Probability and Statistics in Engineering, Fall 2016

Exercise #1

1. A study of 250 households in Taipei showed that a household produced an average of 4 pounds of garbage per day.
   (1) What is the sample in this study?
   (2) How many observations are there?
   (3) What is the variable of interest?
   (4) What is the population?
   (5) Discuss the role of statistical inference in the context of this example.

Sol:

(1) the 250 households

(2) $n = 250$

(3) garbage per day

(4) all of the households in Taipei

(5) 我們用 250 個家戶單位每日垃圾量的樣本平均數 $\overline{X} = 4$，統計推論台北市全體家戶單位每日垃圾量的母體平均數 μ。

2. For each of the following examples of data, indicate the measurement scale that is appropriate.
   (1) the starting salaries of graduates from a Mechanical Engineering program
   (2) the month of highest sales for each firm in a sample
   (3) the weekly closing price of gold throughout the year
   (4) the size of soft drink (small, medium, or large) ordered by a sample of Burger King customers
   (5) method of payment (cash, check, credit card)
   (6) the time to start a baseball game
   (7) the final letter grades received by students in a statistics course
   (8) the amount of crude oil imported monthly by the U.S.
   (9) the number of miles driven annually by employees in company cars
   (10) the marks achieved by the students in a statistics course final exam in which there are ten questions each worth 10 marks

Sol:

(1) ratio scale，薪資具絕對零點。

(2) nominal scale，月份僅具分類與辨識功能。

(3) ratio scale，價格具絕對零點。

(4) ordinal scale，飲料大、中、小杯可排序。

(5) nominal scale，付費方式僅爲分類。

(6) interval scale，時刻爲等距尺度，因爲午夜零時並非絕對零點。

(7) ordinal scale，成績 A, B, C, D 與 F 僅具排序功能。

(8) ratio scale，原油進口量具絕對零點。

(9) ratio scale，汽車所跑之哩程具絕對零點。

(10) ratio scale，統計學考試成績具絕對零點。

3. Let $\overline{X} = 3600.7$ and $S^2 = 14655$ be the sample mean and variance of the weights of nineteen neonates in grams. Suppose that there was a typewriting error in the eighth observation, the accurate weight of the eighth neonate, 4210 was recorded as 4120. After we correct it, what are the accurate sample mean and variance?

Sol:

$$\sum_{i=1}^{19} X_i = 19\overline{X} = 19 \times 3600.7 = 68413.3$$

$$\Rightarrow \sum_{i=1}^{19} Y_i = 68413.3 - 4120 + 4210 = 68503.3 \quad \therefore \overline{Y} = \frac{1}{19}\sum_{i=1}^{19} Y_i = 3605.437$$

$$\sum_{i=1}^{19} X_i^2 = (19-1)S_X^2 + 19\overline{X}^2 = 246599559.3$$

$$\Rightarrow \sum_{i=1}^{19} Y_i^2 = \sum_{i=1}^{19} X_i^2 - 4120^2 + 4210^2 = 247349259.3$$

$$\therefore S_Y^2 = \frac{1}{19-1}\left[\sum_{i=1}^{19} Y_i^2 - \frac{\left(\sum_{i=1}^{19} Y_i\right)^2}{19}\right] = 20274.316$$

4. A random sample of the Statistics scores for 100 freshmen in a university was selected, and the frequency distribution was obtained as follows. Given the mean and the standard deviation of the sample equal to 71.4 and 11.5, respectively, find the following statistics.
   (1) the median of the data set
   (2) the mode of the data set using Pearson's method, King's method, and Czuber's method
   (3) the Pearson coefficient of skewness

Score boundaries	Frequency
30-40	1
40-50	3
50-60	13
60-70	21
70-80	40
80-90	20
90-100	2

Sol:

(1) $m_e = \mathrm{L} + \left(\frac{\frac{n}{2} - \mathrm{F}_{-1}}{f_{m_e}} \right) \times h = 70 + \left(\frac{\frac{100}{2} - 38}{40} \right) \times 10 = 73$

(2) ① Pearson's method： $m_o = 3m_e - 2\overline{\mathrm{X}} = 3 \times 73 - 2 \times 71.4 = 76.2$

② King's method： $m_o = \mathrm{L} + \left(\frac{f_{+1}}{f_{-1} + f_{+1}} \right) \times h = 70 + \left(\frac{20}{21 + 20} \right) \times 10 = 74.878$

③ Czuber's method： $m_o = \mathrm{L} + \left(\frac{\Delta_{-1}}{\Delta_{-1} + \Delta_{+1}} \right) \times h = 70 + \left(\frac{19}{19 + 20} \right) \times 10 = 74.872$

(3) $sk_{\mathrm{P}} = \frac{3(\overline{\mathrm{X}} - m_e)}{\mathrm{S}} = \frac{3 \times (71.4 - 73)}{11.5} = -0.41739$

5. A large manufacture company believes that their hourly wages follow a normal probability distribution. To confirm this, 300 workers were sampled and the results organized into the following frequency distribution.
   (1) What are the mean, median, and standard deviation of these data.
   (2) Find the coefficient of skewness for these data and interpret it.
   (3) Find the coefficient of kurtosis for these data and interpret it.
   (4) Based on what you have got from the above, do you agree that the hourly wages follow a normal distribution?

Sol:

(1) $n=300$, $\overline{X}=\frac{1}{300}\sum_{i=1}^{5} f_i X_i = 8.1$

$$m_e = L+\left(\frac{\frac{n}{2}-F_{-1}}{f_{m_e}}\right)\times h = 7.5+\left(\frac{\frac{300}{2}-74}{130}\right)\times 1 = 8.0846$$

$$S^2=\frac{1}{300-1}\left(\sum_{i=1}^{5} f_i X_i^2 - 300\overline{X}^2\right)=1.0401 \quad \therefore S=\sqrt{S^2}=1.0199$$

(2) $m_1'=\overline{X}=8.1$, $m_2'=\frac{1}{300}\sum_{i=1}^{5} f_i X_i^2 = 66.6467$, $m_3'=\frac{1}{300}\sum_{i=1}^{5} f_i X_i^3 = 556.58$

$$m_3 = m_3'-3m_2'm_1'+2(m_1')^3 = -0.0520 \quad \therefore \alpha_3=\frac{m_3}{S^3}=-0.0490<0,\ \text{左偏分配}$$

(3) $m_4'=\frac{1}{300}\sum_{i=1}^{5} f_i X_i^4 = 4714.0067$

$$m_4 = m_4'-4m_3'm_1'+6m_2'(m_1')^2-3(m_1')^4 = 2.9252 \quad \therefore \alpha_4=\frac{m_4}{S^4}=2.7035<3,\ \text{低闊峰}$$

(4) The data distribution is a negatively skewed and platykurtic distribution. It is not a like a normal distribution.

6. Consider a sample of 12 measurements:

   1, 1, 0, 15, 2, 3, 4, 0, 1, 3, 1, 5

   (1) Calculate the range, sample mean, variance, and standard deviation.

   (2) Find the median, lower quartile, and upper quartile for the data.

   (3) Construct a box-and-whisker plot for the data and identify any outliers.

Sol:

(1) $R=15-0=15$, $\overline{X}=\frac{1}{12}\sum_{i=1}^{12} X_i = 3$

$$S^2=\frac{1}{12-1}\left[\sum_{i=1}^{12} X_i^2 - 12\overline{X}^2\right]=16.7273 \quad \therefore S=\sqrt{S^2}=4.0899$$

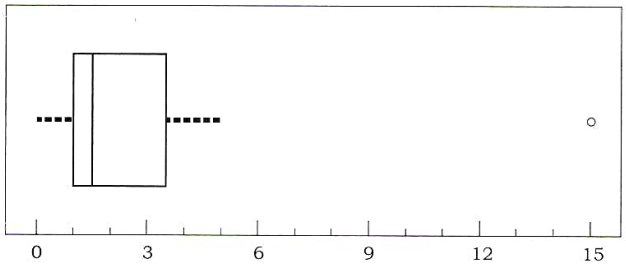
(2) 排序：0, 0, 1, 1, 1, 1, 2, 3, 3, 4, 5, 15

$$m_e=\frac{X_{(6)}+X_{(7)}}{2}=\frac{1+2}{2}=1.5$$

$$i=12\times\frac{1}{4}=3 \Rightarrow Q_1=\frac{X_{(3)}+X_{(4)}}{2}=\frac{1+1}{2}=1$$

$$i=12\times\frac{3}{4}=9 \Rightarrow Q_3=\frac{X_{(9)}+X_{(10)}}{2}=\frac{3+4}{2}=3.5$$

(3)



其中 $X_{(12)}=15$ 爲 extreme outlier

7. Example 2.4 in Chapter 2 - Descriptive Statistics.

Sol:

- 計算報酬率的幾何平均數時, 應先將資料轉換爲**毛報酬率 (gross rate of return)**,[17] 再計算其毛報酬率的幾何平均數, 最後減 1 轉換回淨報酬率的幾何平均數。當報酬率爲時間序列的年資料時, 計算淨報酬率的幾何平均數, 即代表資料所跨時間的**年化持有期間報酬率 (annualized holding period return)**。

  這位基金經理人所採用的是他過去三年報酬率表現的算術平均數

  $$\overline{R}_A = \frac{1}{3}\sum_{i=1}^{3} R_i = \frac{25\% + (-45\%) + 90\%}{3} = 23\%$$

  事實上, 連續數年報酬率的平均數採用算術平均數將會有高估的現象, 宜採用幾何平均數

  $$\overline{R}_G = \sqrt[3]{\prod_{i=1}^{3}(1+R_i)} - 1 = \sqrt[3]{1.25 \times 0.55 \times 1.9} - 1 = 1.093 - 1 = 0.093 = 9.3\%$$

  他過去三年的平均報酬率爲 9.3% (年化持有期間報酬率), 並未有擊敗市場大盤的表現。