Stat 110 Homework 8, Fall 2013

Due: Friday 11/8 at 1:10 pm, in the Stat 110 dropbox outside of SC 105. No late homework will be accepted. Please write your name, and staple your homework. Show your work and make sure to give clear, careful, convincing explanations.

Reading: Sections 7.1-7.4, 8.1-8.2

1. With k people, under the usual assumptions from the birthday problem, find the correlation between how many were born on May 1 and how many were born on May 2.

2. A certain college has m freshmen, m sophomores, m juniors, and m seniors. A certain class there is a simple random sample of size n students, i.e., all sets of n of the $4m$ students are equally likely. Let $X_1, \dots, X_4$ be the numbers of freshmen, $\dots$, seniors in the class.

(a) Find the joint PMF of X_1, X_2, X_3, X_4.

(b) Give both an intuitive explanation and a mathematical justification for whether or not the distribution from (a) is Multinomial.

(c) Find $\text{Cov}(X_1, X_3)$, fully simplified.
Hint: take the variance of both sides of $X_1 + X_2 + X_3 + X_4 = n$.

3. A beautiful visual interpretation of covariance is at `http://goo.gl/F3xaEZ` (please read the post there by whuber, with the red and blue rectangles). This problem explores the interpretation given in that post. Data are collected for $n \geq 2$ individuals, where for each individual two variables are measured (e.g., height and weight).

Let $(x_1, y_1), \dots, (x_n, y_n)$ be the n data points, where we assume independence *across* individuals (e.g., (x_1, y_1) gives no information about (x_2, y_2)) but not *within* an individual data point (x_i and y_i may be dependent). The data are considered here as fixed, known numbers – they are the observed values after performing an experiment.

(a) The *sample covariance* of the data is defined to be

$$r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ are the *sample means*. (It is traditional to divide by $n - 1$ rather than n in the definition of sample covariance because of a property called *unbiasedness*, but that is better left for Stat 111 and need not concern us in this problem.)

Let (X, Y) be one of the (x_i, y_i) pairs, chosen uniformly at random. Determine precisely how $\text{Cov}(X, Y)$ is related to the sample covariance.

(b) Let (X, Y) be as in (a), and $(\tilde{X}, \tilde{Y})$ be an independent draw from the same distribution. Color rectangles red and blue as described in the post linked above. Express the net amount of red in the plot (treating blue as negative values) as a constant times the expected value of a simple function of $X, \tilde{X}, Y, \tilde{Y}$. Then use this to prove that the sample covariance of the data is a constant times the net amount of red in the plot, as suggested by the post.

(c) Based on the interpretation from the post, give intuitive explanations of why for any r.v.s W_1, W_2, W_3 and constants a_1, a_2, covariance has the following properties:

(i) $\text{Cov}(W_1, W_2) = \text{Cov}(W_2, W_1)$;

(ii) $\text{Cov}(a_1 W_1, a_2 W_2) = a_1 a_2 \text{Cov}(W_1, W_2)$;

(iii) $\text{Cov}(W_1 + a_1, W_2 + a_2) = \text{Cov}(W_1, W_2)$;

(iv) $\text{Cov}(W_1, W_2 + W_3) = \text{Cov}(W_1, W_2) + \text{Cov}(W_1, W_3)$.

4. A random point is chosen uniformly in the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$. Let R be its distance from the origin.

(a) Find $E(R)$ using 2D LOTUS.

Hint: to do the integral, convert to polar coordinates (see the math appendix).

(b) Find the CDFs of R^2 and of R *without using calculus*, using the fact that for a Uniform distribution on a region, probability within that region is proportional to area. Then get the PDFs of R^2 and of R, and find $E(R)$ in two more ways: using the definition of expectation, and using a 1D LOTUS by thinking of R as a function of R^2.

5. Let X and Y be i.i.d. $\mathcal{N}(0, 1)$ r.v.s. Find the distribution of $\arctan(Y/X)$.

6. Let X and Y be i.i.d. $\mathcal{N}(0, 1)$ r.v.s, and (R, θ) be the polar coordinates for the point (X, Y). Find the joint PDF of R^2 and θ. Also find the marginal distributions of R^2 and θ, giving their names (and parameters) if they are distributions we have studied before.