Statistics I, Fall 2012
Suggested Solution for Midterm 1

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. (a) False. It is of the ratio level.
   (b) True. It is of the ratio level.
   (c) False. The average lifetime of the 100 cameras is a statistic, not a parameter.
   (d) False. This is always true even without independence.
   (e) False. This may still sometimes be true without independence. For example, consider two random variables X_1 and X_2 such that $\Pr(X_1 = -1, X_2 = -1) = \frac{1}{4}$, $\Pr(X_1 = 0, X_2 = 1) = \frac{1}{2}$, and $\Pr(X_1 = 1, X_2 = -1) = \frac{1}{4}$. We have $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \mathbb{E}[X_1 X_2] = 0$ while X_1 and X_2 are not independent.
   (f) True. For example, $\text{Uni}(-1, 0)$.
   (g) False. They must be independent and have the same probability.
   (h) False. No probability can be greater than 1.
   (i) False. This has been discussed in Homework 5.
   (j) True. This has been discussed in Homework 5.

2. Suppose we define $Z = \frac{X-\mu}{\sigma}$, we know $Z \sim \text{ND}(0, 1)$ and has pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

   Because the random variable Y is nothing but $|Z|$, its pdf $f(x)$ can be found by moving the negative part to the positive part. In other words, its pdf is

$$f(x) = \begin{cases} 2\phi(x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-x^2/2} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$

3. (a) The scatter plot is depicted in Figure 1.

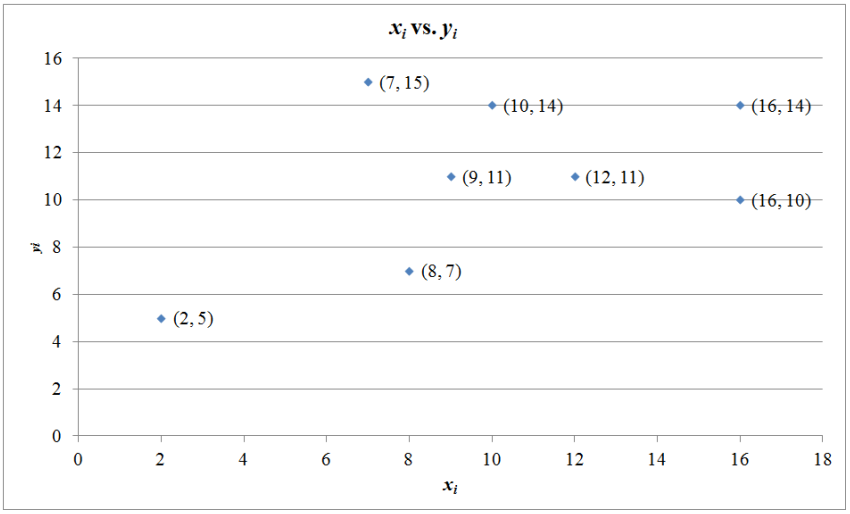

Figure 1: The scatter plot for Problem 3a.

(b) First we calculate the means for $\{x_i\}_{i=1,\dots,8}$ and $\{y_i\}_{i=1,\dots,8}$ as

$$\mu_x = \frac{1}{8}\sum_{i=1}^{8} x_i = 10 \quad \text{and} \quad \mu_y = \frac{1}{8}\sum_{i=1}^{8} y_i = 10.875,$$

respectively. The standard deviations for $\{x_i\}_{i=1,\dots,8}$ and $\{y_i\}_{i=1,\dots,8}$ are thus

$$\sigma_x = \sqrt{\frac{1}{8}\sum_{i=1}^{8}(x_i-\mu_x)^2} \approx 4.387 \quad \text{and} \quad \sigma_y = \sqrt{\frac{1}{8}\sum_{i=1}^{8}(y_i-\mu_y)^2} \approx 3.295,$$

respectively.

(c) The covariance for $\{(x_i, y_i)\}_{i=1,\dots,8}$ is

$$\frac{1}{8}\sum_{i=1}^{8}(x_i-\mu_x)(y_i-\mu_y) = 7.$$

The correlation coefficient is $\frac{7}{\sigma_x\sigma_y} \approx 0.484$. There is a moderately weak positive correlation between the two sets of data.

4. (a) Nominal.

(b) The bar chart is depicted in Figure 2.

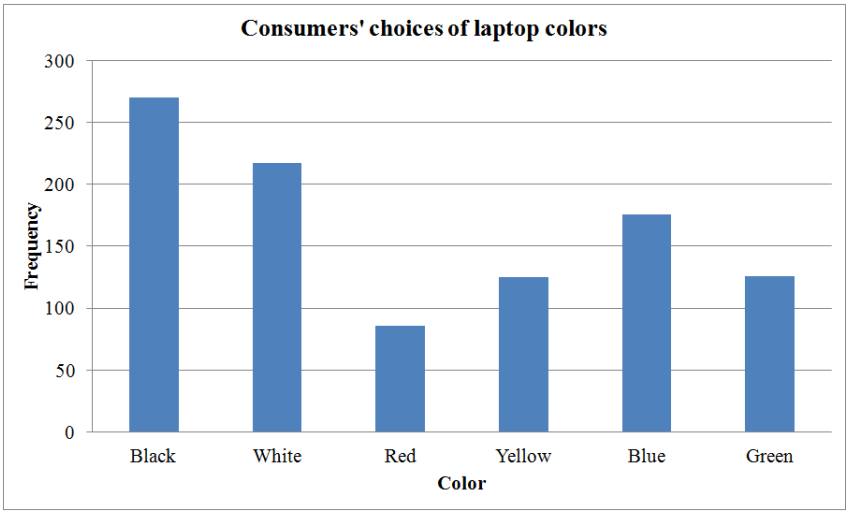

Figure 2: The bar char for Problem 4b.

(c) The Pareto chart is depicted in Figure 3.

5. (a) The frequency distribution is demonstrated in Table 1.

(b) The histogram is depicted in Figure 4.

(c) Let M_i be the class midpoint and f_i be the frequency of the ith class, the grouped mean is

$$\mu_{\text{grouped}} = \frac{\sum_{i=1}^{9} M_i f_i}{100} = 1.81$$

and the grouped variance is

$$\sigma^2_{\text{grouped}} = \frac{\sum_{i=1}^{9}(M_i-\mu_{\text{grouped}})^2 f_i}{100-1} \approx 2.277.$$

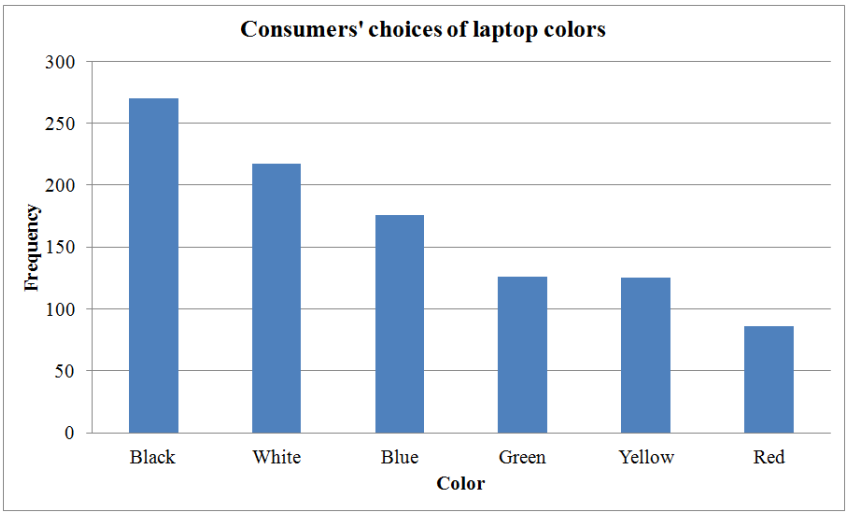

Figure 3: The Pareto char for Problem 4c.

Class	Frequency	Relative frequency	Cumulative frequency
$[0, 1)$	39	0.39	39
$[1, 2)$	28	0.28	67
$[2, 3)$	13	0.13	80
$[3, 4)$	9	0.09	89
$[4, 5)$	7	0.07	96
$[5, 6)$	3	0.03	99
$[6, 7)$	0	0	99
$[7, 8)$	1	0.01	100

Table 1: The frequency distribution for Problem 5.

(d) Skewed to the right.

(e) As we know, the grouped mean is not as accurate as the ungrouped mean because it uses the class midpoint to represent all the values in that class. For an exponentially distributed set of data, most data will be overestimated by its class midpoint while only a small amount of data will be underestimated. Therefore, the grouped mean should overestimate the upgrouped mean. As the grouped mean 1.81 is indeed larger than the ungrouped mean 1.77, this supports our conjecture that the data follow an exponential distribution.

6. Let X be the number of arrivals in a minute, then X follows a Poisson distribution with rate $\frac{1}{10} = 0.1$ arrival per minute.

(a) Let Y be the number of arrivals in an hour, then Y follows a Poisson distribution with rate $60 \times 0.1 = 6$ arrivals per hour. The probability that in an hour there are at least 10 arrivals is

$$\Pr(Y \geq 10 | \lambda = 6) = 1 - \Pr(Y \leq 9 | \lambda = 6) \approx 0.084.$$

(b) The probability that in an hour there are no more than 5 arrivals is

$$\Pr(Y \leq 5 | \lambda = 6) \approx 0.446.$$

(c) Let Z be the number of arrivals in 30 minutes, then Z follows a Poisson distribution with rate $30 \times 0.1 = 3$ arrivals per 30 minutes. The expected number of arrivals in 30 minutes is thus $\mathbb{E}[Z] = 3$ arrivals.

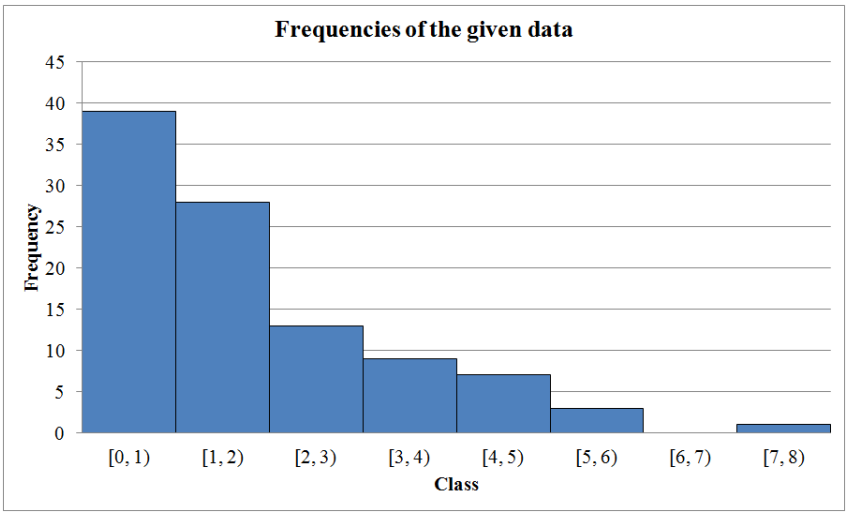

Figure 4: The histogram for Problem 5.

7. (a) We have

$$\begin{aligned}\Pr(X_1 + X_2 = 0) &= \Pr(X_1 = 0, X_2 = 0) = \Pr(X_1 = 0)\Pr(X_2 = 0)\\ &\approx (9.766 \times 10^{-4})(6.047 \times 10^{-3}) \approx 5.905 \times 10^{-6},\end{aligned}$$

where the first equality is due to the nonnegativity of binomial random variables and the second is due to the independence of X_1 and X_2.

(b) We have

$$\begin{aligned}\Pr(X_1 X_2 = 0) &= \Pr(X_1 = 0) + \Pr(X_2 = 0) - \Pr(X_1 = 0, X_2 = 0)\\ &\approx 9.766 \times 10^{-4} + 6.047 \times 10^{-3} - 5.905 \times 10^{-6}\\ &\approx 7.017 \times 10^{-3}.\end{aligned}$$

(c) We have

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 10 \times 0.5 \times 0.5 + 10 \times 0.4 \times 0.6 = 2.5 + 2.4 = 4.9,$$

where the first equality is due to the independence of X_1 and X_2.

(d) We have

$$\begin{aligned}\Pr(X_1 + X_2 = 4) &= \sum_{i=0}^{4} \Pr(X_1 = i, X_2 = 4 - i)\\ &= \sum_{i=0}^{4} \Pr(X_1 = i)\Pr(X_2 = 4 - i) \approx 0.0136,\end{aligned}$$

where the second equality is due to the independence of X_1 and X_2.

8. By the definition of conditional probability, we have

$$\Pr(X > t + s | X > t) = \frac{\Pr(X > t + s, X > t)}{\Pr(X > t)}.$$

Note that the joint event of the two events $X > t + s$ and $X > t$ is exactly $X > t + s$. It then follows that

$$\Pr(X > t + s | X > t) = \frac{\Pr(X > t + s)}{\Pr(X > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \Pr(X > s).$$

9. (a) Because $\frac{30}{2} = 15$ is a whole number, the median is the average of the 15th and 16th values, which is $\frac{55+61}{2} = 58$.

   (b) Because $\frac{30}{4} = 7.5$ is a fractional number, the first quartile is the 8th value, which is 34. Because $\frac{3\times 30}{4} = 22.5$ is a fractional number, the third quartile is the 23rd value, which is 83.

10. (a) To calculate $\Pr(Y_1 = 3)$, we may consider a related experiment as follows. Let $W_i = 1$ if the outcome of the original experiment is 1. Let $W_i = 0$ otherwise. Then we know $\Pr(W_i = 1) = 0.5$ and thus $Y_1 = \sum_{i=1}^{10} W_i$ follows a binomial distribution with 10 trials and probability 0.5. Therefore, $\Pr(Y_1 = 3) = \binom{10}{3}(0.5)^3(0.5)^7 \approx 0.117$.

    (b) To calculate $\Pr(Y_2 = 2)$, we may consider a related experiment as follows. Let $Z_i = 1$ if the outcome of the original experiment is 2. Let $Z_i = 0$ otherwise. Then we know $\Pr(Z_i = 1) = 0.1$ and thus $Y_2 = \sum_{i=1}^{10} Z_i$ follows a binomial distribution with 10 trials and probability 0.1. Therefore, $\Pr(Y_2 = 2) = \binom{10}{2}(0.1)^2(0.9)^8 \approx 0.194$.

    (c) We may apply the definition of conditional probability to solve this problem. Nevertheless, if we utilize the additional information of $Y_1 = 2$, the problem can be solved in an easier way. Suppose we know $Y_1 = 2$, this means $Y_2 + Y_3 = 8$. In other words, we must get two 3s out of 8 trials. For each of these eight trials, we know the outcome is either 2 or 3, so the probability of seeing 2 is 0.2 and that of seeing 3 is 0.8. We thus know that "Y_3 given $Y_1 = 2$" follows a binomial distribution with 8 trials and probability 0.8. Therefore, $\Pr(Y_3 = 2|Y_1 = 2) = \binom{8}{2}(0.8)^2(0.2)^8 \approx 0.00115$.

    Let's try to follow the definition directly. The desired probability is

$$\Pr(Y_3 = 2|Y_1 = 2) = \frac{\Pr(Y_3 = 2, Y_1 = 2)}{\Pr(Y_1 = 2)} = \frac{\binom{10}{2}(0.4)^2\binom{8}{2}(0.5)^2(0.1)^6}{\binom{10}{2}(0.5)^2(0.5)^8},$$

    where in the numerator $\binom{10}{2}(0.4)^2$ is the probability of seeing two 3s, $\binom{8}{2}(0.5)^2$ is that of seeing two 1s out of the remaining 8 trials, and $(0.1)^6$ is that of seeing six 2s out of the remaining 6 trials. Straightforward arithmetic then yields

$$\Pr(Y_3 = 2|Y_1 = 2) = \binom{8}{2}\frac{(0.4)^2(0.1)^6}{(0.5)^8} = \binom{8}{2}(0.8)^2(0.2)^6,$$

    which is exactly what we have calculated with method 1.

11. (a) Let X_1 be the number of accidents in the first week. Because $\Pr(X = 2|\lambda = 1) \approx 0.184$, $\Pr(X = 2|\lambda = 2) \approx 0.271$, $\Pr(X = 2|\lambda = 3) \approx 0.224$, and $\Pr(\lambda = i) = \frac{1}{3}$ for $i = 1, 2, 3$ (this is the prior belief), we have the posterior probability

$$\Pr(\lambda = 1|X_1 = 2) \approx \frac{\frac{1}{3}\times 0.184}{\frac{1}{3}\times 0.184 + \frac{1}{3}\times 0.271 + \frac{1}{3}\times 0.224} \approx 0.271.$$

    The other two posterior probabilities $\Pr(\lambda = 2|X_1 = 2) \approx 0.399$ and $\Pr(\lambda = 3|X_1 = 2) \approx 0.330$ can be found in a similar way.

    (b) We may simply use the posterior belief at the end of the first week to be the prior belief at the beginning of the second week. Let X_2 be the number of accidents in the second week. Because $\Pr(X = 0|\lambda = 1) \approx 0.368$, $\Pr(X = 0|\lambda = 2) \approx 0.135$, and $\Pr(X = 0|\lambda = 3) \approx 0.05$, we have

$$\Pr(\lambda = 1|X_1 = 2, X_2 = 0) \approx \frac{0.271 \times 0.368}{0.271 \times 0.368 + 0.399 \times 0.135 + 0.330 \times 0.05} \approx 0.586.$$

    The other two posterior probabilities $\Pr(\lambda = 2|X_1 = 2, X_2 = 0) \approx 0.317$ and $\Pr(\lambda = 3|X_1 = 2, X_2 = 0) \approx 0.097$ can be found in a similar way.

12. Let $\mu = 25$ minutes and $\sigma = 7$ minutes be the mean and standard deviation. Let $X \sim \mathrm{ND}(25, 7)$ be the waiting time and $Z \sim \mathrm{ND}(0, 1)$.

    (a) The coefficient of variation is $\frac{\sigma}{\mu} = 0.28$ (no unit of measurement).

(b) For a normal distribution, the median is its mean $\mu = 25$ minutes.

(c) The variance is $\sigma^2 = 49$ square minutes.

(d) $\Pr(X < 20) = \Pr(Z < \frac{20-25}{7}) \approx \Pr(Z < -0.714) \approx 0.238$.

(e) $\Pr(23 < X < 33) = \Pr(\frac{23-25}{7} < Z < \frac{33-25}{7}) \approx \Pr(-0.286 < Z < 1.143) \approx 0.486$.

(f) $\Pr(X > x) = 0.03 \Leftrightarrow \Pr(Z > \frac{x-25}{7}) = 0.03 \Leftrightarrow \frac{x-25}{7} \approx 1.88 \Leftrightarrow x \approx 38.16$.

(g) $\Pr(15 < X < 35) = \Pr(|X - 25| < \frac{10}{7}\sigma) \geq 1 - \frac{1}{(\frac{10}{7})^2} = 1 - \frac{49}{100} = 0.51$.

13. (a) $\Pr(C \cup A) = 0.2 + 0.1 + 0.2 + 0.4 = 0.9$.

(b) $\Pr(A|E) = \frac{0.2}{0.2} = 1$.

(c) $\Pr(A) = 0.2 + 0.1 + 0.2 = 0.5$.

14. (a) The cdf is calculated below in five regions:

- For $x < 0$, the cdf is 0.
- For $x \in [0, 1]$, the cdf is $0 + \int_0^x y^2 dy = \frac{x^3}{3}$.
- For $x \in (1, 2)$, the cdf is $\int_0^1 y^2 dy + 0 = \frac{1}{3}$.
- For $x \in [2, 3]$, the cdf is $\frac{1}{3} + \int_2^x \frac{2}{3} dy = \frac{2}{3}x - 1$.
- For $x > 0$, the cdf is $\frac{1}{3} + \int_2^3 \frac{2}{3} dy = 1$.

(b) The expectation $\mathbb{E}[X]$ is

$$\int_0^1 x \cdot x^2 dx + \int_2^3 x \cdot \frac{2}{3} dx = \frac{1}{4}x^4\Big|_0^1 + \frac{1}{3}x^2\Big|_2^3 = \frac{1}{4} + 3 - \frac{4}{3} = \frac{23}{12}.$$

(c) $\mathbb{E}\left[\left|X - \frac{3}{2}\right|\right]$ is

$$\int_0^1 \left(\frac{3}{2} - x\right) \cdot x^2 dx + \int_2^3 \left(x - \frac{3}{2}\right) \cdot \frac{2}{3} dx = \left(\frac{1}{2}x^3 - \frac{1}{4}x^4\right)\Big|_0^1 + \left(\frac{1}{3}x^2 - x\right)\Big|_2^3$$
$$= \frac{1}{2} - \frac{1}{4} + \frac{5}{3} - 1 = \frac{11}{12}.$$