Solution for CENG/NANO 114 Problem Set 4

1 # of data = 18

96, 96, 102, 102, 102, 104, 104, 108, 126, 126, 128, 128, 140, 156, 160, 160, 164, 170

↑ 1st quartile ↑ 3rd quartile

i. $x = \frac{18+1}{4} = 4.75 \approx 5$

$IQR = 156 - 102 = 54$

Mode is 102 (most frequently occurring value).

ii Sample mean: $\bar{X} = \frac{\sum_{i=1}^{18} X_i}{N} = \frac{2272}{18} = 126.22$

Sample standard deviation $SS = \Sigma(X-\bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$

$= 298392 - \frac{2272^2}{18} = 11615.11$

$\sigma^2 = \frac{SS}{N-1}$

$\sigma = \sqrt{\frac{SS}{N-1}} = \sqrt{\frac{11615.11}{17}} = 26.14$

2. i. Construct a frequency distribution diagram

Min = 31 Max = 84, range (31, 84)

If desired number of classes is 9.

we have $\frac{84-31}{9} = 5.89 \approx 6$

So the length of interval should be 6.

Class	(30, 36]	(36, 42]	(42, 48]	(48, 54]	(54, 60]	(60, 66]	(66, 72]	(72, 78]	(78, 84]
Freq	1	1	1	3	4	4	11	6	5
Rel Freq	0.0278	0.0278	0.0278	0.0833	0.1111	0.1111	0.3056	0.1667	0.1389
Cumulative Freq	1	2	3	6	10	14	25	31	36
Rel Cumulative Freq	0.0278	0.0556	0.0833	0.1667	0.2778	0.3889	0.6944	0.8611	1

2 ii Freq : # of value in the interval

Rel Freq : $\frac{\text{\# of value in the interval (Freq)}}{\text{Total number of sample (36)}}$

Cumulative Freq: number of value in each class and all lower ranked classes

Rel Cumulative Freq : $\frac{\text{Cumulative Freq}}{\text{Total number of sample (36)}}$

All data are shown in the frequency distribution diagram (Freq, Rel Freq, Cumulative Freq, Rel Cumulative Freq).

3 Define dollars as X, satisfaction as Y.

From the data in form, we get

$\Sigma X = 153 \quad \Sigma Y = 63 \quad N = 10$

$\Sigma X^2 = 2679 \quad \Sigma Y^2 = 471 \quad \Sigma XY = 976$

$SP = \Sigma(X-\bar{X})(Y-\bar{Y}) = \Sigma XY - \frac{\Sigma X \Sigma Y}{N} = 976 - \frac{153 \times 63}{10} = 12.1$

$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 2679 - \frac{153^2}{10} = 338.1$

$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 471 - \frac{63^2}{10} = 74.1$

$r = \frac{cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{SP}{\sqrt{SS_x\,SS_y}} = \frac{12.1}{\sqrt{338.1 \times 74.1}} = 0.0764 < 0.5$

Since the correlation coefficient r is much smaller than 0.5, it shows that there is not any strong relationship between the cost of internet service and degree of customer satisfaction.

4 i. Ten data are:

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

sample mean $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 7.067$

ii. If use "sample mean" to calculate standard deviation

Sample standard deviation : $\sigma = \sqrt{\frac{SS}{N-1}}$

where $SS = \Sigma(X-\bar{X})^2$

$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 499.6399 - \frac{70.67^2}{10} = 0.215$

$\sigma^2 = \frac{SS}{N-1} = \frac{0.215}{9} = 0.02389$

$\sigma = 0.1546$

iii. If use "true value" of mean (population mean) to solve it.

$SS = \Sigma(X-\mu)^2$ μ: population mean

$\sigma = \sqrt{\frac{SS}{N}}$

Firstly, get μ for SS from equation $T = 2\pi\sqrt{\frac{L}{g}}$

μ is 5 times of period from the question.

$\therefore \mu = 5 \times 2\pi\sqrt{\frac{L}{g}} = 5 \times 2\pi\sqrt{\frac{0.5}{9.8}} = 7.096$ s

$SS = \Sigma(X-\mu)^2 = \sum_{i=1}^{10}(X_i - 7.096)^2 = 0.2234$

$\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{0.2234}{10}} = 0.1495$

5 i. sample mean $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 10092.8$

$\bar{Y} = \frac{1}{10}\sum_{i=1}^{10} Y_i = 999.4$

mean for $X+Y = \bar{X} + \bar{Y} = 10092.8 + 999.4 = 11092.2$

ii. $Var(X+Y) = Var(X) + Var(Y) = \sigma_x^2 + \sigma_y^2$

$\sigma_{X+Y} = \sqrt{Var(X+Y)} = \sqrt{\sigma_x^2 + \sigma_y^2} = \sqrt{250^2 + 20^2} = 250.8$

Note: If you are not given the data of population mean (μ_x, μ_y) and population standard deviation (σ_x, σ_y), this problem is also doable.

When you are only given ten groups of sample values, you can solve it as the following process:

sample data $\Rightarrow$ get $\bar{X}$, $\bar{Y}$, SS_x, SS_y $\Rightarrow$ get standard deviation of X & Y (σ_x, σ_y)

$\sigma_x^2 = \frac{SS_x}{N-1}$, $\sigma_y^2 = \frac{SS_y}{N-1}$ $\Rightarrow$ From central limit theorem, we can get that sample average $S_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, $Var = \frac{\sigma^2}{n}$

$\Rightarrow \sigma_{X+Y} = \sqrt{Var(X) + Var(Y)} = \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}}$

Here n is the number of sample, σ_x, σ_y are standard deviation based on finite samples.

iii. $Z = \frac{(X+Y) - \mu_{X+Y}}{\sigma_{X+Y}}$, $\mu_{X+Y} = \mu_x + \mu_y = 10000 + 1000 = 11000$

$= \frac{11039 - 11000}{250.8}$

$= 0.1555$

Note: Since you are given population mean, just use them.

If you don't know population mean, use sample mean to replace population mean for calculation.

6. # of data is 82

Min = 83.4 Max = 100.3

i. After ranking the data, we can construct the stem and leaf diagram.

#	stem	leaves (x0.1)
1	83	4
2	84	3 3
1	85	3
3	86	7 7 7
6	87	4 5 6 7 8 9
11	88	2 3 3 3 4 5 5 6 6 7 9
10	89	0 2 3 3 6 7 8 8 9 9
13	90	0 1 1 1 3 4 4 4 5 6 7 8 9
13	91	0 0 0 1 1 1 2 2 5 6 6 8 8
8	92	2 2 2 3 6 7 7 7
6	93	0 2 3 3 4 7
4	94	2 2 4 7
2	96	1 5
1	98	8
1	100	3

6 ii. length of each class $= \frac{100.3 - 83.4}{8} = 2.1125$

class	[83.4, 85.51)	[85.51, 87.63)	[87.63, 89.74)	[89.74, 91.85)
Freq	4	6	20	30
Rel Freq	0.049	0.073	0.244	0.366
Cumulative Freq	4	10	30	60
Rel Cumulative Freq	0.049	0.122	0.366	0.732
class	[91.85, 93.96)	[93.96, 96.08)	[96.08, 98.19)	[98.19, 100.3]
Freq	14	4	2	2
Rel Freq	0.171	0.049	0.024	0.024
Cumulative Freq	74	78	80	82
Rel Cumulative Freq	0.902	0.951	0.976	1

iii. histogram is shown on next page.

7 i. $x = \frac{82+1}{4} = 20.75 \approx 21$

1st quartile = 88.6

3rd quartile = 92.2

IQR = 92.2 − 88.6 = 3.6

ii. mean = 90.53 median = 90.4 (position #41, position #42: $\frac{90.4 + 90.4}{2} = 90.4$)

mode : 86.7, 88.3, 90.1, 90.4, 91.0, 91.1, 92.2, 92.7

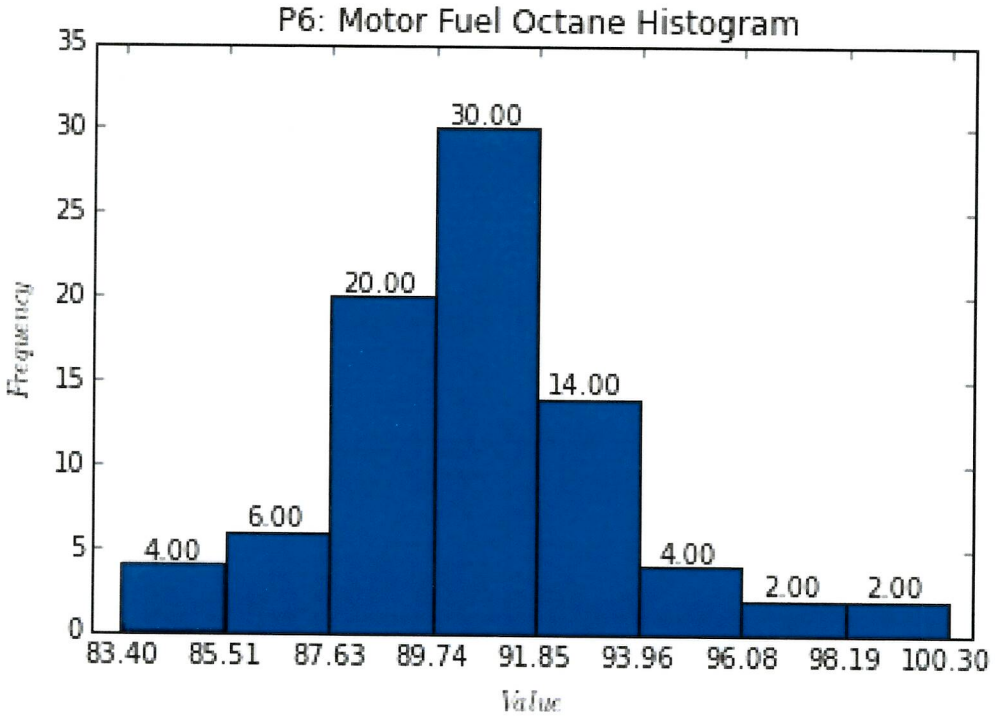

Diagram of Problem 6 (iii).

7. iii. mean = 90.4 , smallest mode = 86.7

	90.4	86.7
Rel. Freq	0.0366	0.0366
Cumulative Freq	42	7
Rel Cumulative Freq	0.512	0.0854

iv. mode : describe the value of the most frequently occuring

median : middle value when observations are ordered from least to most (50th percentile ranked value).

mean : sum of all values divided by number of values.

8. i. $\bar{X} = 939 \quad \bar{Y} = 48.615$

$\overline{X^2} = 886878.4 \quad \overline{XY} = 46007.5$

$Y = bX + a \quad , \quad b = \frac{SP}{SS_x} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\overline{X^2} - \bar{X}^2} = \frac{46007.5 - 939 \times 48.615}{886878.4 - 939^2}$

$= 0.0694$

$a = \bar{Y} - b\bar{x} = 48.615 - 0.0694 \times 939 = -16.5$

So the linear regression equation should be:

$Y = 0.0694X - 16.5$

ii. $S_{Y|X} = \sqrt{\frac{SS_y(1-r^2)}{n-2}}$

$r = \frac{SP}{\sqrt{SS_x SS_y}}$ $\quad SP = \Sigma XY - \frac{\Sigma X \Sigma Y}{N} = 598098 - \frac{12207 \times 632}{13} = 4650$

$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 1152949 - \frac{12207^2}{13} = 67046$

$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 31128 - \frac{632^2}{13} = 403.1$

$\therefore r = \frac{4650}{\sqrt{67046 \times 403.1}} = 0.8945$

$S_{Y|X} = \sqrt{\frac{403.1 \times (1 - 0.8945^2)}{13-2}} = 2.706$

Scatter data points and fitting linear diagram are shown on next page.

iii. $X = 915 \quad , \quad Y = 0.0694 \times 915 - 16.5 = 47.0$

From the data chart: $Y' = 46$

$Y - Y' = 47.0 - 46 = 1$ (associated residual)

iv. $X = 940 \quad , \quad Y = 0.0694 \times 940 - 16.5 = 48.7$

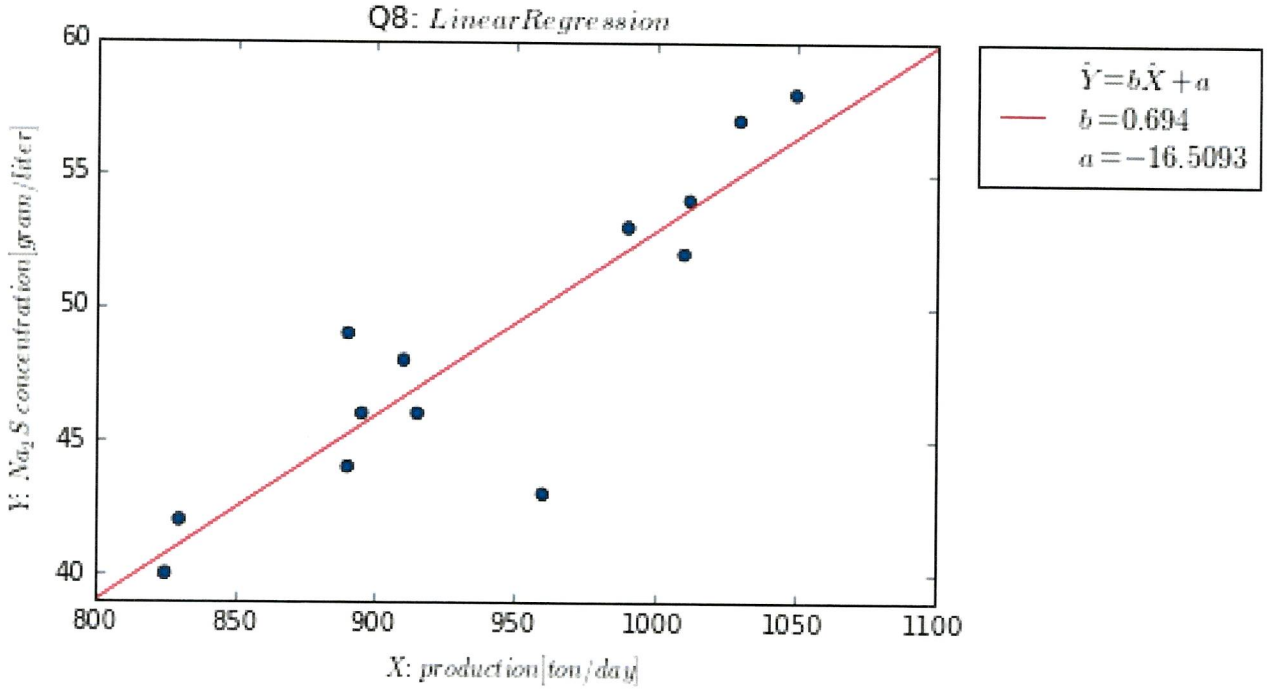

Diagram for Problem 8 (ii).

Solutions for PS4 Q9 and Q10

February 22, 2015

1 Q9

To linearize this equation, we first need to take the natural log of the equation, resulting in

$$\ln r = \ln c - \frac{E}{RT} + s \ln [A]$$

Let us now define the following variables:

$$y = \ln r$$
$$x_1 = \ln [A]$$
$$x_2 = \frac{1}{T}$$
$$b_1 = s$$
$$b_2 = -\frac{E}{R}$$
$$b_3 = \ln c$$

This will allow us to rewrite the original rate equation in the following form:

$$y = [x_1, x_2, 1] \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

Given a series of n observations of reaction rates with temperature and concentration of A, we can then construct the relationship as

$$Y = X\beta + \epsilon$$

with

$$X = \begin{pmatrix} x_{11} & x_{12} & 1 \\ x_{21} & x_{22} & 1 \\ ... & ... & 1 \\ x_{n1} & x_{n2} & 1 \end{pmatrix}$$

We can then estimate β using

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

and substitute the values in the original defintions to obtain c, E and s.

2 Q10

Let us just enter the data into a Python Data Analysis (pandas) library DataFrame object. For the sake of understanding, we will laboriously go through each of the steps to answer the questions. We will see a demonstration of how to do the analysis much more quickly in the last part. Note that using such statistical packages is a necessity in real life when dealing with lots of data. I recommend learning basic Python + Pandas + iPython notebooks. It is free and far more powerful. But you could have just as easily used commercial Matlab or R (or horrors, Excel) to do the same thing. But you must understand the concepts and you will not have access to these tools in the exams.

```
In [1]: data_str = """Year        CO2        SeaLvlChg
        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"""
        from cStringIO import StringIO
        from pandas import read_table
        from math import sqrt
        from IPython.display import Latex

        data = read_table(StringIO(data_str))
        print data
```

```
Year     CO2  SeaLvlChg
0  2001  371.18      18.88
1  2002  373.71      22.22
2  2003  375.93      24.47
3  2004  377.45      29.10
4  2005  379.92      32.56
5  2006  381.79      32.91
6  2007  383.89      32.82
7  2008  385.56      36.38
8  2009  387.31      42.84
9  2010  389.73      36.78
```

Let X, Y and Z be the CO2 level, Sea level change, and year respectively. Let's now compute the necessary X^2, Y^2, Z^2, XY and XZ first. We will be using these shortly.

```
In [2]: data["X^2"] = data["CO2"] ** 2
        data["Y^2"] = data["SeaLvlChg"] ** 2
        data["Z^2"] = data["Year"] ** 2
        data["XY"] = data["CO2"] * data["SeaLvlChg"]
        data["XZ"] = data["CO2"] * data["Year"]

        print data
```

```
Year     CO2  SeaLvlChg          X^2        Y^2      Z^2          XY  \
0  2001  371.18      18.88  137774.5924   356.4544  4004001   7007.8784
1  2002  373.71      22.22  139659.1641   493.7284  4008004   8303.8362
```

```
2  2003  375.93      24.47  141323.3649   598.7809  4012009   9199.0071
3  2004  377.45      29.10  142468.5025   846.8100  4016016  10983.7950
4  2005  379.92      32.56  144339.2064  1060.1536  4020025  12370.1952
5  2006  381.79      32.91  145763.6041  1083.0681  4024036  12564.7089
6  2007  383.89      32.82  147371.5321  1077.1524  4028049  12599.2698
7  2008  385.56      36.38  148656.5136  1323.5044  4032064  14026.6728
8  2009  387.31      42.84  150009.0361  1835.2656  4036081  16592.3604
9  2010  389.73      36.78  151889.4729  1352.7684  4040100  14334.2694
```

```
          XZ
0  742731.18
1  748167.42
2  752987.79
3  756409.80
4  761739.60
5  765870.74
6  770467.23
7  774204.48
8  778105.79
9  783357.30
```

2.1 Q10i

To solve part i, we need to do a regression of Z (Year) vs X (CO2 level), based on

$$X = bZ + a$$

Let's first compute the relevant SP_{xz} and SS_z. We then have

$$b = \frac{SP_{xz}}{SS_z}$$

$$a = \bar{X} - b\bar{Z}$$

In [3]:
```
sums = data.sum()
SS_z = sums["Z^2"] - sums["Year"] ** 2 / 10
SP_xz = sums["XZ"] - sums["CO2"] * sums["Year"] / 10
b = SP_xz / SS_z
a = sums["CO2"] / 10 - b * sums["Year"] / 10
Latex("$$SS_z = %.1f$$$$SS_{xz} = %.1f$$$$b = %.2f$$$$a = %.1f$$" % (SS_z, SP_xz, b, a))
```

Out[3]:

$$SS_z = 82.5$$

$$SS_{xz} = 165.7$$

$$b = 2.01$$

$$a = -3648.5$$

The projected CO2 in year 2020 is then given by

In [4]:
```
Latex("CO$_2$ in 2020 = %.1f" % (b * 2020 + a))
```

Out[4]:

CO_2 in 2020 = 409.8

2.2 Q10ii

For part ii, we do the same linear regression, but now with X and Y instead of X and Z.

```
In [5]: SS_x = sums["X^2"] - sums["CO2"] ** 2 / 10
        SP_xy = sums["XY"] - sums["CO2"] * sums["SeaLvlChg"] / 10
        b = SP_xy / SS_x
        a = sums["SeaLvlChg"] / 10 - b * sums["CO2"] / 10
        Latex("$$SS_x = %.1f$$$$SS_{xy} = %.1f$$$$b = %.2f$$$$a = %.1f$$" % (SS_x, SP_xy, b, a))
```

Out[5]:

$$SS_x = 333.6$$
$$SS_{xy} = 377.3$$
$$b = 1.13$$
$$a = -399.6$$

2.3 Q10iii

The coefficient of determination is given by the square of the correlation coefficient r, where

$$r = \frac{SP}{\sqrt{SS_x SS_y}}$$

```
In [6]: SS_y = sums["Y^2"] - sums["SeaLvlChg"] ** 2 / 10
        r = SP_xy / sqrt(SS_x * SS_y)
        r2 = r ** 2
        Latex("$$r = %.3f$$$$r^2=%.3f$$" % (r, r2))
```

Out[6]:

$$r = 0.941$$
$$r^2 = 0.885$$

The standard error of the estimate is given by

$$SEE = \sqrt{\frac{SS_y(1 - r^2)}{n - 2}}$$

```
In [7]: SS_y = sums["Y^2"] - sums["SeaLvlChg"] ** 2 / 10
        SEE = sqrt(SS_y * (1 - r2) / (10 - 2))
        Latex("$$SEE = %.2f$$" % SEE)
```

Out[7]:

$$SEE = 2.63$$

2.4 Q10iv

The estimated sea level for CO2 = 500ppm can be calculated simply by plugging the value into our regression equation.

```
In [8]: est_sea_level = b * 500 + a
        print "Estimated Sea Level at 500 ppm CO2 = %.1f mm" % est_sea_level
```

```
Estimated Sea Level at 500 ppm CO2 = 165.9 mm
```

2.5 Prologue

Using the Python Data Analysis (pandas) library, you can easily perform the same analysis without having to go through all the math. This is a necessity when you are working with hundreds or even thousands of data points. To demonstrate, this is how you use pandas to get all the results in part ii-iv. It is practically three lines of code to do the same analysis!!

```
In [9]: from pandas import ols
        model = ols(y=data["SeaLvlChg"], x=data["CO2"])
        print model
```

```
-------------------------Summary of Regression Analysis-------------------------

Formula: Y ~ <x> + <intercept>

Number of Observations:         10
Number of Degrees of Freedom:   2

R-squared:         0.8852
Adj R-squared:     0.8708

Rmse:              2.6303

F-stat (1, 8):    61.6789, p-value:     0.0000

Degrees of Freedom: model 1, resid 8

-----------------------Summary of Estimated Coefficients------------------------
      Variable       Coef    Std Err     t-stat    p-value    CI 2.5%   CI 97.5%
--------------------------------------------------------------------------------
             x     1.1310     0.1440       7.85     0.0000     0.8487     1.4132
     intercept  -399.6056    54.8222      -7.29     0.0001  -507.0571  -292.1540
---------------------------------End of Summary---------------------------------
```

You can verify for yourself that this gave the same answer. Note that the SEE is called the RMSE in the model summary. The coefficients are in the second column of the table.