and* y *individually* as follows:

$$p(x, y) = p(x|y)\, p(y) = p(y|x)\, p(x). \tag{3}$$

If you read this out loud from left to right I think this makes most intuitive sense... but usually, this equality is expressed slightly more opaquely in the form of ***Bayes' theorem***:

$$p(x|y) = \frac{p(y|x)\, p(x)}{p(y)}. \tag{4}$$

Both equations 3 and 4 hold when the probability density functions p are replaced by the probability distribution functions P. Now, please, answer the questions below — no geological knowledge is required.

A fragment of a hitherto unknown dinosaur species has been found in a stream bed in New Jersey, and a paleontologist wants to go out and search for more remains. Unfortunately, the source of the fragment cannot be identified with certainty, because it was found below the junction of two dry stream tributaries. Still, the fossil must have come from either of their two drainage basins, B_1 or B_2. We know the area of $B_1 = 18$ km^2 while the area of $B_2 = 10$ km^2. The fossil found was a marine specimen; historically, 35% of the many fossils found in basin B_1 have been marine, compared to 80% of those ever found in B_2. The paleontologist, on a budget, wants to sample only one basin: obviously, he wants to go to the one that has the highest probability of being the source of the fossil fragment.

(a) What are the values of $P(B_1)$, the overall *a priori* probability that a fossil found in the stream bed originates in the first basin, and of $P(B_2)$, in the second? If you need to make assumptions, argue them convincingly. What is their sum, and why?

(b) What are the values of the *conditional* probabilities $P(M|B_1)$ that, given that a fossil comes from basin B_1, it is marine, and $P(M|B_2)$, that a fossil from B_2 is marine? What is their sum, and why?

(c) Write the expressions for the conditional probabilities $P(B_1|M)$, that, given that the fossil fragment found is marine, it comes from B_1; and $P(B_2|M)$, that it comes from B_2, instead. What is their sum, and why?

(d) Write the expression for $P(M)$, the overall probability that a fossil found in either of the two basins is marine. Evaluate this based on what you have so far.

(e) Finally, evaluate the chances that this particular marine fossil comes from basin B_1, or rather from B_2. Where should the search party go look for more remains?

(f) What is the probability that the fossil from which the original fragment was derived belongs to a *mammal*? Motivate your answer statistically.

4. *[Optional]* What is Cheryl's birthday, and why?