MA3466 Tutorial Sheet 1, outline solutions[1]

16 Febuary 2010

1. For two random variables with numerical outcomes, find $p(x,y)$ so that there is zero correllation

$$\langle x'y'\rangle = 0 \tag{1}$$

but X and Y aren't independent. $x' = x - EX$ and $y' = y - EY$.

Solution:

I tried this first for the case where $|\mathcal{X}| = |\mathcal{Y}| = 2$ and found in this case $C = 0$ implies $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. In fact, a simple parameter counting might be enough here, but a full argument goes like, consider $p(x,y)$ given by

$$\begin{array}{c|cc} & 0 & 1 \\ \hline 0 & a & b \\ 1 & c & d \end{array} \tag{2}$$

with X across and Y down. Now, working out the marginal distributions we have $p_X(0) = a + c$, $p_X(1) = b + d$ so that $\langle X\rangle = b + d$; $p_Y(0) = a + b$ and $p_Y(1) = c + d$ so $\langle Y\rangle = c + d$. Substituting into the formula for C we get

$$\begin{aligned} C &= a(-b-d)(-c-d) + b(1-b-d)(-c-d) \\ &\quad +c(-b-d)(1-c-d) + d(1-b-d)(1-c-d) \\ &= d - (b+d)(c+d) \end{aligned} \tag{3}$$

so $C = 0$ means $d = (b + d)(c + d)$. Of course, since these are probabilities, we also have $a + b + c + d = 1$. Now, this probability table corresponds to independent distributions, $p(x,y) = p(x)p(y)$, if the columns of the table are proportional to each other, since $p(x, y = 0) = p(x)p(y = 0)$ and $p(x, y = 1) = p(x)p(y = 1) = p(x, y = 0)p(y = 1)/p(y = 0)$. This implies $ac = bd$, substituting $a = 1 - b - c - d$ into this equation gives $d = (c + d)(b + d)$, precisely the condition for $C = 0$, zero correlation implies statistical independence for two by two probability tables.

Next, lets set $|\mathcal{X}| = 2$ and $|\mathcal{Y}| = 3$. Rather than look at the general case, consider

$$\begin{array}{c|cc} & -1 & 1 \\ \hline -1 & a & c \\ 0 & b & d \\ 1 & a & c \end{array} \tag{4}$$

[1] Conor Houghton, `houghton@maths.tcd.ie`, see also `http://www.maths.tcd.ie/~houghton/MA3466`

It is easy to check that this already has $C = 0$, but the two columns are not proportional provided $ad \neq bc$, showing that there are distributions with zero correlation which are statistically independent. As a concrete example take

$$\begin{array}{c|cc} & -1 & 1 \\ \hline -1 & 1/8 & 1/4 \\ 0 & 1/4 & 0 \\ 1 & 1/8 & 1/4 \end{array} \tag{5}$$

2. Work out the marginal distributions and the $x = a$ conditional distribution for

$$\begin{array}{c|cc} & a & b \\ \hline 1 & \frac{1}{3} & \frac{1}{6} \\ 2 & 0 & \frac{1}{4} \\ 3 & \frac{1}{8} & \frac{1}{8} \end{array}$$

*Solution:*Calculating the marginal distribution just requires adding along the rows or columns, if we take X to be the random variable going across, so $\mathcal{X} = \{a, b\}$ and Y the random variable going down, so $\mathcal{Y} = \{1, 2, 3\}$ then the two marginal distributions are, for X

$$\begin{array}{c|cc} & a & b \\ \hline & \frac{11}{24} & \frac{13}{24} \end{array}$$

and for Y

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{array}$$

The conditional distribution, conditioned on $x = a$ is calculated using $p(y|x)p(x) = p(x, y)$ so we divide the $p(x = a, y)$ column by $p(x = a) = 11/24$, hence

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline & \frac{8}{11} & 0 & \frac{3}{11} \end{array}$$

3. (C&T 2.1) A fair coin is flipped until the first head occurs. Let X denote the number of flips required.

   (a) Find the entropy $H(X)$ in bits. The following expressions may be useful:

$$\begin{aligned} \sum_{n=0}^{\infty} r^n &= \frac{1}{1-r} \\ \sum_{n=0}^{\infty} n r^n &= \frac{r}{(1-r)^2} \end{aligned} \tag{6}$$

(b) A random variable X is drawn according to this distribution. Find an efficient sequence of yes-no questions of the form, 'Is X contained in the set S?'. Compare $H(X)$ to the expected number of questions required to determine X.

*Solution:*So, here the set of possible outcomes is $\mathcal{X} = \{1, 2, 3, \ldots\}$ and we need to start by working out $p_X n$ the chance of throwing n flips before getting a head. To get $X = n$ you need to throw $n - 1$ tails, this has probability $1/2^{n-1}$ followed by a head, which has probability $1/2$, hence

$$p_X(n) = \frac{1}{2^n} \tag{7}$$

It is easy to check that

$$\sum_{n=1}^{\infty} p_X(n) = \sum_{n=1}^{\infty} \frac{1}{2^n} = \sum_{n=0}^{\infty} \frac{1}{2^n} - 1 = 1 \tag{8}$$

Now, to calculate the entropy, just use the formula:

$$\begin{aligned} H(X) &= -\sum_{n\in\mathcal{X}} p_X(x) \log p_X(x) \\ &= -\sum_{n=1}^{\infty} \frac{1}{2^n} \log 2^{-n} = \sum_{n=1}^{\infty} \frac{n}{2^n} \\ &= 2 \end{aligned} \tag{9}$$

It is easy to see that this is the same as the average number of questions asked in order starting with $n = 1$ of the form 'is the answer n?' it would take to find X .

4. (C&T 2.3) What is the minimum value of $H(p_1, p_2, \ldots, p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges overall possible vectors. Find the $\mathbf{p}$ which achieve this bound. $H(p_1, p_2, \ldots, p_n)$ is a common notation for $H(X)$ where X has n-outcomes $\{x_1, x_2, \ldots, x_n\}$ and $p_1 = p(x_1)$, $p_2 = p(x_2)$ and so on.

   *Solution:*First, let's extremize

$$H = -\sum p_i \log p_i \tag{10}$$

subject to the contraint $\sum p_i = 1$: we will ignore the inequality constraints $p_i \geq 0$ for now. Using a Lagrange multiplier, we define

$$L = -\sum p_i \log p_i + \lambda \left(\sum p_i - 1\right) \tag{11}$$

and so, rembering that we have logs to the base two

$$\frac{\partial L}{\partial p_i} = -\log p_i - \log_2 e + \lambda \tag{12}$$

Setting $\partial L/\partial p_i = 0$ shows the p_i are all the same and hence equal to $1/n$, solving for λ gives the same answer. Hence, an extreme value of H is given by $H = \log n$, this, however, is the maximum. The minima don't have $\partial L/\partial p_i = 0$, rather they lie in the corners of the inequality constraint region, by linear programming, or by examining H it is easy to see $H = 0$ if and only if $p_i = 1$ for some i and $p_j = 0$ for $i \neq j$; since $H \geq 0$ these are the only minima, and there are n of them.