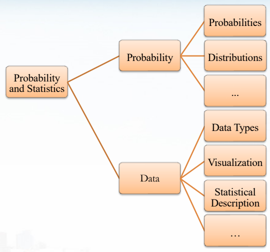
Lecture Notes on Pattern Recognition

Session 01(A): Probability and Statistics

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

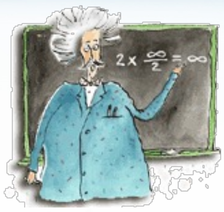
1

Road map



2

Probability Terms



- Random Variables
- Probability
- Tree Diagrams
- Marginal, Joint and Conditional Probability
- Independence and Conditional Independence

3

Random Phenomena Vocabulary

- A random phenomenon is a situation in which we know what outcomes *could* happen,
  - but we don't know which particular outcome *will* happen.
- Trial - Each occasion which we observe a random phenomena
- Outcome - The value of the trial for the random phenomena
- Event - The combination of the trial's outcomes
- Sample Space - The collection of all possible outcomes
- Example: Rolling a 6-sided die
  - Trial: e.g. A single roll could be one trial
  - Outcomes: Could roll a 1, 2, 3, etc
  - Events: Rolling a 1, but also "rolling a number larger than 3" is an event involving combined outcomes
  - Sample space: {1, 2, 3, 4, 5, 6}

- How about 2?
  - Trial: e.g. A single roll (of both dice) is one trial
  - Outcomes: 1 and 2, 3 and 4, two 6s, etc.
  - Events: Rolling a total of 7, rolling a 1 and a 2, rolling two 1s etc.
  - Sample space: (die1,die2):
  - {(1,1),(1,2),(1,3),(1,4),(1,5), (1,6), (2,1), (2,2),…, (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

4

Probability (contd..)

- The probability of an event is the number of outcomes in the event divided by the total number of possible outcomes. For a particular event, A:
- $P(A) = \frac{\#\ of\ outcomes\ in\ A}{\#\ of\ possible\ outcomes}$
  - Eg: Probability of rolling 3 on two dice…

(die1,die2):
{(1,1),(1,2),(1,3),(1,4),(1,5), (1,6),
(2,1), (2,2),…
…
(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

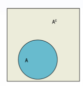
$P(rolling\ a\ 3) = \frac{2}{36}$

5

Formal Probability

- Two requirements for a probability:
  - A probability is a number between 0 and 1 (which can be expressed as a decimal, fraction or percentage).

    For any event A, $0 \le P(A) \le 1$.
  - Probability Assignment Rule:
    - The probability of the set of all possible outcomes of a trial must be 1.
    - $P(\mathbf{S}) = 1$ (S represents the set of all possible outcomes)
      - Eg: S = {1,2,3,4,5,6} when we roll a single die.
- Complement Rule:
  - The set of outcomes that are *not* in the event **A** is called the complement of **A**, denoted $\mathbf{A}^C$.
  - The probability of an event occurring is: 1 minus the probability that it doesn't occur
    $P(\mathbf{A}) = 1 - P(\mathbf{A}^C)$

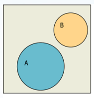


The set **A** and its complement.

6

Addition rule (for disjoint events)

- Addition Rule (for **disjoint** events):
  - Events that have **no outcomes in common** (and, thus, cannot occur together) are called disjoint (or mutually exclusive).



Two disjoint sets, A and B.

  - For two disjoint events **A** and **B**, the probability that one *or* the other occurs is the sum of the probabilities of the two events.
  - $P(\mathbf{A} \text{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$, provided that **A** and **B** are disjoint.

TULIP Team for Universal Learning and Intelligent Processing

7

Multiplication rule (for independent events)

- Multiplication Rule (for **independent** events):
  - For two **independent** events **A** and **B**, the probability that both **A** and **B** occur is the product of the probabilities of the two events.
  - $P(\mathbf{A} \text{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$, provided that **A** and **B** are **independent**.
  - Many Statistics topics require an **Independence Assumption**
    - but *assuming* independence doesn't make it true.
    - Always think carefully about whether that assumption is reasonable before using the Multiplication Rule.

Notation:
- In this text we use the notation $P(\mathbf{A} \text{ or } \mathbf{B})$ and $P(\mathbf{A} \text{ and } \mathbf{B})$.
- In other situations, you might see the following:
  - $P(\mathbf{A} \cup \mathbf{B})$ instead of $P(\mathbf{A} \text{ or } \mathbf{B})$
  - $P(\mathbf{A} \cap \mathbf{B})$ instead of $P(\mathbf{A} \text{ and } \mathbf{B})$

TULIP Team for Universal Learning and Intelligent Processing

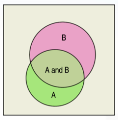
8

The General Addition Rule

- When two events **A** and **B** are disjoint, we can use the addition rule for disjoint events seen previously:

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

- However, when our events are not disjoint, this earlier addition rule will double count the probability of *both* **A** and **B** occurring. Thus, we need the General Addition Rule.
- General Addition Rule:
  - For any two events **A** and **B**,
    $P(\mathbf{A} \text{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$
    $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$
  - The Venn diagram shows a situation in which we would use the general addition rule:



TULIP Team for Universal Learning and Intelligent Processing

9

Conditional probability

- When we want the probability of an event from a *conditional* distribution, we write $P(\mathbf{B} | \mathbf{A})$ and pronounce it **"the probability of B *given* A."**
  - E.g. the probability that a random person from the Titanic passengers survived (B), given that they were male (A).
- A probability that takes into account a given condition is called a conditional probability.
- Example: If I am rolling a dice
  - What is the probability of getting an even number?
    P(even number) = **1/2**
  - What is the probability of getting a '2'?
    P(getting '2') = **1/6**
  - *If you are told that the number that I obtained is an* ***even number***, then what is the probability of getting a '2'?
    P(getting '2' | even number) = **1/3**

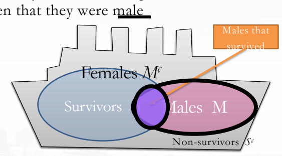
    - Note that P(getting '2' | even number) > P(getting '2')
    That is, knowing the information that an 'even number' has occurred, improved the probability of the event 'getting a '2'.

TULIP Team for Universal Learning and Intelligent Processing

10

Conditional probability (cont.)

- To find the probability of the event **B** *given* the event **A**, we restrict our attention to the outcomes in **A**. We then find in what fraction of *those* outcomes **B** also occurred.
  - E.g. the probability that a random person chosen from the Titanic passengers survived, given that they were male.



TULIP Team for Universal Learning and Intelligent Processing

11

Conditional probability (cont.)

- To find the probability of the event **B** *given* the event **A**, we restrict our attention to the outcomes in **A**. We then find in what fraction of *those* outcomes **B** also occurred.

$$P(B/A) = \frac{P(B \textbf{ and } A)}{P(A)}$$

- Note: $P(\mathbf{A})$ cannot equal 0, since we know that **A** has occurred.
  - In the Titanic example, we write

$$P(Survived\ /\ Male) = \frac{P(Survived \textbf{ and } Male)}{P(Male)}$$

TULIP Team for Universal Learning and Intelligent Processing

12

Example: Conditional probability

A table that displays the results of two categorical variables is called a **contingency table**.

For this table of values given, compute the following probabilities.

- P(girl) = 251/478 = 0.525
- P(girl *and* popular) = 91/478 = 0.190
- P(sports) = 90/478 = 0.188

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

TULIP Team for Universal Learning and Intelligent Processing

13

Conditional probability

- What if we knew the chosen person was a girl? Would that change the probability that the girl's goal was sports?
  - Yes! We write P(sports | girl)
  - Only look at Girl row: P(sports | girl) = 30/251 = 0.120
  - Find the probability of selecting a boy given the goal is grades.
  - P(boy | grades) = 117/247 = 0.474

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

TULIP Team for Universal Learning and Intelligent Processing

14

Conditional Probability Formula

- Conditional Probability Formula: Probability of **B** *Given* **A**:

$$\boxed{P(\mathbf{B} \mid \mathbf{A}) = \frac{P(\mathbf{A} \textit{ and } \mathbf{B})}{P(\mathbf{A})}}$$

- Example: P(girl | popular) =

$$= \frac{P(\text{girl } \textit{and} \text{ popular})}{P(\text{popular})} = \frac{91/478}{141/478} = \frac{91}{141} = 0.65$$

- NOTE:
  - You can get the same results for P(girl | popular) if you use the table values directly as done before:

P(girl | popular) = 91/141

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

Sex	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

TULIP Team for Universal Learning and Intelligent Processing

15

The General Multiplication Rule (cont.)

- When two events **A** and **B** are independent, we can use the multiplication rule for independent events:

P(**A** *and* **B**) = P(**A**) x P(**B**)

- However, when our events are not independent, this earlier multiplication rule does not work. Thus, we need the General Multiplication Rule.
- We encountered the general multiplication rule in the form of conditional probability.

$$\boxed{P(\mathbf{B}|\mathbf{A}) = \frac{P(\text{A and B})}{P(\mathbf{A})}}$$

TULIP Team for Universal Learning and Intelligent Processing

16

The General Multiplication Rule (cont.)

- We encountered the general multiplication rule in the form of conditional probability.

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

- Rearranging the equation in the definition for conditional probability, we get the General Multiplication Rule:
  - For any two events **A** and **B**,

P(**A** *and* **B**) = P(**A**) x P(**B** | **A**)

or

P(**A** *and* **B**) = P(**B**) x P(**A** | **B**)

TULIP Team for Universal Learning and Intelligent Processing

17

Independence

- Independence of two events means that the outcome of one event does not influence the probability of the other.
- With our new notation for conditional probabilities, we can now formalise this definition:
  - Events **A** and **B** are independent whenever P(**B** | **A**) = P(**B**).
    (Equivalently, events **A** and **B** are independent whenever P(**A** | **B**) = P(**A**))

TULIP Team for Universal Learning and Intelligent Processing

18

Independent ≠ Disjoint

Disjoint events cannot be independent! Why not?

- Since we know that disjoint events have no outcomes in common, knowing that one occurred means the other didn't.
- Thus, the probability of the second occurring changed based on our knowledge that the first occurred.
- It follows, then, that the two events are ***not* independent**.

Example:

Consider that you can only get one of the two grades for a course that you are doing, that is either grand 'A' or grade 'B' but not both

- Here A and B are Disjoint: because you can't get both.
- Not independent: Why?
  - $P(\mathbf{A}) = ½$,
  - $P(\mathbf{A} \mid \mathbf{B}) = 0$ because given that you got a grade B, then definitely you cannot get the grade 'A'.
  - Therefore, $P(\mathbf{A} \mid \mathbf{B}) \neq P(\mathbf{A})$. Hence they are not independent.
- Here, A and B are disjoint (also called mutually exclusive) but not independent.

TULIP Team for Universal Learning and Intelligent Processing

19

Conditional probability formulae

- The **Conditional probability** formulae

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

$$\therefore\ P(A \cap B) = P(A/B)P(B)$$

– P(A and B) or $P(A \cap B)$ is called as joint probability, denoted as P(A,B)

$$P(A/B) = \frac{P(A,B)}{P(B)}$$

TULIP Team for Universal Learning and Intelligent Processing

20

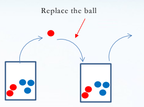
Drawing 'with' or 'Without' Replacement

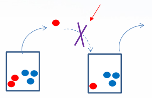
- Consider that a box contains 2 red and 3 blue balls.
- Consider a scenario as follows (Scenario-1):

  In the first instance, I take a ball from the box randomly, note down its color and then put it back in the box. Then, I am taking a ball again for the second time from the box.
  - This experiment is called "Drawing **with** replacement" as I have put the first ball back in the box before making the second attempt.
  - In this case the probability of getting a , say, red ball will be same for both the attempts. The probability of getting a red ball here is 2/5.

- Consider an alternative scenario as follows (Scenario-2):.

  In the first instance, I take a ball from the box randomly, and throw it away (that is, I am not putting it back in the box again). Then, I am taking a ball again for the second time from the box.
  - This experiment is called "Drawing **without** replacement" as I have NOT put the first ball back in the box before making the second attempt.
  - In this case the probability of getting a , say, red ball in the first instance (first attempt) will be 2/5
  - However, the probability of getting a red ball again in the second attempt will depend on the results/outcome from the first attempt
    - If I had taken a Red ball in the first attempt, then the probability of getting a Red in the second attempt =1/4 (as there will be 1 red and 3 blue balls in the box for the second attempt).
    - If I had taken a Blue ball in the first attempt, then the probability of getting a Red in the second attempt = 2/4 (as there will be 2 red and 2 blue balls in the box for the second attempt).
- **Drawing without replacement** is just another instance of working with **conditional probabilities**.





TULIP Team for Universal Learning and Intelligent Processing

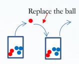
21

Tree Diagrams

- A tree diagram helps us think through conditional probabilities by showing sequences of events as paths that look like branches of a tree.
- The branches can show the probabilities associated with each path
- For the above example, for each scenario, the following tree diagram can be drawn.
  Let R1 and B1 represent the red and blue ball that is selected at the first attempt and R2 and B2 represent the red and blue ball that is selected in the second attempt.
- Scenario-1: Drawing with replacement

P(R1) =2/5 → R1
- P(R2/R1)= 2/5 → R2 P(R1 and R2) = P(R1, R2) = 2/5*2/5 = 4/25
- P(B2/R1)= 3/5 → B2 P(R1, B2) = 2/5*3/5 = 6/25

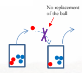
P(B1) =3/5 → B1
- P(R2/B1) = 2/5 → R2 P(B1, R2) = 3/5*2/5 = 6/25
- P(B2/B1) = 3/5 → B2 P(B1, B2) = 3/5*3/5 =9/25



- Scenario-2: Drawing without replacement

P(R1) =2/5 → R1
- P(R2/R1)= 1/4 → R2 P(R1, R2) = 2/5*1/4 = 1/10
- P(B2/R1)= 3/4 → B2 P(R1, B2) = 2/5*3/4 = 3/10

P(B1) =3/5 → B1
- P(R2/B1) = 2/4 → R2 P(B1, R2) = 3/5*2/4 = 3/10
- P(B2/B1) = 2/4 → B2 P(B1, B2) = 3/5*2/4 =3/10

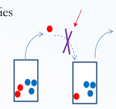


TULIP Team for Universal Learning and Intelligent Processing

22

Tree diagram…

- Using the tree diagram, the following probabilities can be calculated.
- Use the scenario-2, in the above example, to calculate the following probabilities
  - (i) getting two red balls
  - (ii) getting one red ball
  - (iii) getting at least one red balls



Answer:

(i) P(two red balls) = P(R1 and R2) = 2/5*1/4 = 1/10

(ii) P(one red) = P(R1 and B2) + P(B1 and R2) =3/10 + 3/10 = 6/10 = 3/5

(iii) P(at least one red) = 1 - P(no red) = 1 - P(B1 and B2) = 1 - 3/10 = 7/10

TULIP Team for Universal Learning and Intelligent Processing

23

Marginal distribution

- A contingency table is a table that displays two categorical variables and their relationships.
- There were 528 third-class ticket holders who died.
- The margins of the table, both on the right and at the bottom, give totals
- The right column of the table is the frequency distribution of the of the variable ***Survival***
- The bottom row of the table is just the frequency distribution of the variable *ticket* ***Class***
- When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its marginal distribution

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

Survival	Class: First	Second	Third	Crew	Total	Marginal distribution of Survival
Alive	203	118	178	212	711	711/2201=0.3230
Dead	122	167	528	673	1490	1490/2201=0.6770
Total	325	285	706	885	2201	

P(*Alive*) P(*Dead*)

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201
Marginal distribution of Class	325/2201=0.1477	285/2201=0.1295	706/2201=0.3208	885/2201=0.4021	

P(*First*) P(Third)

TULIP Team for Universal Learning and Intelligent Processing

24

Conditional Distributions

- A conditional distribution provides the percent of one variable satisfying the conditions of another.
- The "Condition" can either be based on rows or columns.
- In this example, the "Condition" is based on columns, i.e., Each column represents the conditional distribution **of** ***Survival*** for a given category of ticket *Class*.
  - E.g., 25.2% of all third-class ticket holders survived.
- In this example, the condition is based on rows, i.e., the conditional distribution **of ticket *Class*** conditioned on each value of *Survival*: Alive and Dead.
  - E.g., This table shows that the highest percent of survivors were crew members. The highest percent of the dead were also crew members.

Survival	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

$P(\text{Survial} = Alive \mid \text{Class} = First) = P(Alive \mid First) = \frac{203}{325} \times 100 = 62.5\%$

Survival		First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	711
	% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
Dead	Count	122	167	528	673	1490
	% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
Total	Count	325	285	706	885	2201
		100%	100%	100%	100%	100%

Survival	First	Second	Third	Crew	Total
Alive	203	118	178	212	711
	28.6%	16.6%	25.0%	29.8%	100%
Dead	122	167	528	673	1490
	8.2%	11.2%	35.4%	45.2%	100%

$P(First \mid Dead) = \frac{122}{1490} \times 100 = 8.2\%$

25

Marginal and Joint Probabilities

- Consider a Facebook-Twitter example:
  71% use Facebook, 18% Twitter, 15% both

  Draw a partial table.
  - 0.71 and 0.18 are called marginal probabilities.
  - 0.15 is a joint probability .

Use Twitter	Use Facebook: Yes	No	Total
Yes	0.15		0.18
No			
Total	0.71		1.00

- How can we complete the table?
  - The sum must add up
    - 0.15 + ? = 0.71
    - 0.18 + ? = 1.00
- Are Facebook and Twitter mutually exclusive?
  A = {uses Facebook}, **B** = {uses Twitter}
  $P(\mathbf{A}) = 0.71$, $P(\mathbf{B}) = 0.18$

$P(\mathbf{A} \text{ and } \mathbf{B}) = 0.15 \neq 0$. Facebook and Twitter are **not mutually exclusive**
- Are uses Facebook and uses Twitter independent?
  $P(\mathbf{B} \mid \mathbf{A}) = 0.15/0.71 \approx 0.21$, $P(\mathbf{B}) = 0.18$
  $P(\mathbf{B} \mid \mathbf{A}) \neq P(\mathbf{B})$. Therefore, **not independent**.

Since the respondents who use Facebook are more likely to tweet, they are not independent. Note: Alternatively, you can check if $P(\mathbf{A,B}) = P(\mathbf{A}) \times P(\mathbf{B})$ for independence check.

Use Twitter	Use Facebook: Yes	No	Total
Yes	0.15	0.03	0.18
No	0.56	0.26	0.82
Total	0.71	0.29	1.00

26

Conditional Independence

- Random variables are *conditionally independent* when their conditional distributions are unrelated.
- Consider three random variables A, B and C. If A and B are conditionally independent given C (denoted as $A \perp B|C$ or $A \perp\!\!\!\perp B|C$)

$$\boxed{\begin{array}{c} P(A \mid B,C) = P(A|C) \\ P(B \mid A,C) = P(B|C) \\ P(A,B| C) = P(A|C)P(B|C) \end{array}}$$

- Proof: $P(A,B| C) = P(A|C)P(B|C)$

$$\begin{aligned} P(A,B \mid C) &= \frac{P(A,B,C)}{P(C)} \\ &= \frac{P(A,B,C)}{P(B,C)} \times \frac{P(B,C)}{P(C)} \\ &= P(A \mid B,C) \times P(B \mid C) \\ &= P(A \mid C) \times P(B \mid C) \quad [\text{Note: } P(A \mid B,C) = P(A \mid C)] \end{aligned}$$

- Example: Consider a relationship between a person's arm length and his reading skills
  - One might observe that people with longer arms tend to have higher levels of reading skills.
  - This can be explained by presence of a confounding factor, age. A young child tend to have shorter arm and lacks reading skills of an adult.
  - If the age of the person is fixed/known, then observed relationship between arm length and reading skill disappears.
  - Hence arm length and reading skills are conditionally independent when the age variable is fixed (known).

27

Bayes' Rule

- Bayes' rule
- Prior, Likelihood, and Posterior probabilities, Maximum likelihood

28

Bayes' Rule

- Consider the Conditional probability formulae

$$P(A/B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$
$$\therefore P(A \cap B) = P(A/B)P(B)$$

Similarly, $P(B \cap A)$ can be written in the conditional form as follows.
$$P(B \cap A) = P(B/A)P(A)$$

Note that $P(B \cap A) = P(A \cap B)$
$$\therefore P(B/A)P(A) = P(A/B)P(B)$$
$$\therefore P(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

This is called the Bayes' rule.

$$\boxed{P(B/A) = \frac{P(A/B)P(B)}{P(A)}}$$

- Bayes' rule provides a formulae to *reverse* the conditioning from $P(A/B)$ to $P(B/A)$

29

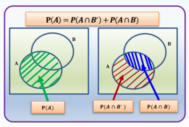
Bayes' Rule…

- Note that P(A) can be written as follows, where B' is the complement of B.

$$P(A) = P(A \cap B) + P(A \cap B')$$
$$\therefore P(A) = P(A/B)P(B) + P(A/B')P(B')$$

- The Bayes' rule can be written as follows

$$\boxed{\begin{array}{c} P(B/A) = \dfrac{P(A/B)P(B)}{P(A)} \\ P(B/A) = \dfrac{P(A/B)P(B)}{P(A/B)P(B) + P(A/B')P(B')} \end{array}}$$



- The above formulae is for two events A and B. This can be extended to more than two events, which we will see later.

30

Example-1

According to a study, 44% of college students engage in binge drinking, 37% drink moderately, and 19% abstain entirely. Another study finds that among binge drinkers, 17% have been involved in an alcohol-related automobile accident, while among moderate drinkers, only 9% have been involved in such accidents. No accidents were observed for those abstained from drinking.

a) What is the probability that a randomly selected college student will be a binge drinker and has had an alcohol related car accident?

b) If we know a student has had an alcohol-related accident, what is the probability that the student is a binge drinker?

a) $P(Binge\ and\ Accident) = P(Binge) \times P(Accident\ |Binge)$
$= 0.44 \times 0.17 = 0.075$

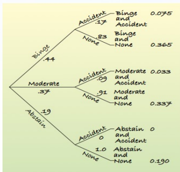
b) $P(Binge|\ Accident\ ) = \frac{P(Binge\ and\ Accident)}{P(Accident)}$

$P(Accident) = P(Binge\ and\ Accident) + P(Moderate\ and\ Accident) + P(Abstain\ and\ Accident)$

$\therefore P(Binge|\ Accident\ ) = \frac{0.44 \times 0.17}{(0.44 \times 0.17) + (0.37 \times 0.09) + (0.19 \times 0)}$

$= \frac{0.075}{(0.075)+(0.033)+(0)}$

$= \frac{0.075}{0.108} = 0.694$



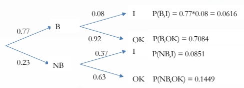
TULIP Team for Universal Learning and Intelligent Processing

31

Example-2

A recent Maryland highway safety study found that in 77% of all accidents the driver was wearing seatbelt. Accident reports indicated that 92% of those drivers escaped serious injury (defined as hospitalised or death), but only 63% of the nonbelted drivers were so fortunate. What is the probability that a driver who was seriously injured wasn't wearing a seatbelt?

- Let **B** = the driver was wearing a seatbelt, and **NB** = no belt.
- Let **I** = serious injury or death, and **OK**=not seriously injured
- P(B) = 0.77, so P(NB) = 1-0.77 = 0.23
- P(OK|B) = 0.92, so P(I|B) = 1-0.92 = 0.08
- P(OK|NB) = 0.63, so P(I|NB) = 1-0.63 = 0.37



Alternatively, using Bayes rule,

$P(NB/\ I) = \frac{P(NB, I)}{P(I)}$

$= \frac{P(I|NB) \times P(NB)}{P(I)}$

$= \frac{P(I|NB) \times P(NB)}{P(I|NB) \times P(NB) + P(I|B) \times P(B)}$

$= \frac{0.37 \times 0.23}{0.37 \times 0.23 + 0.08 \times 0.77}$

$= \frac{0.0851}{0.0851 + 0.0616} = 0.58$

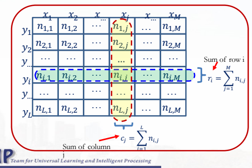
Using the tree diagram: $P(NB/\ I) = \frac{P(NB, I)}{P(I)} = \frac{0.0851}{0.0616+0.0851} = 0.58$

TULIP Team for Universal Learning and Intelligent Processing

32

Generalising

- Consider two random variables, X, which takes the values $\{x_j\}$, where j $= 1,2,3, \dots, M$, and Y which takes the values $\{y_i\}$, where i $= 1,2,3, \dots, L$.
- If we consider a total number of N of instances of these variables, then we denote the number of instances where $X = x_j$ and $Y = y_i$ by n_{ij}, which is the number of points in the corresponding cell of the array.
- The number of points in column j, corresponding to $X = x_j$, is denoted by c_j, and the number of points in row i, corresponding to $Y = y_i$, is denoted by r_i.



$Total\ count = N = \sum_{i=1}^{L} \sum_{j=1}^{M} n_{i,j}$

Example: Titanic data

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	711
Dead	122	167	528	673	1490
Total	325	285	706	885	2201

TULIP Team for Universal Learning and Intelligent Processing

33

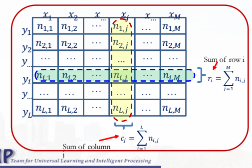
Generalising

- Marginal probability $p(Y = y_i) = \frac{r_i}{N}$
- Joint probability $p(X = x_j, Y = y_i) = \frac{n_{i,j}}{N}$

Conditional Probability $p(X = x_j \mid Y = y_i) = \frac{n_{i,j}}{r_i}$

Product Rule

$p(X = x_j, Y = y_i) = \frac{n_{i.j}}{N} = \frac{n_{i.j}}{r_i} \frac{r_i}{N}$
$= p(X = x_j \mid Y = y_i)$
$p(Y = y_j)$



- Sum Rule

$p(Y = y_i) = \frac{r_i}{N} = \frac{1}{N} r_i = \frac{1}{N} \sum_{j=1}^{M} n_{i,j}$
$= \sum_{j=1}^{M} p(X = x_j, Y = y_i)$

TULIP Team for Universal Learning and Intelligent Processing

34

Sum and Product Rule

- Sum Rule $p(X) = \sum_Y p(X, Y)$
- Sum Rule $p(Y) = \sum_X p(X, Y)$
- Product Rule $p(X, Y) = p(Y|X)p(X)$

Summing over all possible states of Y. This is called marginalisation over Y

- The product rule can be applied multiple times to yield the chain rule of probability

$$\boxed{p(X_1, X_2, X_3, \dots, X_D) = p(X_1)\, p(X_2/X_1)\, p(X_3/X_2X_1)\, p(X_4/X_1X_2X_3) \dots p(X_D/\ X_{1:D-1})}$$

  - Where $1{:}D$ denote the set $\{1,2,3, \dots, D\}$

TULIP Team for Universal Learning and Intelligent Processing

35

Chain rule of probability

Derivation of Chain rule of probability

$p(X_1, X_2, X_3) = p(X_3, X_2, X_1)$

$= p(X_3|X_2, X_1) \quad p(X_2, X_1)$ *[Product rule]*

$= p(X_3|X_2, X_1)\, p(X_2|X_1)\, p(X_1)$ *[Product rule]*

$= p(X_1) p(X_2|X_1)\, p(X_3|X_2, X_1)$

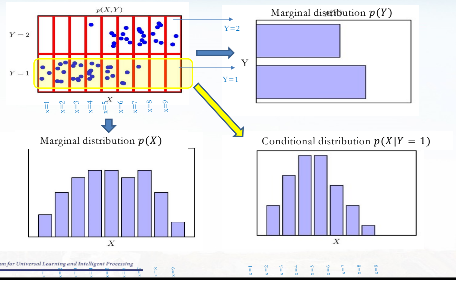
In general,

$p(X_1, X_2, X_3, \dots, X_D) = p(X_1)\, p(X_2/X_1)\, p(X_3/X_2X_1)\, p(X_4/X_1X_2X_3) \dots p(X_D/\ X_{1:D-1})$

TULIP Team for Universal Learning and Intelligent Processing

36

Example: Joint, marginal and conditional distribution



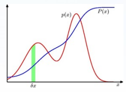
TULIP Team for Universal Learning and Intelligent Processing

37

Probabilities - Continuous variables

- So far we have considered probabilities defined over discrete sets of events. Similarly, we also can consider probabilities with respect to continuous variables.
- Probability that x will lie in the interval (a,b) is:

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$



- Probability density $p(x)$ should satisfy the following conditions

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

- Probability that x will lie in the interval $(-\infty, z)$is given by the cumulative distribution function defined by: $P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$

TULIP Team for Universal Learning and Intelligent Processing

38

Sum and Product Rule – continuous variables

- In the continuous variable case, the sum and product rule will become as follows
  - Sum Rule $p(x) = \int p(x,y)dy$
    $p(y) = \int p(x,y)dx$
  - Product Rule $p(x,y) = p(y|x)p(x)$

TULIP Team for Universal Learning and Intelligent Processing

39

Probability Formula

Important Probability formula

- Conditional probability formulae:
- $P(A/B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A\cap B)}{P(B)} = \frac{P(A,B)}{P(B)}$, $\therefore P(A,B) = P(A/B)P(B) = P(B/A)P(A)$
- Bayes' rule:
- $P(B/A) = \frac{P(A/B)P(B)}{P(A)}$, $P(B/A) = \frac{P(A/B)P(B)}{P(A/B)P(B) + P(A/B')P(B')}$
- SUM Rule:
- $P(A) = \sum_B P(A,B)$, $P(A) = \sum_{B,C,D} P(A,B,C,D)$
- In here the summation is over all values of B, C and D. $P(A,B) = \sum_C P(A,B,C)$
- PRODUCT Rule:
- $P(A,B) = P(A/B)P(B)$
- Chain Rule:
- $P(A,B,C) = P(A|B,C)\,P(B,C) = P(A|B,C)P(B|C)P(C)$
- $P(A,B,C,D,E) = P(A)\,P(B/A)\,P(C/A,B)\,P(D/A,B,C)\,P(E/A,B,C,D)$
- $P(A,B,C,D,E) = P(E)\,P(D/E)\,P(C/D,E)\,P(B/C,D,E)\,P(A/B,C,D,E)$

Important Probability formula

- INDEPENDENCE
  - If A and B are independent (denoted as $A \perp B$)

$$P(A/B) = P(A),\ \ P(B/A) = P(B), P(A,B) = P(A)P(B)$$

- Conditionally Independent
  - Random variables are *conditionally independent* when their conditional distributions are unrelated. Consider three random variables A, B and C.
  - If A and B are conditionally independent given C (denoted as $A \perp B|C$)

$$P(A,B|C) = P(A|C)P(B|C),\ \ P(A/B,C) = P(A/C),$$
$$P(B/A,C) = P(B/C)$$

- Note: In the case of continuous random variables, probability masses P() are replaced by the density function f() and the summation is replaced by the integration.

TULIP Team for Universal Learning and Intelligent Processing

40

Bayes Rule: Prior, Likelihood, Posterior

- Recall that Bayes rule provides a mechanism to ***reverse*** the conditioning.
  - If we know **P(A/B)** then we can find **P(B/A)**
- In many situations, we would like to find a model, e.g., a probability distribution function, that results in a given data that we have observed/collected.
  - If we know the family of distribution (e.g, exponential distribution $p(x) = \theta e^{-\theta x}$) from which the data is generated, but we do not know a *parameter of the distribution* (e.g., the θ in the exponential distribution), then we would like to find/estimate that parameter value (θ) using the data we observed.
  - There are two formal methods available to find/estimate this parameter from data
    1) Frequentist method
       - We use Maximum likelihood estimate (MLE) to find the parameter
    2) Bayesian method
       - We use Bayes rule to find the parameter
- Example:
  - Bacteria count estimation in a beach:
    - You want to determine *the mean bacteria count per liter of water* in a beach in Australia . Assuming that the distribution of number of bacteria per liter of water is *normal (Gaussian)* with an unknown mean μ (called as parameter) and known variance, you want an estimate of μ (*the mean bacteria count per liter of water*). You collected several water samples from the beach and use as the data. How can you estimate the parameter μ?
      - You may use some knowledge about the bacteria count from other beaches in the world as *prior knowledge* to assist the estimation.
      - Here, **Bayesian method** can be used to find/estimate *the mean bacteria count per liter of water.*

TULIP Team for Universal Learning and Intelligent Processing

41

Bayes Rule: Prior, Likelihood, Posterior

- Consider a problem of the form:
  - tell me something about the variable or parameter θ given that I have observed data $\mathcal{D}$ and have some knowledge of the underlying data generating mechanism (i.e., the effect of the observed data $\mathcal{D}$ through the conditional probability $p(\mathcal{D}|\theta)$).
  - Hence, we are interested In finding the quantity $p(\theta/\mathcal{D})$
    - $p(\theta/\mathcal{D})$ is a probability density distribution (pdf). Using this we can obtain an estimate of θ, for example, using the mean or mode from the pdf.
- *Using Bayes' theorem we can form the equation*

Posterior likelihood Prior

$$p(\theta/\mathcal{D}) = \frac{\overbrace{p(\mathcal{D}/\theta)}^{\text{likelihood}}p(\theta)}{p(\mathcal{D})}$$

Normalising constant

- $p(\theta)$ is called the *prior*, $p(\mathcal{D}/\theta)$ is called the *likelihood*
- $p(\theta/\mathcal{D})$ is called the *posterior*
- *Hence, Bayes' theorem can be stated in words as* posterior $\propto$ likelihood $\times$ prior
- Prior, likelihood and posterior quantities are viewed as functions of θ
- The denominator $p(D)$ is the normalisation constant, which ensures that the posterior distribution is a valid probability density and integrates to one.

$$p(\mathcal{D}) = \int p(\mathcal{D},\theta)d\theta = \int p(\mathcal{D}/\theta)p(\theta)\,d\theta$$

TULIP Team for Universal Learning and Intelligent Processing

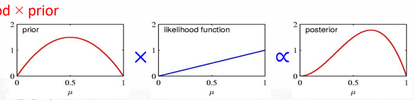
42

Bayes Rule: Prior, Likelihood, Posterior

For example:

- If *prior* $p(\theta)$ has a **pdf** as shown left:
- and *likelihood* $p(\mathcal{D}/\theta)$ has a **pdf** as shown the middle:
- *then the posterior* $p(\theta/\mathcal{D})$ *is obtained by multiplying the prior pdf and the likelihood* ***pdf****. The resulting posterior pdf will be as shown right:*

posterior ∝ likelihood × prior



43

Computing Posterior pdf

posterior ∝ likelihood × prior

$$\underbrace{p(\mu/\mathcal{D})}_{\text{Posterior}} = \frac{\overbrace{p(\mathcal{D}/\mu)}^{\text{likelihood}}\ \overbrace{p(\mu)}^{\text{Prior}}}{\underbrace{p(\mathcal{D})}_{\text{Normalising constant}}}$$

You can use a particular **pdf** as a ***prior*** distribution, collect data of a specific flavor (and find the ***likelihood*** **pdf**) , and then derive the ***posterior*** **pdf**.

What are the ways you can compute the **posterior pdf**?

- There are two ways
  1. Use conjugate distributions
     - In some special cases, the pdf of the prior and posterior are the same (family of) probability density function, but their parameters may differ.
     - The prior distribution with the above property is called a **conjugate prior**, and the effect of the data can then be interpreted in terms of changes in parameter values.
     - Using a conjugate prior gives closed form expressions for posterior, which is computationally efficient
  2. Use numerical integration to combine prior pdf and likelihood pdf to compute posterior pdf, such as Markov Chain Monte Carlo (MCMC) Methods.

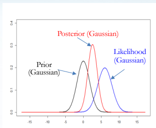
44

Conjugate distributions

- For example, consider that
  - the prior $p(\mu)$ is a **Gaussian** distribution
  - the likelihood $p(\mathcal{D}/\mu)$is a also a **Gaussian** distribution
  - Then, the posterior is also a **Gaussian** distribution:

  $p(\mu/\mathcal{D}) \propto p(\mathcal{D}/\mu) \times p(\mu)$

  $= Gaussian \times Gaussian = Gaussian\ distributed$
  - Note that if two Gaussian distributions are multiplied, it results in a Gaussian distribution again!
- In this case we can notice that the prior $p(\mu)$ is a Gaussian, and the resulting posterior $p(\mu|\mathcal{D})$ is also a Gaussian



- In Bayesian statistics, for *certain* ***likelihood*** *functions* if a certain **prior** is chosen, it results in the **posterior** having the **same distribution** ***as the prior***.
- Conjugate distributions (or conjugate pairs):
  - If the posterior distribution $p(\mu|\mathcal{D})$ is in the same family (in this case Gaussian) as the prior probability distribution$p(\mu)$ (in this case Gaussian), the prior and posterior (in this case Gaussian again; same as prior distribution) distributions are then called conjugate distributions
  - And the prior is called a conjugate prior for the likelihood function

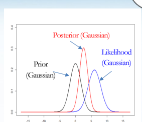
45

Conjugate pairs

- Recall the example with Gaussian,
  - $p(\mu/\mathcal{D}) \propto p(\mathcal{D}/\mu) \times p(\mu)$
  - consider that the prior $p(\mu)$ is a **Gaussian** distribution, the likelihood $p(\mathcal{D}/\mu)$is also a **Gaussian** distribution, then the posterior $p(\mu|\mathcal{D})$ will become a **Gaussian distribution**
  - Note that if two Gaussian distributions are multiplied, it results in another Gaussian distribution again!

  $p(\mu/\mathcal{D}) \propto p(\mathcal{D}/\mu) \times p(\mu)$

  $p(\mu/\mathcal{D}) = Gaussian \times Gaussian = Gaussian\ distributed$



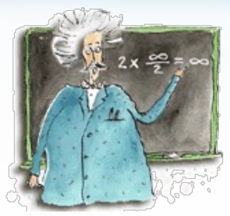
- We can easily find the posterior distribution parameters by simply combing (as a function of) prior and likelihood distribution parameters: If Prior $\sim N(\mu_0, \sigma_0^2)$ and *Likelihood* $\sim N(\mu, \sigma^2)$, then *Posterior* $\sim N(\mu_N, \sigma_N^2)$, where μ_N and σ_N can be found as follows:

$$\boxed{\begin{aligned} \mu_N &= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}. \end{aligned}}$$

- Example conjugate-pairs based models
  - Gaussian –Gaussian model (Gaussian prior, Gaussian likelihood => Gaussian posterior)
  - Dirichlet-Multinomial model (Dirichlet prior, Multinomial likelihood => Dirichlet posterior)
  - See [https://en.wikipedia.org/wiki/Conjugate_prior] for more **conjugate pairs**
- Note that, in situations where we cannot use conjugate pairs, i.e., using non-conjugate priors, then numerical methods are used to compute the posterior.

46

Common Distributions



- Discrete Distributions
  - Binomial
  - Multinomial
- Continuous Distributions
  - Gaussian
  - Gamma
  - Beta
  - Dirichlet

47

Common distribution for Bayesian learning

- *Before we learn* further about the *possible conjugate prior pdfs and the* Frequentists and Bayesian estimation methods (next week), *first we will learn about some of the useful probability distributions* (both discrete distributions and continuous distributions).
- In particular, we will look at
  - What are the shapes of the distributions
  - What parameters are used for each distribution
  - How the distribution function looks like
  - How to find (**look up for**) the mean and standard deviation/variance of those distributions

48

Common Distributions

What is probability distribution

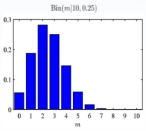
- A function that gives the probabilities of occurrence (i.e., how frequently occurs) for different possible outcome (values) of one or more variables from experiment.
- A distribution describes the complete pattern of behaviour of a variable,
  - This will help to find some answers, for example, 'what values are most likely to occur', and 'what is the average value of the variable'
- If you know the distribution, you will be able to compute some quantities, such as the centre of the distribution (e.g., mean, median) or the variance of the distribution (standard deviation, variance)

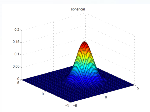
Discrete Distributions

- The variable takes discrete values (e.g., head or tail)
- Useful discrete distributions
  - Bernoulli Distribution
  - Binomial Distribution
  - Multinoulli Distribution
  - Multinomial Distribution

Continuous Distributions

- The variable takes continuous values (e.g., length)
- Useful continuous distributions
  - Gaussian (Normal) distribution and their properties
  - Gamma distribution
  - Beta distribution
  - Dirichlet Distribution





TULIP Team for Universal Learning and Intelligent Processing

49

Discrete Distributions

- The below discrete distributions can be observed with following experiments.

Experiment	Resulting distribution
Tossing a (2 sided) coin once	Bernoulli Distribution
Tossing a (2 sided) coin N times	Binomial Distribution
Tossing a K sided dice once	Multinoulli or Categorical Distribution
Tossing a K sided dice N times	Multinomial Distribution

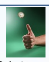
TULIP Team for Universal Learning and Intelligent Processing

50

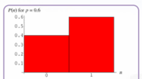
Discrete distributions: Bernoulli

- Consider tossing a Coin only *once*.
- The outcomes are either a head=1 or a tail=0.
  Let $X \in \{0,1\}$ be a (binary) random variable representing the number of heads.
- Probability of getting a "success" or "head" (1) is μ. That is, $p(x=1|\mu)=\mu$
- Then X has a distribution called Bernoulli Distribution $\mathrm{Bern}(x|\mu)=\begin{cases}\mu & \text{if } x=1\\ 1-\mu & \text{if } x=0\end{cases}$
- The above equation can be combined and written as a single function as follows

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x}\\ \mathbb{E}[x] &= \mu \quad \longleftarrow \text{Mean}\\ \mathrm{var}[x] &= \mu(1-\mu) \longleftarrow \text{variance}\end{aligned}$$



Coin tossed only once



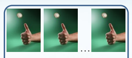
Bernoulli distribution with probability of success p=0.6 and n={0,1}.
https://mathworld.wolfram.com/BernoulliDistribution.html

TULIP Team for Universal Learning and Intelligent Processing

51

Binomial distribution: Binomial probabilities - an example

- Suppose a coin is tossed 3 times
  - TTT, HTT, THT, TTH, HHT, HTH, THH, HHH
  - each with probability $\frac{1}{8}$
- Let X represent be the number of heads thrown. Possible values for X here are :
  - 0 (no head), 1 (one head), 2 (two heads) and 3 (three heads)
- The binomial distribution can be used to obtain these probabilities more easily,
  - e.g., $P(X=2)=\frac{3}{8}$



Coin is tossed multiple (N) times

X	Outcomes	P(X=x)
0	TTT	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$
1	TTH, THT, HTT	$\left(\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\right)\times 3=\frac{3}{8}$
2	THH, HTH, HHT	$\left(\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\right)\times 3=\frac{3}{8}$
3	HHH	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$

TULIP Team for Universal Learning and Intelligent Processing

52

Binomial experiment

The above coin toss (multiple times) example is called a binomial experiment

A binomial experiment satisfies 3 criteria
(i) there are a fixed number of identical trials, n
E.g., throwing a coin 10 times
(ii) for each trial, there are exactly 2 possible outcomes, each with a fixed probability μ and q=1- μ
i.e., the probability of success is μ, and
the probability of failure is q=1-μ.
(iii) the trials are independent

Other examples of binomial experiment:
Which of the following satisfy the binomial model?

(i) on average, a tennis player gets the 1st serve in 60% of the time. What is the probability that 4 of the next 5 serves are in?
Yes. This is a Binomial, with n=5, μ =0.6, q=1-0.6 = 0.4.

(ii) A drawer contains 5 red socks and 4 blue socks. What is the probability that they are both red?
No. This is Not a binomial, as trials are not independent.
i.e., probabilities are not fixed.
e.g., with success defined as 'selecting a red sock', on the first trial μ = 5/8. on the second trial, μ =4/7 or 5/7.

TULIP Team for Universal Learning and Intelligent Processing

53

Binomial Probabilities

A function/formulae to calculate Binomial Probabilities

- For a binomial experiment with n trials, and probability μ of success at each trial, the probability of exactly r successes is given by

$$P(X=r)=\binom{n}{r}\mu^r(1-\mu)^{n-r}$$

Combinations: $\binom{n}{r}$ or ${}^nC_r$
$\binom{n}{r}$ is the number of ways of choosing r objects from n objects.

$$\binom{n}{r}={}^nC_r=\frac{n!}{r!\,(n-r)!}$$

Where $n! = n\times(n-1)\times(n-2)\times\cdots\times3\times2\times1$. This is called the *n factorial.*

e.g., if a coin is tossed 3 times, the probability of getting 2 heads is

$$P(X=2)=\binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1=3\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)=\frac{3}{8}$$

TULIP Team for Universal Learning and Intelligent Processing

54

Previous example:

For the previous “coin toss (multiple times)” example

- a coin is tossed 3 times, and X represent the number of heads thrown, and μ is the probability of getting a head (here, $\mu = \frac{1}{2}$)
- then the probabilities can be found as follows using the binomial probability formulae:

$$P(X = r) = \binom{n}{r}\mu^r(1-\mu)^{n-r}$$

- (as shown in the last column below)

X	Outcomes	P(X=x)	P(X=x) Using Binomial model formulae with n=3, μ=0.5
0	TTT	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$	${}^3C_0\,(0.5)^0(0.5)^3 = \frac{1}{8}$
1	TTH, THT, HTT	$\left(\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\right)\times 3 = \frac{3}{8}$	${}^3C_1\,(0.5)^1(0.5)^2 = \frac{3}{8}$
2	THH, HTH, HHT	$\left(\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\right)\times 3 = \frac{3}{8}$	${}^3C_2\,(0.5)^2(0.5)^1 = \frac{3}{8}$
3	HHH	$\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}$	${}^3C_3\,(0.5)^3(0.5)^0 = \frac{1}{8}$

TULIP Team for Universal Learning and Intelligent Processing

55

Binomial distribution

Binomial distribution:

- Consider a coin is tossed N times.
- Let $m \in \{0,1,2,\ldots,N\}$ be the random variable representing the number of heads obtained.
- The probability of heads is μ, then m has a Binomial Distribution

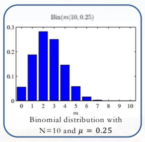

Coin is tossed N times

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu \quad \text{Mean}$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N}(m-\mathbb{E}[m])^2\,\mathrm{Bin}(m|N,\mu) = N\mu(1-\mu) \quad \text{Varianc e}$$

$$\binom{N}{m} = {}^N C_m = \frac{N!}{m!\,(N-m)!}$$



Binomial distribution with N=10 and $\mu = 0.25$

TULIP Team for Universal Learning and Intelligent Processing

56

An Example

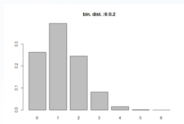
Example: Assuming a binomial distribution, find the following probabilities.

On average, 20% of all doctors' patients have flue symptoms. If 6 people consult a doctor, find the probability that

(i) 1 has flue symptoms
(ii) 4 have flue symptoms
(iii) at least 5 have flue symptoms
(iv) Only first patient and the third patient have flue symptoms
(v) The third patient is the first one with the flue symptom.
(vi) Plot the distribution

Answers:

(i) $P(X=1) = \binom{6}{1}(0.2)^1(0.8)^5 = 6\times0.2\times0.32768 = 0.393216$
(ii) $P(X=4) = \binom{6}{4}(0.2)^4(0.8)^2 = 15\times0.0016\times0.64 = 0.01536$
(iii) $P(X\geq5) = P(X=5)+P(X=6)$
$= \binom{6}{5}(0.2)^5(0.8)^1 + \binom{6}{6}(0.2)^6(0.8)^0 = 0.001536 + 0.000064 = 0.0016$
(iv) $P(\text{1st and the 3rd only}) = 0.2\times0.8\times0.2\times0.8\times0.8\times0.8 = 0.01638$
(v) P(3rd one is the first to have the flue symptom) $= 0.8\times0.8\times0.2 = 0.128$
(vi) The probabilities are:
$P(X=0) = 0.262144$
$P(X=1) = 0.393216$,
$P(X=2) = 0.245760$
$P(X=3) = 0.081920$
$P(X=4) = 0.015360$
$P(X=5) = 0.001536$
$P(X=6) = 0.000064$



TULIP Team for Universal Learning and Intelligent Processing

57

Multinoulli (Categorical) Distribution

- Consider rolling a K sided die once. Let $x = (x_1, x_2, \ldots, x_K)$ be the random vector, where $x_k \in \{0,1\}$ represent the occurrence of side k of the die. Since the die is rolled only once, x will be a vector of 0s and 1s (a bit vector), in which one of the element x_k equals 1, and all remaining elements equal 0 (This is sometimes called as 1-of-K coding).
- For example, for a variable that can take K=6 states (say, a six-sided die) and a particular observation of the variable happens to corresponds to the state $x = (0,0,1,0,0,0)$ where $x_3 = 1$.


Rolling a K sided die once

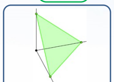
If the probability of $x_k = 1$ is μ_k (that is the probability that side k shows up), then the distribution of x is a Multinoulli distribution (or categorical distribution)

$$p(X|\mu) = \begin{cases}\mu_1 \text{ if } x_1 = 1\\ \mu_2 \text{ if } x_2 = 1\\ \vdots \\ \mu_K \text{ if } x_K = 1\end{cases} \rightarrow p(X|\mu) = \mu_1^{x_1}\mu_2^{x_2}\ldots\mu_K^{x_K} \rightarrow p(X|\mu) = \prod_{k=1}^{K}\mu_k^{x_k}$$

$$\boxed{p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K}\mu_k^{x_k} \qquad \forall k: \mu_k \geqslant 0 \text{ and } \sum_{k=1}^{K}\mu_k = 1}$$

The Multinoulli distribution can be regarded as the *generalisation* of the “Bernoulli distribution” to more than two outcomes.

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1,\ldots,\mu_K)^{\mathrm{T}} = \boldsymbol{\mu} \quad \text{Mean}$$


The possible probabilities for the categorical distribution with K = 3 are the 2-simplex $\mu_1 + \mu_2 + \mu_3 = 1$ embedded in 3-space.
https://en.wikipedia.org/wiki/Categorical_distribution

TULIP Team for Universal Learning and Intelligent Processing

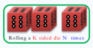
58

The Multinomial Distribution

- Consider rolling a K sided die N times
- Let $m = (m_1, m_2, \ldots, m_K)$ be the random vector, where m_k represent the number of times side k of the die occurs.
- The distribution of m is a Multinomial distribution

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k}$$


Rolling a K sided die N times

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K}\prod_{k=1}^{K}\mu_k^{m_k}$$

$$\mathbb{E}[m_k] = N\mu_k \quad \text{Mean}$$

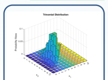
$$\mathrm{var}[m_k] = N\mu_k(1-\mu_k) \quad \text{Variance}$$

$$\mathrm{cov}[m_j m_k] = -N\mu_j\mu_k \quad \text{Covariance}$$

μ_k is the probability that side k shows up (the probability of the kth event occurs in any given trial), and $\binom{N}{m_1, m_2, \ldots, m_k} = \frac{N!}{m_1!m_2!\ldots m_k!}$

This is the multinomial coefficient - the number of ways to divide a set of size $N = \sum_{k=1}^{K} m_k$ into subsets with sizes m_1 up to m_K.

Note that when N=1, $Multi(m_1, m_2, \ldots, mk|\mu, 1) =$ Multinoulli distribution



Multinomial distribution with a sample size of N = 10. The probabilities are $\mu_1 = 1/2$ for outcome 1, $\mu_2 = 1/3$ for outcome 2, and $\mu_3 = \frac{1}{6}$ for outcome 3. Note that the visualization shows m_1 and m_2 and does not show m_3, which is determined by the constraint $m_1 + m_2 + m_3 = N$.
https://www.mathworks.com/help/stats/mnpdf.html

TULIP Team for Universal Learning and Intelligent Processing

59

Summary of Discrete distributions so far

Experiment	Distribution name	Distribution	Mean
Tossing a **coin once**	Bernoulli	$Bern(x\|\mu) = \mu^x(1-\mu)^{1-x}$ $x \in \{0,1\}$	μ
Tossing a **coin N times**	Binomial	$\mathrm{Bin}(m\|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$ $m \in \{0,1,\ldots,N\}$	$N\mu$
Tossing a **K sided dice once**	Multinoulli or Categorical	$p(\mathbf{x}\|\boldsymbol{\mu}) = \prod_{k=1}^{K}\mu_k^{x_k}$ $\quad x \in \{0,1\}^K$	μ
Tossing a **K sided dice N times**	Multinomial	$\mathrm{Mult}(m_1, m_2, \ldots, m_K\|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K}\prod_{k=1}^{K}\mu_k^{m_k}$ $m \in \{0,1,\ldots,N\}^k$	$N\mu_k$

TULIP Team for Universal Learning and Intelligent Processing

60

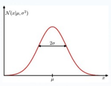
Gaussian Distribution

Univariate Gaussian

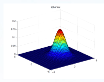
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- *The parameters are:* μ - mean, σ - standard deviation
- We write $X \sim N(\mu, \sigma^2)$ to denote $P(X=x) = N(x|\mu, \sigma^2)$.
- $1/\sigma^2$ is called the Precision
- $X \sim N(0, 1)$ is called the standard normal distribution

Univariate Gaussian



Multivariate Gaussian



Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

D – dimensions.
The parameters are:
μ - mean (vector),
Σ - covariance matrix (Σ^{-1} - inverse of the covariance matrix)

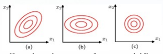


Figure shows the contours of constant probability density for a Gaussian distribution in 2 dimensions, in which covariance matrix is (a) general form (b) diagonal, in which ellipse contours are aligned with the axes, and (c) proportional to the identity matrix, in which the contours are concentric circles

TULIP Team for Universal Learning and Intelligent Processing

61

Gaussian Distribution – Properties: Marginal and Conditional distribution of Gaussians

Marginal and Conditional distributions of a Gaussian are Gaussians

- f(x1 , x2) – Bivariate Gaussian distribution
- f(x1) - Marginal distribution of x 1 (A Gaussian)
- f(x2) - Marginal distribution of x2 (A Gaussian)
- f(x1 | x2=b) - Conditional distribution of x1 given X2=b. (A Gaussian)

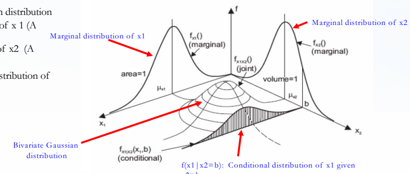


Image fromhttps://www.researchgate.net/publication/260942817_UNCERTAINTY_TREATMENT_IN_CIVIL_ENGINEERING_NUMERICAL_MODELS/figures?lo=1

TULIP Team for Universal Learning and Intelligent Processing

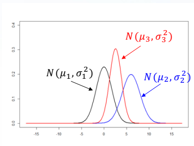
62

Gaussian Distribution – Properties: Multiplication of two Gaussian

- Given two Gaussian distributions $a \sim N(\mu_1, \sigma_1^2)$ and $b \sim N(\mu_2, \sigma_2^2)$, then the distribution of $c = a \times b$ (the multiplication of a and b) is a Gaussian with $c \sim N(\mu_3, \sigma_3^2)$, where

$$\frac{1}{\sigma_3^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \Rightarrow \sigma_3^2 = \left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right]^{-1} = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

$$\mu_3 = \sigma_3^2 \left[\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right]$$



TULIP Team for Universal Learning and Intelligent Processing

63

Gamma Distribution

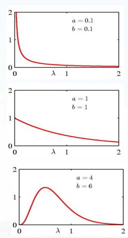
- The Gamma distribution is a flexible distribution for positive real valued random variable (λ).
  - With two hyperparameters, called the shape $a > 0$ and the rate $b > 0$.

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

where, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ is called the gamma function

Mean: $\mathbb{E}[\lambda] = \frac{a}{b}$ Variance: $\mathrm{var}[\lambda] = \frac{a}{b^2}$

- There are several distributions which are special cases of Gamma distribution.
  - Exponential distribution:
    $Expon(x|\lambda) = Gam(x|1, \lambda) = \lambda e^{-\lambda x}$
  - Chi-squared distribution: $\chi^2(x|\nu) = Gam(x|\frac{\nu}{2}, \frac{1}{2})$



TULIP Team for Universal Learning and Intelligent Processing

64

Beta Distribution

- Beta distribution is defined over the interval $\mu \in [0,1]$ as

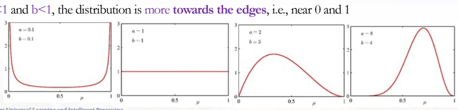
$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

where, $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ is the gamma function

$$\text{Mean:} \quad \mathbb{E}[\mu] = \frac{a}{a+b}$$

$$\text{Variance:} \quad \mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

- Note
  - Different shapes are produced for different values of hyperparameters **a** and **b**.
  - when a=1, b=1, the Beta distribution gives a uniform distribution
  - When a <1 and b<1, the distribution is more **towards the edges**, i.e., near 0 and 1



TULIP Team for Universal Learning and Intelligent Processing

65

The Dirichlet Distribution

- Dirichlet distribution

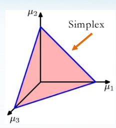
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k \qquad 0 \le \mu_k \le 1, \ \sum_{k=1}^{K} \mu_k = 1$$

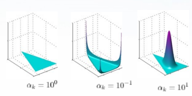
Mean: $E(\mu_k) = \frac{\alpha_k}{\alpha_0}$

Variance: $\mathrm{var}(\mu_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$

$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ is called the gamma function



- Samples from the Dirichlet distribution lie in the (K-1)-dimensional simplex.
- when K=3, it defines a distribution over the simplex, which can be represented by the triangular surface (as shown in the figure).
- Points in this surface satisfy $0 \le \mu_k \le 1$ and $\sum_{k=1}^{3} \mu_k = 1$
- $\alpha_1, \alpha_2, \ldots \alpha_k, \ldots, \alpha_K$ are the hyperparameters. $\alpha_0 = \alpha_1 + \alpha_2 + \cdots + \alpha_K$
- α_0 controls the strength of the distribution (how peaked it is)
- α_k control where the peak occurs
- $\boldsymbol{\alpha}$ = (1,1,1) gives a uniform distribution
- $\boldsymbol{\alpha}$ = (0.1, 0.1, 0.1), that is if α_k <1 for all k, gives spikes at the corner of the distribution
- $\boldsymbol{\alpha}$ = (10,10,10) gives narrow distribution centred at (1/3,1/3,1/3)



TULIP Team for Universal Learning and Intelligent Processing

66

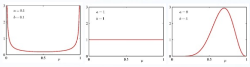
Dirichlet distribution is a *multivariate generalisation of Beta distribution*

- Beta distribution is defined over the interval $\mu \in [0,1]$ as

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \quad \text{where, } \Gamma(x)=\int_0^\infty u^{x-1}e^{-u}du \text{ is the gamma function}$$

$$\text{Mean:} \quad \mathbb{E}[\mu] = \frac{a}{a+b}$$

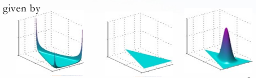
$$\text{Variance:} \quad \mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$



- Dirichlet distribution is a *multivariate generalisation of this Beta distribution*, which has a support over the probability simplex given by

$0 \leq \mu_k \leq 1$ and $\sum_{k=1}^{K} \mu_k = 1$

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \qquad \alpha_0 = \sum_{k=1}^{K}\alpha_k$$



TULIP Team for Universal Learning and Intelligent Processing

67

Data and Data Descriptions

- Different data types
- Displaying data
- Descriptive statistics:
  - Histogram
  - Contingency Table
  - Shape, Skewness
  - Summarizing data

TULIP Team for Universal Learning and Intelligent Processing

68

Data

- Information is all around us
  - Sports data
  - Health/medical data
  - Mobile devices
    - Mobile phones, notepads
    - Wearable devices
  - Stock markets
  - Social networks
  - Internet of Things and Sensor networks
    - Environmental monitoring
    - Infrastructure monitoring
  - Retail/business data

TULIP Team for Universal Learning and Intelligent Processing

69

Data Types

- Data type determines which table, graphs, numerical summaries and analysis techniques are possible or suitable.
  - Data can be numbers, names, or other labels. Data are useless without their context, or unit.
  - Categorical (or Qualitative)
    - The variables have values that fall into distinct categories
    - e.g. gender (male, female)
    - employment status (full-time, part-time, casual, unemployed)
  - Numerical (or Quantitative)
    - The variables have number values, with units, in either of two forms:
      - Discrete – between two sequential values there may be no permissible values, e.g. number of members in a family, shoe size
      - Continuous – an infinite number of values with a defined range are possible, e.g. measurements of weight, time etc.
    - Not all data represented by numbers are numerical data
      - (e.g., 1=male, 2=female; this is a categorical data).

TULIP Team for Universal Learning and Intelligent Processing

70

Penguin example

- This dataset contains the penguins' profile
  - Gender is a categorical variable with two categories 'M' and 'F'
  - Age, height, weight Fat1 and Fat2 are numerical variables.

ID	gender	Age (months)	Height (cm)	Weight (kg)	Fat 1 (mm)	Fat 2 (mm)
342	M	12	67	25.3	15.1	16.2
12	F	23	73	26.4	14.1	13.9
432	M	5	24	0.5	7.1	8.5
14	M	62	79	28.9	12.1	11.5
98	F	28	76	31.0	13.5	14.9
987	F	31	81	30.7	14.6	14.7

TULIP Team for Universal Learning and Intelligent Processing

71

Displaying Qualitative/Categorical Data

- Displaying a single categorical variable
  - Frequency table and relative frequency table
  - Bar chart and relative frequency bar chart
  - Pie chart
- Two or more categorical variable
  - Contingency table
  - Segmented bar chart

TULIP Team for Universal Learning and Intelligent Processing

72

Single categorical variable

The *titanic data*[1] shown below lists the category of passengers (class of passengers) and their counts (number of passengers) who have travelled in the ship. Frequency tables and their corresponding bar charts are shown below

Frequency table

Class	Count
First	325
Second	285
Third	706
Crew	885

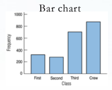
Relative frequency table

Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

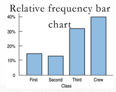
Total count = 2201

Relative frequency for the 'First' class = 325x100/2201 = 14.77%

Bar chart



Relative frequency bar chart



[1] http://en.wikipedia.org/wiki/Passengers_of_the_RMS_Titanic

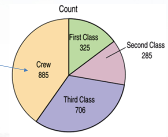
TULIP Team for Universal Learning and Intelligent Processing

73

Pie chart

- When you are interested in parts of the whole, a pie chart might be your display of choice.
- Pie charts show the whole group of cases as a circle.
- They slice the circle into pieces whose size is proportional to the fraction of the whole in each category.

The angle corresponding to the 'crew' category is calculated as 885x360º/2201 = 145º



TULIP Team for Universal Learning and Intelligent Processing

74

Contingency Tables

A contingency table allows us to look at two categorical variables together.

- It shows how individuals are distributed along each variable, contingent on the value of the other variable.
  - Example: we can examine the class of ticket and whether a person survived the *Titanic*:

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

TULIP Team for Universal Learning and Intelligent Processing

75

Displaying Quantitative/Numerical Data

- Summarising the data will help us when we look at large sets of quantitative data.
- Without summaries of the data, it's hard to grasp what the data tell us.
- Displaying numerical data
  - Histogram
  - Stem and leaf plot

TULIP Team for Universal Learning and Intelligent Processing

76

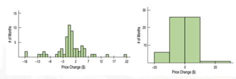
Histogram: construction

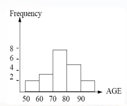
- Consider the data representing ages of people at a nursing home:

82 74 88 66 58 74 78 84 96 76
62 68 72 92 86 76 52 76 82 78

  - First decide on suitable bins/intervals
  - Then count how many ages fall within each bin
  - Form a frequency table before creating the histogram.
    - Choose a bin width appropriate to the data.
    - Changing the bin width changes the appearance of the histogram

Bin	no. people
50-59	2
60-69	3
70-79	8
80-89	5
90-99	2
Total	20



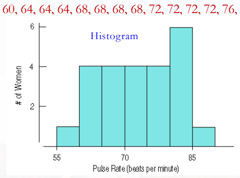


TULIP Team for Universal Learning and Intelligent Processing

77

Stem-and-Leaf Displays

- Stem-and-leaf displays show the distribution (shape) of a quantitative variable, like histograms do, while preserving the individual values.
  - Example data (Pulse rate of women):
    - 56, 60, 64, 64, 64, 68, 68, 68, 68, 72, 72, 72, 72, 76, 76, 76, 76, 80, 80, 80, 80, 84, 84,88



Stem and leaf plot

5	6
6	0444
6	8888
7	2222
7	6666
8	000044
8	8

Pulse Rate
(5|6 means 56 beats/min)

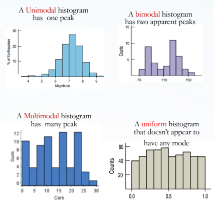
Key: 56 → 5|6

TULIP Team for Universal Learning and Intelligent Processing

78

Describing the shape of a distribution

- Does the histogram have a single, central hump/peak or several separated peaks?
  - Humps in a histogram are called modes.
  - A histogram with one main peak is called unimodal; histograms with two peaks are bimodal; histograms with three or more peaks are called multimodal.
- Is the histogram symmetric?
- Do any unusual features stand out (outliers)?

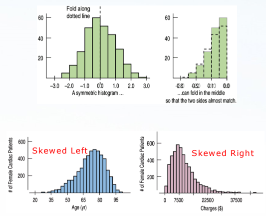


79

Symmetry, Skew

Symmetry
- Is the histogram symmetric?
- If you can fold the histogram along a vertical line through the middle and have the edges match pretty closely, the histogram is symmetric.

Skew
- The (usually) thinner ends of a distribution are called the tails. If one tail stretches out more than the other, the histogram is said to be skewed to the side of the longer tail.
- In the figure below, the histogram on the left is skewed left (negatively skewed), while the histogram on the right is skewed right (positively skewed).



80

Anything Unusual (Outlier)?

- Do any unusual features stick out?
  - Sometimes it's the unusual features that tell us something interesting or exciting about the data.
  - You should always mention any stragglers, or outliers, that stand off away from the body of the distribution.
  - Are there any gaps in the distribution? If so, we might have data from more than one group or may have missed a group in our sampling.
- The following histogram has outliers—there are three cities in the leftmost bar:

81

Summarising Quantitative Data

- Centre of Distributions
- Spread of Distributions

82

Centre of a Distribution

- Mean or Average

$$\bar{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

- E.g.: the average of 12, 14, 6 and 8 is the sum of all the values (40) divided by the number of values (4), which gives 10.

  Mean = (12+14+6+8)/10 = 40/10 = 4

- The median is the value with exactly half the data values below it and half above it.
  - It is the middle data value (once the data have been sorted) when the number of observations is odd
  - It is the average of the middle two data values (once the data have been sorted) when the number of observations is even
  - E.g. for 4 4 **8** 9 11, the median is 8
  - E.g. for 1 2 4 **4** **6** 8 9 11, the median is (4+6)/2 = 7

83

Centre of a Distribution

- In symmetric distributions, the mean and median are approximately the same in value, so either measure of centre may be used.
- For skewed data, it's better to report the median than the mean as a measure of centre, as the mean is 'pulled' in the direction of the skewness.

84

Spread of the distribution

- Measure of spread are
  - Range, Interquartile range (IQR), Standard deviation
  - Always report a measure of spread along with a measure of centre when describing a distribution numerically.
- Range
  - The range of the data is the difference between the maximum and minimum values:

$$range = max – min$$

  - A disadvantage of the range is that a single extreme value can make it very large and so it is not representative of the data overall.

TULIP Team for Universal Learning and Intelligent Processing

85

Interquartile Range (IQR)

- Interquartile Range (IQR)
  - let us ignore extreme data values and concentrate on the middle of the data.
  - To find the IQR, we first need to know what quartiles are…
  - Quartiles divide the data into four equal sections.
    - One quarter of the data lies below the lower quartile (first quartile), Q1
    - One quarter of the data lies above the upper quartile (third quartile), Q3.
  - The difference between the third and the first quartile is the IQR, so IQR = Q3 - Q1
  - The quartiles can be thought of as the median of each half of the data.
  - E.g. find the quartiles for the following data:

3.58	3.80	4.01	4.01	4.05	4.05	4.12	4.18	4.20	4.21
4.27	4.28	4.30	4.32	4.33	4.35	4.35	4.41	4.42	4.45

Q1 = (4.05 + 4.05)/2 = 4.05
Q3= (4.33+4.35)/2 = 4.34
IQR = Q3 - Q1 = 4.34-4.05 = 0.29

- NOTE: median is known as the second Quartile Q2

TULIP Team for Universal Learning and Intelligent Processing

86

Standard Deviation

- A more powerful measure of spread than the IQR is the standard deviation, which takes into account how far *each* data value is from the mean.
- A deviation is the distance that a data value is from the mean.
  - Since adding all deviations together would total zero, we square each deviation and find an average of sorts for the deviations.
- The variance s^2, is found by summing the squared deviations and dividing by n -1:

$$s^2 = \frac{\sum(y-\bar{y})^2}{n-1}$$

- The standard deviation, s, is just the square root of the variance and is measured in the same units as the original data.

$$s = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$$

TULIP Team for Universal Learning and Intelligent Processing

87

Shape, Centre and Spread

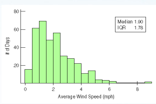
- Always report the *shape* of a distribution, along with a *centre* and a *spread.*
  - If the shape is *skewed or has outliers*, report the *median* and *IQR.*
  - If the shape is *symmetric*, report the *mean* and *standard deviation* and possibly the median and IQR as well.

TULIP Team for Universal Learning and Intelligent Processing

88

Five number summaries and boxplots

- The five-number summary of a distribution reports its median (Q2), quartiles (Q1, Q3), and extremes (maximum and minimum).
  - Example: a histogram of the Average Wind Speed for every day in 1989 is shown.
    - The distribution is unimodal and skewed to the right.
    - The high value may be an outlier
    - Median daily wind speed is about 1.90 mph and the IQR is reported to be 1.78 mph.
    - Can we say more?
  - The five-number summary for the daily wind speed is as shown:



Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20

TULIP Team for Universal Learning and Intelligent Processing

89

Making Boxplots

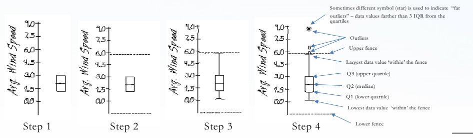
- A boxplot is a graphical display of the five-number summary.
  - Boxplots are particularly useful when comparing groups, and for identifying outliers.
- Steps to draw a box plot is:
  1. Draw a single vertical axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.
  2. Erect "fences" around the main part of the data. (Note that for the above example IQR = Q3 – Q1 = 2.93-1.15=1.78)
     - The upper fence is 1.5 IQRs above the upper quartile. (Q3 + 1.5*IQR) = 2.93 + 1.5*1.78 = 5.6
     - The lower fence is 1.5 IQRs below the lower quartile. (Q1 - 1.5*IQR) = 1.15 - 1.5*1.78 = -1.2
     - Note: the fences only help with constructing the boxplot and should not appear in the final display.
  3. Use the fences to draw "whiskers."
     - Draw lines from the ends of the box up and down to the most extreme data values found within the fences.
     - If a data value falls outside one of the fences, we do not connect it with a whisker.
  4. Add the outliers by displaying any data values beyond the fences with special symbols.
     - We often use a different symbol for "far outliers" that are farther than 3 IQRs from the quartiles (Q3 + 3*IQR) or (Q1 - 3*IQR).

Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20

TULIP Team for Universal Learning and Intelligent Processing

90

Making Boxplots



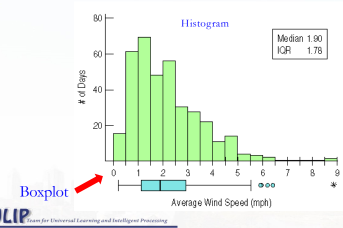
Max	8.67
Q3	2.93
Median	1.90
Q1	1.15
Min	0.20

TULIP Team for Universal Learning and Intelligent Processing

91

Box plots…

- Compare the histogram and boxplot for daily wind speeds:



TULIP Team for Universal Learning and Intelligent Processing

92

Parallel Box Plots

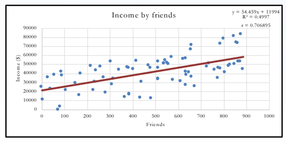
- E.g., Assets of larger companies by market sector

TULIP Team for Universal Learning and Intelligent Processing

93

Scatter plots

- **Scatterplots** have the
  - explanatory/independent variable on the horizontal (x) axis
  - and the response/dependent variable on the vertical (y) axis
  - Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two *quantitative* variables.
  - In a scatterplot, you can see patterns, trends, relationships, and possible outliers
- **Correlation** numerically measures the strength of a linear relationship between two quantitative variables
- **Linear regression** produces a linear equation that further numerically describes a linear relationship between two quantitative variables



TULIP Team for Universal Learning and Intelligent Processing

94

Correlation coefficient and Coefficient of Determination

- The correlation coefficient (r) gives us a numerical measurement of the strength of the "linear relationship" between the explanatory and response variables.
- Formula

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}$$

x, y – mean of x and y

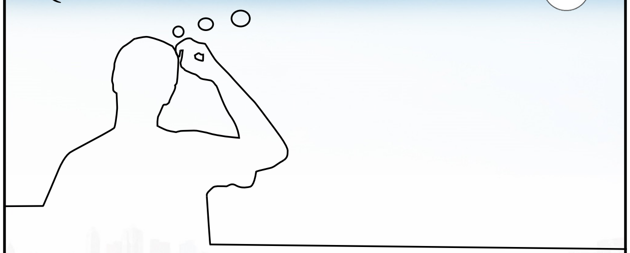
s_x, s_y – Std. deviation of x and y

- Correlation is always between -1 and +1.
  - Correlation *can* be exactly equal to -1 or +1, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
  - A correlation near zero corresponds to a weak linear association.
  - $r<0$ stands for negatively correlated and $r>0$ stands for positively correlated
- Coefficient of determination (r^2):
  - The squared correlation, r^2, gives the fraction of the data's variance accounted for by the model.

TULIP Team for Universal Learning and Intelligent Processing

95

Questions?



TULIP Team for Universal Learning and Intelligent Processing

97