ETC5242 Class 2

Probability & MLE

授课老师：Joe

ETC5242平时班 – Class 1

- **Week 6**
  - **Probability**

  - **Maximum likelihood estimate (MLE)**

  - **Bootstrapping for model parameters**

ETC5242平时班 – Class 2

There are three fundamental rules of probability.

1. If $\Pr(A)$ is the probability associated with event A, then $0 \leq \Pr(A) \leq 1$
2. The total probability of all outcomes in the sample space is 1
3. If $A_1, A_2, \ldots$ is a sequence of **mutually exclusive** events, then

$$\Pr(A_1 \cup A_2 \cup \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots$$

ETC5242平时班 – Class 2

Mutually exclusive events are sometimes referred to as **disjoint** events.

These are events that cannot happen simultaneously (their intersection is empty)

- From the third axiom, if events A_1 and A_2 are **mutually exclusive**
- Then $\Pr(A_1 \text{ or } A_2) = \Pr(A_1) + \Pr(A_2)$

Non-disjoint events

- Outcomes that overlap are called **non-disjoint** events
  - We need a more general rule for working out their probabilities

Example: Consider events $(X > 2)$ and $(X < 4)$

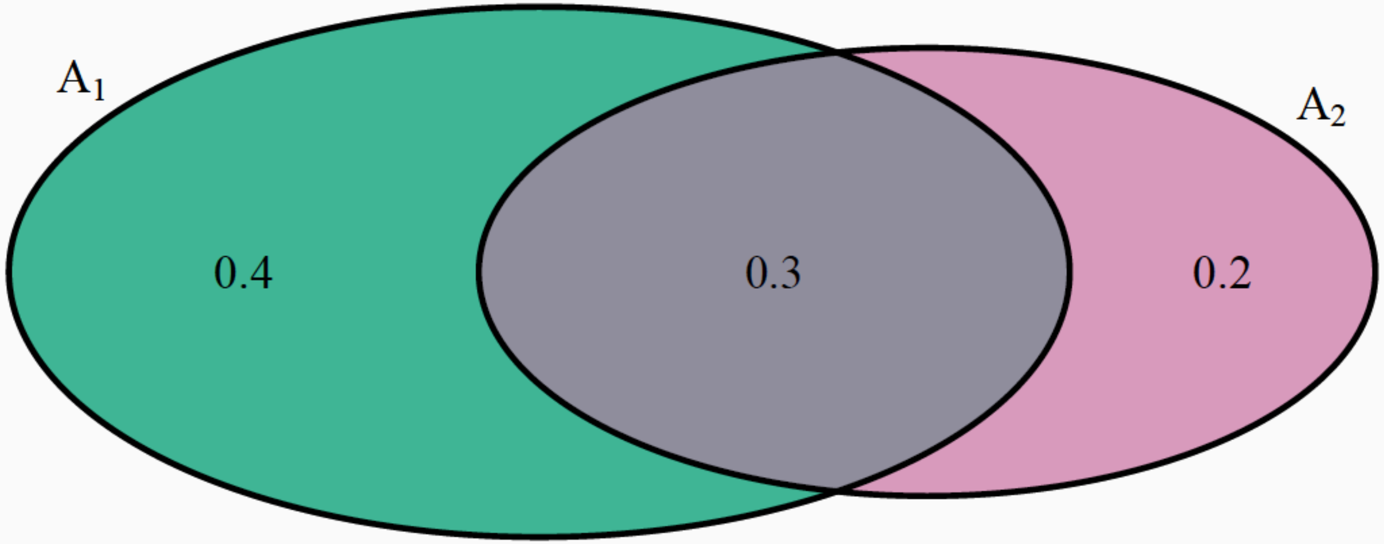
- These are NOT disjoint!
- No "double counting" of probability allowed!!
- Need to take out the "double counted" (overlap) part:

$$\Pr(X > 2 \cup X < 4) = \Pr(X > 2) + \Pr(X < 4) - \Pr(X > 2 \cap X < 4)$$

ETC5242平时班 – Class 2

Suppose $\Pr(A_1) = 0.7$ and $\Pr(A_2) = 0.5$ and $\Pr(A_1 \cap A_2) = 0.3$

Then $\Pr(A_1 \cup A_2) = 0.7 + 0.5 - 0.3 = 0.9$



Example

$X = x$	$\Pr(X = x)$
$X = 1$	1/2
$X = 2$	1/8
$X = 3$	1/4
$X = 4$	1/8

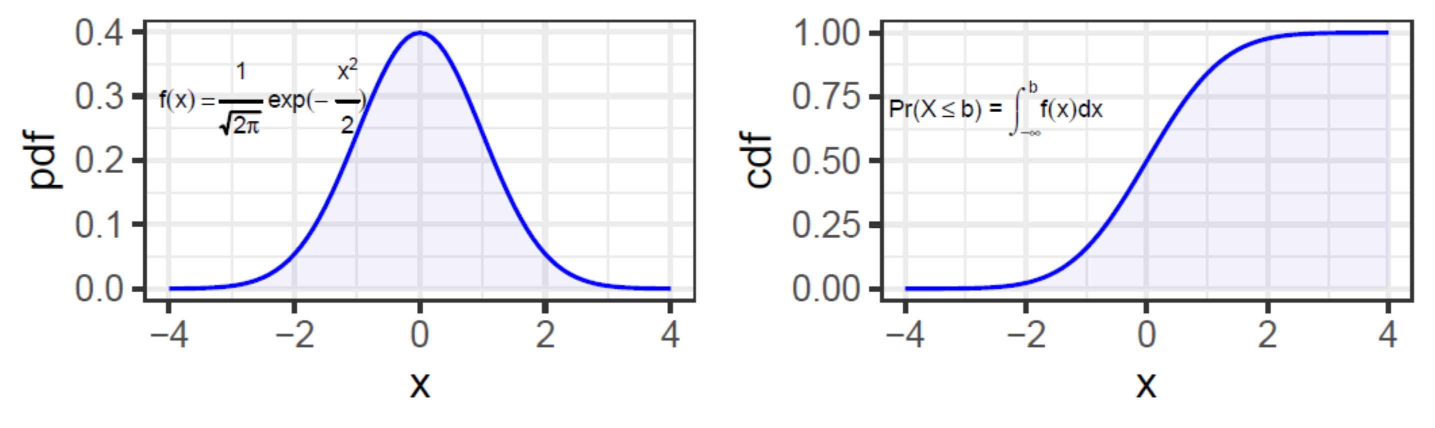
Find probabilities for given events:

1. $\Pr(X = 2)$
2. $\Pr(X \leq 2)$
3. $\Pr(X \text{ is even})$
4. $\Pr(X < 4)$
5. $\Pr(X > 2 \text{ and } X < 3) = \Pr(X > 2 \cap X < 3)$
6. $\Pr(X > 2 \text{ or } X < 3) = \Pr(X > 2 \cup X < 3)$

ETC5242平时班 – Class 2

Example: Continuous random variable over an infinite sample space

If $X \sim N(0, 1)$



Find

1. $\Pr(X = 1)$
2. $\Pr(X < 1)$
3. $\Pr(X \text{ is even})$
4. $\Pr(X < -\frac{1}{2})$
5. $\Pr(X > 2 \text{ and } X < 3) = \Pr(X > 2 \cap X < 3)$
6. $\Pr(X > 2 \text{ or } X < 3) = \Pr(X > 2 \cup X < 3)$

- **Two** processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other

Multiplication Rule for independent processes

- If A and B are simple events from two **different** and **independent** processes
  - two compound processes but "simple" relationship between them due to assumed independence
- Then the event that **both** A and B occur corresponds to an **intersection**
  - the joint probability can calculated as the product of the individual probabilities:

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$$

- Similarly, if there are k simple events $A_1, A_2, \ldots, A_k$ from k independent processes
  - Then the probability that **all events** will occur is given by

$$\Pr(A_1) \times \Pr(A_2) \times \cdots \times \Pr(A_k)$$

ETC5242平时班 – Class 2

Example: Two independent coin tosses

- For $i = 1$ and $i = 2$

	$X_i = x$	$\Pr(X_i = x)$
x	Head	0.5
	Tail	0.5

- If two fair coin tosses are **independent**, then any **joint probability** about **each outcome** will be the **product of the two marginal probabilities** about each outcome.

1. *What is the probability of a "Head" on the **first** toss and a "Tail" on the **second** toss?*

$$
\begin{aligned}
\Pr(X_1 &= \text{Head and } X_2 = \text{Tail}) \\
&= \Pr(X_1 = \text{Head}) \times \Pr(X_2 = \text{Tail}) \\
&= (0.5) \times (0.5) \\
&= 0.25
\end{aligned}
$$

ETC5242平时班 – Class 2

Example: Left-handedness

- About 9% of people in the population are left-handed
- Suppose 2 people are selected at random from the Australian population
  - (Assume population is so large that the outcomes for the two selected are independent)

1. What is the probability that both people selected are left-handed?

$(0.09)(0.09) = 0.0081$

8. What is the probability that both people selected are right-handed?

- $(1-0.09)(1-0.09) = (0.91)^2 = 0.8281$

9. What is the probability that one person selected is left-handed, and the other is right-handed?

- Note probabilities of all events must sum to one
- $1-(0.0081+0.8281) = 0.1638$

ETC5242平时班 – Class 2

Example: Travel survey data

- Random sample survey of 100 people with particular credit card
- *Are you planning to travel abroad next year?*
- Take these as proportional to "true probabilities"

		Age group			
		25 or less	26-40	41 or more	Total
Response	Yes	2	12	15	29
	Undecided	5	10	16	31
	No	10	15	15	40
	Total	17	37	46	100

1. $\Pr($Card holder intends to travel over next 12 months$)$?
2. $\Pr($Card holder intends to travel over next 12 months OR is undecided$)$?
3. $\Pr($Card holder intends to travel over next 12 months AND is 25 years old or less$)$?

ETC5242平时班 – Class 2

- Joint probability
  - probability of outcomes for two or more variables or processes

- Marginal probability
  - probability of outcomes for a single variable or process

- Conditional probability
  - probability of outcomes for a single variable or process **given information about a second variable or process**

Example: Travel survey data (revisited)

Probability for all possible pairs

Age group AND Response combination	Prob
Yes response AND (25 or less)	0.02
Yes response AND (26-40)	0.12
Yes response AND (41 or more)	0.15
Undecided response AND (25 or less)	0.05
Undecided response AND (26-40)	0.10
Undecided response AND (41 or more)	0.16
No response AND (25 or less)	0.10
No response AND (26-40)	0.15
No response AND (41 or more)	0.15
	1.0

ETC5242平时班 – Class 2

The conditional probability for a single **outcome of interest** A, given **conditioned on an event** B, is defined as

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

General Multiplication Rule

- If A and B represent two outcomes or events, then

$$\Pr(A \text{ and } B) = \Pr(A \mid B) \times \Pr(B)$$

- Here A is the outcome of interest, and B is the event being conditioned upon
- Alternatively,

$$\Pr(A \text{ and } B) = \Pr(B \mid A) \times \Pr(A)$$

ETC5242平时班 – Class 2

Expected value & variance of discrete random variable

- If X takes outcomes $x_1, \ldots, x_k$ with probabilities $\Pr(X = x_1), \ldots, \Pr(X = x_k)$, respectively, then the **expected value** of X is

$$E[X] = x_1 \Pr(X = x_1) + \cdots + x_k \Pr(X = x_k) = \sum_{i=1}^{k} x_i \Pr(X = x_i)$$

$$Var(X) = (x_1 - \mu)^2 \Pr(X = x_1) + \cdots + (x_k - \mu)^2 \Pr(X = x_k) = \sum_{i=1}^{k} (x_i - \mu)^2 \Pr(X = x_i)$$

ETC5242平时班 – Class 2

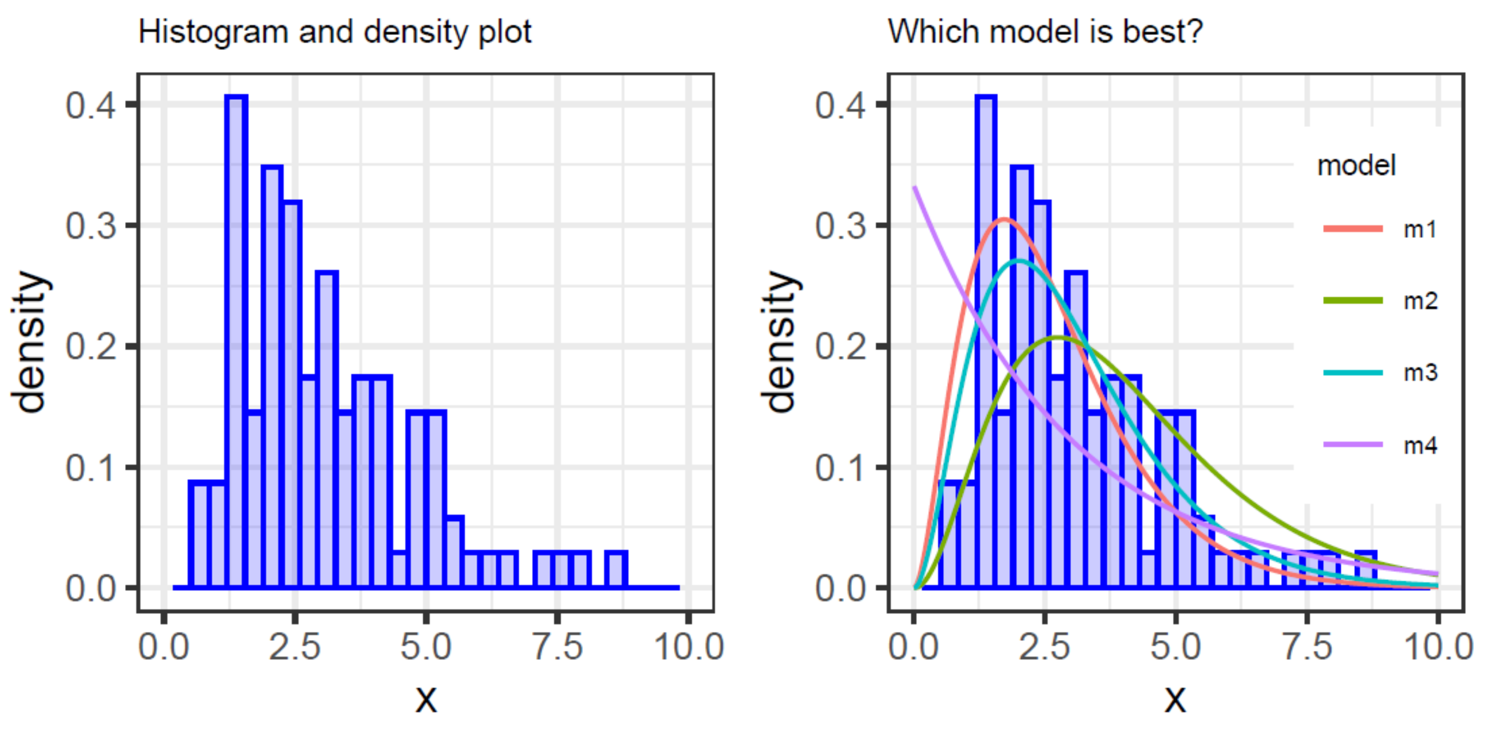
Expected value and variance of continuous random variable

$$E[X] = \mu = \int_{-\infty}^{\infty} x\, f(x) dx$$

$$Var(X) = E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx$$

ETC5242平时班 – Class 2

- Which distributions might fit this data?
  - A normal distribution? An exponential? A gamma distribution? Something else?



- Assuming the data are a random sample, we need to **choose a model** $F_X(x \mid \theta)$
  - We fit models using the sample and well-established distributional families
- Once we choose a model, we'll need to **estimate** the parameter θ
  - use the **maximum likelihood estimation** (MLE) method
- A fitted model will imply an estimate of the population mean
  - and other features

ETC5242平时班 – Class 2

Likelihood Function

If $x_1, x_2, ..., x_n \overset{i.i.d.}{\sim} F_X(x \mid \theta)$, then the likelihood function is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

And the **MLE** for θ is

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$$

ETC5242平时班 – Class 2

Gaussian density function (normal distribution)

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Likelihood(Probability) of observing the three data points, 9, 9.5 and 11 given a particular gaussian density function, But we don't know the two parameters yet

We want to maximise this joint probability

$$P(9, 9.5, 11;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

ETC5242平时班 – Class 2

Optimising the likelihood function

- It is often easier to maximise the **log-likelihood function**

$$\ell_n(\theta) = \ln \mathcal{L}_n(\theta) = \left[\sum_{i=1}^{n} \ln f_X(x_i|\theta) \right]$$

- The **same** $\hat{\theta}_{MLE}$ maximises $\mathcal{L}_n(\theta)$ and $\ell_n(\theta)$
- In simple cases we can solve for $\hat{\theta}_{MLE}$ through differentiation
  - set first derivative of $\ell_n(\theta)$ equal to zero and solve
  - then check the second derivative of $\ell_n(\theta)$ is negative at $\hat{\theta}_{MLE}$
- More generally MLE is found using numerical optimisation on a computer

Very handy in R

```
fit <- fitdistr(x, "gamma")
fit
```

```
     shape       rate   
  3.4697     1.1235  
 (0.4690)   (0.1634)
```

ETC5242平时班 – Class 2

Bootstrapping for confidence interval of model parameters

1. Generate a Bootstrap sample of B potential $\hat{\theta}$ values
   - For each b in $1 : B$
     - resample n draws from the observed data values, with replacement
     - label these values as $\{x_1^{[b]}, x_2^{[b]}, \ldots, x_n^{[b]}\}$
     - compute the MLE $\hat{\theta}^{[b]}$ by maximising $\mathcal{L}_n^{[b]}(\theta)$, constructed from the bootstrap sample
   - Bootstrap sample: $\{\hat{\theta}^{[1]}, \hat{\theta}^{[2]}, \ldots, \hat{\theta}^{[B]}\}$
2. Use the empirical distribution from this Bootstrap sample to approximate the sampling distribution of $\hat{\theta}_{MLE}$
3. Construct an approximate 95% confidence interval by selecting interval from 2. with (empirical) probability (at least) 95%
   - (lower) 2.5% quantile to 97.5% quantile