Statistics I, Fall 2012
Suggested Solution for Homework 03

Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. (a) The ogive is depicted in Figure 1.

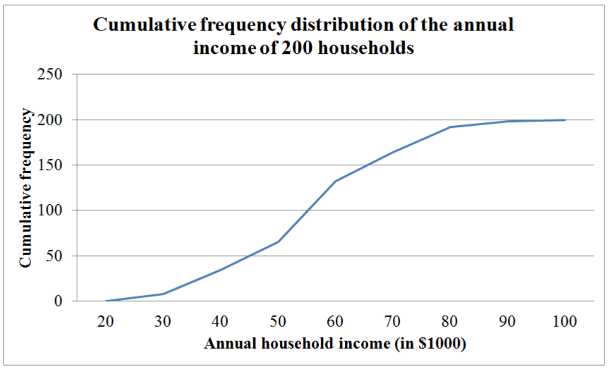

Figure 1: The ogive for Problem 1a.

(b) Table 1 summarizes the calculations, where

$$\bar{x} = \frac{25 \times 8 + 35 \times 26 + \cdots + 95 \times 2}{200} = 55.35$$

and

$$s^2 = \frac{(25 - 55.35)^2 \times 8 + (35 - 55.35)^2 \times 26 + \cdots + (95 - 55.35)^2 \times 2}{200 - 1} \approx 219.47.$$

Class (in $1000)	Frequency	Class midpoint M_i (in $1000)	$(M_i - \bar{x})^2$ (in 1000000 square dollars)
$[20, 30)$	8	25	921.1225
$[30, 40)$	26	35	414.1225
$[40, 50)$	31	45	107.1225
$[50, 60)$	67	55	0.1225
$[60, 70)$	32	65	93.1225
$[70, 80)$	28	75	386.1225
$[80, 90)$	6	85	879.1225
$[90, 100)$	2	95	1572.1225
Weighted average		$\bar{x} = 55.35$	$s^2 \approx 219.47$

Table 1: Calculations for Problem 1b.

(c) The mode is the 55 (in $1000), the class midpoint of the class with the highest frequency. The standard deviation is $\sqrt{219.47} \approx 14.82$ (in $1000).

(d) For the median, first note that the class $[50, 60)$ contains the $\frac{200}{2} = 100$th term and is the median class. Within the median class, the 100th term is the 35th, as $100 - (8 + 26 + 31 + 67) = 35$. Then we do an interpolation

$$50 + \frac{35}{67}(60 - 50) \approx 55.22.$$

Therefore, the median is 55.22 (in $1000).

(e) As we may observe, the mode is smaller than the median, which is smaller than the mean. This suggests that the data are skewed to the right.

2. (a) Table 2 lists the ranges $[\bar{x} - ks, \bar{x} + ks]$, $k = 1, 2, 3$, number of values in each range, proportion of values in each range, and the estimates based on the empirical rule.

k	Range from the empirical rule	Number of values in the range	Proportion of values in the range	Estimates from the empirical rule
1	$[7020.62, 24187.70]$	133	0.665	0.68
2	$[-1562.92, 32771.24]$	193	0.965	0.95
3	$[-10146.46, 41354.78]$	200	1	1.00

Table 2: Comparisons for Problem 2a.

(b) By comparing the last two columns, we may conclude that the empirical rule provides a good approximation for this set of data. The reason is that the data is approximately bell-shaped, as illustrated in Figure 2.

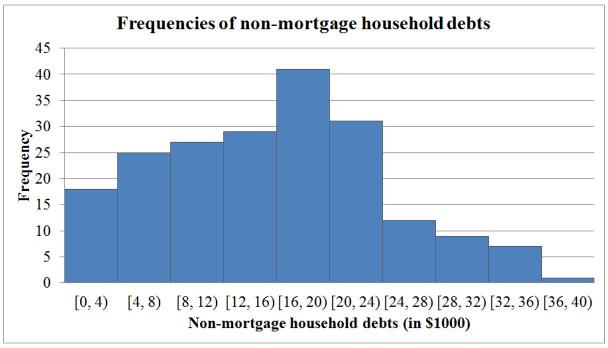

Figure 2: The histogram for Problem 2b.

3. Table 3 summarizes the calculations for the covariance, where

$$\sigma_{xy} = \frac{-0.99 + 6.21 + \cdots + 19.61}{10} = 3.99.$$

i	x_i	y_i	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
1	7	5	0.3	−3.3	−0.99
2	4	6	−2.7	−2.3	6.21
3	2	9	−4.7	0.7	−3.29
4	12	6	5.3	−2.3	−12.19
5	10	15	3.3	6.7	22.11
6	7	6	0.3	−2.3	−0.69
7	8	9	1.3	0.7	0.91
8	8	15	1.3	6.7	8.71
9	6	9	−0.7	0.7	−0.49
10	3	3	−3.7	−5.3	19.61
Average	$\mu_x = 6.7$	$\mu_y = 8.3$	–	–	$\sigma_{xy} = 3.99$

Table 3: Calculations for Problem 3.

4. The first step of writing a proof is always to define the notations clearly. Let the two-dimensional data be $\{(x_i, y_i)\}_{i=1,\ldots,N}$ with means $\mu_x = \frac{\sum_{i=1}^N x_i}{N}$ and $\mu_y = \frac{\sum_{i=1}^N y_i}{N}$, variances $\sigma_x^2 = \frac{\sum_{i=1}^N (x_i - \mu_x)^2}{N}$ and $\sigma_y^2 = \frac{\sum_{i=1}^N (y_i - \mu_y)^2}{N}$, covariance σ_{xy} and correlation coefficient ρ.

According to the Cauchy-Schwarz inequality, we have

$$\left|\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)\right|^2 \leq \sum_{i=1}^{N}(x_i-\mu_x)^2\sum_{i=1}^{N}(y_i-\mu_y)^2.$$

Note that both sides are nonnegative, so it is safe to take the square root for both sides. By doing so and then dividing both side by N, we have

$$|\sigma_{xy}| \equiv \left|\frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}\right| \leq \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu_x)^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(y_i-\mu_y)^2}{N}} \equiv \sigma_x\sigma_y.$$

Suppose the right-and-side (RHS) is zero, then $x_1 = x_2 = \cdots = x_N$ and $y_1 = y_2 = \cdots y_N$, which implies that $\rho = 0$. Suppose the RHS is positive, we may take it to the left-hand-side and yield

$$\frac{|\sigma_{xy}|}{\sigma_x\sigma_y} \leq 1 \quad \Leftrightarrow \quad \left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right| \leq 1 \quad \Leftrightarrow \quad |\rho| = 1.$$

This then implies that $-1 \leq \rho \leq 1$. Note that the first $\Leftrightarrow$ holds because $\sigma_x\sigma_y > 0$.

5. (a) The mean for y_is is

$$\mu_y \equiv \frac{\sum_{i=1}^{N} y_i}{N} = \frac{\sum_{i=1}^{N}(a+bx_i)}{N} = \frac{Na + b\sum_{i=1}^{N} x_i}{N} = a + b\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) = a + b\mu_x.$$

(b) The variance for y_is is

$$\sigma_y^2 \equiv \frac{\sum_{i=1}^{N}(y_i-\mu_y)^2}{N} = \frac{\sum_{i=1}^{N}[a+bx_i-(a+b\mu_x)]^2}{N} = \frac{\sum_{i=1}^{N} b^2(x_i-\mu_x)^2}{N} = b^2\sigma_x^2.$$

(c) The proof is wrong. First of all, if $b = 0$, it is straightforward to show that $\sigma_{xy} = 0$. Then $\rho = \frac{0}{0}$, which is undefined mathematically (in practice we say $\rho = 0$ in this case, but anyway it is not 1). Now assume that $b \neq 0$. In the last step

$$\rho \equiv \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{b\sigma_x^2}{\sigma_x(b\sigma_x)} = 1,$$

$\sigma_y^2 = b^2\sigma_x^2$ does not imply $\sigma_y = b\sigma_x$! In general, $\sqrt{x^2}$ is not always x. In fact, we have $\sqrt{x^2} = -x$ if $x < 0$. What is generally true is $\sqrt{x^2} = |x|$. Therefore, to fix the proof, we should replace the last step by

$$\rho \equiv \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{b\sigma_x^2}{\sigma_x|b\sigma_x|} = \left(\frac{b}{|b|}\right)\left(\frac{\sigma_x^2}{\sigma_x\sigma_x}\right) = \frac{b}{|b|} = \begin{cases} 1 & \text{if } b > 0 \\ -1 & \text{if } b < 0 \end{cases}.$$

In conclusion, when $y_i = a + bx_i$ for all i, $\rho = 1$ if $b > 0$, $\rho = -1$ if $b < 0$, and we define $\rho = 0$ if $b = 0$. Unless $b = 0$, there is the strongest correlation between x_is and y_is. Do you think that makes sense? Why or why not?

6. (a) $A \cup C = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
   (b) $A \cap B = \{7, 9\}$.
   (c) $A \cap B \cap C = \emptyset$.
   (d) $(A \cup B) \cap C = \{1, 2, 3, 4, 5, 7, 8, 9\} \cap C = \{1, 2, 3, 4\}$.
   (e) $(B \cap C) \cup (A \cap B) = \{2, 4\} \cup \{7, 9\} = \{2, 4, 7, 9\}$.

7. We shall first construct the joint probability table, as shown in Table 4.
   (a) $\Pr(A) = 0.392$.
   (b) $\Pr(A \cap F) = 0.089$.

	D	E	F	G	Total
A	0.038	0.114	0.089	0.152	0.392
B	0.101	0.051	0.101	0.051	0.304
C	0.114	0.063	0.038	0.089	0.304
Total	0.253	0.228	0.228	0.291	1.000

Table 4: The joint probability table for Problem 7.

	Freshman	Sophomore	Junior	Senior	Total
Female	0.05	0.075	0.06	0.09	0.275
Male	0.2	0.175	0.19	0.16	0.725
Total	0.25	0.25	0.25	0.25	1

Table 5: The joint probability table for Problem 8a.

(c) $\Pr(A|F) = \frac{0.089}{0.228} \approx 0.389$.

(d) $\Pr(B \cup E) = 0.304 + 0.228 - 0.051 = 0.481$.

(e) $\Pr(D \cup G|C) = \frac{0.114+0.089}{0.304} \approx 0.667$.

(f) They are not independent because, e.g., $\Pr(A)\Pr(D) \approx 0.099$, which is not $\Pr(A \cap D) \approx 0.038$.

8. (a) The joint probability table is shown in Table 5.

   (b) The proportion of girls with respect to the whole department is 0.275.

   (c) The proportion of girls with respect to the sophomore class is $\frac{0.075}{0.25} = 0.3$.

   (d) For (b), it is a marginal probability. For (c), it is a conditional probability.

   (e) The two variables are not independent. This is because knowing that one is a sophomore gives us additional information regarding the probability that she is a girl.

9. (a) This probability is the product of 78% (the proportion of people living in urban areas) and 13% (among them, the proportion of people taking care of ill relatives), i.e., $0.78 \times 0.13 = 0.1014$.

   (b) The joint probability table is shown in Table 6.

	Taking care	Not taking care	Total
Urban	0.1014	0.6786	0.78
Nonurban	0.0786	0.1414	0.22
Total	0.18	0.82	1

Table 6: The joint probability table for Problem 9b.

   (c) The conditional probability is $\frac{0.0786}{0.18} \approx 0.437$.