STAT 311: Quantitative Bivariate Data

Y. Samuel Wang

Summer 2016

Questions?

Any questions so far?

- Class organization
- Material covered so far
- Labs

Describing data

Last lecture, we considered describing one variable at a time.

- Numerical summaries (mean, standard deviation, five number summary)
- Graphical summaries (histogram, boxplot, etc)

Bivariate data

Often individual variables are related or associated, so we need ways to describe two variables at once (bivariate data).

- Education vs Income
- Precipitation vs Temperature
- Hair color vs Eye color
- Hours studying vs Grade on exam

Bivariate data

Often individual variables are related or associated, so we need ways to describe two variables at once (bivariate data).

- Education vs Income
- Precipitation vs Temperature
- Hair color vs Eye color
- Hours studying vs Grade on exam

How we describe the data depends on how the variables are measured. Today we will focus on numeric data.

Bivariate Data

Often we assign labels to each of the individual variables-

X	Y
Explanatory	Response
Independent	Dependent
Predictor	Outcome

Bivariate Data

Often we assign labels to each of the individual variables-

X	Y
Explanatory	Response
Independent	Dependent
Predictor	Outcome

The terminology suggests a causal relationship, but **be careful**.
Often we cannot determine a causal relationship, only a correlation.

Time Series Data

When the time variable is the "X" variable, we typically call the data **time series**. This is a special type of bivariate data.

- Measurements can be taken at regular or irregular intervals
- Time scale (seconds, minutes, days, years) depends on data
- Examples:

Time Series Data

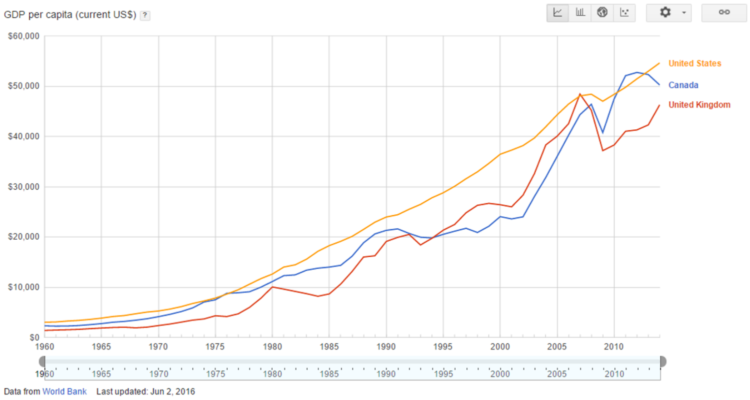
When the time variable is the "X" variable, we typically call the data **time series**. This is a special type of bivariate data.

- Measurements can be taken at regular or irregular intervals
- Time scale (seconds, minutes, days, years) depends on data
- Examples: stock prices, temperature, population, GDP, bandwith speed, glucose levels, etc

Example Time Series

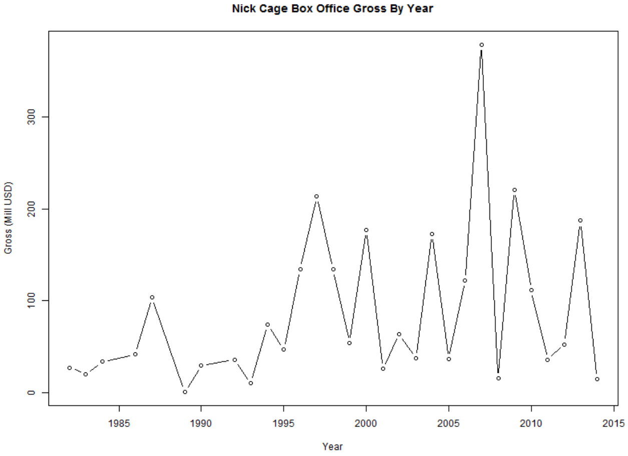
Some time series exhibit clear long term trends

GDP Per Capita



Example Time Series

Some are not as clear
Nicholas Cage box office Revenue vs Year



Example Time Series

Some exhibit seasonal trends: Nordstrom's Quarterly Revenue

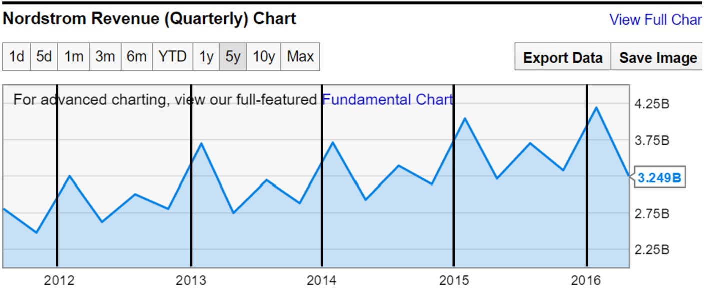
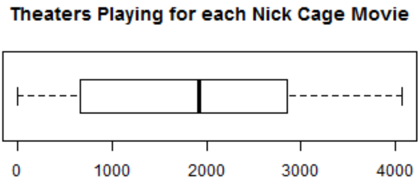
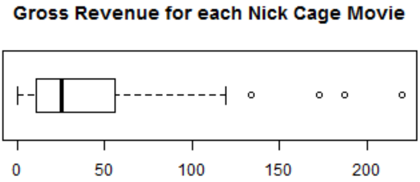


Figure: Data from ycharts.com

Marginal Distribution

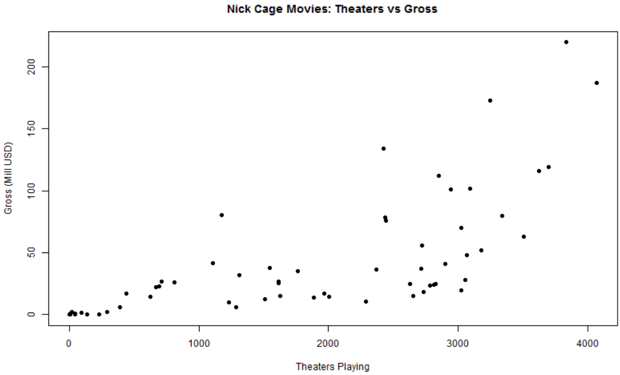
Recall from last lecture how we described single variables at a time. We call these **marginal distributions**.





Scatterplot

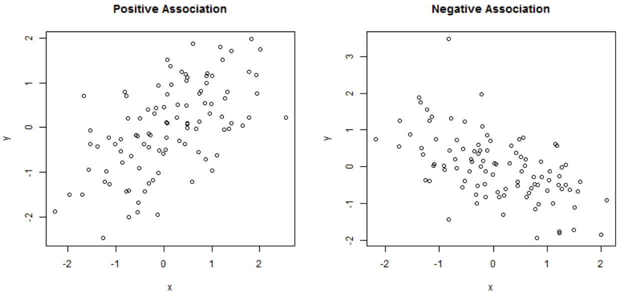
Scatterplots give us a way to describe general bivariate data. When we consider two variables together, we call this the **joint distribution**. Each **observational unit** is represented by a point in the graph



How to describe a scatterplot

Variables can have a positive or negative association

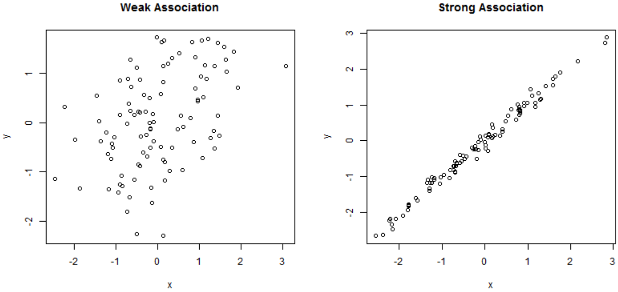
- **Positive association**: An increase (decrease) in one variable generally corresponds to an increase (decrease) in the other variable
- **Negative association**: An increase (decrease) in one variable generally corresponds to an decrease (increase) in the other variable



How to describe a scatterplot

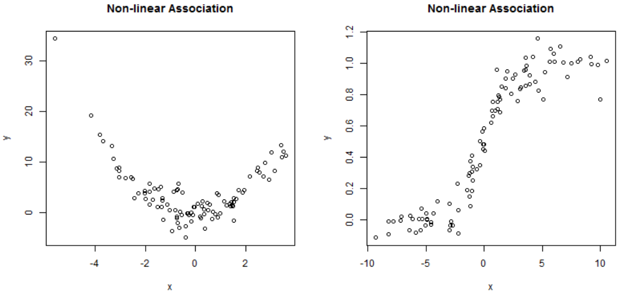
We can also think about the strength of the association

- **Weak association**: The points are scattered around the general pattern
- **Strong association**: The points are closely clustered around the general pattern



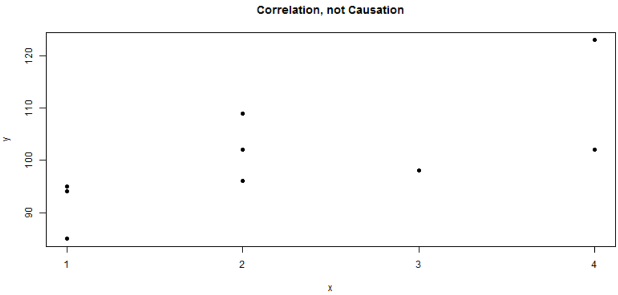
How to describe a scatterplot

So far the scatterplots we have looked at have mostly been **linear** and can be mostly described by a straight line. But variables can have a **non-linear** association as well.



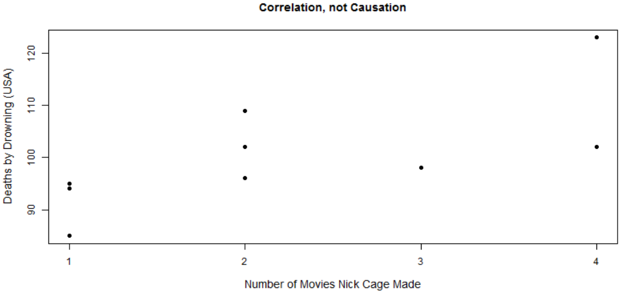
How to describe a scatterplot

Notice we have used the word association or correlation here, **not causation**!



How to describe a scatterplot

Notice we have used the word association or correlation here, **not causation**!



How to describe an association numerically

To numerically summarize how two variables are linearly associated we use

- Covariance
- Correlation

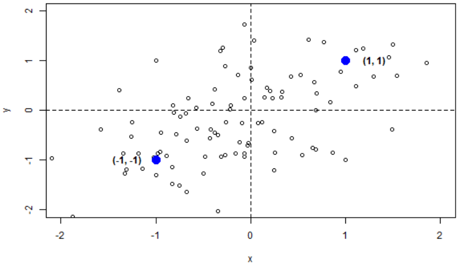
Note that these measures will not measure non-linear relationships well.

Basic Idea

We consider the product of the "deviations from the mean" for both the x and y coordinates of each observation.

$$(x - \bar{x})(y - \bar{y}) \tag{1}$$

Points in the upper right or lower left quadrants yield positive values



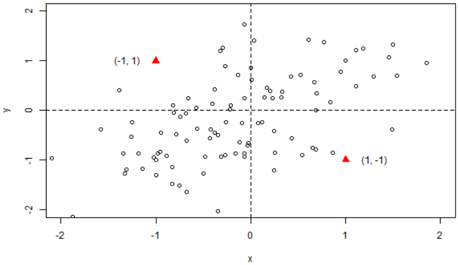
$$\begin{aligned}
\bar{x} &= 0 \\
\bar{y} &= 0 \\
(x_1 - \bar{x})(y_1 - \bar{y}) &= (1-0)(1-0) = 1 \\
(x_2 - \bar{x})(y_2 - \bar{y}) &= (-1-0)(-1-0) = 1
\end{aligned} \tag{2}$$

Basic Idea

We consider the product of the "deviations from the mean" for both the x and y coordinates of each observation.

$$(x - \bar{x})(y - \bar{y}) \tag{3}$$

Points in the lower right or upper left quadrants yield negative values



$$\begin{aligned} \bar{x} &= 0 \\ \bar{y} &= 0 \\ (x_1 - \bar{x})(y_1 - \bar{y}) &= (-1 - 0)(1 - 0) = -1 \\ (x_2 - \bar{x})(y_2 - \bar{y}) &= (1 - 0)(-1 - 0) = -1 \end{aligned} \tag{4}$$

Covariance

$$\text{Covariance}(x, y) = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{5}$$

where $\bar{x}$ and $\bar{y}$ are the average x and y values and n is the number of observations

Covariance

$$\text{Covariance}(x, y) = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \tag{5}$$

where $\bar{x}$ and $\bar{y}$ are the average x and y values and n is the number of observations

- If covariance is positive (negative), that means there is a positive (negative) linear association
- Covariance can range from $(-\infty, \infty)$
- Cov(x,y) $=$ Cov(y,x)
- Cov(x, x) $=$ Var(x) $=$ sd(x)^2

Correlation

$$\text{correlation}(x, y) = \rho_{xy} = \frac{1}{n-1} \sum_i \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} \tag{6}$$

where $\bar{x}$ and $\bar{y}$ are the average x and y values, n is the number of observations and s_x and s_y are the standard deviations of x and y

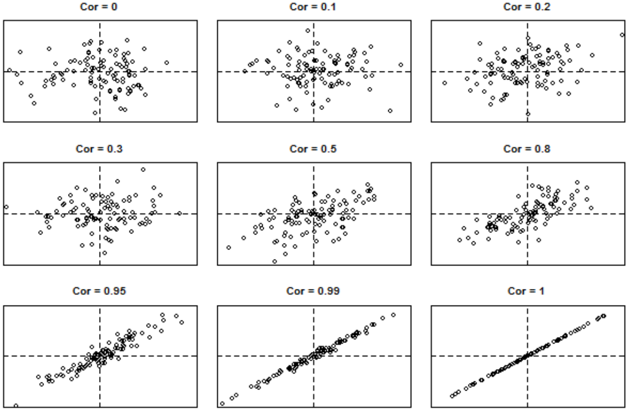
Correlation

$$\text{correlation}(x, y) = \rho_{xy} = \frac{1}{n-1} \sum_i \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y} \qquad (6)$$

where $\bar{x}$ and $\bar{y}$ are the average x and y values, n is the number of observations and s_x and s_y are the standard deviations of x and y

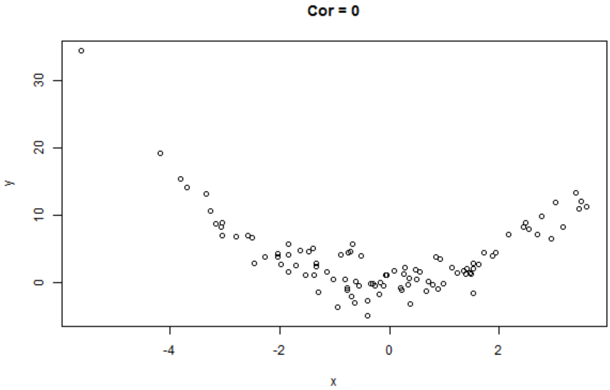
- If correlation is positive (negative), that means there is a positive (negative) linear association
- Correlation can range from $(-1, 1)$
- cor(x,y) = cor(y,x)

Correlation can give us a way to measure strength of **linear** association

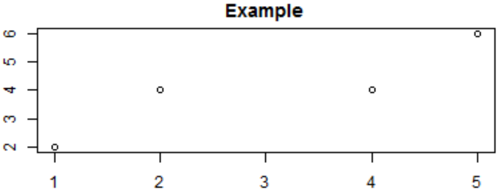


Correlation

The correlation may not accurately represent the strength of association if it is non-linear



Example



X	Y	(x- $\bar{x}$)	(y - $\bar{y}$)	(x- $\bar{x}$) (y - $\bar{y}$)
1	2			
2	4			
4	4			
5	6			

Example