Suggested Solution for Homework 2

Statistics and Data Analysis, Fall 2015

1. (25 points; 5 points each)

(a) $1 - (0.2 + 0.1 + 0.3 + 0.2 + 0.1) = \mathbf{0.1}$

(b) $\mathrm{E[X]} = 0.2\times0 + 0.1\times5 + 0.3\times10 + 0.2\times11 + 0.1\times12 + 0.1\times15 = \mathbf{8.4}$

(c) $\mathrm{Var(X)} = 0\times(0 - 8.4)^2 + 0.1\times(5 - 8.4)^2 + 0.3\times(10 - 8.4)^2 + 0.2\times(11 - 8.4)^2 +$

$0.1\times(12 - 8.4)^2 + 0.1\times(15 - 8.4)^2 = \mathbf{23.04}$

(d)

y	5	10	15	16	17	20
$\Pr(X{=}y)$	0.2	0.1	0.3	0.2	0.1	0.1

(e) $\mathrm{E}[Y] = 0.2\times5 + 0.1\times10 + 0.3\times15 + 0.2\times16 + 0.1\times17 + 0.1\times20 = \mathbf{13.4}$

$\mathrm{Var}(Y) = \mathbf{23.04}$

2. (25 points; 5 points each)

(a) $0.001\times(10000 - 100) + 0.05\times(1000 - 100) + (1 - 0.001 - 0.05)\times(0 - 100) = \mathbf{-40}$

(b) $\mathrm{Var}(X) = 0.001\times(9900 + 40)^2 + 0.05\times(900 + 40)^2 + 0.949\times(-100 + 40)^2 = 146400$

$\mathrm{Std}(X) = \mathbf{382.6225}$

(c) $0.001\times(10000 - 150) + 0.05\times(1000 - 150) + (1 - 0.001 - 0.05)\times(50 - 150) = \mathbf{-42.55}$

(d) $\mathrm{Var}(Y) = 0.001\times(9850 + 42.55)^2 + 0.05\times(850 + 42.55)^2 + 0.949\times(-100 + 42.55)^2 =$

140826.9975

$\mathrm{Std}(Y) = \mathbf{375.2692}$

(e) No, since the expected profit of not buying the insurance is higher. However, if you consider the variance, you may buy the insurance to get a smaller variability.

3. (25 points; 5 points each)

(a) $X_1 \sim \mathrm{ND}(5.5, 0.5)$

$\Pr(X_1 < 5.2) = \mathbf{0.27425}$

(b) $X_2 \sim \mathrm{ND}(5.5, 0.5)$

$\Pr(X_1 < 5.2 \ \& \ X_2 < 5.2\ ) = \Pr(X_1 < 5.2) \times \Pr(X_2 < 5.2) = \mathbf{0.0752}$

(c) $\Pr(X_1 < t) = 0.05$

$t = \mathbf{4.67757}$

(d) $\Pr(X_1 < t \ \& \ X_2 < t\ ) = \Pr(X_1 < t) \times \Pr(X_2 < t) = 0.01$

$\Pr(X_1 < t) = \Pr(X_1 < t) = 0.1$

$t = \mathbf{4.8592}$

(e) Since the standard deviation is fixed ($\sigma = 0.5$), the curve of the normal distribution shifted but not transformed.

$5.5 - 4.8592 = \mu - 5.2$

$\mu = (5.5 - 4.8592) + 5.2 = \mathbf{5.8408}$

4. (25 points; 5 points each)

(a) From the probability distribution table below, we could see that "education" and "buy" are not quite independent. Take column "basic.6y" and "university.degree" for instance, the proportion of "no" in "university.degree" is about five times larger than the proportion in "basic.6y", and "yes" in "university.degree" is about ten times larger than in "basic.6y".

	illiterate	basic.4y	basic.6y	basic.9y	high.school	prof.course[1]	uni.degree[2]	unknown	Total
Pr(no)	0.0002	0.0949	0.0512	0.1289	0.2000	0.1141	0.2668	0.0342	0.8905
Pr(yes)	0	0.0090	0.0041	0.0104	0.0235	0.0157	0.0400	0.0063	0.1094
Total	0.0002	0.1041	0.0553	0.1393	0.2235	0.1298	0.3060	0.0405	1

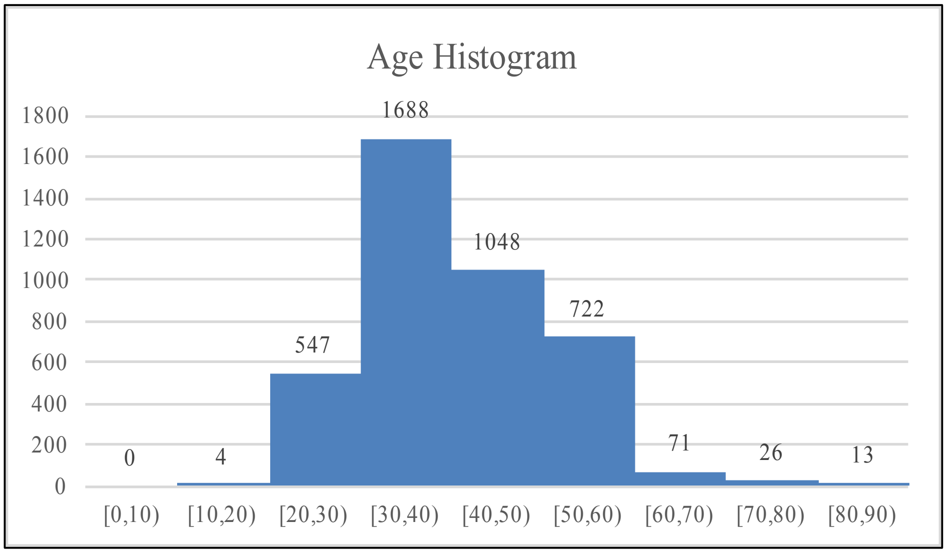
[1] professional.course

[2] university.degree

(b) Probability distribution of “job” :

Job	Pr(x)
admin.	0.245690702
blue-collar	0.214615198
entrepreneur	0.035931051
housemaid	0.026705511
management	0.078659869
retired	0.040301044
self-employed	0.038601602
services	0.095411508
student	0.019907745
technician	0.167759165
unemployed	0.026948288
unknown	0.009468318
total	1

(c) Histogram for “age” :



(d) Observed and theoretical frequencies table for “age”:

	Observed	Theoretical	|Difference|
[0,10)	0	7.0051	7.0051
[10,20)	4	98.1241	94.1241
[20,30]	547	567.6546	20.6546
[30,40)	1688	1368.4063	319.5936
[40,50)	1048	1381.9847	333.9847
[50,60]	722	584.7586	137.2413
[60,70]	71	103.1208	32.1208
[70,80]	26	7.5121	18.4878
[80,90]	13	0.2237	12.7762

Total difference is 975.9886.

(e) If one considers the distribution of “age” data tends to be close to the center, and its peak locates at its mean (expected value) and its mean equals its median, one may think it is reasonable to use normal distribution, thus, estimate “age” as a normal random variable is good and useful. Otherwise, one may think the distribution of “age” does not tend to fit normal distribution in real world, maybe the amount of the elders is quite the same as the middle-aged, one may consider it is not good to estimate “age” as a normal random variable.

5. (0 point; do need to submit your answers for this problem)

(a) $\bar{x} \sim \text{ND}(\ \mu, \sigma\sqrt{2})$, $\mu_{\bar{x}} = \mu = 5.5$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{2}} = 0.3535$

(b) $\bar{x} \sim \text{ND}\left(\ \mu, \frac{\sigma}{\sqrt{n}}\ \right)$, $\mu_{\bar{x}} = \mu = 5.5$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{n}}$

(c) Let $t = 5.2$ and $n = 2$, $\bar{x} \sim \text{ND}\left(\ \mu, \frac{\sigma}{\sqrt{2}}\ \right)$, $\Pr(\bar{x} < 5.2) = 0.1980$

The probability distribution and probability of rejection would not be the same as Problem 3b. Here we draw only once; and sample size is 2. But in Problem 3b, we draw twice; and sample size is 1. Apparently the two questions are not the same.