Statistics I, Fall 2012
Suggested Solution for Final Exam

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

1. (a) True. A Pareto chart is a bar chart with bars sorted. A bar chart is not a Pareto chart if its bars are not sorted.

   (b) False. The Poisson distribution is discrete but it is skewed to the right.

   (c) True. This can be proved with moment generating functions.

   (d) True. We have $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.

   (e) True. Though we typically use the z test in this case, we may still use the t test. We still have $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ in this case.

   (f) False. The two-tailed confidence interval for a the population variance is asymmetric (because the chi-square distribution is asymmetric).

   (g) False. When the p-value becomes smaller, the probability that there is a significant difference becomes larger.

   (h) True. Both error probabilities will decrease.

   (i) False. When we do not reject H_0 in a test, we can make no conclusion.

   (j) False. When we reject H_0 in a test whose significance level is α, we do not know the probability that H_0 is true. What we know is the conditional probability of rejecting H_0 given that H_0 is true.

2. (a) We have

$$\begin{aligned}\Pr(X_1+X_2=0) &= \Pr(X_1=0, X_2=0) = \Pr(X_1=0)\Pr(X_2=0)\\ &= 0.1678 \times 0.5341 = 0.0892,\end{aligned}$$

   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

$$\begin{aligned}\Pr(X_1X_2=0) &= \Pr(X_1=0)+\Pr(X_2=0)-\Pr(X_1=0,X_2=0)\\ &= 0.1678+0.5341-0.0892 = 0.6101.\end{aligned}$$

   (c) We have

$$\begin{aligned}\mathrm{Var}(X_1+X_2) &= \mathrm{Var}(X_1)+\mathrm{Var}(X_2)\\ &= 8\times 0.2\times 0.8 + 6 \times 0.1 \times 0.9 = 1.28+0.54 = 1.82,\end{aligned}$$

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

   (d) We have

$$\begin{aligned}\Pr(X_1+X_2=3) &= \sum_{i=0}^{3}\Pr(X_1=i, X_2=3-i)\\ &= \sum_{i=0}^{3}\Pr(X_1=i)\Pr(X_2=3-i) \approx 0.2175,\end{aligned}$$

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

3. Let $Z = 1$ if the shipping company is strong and 0 if it is weak. The prior belief is $\Pr(Z = 1) = 0.4$ and $\Pr(Z = 0) = 0.6$.

   (a) Let $Y_1 = 2$ be the shipping time of the first order. We have

   $$\begin{aligned}\Pr(Z = 1|Y_1 = 2) &= \frac{\Pr(Z = 1)\Pr(Y_1 = 2|Z = 1)}{\Pr(Z = 1)\Pr(Y_1 = 2|Z = 1) + \Pr(Z = 0)\Pr(Y_1 = 2|Z = 0)} \\ &= \frac{0.4 \times 0.3}{0.4 \times 0.3 + 0.6 \times 0.2} = 0.5.\end{aligned}$$

   It then follows that $\Pr(Z = 0|Y_1 = 2) = 1 - 0.5 = 0.5$. Therefore, the posterior belief after the order is that the company is strong with probability 50% and weak with probability 50%.

   (b) We may use the posterior belief in Part (a) as the prior belief here. We have

   $$\begin{aligned}\Pr(Z = 1|Y_1 = 2) &= \frac{\Pr(Z = 1)\Pr(Y_1 = 2|Z = 1)}{\Pr(Z = 1)\Pr(Y_1 = 2|Z = 1) + \Pr(Z = 0)\Pr(Y_1 = 2|Z = 0)} \\ &= \frac{0.5 \times 0.3}{0.5 \times 0.3 + 0.5 \times 0.2} = 0.6.\end{aligned}$$

   It then follows that $\Pr(Z = 0|Y_1 = 2) = 1 - 0.6 = 0.4$. Therefore, the posterior belief after the two orders is that the company is strong with probability 60% and weak with probability 40%.

   (c) Let W be the number of orders that are fulfilled in one day among the six orders. If the shipping company is strong, we have $W \sim \text{Bi}(6, 0.5)$ and $\Pr(W = 4|Z = 1) = 0.0938$; otherwise, $W \sim \text{Bi}(6, 0.2)$ and $\Pr(W = 4|Z = 0) = 0.0092$. It then follows that

   $$\begin{aligned}\Pr(Z = 1|W = 4) &= \frac{\Pr(Z = 1)\Pr(W = 4|Z = 1)}{\Pr(Z = 1)\Pr(W = 4|Z = 1) + \Pr(Z = 0)\Pr(W = 4|Z = 0)} \\ &= \frac{0.4 \times 0.0938}{0.4 \times 0.0938 + 0.6 \times 0.0092} = 0.9105\end{aligned}$$

   and $\Pr(Z = 0|W = 4) = 1 - 0.9105 = 0.0895$. The posterior belief after these six orders is that the company is strong with probability 91.05% and weak with probability 8.95%.

4. (a) We have $\mathbb{E}\left[\overline{X}\right] = \mathbb{E}[X] = 1 \times 0.2 + 2 \times 0.3 + 3 \times 0.5 = 2.3$.

   (b) The sample proportion of 1 $\widehat{P}$ We have $\text{Var}\left(\widehat{P}\right) = \frac{0.2\times0.8}{3} = 0.0533$.

   (c) To find the sampling distribution of M, we need to write down all the results of ordering the three sampled values and the associated probabilities. Let $X_{(k)}$ be the kth small sampled value, Table 1 summarizes all the information we need. For example, row 1 means the probability of obtaining three 1s in the sample is $0.2^3 = 0.008$ and the median in this case is 1; row 2 means the probability of obtaining two 1s and one 2 in the sample is $\binom{3}{1}(0.2)^2(0.3) = 0.036$ and the median in this case is also 1. We may then further combine the information contained in Table 1 to obtain the sampling distribution of M:

   $$\Pr(M = 1) = 0.104, \quad \Pr(M = 2) = 0.396, \quad \text{and } \Pr(M = 3) = 0.5.$$

   (d) $\mathbb{E}[M] = 1 \times 0.104 + 2 \times 0.396 + 3 \times 0.5 = 2.396$.

5. (10 points) Consider the experiment of rolling one fair dice. Let X be the outcome.

   (a) $\Pr(X = i) = \frac{1}{6}$ for $i \in \{1, 2, ..., 6\}$.

   (b) Let $m(t)$ be the moment generating function of X. We have

   $$m(t) = \mathbb{E}\left[e^{tX}\right] = \frac{1}{6}\left(e^t + e^{2t} + \cdots + e^{6t}\right).$$

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	Median	Probability
1	1	1	1	0.008
1	1	2	1	0.036
1	1	3	1	0.060
1	2	2	2	0.054
1	2	3	2	0.180
1	3	3	3	0.150
2	2	2	2	0.027
2	2	3	2	0.135
2	3	3	3	0.225
3	3	3	3	0.125

Table 1: Calculations for the sampling distribution of M.

(c) The first-order derivative of $m(t)$ is $m'(t) = \frac{1}{6}\big(e^t + 2e^{2t} + \cdots + 6e^{6t}\big)$. Therefore, we have

$$\mathbb{E}[X] = m'(0) = \frac{1}{6}(1 + 2 + \cdots + 6) = \frac{7}{2} = 3.5.$$

(d) The second-order derivative of $m(t)$ is $m''(t) = \frac{1}{6}\big(e^t + 4e^{2t} + \cdots + 36e^{6t}\big)$. Therefore, we have

$$\mathbb{E}\big[X^2\big] = m''(0) = \frac{1}{6}(1 + 4 + \cdots + 36) = \frac{91}{6}$$

and $\mathrm{Var}(X) = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} = 2.9167$.

(e) $\mathbb{E}[X^n] = \frac{d^n}{dt^n} m(t)|_{t=0} = \frac{1}{6}(1 + 2^n + \cdots + 6^n)$.

6. Let $\alpha = 0.1$ be the significance level. In Parts (a) to (c), because the sample is adopted from a finite population, we need to do the finite population correction. The correction coefficient is $\sqrt{\frac{N-n}{N-1}}$, where $N = 55$ is the population size and n is the sample size (different in different parts).

(a) Let μ_I be the average rating for IBM in our class and $\overline{X}_I$ be the average rating for IBM in the sample. The hypothesis is

$$\begin{aligned} H_0 &\colon \mu_I = 3.2 \\ H_0 &\colon \mu_I > 3.2. \end{aligned}$$

Because the sample size is large and the population variance is unknown, we adopt the z test and use the sample variance as a substitute. In this right-tailed test, the p-value is

$$\Pr(\overline{X}_I > 3.394) = \Pr\left(Z > \frac{3.394 - 3.2}{\frac{1.059}{\sqrt{33}}\sqrt{\frac{55-33}{55-1}}}\right) = \Pr(Z > 1.649) = 0.05.$$

As the p-value is smaller than $\alpha = 0.1$, we reject H_0. With a 10% significance level, there is a strong evidence showing that the average rating for IBM in our class is higher than 3.2. Jack should write the proposal.

(b) Let μ_M be the average rating for Microsoft in our class and $\overline{X}_M$ be the average rating for Microsoft in the sample. The hypothesis is

$$\begin{aligned} H_0 &\colon \mu_M = 3.2 \\ H_0 &\colon \mu_M < 3.2. \end{aligned}$$

Because the sample size is large and the population variance is unknown, we adopt the z test and use the sample variance as a substitute. In this left-tailed test, the p-value is

$$\Pr(\overline{X_M} < 3.394) = \Pr\left(Z < \frac{3.125 - 3.2}{\frac{0.887}{\sqrt{32}}\sqrt{\frac{55-32}{55-1}}}\right) = \Pr(Z < -0.733) = 0.768.$$

As the p-value is larger than $\alpha = 0.1$, we do not reject H_0. With a 10% significance level, there is no strong evidence showing that the average rating for Microsoft in our class is lower than 3.2. Jack cannot discourage his friend.

(c) Let p_I be the proportion of people in our class rating IBM as 4 or above and $\widehat{P}_I$ be the proportion of people in the sample rating IBM as 4 or above. The hypothesis is

$$H_0 \colon p_I = 0.5$$
$$H_0 \colon p_I < 0.5.$$

Because the sample size is large, we adopt the z test. In this left-tailed test, the p-value is

$$\Pr(\widehat{P}_I < 0.485) = \Pr\left(Z < \frac{0.485 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{33}}\sqrt{\frac{55-33}{55-1}}}\right) = \Pr(Z < -0.273) = 0.393.$$

As the p-value is larger than $\alpha = 0.1$, we do not reject H_0. With a 10% significance level, there is no strong evidence showing that the proportion of people in our class rating IBM as 4 or above is below one half.

(d) There is no evidence supporting that the population is normal. Without normality, we cannot apply the chi-square test, the only instrument we have for testing the population variance. Therefore, Jack can conclude nothing. If you failed to recognize this and did a chi-square test, you should do the following:

Let σ_G^2 be the the variance of ratings for Google in our class. The hypothesis is

$$H_0 \colon \sigma_G^2 = 0.3$$
$$H_0 \colon \sigma_G^2 \neq 0.3.$$

We use the chi-square test. In this two-tailed test, the two critical chi-square values are $\chi^2_{32,0.95} = 20.072$ and $\chi^2_{32,0.05} = 46.194$. The observed chi-square value is $\chi^2 = \frac{(33-1)(0.318)}{0.3} = 33.939$. As the observed chi-square value is between the two critical chi-square values, we do not reject H_0. With a 10% significance level, there is no strong evidence showing that the the variance of the ratings for Google in our class is not 0.3.

You may still get some partial credits if your calculations are correct.

7. (a) We cannot be 100% sure that $\mu_1 > \mu_2$. It is possible that $\mu_1 < \mu_2$ but $\bar{x}_1$ happens to be large while $\bar{x}_2$ happens to be small. As sample means are random, they do not imply any relationship among population means.

(b) He cannot be 95% sure that $\mu_1 > \mu_2$. What he can guarantee is the following: Based on the random sample mean, he will reject the null hypothesis $\mu_1 \leq \bar{x}_2$ if it is true with a 95% probability. In other words, he can be 95% confident that, if $\mu_1 \leq \bar{x}_2$, his decision is correct.

(c) He cannot be 95% sure that $\mu_1 > \mu_2$. The only two things he can guarantee is (1) if $\mu_1 \leq 100$, he will reject it with a 5% probability and (2) if $\mu_2 \geq 100$, he will reject it with a 5% probability. Combining these two, he can guarantee that if $\mu_1 \leq 100 \leq \mu_2$, he will conclude that $\mu_1 > \mu_2$ with a 0.25% probability.

8. (17 points) Consider a uniformly distributed population with lower bound a and upper bound $a+6$. You know the population is uniformly distributed but you do not know the value of a. Let μ be the population mean. Let $\overline{X}$ be the mean of a random sample whose sample size is $n = 2$.

(a) For a uniform distribution, we know its mean is the average of the two bounds. Therefore, $\mu = \frac{a+a+6}{2} = a + 3$.

(b) Intuitively, the curve should be a symmetric triangle. While the base is 6, the height should be $\frac{1}{3}$ so that the area of the triangle (i.e., the probability) can be 1. It then follows that the sampling distribution can be described by the pdf $f(x)$ over $[a, a+6]$:

$$f(x) = \begin{cases} \frac{1}{9}(x-a) & \text{if } x \in [a, a+3] \\ \frac{2}{3} - \frac{1}{9}(x-a) & \text{if } x \in [a+3, a+6] \end{cases}.$$

(c) The interval length is independent from the observed sample mean and can be calculated as follows. First, we need to determine a quantity $b \in [0, 6]$ such that $\int_a^{a+b} f(x)dx = 0.05$. $a + b$ will then be the lower bound of the confidence interval. As it is clear that $b < 3$, we have

$$\int_a^{a+b} f(x)dx = \int_a^{a+b} \frac{1}{9}(x-a)dx = \frac{1}{18}(x-a)^2\bigg|_a^{a+b} = \frac{b^2}{18} = 0.05,$$

which implies that $b = \sqrt{0.9} = 0.9487$. With this in mind, we know the leg length of the 90% confidence interval should be $3 - 0.9487 = 2.0513$. Now, because the observed sample mean is 8, the 90% confidence interval is $[8 - 2.0513, 8 + 2.0513] = [5.9487, 10.0513]$.

(d) Because $\overline{X}$ is an unbiased estimator of μ, which is $a + 3$, it is natural to conjecture that $\overline{X} - 3$ is an unbiased estimator of a. This is true as

$$\mathbb{E}\big[\overline{X} - 3\big] = \mathbb{E}\big[\overline{X}\big] - 3 = \mu - 3 = a.$$

(e) As the sample size is so large, the sample mean will (approximately) follow a normal distribution according to the central limit theorem. Its mean is the population mean $\mu = a + 3$ and its variance is the population variance

$$\int_a^{a+6} \frac{1}{6}(x-a-3)^2dx = \frac{1}{18}(x-a-3)^3\bigg|_a^{a+6} = 3$$

divided by the sample size 100. As $\sqrt{\frac{3}{100}} = 0.1732$, we have $\overline{Y} \sim \text{ND}(a + 3, 0.1732)$.

(f) We can solve this question through the usual interval estimation technique. As the sample mean follows the normal distribution, we calculate $z_{0.05} = 1.645$. The 90% confidence interval is thus

$$\Big[8 - 1.645 \times 0.1732, 8 + 1.645 \times 0.1732\Big] = [7.7151, 8.2849].$$

It is smaller than the interval found in Part (c) because the sample size is larger.

9. (a) With a 5% significance level, there is no strong evidence showing that $\mu > 10$.

(b) This claim is false. As μ is a fixed number, whether $\mu > 10$ is also fixed: It must be either true or false, though we do not know. There is no randomness associated with this statement. What we can guarantee is that if $\mu \leq 10$, we will claim $\mu > 10$ with a 5% probability.

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

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

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

(c) $\Pr(X > x) = 0.1 \Leftrightarrow \Pr(Z > \frac{x-25}{7}) = 0.1 \Leftrightarrow \frac{x-25}{7} \approx 1.282 \Leftrightarrow x \approx 33.971$.

11. (a) This is correct for this study. Typically the second one is wrong. However, because in this study there are only two possibilities of p: 0.2308 and 0.2424, so writing $p \neq 0.2308$ still leaves us a default position ($p = 0.2424$). In fact, writing

$$\begin{array}{lcl} H_0\colon p = 0.2308 & \text{or} & H_0\colon p = 0.2424 \\ H_a\colon p = 0.2424 & & H_a\colon p = 0.2308 \end{array}$$

is also correct.

(b) Let $\widehat{P}$ be the proportion of coins that are counterfeit in a sample and $\alpha = 0.01$ as the significance level. Because the sample size is large, we adopt the z test. In this right-tailed test, $z_{0.01} = 2.3263$ and thus the critical value is

$$0.2308 + 2.3263 \times 0.0596 = 0.3694,$$

where $0.0596 = \sqrt{(0.2308)(1 - 0.2308)/50}$ is the standard error. As the observed sample proportion 0.24 is lower than the critical value, we do not reject H_0. With a 1% significance level, there is no strong evidence showing that the proportion of coins that are counterfeit is higher than 0.2308. They cannot conclude that the 200 counterfeit coins were not destroyed.

(c) Making a Type II error means not rejecting a false null hypothesis. In this test, the null hypothesis is $p = 0.2308$. If it is false, it must be that $p = 0.2424$. Based on this, the probability of not rejecting the null hypothesis is

$$\Pr\left(\widehat{P} < 0.239\right) = \Pr\left(Z < \frac{0.3694 - 0.2424}{0.0606}\right) = \Pr(Z < 2.0948) = 0.9819,$$

where $0.0606 = \sqrt{(0.2424)(1 - 0.2424)/50}$ is the standard error. In summary, the probability of making a Type II error is 98.19%.