Suggested Solution for Midterm Exam

Statistics and Data Analysis, Fall 2015

1. (20 points)

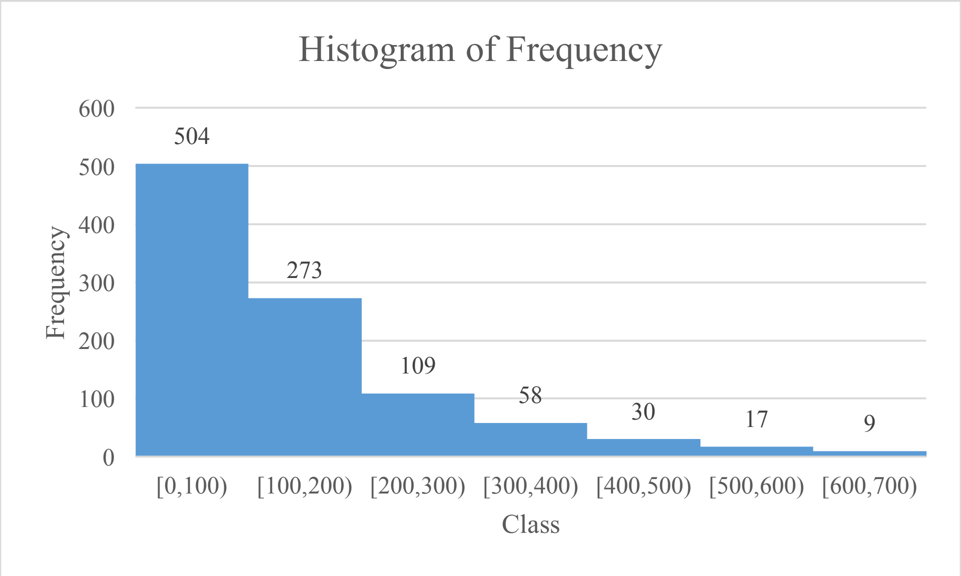
(a) (3 points)

Frequency distribution with classes [0,100), [100,200), ..., and [600,700).

Class	Frequency
[0, 100)	504
[100, 200)	273
[200, 300)	109
[300, 400)	58
[400, 500)	30
[500, 600)	17
[600, 700)	9

(b) (3 points)

Histogram with classes [0,100), [100,200), ..., and [600,700).



(c) (4 points)

Mean is 136.215. Median is 98.5. Mode is 50.

By comparing the mean, median, and mode, the data distribution is said to be skewed to the left if mode > median > mean, and skewed to the right if mode < median < mean. For the dataset here, the data is skewed to the right (mean > median > mode).

(d) (3 points)

Variance is 16951.9807. Standard deviation is 130.1997. Coefficient of variance is 0.9558.

(e) (3 points)

There are 23 values above $\mu + 3\sigma$, 0 value below $\mu - 3\sigma$. After removing the values, the new standard deviation is 109.9945 which is smaller than the original one 130.1997. The new standard deviation would become smaller because we remove the data with high variance (3σ from μ).

(f) (4 points)

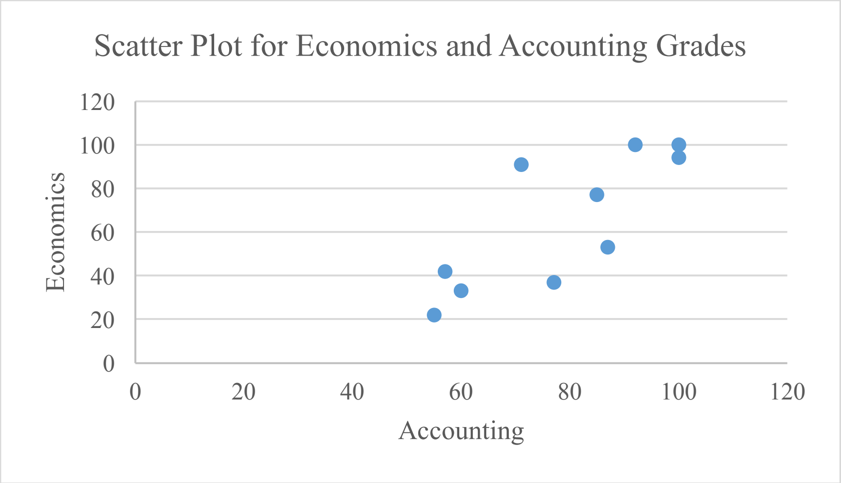
The mean absolute deviation is 97.5036. It is less than standard deviation. Standard deviation squares the deviations and sum these squares. Due to the square, more weight is given to high deviations, hence the sum of these squares will be different from the sum of absolute value.

2. (15 points)

(a) (3 points)

Consider a set of sample data, the standard deviation of Accounting, Statistics and Economics are14.0228, 10.7305 and 29.0302. By comparing these standard deviations, Economics has the largest variability.

(b) (2 points)



(c) (5 points)

The correlation between Statistics and Economics is 0.8745, followed by the correlation between Accounting and Statistics is 0.6505, followed by the correlation between Accounting and Economics is 0.5916. Statistics tends to have greater impact on Accounting than Economics. Economics tends to have grater impact on Statistics than Accounting. Statistics tends to have greater impact on Economics than Accounting.

(d) (5 points)

The means of Accounting and Statistics are 81.6734 and 70.9795.

We have (1) 16 number of students, (2) 6 number of students, (3) 7 number of students, (4) 20 number of students. Hence, the proportion of students who are below and above average in both subjects to all students is 0.7346. It shows whether one is above or below average in Accounting when one is above or below average in Statistics. The relationship shares similar idea of correlation between Accounting and Statistics.

3. (10 points)

(a) (3 points)

Given $n = 5$, $\mu_x = np = 2$, $\sigma_x = \sqrt{np(1-p)} = 1.0954$.

(b) (3 points)

Given $n = 50$, we apply Central Limit Theorem, $\Pr(X_n \leq 25) = \Pr(\hat{p} \leq 0.5)$.

$\mu_{\bar{x}} = p = 0.4$, $\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}} = 0.0692$, $\Pr(\hat{p} \leq 0.5) = 0.9255$.

(c) (4 points)

Given $n = 10$, $\Pr(X_n = 0) = 0.6^{10} = 0.0060$.

4. (15 points)

(a) (3 points)

$$\frac{2}{3} \times \frac{1}{3} = 0.2222$$

(b) (3 points)

$$(\frac{2}{3})^2 \times \frac{1}{3} = 0.1482$$

(c) (4 points)

$$(\frac{2}{3})^{k-1} \times \frac{1}{3}$$

(d) (5 points)

Let $A = 1 \times (\frac{2}{3})^0 \times \frac{1}{3} + 2 \times (\frac{2}{3})^1 \times \frac{1}{3} + 3 \times (\frac{2}{3})^2 \times \frac{1}{3} + \cdots$

$= \frac{1}{3}(1 \times (\frac{2}{3})^0 + 2 \times (\frac{2}{3})^1 + 3 \times (\frac{2}{3})^2 + \cdots)$ [1]

$\frac{2}{3}A = \frac{1}{3}(1 \times (\frac{2}{3})^1 + 2 \times (\frac{2}{3})^2 + 3 \times (\frac{2}{3})^3 + \cdots)$ [2]

$\frac{1}{3}A = \frac{1}{3}((\frac{2}{3})^0 + (\frac{2}{3})^1 + (\frac{2}{3})^2 + (\frac{2}{3})^3 + \cdots)$ [3] = [1] – [2]

$A = (\frac{2}{3})^0 + (\frac{2}{3})^1 + (\frac{2}{3})^2 + (\frac{2}{3})^3 + \cdots = \frac{1}{1-\frac{2}{3}} = 3$ [4] = [3] × 3

Note that if you counted out $1\times(\frac{2}{3})^0\times\frac{1}{3}+2\times(\frac{2}{3})^1\times\frac{1}{3}+3\times(\frac{2}{3})^2\times\frac{1}{3}+\cdots+$ $k\times(\frac{2}{3})^{k-1}\times\frac{1}{3}$ for a large enough *k*, you may get a number approximate to 3. (You get full points as well if you clearly explained by doing so!)

5. (15 points)

(a) (5 points)

Class	Observed frequency	Theoretical frequency
[0, 100)	504	142.8571
[100, 200)	273	142.8571
[200, 300)	109	142.8571
[300, 400)	58	142.8571
[400, 500)	30	142.8571
[500, 600)	17	142.8571
[600, 700)	9	142.8571

Not appropriate. The observed frequency is larger in the first class and then decrease, but the theoretical frequency is the same for every class.

(b) (10 points)

Class	Observed frequency	Theoretical frequency
[40, 50)	1	1.0486
[50, 60)	8	6.6593
[60, 70)	11	16.7258
[70, 80)	17	16.7258
[80, 90)	10	6.6593
[90, 100)	2	1.0486

By comparing them, it is okay to say that Statistics grades follow a normal distribution with mean 70 and standard deviation 10 since the two frequencies are similar.

6. (25 points)

(a) (3 points)

Mean $= \frac{1+2+3+4+5+6}{6} = 3.5$

Std $= \sqrt{\frac{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{6}} = 1.7078$

(b) (3 points)

X_i	Probability
1	0.1667
2	0.1667
3	0.1667
4	0.1667
5	0.1667
6	0.1667

(c) (3 points)

Mean = 3.5

Std $= \frac{1.7078}{\sqrt{2}} = 1.2076$

(d) (5 points)

$\overline{x}_2$	Probability
1	0.0277
1.5	0.0555
2	0.0833
2.5	0.1111
3	0.1388
3.5	0.1666
4	0.1388
4.5	0.1111
5	0.0833
5.5	0.0555
6	0.0277

(e) (3 points)

$\text{Std} = \frac{1.7078}{\sqrt{100}} = 0.17078$

By using Central Limit Theorem, since $n = 100 > 30$,

we have $\overline{x}_{100} \sim \text{ND}(3.5, 0.17078)$.

(f) (3 points)

$\Pr(\overline{x}_{100} > 3.6) = 1 - \Pr(\overline{x}_{100} \leq 3.6) = 0.2791$

(g) (5 points)

$\Pr(\overline{x}_{100} \leq 2.5) = 0.000000009$. The probability is really small to get a sample mean not larger than 2.5, so the dice might be unfair.

7. (25 points)

(a) True.

(b) False.

(c) False.

(d) True.

(e) True.