ITCS 121 Statistics

Lecture 3

Describing Bivariate Data

Bivariate data

- When each experimental unit is associated with two variables we have **bivariate data**.
- E.g. every person has a height and a weight.
- We can investigate each variable individually (just pretend the other variable doesn't exist).
- Sometimes we are interested in the relationship between the two variables.
- E.g. Is a person's height usually related to their weight?
- We can describe this relationship with graphical and numerical methods.

Pie charts for two qualitative variables

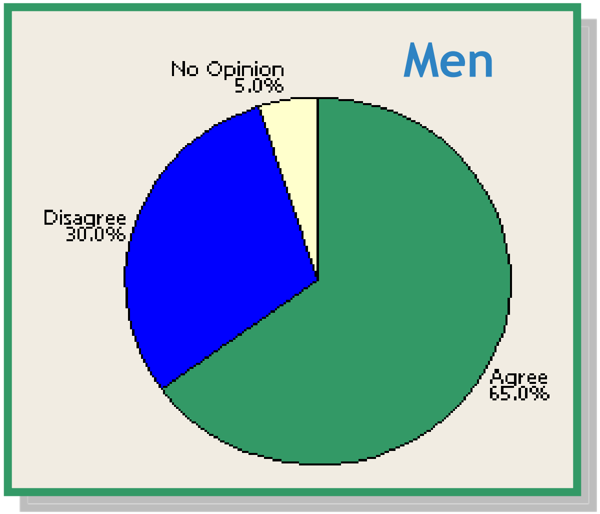
- When both variables are qualitative we can use comparative pie charts.
- E.g. Suppose we ask the men and women in a company whether they agree that male and female employees are treated equally in the company.
- Variable 1 – Male | Female
- Variable 2 – Agree | Disagree | No opinion

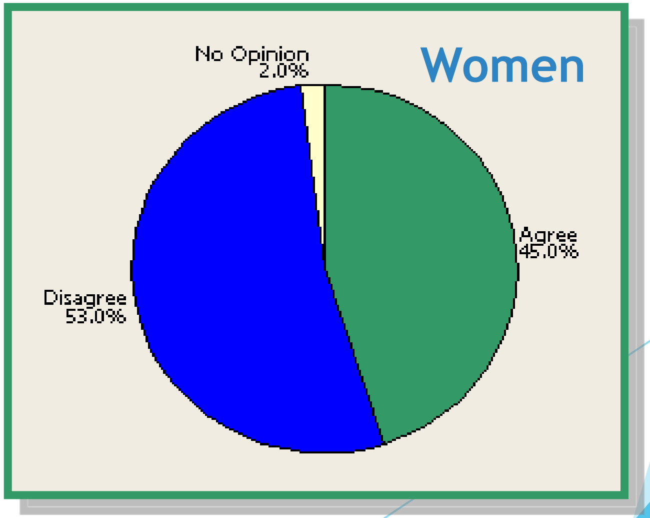
Example: pie charts

- Suppose we asked 140 men and 100 women. Suppose we get results as follows:

	Agree	Disagree	No opinion
Men	91	42	7
Women	45	53	2

- We can represent this information with two pie charts.





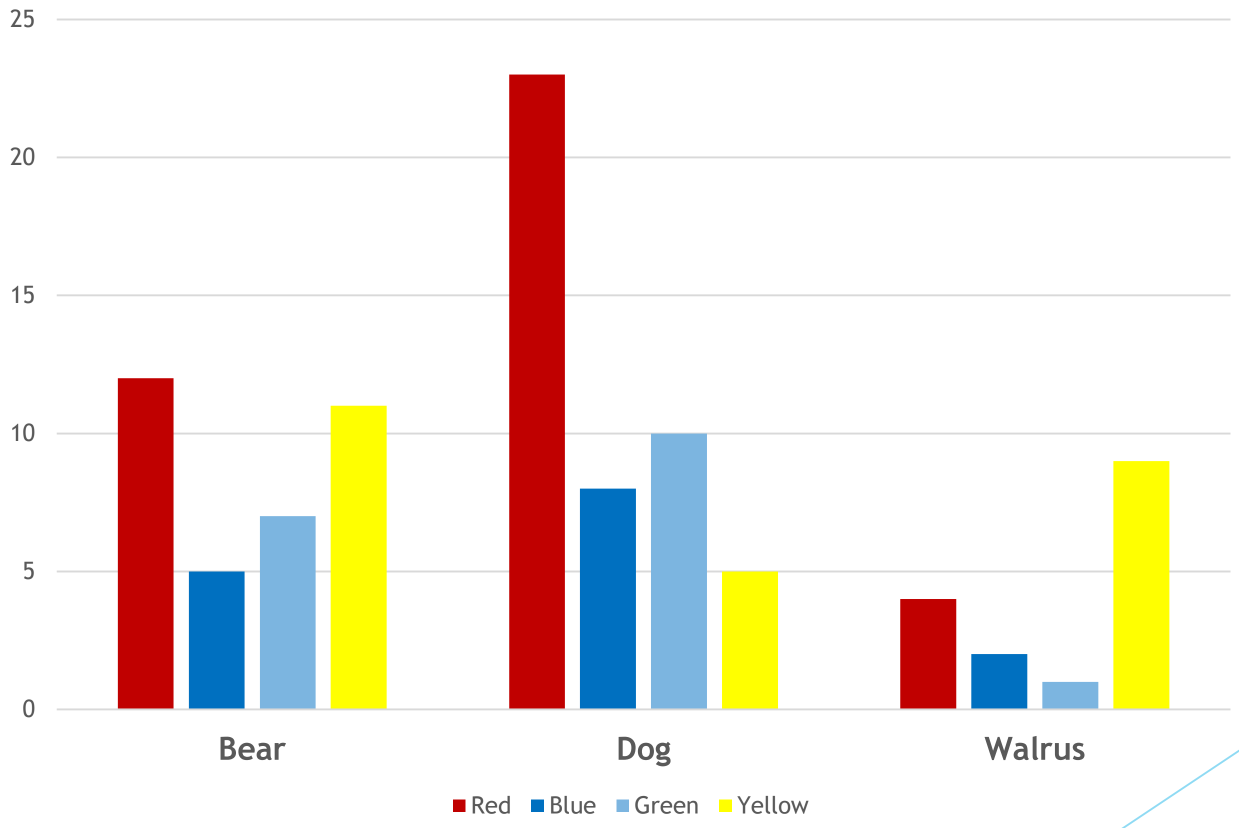
Bar charts for two qualitative variables

- We can also use comparative bar charts.
- E.g. We could interview people and record their favourite animal and their favourite colour.
- We might get results like this:

	Red	Blue	Green	Yellow
Bear	12	5	7	11
Dog	23	8	10	5
Walrus	4	2	1	9

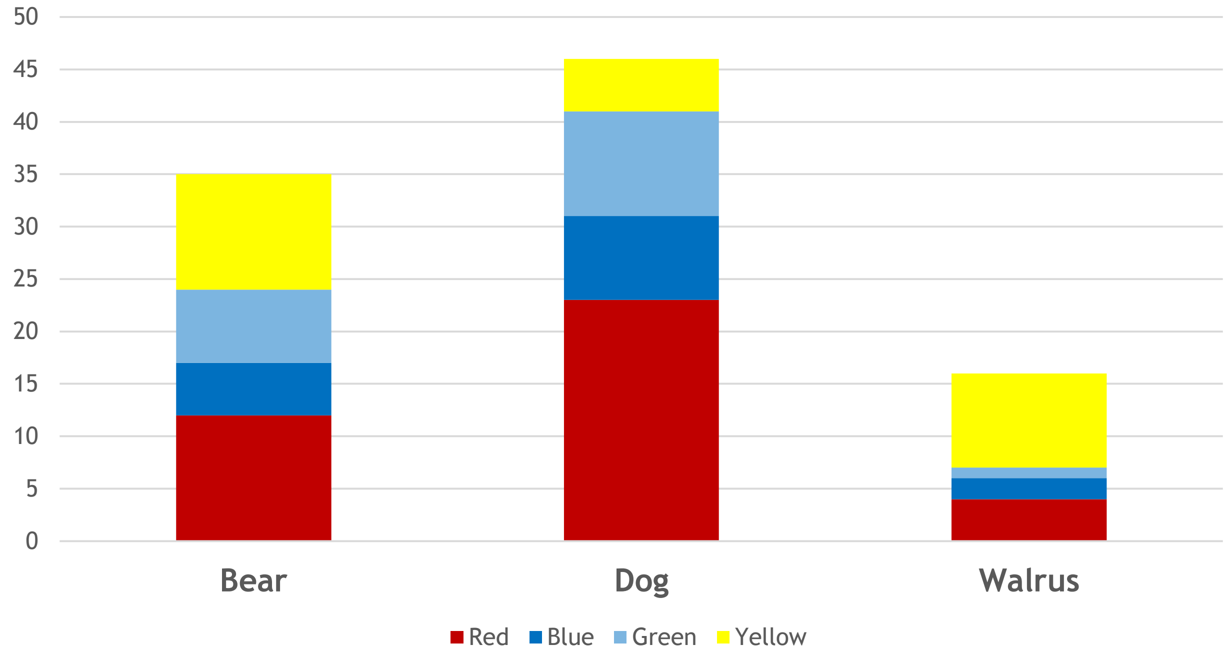
Example: side-by-side

- We can represent this data like this:



Example: stacked

- We could also use a **stacked bar chart.**



- Stacked charts can be useful because, for example, we can easily see that 35 people chose bear.
- However, it is harder to read from this chart how many people chose ‘bear’ and ‘blue’, for example.

Some comments

- We can also use bar charts when one variable is qualitative and the other is quantitative.
- For example, in the first class there was the M&Ms example, where we counted the number of candies of each colour. We can treat this as data with variables ‘colour’ and ‘frequency’.
- We can also use bar charts with multivariate data in some situations.
- There are other kinds of charts we could use too, but we’re not going to talk about them here.

Class activity 1

A research group surveyed 198 parents and 200 children and recorded their responses to the question, “How much free time does your child have?” or “How much free time do you have?” The responses are shown in the table below.

	Just the right amount	Not enough	Too much	Don’t know
Parents	138	14	40	6
Children	130	48	16	6

a) What is the population of interest here?

b) Describe the variables that have been measured in this survey. Are they qualitative or quantitative? Is the data univariate or bivariate?

c) Use comparative pie charts to compare the responses for parents and children.

d) Use a side-by-side comparative bar char to compare the responses for parents and children.

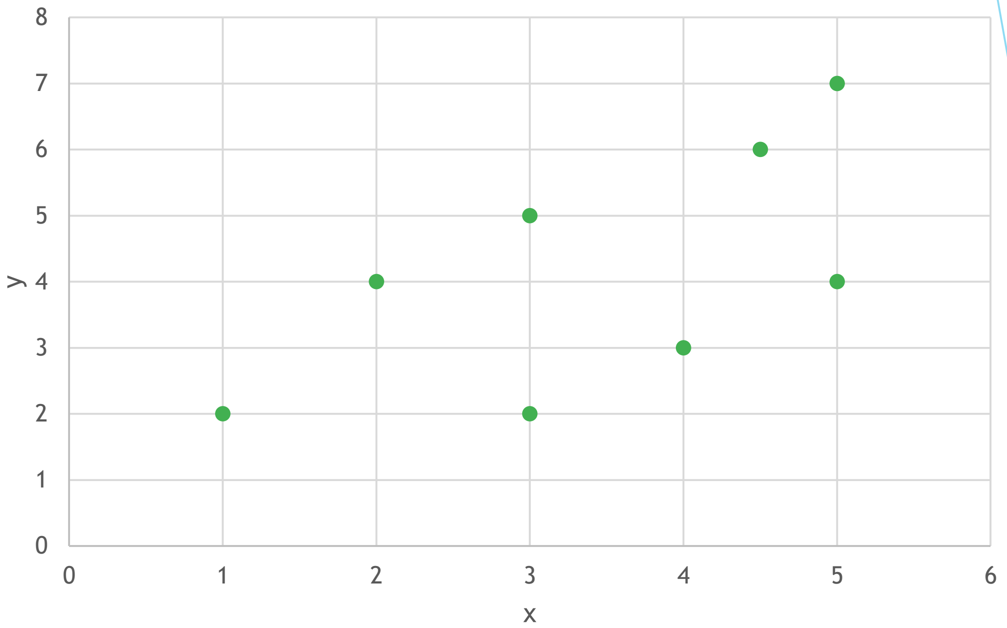
e) What, if any, information can we get from these graphs?

Two quantitative variables

- It is a very common situation that we have two quantitative variables, and we want to investigate the relationship between them.
- An easy way to represent the data is to use a **scatter plot**.
- If we have values $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, a scatter plot draws these as points on a 2-dimensional plane.

Example: Scatter plot

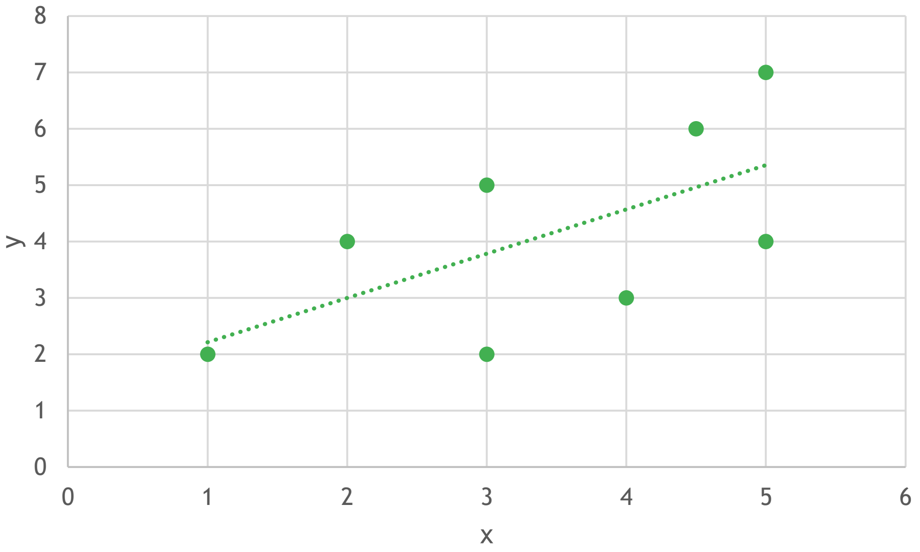
x	y
1	2
5	4
3	2
4	3
4.5	6
3	5
2	4
5	7



- Is there a pattern in this data? Can we see some kind of relationship?

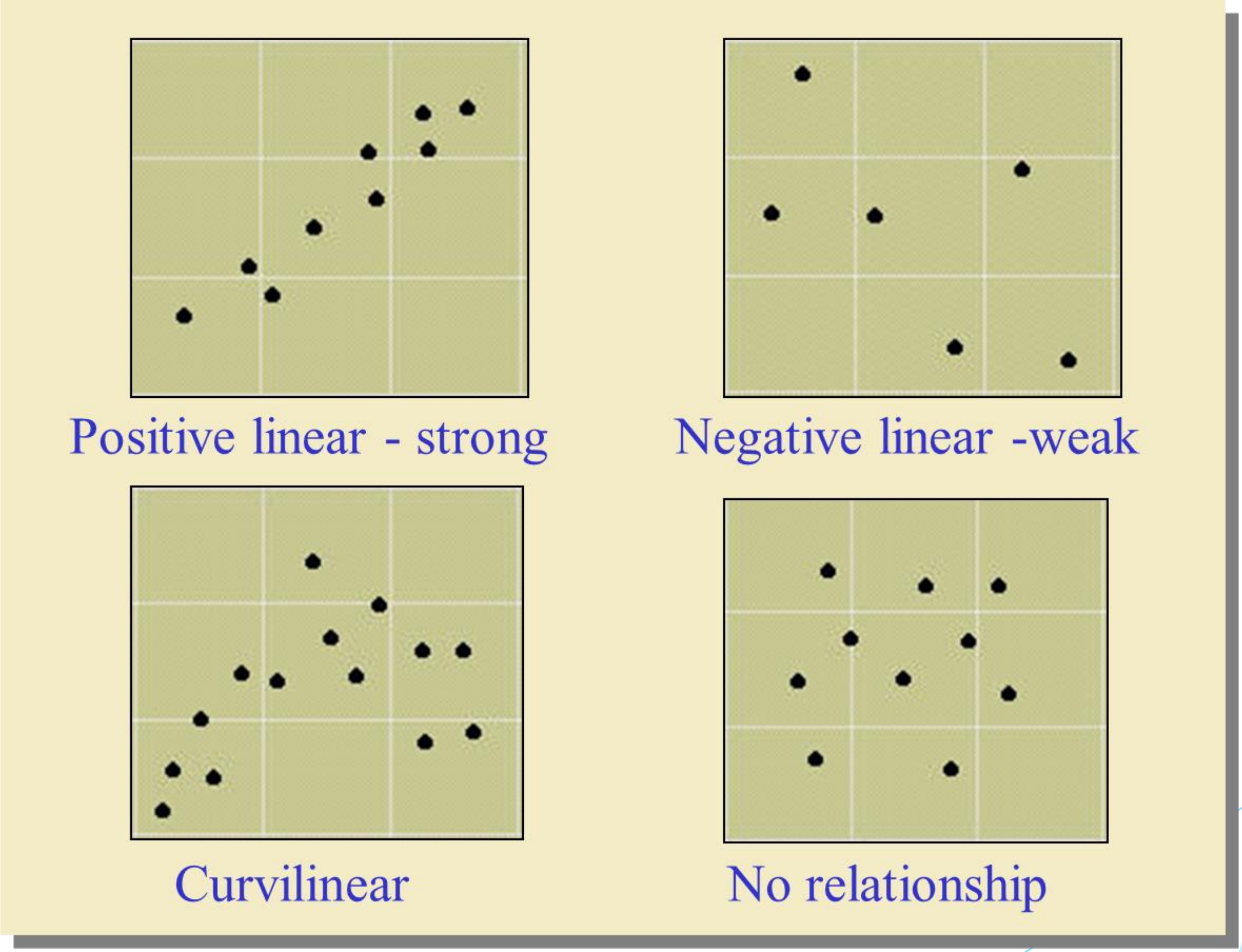
Example: Scatter plot

- There seems to be a positive linear relationship between the x and y values.
- I.e. as x gets bigger, y gets bigger proportionally.



- This isn’t always going to be true. Data may have no strong relationship, or even relationships that are non-linear.

Some relationships



Positive linear - strong

Negative linear -weak

Curvilinear

No relationship

Describing linear relationships

- To say that two variables x and y have a linear relationship is to say that (x, y) pairs in the data lie approximately on a straight line.

- That is, the pairs lie roughly on a line defined by an equation $y = a + bx$.

- If we have a relationship like this, we want to know two things:
  1. Is the relationship positive or negative? I.e. if x gets bigger does y get bigger or smaller?
  2. How strong is the relationship? I.e. how close are the points in the data to the line $y = a + bx$.

Covariance

- The **variance** measures how much a variable deviates, on average, from its mean.
- The **covariance** of two variables x and y is a measure of how much their variation from their means is related.
- To calculate the covariance of two random variables using a sample $\{(x_1, y_1), \dots, (x_n, y_n)\}$ of bivariate data, we use the formula

$$Cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- Suppose $x_i > \bar{x}$. If $y_i > \bar{y}$ then we get a positive contribution to the covariance. If $y_i < \bar{y}$ then we get a negative contribution. The covariance is a kind of average of all these.
- The formula can be rearranged to get

$$Cov(x, y) = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n - 1}$$

The correlation coefficient

- The **correlation coefficient**, denoted by r, is a numerical value designed to give us information about the direction and strength of linear relationships between two variables.

After some algebra!

$$r = \frac{s_{xy}}{s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

$$\left.\begin{aligned} s_x^2 &= \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} \\ s_y^2 &= \frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n-1} \end{aligned}\right\} \text{Sample variances}$$

$$s_{xy} = Cov(x,y) = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n-1}$$

Properties of r

- This is not obvious from the definition, but r has the following important property:

$$-1 \leq r \leq 1$$

- So, r is always between -1 and 1.
- Again, r **is always between -1 and 1.**
- If you calculate r and your answer is not between -1 and 1, then you have made a mistake somewhere.

Understanding the correlation coefficient

- r is a *normalized* version of $Cov(x, y)$.
- In other words, r is $Cov(x, y)$ but scaled so it is not dependent on the sizes of the x and y values.
- Normalization is a very important idea in statistics.
- If we want to compare properties of different populations, we must make sure we are using readings on the same scale.
- The z-score is another example of a normalized measurement.

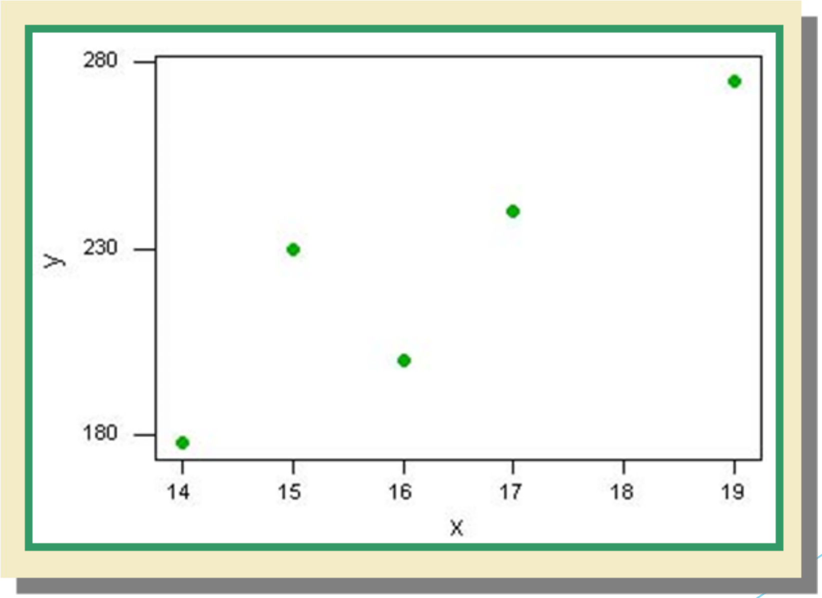
How to interpret r

- The sign of r tells us the direction of the relationship.
  - If r is positive the relationship is positive.
  - If r is negative the relationship is negative.
- The magnitude (absolute value) of r tells us the strength of the direction.
  - If $r = 0$ there is no relationship between the two variables.
  - If r is exactly -1 or 1 the relation ship is perfect. In other words, the data points all lie on the same straight line.
  - r values close to 1 or -1 represent strong relationships.
  - r values close to zero represent weak relationships.

Example: Homes

Living area vs selling price of 5 homes:

Residence	1	2	3	4	5
x (thousand sq ft)	14	15	17	19	16
y (thousand dollars)	178	230	240	275	200



Example: Homes

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

- What is the value of r here?

x	y	x^2	y^2	xy
14	178			
15	230			
17	240			
19	275			
16	200			
81	**1123**			

Example: Homes

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

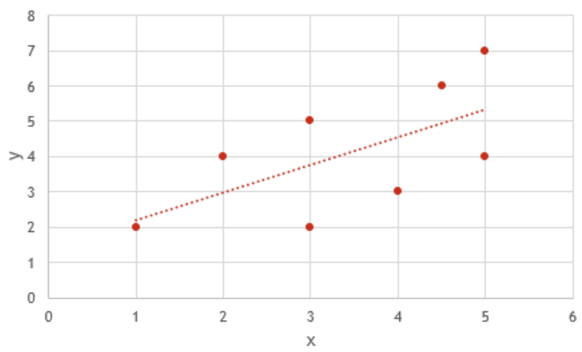
- What is the value of r here?

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x^2}$	$\boldsymbol{y^2}$	$\boldsymbol{xy}$
14	178	196	31684	2492
15	230	225	52900	3450
17	240	289	57600	4080
19	275	361	75625	5225
16	200	256	40000	3200
81	**1123**	**1327**	**257809**	**18447**

$\sum x$ $\sum y$ $\sum x^2$ $\sum y^2$ $\sum xy$

$$r = \frac{5(18447) - (81)(1123)}{\sqrt{5(1327) - 81^2}\sqrt{5(257809) - 1123^2}} = 0.885$$

The regression line



- For any set of points in a plane, we can draw something called the **regression line**, which is also known as the **best fit line**.
- This is the straight line that fits the data ‘as well as possible’.
- You will see exactly what this means in the numerical methods course next year.
- This is particularly useful when the variable y depends on the variable x.
- We can use the regression line to predict values of y given values of x.
- E.g. Predict the selling price of a home from its square footage.

Finding the regression line

- We want to find the equation $y = a + bx$ for the regression line.
- You will see a systematic way to do this in the numerical methods class next year.
- In turns out we can find a and b using r and the means and standard deviations of x and y.

$$b = r\frac{s_y}{s_x} \qquad a = \bar{y} - b\bar{x}$$

Example: Homes again

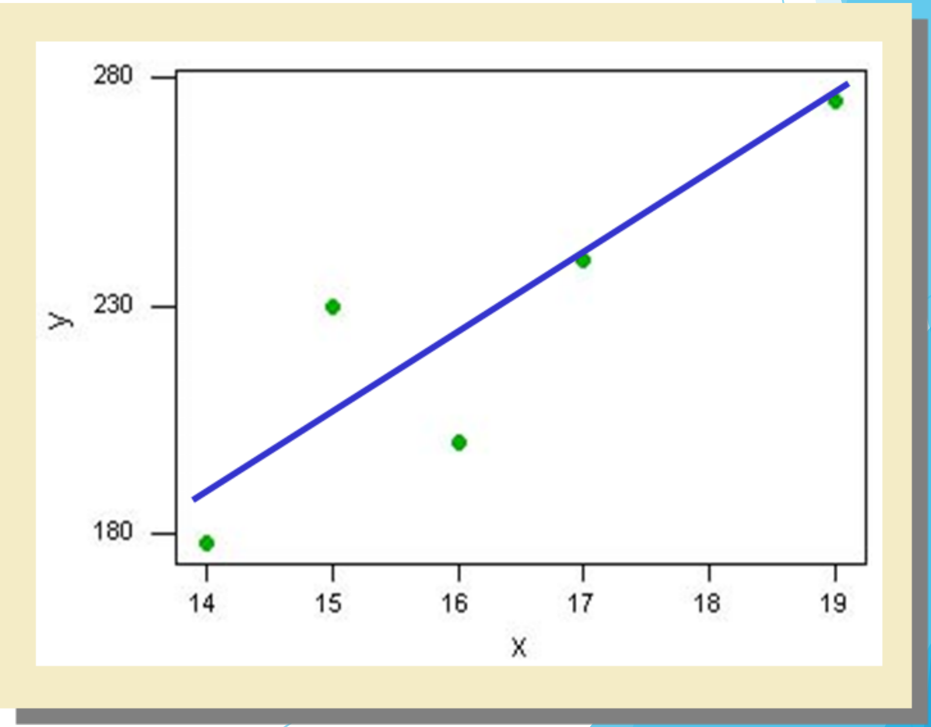
- We already found $r = 0.885$.
- We can calculate $\bar{x} = 16.2$ and $\bar{y} = 224.6$.
- We can also calculate $s_x = 1.9235$ and $s_y = 37.3604$.
- So $b = 0.885\frac{37.3604}{1.9235} = 17.189$.
- And $a = 224.6 - 17.189(16.2) = -53.86$.
- So the regression line is:

$$y = -53.86 + 17.189x$$

How much should a home with 16,000 square feet of living are cost?

$$y = -53.86 + 17.189(16) = 221.16$$

So $221,160.



Class activity 2

A set of bivariate data consists of these measurements on two variables, x and y:

(3,6) (5,8) (2,6) (1,4) (4,7) (4,6)

a) Draw a scatter plot to describe the data.

b) Does there appear to be a relationship between x and y? If so, how would you describe it?

c) Calculate the correlation coefficient, r.

d) Find the best-fit line. Graph the line on the scatter plot from part a).

Class activity 3

Consider this set of bivariate data.

x	1	2	3	4	5	6
y	5.6	4.6	4.5	3.7	3.2	2.7

a) Draw a scatter plot to describe the data.

b) Does there appear to be a relationship between x and y? If so, how do you describe it?

c) Calculate the correlation coefficient, r. Does the value of r confirm your conclusions in part b)? Explain.

Class activity 4

The car maker Lexus have steadily increased their sales since their U.S. launch in 1989. However, the rate of increase changed in 1996 when Lexus introduced a new line of trucks. The sales of Lexus from 1996 to 2003 are shown in the table.

Year	1996	1997	1998	1999	2000	2001	2002	2003
Sales (thousand dollars)	80	100	155	180	210	225	230	260

a) Plot the data using a scatter plot. How would you describe the relationship between year and sales of Lexus?

b) Find the regression line relating the sales of Lexus to the year being measured.

c) Predict the sales of Lexus in the year 2015.

d) What problems might your prediction have?