Statistics I, Fall 2012
Suggested Solution for Homework 05

Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. (a) First, we need to calculate μ_X and μ_Y. Straightforward calculations lead to $\mu_X = 2.3$ and $\mu_Y = 2.9$. Then for each pair of (x, y), we may calculate $(x - \mu_X)(y - \mu_y$. For example, for $(x, y) = (2, 1)$, $(2 - 2.3)(1 - 2.9) = 0.57$. This quantity should then be weighted based on its probability, $\Pr(X = 2, Y = 1) = 0.4$. This is the first entry (in the intersection of column "1" and row "2") of Table 1. We may repeat the process for all the six pairs and then sum all the six values up to obtain the covariance $\text{Cov}(XY) = 0.33$. As we may find the standard deviations $\sigma_X = 0.21$ and $\sigma_Y = 3.09$, the correlation coefficient is $\frac{0.33}{0.21 \times 3.09} \approx 0.509$. We may say that these two random variables have a moderately strong correlation.

x	y			Total
	1	4	5	
2	0.228	−0.066	−0.063	0.099
3	−0.0665	0.077	0.2205	0.231
Total	0.1615	0.011	0.1575	0.330

Table 1: Calculations for Problem 1a.

w	z			Total
	1	4	5	
2	0.20	−0.02	−0.18	0.00
3	−0.15	0.06	0.09	0.00
Total	0.05	0.04	−0.09	0.00

Table 2: Calculations for Problem 1c.

(b) X and Y are not independent because $\Pr(X = x, Y = y) \neq \Pr(X = x)\Pr(Y = y)$ for all possible x and y. For example, $\Pr(X = 2, Y = 1) = 0.4$, but $\Pr(X = 2)\Pr(Y = 1) = 0.7 \times 0.45 = 0.315$.

(c) We may follow the same procedure used in Part (a) to solve this problem. The means are $\mu_W = 2.4$ and $\mu_Z = 3.5$. The relevant numbers are recorded in Table 2. Both the covariance and the correlation coefficient are zero. The two random variables have no correlation.

(d) W and Z are not independent because $\Pr(W = w, Z = z) \neq \Pr(W = w)\Pr(Z = z)$ for all possible w and z. For example, $\Pr(W = 2, Z = 1) = 0.2$, but $\Pr(X = 2)\Pr(Y = 1) = 0.6 \times 0.3 = 0.18$.

Note. A zero correlation just means "no correlation", which means "no *linear* relationship". It does not imply independence! In general, independence implies zero correlation but the opposite is not true (as shown in Part (c)).

2. (a) Imagine that there are N ball while A are white and $N - A$ are black and the experiment is to draw n balls randomly. When $n = N$, all the balls will be drawn and the number of white balls must be A. In other words, there is no uncertainty in this experiment. The variance is thus zero.

(b) When $n = 1$, there is no difference between sampling with and without replacement because we do not do the second trial. Therefore, their variances are the same.

3. Let X be a hypergeometric random variable with population size N, number of "1"s A, and the sample size n. Let $p = \frac{A}{N}$. In this problem, we will derive the mean and variance of X.

(a) We have

$$\begin{aligned}\Pr(X_2 = 1) &= \Pr(X_2 = 1 | X_1 = 1)\Pr(X_1 = 1) + \Pr(X_2 = 1 | X_1 = 0)\Pr(X_1 = 0) \\ &= \left(\frac{A-1}{N-1}\right)\left(\frac{A}{N}\right) + \left(\frac{A}{N-1}\right)\left(\frac{N-A}{N}\right) = \frac{A}{N} = p.\end{aligned}$$

(b) Note that each trial is a Bernoulli trial with parameter p (they are identical trials but they are not independent). Therefore, following those results for the Bernoulli distribution, we have $\mathbb{E}[X_i] = p$ and $\mathrm{Var}(X_i) = p(1-p)$.

(c) Because the expectation of a linear function is separable, we have

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np.$$

Note that the first equality holds no matter X_is are independent or not.

(d) By definition, we have

$$\begin{aligned}\mathrm{Cov}(X+Y) &= \mathbb{E}\Big[\big[(X+Y) - (\mu_X + \mu_Y)\big]^2\Big] = \mathbb{E}\Big[\big[(X-\mu_X) + (Y-\mu_Y)\big]^2\Big] \\ &= \mathbb{E}\Big[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)\Big] \\ &= \mathbb{E}\Big[(X-\mu_X)^2\Big] + \mathbb{E}\Big[(Y-\mu_Y)^2\Big] + 2\mathbb{E}\Big[(X-\mu_X)(Y-\mu_Y)\Big] \\ &= \mathrm{X} + \mathrm{Y} + 2\mathrm{Cov}(X,Y).\end{aligned}$$

(e) Using the general formula described in Part (d) and the fact that $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_1, X_2)$ for all $i \neq j$, we have

$$\begin{aligned}\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{Cov}(X_i, X_j) \\ &= np(1-p) + n(n-1)\mathrm{Cov}(X_1, X_2).\end{aligned}$$

(f) Recall that

$$\mathrm{Var}(X) = np(1-p) + n(n-1)\mathrm{Cov}(X_1, X_2) \tag{1}$$

as we derived above and this must be true for the special case $n = N$. By plugging $n = N$ into (1), we have $0 = Np(1-p) + N(N-1)\mathrm{Cov}(X_1, X_2)$, which implies $\mathrm{Cov}(X_1, X_2) = -\frac{p(1-p)}{N-1}$. We may now do a substitution in (1) and get

$$\begin{aligned}\mathrm{Var}(X) &= np(1-p) + n(n-1)\mathrm{Cov}(X_1, X_2) = np(1-p) - n(n-1)\frac{p(1-p)}{N-1} \\ &= np(1-p)\left(1 - \frac{n-1}{N-1}\right) = np(1-p)\left(\frac{N-n}{N-1}\right).\end{aligned}$$

4. Because $X \sim \mathrm{Uni}(a, b)$, the pdf is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}.$$

(a) The mean is

$$\mathbb{E}[X] = \int_a^b x\left(\frac{1}{b-a}\right)dx = \frac{1}{b-a}\left(\frac{1}{2}x^2\right)\Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

(b) The variance is

$$\begin{aligned}\mathrm{Var}(X) &= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \left(\frac{1}{b-a}\right)dx \\ &= \frac{1}{b-a}\int_a^b \left[x^2 - (a+b)x + \left(\frac{a+b}{4}\right)^2\right]dx \\ &= \frac{1}{b-a}\left(\frac{1}{3}x^3 - \frac{a+b}{2}x^2 + \frac{a+b}{4}x\right)\Big|_a^b \\ &= \frac{1}{b-a}\left[\frac{1}{3}(b^3 - a^3) - \frac{a+b}{2}(b^2 - a^2) + \left(\frac{a+b}{4}\right)(b-a)\right] \\ &= \frac{1}{3}(a^2 + ab + b^2) - \frac{1}{2}(a^2 + 2ab + b^2) + \frac{1}{4}(a+b) = \frac{(b-a)^2}{12}.\end{aligned}$$

5. (a) Because ke^2 is a pdf over $[1, 2]$, it must satisfy

$$\int_1^2 ke^{2x}dx = k\int_1^2 e^{2x}dx = \frac{k}{2}e^{2x}\Big|_1^2 = \frac{k}{2}\left(e^4 - e^2\right) = 1,$$

which means $k = \dfrac{2}{e^4 - e^2}$.

(b) The mean is

$$\begin{aligned}\mathbb{E}[X] &= \int_1^2 xke^{2x}dx = k\int_1^2 xe^{2x}dx = k\left[\frac{1}{2}xe^{2x}\Big|_1^2 - \int_1^2 \frac{1}{2}e^{2x}dx\right]\\ &= \frac{k}{2}\left[2e^4 - e^2 - \frac{e^4 - e^2}{2}\right] = \frac{k}{2}\left(\frac{3}{2}e^4 - \frac{1}{2}e^2\right) = \frac{k}{4}\left(3e^4 - e^2\right),\end{aligned}$$

where in the third equality we apply integration by parts. Now, because $k = \dfrac{2}{e^4 - e^2}$, we have

$$\mathbb{E}[X] = \frac{3e^4 - e^2}{2\left(e^4 - e^2\right)} = \frac{3e^2 - 1}{2\left(e^2 - 1\right)}.$$

6. (a) Because $f(x)$ is a pdf over $[0, 4]$, it must satisfy

$$\begin{aligned}\int_0^1 f(x)dx &= k\left(\int_0^1 \frac{x}{4}dx + \int_1^4 \frac{4 - x}{12}dx\right) = k\left(\frac{1}{8}x^2\Big|_0^1 + \frac{1}{12}\left(4x - \frac{1}{2}x^2\right)\Big|_1^4\right)\\ &= k\left(\frac{1}{8} + \frac{12 - \frac{15}{2}}{12}\right) = \frac{k}{2} = 1,\end{aligned}$$

which means $k = 2$.

(b) The mean is

$$\begin{aligned}\int_0^1 xf(x)dx &= \int_0^1 \frac{x^2}{2}dx + \int_1^4 \frac{4x - x^2}{6}dx = \frac{1}{6}x^3\Big|_0^1 + \frac{1}{6}\left(2x^2 - \frac{1}{3}x^3\right)\Big|_1^4\\ &= \frac{1}{6} + \frac{30 - 21}{6} = \frac{5}{3}.\end{aligned}$$

(c) First, consider the case that $x \leq 1$. In this case,

$$\int_0^x f(y)dy = \int_0^x \frac{y}{2}dy = \frac{x^2}{4}.$$

If $x > 1$, we need to again split the integration into two parts:

$$\begin{aligned}\int_0^x f(y)dy &= \int_0^1 \frac{y}{2}dy + \int_1^x \frac{4 - y}{6}dy = \frac{1}{4} + \frac{1}{6}\left(4y - \frac{1}{2}y^2\right)\Big|_1^x\\ &= \frac{1}{4} + \frac{1}{6}\left(4x - 4 - \frac{1}{2}x^2 + \frac{1}{2}\right) = -\frac{1}{12}x^2 + \frac{2}{3}x - \frac{1}{3}.\end{aligned}$$

Combining the above two cases, the cdf is

$$F(x) = \begin{cases} \frac{x^2}{4} & \text{if } x \in [0, 1] \\ -\frac{1}{12}x^2 + \frac{2}{3}x - \frac{1}{3} & \text{if } x \in [1, 4] \end{cases}.$$

7. (a) $\Pr(X \leq 3) = \sum_{x=0}^{3} \binom{20}{x}(0.1)^x(0.9)^{20-x} \approx 0.867$.

(b) The Poisson random variable that approximates X should have its rate $\lambda = np = 2$. Let $Y \sim \text{Poi}(2)$, we have $\Pr(Y \leq 3) = \sum_{y=0}^{3} \frac{2^y e^{-2}}{y!} \approx 0.857$. The approximation is moderately good. To understand this, note that $n \geq 20$ and $np \leq 7$ satisfies the empirical rule adopted in the textbook. Nevertheless, as n is not large enough, the approximation is not very good.

8. (a) Let X be the number of typing errors made by the typist on a page, then $X \sim \text{Poi}(2)$. The probability of not getting her salary, i.e., having more than four errors on one page, is

$$\Pr(X > 4) = 1 - \Pr(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{2^x e^{-2}}{x!} \approx 0.053.$$

(b) Let Y be the number of typing errors made by the typist on twenty pages, then $Y \sim \text{Poi}(40)$. The probability of getting the bonus, i.e., making no more than 20 errors on 20 pages, is

$$\Pr(Y \leq 20) = \sum_{y=0}^{20} \frac{40^y e^{-40}}{y!} \approx 0.000368.$$

Therefore, it is quite unlikely that she can get the bonus.

9. (a) Let a be the lower bound and b be the upper bound, we know $\frac{a+b}{2} = 13$ and $\frac{(b-a)^2}{12} = 3$. Solving the two equations yield $a = 10$ and $b = 16$.

(b) Let X be the daily demand in kiloliters, then $X \sim \text{Uni}(10, 16)$. We then have $\Pr(X > 15) = 1 - \Pr(X \leq 15) = 1 - \frac{15-10}{16-10} \approx 0.167$. The probability of running out of gasoline is around 16.7%.

(c) Let h be the desired quantity, then h satisfies

$$\Pr(X \leq h) = 0.99 \quad \Leftrightarrow \quad \frac{h - 10}{6} = 0.99,$$

which implies that $h = 15.94$. The station should prepare 15.94 kiloliters of gasoline to achieve a service level of 99%.