

# Statistic Inference

3007/7059 Artificial Intelligence

Slides by Lingqiao Liu, Wei Zhang

School of Computer Science  
University of Adelaide

# Statistic Inference

- **Inference** is the process of drawing a **conclusion** by applying rules (of logic, statistics etc.) to **observations** or hypotheses.
- Here we are mainly interested in **statistical inference**, i.e. our knowledge is uncertain, encoded using probability

# Probability

- ▶ Begin with a set  $\Omega$  called the **sample space**.  
E.g., 6 possible rolls of a die,  $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - ▶  $\omega \in \Omega$  is a **sample point**.  
E.g.,  $\omega = 1$ ,  $\omega = 2$ ,  $\omega = 3$ ,  $\omega = 4$ ,  $\omega = 5$ ,  $\omega = 6$
  - ▶ A **probability space** or **probability model** is a sample space with an assignment  $P(\omega)$  for every  $\omega \in \Omega$  s.t.
    - ▶  $0 \leq P(\omega) \leq 1$
    - ▶  $\sum_{\omega} P(\omega) = 1$
- E.g.,  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$

# Event

- ▶ An **event**  $a$  is any subset of  $\Omega$   
E.g.,  $a$  is the event of rolling a die and obtaining less than 4, i.e.  $a = \{1, 2, 3\}$ .
- ▶ The probability of an event is the sum of the probabilities of the sample points contained in the event, i.e.

$$P(a) = \sum_{\forall \omega \in a} P(\omega)$$

E.g.  $P(a) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

- ▶ An **atomic event** contains only one sample point.  
E.g.,  $b$  is the event of rolling a die and obtaining 6, i.e.  $b = \{6\}$ .

# Disjunctions & Conjunctions

- ▶ The **disjunction** of two events  $a$  and  $b$ , written as " $a \vee b$ ", is the event where the outcomes satisfy *either  $a$  or  $b$* .
- ▶ The **conjunction** of two events  $a$  and  $b$ , written as " $a \wedge b$ ", is the event where the outcomes satisfy *both  $a$  and  $b$* .
- ▶ E.g.,  
 $a$  is the event of rolling a die and obtaining less than 5,  
 $b$  is the event of rolling a die and obtaining more than 2,  
then
$$a \vee b = \{1, 2, 3, 4, 5, 6\}$$
and
$$a \wedge b = \{3, 4\}.$$



# Axioms of Probability

1. All probabilities are between 0 and 1.

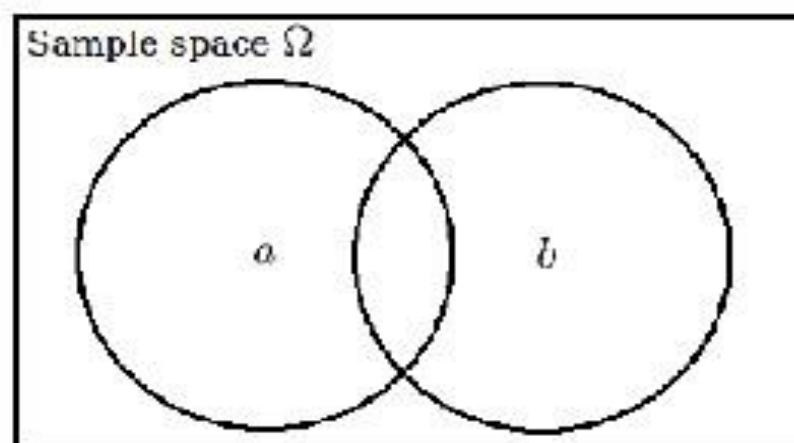
$$0 \leq P(a) \leq 1$$

2. Necessarily true propositions have probability 1, and necessarily false propositions have probability 0.

$$P(\text{true}) = 1, \quad P(\text{false}) = 0$$

3. The probability of a **disjunction** of two events  $a$  and  $b$  is

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$



# Axioms of Probability

- ▶ Example:

$a$  is the event of rolling a die and obtaining less than 5,  
 $b$  is the event of rolling a die and obtaining more than 2,  
then

$$P(a \wedge b) = P(3) + P(4) = 1/3$$

$$P(a \vee b) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1$$

- ▶ This can also be obtained using the third axiom as

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b) = 2/3 + 2/3 - 1/3 = 1$$

# Random Variable

- ▶ A **random variable** is a function from sample points to some range, e.g., the reals or Booleans.

E.g.,  $\Omega = \{1, 2, 3, 4, 5, 6\}$  with random variable *Odd*.

$Odd(1) = true, Odd(2) = false, Odd(3) = true, \dots$

- ▶  $P$  induces a **probability distribution** for any random variable  $X$ :

$$P(X = x_i) = \sum_{\{\omega: X(\omega) = x_i\}} P(\omega)$$

Example:

$$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$$

- ▶ When it is clear from the context which random variable is in question, we can directly write

$$P(X = x_i) = P(x_i)$$



# Types of random variables

- ▶ **Boolean** random variables

E.g.,  $Cavity = true$  or  $Cavity = false$

$Cavity = true$  is a proposition

- ▶ **Discrete** random variables

E.g.,  $Weather$  is one of  $\{sunny, rain, cloudy, snow\}$

$Weather = rain$  is a proposition

Values must be **exhaustive and mutually exclusive**

Boolean random variables are also discrete random variables

- ▶ **Continuous** random variables

E.g.,  $Temp = 21.6$  is a proposition (but one that has vanishingly small probability of being true)

More normally we consider *inequality propositions* for continuous random variables, e.g.,  $Temp < 22.0$ .

# Types of random variables

- Note: We usually write random variables in uppercase (e.g. *Cavity*, *Weather*, *Temp*) and values/instantiations of random variables in lowercase (e.g. *true*, *false*, *sunny*, *rain*, *cloudy*, *snow*).

# Discrete probability distribution

- ▶ Probabilities of values of a discrete random variable define a **discrete probability distribution**.

- ▶ Examples:

- ▶  $P(Cavity = true) = 0.1, P(Cavity = false) = 0.9$

$$P(Cavity) = \langle \underbrace{0.1}_{true}, \underbrace{0.9}_{false} \rangle$$

- ▶  $P(Weather = sunny) = 0.72, P(Weather = rain) = 0.1,$   
 $P(Weather = cloudy) = 0.08, P(Weather = snow) = 0.1$

define

$$P(Weather) = \langle \underbrace{0.72}_{sunny}, \underbrace{0.1}_{rain}, \underbrace{0.08}_{cloudy}, \underbrace{0.1}_{snow} \rangle$$

- ▶ Probability distributions are **normalised**, i.e., they sum to 1.
- ▶ Discrete probability distributions are also called **probability mass functions**.

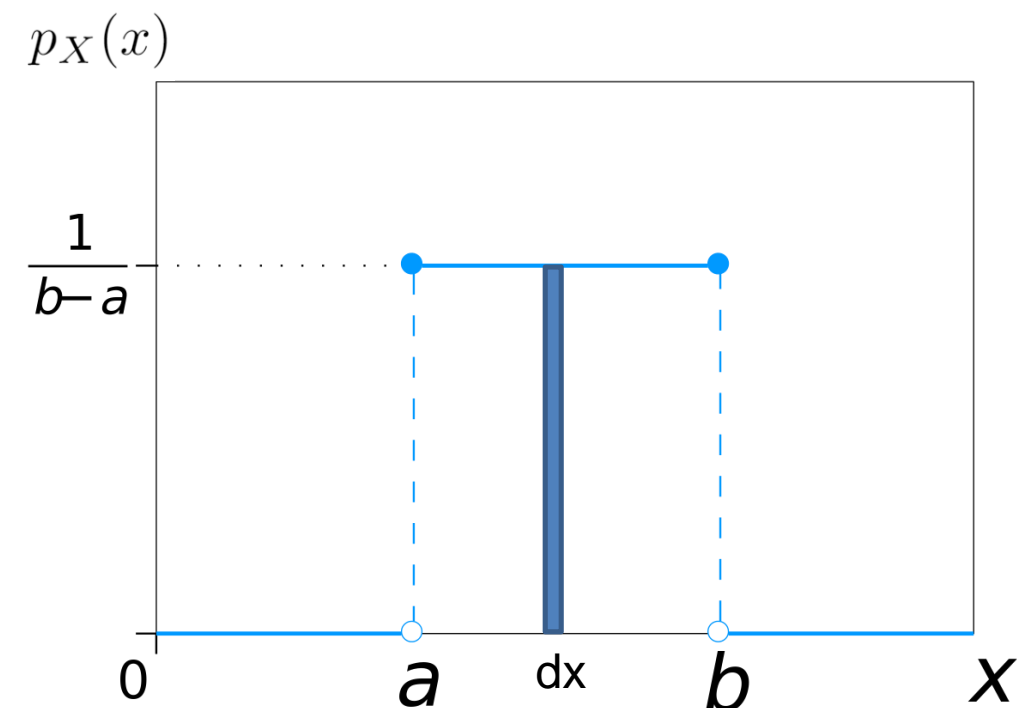
# Probability Density Function

- Probability density function (PDF)

$$P(a \leq X \leq b) = \int_a^b p_X(x) dx$$

- A simple example is the **uniform distribution**

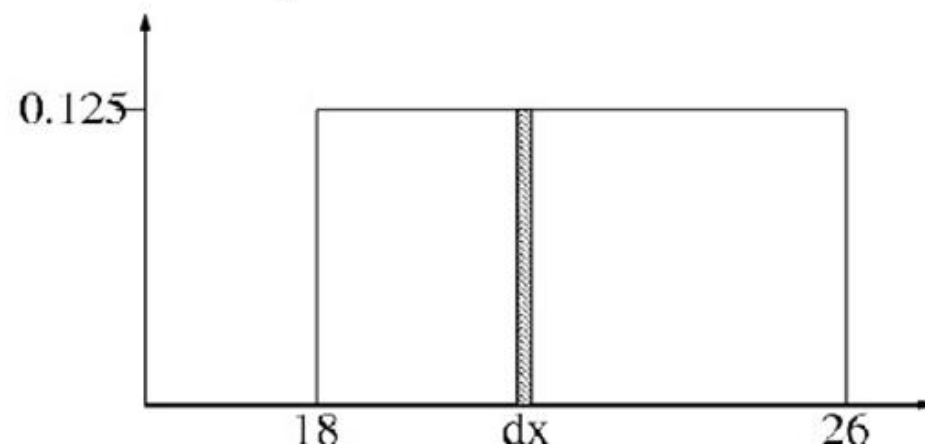
$$p_X(x) = \begin{cases} 0, & x < a \text{ or } x > b \\ \frac{1}{b-a}, & a \leq x \leq b \end{cases}$$





# Probability Density Function

$$p_X(x) = \begin{cases} 0.125 & \text{if } 18 \leq x \leq 26 \\ 0 & \text{otherwise} \end{cases}$$



A PDF is a density: To obtain the probability of an event we have to integrate over the sample points belonging to the event.

Example: With the above uniform distribution

$$P(15 \leq X \leq 20) = 0 + \int_{18}^{20} p(x)dx = 0.25$$

Also,  $p_X(20.5)$  is obtained by taking the limit

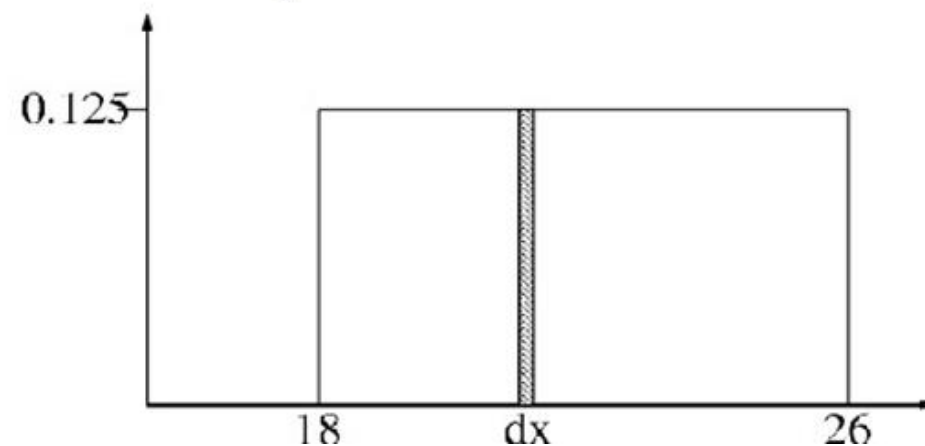
$$\lim_{dx \rightarrow 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

# Probability Density Function

A PDF integrates to 1, i.e.

$$\int_x p_X(x) dx = 1$$

$$p_X(x) = \begin{cases} 0.125 & \text{if } 18 \leq x \leq 26 \\ 0 & \text{otherwise} \end{cases}$$

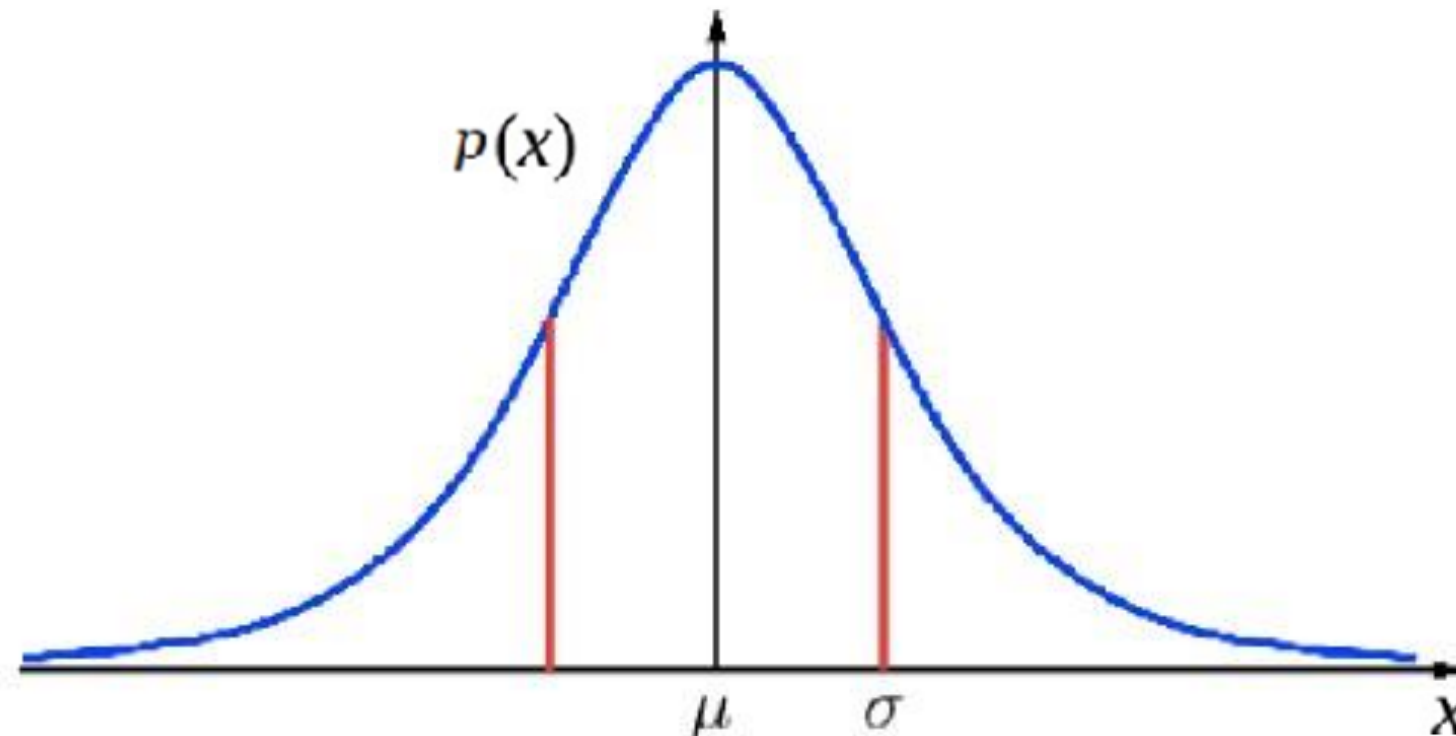


# The Gaussian Distribution

An important probability distribution for continuous random variables is the **Gaussian distribution**, also called the **Normal distribution**.

$$X \sim N(\mu, \sigma) \quad p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

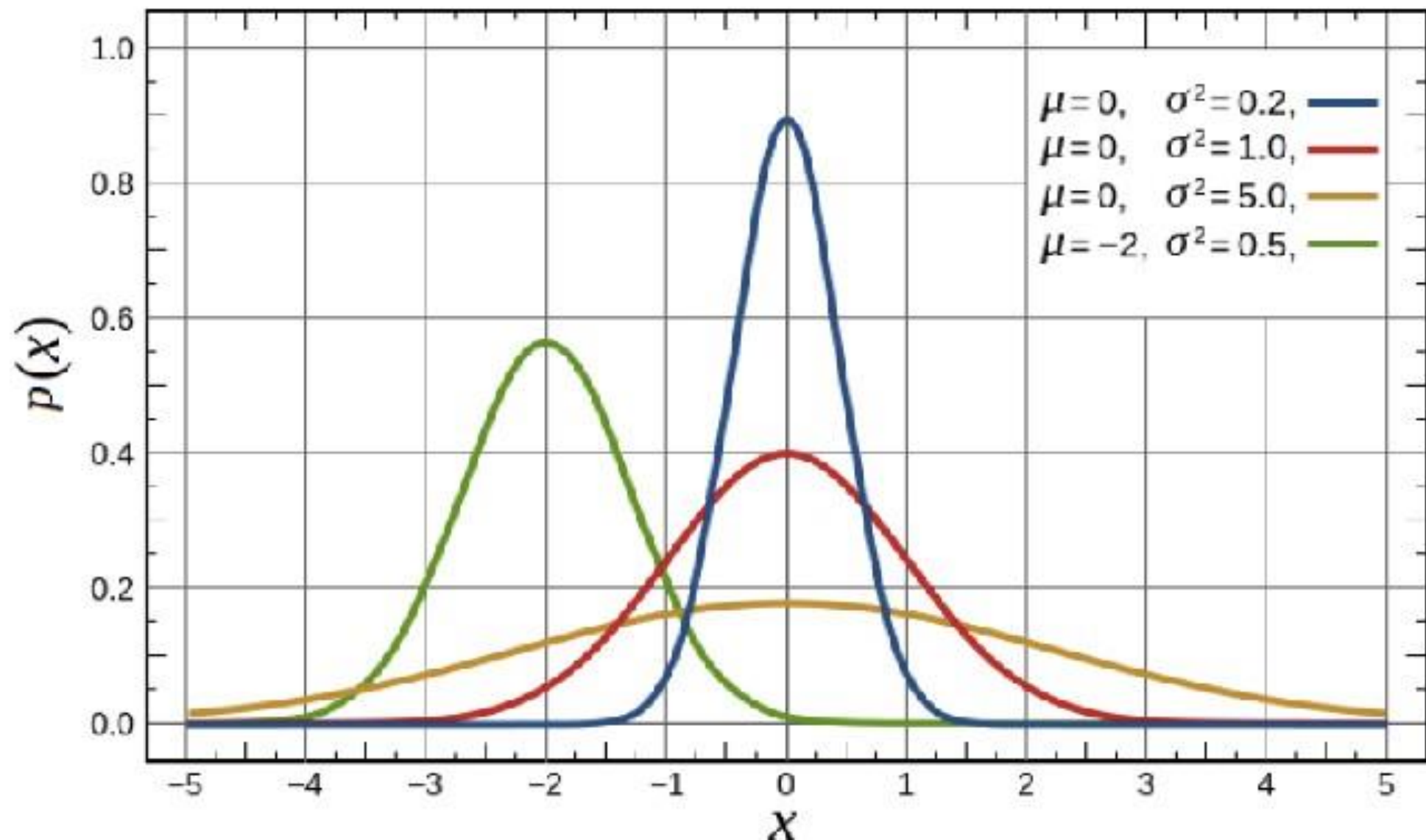
A Gaussian contains two parameters: the **mean**  $\mu$  and **standard deviation**  $\sigma$ .



# The Gaussian Distribution

The  $\mu$  specifies the “location” of the Gaussian, while the  $\sigma$  controls the spread.

Example:





# The Joint Probability Distribution

- ▶ **Joint probability distribution** for a set of random variables gives the probability of every possible combination of atomic event on those random variables (i.e., every sample point).
- ▶ The joint probability of a set of random variables  $X_1, \dots, X_n$  is written

$$P(X_1, \dots, X_n)$$

- ▶ E.g., the joint probability of *Weather* and *Cavity*  $P(\text{Weather}, \text{Cavity})$  is a  $4 \times 2$  matrix of values:

<i>Weather</i> =	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity</i> = <i>true</i>	0.144	0.02	0.016	0.02
<i>Cavity</i> = <i>false</i>	0.576	0.08	0.064	0.08

$P(\text{Weather} = \text{rain}, \text{Cavity} = \text{true})$  gives the probability that the weather is raining *and* I have cavity.

- ▶ Notice that the **values in the table sum to 1**.

# Marginalisation

To get the joint probability of a subset of the variables, we **marginalise** (sum out) the other variables we are not interested in:

$$P(x_1, x_2, \dots, x_{N-1}) = \sum_{\forall \text{ values of } X_N} P(x_1, x_2, \dots, x_{N-1}, X_N)$$

Example:

<i>Weather =</i>	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity = true</i>	0.144	0.02	0.016	0.02
<i>Cavity = false</i>	0.576	0.08	0.064	0.08

$P(\text{Cavity} = \text{false})$

# Marginalisation

To get the joint probability of a subset of the variables, we **marginalise** (sum out) the other variables we are not interested in:

$$P(x_1, x_2, \dots, x_{N-1}) = \sum_{\forall \text{ values of } X_N} P(x_1, x_2, \dots, x_{N-1}, X_N)$$

Example:

<i>Weather =</i>	<i>sunny</i>	<i>rain</i>	<i>cloudy</i>	<i>snow</i>
<i>Cavity = true</i>	0.144	0.02	0.016	0.02
<i>Cavity = false</i>	0.576	0.08	0.064	0.08

$$\begin{aligned} &P(\text{Cavity} = \text{false}) \\ &= \sum_{\forall \text{ values of } \text{Weather}} P(\text{Cavity} = \text{false}, \text{Weather}) \\ &= P(\text{Cavity} = \text{false}, \text{Weather} = \text{sunny}) \\ &\quad + P(\text{Cavity} = \text{false}, \text{Weather} = \text{rain}) \\ &\quad + P(\text{Cavity} = \text{false}, \text{Weather} = \text{cloudy}) \\ &\quad + P(\text{Cavity} = \text{false}, \text{Weather} = \text{snow}) \\ &= 0.8 \end{aligned}$$



# Conditional Probability

A **conditional probability** is the probability of an event  $a$  given the occurrence/observation of some other event  $b$ .

This is expressed as

$$P(a|b)$$

or the probability of  $a$  given  $b$ .

Example:

$$P(Cavity = true | Toothache = true) = 0.8$$

Given the observation that the patient has toothache, the probability that he/she has cavity is 0.8.



# Conditional Probability

Formal definition of conditional probability:

$$P(A|B) = \frac{P(A, B)}{P(B)} \text{ if } P(B) \neq 0$$

The **product rule** gives an alternative formulation:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

The **chain rule** is derived by successive application of product rule:

$$\begin{aligned} & P(X_1, \dots, X_n) \\ = & P(X_1, \dots, X_{n-1})P(X_n|X_1, \dots, X_{n-1}) \\ = & P(X_1, \dots, X_{n-2})P(X_{n-1}|X_1, \dots, X_{n-2})P(X_n|X_1, \dots, X_{n-1}) \\ = & \dots \\ = & \prod_{i=1}^n P(X_i|X_1, \dots, X_{i-1}) \end{aligned}$$

# Conditioning

Combining marginalisation and the product rule yields the **conditioning rule**:

$$\begin{aligned} P(A) &= \sum_{\forall \text{ values of } B} P(A, B) \\ &= \sum_{\forall \text{ values of } B} P(A|B)P(B) \end{aligned}$$

# Statistic Inference

Statistical inference (or probabilistic inference) is the computation from observed evidence of probabilities for query propositions.

The joint distribution of the variables involved is used as the “knowledge base” from which the inference is conducted.

In other words, our knowledge about the domain is encoded in the joint distribution of variables.



# Example: Toothache and Cavity

Observation: The patient complains of toothache.

Query proposition: He has cavity.

Probability of query proposition:

$$P(Cavity = true | Toothache = true)$$

Knowledge base of dentist:

A joint distribution of the variables *Cavity*, *Toothache* and *Catch* (the dentist's steel probe catches in a tooth).

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

Here we use a shorthand: *Cavity* = *true* is written as *cavity*, while *Cavity* = *false* is written as  $\neg$ *cavity*. Similarly for the other variables.



# Example: Toothache and Cavity

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

$$\begin{aligned}
 P(\text{cavity}|\text{toothache}) &= \frac{P(\text{cavity}, \text{toothache})}{P(\text{toothache})} \quad (\text{product rule}) \\
 &= \frac{\sum_{\forall \text{Catch}} P(\text{cavity}, \text{toothache}, \text{Catch})}{\sum_{\forall \text{Catch}} \sum_{\forall \text{Cavity}} P(\text{Cavity}, \text{toothache}, \text{Catch})} \quad (\text{marginalise}) \\
 &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} \\
 &= 0.6
 \end{aligned}$$

What is the probability that there is no cavity given the patient has toothache? Prove it.

# Example: Toothache and Cavity

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

$$P(\neg \text{cavity} \mid \text{toothache}) = \frac{P(\neg \text{cavity}, \text{toothache})}{P(\text{toothache})}$$

$$= \frac{\sum_{\forall \text{Catch}} P(\neg \text{cavity}, \text{toothache}, \text{Catch})}{\sum_{\forall \text{Catch}} \sum_{\forall \text{Cavity}} P(\text{Cavity}, \text{toothache}, \text{Catch})}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

# General Rule for Statistical Inference

Notice that when we compute  $P(\text{cavity}|\text{toothache})$  and  $P(\neg\text{cavity}|\text{toothache})$  the denominator  $P(\text{toothache})$  is fixed—  
The denominator is the **probability of the observed evidence**.

Moreover, the denominator provides a **normalisation constant** which ensures that  $P(\text{cavity}|\text{toothache})$  and  $P(\neg\text{cavity}|\text{toothache})$  sum to 1.

We can express this normalisation constant as  $\alpha = \frac{1}{P(\text{toothache})}$  and solve for it at the end.

This allows us to rewrite the preceding inference procedure as

$$\begin{aligned} P(\text{Cavity} \mid \text{toothache}) &= \alpha P(\text{Cavity}, \text{toothache}) \\ &= \alpha [P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg\text{catch})] \\ &= \alpha [P(\text{cavity}, \text{toothache}, \text{catch}) + P(\neg\text{cavity}, \text{toothache}, \text{catch}) \\ &\quad + P(\text{cavity}, \text{toothache}, \neg\text{catch}) + P(\neg\text{cavity}, \text{toothache}, \neg\text{catch})] \\ &= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] \\ &= \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle \end{aligned}$$

$P(\text{cavity}, \text{toothache})$

$P(\neg\text{cavity}, \text{toothache})$

# General Rule for Statistical Inference

In a general problem domain, let

- ▶  $X$  represent the query variable (e.g. *Cavity*)
- ▶  $E$  be the set of evidence variables (e.g. *Toothache*)
- ▶  $e$  be the observed values of  $E$  (e.g. *toothache*)
- ▶  $Y$  be the remaining unobserved variables (e.g. *Catch*)

The **general rule of statistical inference** can be written as

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$



# Open Questions

Perform statistical inference for the following propositions:

1. The patient has toothache given that he has cavity.
2. The probe did not catch the patient's tooth given that he has toothache.

Calculate the probability distribution of

1. Toothache given that the probe caught the patient's tooth.

