INTRODUCTION TO STATISTICS

LECTURE 4

BY MR THUO

4. Bivariate Data

4.1 Introduction

So far we have confined our discussion to the distributions involving only one variable. Sometimes, in practical applications, we might come across certain set of data, where each item of the set may comprise of the values of two or more variables.

A Bivariate Data is a a set of paired measurements which are of the form

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Examples

i. Marks obtained in two subjects by 60 students in a class.
ii. The series of sales revenue and advertising expenditure of the various branches of a company in a particular year.
iii. The series of ages of husbands and wives in a sample of selected married couples.

In a bivariate data, each pair represents the values of the two variables. Our interest is to find a relationship (if it exists) between the two variables under study.

4.2 Scatter Diagrams and Correlation

A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The pattern of their intersecting points can graphically show relationship patterns. Most often a scatter diagram is used to prove or disprove cause-and-effect relationships. While the diagram shows relationships, it does not by itself prove that one variable *causes* the other. In brief, the easiest way to visualize Bivariate Data is through a Scatter Plot.

"Two variables are said to be correlated if the change in one of the variables results in a change in the other variable".

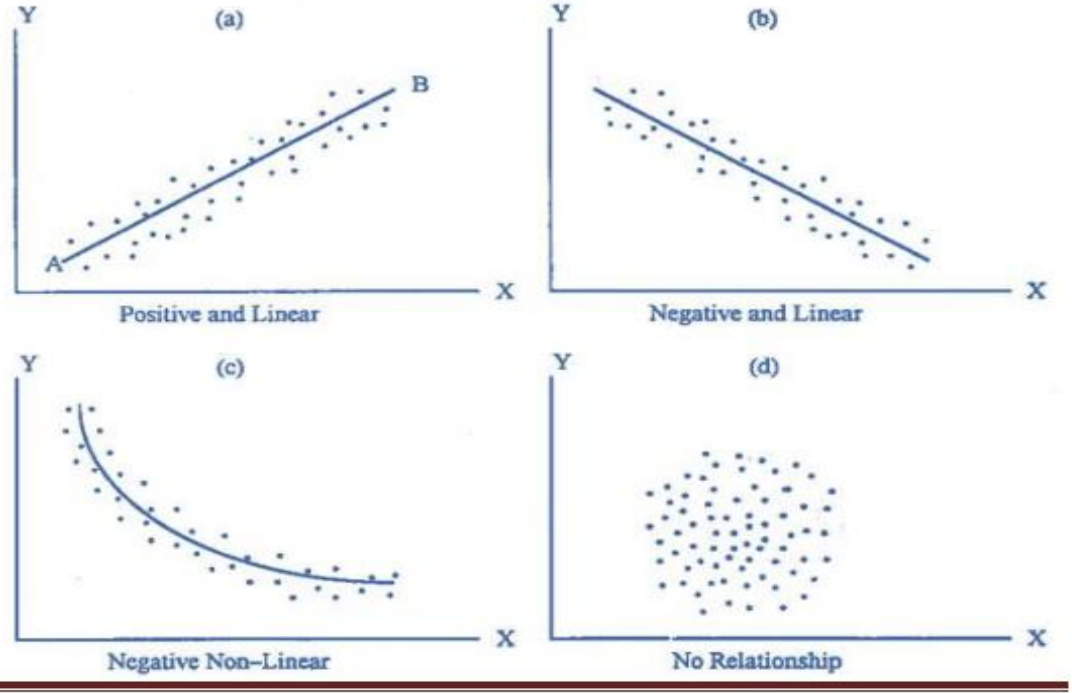
4.2.1: Positive and Negative Correlation

If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.

Some examples of series of positive correlation are:

i. Heights and weights;
ii. Household income and expenditure;
iii. Price and supply of commodities;
iv. Amount of rainfall and yield of crops.

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results in corresponding decrease (or increase) in the values of the other variable. Eg Price and demand of goods.



4.2.2 Interpreting a Scatter Plot

Scatter diagrams will generally show one of six possible correlations between the variables:

i. *Strong Positive Correlation* The value of Y clearly increases as the value of X increases.
ii. *Strong Negative Correlation* The value of Y clearly decreases as the value of X increases.
iii. *Weak Positive Correlation* The value of Y increases slightly as the value of X increases.
iv. *Weak Negative Correlation* The value of Y decreases slightly as the value of X increases.
v. *Complex Correlation* The value of Y seems to be related to the value of X, but the relationship is not easily determined.
vi. *No Correlation* There is no demonstrated connection between the two variables

4,3 Correlation Coefficient

Correlation coefficient measures the degree of linear association between 2 paired variables It takes values from + 1 to – 1.

i. If r = +1,we have **perfect positive** relationship
ii. If r = -1,we have **perfect negative** relationship
iii. If r = 0 there is **no** relationship ie the variables are **uncorrelated.**

4,3 .1 Pearson's Product Moment Correlation Coefficient

Pearson's product moment correlation coefficient, usually denoted by r, is one example of a correlation coefficient. It is a measure of the linear association between two variables that have been measured on interval or ratio scales, such as the relationship between height and weight. However, it can be misleadingly small when there is a relationship between the variables but it is a non-linear one. The correlation coefficient r is given by

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$$ It can be shown that $$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Example:: A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study. Let us determine the coefficient of correlation for this set of data. The first column represents the serial number and the second and third columns represent the weight and blood pressure of each patient.

Weight	78	86	72	82	80	86	84	89	68	71
Blood Pressure	140	160	134	144	180	176	174	178	128	132

Solution

											Totals
X	78	86	72	82	80	86	84	89	68	71	**796**
Y	140	160	134	144	180	176	174	178	128	132	**1546**
xy	10920	13760	9648	11808	14400	15136	14616	15842	8704	9372	**124206**
x^2	6084	7396	5184	6724	6400	7396	7056	7921	4624	5041	**63826**
y^2	19600	25600	17956	20736	32400	30976	30276	31684	16384	17424	**243036**

Thus $r = \dfrac{10(124206) - (796)(1546)}{\sqrt{[(10)63776 - (796)^2 (10)][(243036) - (1546)^2]}} = \dfrac{11444}{\sqrt{(1144)(40244)}} = 0.5966$

Example 2:

Obtain the correlation coefficient of the following data

Mean Temp. (x)	14.2	14.3	14.6	14.9	15.2	15.6	15.9
Pirates (y)	35000	45000	20000	15000	5000	400	17

Solution

Mean Temp. (x)	Pirates (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
14.2	35000	-0.76	17797.57	0.57	316753548	-13475
14.3	45000	-0.66	27797.57	0.43	772704977	-18266
14.6	20000	-0.36	2797.57	0.13	7826405	-999
14.9	15000	-0.06	-2202.43	0	4850691	125
15.2	5000	0.24	-12202.43	0.06	148899263	-2963
15.6	400	0.64	-16802.43	0.41	282321605	-10801
15.9	17	0.94	-17185.43	0.89	295338955	-16203
Tot.=104.7	120417	0	0	2.5	1828695447	-62583
$\bar{x} = 14.96$	$\bar{y} = 17202.43$			S_{xx}	S_{yy}	S_{xy}

We then have that $r = \dfrac{-62583}{\sqrt{2.5(1828695447)}} \approx -0:93$

4.3 .2 Spearman rank correlation coefficient

Data which are arranged in ascending order are said to be in **ranks** or **ranked data.**. The coefficient of correlation for such type of data is given by **Spearman rank difference correlation coefficient** and is denoted by R.

R is given by the formula $R = 1 - \dfrac{6\sum d^2}{n(n^2 - 1)}$

Example

The data given below are obtained from student records.(Grade Point Average (x) and Graduate Record exam score (y)) Calculate the rank correlation coefficient 'R' for the data.

Subject	1	2	3	4	5	6	7	8	9	10
X	8.3	8.6	9.2	9.8	8.0	7.8	9.4	9.0	7.2	8.6
y	2300	2250	2380	2400	2000	2100	2360	2350	2000	2260

Solution

Note that in the x row, we have two students having a grade point average of 8.6 also in the y row; there is a tie for 2000.

Now we arrange the data in descending order and then rank 1,2,3,. . . . .10 accordingly. In case of a tie, the rank of each tied value is the mean of all positions they occupy. In x, for instance, 8.6 occupy ranks 5 and 6. So each has a rank $\dfrac{5+6}{2} = 5.5$

Similarly in 'y' 2000 occupies ranks 9 and 10, so each has rank 9.5

Now we come back to our formula $R = 1 - \dfrac{6\sum d^2}{n(n^2 - 1)}$

We compute d, square it and substitute its value in the formula

X	8.3	8.6	9.2	9.8	8.0	7.8	9.4	9.0	7.2	8.6
y	2300	2250	2380	2400	2000	2100	2360	2350	2000	2260
Rank(X)	7	5.5	3	1	8	9	2	4	10	5.5
Rank(Y)	5	7	2	1	9.5	8	3	4	9.5	6
d	2	-1.5	1	0	-1.5	1	-1	0	0.5	-0.5
d^2	4	2.25	1	0	2.25	1	1	0	0.25	0.25

So here, n = 10 and $\sum d^2 = 12$. So $R = 1 - \frac{6(12)}{10(100-1)} = 1 - 0.0727 = 0.9273$

Note: If we are provided with only ranks without giving the values of x and y we can still find Spearman rank difference correlation R by taking the difference of the ranks and proceeding in the above shown manner.

4,4 Regression

If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other.

Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical.

Regression analysis was explained by M. M. Blair as follows:

"Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data."

3.4.1 Regression Equation

Regression analysis can be thought of as being sort of like the flip side of correlation. It has to do with finding the equation for the kind of straight lines you were just looking at Suppose we have a sample of size n and it has two sets of measures, denoted by x and y. We can predict the values of y given the values of x by using the equation, $y^* = a + bx$

Where the coefficients 'a' and 'b' are real numbers given by

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \quad \text{and} \quad a = \frac{\sum y - b\sum x}{n}$$

The symbol y^* refers to the predicted value of y from a given value of x from the regression equation.

Example: Scores made by students in a statistics class in the mid-term and final examination are given here. Develop a regression equation which may be used to predict final examination scores from the mid – term score.

Student	1	2	3	4	5	6	7	8	9	10
Mid term	98	66	100	96	88	45	76	60	74	82
Final	90	74	98	88	80	62	78	74	86	80

Solution:

We want to predict the final exam scores from the mid term scores. So let us designate 'y' for the final exam scores and 'x' for the mid term exam scores.

We obain the following table for the calculations.

Student	1	2	3	4	5	6	7	8	9	10	**Totals**
X	98	66	100	96	88	45	76	60	74	82	785
Y	90	74	98	88	80	62	78	74	86	80	810
X^2	9604	4356	10000	9216	7744	2025	5776	3600	5476	6724	64521
XY	8820	4884	9800	8448	7040	2790	5928	4440	6364	6560	65074

$$b=\frac{10\,(65{,}074)-\,785\,(810)}{10(\,64,521)-\,(785)^2}=\frac{14{,}860}{28{,}985}=0.5127 \quad \text{and} \quad a=\frac{810-785(\,0.5127)}{10}=40.7531$$

Thus, the regression equation is given by $y^{*}=40.7531+(0.5127)\,x$ We can use this to find the projected or estimated final scores of the students. Eg for the midterm score of 50 the projected final score is $y^{*}=40.7531+(0.5127)\,50=66.3881$, which is a good estimation. To give another example, consider the midterm score of 70. Then the projected final score is $y^{*}=40.7531+(0.5127)\,70=76.6421$, which is again a very good estimation.

Practice Problems:

1. Consider the following data and draw a scatter plot

X	1.0	1.9	2.0	2.9	3.0	3.1	4.0	4.1	5
Y	10	99	100	999	1,000	1,001	10,000	10,001	100,000

2. . Let variable X is the number of hamburgers consumed at a cook-out, and variable Y is the number of beers consumed. Develop a regression equation to predict how many beers a person will consume given that we know how many hamburgers that person will consume.

Hamburgers	5	4	3	2	1
Beers	8	10	4	6	2

3. A horse owner is investigating the relationship between weight carried and the finish position of several horses in his stable. Calculate r and R for the data given

Weight carried	110	113	120	115	110	115	117	123	106	108	110	110
Position Finished	2	6	3	4	6	5	4	2	1	4	1	3

4. The top and bottom number which may appear on a die are as follows Calculate r and R for these values. Are the results surprising?

Top	1	2	3	4	5	6
Bottom	5	6	4	3	1	2

5. The ranks of two sets of variables (Heights and Weights) are given below. Calculate the Spearman rank difference correlation coefficient R.

Heights	2	6	8	4	7	4	9.5	4	1	9.5
Weights	9	1	9	4	5	9	2	7	6	3

6. Researchers interested in determining if there is a relationship between death anxiety and religiosity conducted the following study. Subjects completed a death anxiety scale (high score = high anxiety) and also completed a checklist designed to measure an individuals degree of religiosity (belief in a particular religion, regular attendance at religious services, number of times per week they regularly pray, etc.) (high score = greater religiosity . A data sample is provided below:

X	38	42	29	31	28	15	24	17	19	11	8	19	3	14	6
y	4	3	11	5	9	6	14	9	10	15	19	17	10	14	18

   a) What is your computed answer?
   b) What does this statistic mean concerning the relationship between death anxiety and religiosity?
   c) What percent of the variability is accounted for by the relation of these two variables?

7. The data given below are obtained from student records.(Grade Point Average (x) and Graduate Record exam score (y)) Calculate the regression equation and compute the estimated GRE scores for GPA = 7.5 and 8.5..

X	8.3	8.6	9.2	9.8	8.0	7.8	9.4	9.0	7.2	8.6
y	2300	2250	2380	2400	2000	2100	2360	2350	2000	2260

8. A horse was subject to the test of how many minutes it takes to reach a point from the starting point. The horse was made to carry luggage of various weights on 10 trials.. The data collected are presented below in the table. Find the regression equation between the load and the time taken to reach the goal. Estimate the time taken for the loads of 35 Kgs , 23 Kgs, and 9 Kgs. Are the answers in agreement with your intuitive feelings? Justify.

Trial Number	1	2	3	4	5	6	8	8	9	10
Weight (in Kgs)	11	23	16	32	12	28	29	19	25	20
Time taken (in mins)	13	22	16	47	13	39	43	21	32	22

9. A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study.

Weight	78	86	72	822	80	86	84	89	68	71
Blood Pressure	140	160	134	144	180	176	174	178	128	132

10. It is assumed that achievement test scores should be correlated with student's classroom performance. One would expect that students who consistently perform well in the classroom (tests, quizes, etc.) would also perform well on a standardized achievement test (0 - 100 with 100 indicating high achievement (x)). A teacher decides to examine this hypothesis. At the end of the academic year, she computes a correlation between the students achievement test scores (she purposefully did not look at this data until after she submitted students grades) and the overall g.p.a.(y) for each student computed over the entire year. The data for her class are provided below.

X	98	96	94	88	01	77	86	71	59	63	84	79	75	72	86	85	71	93	90	62
y	3.6	2.7	3.1	4.0	3.2	3.0	3.8	2.6	3.0	2.2	1.7	3.1	2.6	2.9	2.4	3.4	2.8	3.7	3.2	1.6

a) Compute the correlation coefficient.
b) What does this statistic mean concerning the relationship between achievement test prformance and g.p.a.?
c) What percent of the variability is accounted for by the relationship between the two variables and what does this statistic mean?
d) What would be the slope and y-intercept for a regression line based on this data?
e) If a student scored a 93 on the achievement test, what would be their predicted G.P.A.? If they scored a 74? A 88?

11. 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.

Cost of internet (in $)	11	18	17	15	9	5	12	19	22	25
satisfaction	6	8	10	4	9	6	3	5	2	10

a) Compute the correlation coefficient.
b) What does this statistic mean concerning the relationship between amount of money spent per month on internet provider service and level of customer satisfaction?
c) What percent of the variability is accounted for by the relationship between the two variables and what does this statistic mean?

12. It is hypothesized that there are fluctuations in norepinephrine (NE) levels which accompany fluctuations in affect with bipolar affective disorder (manic-depressive illness). Thus, during depressive states, NE levels drop; during manic states, NE levels increase. To test this relationship, researchers measured the level of NE by measuring the metabolite 3-methoxy-4-hydroxyphenylglycol (MHPG in micro gram per 24 hour) in the patient's urine experiencing varying levels of mania/depression. Increased levels of MHPG are correlated with increased

metabolism (thus higher levels) of central nervous system NE. Levels of mania/depression were also recorded on a scale with a low score indicating increased mania and a high score increased depression. The data is provided below.

MHPG	980	1209	1403	1950	1814	1280	1073	1066	880	776
Affect	22	26	8	10	5	19	26	12	23	28

a) Compute the correlation coefficient.
b) What does this statistic mean concerning the relationship between MHPG levels and affect?
c) What percent of the variability is accounted for by the relationship between the two variables?
d) What would be the slope and y-intercept for a regression line based on this data?
e) What is the predicted affect score for an individual with MHPG level of; 1100, 950, 700?

13. The table below contains 25 cases -- the mother's weight in kilograms and the infant's birth weight in grams. Does this data suggest some relationship between the mother's weight and the infant's birth weight Why would such a relationship be important

. **Prepregnancy Weights of Mothers and Birthweights of their Infants**

Case Number	Mother's Weight (kg)	Infant's Birthweight (g)
1	49.4	3515
2	63.5	3742
3	68.0	3629
4	52.2	2680
5	54.4	3006
6	70.3	4068
7	50.8	3373
8	73.9	4124
9	65.8	3572
10	54.4	3359
11	73.5	3230
12	59.0	3572
13	61.2	3062
14	52.2	3374
15	63.1	2722
16	65.8	3345
17	61.2	3714
18	55.8	2991
19	61.2	4026
20	56.7	2920
21	63.5	4152
22	59.0	2977
23	49.9	2764
24	65.8	2920
25	43.1	2693