Exploratory Data Analysis and Visualization

3. EDA

Li Xinke

DS

City University of Hong Kong

1. What is data
    - Observations
    - Categorical Variables
    - Numerical Variables

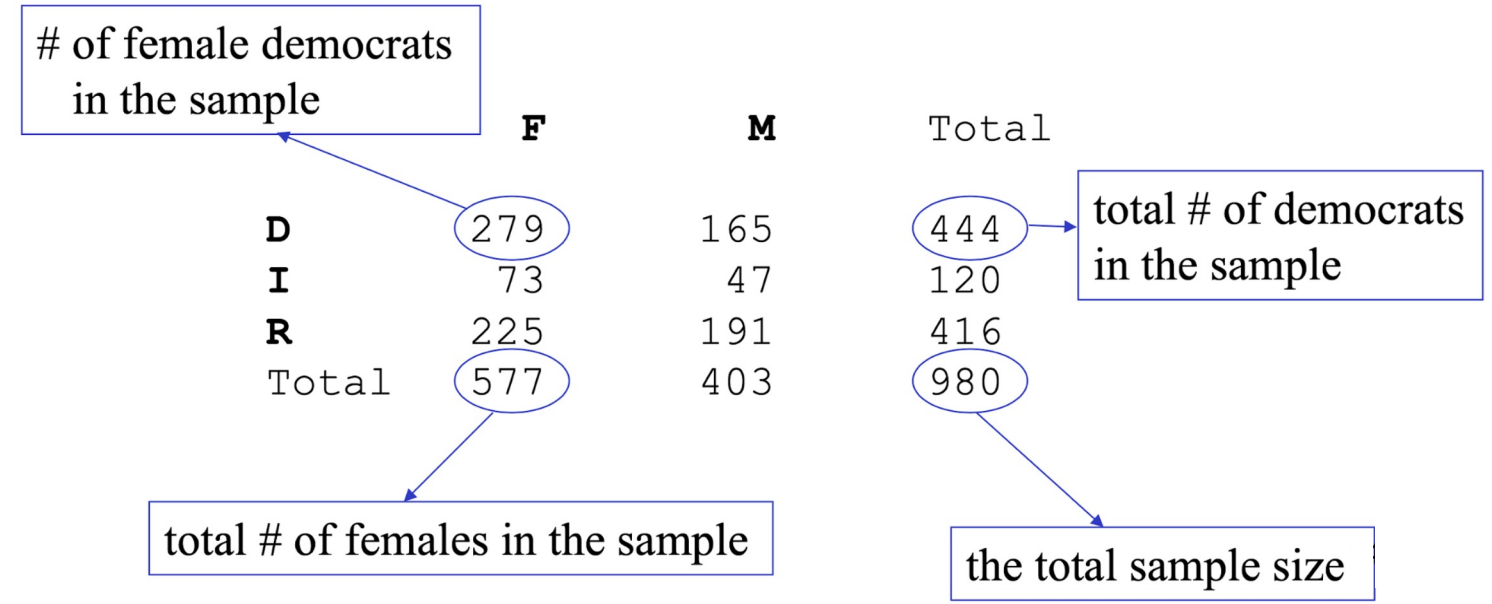
2. EDA

	Categorical Variables	**Numerical Variables**
Univariate Analysis	Count/Percentage	Mean, Median, Mode, Quantiles, Variance, SD
	Bar chart/Pie chart	Histogram & Density Curve, Box plot
Multivariate Analysis	Two-way tables	

Two-way tables

- E.g., Cross-classification of a sample of 980 Americans by gender and party identification
- Variables:

rows: Party (D,I,R) *columns: Gender (F,M)*



Another Example

Two-way table:

	Hospital A	Hospital B
Died	300	50
Survived	2700	950

Death status distributions conditional on hospitals are specified by the % died:

Hospital A: 300/3000=10%
Hospital B: 50/1000 = 5%

Another Example

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%

Third variable

Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

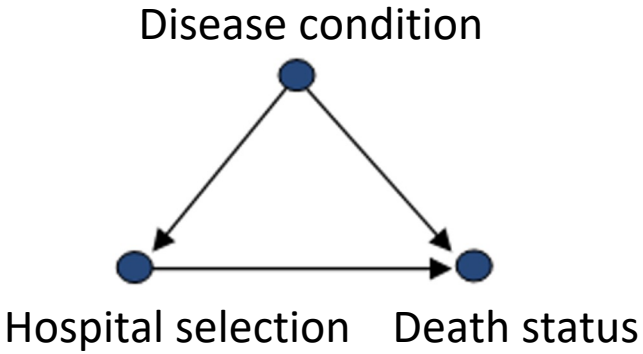
Hospital A has more bad condition patients

Simpson's paradox

Association between two variables has a different direction from the association conditional on a third variable (lurking variable (hidden) or confounding variable (considered))

What is the lurking variable in our example?

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%



Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Simpson's Paradox

- A change in the direction of association between two variables when data are separated into groups defined by a third variable
- Berkeley Sex discrimination case (1977): https://homepage.stat.uiowa.edu/~mbognar/1030/Bickel-Berkeley.pdf

Department	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%
Total	**4526**	**39%**	**2691**	**45%**	**1835**	**30%**

Legend:

greater percentage of successful applicants than the other gender

greater number of applicants than the other gender

bold - the two 'most applied for' departments for each gender

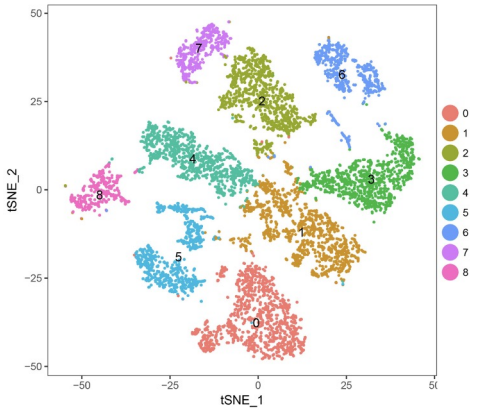
multivariate analysis——

Scatterplot

- Shows the relationship between **two numerical variables** measured on the same individuals
- The values of one variable -> horizontal axis
- The values of the other variable -> vertical axis
- Each individual (observation) appears as a point in the plot

To add a categorical variable to the scatterplot, you can use a different color or symbol for each category

https://seaborn.pydata.org/tutorial/categorical.html



Scatterplot

What can we see from a scatterplot?

- **form:** clusters, linear association, etc.
- **direction:** positive association, negative association.
- **strength:** how close the data points follow the form.

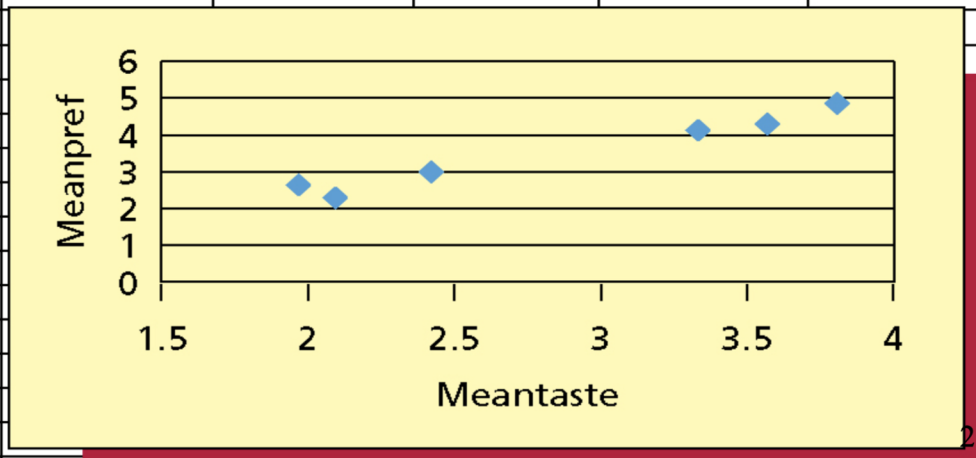
- Positive association: above-average values of one variable accompany above-average values of the other, and below-average values also tend to occur together.

- Negative association: above-average values of one variable accompany below-average values of the other and vice versa

Scatter plot

Restaurant Ratings: Mean Preference vs Mean Taste

	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee's	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald's	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy's	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



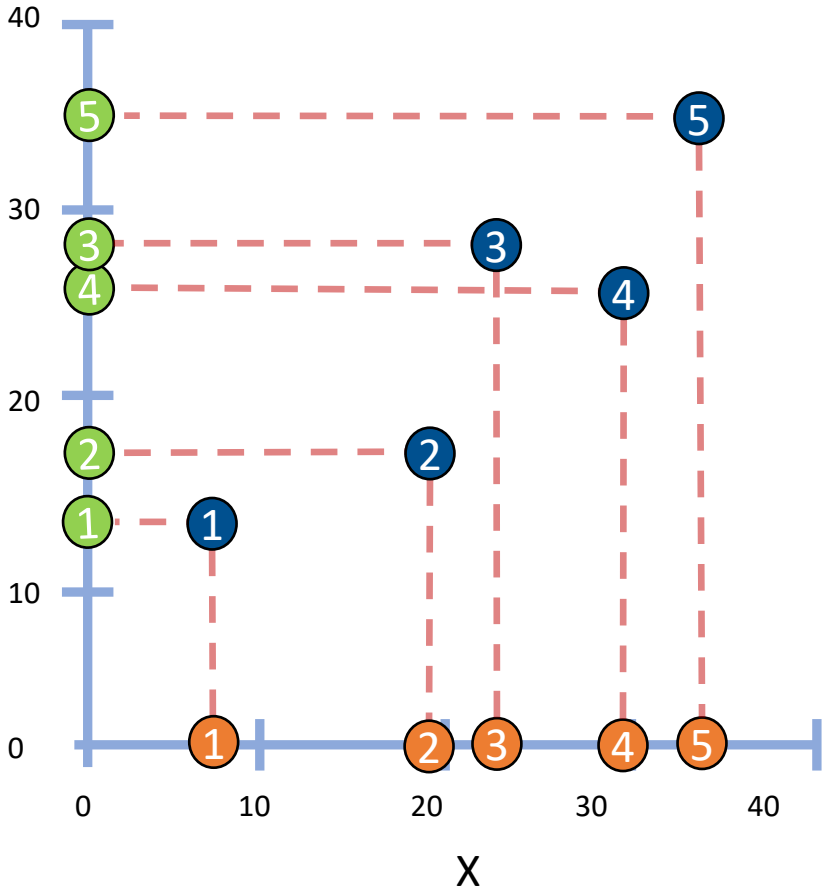
Example:

Five students' performance in class

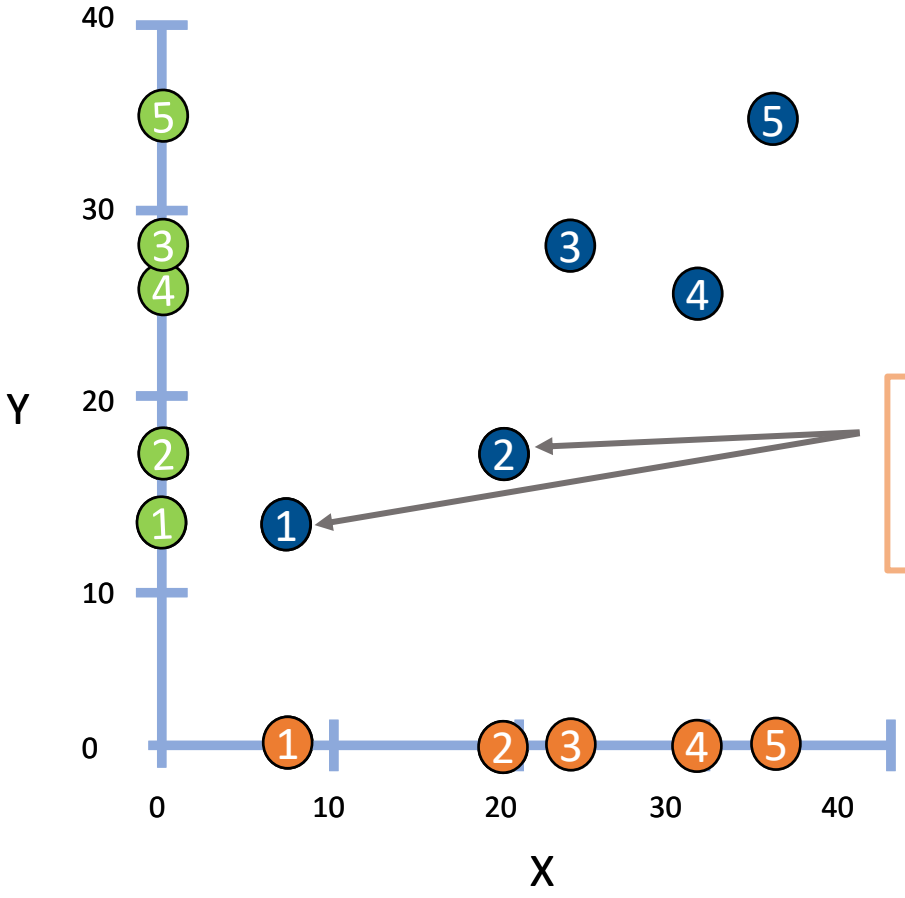
x = attendance score

y = final exam score

	1	2	3	4	5
x	7	19	23	29	33
y	14	16	28	27	35

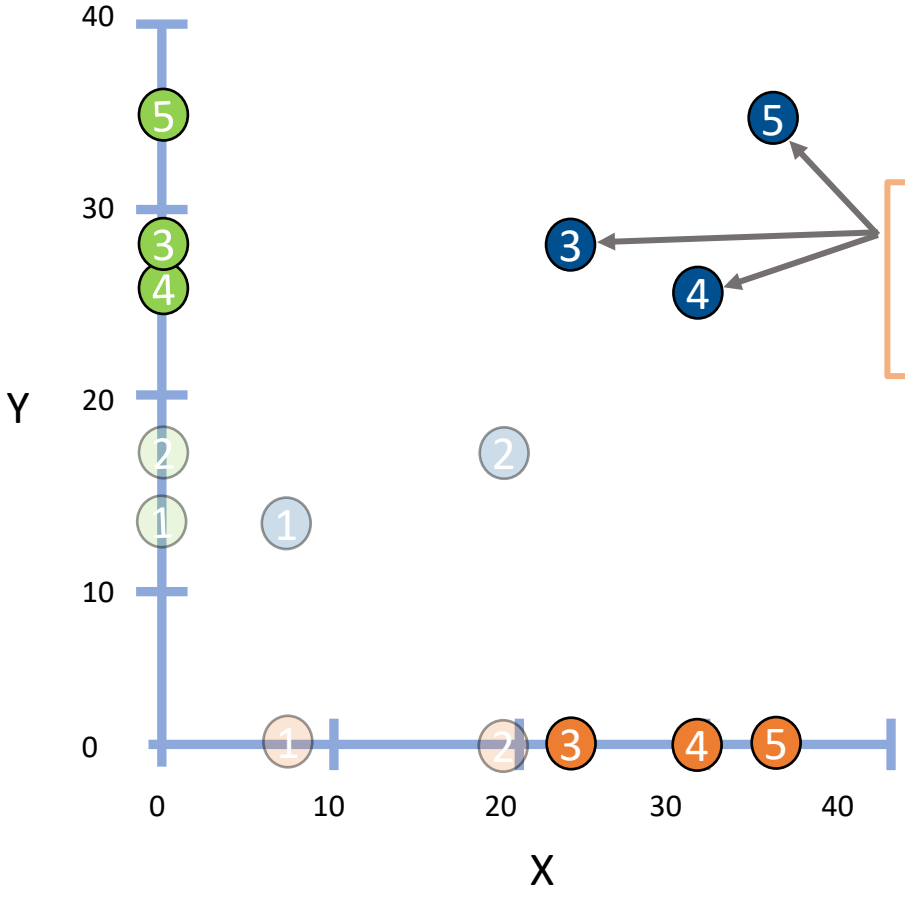


Trend



Relatively low attendance paired with relatively low final score.

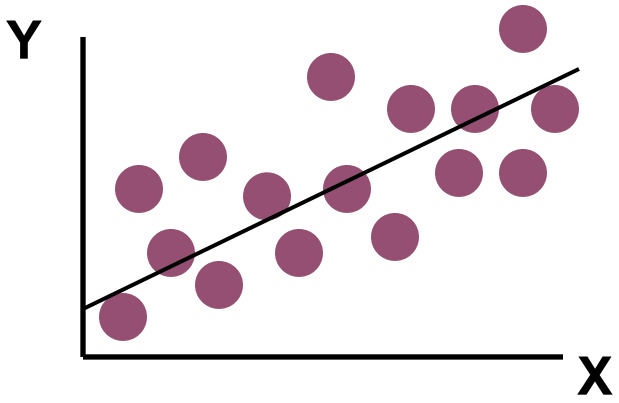
Trend

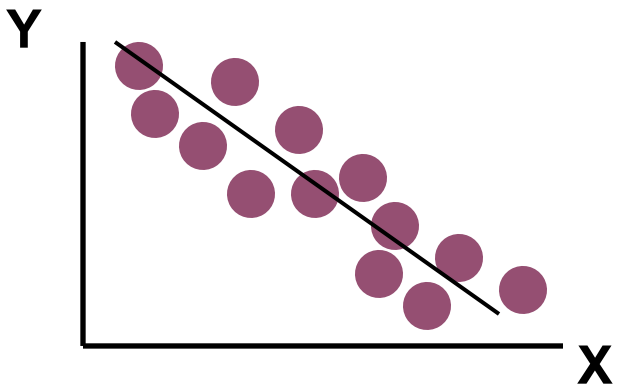


Relatively high attendance paired with relatively high final score.

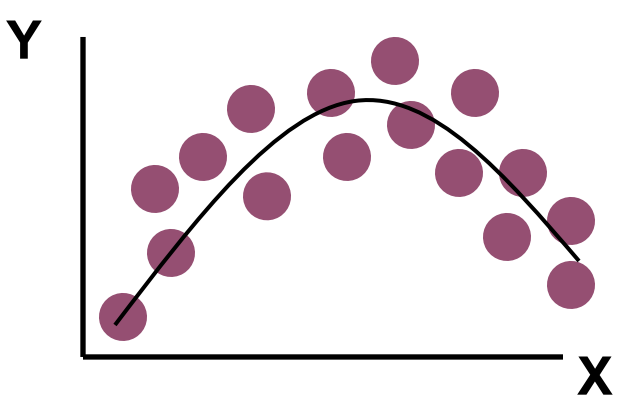
Types of Relationships (Association)

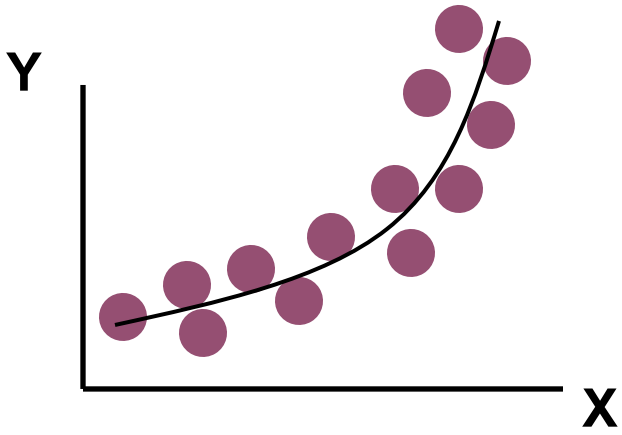
Linear relationships





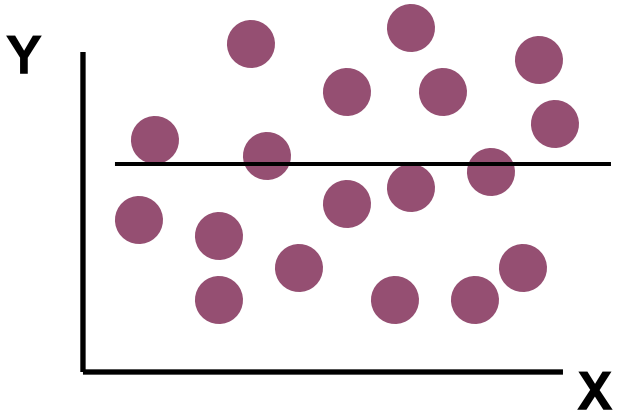
Curvilinear relationships

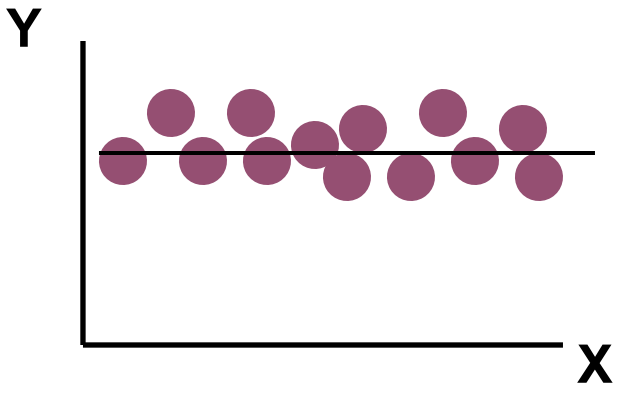




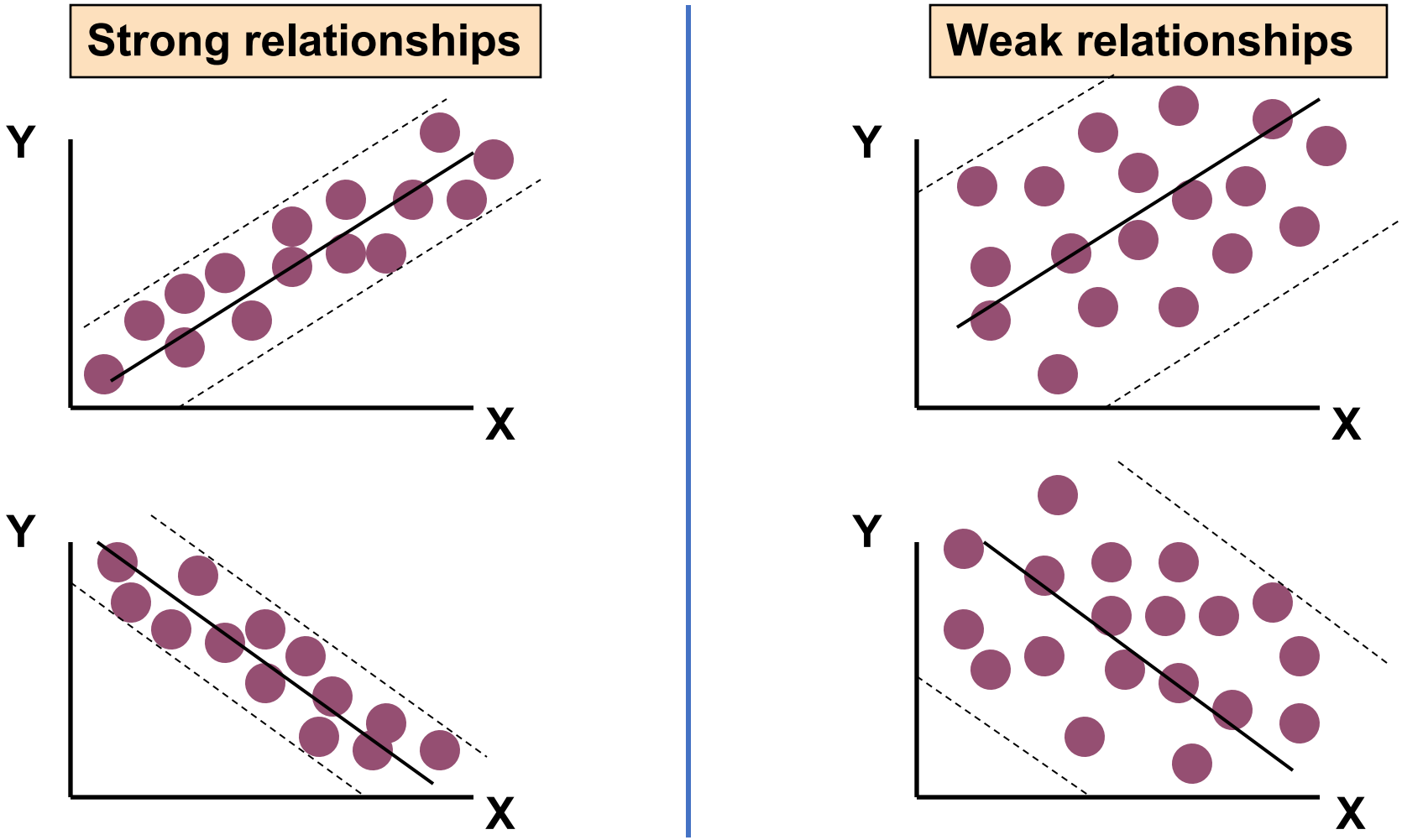
Types of Relationships

No apparent relationship



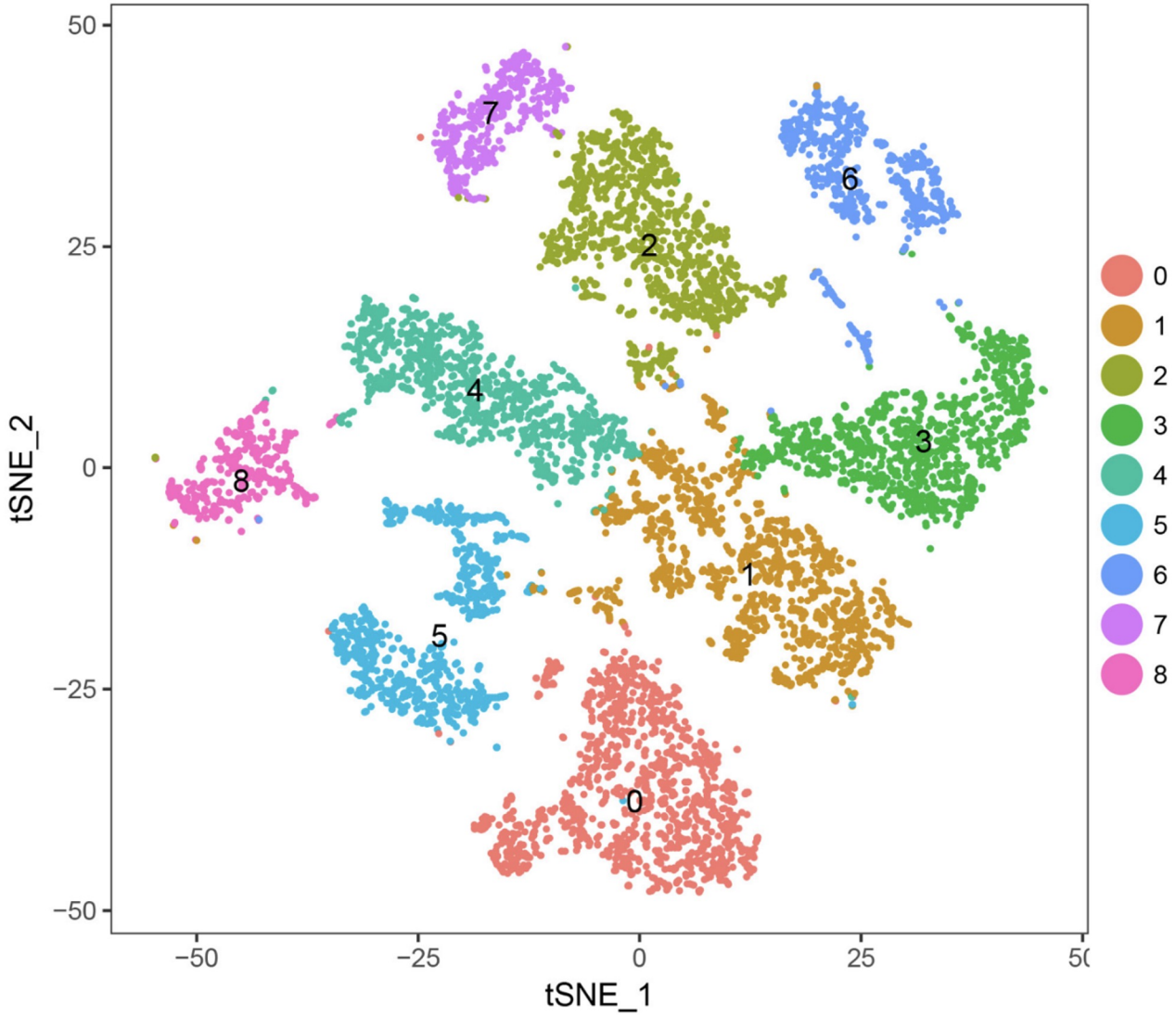


Types of Relationships



Linear relationship can be described by **correlation**

Clusters



Correlation (r)

- Measures the direction and **strength** of the **linear relationship** between two numerical variables
- Is always between –1 and 1
- The strength increases as you move away from 0 to either –1 or 1

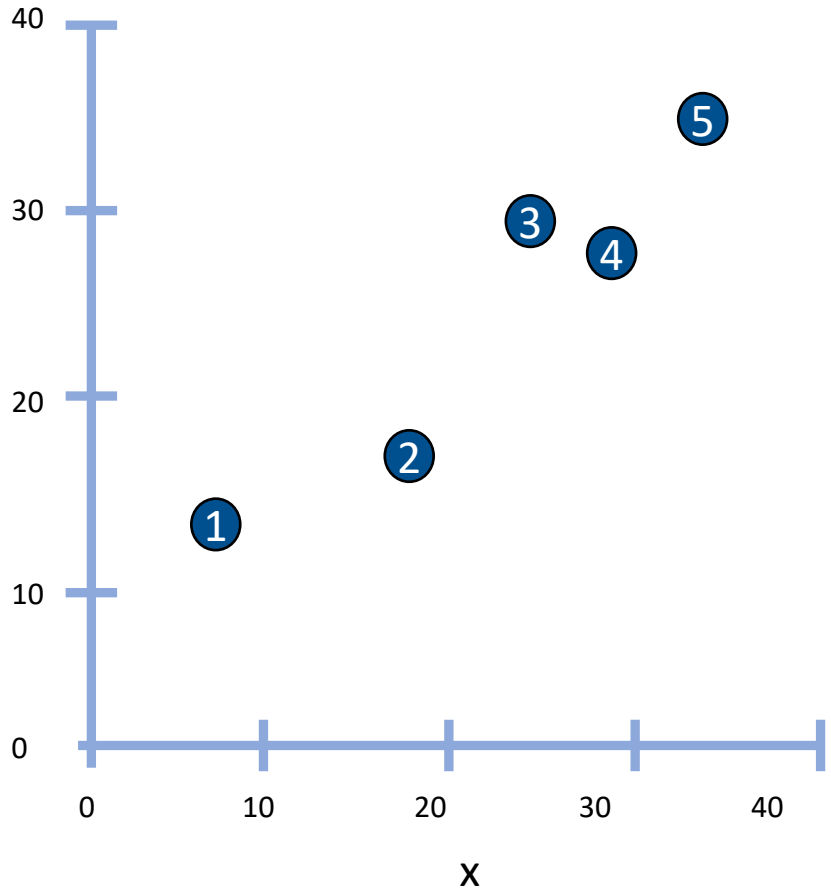
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Example

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

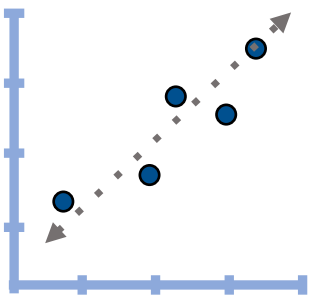
	1	2	3	4	5
x	7	19	23	29	32
y	14	16	28	27	35

r = 0.96

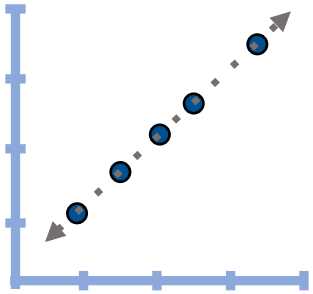


The magnitude of correlation

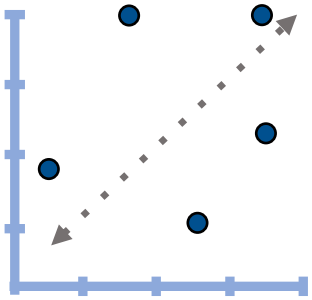
Rank the plots based on the magnitude of correlation:



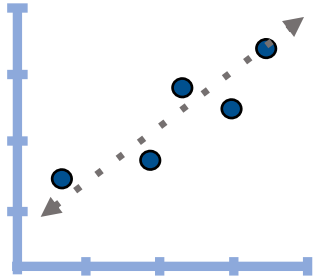
A



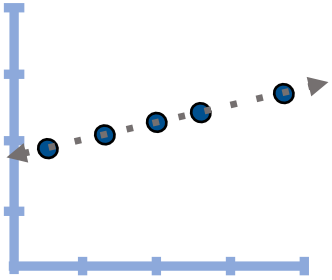
B



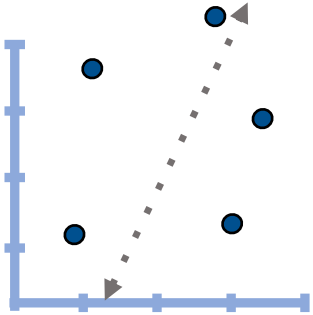
C



A



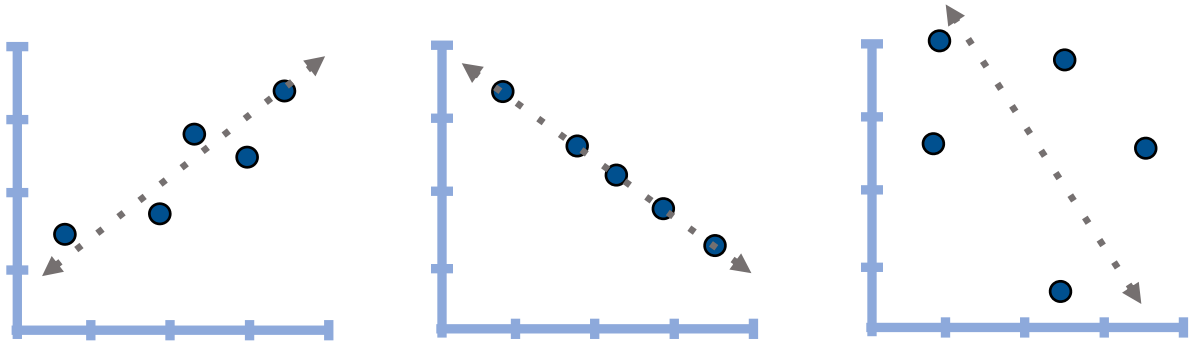
B



C

- The **larger** the absolute value of the correlation, the **stronger** the linear relationship is.
- The strength of linear relationship refers to how well a **line** depicts the relationship between X and Y.
- Perfect linear relationship have the absolute value of the correlation equal to 1.

Correlation



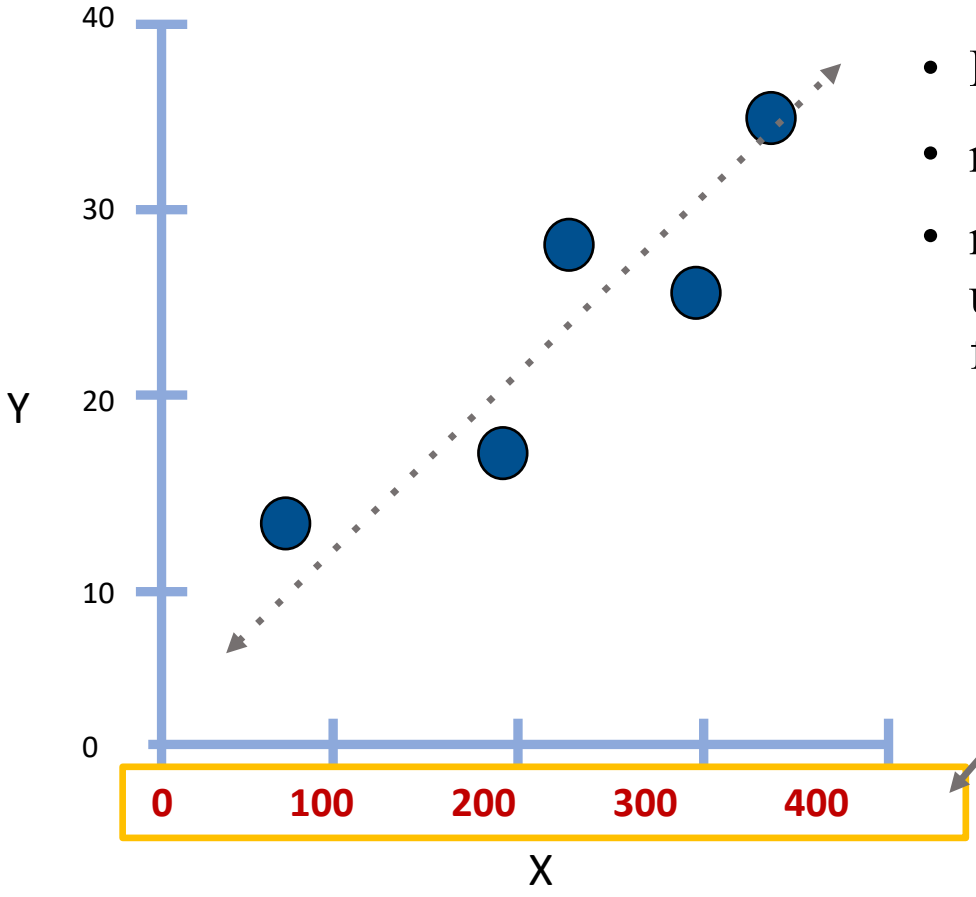
Correlation： A > C > B

The absolute value of the correlation: B> A > C

Magnitude ⟶ the strength of linear relationship

Sign ⟶ the direction of the relationship (increase or decrease)

Notes about correlation

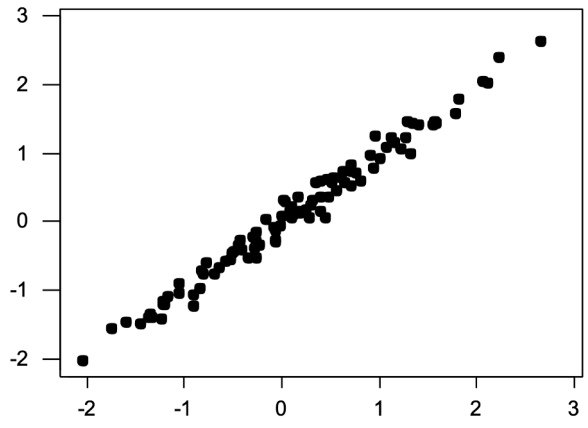


- Both variables have to be numerical
- r has no units of measurement
- r does not change if you change the units of measurement of the data (e.g., from *lbs* to *kg*)

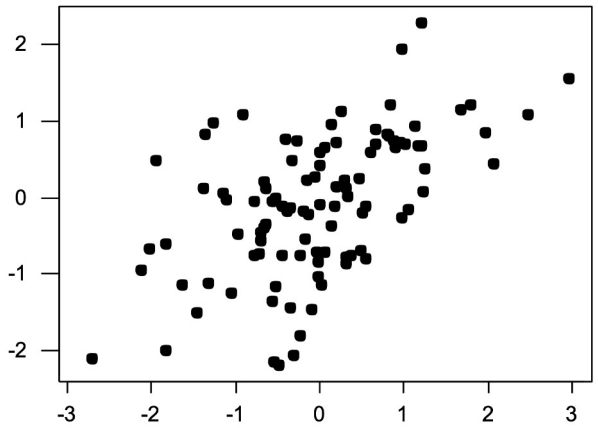
Correlation will not change

Question

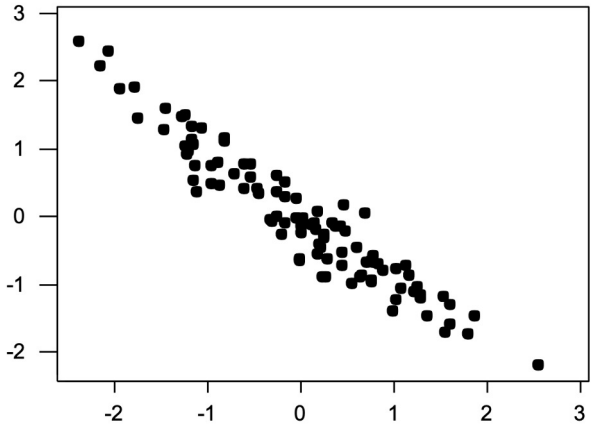
A



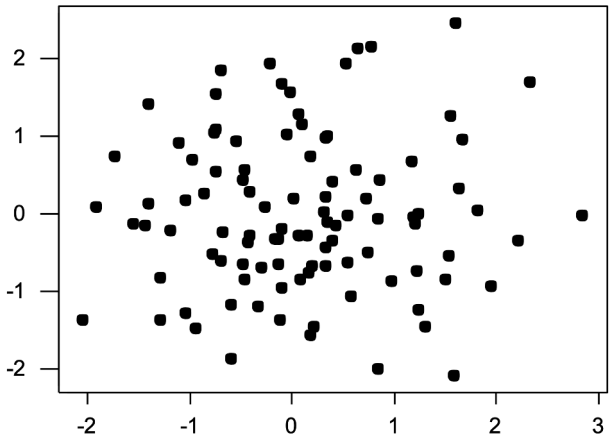
B



C

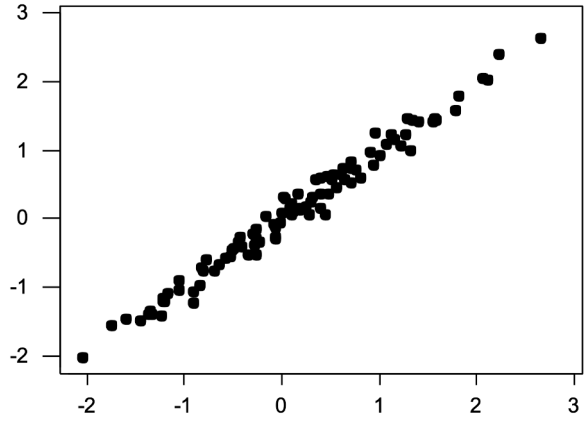


D

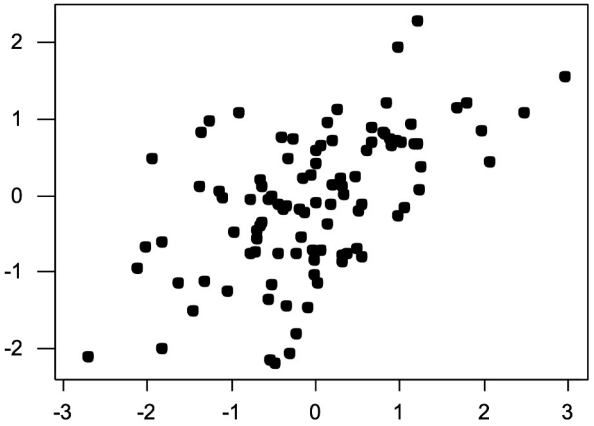


Question

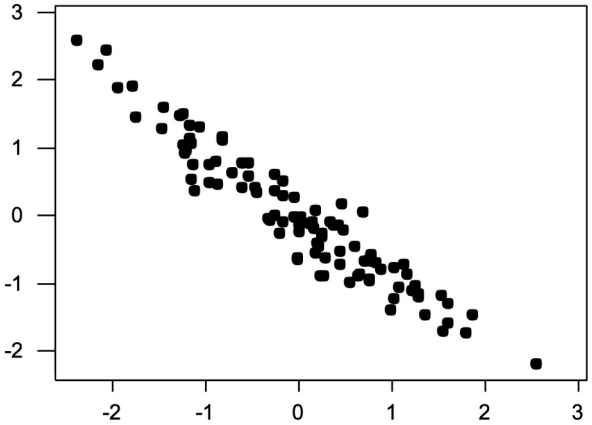
A. R=0.99



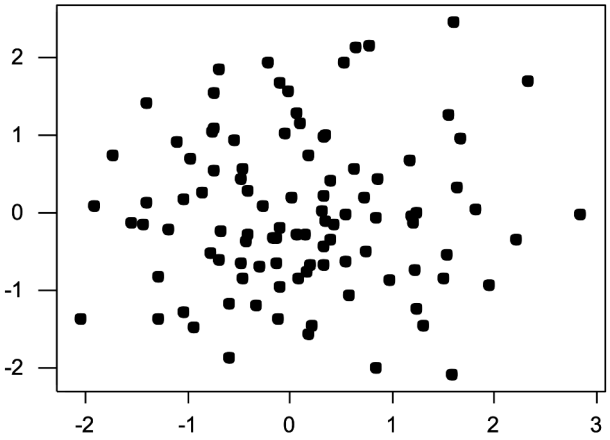
B. R=0.55



C. R=-0.96



D. R=0



A>B>D>C

Example

1971 study: people who drink coffee a lot have higher incidence of bladder cancer

Correlation noticed. Causation?

Example

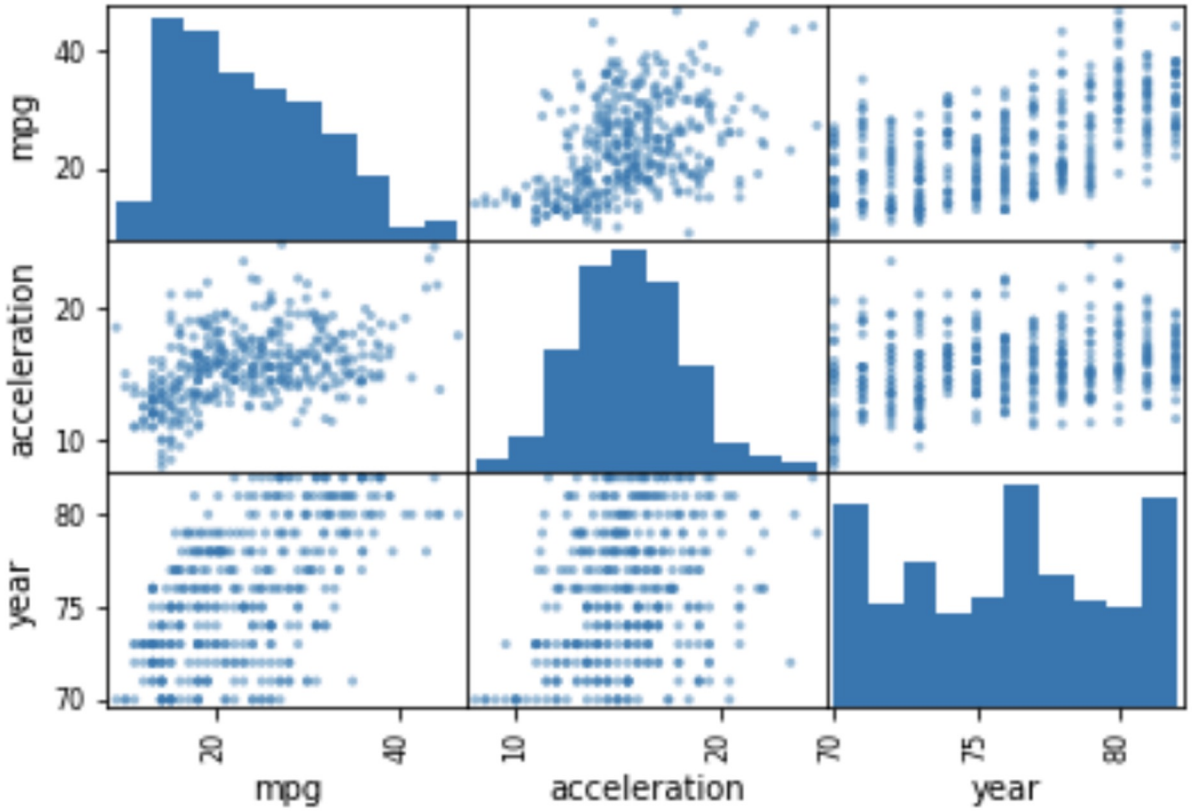
1993: A larger study concluded that after adjusting for the effects of smoking, no evidence was founded for increased risk from coffee.

Correlation does not imply causation

Multiple pairs of variables

Scatter plot matrix

Dataset: auto.csv



Multiple pairs of variables

Correlation matrix

Dataset: auto.csv

```
#note that 'origin' is also not quantitative, we need to remove it
df1 = df.drop(columns=['name','origin']) #remove the selected feature
df1.corr()
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	year
mpg	1.000000	-0.777618	-0.805127	-0.778427	-0.832244	0.423329	0.580541
cylinders	-0.777618	1.000000	0.950823	0.842983	0.897527	-0.504683	-0.345647
displacement	-0.805127	0.950823	1.000000	0.897257	0.932994	-0.543800	-0.369855
horsepower	-0.778427	0.842983	0.897257	1.000000	0.864538	-0.689196	-0.416361
weight	-0.832244	0.897527	0.932994	0.864538	1.000000	-0.416839	-0.309120
acceleration	0.423329	-0.504683	-0.543800	-0.689196	-0.416839	1.000000	0.290316
year	0.580541	-0.345647	-0.369855	-0.416361	-0.309120	0.290316	1.000000

Multiple pairs of variables

Heatmap: correlation

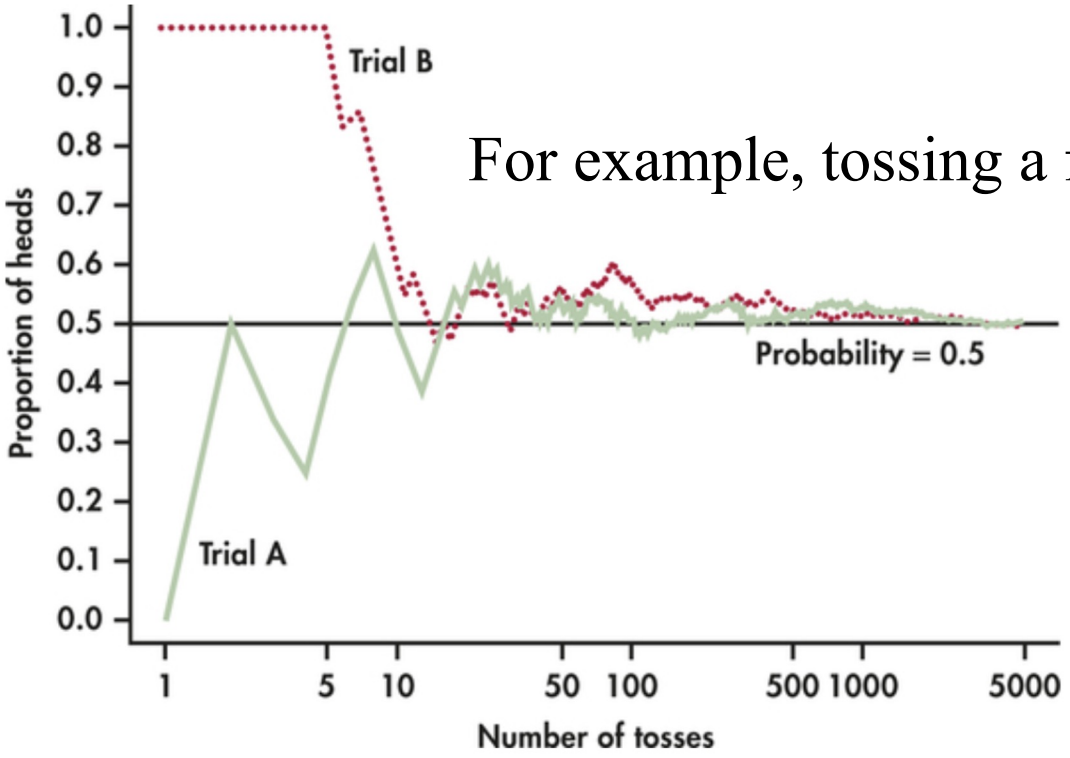
Dataset: auto.csv

Statistics and probability

- *Statistics* enables us to make decisions/inference under uncertainty.
- By using *Probability*, we can make numerical statements about uncertainty.
- By uncertainty, we mean *randomness(defined on the next slide)*.

Randomness

- Randomness ≠ complete chaos!
- A phenomenon is **random** if individual outcomes are uncertain but outcomes have *a regular pattern in a **large** number of repetitions*



For example, tossing a fair coin

Probability Models

- Any process that results in an *outcome* is an *experiment*.
- An experiment may have more than one possible outcome.

S = **sample space** = set of *all* possible outcomes.

E.g. Experiment: toss a coin once;

Outcomes: H, T;

$S = \{H,T\}$

Probability Models

An **event** is a collection of *some* outcomes

E.g., A = (get exactly one head in 3 tosses)
= {HTT, THT, TTH}

Each event is assigned a **probability**, *i.e.,* a number between 0 and 1.

If A is an event, then ***P(A)*** denotes the probability of A.

Equally-likely Case

When all possible outcomes are equally likely,

$$P(A) = \frac{\#\,\text{outcomes in } A}{\#\,\text{outcomes in } S}$$

E.g. tossing a coin once, with $S = \{H, T\}$.

If $A = \{H\}$, then

$$P(A) = \frac{1}{2}$$

Equally-likely Case

E.g. roll two dice. What is the probability of getting a total of at least 11?

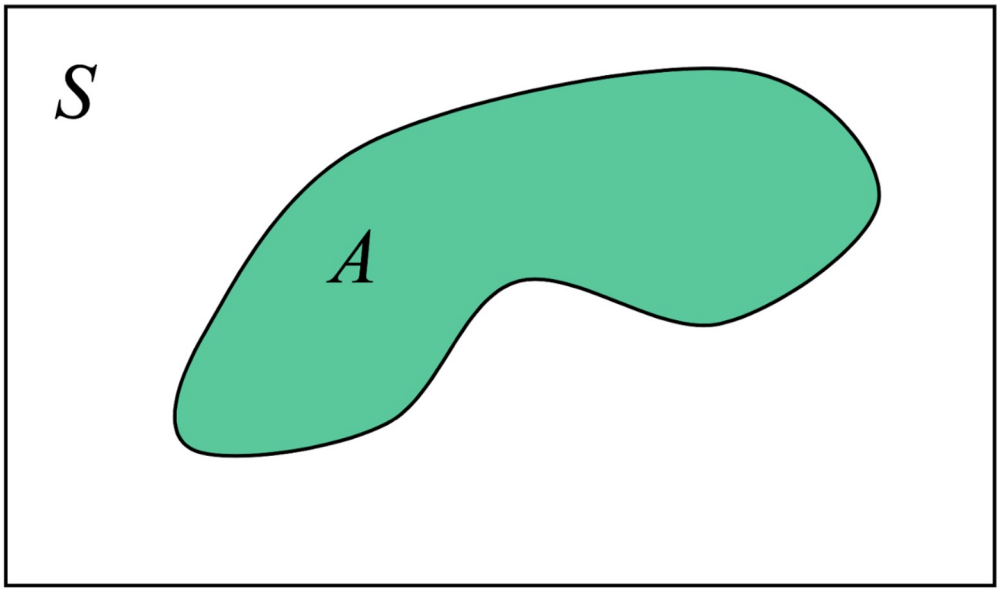
Here is the sample space S

36 outcomes, equally likely, each with 1/36 probability

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

$$P\{\text{total at least } 11\} = 3/36 = 0.083$$

A useful picture/example of probability



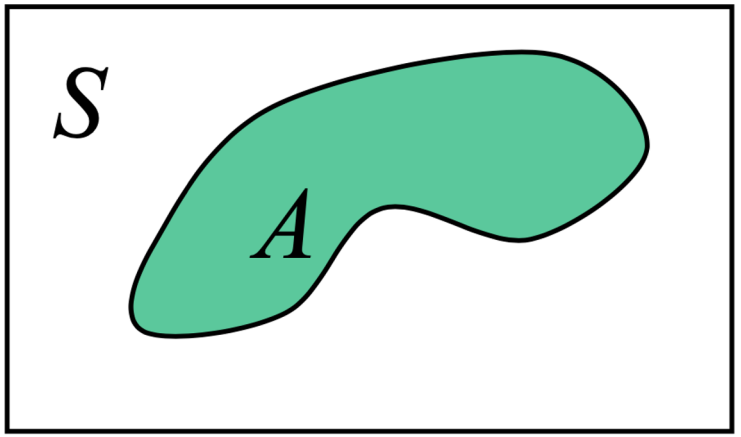
Venn diagram

S is the sample space, A is an event

You are driving and it's about to start raining. Think of S as your windshield. Event A corresponds to statement {the first drop to hit the windshield will hit the set A}.

A useful picture/example

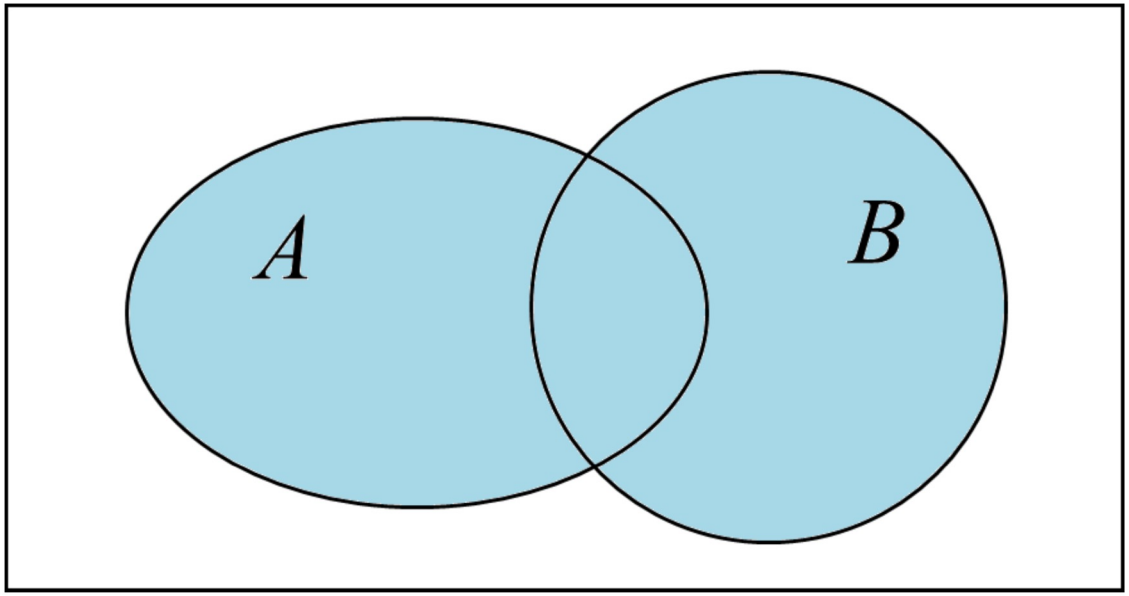
A simple probability measure to model this:



$$P(A) = \frac{\text{area of } A}{\text{area of } S}$$

Note that $0 \le P(A) \le 1$ and $P(S) = 1$

New events from old

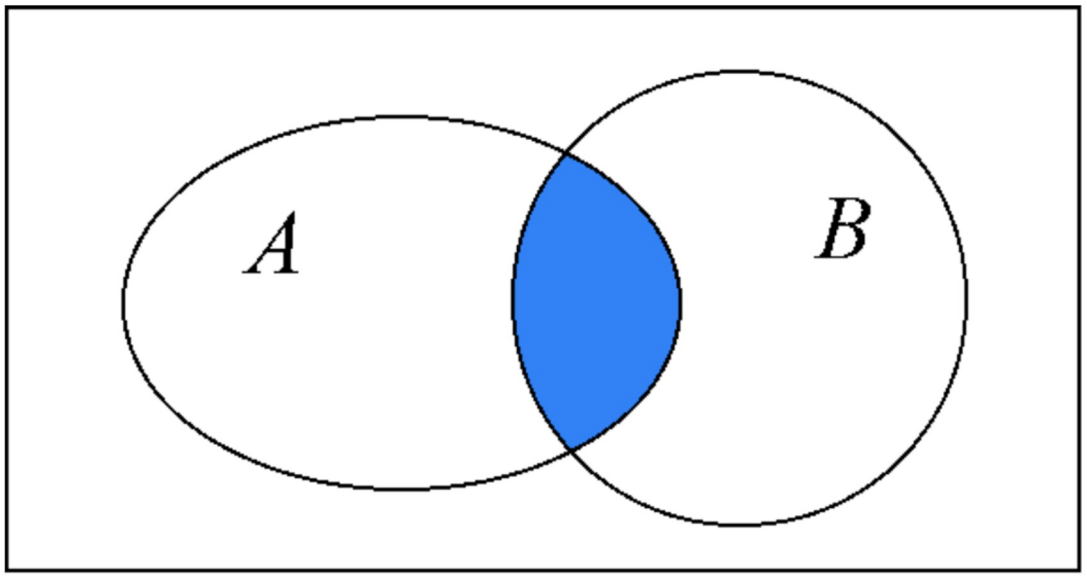


What should we call this?

A and B ?

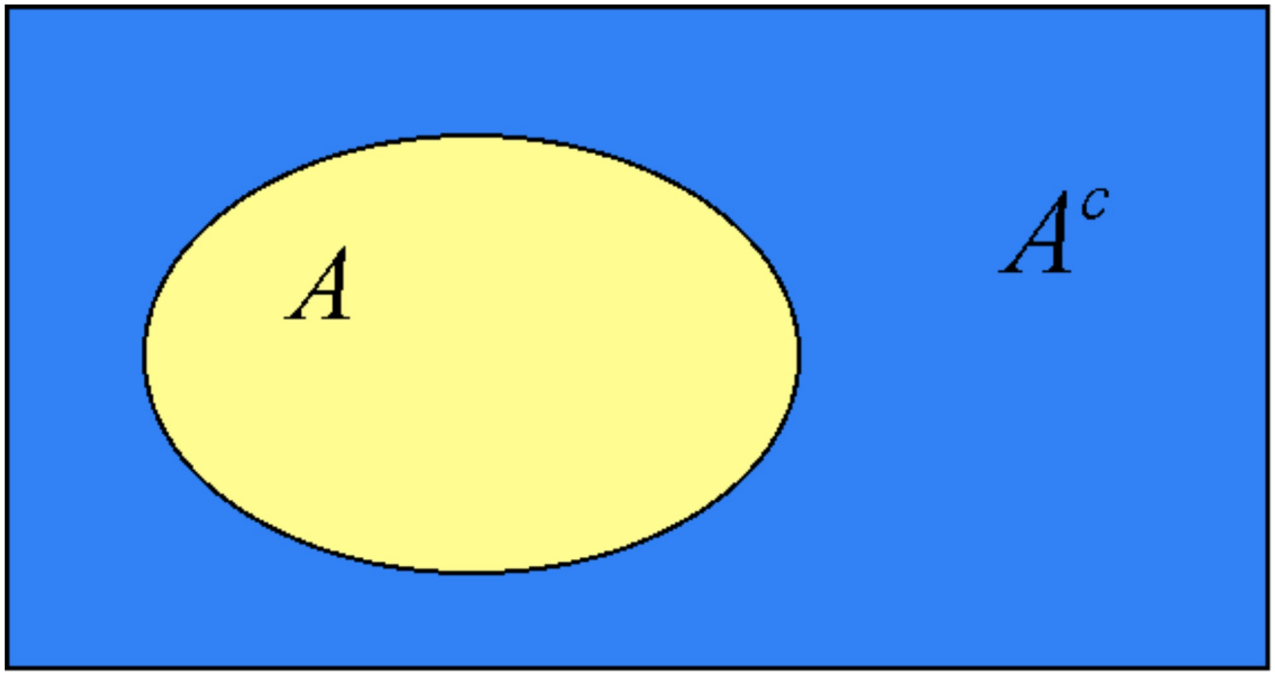
A or B ?

A and B



(raindrop falls in A) *and* (raindrop falls in B)

Complement of A?

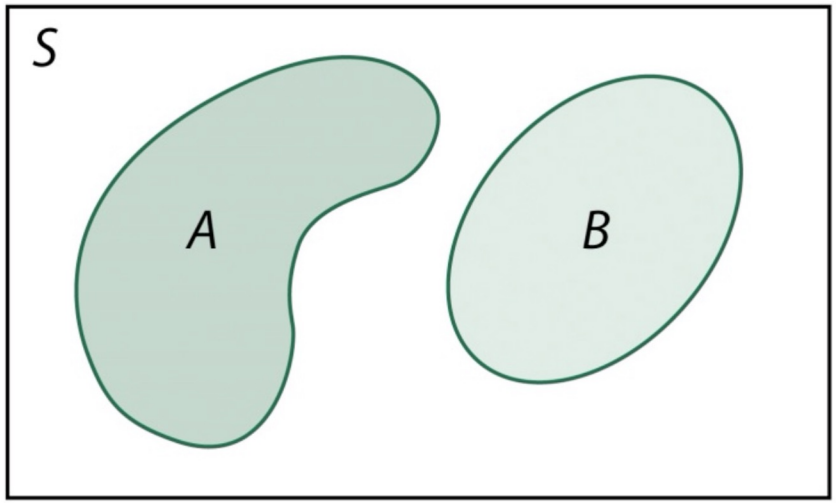


$$\text{"complement of } A\text{"} = \text{"not } A\text{"} = A^c$$

Rules of Probability

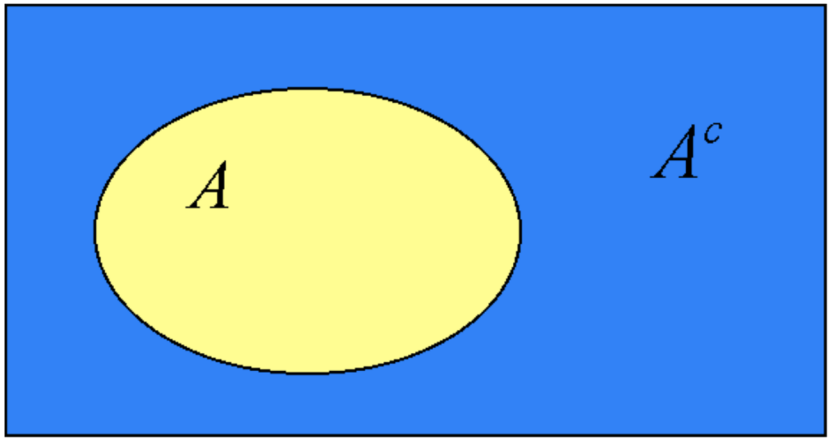
- For every event A, $P(A) \geq 0$ and $P(A) \leq 1$.
- $P(S) = 1$, where S is the sample space.
- If events A and B are **disjoint**, then

$$P(A \text{ or } B) = P(A) + P(B) .$$



Complement rule

$$P(A^c) = 1 - P(A)$$



Why?

$$(A \text{ or } A^c) = S$$

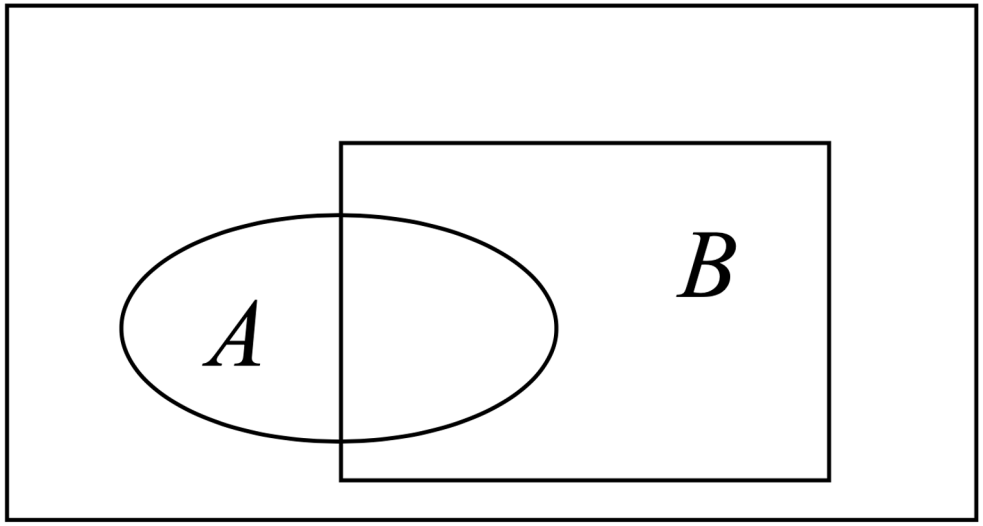
$$\text{So } P(A \text{ or } A^c) = P(S) = 1$$

$$A \text{ and } A^c \text{ are disjoint.}$$

$$\text{So } P(A \text{ or } A^c) = P(A) + P(A^c)$$

$$\text{So } P(A) + P(A^c) = 1.$$

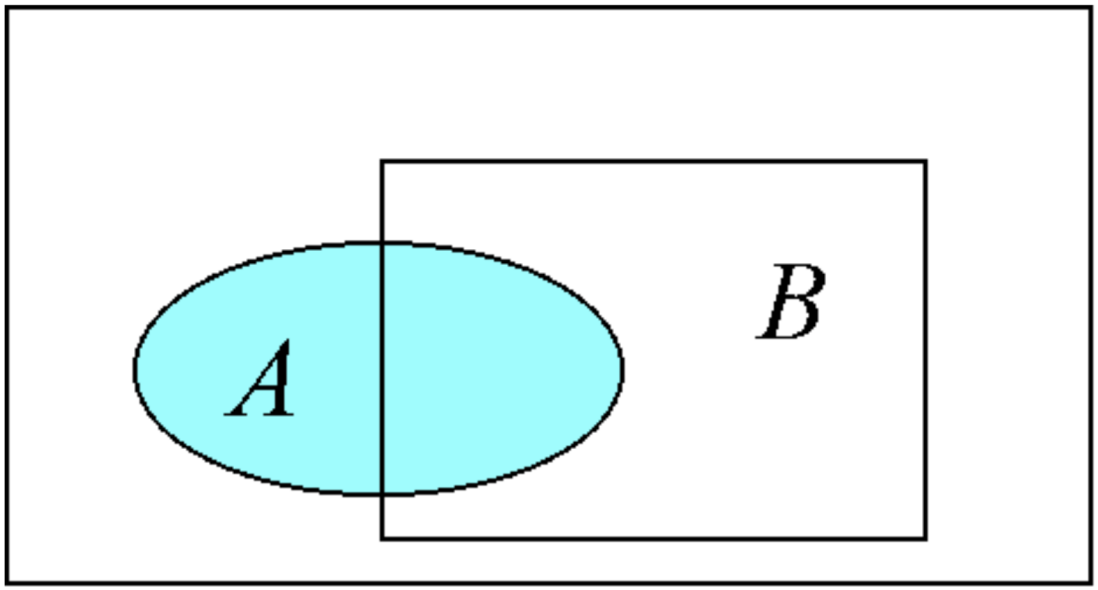
Conditional Probability P(B | A)



Idea of $P(B|A)$: *Given* that A occurs, what is the probability that B also occurs?

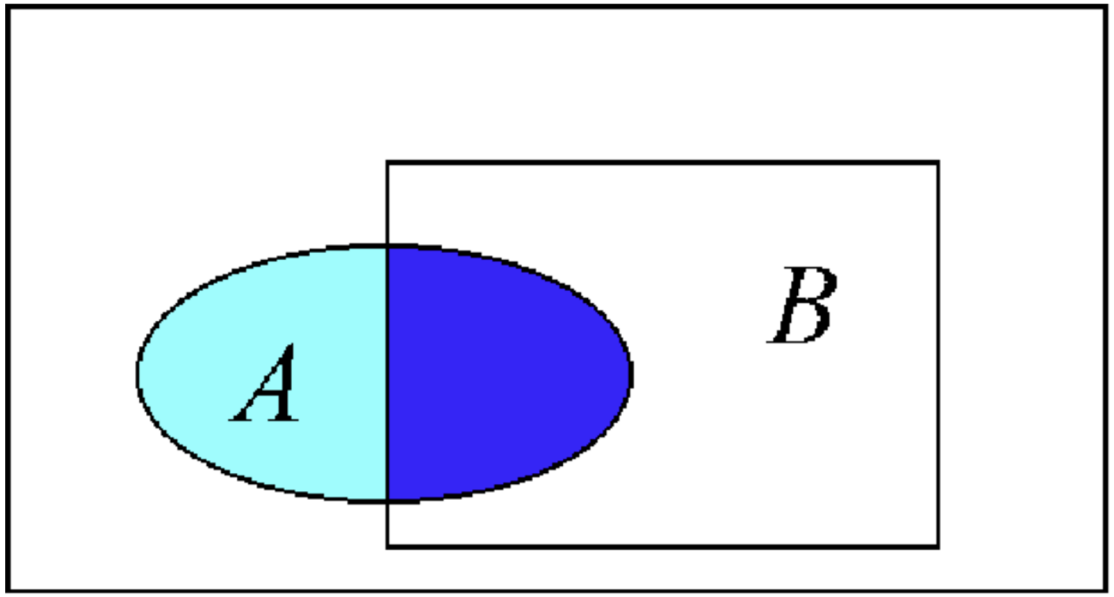
Question: By eyeball, what is $P(B|A)$?

Conditional Probability P(B | A)



Given that the raindrop fell in A, we restrict our attention to the set A. The drop is equally likely to fall anywhere within A.

Definition of P(B | A)



$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

Independence

Two events A and B are **independent** if knowing one event's occurrence does not change the probability of the other event.

i.e. $P(B \mid A) = P(B)$

i.e. $\dfrac{P(A \text{ and } B)}{P(A)} = P(B)$

i.e. $P(A \text{ and } B) = P(A)P(B)$

E.g. Experiment: two tosses of a coin

A=(get H in the first toss)

B=(get H in the second toss)

Multiplication Rule

$$P(A \text{ and } B) = P(A)P(B)$$

the **multiplication rule for independent events**

Random Variable (r.v.)

Random variable is a variable whose value depends on the outcome of an experiment (e.g., coin tossing)

- *Random variable* assigns a value (typically a number) to each possible outcome in the sample space S = {all possible outcomes}

Experiment: Toss a coin 3 times. X = # of "heads".

Outcome	TTT	TTH	THT	HTT	HHT	HTH	THH	HHH
Value of X	0	1	1	1	2	2	2	3

Random Variables

Discrete random variable: Possible values can be listed or counted

- e.g., the number of defective units in a batch of 20

Continuous random variable: May assume any numerical value in one or more intervals

- e.g., the waiting time for a credit card authorization

Discrete Random Variable

- **(probability) distribution** of a discrete random variable is a table, graph, or formula that gives the probability associated with each possible value
- probabilities of all possible values must sum to 1

e.g. X = number of heads in 3 tosses

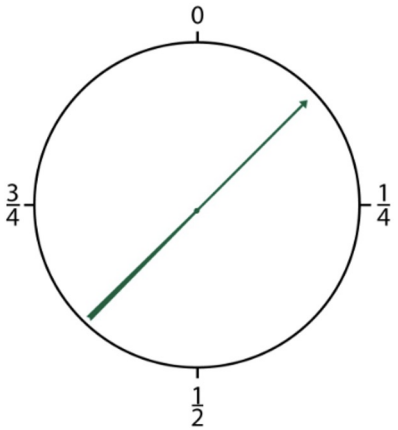
Value	0	1	2	3
Probability	1/8	3/8	3/8	1/8

Example (continuous random variable)

A spinner turns freely on its axis and slowly comes to a stop

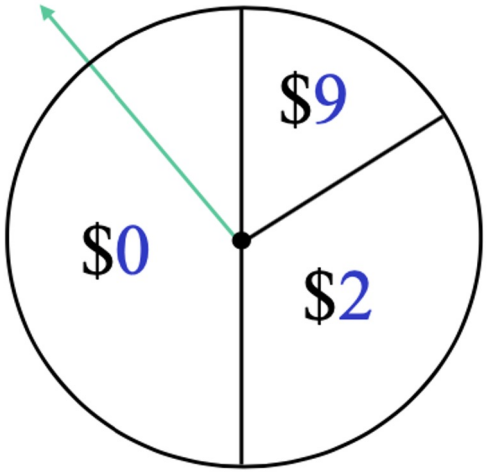
random variable X: location of the pointer
can be anywhere on a circle that is marked from 0 to 1

(does not favor any part of the circle)



Another r.v. defined on the same experiment

E.g. Y = payoff in spinner game



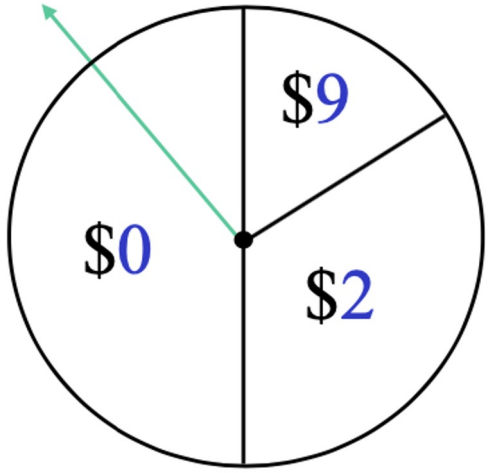
Distribution of Y:

$$Y = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases}$$

Discrete or continuous?

Expected Value (mean) of random variable

E.g. Y = payoff in spinner game



Distribution of Y:

$$Y = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases}$$

notation: expected value (mean) of $Y = E(Y) = \mu$

$$E(Y) = 0(0.5) + 2(0.3) + 9(0.2) = 2.4 \text{ dollars}$$

Why is “mean” defined this way?

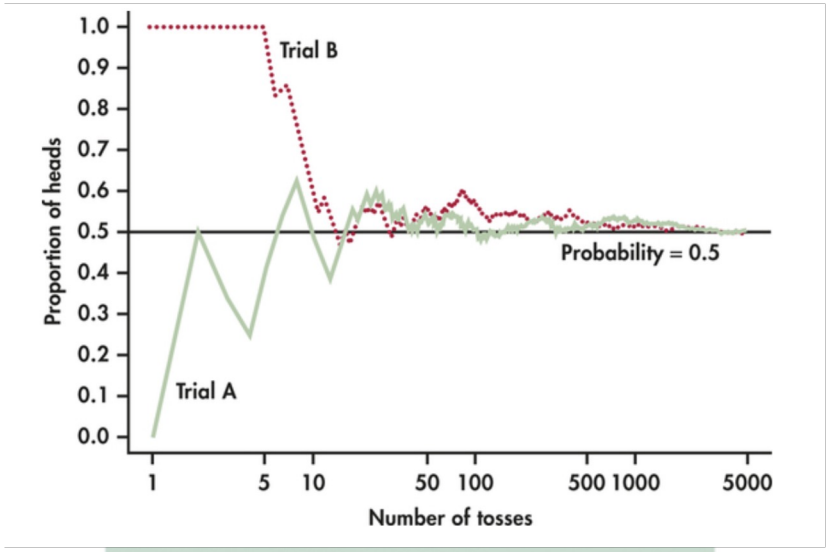
“long run frequency” interpretation of probability:

Suppose P(an event) = p (e.g. 0.5). Define

$$F_n = \frac{\#\text{ times this event occurs in } n \text{ independent trials}}{n}$$

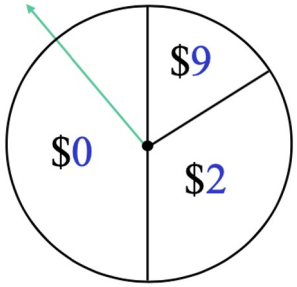
As n increases, this fraction will approach p :

$$F_n \to p$$



Why is “mean” defined this way?

Imagine playing the game n times (n large).
Total payoff:



$$x_1 + x_2 + \cdots + x_n = 0(\#\text{ of }0\text{'s}) + 2(\#\text{ of }2\text{'s}) + 9(\#\text{ of }9\text{'s})$$

Average payoff:

$$\bar{x}_n = 0\left(\frac{\#\text{ of }0\text{'s}}{n}\right) + 2\left(\frac{\#\text{ of }2\text{'s}}{n}\right) + 9\left(\frac{\#\text{ of }9\text{'s}}{n}\right)$$

$$\underbrace{\phantom{\frac{\#\text{ of }0\text{'s}}{n}}}_{\downarrow\ 0.5} \qquad \underbrace{\phantom{\frac{\#\text{ of }2\text{'s}}{n}}}_{\downarrow\ 0.3} \qquad \underbrace{\phantom{\frac{\#\text{ of }9\text{'s}}{n}}}_{\downarrow\ 0.2}$$

i.e. $\quad \bar{x}_n \to 0(0.5) + 2(0.3) + 9(0.2) = E(X)$

Law of large numbers

As we do many independent repetitions of the experiment, drawing more and more numbers from the same distribution, the mean of our sample will approach the mean of the distribution more and more closely:

$$\overline{X}_n \to E(X) \quad \text{as } n \text{ increases}$$

The "information" we can observe

The "truth" we want to learn

General formula for discrete r.v.'s

If X has distribution

value of X	v_1	v_2	…	v_k
probability	p_1	p_2	…	p_k

Then $E(X) = \mu = v_1p_1 + \ldots + v_kp_k$

Example

- Suppose a day trader buys one share of IBM
- Let X represent the change in price of IBM
- She pays \$100 today, and the price tomorrow can be either \$105, \$100, or \$95

Stock Price	Change x	Probability $P(X = x)$
Increases	\$5	0.11
Stays same	0	0.80
Decreases	−\$5	0.09

Question

Suppose you buy one share of IBM today ($100). How much are you expected to earn tomorrow?

A. -5

B. 0

C. 0.1

D. 1

Stock Price	Change x	Probability $P(X = x)$
Increases	$5	0.11
Stays same	0	0.80
Decreases	−$5	0.09

Mean of X

Stock Price	Change x	Probability $P(X = x)$
Increases	\$5	0.11
Stays same	0	0.80
Decreases	−\$5	0.09

$$\mu = -5\,P(-5) + 0\,P(0) + 5\,P(5)$$
$$= -5(0.09) + 0(0.80) + 5(0.11)$$
$$= \$.10$$

SD and Variance of random variable

Notation: $SD(X) = \sigma$ $Var(X) = \sigma^2$

Definition: $Var(X) = E\,(X - \mu)^2$

Expected value and SD: properties

Adding or Subtracting a Constant (c)

- Changes the expected value by a fixed amount:

$$E(X \pm c) = E(X) \pm c$$

- Does not change the standard deviation (SD):

$$SD(X \pm c) = SD(X)$$

Expected value and SD: properties

Multiplying by a Constant (c)

- $E(cX) = c\,E(X)$
- $SD(cX) = |c|\,SD(X)$

Question?

McDonald’s has a monthly return with mean 0.53% and SD 7.6%, and Disney has a monthly return with mean 0.61% and SD 8.3%. As an investor, how would you choose between McDonald's and Disney stocks?

A. McDonald’s
B. Disney
C. Neither as both are risky

The Sharpe Ratio

- Popular in finance for comparing investments: the higher the Sharpe ratio, the better the investment
- Is the ratio of an investment's net expected gain to its standard deviation

$$Sharpe(X) = \frac{\mu - r_f}{\sigma}$$

- μ and σ are the mean and SD of the return on the investment
- r_f stands for the return on a risk-free investment (e.g. interest rate on a savings account)

Example

$$Sharpe(X) = \frac{\mu - r_f}{\sigma}$$

- Summary of monthly returns in 2000-2006:

Company	Random Variable	Mean	SD
Disney	D	0.61%	8.3%
McDonald's	M	0.53%	7.6%

- Suppose the risk free rate is 0.4% per month

- Sharpe(D) = (0.61-0.4)/8.3 = 0.0253
- Sharpe(M) = (0.53-0.4)/7.6 = 0.0171

- Disney is preferred to McDonald's

Density Curves

- Density curves: A curve that
  - lies above the x-axis,
  - has total area 1 under the curve (and above the x-axis).

- Using density curves to describe continuous probability distributions (think about normal distribution)

Note

- Finish **QIZE 2** before 11:59pm.
- Read tutorial
  - It is highly relevant to HW1.
- HW1 will be posted after next week's lecture