Problem Set 2

Statistics - NYU, Summer 2016
Ercan Karadas

Section 1

[1] U.S. population in 1998 was 227,224,682 and in 2012 it was 313,877,061. Furthermore it is projected that the population will be 341,387,000 in 2020.

a) Calculate the average population growth rate between 1998 and 2012.

b) Now suppose that estimated population growth rate for 2013 is 0.9 %. What should be the average population growth rate between 2014 and 2020 in order the projection to hold?

[2] Recall that in Problem Set 1, we worked with the following data- Education and annual Family Income of 10 individuals:

Individual #	1	2	3	4	5	6	7	8	9	10
Education	18	14	15	11	16	14	17	16	12	18
Family Income	110	80	75	60	90	70	120	150	60	70

a) Calculate the variance for both of the variables.

b) Calculate the coefficient of variation for Education and Family Income, and discuss which variable varies more around its mean.

c) Calculate covariance and correlation coefficient between the two variables.

d) Interpret the relation between education and family income based on the previous part.

e) Is there a causal relation between these two variables? For example, based on these results, can we say that individuals with high family income are tend to be more educated? Explain.

[3] Consider the following paired (X, Y) data set:

	1	2	3	4	5	6	7	8	9	10	*(pair index)*
X	53	34	10	63	28	58	28	50	39	35	
Y	37	26	29	55	36	48	41	42	21	46	

a) Calculate $\overline{X}$ and $\overline{Y}$.

b) Calculate the standard deviations, s_x and s_y, and and the covariance s_{xy}.

c) Calculate the sample correlation coefficient, r_{xy}.

d) In general, the covariance or the sample correlation coefficient is a more useful measure of the relationship between two variables?

[4] For the paired data in the previous problem consider the following data transformations:

$$X \longrightarrow X + 3 \qquad \text{and} \qquad Y \longrightarrow 2Y + 3$$

In another words, we are defining two new variables:

$$X_{new} = X + 3 \qquad \text{and} \qquad Y_{new} = 2Y + 3$$

a) Complete the following table

	1	2	3	4	5	6	7	8	9	10	*(pair index)*
X_{new}	.	.	.	.	.	.	.	.	.	.	
Y_{new}	.	.	.	.	.	.	.	.	.	.	

b) Calculate $\overline{X}_{new}$ and $\overline{Y}_{new}$ and explain how they are related to $\overline{X}$ and $\overline{Y}$?

c) Calculate $s_{x_{new}}$ and $s_{y_{new}}$ and explain how they are related to s_x and s_y?

d) Similarly, calculate $s_{x_{new}y_{new}}$ and $r_{x_{new}y_{new}}$ and explain the relationship between these two statistics and s_{xy} and r_{xy} of the previous problem.

[5] Consider the sample X consisting n observations

$$X = \{x_1, x_2, \ldots x_n\}$$

and suppose that we generate a new sample Y as

$$Y = aX + b$$

a) Show that

$$\overline{Y} = a\overline{X} + b$$

b) Show that

$$s_Y^2 = a^2 s_X^2$$

$$\big(\text{i.e.,} \quad Var(Y) = a^2 Var(X)\big)$$

[6] The president of Floor Coverings Unlimited wants information concerning the relationship between retail experience (years) and weekly sales (in hundreds of dollars). He obtained the following random sample on experience and weekly sales:

$$(2,5) \quad (4,10) \quad (3,8) \quad (6,18) \quad (3,6) \quad (5,15) \quad (6,20) \quad (2,4)$$

The first number for each observation is years of experience, and the second number is weekly sales. Compute the covariance and the correlation coefficient.

[7] How much time (in minutes) do people spend on a typical visit to a local mall? A random sample of n = 104 shoppers was timed and the results (in minutes) are stored in the data file Shopping Times.xlsx (download from NYUClasses/Resources/Data). We are going to describe the shape of the distribution of shopping times both graphically and numerically.

a) Construct a histogram of these shopping times.

b) Find the mean shopping time.

c) Find the variance and standard deviation in shopping times.

d) Find the coefficient of variation.

Section 2

[8] What is the average growth rate of the sales over 5 years if

a) sales have grown 25%?

b) sales have grown 5% in the first two years and then 20% in the next three years?

[9] We can generate even more interesting new data from the original data. Again consider the data in Problem 3 and suppose that the pair (x_k, y_k) represents the number of facebook and twitter status updates of individual k per month. Furthermore, suppose that it is known that people spend 3 minutes when they update their status on facebook compare to 75 seconds on twitter.

a) Define a new variable, say Z, to denote the total time spent in status updating per month for these 10 people. In another words, after expressing Z in terms of X and Y, complete the table

	1	2	3	4	5	6	7	8	9	10	*(individual index)*
Z	.	.	.	.	.	.	.	.	.	.	

b) Suppose that another data set, say W, is generated from (x_k, y_k) pairs as

$$w_k = 6x_k + 2.5y_k$$

Calculate the covariance between Z and W. (*You do not need to do any calculation actually, just use the results obtained in the previous problems*)

[10] A corporation administers an aptitude test to all new sales representatives. Management is interested in the extent to which this test is able to predict weekly sales of new representatives. Aptitude test scores range from 0 to 30 with greater scores indicating a higher aptitude. Weekly sales are recorded in hundreds of dollars for a random sample of 10 representatives. Test scores and weekly sales are as follows:

Test Score, x	12	30	15	24	14	18	28	26	19	27
Weekly Sales, y	20	60	27	50	21	30	61	54	32	57

a) Compute the covariance between test score and weekly sales.

b) Compute the correlation between test score and weekly sales.