SM-4331 Advanced Statistics
Lecture 0 (Introduction)

Dr Haziq Jamil

FOS M1.09

Universiti Brunei Darussalam

Semester II 2019/20

Outline

Admin

- Lecturer information

Dr. Haziq Jamil
Assistant Professor in Statistics
Room FOS M1.09
haziq.jamil@ubd.edu.bn

- Class times
  - Tuesdays 2.10pm–4.00pm
  - Wednesdays 11.50am–1.40pm
- Slides, materials and announcements will be on Canvas.
- Prerequisites: SM-2205 Intermediate statistics, or equivalent knowledge thereof.

Course contents

- Chapter 1: Estimation theory
  - Inequalities for probabilities and expectations
  - Convergence in probability, convergence in distribution
  - Point estimation
  - Desirable properties of estimators
  - Maximum likelihood estimation
  - Confidence intervals
- Chapter 2: Sampling
  - Sampling as part of survey methodology
  - Simple random sampling
  - Stratified sampling
  - Cluster sampling

Course contents

- Chapter 3: Hypothesis testing
  - p-values
  - Procedure for statistical testing
  - Wald test, t-tests, likelihood ratio tests, χ^2 tests
- Chapter 4: Non-parametric methods
  - Tests of normality
  - Testing in cases where data are non-normal
  - Bootstrap
- Chapter 5: Multivariate methods
  - Bivariate normal distributions and its properties
  - Multivariate normal distributions and its properties

Format

- Lectures
  - "Power-point" style presentations
  - You are encouraged to write your own notes
  - Supplement lectures with individual readings
- Office hours
  - These provide an opportunity for you to meet with me to ask any question you might have regarding the course
  - You may come in groups or individually
  - You are encouraged to make use of my office hours
- Canvas discussion
  - If you have a question and would prefer to e-mail me, I would rather you submit this as a discussion item on Canvas for everyone's benefit

Assessment

- Formative assessment
  - Tutorials (no need to hand in)–**attempt questions prior to tutorial sessions**
  - Mini quizzes on Canvas (from time to time)
- Summative assessment
  - 2 $\times$ class tests (20% each)
  - Exam (60%)

Schedule

- First half of semester

	Day 1	Day 2	Day 3
Week 1 (start 6/1/20)	Lecture		Lecture
Week 2 (start 13/1/20)	Lecture	Lecture	Office Hours
Week 3 (start 20/1/20)	Tutorial	Lecture	Office Hours
Week 4 (start 27/1/20)	Lecture	Tutorial	Office Hours
Week 5 (start 3/2/20)	Lecture	Lecture	Office Hours
~~Week 6 (start 10/2/20)~~			
Week 7 (start 17/2/20)	**Test**		
	SEMESTER BREAK		

- Office hours time will be announced later

Schedule

- Second half of semester

	Day 1	Day 2
Week 8 (start 2/3/20)	Lecture	Tutorial
Week 9 (start 9/3/20)	Lecture	Office Hours
Week 10 (start 16/3/20)	Lecture	Tutorial
Week 11 (start 23/3/20)	Lecture	Office Hours
Week 12 (start 30/3/20)	Lecture	Tutorial
Week 13 (start 6/4/20)	Lecture	Office Hours
Week 14 (start 13/4/20)	Lecture	**Test**
Revision Week (start 20/4/20)	Revision Class	Office Hours

- Office hours time will be announced later

What is statistics?

Statistics is a scientific subject on collecting and analysing data.

- **Collecting** means designing experiments, designing questionnaires, designing sampling schemes, administration of data collection.
- **Analysing** means modelling, estimation, testing, forecasting.

What is statistics?

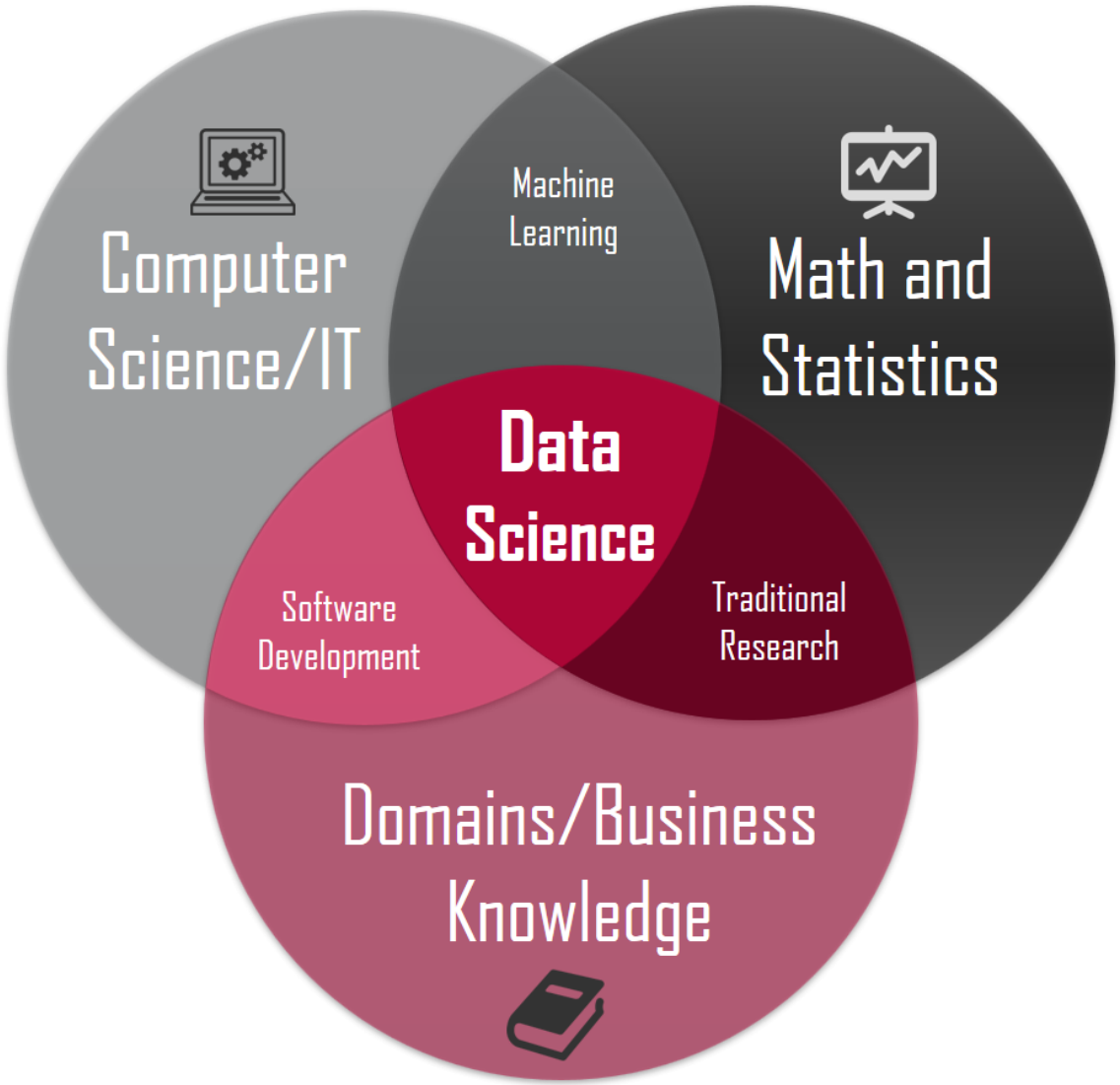
Statistics is a scientific subject on collecting and analysing data.

- **Collecting** means designing experiments, designing questionnaires, designing sampling schemes, administration of data collection.
- **Analysing** means modelling, estimation, testing, forecasting.

Statistics is an application-oriented mathematical subject; it is particularly useful or helpful in answering questions such as:

- Does a certain new drug prolong life for AIDS sufferers?
- Is global warming really happening?
- Are O-level and A-level examinations standard declining?
- Is the house market in Brunei oversaturated?
- Is the Chinese yuan undervalued? If so, by how much?

What is statistics? (cont.)



What is statistics? (cont.)

What to learn in statistics: **basic ideas**, methods and theory.

Some guidelines for learning/applying statistics:

- Understand what data say in each specific context. All the methods are just tools to help understand data.
- Concentrate on what to do and why, rather than concrete calculation and graphing.
- It may take a while to catch the basic idea of statistics... keep at it!

What is statistics? (cont.)

What to learn in statistics: **basic ideas**, methods and theory.

Some guidelines for learning/applying statistics:

- Understand what data say in each specific context. All the methods are just tools to help understand data.
- Concentrate on what to do and why, rather than concrete calculation and graphing.
- It may take a while to catch the basic idea of statistics... keep at it!

This course covers the necessary topics for anyone calling themselves statisticians to survive. These include sampling, estimation theory and hypothesis testing, among others.

Population vs sample

Two practical situations:

- A new type of tyre was designed to increase its lifetime. The manufacturer tested 120 new tyres and obtained the average lifetime (over those 120 tyres) as 56,956 km. So the manufacturer claims that the mean lifetime of the new tyres is 56,956 km.
- A newspaper sampled 1,000 of their readers, and 50 of them lived in Belait district. It claims that the proportion of Bruneians living in Belait is $50/1000 = 5\%$.

Population vs sample

Two practical situations:

- A new type of tyre was designed to increase its lifetime. The manufacturer tested 120 new tyres and obtained the average lifetime (over those 120 tyres) as 56,956 km. So the manufacturer claims that the mean lifetime of the new tyres is 56,956 km.
- A newspaper sampled 1,000 of their readers, and 50 of them lived in Belait district. It claims that the proportion of Bruneians living in Belait is $50/1000 = 5\%$.

In both cases, the conclusion is drawn on a **population** (i.e. all of the subjects concerned) based on the information from a **sample** (i.e. a subset of the population).

Population vs sample

Two practical situations:

- A new type of tyre was designed to increase its lifetime. The manufacturer tested 120 new tyres and obtained the average lifetime (over those 120 tyres) as 56,956 km. So the manufacturer claims that the mean lifetime of the new tyres is 56,956 km.
- A newspaper sampled 1,000 of their readers, and 50 of them lived in Belait district. It claims that the proportion of Bruneians living in Belait is $50/1000 = 5\%$.

In both cases, the conclusion is drawn on a **population** (i.e. all of the subjects concerned) based on the information from a **sample** (i.e. a subset of the population).

In the first case, it is **impossible** to measure the entire population. In the second case, it is (economically) unfeasible to measure the entire population. Therefore, errors are inevitable!

Population vs sample (cont.)

The **population** is an entire set of the objects concerned, and those objects are typically represented by some numbers. We do not know the entire population in practice.

Population vs sample (cont.)

The **population** is an entire set of the objects concerned, and those objects are typically represented by some numbers. We do not know the entire population in practice.

A **sample** is a randomly selected subset of a population, and is a set of known data in practice.

Parametric models

For a given problem, we typically assume a population to follow a **probability distribution** with pdf/pmf $f(x|\theta)$.

- The form of the distribution i.e. $f(\cdot|\theta)$ is known (e.g. normal, Poisson, exponential, etc.).
- The "specifics" of the distribution is (assumed to be) **not known**, but knowable if data were available.
- The unknown characteristics of the distribution is represented by θ (such as the mean, variance, rate, etc.). We call θ the parameter(s) of the model.

Such an assumed distribution is called a **parametric model**.

Parametric models (cont.)

In the two examples given earlier,

- Assume the population of tyres to follow a $\mathrm{N}(\mu, \sigma^2)$ distribution. Here $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, where μ is the 'true' lifetime. Let X= lifetime of the tyre. Then $X \sim \mathrm{N}(\mu, \mathit{sigma}^2)$.

Parametric models (cont.)

In the two examples given earlier,

- Assume the population of tyres to follow a $\mathrm{N}(\mu, \sigma^2)$ distribution. Here $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, where μ is the 'true' lifetime. Let X= lifetime of the tyre. Then $X \sim \mathrm{N}(\mu, sigma^2)$.
- For the proportion of Belait district residents example, the population is a Bernoulli distribution $p = \mathrm{P}(\text{lives in Belait})$. Let X=someone in Brunei lives in Belait district. Then $X \sim \mathrm{Bern}(p)$.

A sample: a set of data or random variables?–A duality

A sample of size n $\{X_1, \ldots, X_n\}$ is also called a *random* sample. It consists of n concrete numbers in a practical problem.

A sample: a set of data or random variables?–A duality

A sample of size n $\{X_1, \ldots, X_n\}$ is also called a *random* sample. It consists of n concrete numbers in a practical problem.

The word 'random' captures the characteristic of the sample (of the same size) may be different, it it were taken by different people/entities, or at different times, or under different circumstances, etc. Essentially, they would be different subsets of a population.

A sample: a set of data or random variables?–A duality

A sample of size n $\{X_1, \ldots, X_n\}$ is also called a *random* sample. It consists of n concrete numbers in a practical problem.

The word 'random' captures the characteristic of the sample (of the same size) may be different, it it were taken by different people/entities, or at different times, or under different circumstances, etc. Essentially, they would be different subsets of a population.

Furthermore, a sample is also viewed as n independent and identically distributed (iid) random variables, when we assess the performance of a statistical method.

A sample: a set of data or random variables? (cont.)

For the tyre lifetime example, the sample of size $n = 120$ used gives the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = 56,956$$

A sample: a set of data or random variables? (cont.)

For the tyre lifetime example, the sample of size $n = 120$ used gives the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = 56,956$$

A different sample may well give a different sample mean, e.g. 57,062.

A sample: a set of data or random variables? (cont.)

For the tyre lifetime example, the sample of size $n = 120$ used gives the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = 56,956$$

A different sample may well give a different sample mean, e.g. 57,062.

QUESTION: Is the sample mean $\bar{X}_n$ a good estimator for the unknown 'true' lifetime μ? Obviously, we cannot use the concrete number 56,956 to assess how good this estimator is, as a different sample may give a different average value.

Key idea

By treating $X_1, \ldots, X_n$ as random variables, $\bar{X}_n$ is also a random variable, so has a distribution. If the distribution of $\bar{X}_n$ concentrates closely around the unknown μ, then it is a good estimator!

A statistic

Definition 1 (Statistic)

Any known function of a random sample is called a **statistic**. These are used for statistical inference such as estimation, testing and constructing confidence sets.

A statistic

Definition 1 (Statistic)

Any known function of a random sample is called a **statistic**. These are used for statistical inference such as estimation, testing and constructing confidence sets.

Example 2

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a sample from the population $\mathrm{N}(\mu, \sigma^2)$. Then,

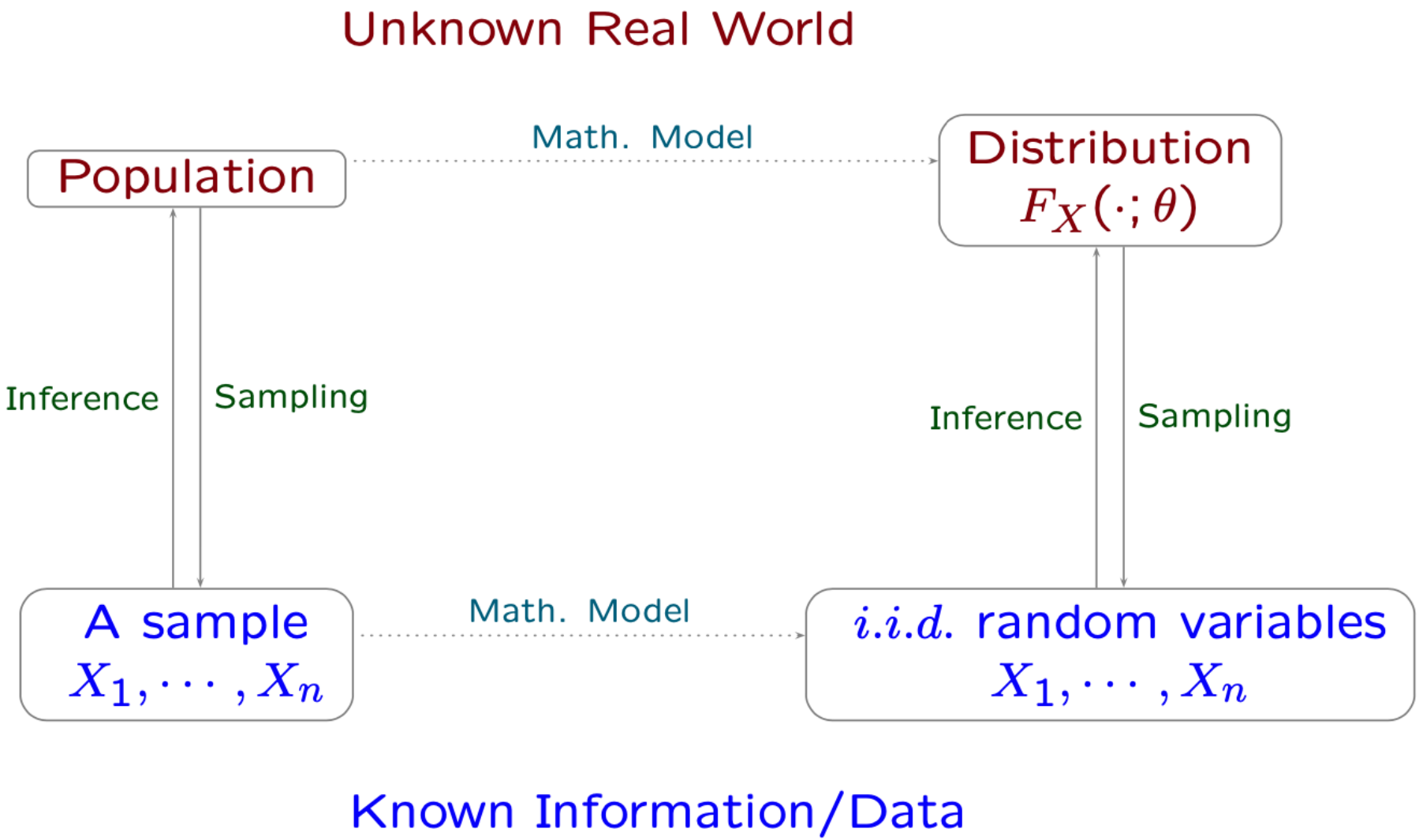
$$T_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad T_2(\mathbf{X}) = X_1 + X_n^2 \qquad T_3(\mathbf{X}) = \sin(X_3) + 6$$

are all statistics. But

$$T_4(\mathbf{X}) = \frac{X_1 - \mu}{\sigma}$$

is not a statistic, as it depends on unknown quantities μ and σ^2.

The statistical paradigm



Probability vs statistics

Probability is a mathematical subject.

Statistics is an applied subject which uses probability heavily.

Consider this problem: Let

$$X = \text{number of SM-4331 lectures attended by a student}$$

Then model $X \sim \text{Bin}(16, p)$.

Probability vs statistics (cont.)

Probability questions (treating p as known):

- What is $\mathrm{E}(X)$, the average lectures attended?
- What is $P(X \geq 8)$, the probability that students attend at least half of all lectures?
- What is $P(X = 16)/\big(1 - P(X = 16)\big)$, the odds that a student attends all their lectures?

Probability vs statistics (cont.)

Probability questions (treating p as known):

- What is $\mathrm{E}(X)$, the average lectures attended?
- What is $P(X \geq 8)$, the probability that students attend at least half of all lectures?
- What is $P(X = 16)/\big(1 - P(X = 16)\big)$, the odds that a student attends all their lectures?

Statistics questions :

- What is p, the average attendance rate?
- How confident am I that p lies in $(0.4, 0.6)$?
- Is p not smaller than 0.9?

Things you should know already

- Basic probability calculations and results (e.g. conditional probabilities, Bayes' law, independence, etc.) [SM-2205].
- Basic statistical calculations (e.g. sample mean, variance, covariance, correlation).
- Random variables.
- Common probability distributions, including Bernoulli, binomial, Poisson, uniform, normal, exponential.

Warm-up 1

- Ashley and Allison want to go to the movies to see either Batman or Superman. They are willing to separate and go to different movies. Each of them independently flips a fair, standard coin to decide which movie they will go to. What is the probability that they go to the same movie?
- Write down the formulae for the sample mean and variance.
- What does the term 'correlation' mean? What values can correlation take?
- Prove Bayes' law: for two events A and B in the sample space,

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

Warm-up 2

- Show that for any two random variables X and Y, and a constant $a \in \mathbb{R}$, $\mathrm{E}(X + aY) = \mathrm{E}(X) + a\,\mathrm{E}(Y)$.
- Must X and Y be independent for the result to hold?
- Show that for any r.v. $Y_1, \ldots, Y_n$ and constants $a_1, \ldots, a_n$,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i\,\mathrm{E}(Y_i)$$

Warm-up 3

List down the support, parameters, mean and variance for each of the following distributions: Bernoulli, binomial, Poisson, uniform, normal, exponential.