CENG/NANO 114

Probability and Statistical Methods for Engineers

Problem Set 4

Note: Questions denoted with a * are meant to be more difficult than usual. Hints are provided for a selected number of questions.

1. An experiment was conducted to test the yield strength of circular tubes with caps welded to the ends. The first yields (in kN) are 96, 96, 102, 102, 102, 104, 104, 108, 126, 126, 128, 128, 140, 156, 160, 160, 164, and 170.
   i. Calculate the IQR and mode.
   ii. Calculate the sample mean and sample standard deviation.

2. The following data are the joint temperatures of the O-rings (° F) for each test firing or actual launch of the space shuttle rocket motor (from *Presidential Commission on the Space Shuttle Challenger Accident*, Vol. 1, pp. 129 – 131): 84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.
   i. Construct a frequency distribution diagram.
   ii. Calculate and show the percentiles with relative frequency and cumulative frequency in your diagram.

3. With the growth of internet service providers, a researcher decides to examine whether there is a correlation between cost of internet service per month (rounded to the nearest dollar) and degree of customer satisfaction (on a scale of 1 - 10 with a 1 being not at all satisfied and a 10 being extremely satisfied). The researcher only includes programs with comparable types of services. A sample of the data is provided below.

dollars	11	18	17	15	9
satisfaction	6	8	10	4	9
dollars	5	12	19	22	25
satisfaction	6	3	5	2	10

Determine the strength of relationship using correlation coefficient r.

4. A group of students want to repeat single pendulum tests to verify the equation

$$T = 2\pi\sqrt{\frac{L}{g}}$$

for the period of oscillation, where T is the period in seconds, L is the length of single pendulum in meters. We assume g is 9.8 ms^{-2} for this experiment. To reduce the error made by the delay of human reaction, these students measured the time based on five periods for their data, i.e., each of the following data is the time of five periods for the pendulum motion. They finished 10 groups of tests and obtained the following data:

7.12s, 7.45s, 7.01s, 6.98s, 7.02s, 7.08s, 6.86s, 7.04s, 7.12s, 6.99s

The length of single pendulum used is 0.5m. For the following questions, the mean and standard deviation refer to the data itself, i.e., for five periods.

i. What is the sample mean for the data?
ii. What is the standard deviation if they use the "sample mean" for calculation?
iii. What is the standard deviation if they use the "true value" of mean calculated from the period equation above?

5. Suppose that X and Y represent the thickness in micrometers of a substrate and an active layer of a chemical product, respectively. Assume that X and Y are independent and normally distributed with $\mu_x = 10000$, $\mu_y = 1000$, $\sigma_x = 250$, $\sigma_y = 20$, respectively. Now you want to verify the value, so you choose 10 samples of this chemical product and measure their thicknesses of substrate layer and active layer.

X(μm)	10110	10129	9988	9890	10234	10145	9887	10218	10256	10071
Y(μm)	999	978	993	1011	1005	1001	996	984	1012	1015

i. What is the mean for the sample of X and Y? What is the mean for X+Y?
ii. Calculate the standard deviation for X+Y. (Hint: Use the population standard deviation for X and Y given in this problem.)
iii. What is the z-score of 11039 μm for X+Y?

6. An article presented the following data on the motor fuel octane ratings of several blends of gasoline:

88.5 98.8 94.7 88.3 84.3 90.4 90.1 91.2 89.0 90.6 89.8 92.2 91.6 87.7 90.3 91.1 90.0 86.7 91.5 93.4 89.9 96.1 89.6 92.2 90.4 83.4 91.6 91.0 90.7 88.2 88.6 88.5 88.3 93.3 94.2 87.4 85.3 91.1 90.1 90.5 89.3 100.3 91.1 87.6 92.7 88.4 87.9 92.6 93.0 93.7 94.4 96.5 90.4 84.3 91.2 93.2 86.7 88.6 94.2 88.7 90.8 92.7 90.1 89.3 91.8 91.0 87.5 90.9 87.8 89.9 88.3 91.8 89.2 89.7 92.3 92.2 88.9 89.8 92.7 93.3 86.7 91.0

i. Construct the stem and leaf display of these data.
ii. Construct the frequency distribution with 8 bins.
iii. Construct the histogram with 8 bins.

(Hint: data also available in p6_data.txt, try to rank the data on computer)

7. Use data from the above question.
   i. Calculate the range and inter-quartile range (IQR).
   ii. Calculate the sample median, mode and mean.
   iii. Calculate the relative frequency, cumulative frequency and relative cumulative frequency for median and the smallest mode.
   iv. Explain how these three measures of location describe different features of the data.

8. An article in the *Tappi* Journal (March 1986) presented data on green liquor Na2S concentration (in grams per liter) and paper machine production (in tons per day). The data (read from a graph) follow:

y	40	42	49	46	44	48	
x	825	830	890	895	890	910	
y	46	43	53	52	54	57	58
x	915	960	990	1010	1012	1030	1050

i. Fit a simple linear regression model with y = green liquor Na_2S concentration and x = production.
ii. Find the standard error of estimate $S_{Y|X}$. Draw a scatter diagram of the data and the resulting least squares fitted model.
iii. Find the fitted value of y corresponding to x = 915 and the associated residual.
iv. Find the mean green liquor Na_2S concentration when the production rate is 940 tons per day.

9.* The rate of a reaction is given by the following relationship:

$$r = ce^{-\frac{E}{RT}}[A]^s$$

where r is the reaction rate, E is the activation energy, R is gas constant and [A] is the concentration of reactant A, c is a constant and s is known as the order of the reaction. Derive a linearized equation that will allow you to determine the unknown constant c, s and E given a series of data on reaction rates, temperatures and concentrations. State clearly what variables will need to be calculated and how the unknown constants will be derived from the output of a multiple least squares regression. Write down the form of the matrices needed for the multiple least squares. Hint: One of the columns of the matrix is always 1.

10. The following is CO2 levels in ppm (X) and changes in the sea level in mm (Y) for the period 2001-2010 downloaded from NASA's website (http://climate.nasa.gov/vital-signs, only the last data point for Dec is used to make this problem easier to solve without using a computer).

Year	CO2 level (ppm)	Sea Level Change (mm)
2001	371.18	18.88
2002	373.71	22.22
2003	375.93	24.47
2004	377.45	29.1
2005	379.92	32.56
2006	381.79	32.91
2007	383.89	32.82
2008	385.56	36.38
2009	387.31	42.84
2010	389.73	36.78

Answer the following questions:

i. Estimate the CO_2 level in ppm in the year 2020, assuming the current rate of CO_2 increase remains the same.

ii. It can be assumed that there is a linear relationship between CO_2 levels and changes in the sea level.

$$Y = bX + a$$

Determine the parameters b and a.

iii. What is the standard error of the estimate and coefficient of determination for part (ii)? What is the correlation coefficient?

iv. What would be the estimated sea level if CO_2 reaches 500 ppm?