MATH1180 Computational Methods and Numerical Techniques Probability and Statistics<sup>1</sup>

Kayvan Nejabati Zenouz<sup>2</sup>

University of Greenwich

April 27, 2020

"Thus it is left to the reader to put it all together by himself, if he so pleases, but nothing is done for a reader's comfort"

STAGES ON LIFE'S WAY, SØREN KIERKEGAARD 1813 - 1855

 $<sup>^1</sup>$  Use these notes in conjunction with R demos accessible on http://rpubs.com/KayvanNejabati/565054.

<sup>&</sup>lt;sup>2</sup>Office: QM315, Email: K.NejabatiZenouz@greenwich.ac.uk, Student Drop-in Hours: MONDAYS 12:00-13:00 (MATHS ARCADE) AND TUESDAYS 15:00-16:00 (QM315)

# Contents

#### 1

3

#### Introduction

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R
- RStudio GUI
- Reading List and References

#### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

#### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution

5

The Normal Distribution

#### Topic 4: Joint Distribution of Bandom Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

#### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking

6

 Model Fitting with Least Square Method

#### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals
- 8 Topic 7: Introduction to Time Series
  - Times Series Data and Visualisation
  - Definition and Examples
  - Time Series
    Decomposition
  - Lags and Differences
  - Autocorrelation and Autocovariance
  - Stationary Time Series
  - Trend and Seasonality
  - Autoregressive Moving Average
  - Forecasting

# Lecture Contents

#### Introduction

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R
- BStudio GUI
- Reading List and References

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability
- Independent Events

- Bandom Variables
- Probability Mass

- Cumulative Distribution
- Probability Density
- Expected Value
- Variance and Standard

- The Bernoulli and
- The Geometric
- The Poisson Distribution
- The Exponential
- The Normal Distribution

- Joint Probability
- Conditional Probability
- Conditional Expectation:
- Independence
- Functions of Bandom
- Variance and Covariance

- Pearson's Correlation
- Regression Models
- Model Checking
- Model Fitting with Least

- The Law of Large
- The Central Limit
- Confidence Intervals

- Times Series Data and
- Definition and Examples

Time Series

- Lags and Differences
- Autocorrelation and
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving
- Forecasting

MATH1180

# Introduction

#### Aims

The aim of this part of the module is to provide an overview of probability and statistics skills essential for computer scientists. In particular, by the end of this part you will be able to...

- Use common statistical techniques, carry out statistical analysis, and reflect on results.
- Understand the formulation of simple regression and time series models and apply them to data and make forecasts.
- Use a statistical package to analyse data sets, interpret computer outputs of their analysis, and report on findings.

#### Assessment

- Assignment weight 25% due 03/03/2020.
- Closed Book Examination, weight 50%, combined with first term materials, May 2020.

#### Topics to be Covered...

- Review of Statistics and Probability
- **2** Introduction to Random Variables
- **③** Discrete and Continuous Random Variables
- **4** Joint Distribution of Random Variables
- **(**) Pearson's Correlation and Regression
- Approximations, Confidence Intervals
- **@** Introduction to Time Series

# Class Activity with www.menti.com

Please **scan** the barcode with your **phone** in order to take part in the class activity.



https://www.menti.com/dxp9d13ni1

Alternatively, go to <u>www.menti.com</u> on your electronic devices and enter the access code **94 56 92**.

Kayvan Nejabati Zenouz MATH1180

# Programming Package R

- The programming language R was developed around 1993 it is object orientated and open source.
- R has become a powerful tool used by **statistician** and **data scientists** used for
  - Data analysis and visualisation
  - Statistical modelling
  - HTML application development, automated web-browsing.

#### How to Get R and its GUI

- Go To <u>www.r-project.org</u> and to install R click on **download R** in the first paragraph.
- Select a server from the list. Download and install R for your operating system.
- Studio, Select a free version and install.
- The website <u>www.cran.r-project.org</u> offers useful **notes** for programming with R.

# RStudio GUI



# Suggested Reading List and References

For reading list see Akritas (2014); Shumway and Stoffer (2017); Devore (2011); Rice (2006).

Akritas, M.

2014. Probability & Statistics with R for Engineers and Scientists. Pearson.

Devore, J.

2011. Probability and Statistics for Engineering and the Sciences. Cengage Learning.

Rice, J.

2006. *Mathematical Statistics and Data Analysis*. Cengage Learning. Shumway, R. and D. Stoffer

2017. *Time Series Analysis and Its Applications: With R Examples*, Springer Texts in Statistics. Springer International Publishing.

#### **Guidance for Success**

Attend Lectures, Engage with Tutorials, Ask Questions, Read Books, Use Online Resources (Google, YouTube, etc...), Keep Your Work Organised, and Always Ask for Help.

# Topic 1 Review of Statistics and Probability



- Sample Space  $\mathcal{S}$
- Event E
- Probability of E

$$P(E) = \frac{|E|}{|\mathcal{S}|}$$

• Conditional probability

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

	Population	Sample
Location	$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$	$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
Scale	$\sigma^{2} = \frac{1}{N} \sum_{i=1}^{N} (x_{i} - \mu)^{2}$	$S^{2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i} - \overline{x})^{2}$

# Lecture Contents

#### Introdu

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R
- RStudio GUI
- Reading List and References

#### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

#### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

#### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking
- Model Fitting with Least Square Method

#### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals

#### Topic 7: Introduction to Time Series

- Times Series Data and Visualisation
- Definition and Examples

Time Series
 Decomposition

- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

#### MATH1180

The main objective of this session is to review elementary concepts in statistics and probability.

#### By the end of this session you will be able to...

- Understand the basic applications of statistics and probability.
- <sup>2</sup> Learn about population, samples, and sampling techniques.
- **③** Calculate populations and sample parameters.
- O Produce statistical visualisations.
- Analyse sample spaces and compute probabilities of events.
- Learn the rules governing probability of events, conditional probability, and independent events.

#### **Statistics**

Theory of **statistics** is concerned with collecting, processing, summarizing, analysing, and interpreting data.

- Height of students in a lecture room.
- The reason why we require statistics is variability.
- Earliest uses of statistical concepts go back to 450 BC.

### Probability

Theory of **probability** is concerned with likelihood of events.

- Throw a coin, what is the probability of obtain a head?
- Earliest uses of probability date back to Al-Khalil around 8<sup>th</sup>, and modern theory was developed by Cardano, Fermat, Pascal, and Laplace after 16<sup>th</sup>, century.
- Comprehending the theory of statistics requires a sound background in probability.

# Type of Data in Statistics

#### **Exercise 1: Research**

Find applications and uses of probability and statistics in science and industry.

We need to think about what **type of data** we are dealing with.

#### Quantitative Data

This is also refereed to as numerical data; it comes as **continuous** or **discrete**.

#### Qualitative Data

This is also referred to as categorical data; it comes as **binary**, **nominal**, or **ordinal**.

#### Example

- Quantitative: weight or number of people in a room.
- Qualitative: true/false, red/blue/yellow, or good/OK/bad.

- In many case we are interested in understanding certain characteristics of a **collection** of objects/subjects.
- The collection of all items of interest in known as the **population**.
- The variation in characteristic of interest among members of the same population in known as **inherent** or intrinsic **variability**.
- Population is studied through a **census**, the examination of all members.
- Can be **hypothetical** or conceptual, in the sense that not all members are available for examination.

# Methods of Cement Preparation

# Example

- Comparing two or more methods of cement preparation in terms of compressive strength.
- Not all cement prepared according to the same method had the same compressive strength.
- If the hardness differs among preparations of the same cement mixture, then what does it mean to compare the hardness of different cement mixtures?
- A more precise statement of the problem would be to compare the average (or mean) hardness of the different cement mixtures.
- We have two or more populations, one for each type of cement mixture, and the characteristic under investigation is compressive strength.
- Population is conceptual.

### Sample Surveys

Constraints on resources, time, money, and availability usually make a census impractical or infeasible, so **sample surveys** are used to obtain information about a large population by examining only a selected items from the population.

#### Census:

- Aims for complete coverage of the population.
- May use too much time/money/effort.
- Should provide large numbers even in minority groups.
- May be of value to other studies/research projects.

### Sample Survey:

- Saves money/time/effort.
- Can provide a good enough level of accuracy, but involves a margin of error.
- Involves an element of risk.
- May produce only small numbers in minority groups.

# Population and Sample Properties

- Population-level properties/attributes of characteristic(s) are called **population parameters**, e.g., mean, proportion, variance.
- The corresponding sample properties/attributes of characteristics are called **statistics**.
- A sample statistic provides an **estimate** for that of the population.
- New samples collected in the same way will produce new estimates.
- These estimates will **not**, in general, be **equal**, each being based on a different collection of values.
- This is called **sampling variability**.

#### Exercise 2: Research

Find examples of populations and a parameter of interest in them.

# Statistical Inference

- Statistical inference deals with the uncertainties that arise in extrapolating to the population the information contained in the sample.
- Takes the form of **estimation** (both point and interval estimation) of the population parameter(s) of interest.
- **Testing** various hypotheses for the value of the population parameter(s) of interest.
- Finally statistical inference may be used in the problem of **prediction**.
- Main **approaches** to statistical inference can be classified into
  - Parametric
  - Robust
  - Nonparametric
  - Bayesian

Sampling theory is concerned with selecting a sample form a population, so that the statistics produced approximate population parameters.

#### **Exercise 3: Research**

Find sampling techniques and discuss uses/examples.

### Includes

- Simple Random Sampling
- Stratified Sampling
- Cluster Sampling
- Systematics Sampling

# **Definition (Simple Random Sampling)**

A sample of size n is selected from a population and we ensure that each particular sample of size n has the same probability of occurrence.

- If the total population has size N, we may use a random number generator to select n samples.
- Sampling can be done with or without replacement.
- If **x** is a vector from which we need to take a sample of size *n*, without replacement, you can use R code

• The ratio 
$$f = \frac{n}{N}$$
 is called the **sampling fraction**.

#### Histograms

Dividing the range of the data into consecutive intervals, or bins, and constructing a box, or vertical bar, above each bin. The height of each box represents the bin's frequency, which is the number of observations that fall in the bin.

#### Scatterplots

Useful for exploring the relationship between two and three variables, e.g., relationship between height and weight. Also scatter matrix of scatterplots for all pairs of variables in a data set can be used in certain situations.

#### **Boxplots**

Boxplots are a standardized way of displaying the distribution of data based on a five number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

- We are often interested in **quantifiable** aspects of a population known as **parameters** of statistical population.
- Common parameters are **proportion**, **average**, and **variance**.
- Consider a finite population of size N from which we may select a sample of size n.
- As population have true parameters, which may or may not be known in general.
- Sample parameters provide an estimate for population parameters.

### Proportion

If a variable in the population of interest has categorical nature, e.g., types Defective or Non-Defective, we are may be interested in the proportion of items in each category.

• If  $N_i$  items are in category *i*, then the **Population Proportion** of category *i* is

$$p_i = \frac{N_i}{N}.$$

• If  $n_i$  items are in category i, then the **Sample Proportion** of category i is

$$\widehat{p}_i = \frac{n_i}{n}$$

We say  $\hat{p}_i$  is an estimate for  $p_i$ .

#### **Exercise 4: Proportions**

In a bag with 1000 sweets there are 606 blue, 284 green, and 110 red. Find the population proportions for each colour. In a simple random sample of size 100 we find there are 79 blue, 12 green, and 9 red. Find the sample proportion for each colour.

#### Average or Mean

Suppose we study a numerical property of a population and let  $x_1, x_2, ..., x_N$  denote the values for the property of interest for each item. Then the **population average** or mean is

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

If a sample of size n is selected with values  $x_1, x_2, ..., x_n$ , then

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

is the **sample mean**. We say  $\overline{x}$  is an estimate for  $\mu$ .

Mean or average value naturally shows us the "central" value of a set of values, it is a known as a **location parameter**.

# Population and Sample Variance

# Variance and standard deviation quantify the intrinsic variability of the values.

### Variance and Standard Deviation

Suppose we study a numerical property of a population and let  $x_1, x_2, ..., x_N$  denote the values for the property of interest for each item. Then the **population variance** is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2,$$

where  $\sigma$  is known as the **standard deviation**. If a sample of size *n* is selected with values  $x_1, x_2, ..., x_n$ , then

$$S^{2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i} - \overline{x})^{2}$$

is the (unbiased) **sample variance**. We say S is an estimate for  $\sigma$ .

The length of time between documents being send to a network printer has been measurement for a day in minutes with results below.

60, 120, 20, 30, 40, 200, 10, 20, 50

- Calculate the population mean and variance for the items above.
- Sample the population by taking items number 1, 3, 5, 7, 9, and calculate the sample mean and variance.
- What proportion of population has length of time less than 20?
- What proportion of the sample has length of time less than 20?

# Role of Probability

• **Statistics** uses sample-level information to infer/estimate properties of the population.



• **Probability Theory** assumes that all relevant information about the population is known and seeks to assess the chances that a sample will possess certain properties of interest.

# Probability Examples and Motivation

#### Example

- What are the chances of 4 heads, or of 10 heads, or of 18 heads in 20 flips of a fair coin? This can be rephrased in terms of a sample of size 20 taken with replacement from the population {Head, Tail}.
- If 5% of electrical components have a certain defect, what are the chances that a batch of 500 such components will contain less than 20 defective ones?

#### **Statistical Proof**

Statistical proof is used for proving a population has a certain property using a sample and works as follows.

- **()** Assume the population does not have the property.
- **2** Calculate the chances of obtaining the sample we observed.
- If the chances are small enough, say less than 5%, conclude that the assumption in (1) was wrong.

# Introduction to Probability

Probability quantifies **likelihood** of certain events, which are related to sampling experiments.

#### Example

Role a fair die twice and record the number each time.



- What is the set of all possible outcomes?
- How many of the outcomes include
  - sum of two numbers equalling 10?
  - 2 two sixes?
  - **3** sum of two numbers being 9 and one number 3?
  - um of two numbers being 9 or one number 3?

# **Definition (Sample Space)**

The set of all possible **outcomes** of an experiment is called the **sample space** of the experiment and will be denoted by S.

#### Example

- Toss a fair coin:  $S = \{H, T\}.$
- Role a die record number:  $S = \{1, 2, 3, 4, 5, 6\}.$
- A student's satisfaction about a module is recorded on a scale from 1 to 10:  $S = \{1, 2, 3, 4, ..., 10\}$ .

#### **Remark 1: Sample Space and Statistical Population**

The statistical population for a sampling experiment is the collection of all measurements, but the sample space is smaller since each of the possible outcomes is listed only once.

# Events

### **Definition** (Event)

A subset of a sample space of S for an experiments is called an **event**. If an outcome x is in A we write  $x \in A$ .

#### Example

Throw a die, the events that the number shown is even  $A = \{2, 4, 6\}.$ 

### **Exercise 6: Samples Spaces and Events**

- Write the sample space for the following.
  - Tossing a coin twice.
  - The length of time between successive earthquakes in a particular region in hours.
- In tossing a coin twice write the event that we find at least one head.
- The event that length of time between successive earthquakes in a particular region is longer than 2 hours.

# Set Operations for Events

Events are sets, so the usual set operations are relevant for probability theory. **Venn diagrams** are used to illustrate basic set operations.



34

# More on Events

• Two events A and B are called **disjoint** of **mutually exclusive** if they have no outcomes in common, we write

 $A \cap B = \emptyset.$ 

- We say an event A is a **subset** of another event B if all outcomes of A are also outcomes of B.
- The following events laws hold for events A, B, and C.
  Commutative Laws: A ∪ B = B ∪ A, A ∩ B = B ∩ A.
  Associative Laws:

 $(A\cup B)\cup C=A\cup (B\cup C), \ \ (A\cap B)\cap C=A\cap (B\cap C).$ 

**Distributive Laws:** 

 $(A\cup B)\cap C=(A\cap C)\cup (B\cap C), \ (A\cap B)\cup C=(A\cup C)\cap (B\cup C).$ 

De Morgan's Laws:

 $(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c.$ 

Exercise 7: Events

Draw Venn diagrams of the items above.

- We are interested in assessing the likelihood, or chance, of occurrence of an outcome or, more generally, of an event E from sample space S.
- The probability of E, denoted by  $P(E) \in [0, 1]$  is used to quantify the likelihood of occurrence of E.
- In general if  $N_n(E)$  denotes the number of occurrences of E is n repetitions of the experiment, then P(E) may be interpreted as  $\lim_{x \to \infty} \frac{N_n(E)}{n}$ .
# Equally Likely Outcomes

#### Definition (Probability of Equally Likely Outcomes)

If the sample space S of an experiment consists of N = |S|outcomes that are equally likely to occur, then the probability of each outcome is 1/N. If N(E) = |E| denotes the number of outcomes constituting the event E, then the probability of E is

$$P(E) = \frac{N(E)}{N}.$$

#### Example

- In rolling a die the probability of obtain number 1 is 1/6, and the probability of obtaining an odd number is 3/6.
- In tossing a coin twice the probability of obtaining at least one head is

$$P(\text{as least one head}) = \frac{|\{HT, TH, HH\}|}{|\{HT, TH, HH, TT\}|} = \frac{3}{4}.$$

Role a fair die twice and record the number each time. What is the probability of

- sum of two numbers equalling 10?
- **2** two sixes?
- **3** sum of two numbers being 9 and one number 3?
- **4** sum of two numbers being 9 or one number 3?

## **Application of Counting**

In determining the probability of an event E in many case we need use a counting procedure to determine the size of sample space S as well as size of E.

### Example

- How many outcomes there are for choosing a shirt and a coat out of 3 shirts and 2 coats?
- Two cards will be selected from a deck of 52 cards.
  - How many outcomes are there if the first card will be given to player 1 and the second card will be given to player 2?
  - How many outcomes are there if both cards will be given to player 1?

### **Stage Counting**

If a task can be completed in two stages, if stage 1 has  $n_1$  outcomes, and if stage 2 has  $n_2$  outcomes, regardless of the outcome in stage 1, then the task has  $n_1n_2$  outcomes.

### Example

Choosing a shirt and a coat out of 3 shirts and 2 coats we have  $2 \times 3 = 6$  outcomes.

### Remark 2: k Stage Counting

If a task can be completed in k stages and stage i has  $n_i$  outcomes, regardless of the outcomes of the previous stages, then the task has  $n_1n_2\cdots n_k$  outcomes.

### **Exercise 9: Counting**

**9** How many outcomes there are for throwing a die twice?

**2** How many outcomes to form a binary sequence of length 10?

# Sampling Without Replacements

If the k stages of a task involve sampling one unit each, without replacement, from the same group of n units, then the outcomes can be **ordered** if each stage is numbered, and **unordered** otherwise.

• In the ordered case the outcomes are called **permutation** of k units out of n, denoted by  $P_{k,n}$ , and given by

$$P_{k,n} = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

In R you can write factorial(n)/factorial(n-k).

• In the unordered case the outcomes are called **combination** of k units out of n, denoted by  $C_{k,n}$ , and given by

$$C_{k,n} = \binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots\times 2\times 1} = \frac{n!}{(n-k)!k!} = \frac{P_{k,n}}{P_{k,k}}$$

In R you can write choose(n, k).

### **Exercise 10: Combinations Probability**

In drawing three cards out of a set of 52 what is the probability of getting two aces and a king?

### **Definition (Axioms of Probability)**

For an experiment with sample space S, probability is a function that assigns a number P(E) to any event E such that

**Axiom 1:**  $0 \le P(E) \le 1$ 

**Axiom 2:** P(S) = 1

**Axiom 3:** For any sequence of disjoint events  $E_1, E_2, ...,$  we have

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

It follows from the axioms of probability that

• 
$$P(\emptyset) = 0$$
  
•  $P(A^c) = 1 - P(A)$   
• If  $A \subseteq B$ , then  $P(A) \le P(B)$   
•  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$   
•  $P(E) = \sum_{x \in E} P(\{x\})$ 

Exercise 11: Consequences of Axioms of Probability

Find a proof, using the axioms, for each of the items above.

# Conditional Probability

- Conditional probability is concerned with **probability** of an event **given** another event has already taken place.
- For example throw a dice, probability of getting 2 is 1/6. Now given we already know the number is even, probability of getting 2 is 1/3.
- For two events A, B, the **conditional probability** that B occurs given that A occurred and is denoted by  $P(B \mid A)$ .
- If A has already occurred, then our sample space is no longer S, but A, in this can we are concerned with  $A \cap B$ .



• Therefore we have

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

for  $P(A) \neq 0$ .

Kayvan Nejabati Zenouz MATH1180

# Total Law of Probability, Bayes' Theorem

Note using the conditional probability rule for two events A and B we can write

$$P(A \cap B) = P(B \mid A)P(A).$$

This leads to two useful consequences.

### **Total Law of Probability**

Suppose  $A_1, ..., A_k$  are disjoint sets whose union is the sample space S, and B any event. Then we can express P(B) as

$$P(B) = P(A_1)P(B \mid A_1) + \dots + P(A_k)P(B \mid A_k).$$

#### Theorem (Bayes)

Suppose  $A_1, ..., A_k$  are disjoint sets whose union is the sample space S, and B any event. Then for any j = 1, ..., k,

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B \mid A_j)}{\sum_{i=1}^k P(B \mid A_i)}.$$

# Independent Events

- In some cases for two events A and B, the knowledge that B has occurred does not effect the chance that A may occur that is  $P(B \mid A) = P(B)$ .
- For example if role a die twice, information about the first die has no effect on what the second would be.

#### **Definition (Independence)**

Two events A and B are called **independent** if

$$P(A \cap B) = P(A)P(B).$$

Otherwise the events are called **dependent**.

#### **Exercise 12: Combinations Probability**

It is known that 30% of a washing machines require service while under warranty, whereas only 10% of dryers need such service. If someone purchases both a washer and a dryer, what is the probability that both machines will need service?



# Topic 2 Introduction to Random Variables

Random Variable  $X : \mathcal{S} \longrightarrow \mathbb{R}$ .

Probability Mass Function  $p_X(x) = P(X = x)$ . Probability Density Function  $f_X(x)$ 

$$P(a \le X \le b) = \int_{a}^{b} f_X(x) dx.$$

Cumulative Distribution Function  $F_X(x) = P(X \le x)$ . Expected value

$$E(X) = \mu_X = \sum_{x \in S_X} x p_X(x), \quad E(X) = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Variance

$$\sigma_X^2 = \sum_{x \in S_X} (x - \mu_X)^2 p_X(x), \quad \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

# Lecture Contents

1 In

#### Introduction

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R
- RStudio GUI
- Reading List and References

#### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

#### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

#### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking
- Model Fitting with Least Square Method

#### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals

#### Topic 7: Introduction to Time Series

- Times Series Data and Visualisation
- Definition and Examples
- Time Series
   Decomposition
- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

MATH1180

### By the end of this session you will be able to...

- **0** Understand different types of random variables.
- Prind Probability Mass Functions and Cumulative Distribution Functions and Probability Density Functions for random variables.
- 3 Analyse properties of PMF, CDF, PDF.
- **0** Calculate expectation and variance of random variables.

# Introduction

- Last time we looked at populations, samples, and their characteristics.
- Characteristic of interest can be **quantitative**, or **qualitative**, though both can be represented by numbers.
- A characteristic of any type represented by a number is called a variable.
- Categorical variables are a particular kind of **discrete** variables.
- Quantitative variables expressing measurements on a continuous scale, such as length, are examples of **continuous** variables.
- When a population unit is randomly sampled from a population, its value is **not known** a priori.
- A random variable, X, denotes the value of the variable of a population unit that will be sampled.

### Example

• A coin in thrown three times, the sample space is

 $\mathcal{S} = \{HHH, HHT, HTT, HTH, TTT, TTH, THH, THT\}.$ 

One random variables X defined on  $\mathcal{S}$  is the total number of heads, so for example

$$X(HHH) = 3, X(HHT) = 2, ..., X(TTT) = 0.$$

Q Consider the experiment where an electrical product is tested until failure, and let X denote the time to failure. The sample space of this experiment, or of X, is S = [0,∞).

## **Definition** (Random Variable)

A random variable is a rule that assigns to each element of S a real number in  $\mathbb{R}$ . More formally, a random variable is a function

$$X:\mathcal{S}\longrightarrow\mathbb{R}.$$

If X can take on only a finite, or at most a countably infinite number of values, then X is known as a **discrete** random variable, otherwise X is a **continuous** random variable.

#### Example

In a sampling experiment where observations  $X_1, X_2, ..., X_n$  are collected from a population, the observation  $X_1, X_2, ..., X_n$ , together with sample mean  $\overline{X}$ , the sample variance  $S^2$ , and a sample proportion  $\hat{p}$  are random variables.

# Probability Mass Functions PMF

- Given a random variable X, then we may say what would be the probability of X taking a value is a subset  $A \subseteq \mathbb{R}$ .
- We write that as

$$P(X \in A) = P\left(\{a \in \mathcal{S} \mid X(a) \in A\}\right).$$

• For example, if X is number of heads on throwing a coin three times, then

$$P(X = 3) = \frac{1}{8}$$
 and  $P(X \in \{1, 2\}) = \frac{6}{8}$ .

#### **Definition (Probability Mass Function)**

Probability mass function for a discrete random variable X is the probability  $p_X(x) = P(X = x)$ .

### Definition (Cumulative Distribution Function)

The cumulative distribution function, of a random variable X gives the probability of events of the form  $X \leq x$ , for  $x \in \mathbb{R}$ , which is written as

$$F_X(x) = P(X \le x).$$

## Example

A coin in thrown three times. Let be X be the total number of heads. Find

- Probability mass function  $p_X(x)$ .
- **2** Cumulative distribution function  $F_X(x)$ .

Note X is a discrete random variable and the sample space is

 $\mathcal{S} = \{HHH, HHT, HTT, HTH, TTT, TTH, THH, THT\}.$ 

For the **probability mass function**  $p_X(x)$ . For the values of X which 0, 1, 2, 3 and we have

$$p_X(0) = P(X = 0) = \frac{|\{TTT\}|}{|S|} = \frac{1}{8}$$

$$p_X(1) = P(X = 1) = \frac{|\{HTT, THT, TTH\}|}{|S|} = \frac{3}{8}$$

$$p_X(2) = P(X = 2) = \frac{|\{HHT, HTH, THH\}|}{|S|} = \frac{3}{8}$$

$$p_X(3) = P(X = 3) = \frac{|\{HHH\}|}{|S|} = \frac{1}{8},$$

$$p_X(x) = P(X = x) = 0 \text{ if } x \neq 0, 1, 2, 3.$$

# Example Cont II

For the **cumulative distribution function**  $F_X(x)$ , we have

$$\begin{split} F_X(x) &= P(X \le x) = 0 \text{ if } x < 0. \\ F_X(0) &= P(X \le 0) = \frac{|\{TTT\}|}{|\mathcal{S}|} = \frac{1}{8} \\ F_X(1) &= P(X \le 1) = \frac{|\{TTT, HTT, THT, TTH\}|}{|\mathcal{S}|} = \frac{4}{8} \\ F_X(2) &= P(X \le 2) = \\ \frac{|\{TTT, HTT, THT, TTH, HHT, HTH, HTH\}|}{|\mathcal{S}|} = \frac{7}{8} \\ F_X(3) &= P(X \le 3) = \\ \frac{|\{HHH, TTT, HTT, THT, TTH, HHT, HHT, HTH, THH\}|}{|\mathcal{S}|} = \frac{8}{8}, \\ F_X(x) &= P(X \le x) = 1 \text{ if } x \ge 3. \end{split}$$

### We end up with the tables

and

The cumulative distribution function F(x) of a random variable X satisfies the following properties.

• It is non-decreasing: If  $a \leq b$  then  $F(a) \leq F(b)$ .

**2** We have 
$$F(-\infty) = 0$$
 and  $F(\infty) = 1$ .

**3** If 
$$a < b$$
, then  $P(a < X \le b) = F(b) - F(a)$ .

Exercise 1: Cumulative distribution function

Find a proof for each of the items above.

# Connection between the PMF and the CDF

Let  $x_1, x_2, ...$  denote the possible values of the discrete random variable X arranged in increasing order. Then

- The cumulative distribution function F is a step function, with jumps occurring at values  $x \in S$ . The size of each jump at x is p(x) = P(X = x).
- **2** The CDF can be obtained by

$$F(x) = \sum_{x_i \le x} p(x_i).$$

**③** The PMF can be obtained from CDF as

$$p(x_1) = F(x_1)$$
, and  $p(x_i) = F(x_i) - F(x_{i-1})$  for  $i = 2, 3, ...$ 

We have

$$P(a < X \le b) = \sum_{a < x_i \le b} p(x_i),$$

and in terms of CDF we have

$$P(a < X \le b) = F(b) - F(a).$$

Kayvan Nejabati Zenouz MATH1180

# Random Variables: More Examples

### **Definition (Bernoulli Distribution)**

A Bernoulli random variable X takes on only two values: 0 and 1, with probabilities 1 - p and p, respectively. Therefore, P(X = 0) = 1 - p and P(X = 1) = p and we write  $X \sim \text{Ber}(p)$ 

## **Definition** (Uniform Distribution)

A uniform random variable on the interval [0, 1] choosing a number at random between 0 and 1 and we have

P(X in the interval of length l) = l.

We write  $X \sim \text{Uni}(0, 1)$  and the CDF is given by

$$F_X(x) = P(X \le x) = \begin{cases} 0 & x < 0\\ x & 0 \le x \le 1\\ 1 & x > 1 \end{cases}$$

# Probability Density Function PDF

- A continuous random variable **cannot** have a PMF, since P(X = x) = 0 for all x.
- In addition to the CDF, the probability distribution of a **continuous** random variable can be described in terms of its probability density function.

### **Definition** (Probability Density Function)

The probability density function of a continuous random variable X is a nonnegative function  $f_X$ , thus,  $f_X(x) \ge 0$ , for all x, with the property that P(a < X < b) equals the area under it and above the interval [a, b]. Thus

$$P(a < X < b) = \int_{a}^{b} f_X(x) dx.$$

• Note the definition above implies that

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x) dx = 1.$$
Kavvan Neiabati Zenouz
MATH1180

Kavvan Nejabati Zenouz

## Example and Properties

If  $X \sim U(0, 1)$ , then we have

$$f_X(x) = \begin{cases} 0 & x < 0\\ 1 & 0 \le x \le 1\\ 0 & x > 1 \end{cases}$$

#### **Properties PDF**

If X is a continuous random variable with PDF f, the we have P(a < X < b) = P(a \le X \le b) = F(b) - F(a). CDF for X can be obtained by  $F(x) = \int_{-\infty}^{x} f(y) dy.$ 

**9** PDF can be obtained by

$$f(x) = F'(x) = \frac{d}{dx}F(X).$$

Kayvan Nejabati Zenouz MATH1180

# Example

If the life time T, measured in hours, of a randomly selected electrical component has PDF

$$f_T(t) = \begin{cases} 0 & t < 0\\ 0.005e^{-0.005t} & t \ge 0 \end{cases}$$

- Find the probability that the component will last between 300 and 600 hours.
- **2** Find the CDF of F(t).



# Solution

• The probability is

$$P(300 \le t \le 600) = \int_{300}^{600} 0.005e^{-0.005t} dt$$
$$= \left[-e^{-0.005t}\right]_{300}^{600} = -e^{-0.005 \times 600} + e^{-0.005 \times 300} = 0.1733.$$

**2** We have F(t) = 0 for  $t \leq 0$ , and

## **Exercise 2:** CDF for Random Variables

• The PMF of a random variable X is

Find the CDF of  $F_X(x)$ .

**2** A random variable Y is said to have the uniform in [A, B] distribution, denoted by  $Y \sim U(A, B)$ , if its PDF is

$$f_Y(y) = \begin{cases} 0 & y < A \\ \frac{1}{B - A} & A \le y \le B \\ 0 & y > B \end{cases}$$

Find the CDF of  $F_Y(y)$ .

# Parameters of Distributions

- The graph of a PMF or PDF for a random variable X tell us about the **distribution** of X
- For example, if  $X \sim \text{Uni}(0, 1)$ , then we have



- Parameters of random variables are used to create **summaries** of these distributions.
- The parameters we will consider are the mean value, or **expected value**, the **variance**, and **standard deviation**.

# Expected Value I

### **Definition (Discrete Random Variable)**

Let X be a discrete random variable with sample space  $S_X$ , and let p(x) = P(X = x) denote its probability mass function. Then, the **expected value** E(X) or  $\mu_X$  is defined as

$$E(X) = \mu_X = \sum_{x \in S_X} xp(x).$$

This E(X) indicates the values near to which we are "most likely to obtain".

#### Example

Toss a coin twice. Let X be the number of heads. Then we have

1.

## Exercise and Example

### **Exercise 3: Expectation**

Find E(X), where the sample space of X and  $p_X(x)$  is given by

### Example

Suppose X is obtained by simple random sampling from any finite population. Let  $\nu_1, ..., \nu_N$  be the values for underlying statistical population, and  $\mathcal{S}_X = \{x_1, ..., x_m\}$  be the sample space. Suppose  $n_j$  is the number of times  $x_j$  is repeated in the population, so  $P(x = x_j) = \frac{n_j}{N}$ . The we have  $\mu_X = \frac{1}{N} \sum_{i=1}^N \nu_i = \sum_{j=1}^m x_j P(X = x_j).$ 

# Expected Value II

### **Definition (Continuous Random Variable)**

Let X be a continuous random variable with sample space  $S_X$ , and let f(x) denote the probability density. Then, the **expected** value E(X) or  $\mu_X$  is defined as

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx.$$

#### Example

If the PDf of X is given by f(x) = 2x for  $0 \le x \le 1$  and 0 otherwise. Then we have

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{1} x \times 2xdx = \frac{2}{3}.$$

**Exercise 4: Uniform Distribution** 

Find E(X) for  $X \sim \text{Uni}(0, 1)$ .

# Expectation of Functions

Sometimes a random variable is a function of another: say X is a random variable and Y = h(X) for some function h.

### **Expectation of Functions**

Let X be a random variable, h a function on  $S_X$ , and Y = h(X).

• If X is a discrete random variable we have

$$E(Y) = E(h(X)) = \sum_{x \in S_X} h(x) p_X(x).$$

• If X is a continuous random variable we have

$$E(Y) = E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x).$$

• If Y = h(X) = aX + b, for some constants a, b, then

$$E(Y) = aE(X) + b.$$
# Example

A book store purchases three copies of a book at £6.00 each and sells them for £12.00 each. Unsold copies are returned for £2.00 each. Let  $X = \{$ number of copies sold $\}$  and  $Y = \{$ net revenue $\}$ . If the PMF of X is

Find E(X) and E(Y). Solution: For expected value of X we have

$$E(X) = 0 \times 0.1 + 1 \times 0.2 + 2 \times 0.2 + 3 \times 0.5 = 2.1.$$

Now Y = 12X + 2(3 - X) - 18 = 10X - 12, so

$$E(Y) = 10E(X) - 12 = 10 \times 2.1 - 12 = 9.$$

# Variance and Standard Deviation

## Variance

Given a random variable X, the variance of X is defined as

$$\operatorname{Var}(X) = \sigma_X^2 = E\left[ (X - \mu_X)^2 \right].$$

 $\sigma_X$  is known **standard deviation**. In particular if X is discrete, then we have

$$\sigma_X^2 = \sum_{x \in S_X} \left( x - \mu_X \right)^2 p(x)$$

and if X is continuous we have

$$\sigma_X^2 = \int_{-\infty}^{\infty} \left(x - \mu_X\right)^2 f(x) dx.$$

The shortcut formula for varianvce of X is

$$E\left[\left(X-\mu_X\right)^2\right] = E\left(X^2\right) - \mu_X^2.$$

# Example and Exercises

## Example

Toss a coin twice. Let X be the number of heads. Then we have

Recall we had E(X) = 1, now we have

$$Var(X) = (0-1)^2 \times 0.25 + (1-1)^2 \times 0.5 + (2-1)^2 \times 0.25 = 0.5.$$

### **Exercise 5: Variance**

• Find Var(X), where the sample space of X and  $p_X(x)$  is given by

**2** Find Var(X) for  $X \sim Uni(0, 1)$ .

## Variance of Linear Transformation

Let X be a random variable and Y = h(X) = aX + b, for some constants a, b, then

$$\operatorname{Var}(Y) = a^2 \operatorname{Var}(X).$$

# Example

A book store purchases three copies of a book at £6.00 each and sells them for £12.00 each. Unsold copies are returned for £2.00 each. Let  $X = \{$ number of copies sold $\}$  and  $Y = \{$ net revenue $\}$ . If the PMF of X is

Find Var(X) and Var(Y). Solution: For variance of X we have

$$Var(X) = 0 \times 0.1 + 1^{2} \times 0.2 + 2^{2} \times 0.2 + 3^{2} \times 0.5 - 2.1^{2} = 5.5 - 4.41 = 1.09.$$
  
Now  $Y = 12X + 2(3 - X) - 18 = 10X - 12$ , so  
 $Var(Y) = 100Var(X) = 100 \times 1.09 = 109.$ 

Let  $Y \sim \text{Uni}(A, B)$ . Find E(Y) and Var(Y) using the fact that Y = (B - A)X + A for  $X \sim \text{Uni}(0, 1)$ .



# Topic 3 Discrete and Continuous Random Variables



# Lecture Contents

### Intro

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R
- RStudio GUI
- Reading List and References

### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking
- Model Fitting with Least Square Method

### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals

#### Topic 7: Introduction to Time Series

- Times Series Data and Visualisation
- Definition and Examples
- Time Series
  Decomposition
- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

## By the end of this session you will be able to...

- Understand different types random variable including discrete and continuous.
- **2** Learn about their uses in real life examples.

# **Definition (The Bernoulli Distribution)**

A Bernoulli trial or experiment is one whose outcome can be classified as either a success or a failure. The Bernoulli random variable X takes the value 1 if the outcome is a success, and the value 0 if it is a failure.

## Example

- The prototypical Bernoulli experiment is a flip of a coin, with heads and tails being success and failure, respectively.
- In an experiment where a product is selected from the production line, the Bernoulli random variable X takes the value 1 or 0 as the product is defective (success) or not (failure)

# The Bernoulli Distribution PMF, CDF, and Parameters

• If the probability of success is p and that of failure is 1 - p, the PMF and CDF of X are

$$\begin{array}{c|ccc} x & 0 & 1 \\ \hline p(x) & 1-p & p \\ \hline F(x) & 1-p & 1 \\ \end{array}$$

• The parameters of the distribution are given by

$$\mu_X = p, \ \ \sigma_X^2 = (1-p)p.$$

### **Exercise 1: Bernoulli Distribution**

The probability that an electronic product will last more than 5500 time units is 0.1. Let X take the value 1 if a randomly selected product lasts more than 5500 time units and the value 0 otherwise. Find the mean value and variance of X.

## **Definition (The Binomial Distribution)**

Suppose n Bernoulli experiments, each having probability of success equal to p, are performed independently. Taken together, the n independent Bernoulli experiments constitute a binomial experiment. The binomial random variable Y is the total number of successes in the n Bernoulli trials.

## Example

The prototypical binomial experiment consists of n flips of a coin, with the binomial random variable Y being the total number of heads.

# The Binomial Distribution PMF and Parameters

• If  $X_i$  denotes the Bernoulli random variable associated with the *i*th Bernoulli trial, that is  $X_i = 1$  if *i*th experiment results in success and 0 otherwise, then the binomial random variable Y equals

$$Y = \sum_{i=1}^{n} X_i.$$

• The sample space of Y is  $S = \{0, 1, ..., n\}$ . The probability distribution is controlled the number of trials n and the probability of success in each trials, we write

$$Y \sim \operatorname{Bin}(n, p).$$

• The probability mass function is given by

$$P(Y = y) = \binom{n}{y} p^{y} (1 - p)^{n - y} \ y = 0, 1, ..., n.$$

• We have E(Y) = np and Var(Y) = n(1-p)p.

In a medical trial the rate of failure of a particular experiment is know to be 30%. Let Y denote the number of success in 15 experiments. Find the following.

- The probability that we have exactly 5 successes.
- **2** The probability  $P(2 \le Y \le 4)$ .
- $\bullet$  E(X) and Var(X).

# **Definition (The Geometric Distribution)**

A **geometric** experiment is one where independent Bernoulli trials, each with the same probability p of success, are performed until the occurrence of the first success. The Geo(p) random variable X is the total number of trials up to and including the first success in such a geometric experiment.

## Example

The prototypical engineering application of the geometric distribution is that of quality control at the production level: Product items are being inspected as they come off the production line until the first one with a certain defect is found. The geometric random variable X is the total number of items inspected.

# The Geometric Distribution PMF and Parameters

- The sample space of a geometric random variable X is  $S_X = \{1, 2, 3, ...\}.$
- The PMF P(X = x) gives the probability when x products tested, the first x 1 failed and last one succeeded, hence

$$P(X = x) = (1 - p)^{x - 1}p$$
, for  $x = 1, 2, 3, ...$ 

• The CDF is given by

$$F(x) = 1 - (1 - p)^x$$
, for  $x = 1, 2, 3, ...$ 

• The expectation and variance is given by

$$E(X) = \frac{1}{p}, \text{ Var}(X) = \frac{1-p}{p^2}.$$

In a production line the rate of failure of a particular product is know to be 30%. Let Y denote the number of product tested until first success. Find the following.

- The probability that fifth test is a success.
- **2** The probability  $P(2 \le Y \le 4)$ .
- $\bullet$  E(X) and Var(X).

# Definition (The Poisson Distribution)

The Poisson distribution is used to model the probability that a number of certain events occur in a specified period of time. The type of events whose occurrences are thus modelled must occur at random and at a rate that does not change with time. The random variable X is said to be Poisson with parameter  $\lambda$ , denoted by  $X \sim \text{Poi}(\lambda)$ .

## Example

The prototypical engineering application of the geometric distribution is the number of cars arrived at the car park. The Poisson distribution can also be used for the number occurrences of events occurring in other specified intervals such as distance, area, or volume.

# The Poisson Distribution PMF and Parameters

- The sample space of a Poisson random variable X is  $S_X = \{0, 1, 2, 3, ...\}.$
- The PMF P(X = x) is given by

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$$
, for  $x = 0, 2, 3, ...$ 

• The expectation and variance is given by

$$E(X) = \lambda, \ \operatorname{Var}(X) = \lambda.$$

### Example

Let  $X \sim \text{Poi}(4)$ . Find P(X = 2) and  $P(1 \le X \le 3)$ .

A binomial experiment where the number of trials n is large  $(n \ge 100)$ , the probability p of success in each trial is small  $(p \le 0.01)$ , and the product np is not large  $(np \le 20)$ , can be modelled (to a good approximation) by a Poisson distribution with  $\lambda = np$ .

### Example

Due to a serious defect, a car manufacturer issues a recall of n = 10,000 cars. Let p = 0.0005 be the probability that a car has the defect, and let Y be the number of defective cars. Find  $P(Y \ge 10)$  (b) P(Y = 0).

# **Definition (The Poisson Distribution)**

A random variable X is said to be an exponential, or to have the exponential distribution with parameter  $\lambda$ , denoted by  $X \sim \text{Exp}(\lambda)$ , if its PDF is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0\\ 0 & x < 0 \end{cases}$$

We have  $F(x) = 1 - e^{-\lambda x}$  for  $x \ge 0$  and F(x) = 0 otherwise. Also  $E(X) = \frac{1}{\lambda}$  and  $\operatorname{Var}(X) = \frac{1}{\lambda^2}$ .

### Example

The exponential distribution is used in reliability theory as the simplest model for the life time of equipment.

Suppose the useful life time, in years, of a personal computer (PC) is exponentially distributed with parameter  $\lambda = 0.25$ . A student entering a four-year undergraduate program inherits a two-year-old PC from his sister who just graduated. Find the probability the useful life time of the PC the student inherited will last at least until the student graduates.

# The Normal Distribution

## **Definition** (The Normal Distribution)

A random variable is said to have the **standard normal** distribution if its PDF and CDF, which are denoted (universally) by  $\phi$  and  $\Phi$ , respectively, are

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$
 and  $\Phi(z) = \int_{-\infty}^{z} \phi(x) dx$ 

for  $-\infty < z < \infty$ .

A random variable X is said to have the normal distribution, with parameters  $\mu$  and  $\sigma$ , denoted by  $X \sim N(\mu, \sigma^2)$ , if its PDF and CDF are

$$f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$
 and  $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ .

for  $-\infty < x < \infty$ . We have  $E(X) = \mu$  and  $Var(X) = \sigma^2$ .



# Topic 4 Joint Distribution of Random Variables

# Lecture Contents

### l Int

### Introduction

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R
- RStudio GUI
- Reading List and References

### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking
- Model Fitting with Least Square Method

### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals

#### Topic 7: Introduction to Time Series

- Times Series Data and Visualisation
- Definition and Examples

Time Series
 Decomposition

- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

## By the end of this session you will be able to...

- Understand joint random variables.
- **2** Calculate marginal distributions.
- **③** Calculate conditional probabilities and expectations.
- **4** Learn about independent random variables.
- Calculate expectations and variance of functions of random variables.
- Calculate covariance of random variables.

- In some experiments we record **multivariate** observations.
- For example, studies of atmospheric turbulence we may quantifying the degree of relationship between the components X, Y, and Z of wind velocity.
- Studies of car safety may focus on the relationship between the velocity X and stopping distance Y under different road and weather conditions.
- To study these we need the concept of the **joint distribution** of the random variables.
- This will allow use to study **correlation** and **regression**.

## **Definition (Joint PMF)**

The joint, or bivariate, probability mass function (PMF) of the jointly discrete random variables X and Y is defined as

$$p(x,y) = P(X = x, Y = y).$$

## **Remark 1: Properties of Joint Distributions**

• If  $S = \{(x_1, y_1), (x_2, y_2), ...\}$  is the sample space of (X, Y), then axioms of probability imply that

$$p(x_i, y_i) \ge 0$$
 for all  $i$  and  $\sum_{(x_i, y_i) \in S} p(x_i, y_i) = 1.$ 

• Furthermore,

$$P(a < X \le b, \ c < Y \le d) = \sum_{a < x_i \le b, \ c < y_i \le d} p(x_i, y_i).$$

# Marginal Distributions PMF

# **Definition (Marginal Distributions PMF)**

The marginal PMFs of X and Y are obtained as

$$P_X(x) = \sum_{y \in \mathcal{S}_Y} p(x, y) \text{ and } P_Y(y) = \sum_{x \in \mathcal{S}_X} p(x, y).$$

## Example

Let X, Y have the joint PMF as shown in the following table.

			y
	p(x,y)	1	2
	1	0.3	0.13
x	2	0.06	0.26
	3	0.1	0.15

Find the marginal PMF of Y and

 $P(0.5 < X \le 2.5, 1.5 < Y \le 2.5)$  and  $P(0.5 < X \le 2.5)$ .

# Example Cont I

For marginal PMF of Y we have

$$p_Y(1) = p(1,1) + p(2,1) + p(3,1) = 0.3 + 0.06 + 0.1 = 0.46$$
  
$$p_Y(2) = p(1,2) + p(2,2) + p(3,2) = 0.13 + 0.26 + 0.15 = 0.54.$$

For

$$P(0.5 < X \le 2.5, \ 1.5 < Y \le 2.5) = p(1,2) + p(2,2)$$
  
= 0.13 + 0.26 = 0.39.

And for

$$P(0.5 < X \le 2.5) = p(1,1) + p(1,2) + p(2,1) + p(2,2)$$
  
= 0.3 + 0.06 + 0.13 + 0.26 = 0.75.

Calculating marginal distribution for both X and Y, we can create a table

			y	
	p(x,y)	1	2	$p_X(x)$
	1	0.3	0.13	0.43
x	2	0.06	0.26	0.32
	3	0.1	0.15	0.25
	$p_Y(y)$	0.46	0.54	

## **Remark 2: Multivariate PMF Distributions**

If  $X_1, X_2, ..., X_n$  are jointly discrete, their joint or multivariate PMF is defined as

$$p(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

## **Definition (Joint PDF)**

The joint, or bivariate, density function of the jointly continuous random variables X and Y is a nonnegative function f(x, y) with the property that the probability that (X, Y) will take a value in a region A of the x - y plane equals the volume under the surface defined by f(x, y) and above the region A.

## **Remark 3: Properties of Joint Distributions**

• The axioms of probability imply that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$$

• Furthermore,

$$P(a \le X \le b, \ c \le Y \le d) = \int_c^d \int_a^b f(x, y) dx dy.$$

# Marginal Distributions PDF

# **Definition (Marginal Distributions PMF)**

The marginal PDFs of X and Y are obtained as

$$p_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$
 and  $p_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$ .

## Example

Let X, Y have the joint density function

$$f(x,y) = \begin{cases} \frac{12}{7} (x^2 + xy) & 0 \le x, y \le 1\\ 0 & \text{otherwise.} \end{cases}$$

Find the following.

• 
$$P(X \le 0.6, Y \le 0.4)$$

**2** Marginal PDF of 
$$X$$
 and  $Y$ .

# Example Cont

**9** We have

$$P(X \le 0.6, Y \le 0.4) = \int_0^{0.4} \int_0^{0.6} \frac{12}{7} (x^2 + xy) \, dx \, dy$$
$$= \int_0^{0.4} \left[ \frac{12}{7} \left( \frac{x^3}{3} + \frac{x^2 y}{2} \right) \right]_{x=0}^{x=0.6} \, dy$$
$$= \int_0^{0.4} \frac{12}{7} \left( \frac{0.6^3}{3} + \frac{0.6^2 y}{2} \right) \, dy$$
$$= \frac{12}{7} \left[ \frac{0.6^3}{3} y + \frac{0.6^2 y^2}{4} \right]_0^{0.4} = 0.0741$$

**2** We have

$$f_X(x) = \int_0^1 \frac{12}{7} \left(x^2 + xy\right) dy = \frac{6}{7} \left(2x^2 + x\right),$$
  
$$f_Y(y) = \int_0^1 \frac{12}{7} \left(x^2 + xy\right) dx = \frac{2}{7} \left(2 + 3y\right).$$
# **Conditional Probability Functions**

### **Remark 4: Multivariate PDF Distributions**

If  $X_1, X_2, ..., X_n$  are jointly discrete, their joint or multivariate PDF can be defined as  $f(x_1, x_2, ..., x_n)$ , a multivariate function.

### Definition (Conditional PMF)

Conditional PMF of Y given X = x is given by

$$p_{Y|X=x}(y) = rac{p(x,y)}{p_X(x)}, \ y \in \mathcal{S}_Y.$$

For the example on slide 47 we have

	I	y	1
$p(x,y)/p_X(x)$	1	2	$p_X(x)$
$p_{Y X=1}(y)$	$\frac{0.3}{0.43}$	$\frac{0.13}{0.43}$	0.43
$p_{Y X=2}(y)$	$\frac{0.06}{0.32}$	$\frac{0.26}{0.32}$	0.32
$p_{Y X=3}(y)$	$\frac{0.1}{0.25}$	$\frac{0.15}{0.25}$	0.25

Kayvan Nejabati Zenouz MATH1180

# **Conditional Probability Functions**

### **Definition (Conditional PDF)**

Conditional PDF of Y given X = x is given by

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)}.$$

For the example on slide 50 we have

$$f_{Y|X=x}(y) = \frac{2(x^2 + xy)}{(2x^2 + x)}$$
 for  $0 \le y \le 1$ .

And

$$f_{X|Y=y}(x) = \frac{6(x^2 + xy)}{(2+3y)}$$
 for  $0 \le x \le 1$ .

# Exercise 1: Conditional ProbabilityFind $P(Y < 0.4 \mid X = 1).$

Kayvan Nejabati Zenouz MATH1180

### Conditional Expectation: Regression Function

The conditional expected value of Y given that X = x,

$$\mu_{Y|X}(x) = E(Y \mid X = x)$$

when considered as a function of x, is called the **regression** function of Y on X.

**Definition** (Conditional Expectation, Discrete)

Regression function for jointly discrete (X, Y) is given by

$$\mu_{Y|X}(x) = \sum_{y \in \mathcal{S}_Y} y p_{Y|X=x}(y), \ x \in \mathcal{S}_X.$$

For the example on slide 47 we have

	I	y	
$p(x,y)/p_X(x)$	1	2	$E(Y \mid X = x)$
$p_{Y X=1}(y)$	$\frac{0.3}{0.43}$	$\frac{0.13}{0.43}$	$\frac{0.3}{0.43} + 2 \times \frac{0.13}{0.43}$
$p_{Y X=2}(y)$	$\frac{0.06}{0.32}$	$\frac{0.26}{0.32}$	$\frac{0.06}{0.32} + 2 \times \frac{0.26}{0.32}$
$p_{Y X=3}(y)$	$\frac{0.1}{0.25}$	$\frac{0.15}{0.25}$	$\frac{0.1}{0.25} + 2 \times \frac{0.15}{0.25}$
Kayyan Nejabati	Zenouz	MATH1180	

111

### **Definition (Conditional Expectation Continuous)**

Regression function for jointly continuous (X, Y) is given by

$$\mu_{Y|X}(x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy, \ x \in \mathcal{S}_X.$$

For the example on slide 50 we have

$$\mu_{Y|X}(x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy = \int_{0}^{1} y \frac{2(x^{2} + xy)}{(2x^{2} + x)} dy \text{ for } 0 \le x \le 1.$$

And

$$\mu_{X|Y}(y) = \int_{-\infty}^{\infty} x f_{X|Y=y}(x) dx = \int_{0}^{1} x \frac{6(x^{2} + xy)}{(2+3y)} dx \text{ for } 0 \le y \le 1.$$

- The expected value of Y can be obtained as the expected value of the regression function.
- This is called the **Law of Total Expectation**.

$$E(Y) = E\left[E(Y \mid X)\right].$$

• Therefore,

$$E(Y) = \sum_{x \in \mathcal{S}_X} E(Y \mid X = x) p_X(X = x) = \sum_{x \in \mathcal{S}_X} \mu_{Y|X}(x) p_X(X = x)$$

or in the continuous case

$$E(Y) = \int_{-\infty}^{\infty} E(Y \mid X = x) f_X(x) dx = \int_{-\infty}^{\infty} \mu_{Y|X}(x) f_X(x) dx.$$

### Independence

- The random variables X and Y are independent if any event defined in terms of X is independent of any event defined in terms of Y.
- In particular, X and Y are independent if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B),$$

holds for any two sets (subsets of the real line) A and B.For the discrete case we have

$$p_{X,Y}(x,y) = p_X(x)p_Y(y),$$

holds for all x, y.

• For the continuous case

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

holds for all x, y.

Kayvan Nejabati Zenouz MATH1180

# Example

**\bigcirc** Random variable X and Y in the table

			y	
	p(x,y)	1	2	$p_X(x)$
	1	0.3	0.13	0.43
x	2	0.06	0.26	0.32
	3	0.1	0.15	0.25
	$p_Y(y)$	0.46	0.54	

are not dependent, since  $0.3 \neq 0.43 \times 0.46$ .

**2** Random variable X and Y in the table

			y	
	p(x,y)	1	2	$p_X(x)$
x	1	0.08	0.02	0.1
	2	0.72	0.18	0.9
	$p_Y(y)$	0.8	0.2	

are independent.

• The random variables in example on slide 50 are not independent since  $\frac{12}{7}(x^2 + xy) \neq \frac{6}{7}(2x^2 + x) \times \frac{2}{7}(2 + 3y)$ .

# Consequences of Independence

Let X and Y be independent jointly discrete random variables (similar holds when PDFs replacing PMFs). Then we have the following.

$$p_{Y|X=x}(y) = \frac{p(x,y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y),$$

similarly,  $p_{X|Y=y}(x) = p_X(x)$ .

• The regression function is constant,

$$\mu_{Y|X}(x) = \sum_{y \in \mathcal{S}_Y} y p_{Y|X=x}(y) = \sum_{y \in \mathcal{S}_Y} y p_Y(y) = E(Y), \ x \in \mathcal{S}_X,$$

similarly  $\mu_{X|Y}(y) = E(X).$ 

- For any functions g and h, we have g(X) and h(Y) are independent.
- We have

$$E\left[g(X)h(Y)\right] = E\left[g(X)\right]E\left[h(Y)\right].$$

### **Remark 5: Multivariate Distributions**

The jointly discrete random variables  $X_1, X_2, ..., X_n$  are independent if and only if

$$p(x_1, x_2, ..., x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n),$$

and the jointly continuous  $X_1, X_2, ..., X_n$  are independent if and only if

$$f(x_1, x_2, ..., x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

### **Definition (IID Variables)**

If  $X_1, X_2, ..., X_n$  are independent and also have the same distribution (which is the case of a simple random sample from an infinite/hypothetical population) they are called **independent and identically distributed**, or iid for short.

### Functions of Random Variables

• Let (X, Y) be discrete with joint PMF p(x, y). The expected value of a function, h(X, Y), of (X, Y) is computed by

$$E[h(X,Y)] = \sum_{x \in S_X} \sum_{y \in S_Y} h(x,y)p(x,y).$$

• Let (X, Y) be continuous with joint PDF f(x, y). The expected value of a function, h(X, Y), of (X, Y) is computed by

$$E[h(X,Y)] = \int_{\infty}^{\infty} \int_{\infty}^{\infty} h(x,y) f(x,y) dx dy.$$

• The variance of h(X, Y) is computed by

$$\sigma_{h(X,Y)}^{2} = E\left[h^{2}(X,Y)\right] - \left[E\left[h(X,Y)\right]\right]^{2}.$$

### Example

Show that for any two random variables X and Y we have

$$E(X+Y) = E(X) + E(Y).$$

• Let  $X_1, ..., X_n$  be any *n* random variables (i.e., they may be discrete or continuous, independent or dependent), with marginal means  $E(X_i) = \mu_i$ . Then

$$E\left(a_1X_1 + \dots + a_nX_n\right) = a_1\mu_1 + \dots + a_n\mu_n,$$

where  $a_i$  are constants. In particular, if  $\mu_1 = \mu_2 = \cdots = \mu_n = \mu$ , then

$$E(X_1 + \dots + X_n) = n\mu$$
, and  $E(\overline{X}) = \mu$ .

### Variance of Sum and Covariance

• Let's consider the variance of X + Y, then we have

$$Var (X + Y) = E \left\{ [X + Y - E (X + Y)]^2 \right\}$$
  
= Var(X) + Var(Y)  
+ 2E [(X - E(X)) (Y - E(Y))].

• If X and Y are independent, then we have

$$E[(X - E(X))(Y - E(Y))] = 0.$$

The quantity E [(X - E(X)) (Y - E(Y))] is known as the covariance of X and Y denoted by

$$\operatorname{Cov}(X, Y) = \sigma_{X,Y} = E\left[(X - E(X))\left(Y - E(Y)\right)\right]$$
$$= E(XY) - \mu_X \mu_Y.$$

Calculate Cov(X, Y) for the discrete random variables X, Y with p(x, y) given by table below

			y	
	p(x,y)	1	2	$p_X(x)$
	1	0.3	0.13	0.43
x	2	0.06	0.26	0.32
	3	0.1	0.15	0.25
	$p_Y(y)$	0.46	0.54	

**2** Calculate Cov(X, Y) when X, Y have the joint density function

$$f(x,y) = \begin{cases} \frac{12}{7} (x^2 + xy) & 0 \le x, y \le 1\\ 0 & \text{otherwise.} \end{cases}$$

### Properties of Variance

• Let  $\sigma_1^2$  and  $\sigma_2^2$  denote the variances of  $X_1, X_2$ , respectively. Then

• If  $X_1, X_2$  are independent (or just  $Cov(X_1, X_2) = 0$ ),

$$\operatorname{Var}(X_1 + X_2) = \sigma_1^2 + \sigma_2^2, \ \operatorname{Var}(X_1 - X_2) = \sigma_1^2 + \sigma_2^2$$

**2** If  $X_1, X_2$  are dependent,

$$Var(X_1 + X_2) = \sigma_1^2 + \sigma_2^2 + 2Cov(X, Y), Var(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 - 2Cov(X, Y).$$

• If  $X_1, ..., X_n$  are independent (or just  $Cov(X_i, X_j) = 0$  for all  $i \neq j$ ) with variances  $\sigma_1^2, ..., \sigma_n^2$ ,

$$\operatorname{Var}(a_1X_1 + \dots + a_nX_n) = a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2.$$

where  $a_i$  are constants.

• If  $X_1, ..., X_n$  are not independent, then

$$\operatorname{Var}(a_1X_1 + \dots + a_nX_n) = a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2 + \sum_i \sum_{i \neq j} a_i a_j\sigma_{ij}.$$

Kayvan Nejabati Zenouz MATH1180

# Properties of Variance and Covariance

• Let  $X_1, ..., X_n$  be iid (i.e., a simple random sample from an infinite population) with common variance  $\sigma^2$ . Then,

$$\operatorname{Var}(X_1 + \dots + X_n) = n\sigma^2$$
, and  $E(\overline{X}) = \frac{\sigma^2}{n}$ .

- We have
  - Cov(X, Y) = Cov(Y, X).
    Cov(X, X) = Var(X).
    If X, Y are independent, then Cov(X, Y) = 0.
    Cov(aX + b, cY + d) = acCov(X, Y) for any real numbers a, b, c, and d.

What we did today...



# Topic 5 Pearson's Correlation and Regression

# Lecture Contents

1 Int

#### Introduction

- Module Aims and Assessment
- Topics to be Covered
- Programming Package R

RStudio GUI

Reading List and References

#### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

#### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

#### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking

6

 Model Fitting with Least Square Method

#### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals

#### Topic 7: Introduction to Time Series

- Times Series Data and Visualisation
- Definition and Examples

Time Series
 Decomposition

- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

### By the end of this session you will be able to...

- Understand how to quantify dependence of random variables.
- **2** Calculate Pearson's correlation coefficient.
- **③** Learn the concepts behind linear regression models.
- **4** Interpret R output for a linear regression.
- Preform model checking on linear models.

# Introduction

- In the previous topic we studied **joint distribution** of the random variables.
- For X and Y random variables which are jointly distributed we defined the concepts of independence.

 $p_{X,Y}(x,y) = p_X(x)p_Y(y), \ f_{X,Y}(x,y) = f_X(x)f_Y(y).$ 

- When two random variables X and Y are not independent, they are dependent: **correlation** is a mean for quantifying dependence.
- Variables X and Y are **positively dependent**, or positively correlated, if "large" values of X are associated with "large" values of Y and "small" values of X are associated with "small" values of Y.
- Similarly you can have **negatively dependent** or negatively correlated variables.
- An example of negatively dependent variables is X stress applied and Y time to failure.

• Recall for X and Y, then the regression function

$$\mu_{Y|X}(x) = E(Y \mid X = x)$$

is a function of x.

- Now if X and Y are **positively dependent**, then  $\mu_{Y|X}(x)$  is an increasing function of x.
- For example, consider X height and Y weight then, due to the positive dependence of these variables, we have  $\mu_{Y|X}(1.82) < \mu_{Y|X}(1.90)$ , that is, the average weight of men 1.82 meters tall is smaller than the average weight of men 1.90 meters tall.
- It turns out that **dependence** is positive or negative if the **covariance** takes a positive or negative value, respectively.

# Covariance as a Measure of Dependence

- Consider a finite population of N units.
- Let  $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$  denote the values of a bivariate characteristic of interest for each of the N units, and let (X, Y) denote the bivariate characteristic of a randomly selected unit.
- Then (X, Y) has a discrete distribution taking each of the possible values  $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$  with probability 1/N.
- In this case the covariance formula in definition can be written as

$$\sigma_{X,Y} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X) (y_i - \mu_Y).$$

• The value of  $\sigma_{X,Y}$  will tend to be positive or negative based on dependence of X and Y.

### **Definition (Pearson's Correlation)**

The Pearson's (or linear) correlation coefficient of X and Y, denoted by  $\operatorname{Corr}(X, Y)$  or  $\rho_{X,Y}$ , is defined as

$$\rho_{X,Y} = \operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

where  $\sigma_X$  and  $\sigma_Y$  are the marginal standard deviations of X, Y, respectively.

### **Remark 1: Properties of Correlation Coefficient**

• If a and c are either both positive or both negative, then

$$\operatorname{Corr}(aX + b, cY + d) = \operatorname{Corr}(X, Y).$$

If a and c are of opposite signs, then  $\operatorname{Corr}(aX + b, cY + d) = -\operatorname{Corr}(X, Y).$ 

- We have  $-1 \leq \rho_{X,Y} \leq 1$ .
- If X and Y are independent, then  $\rho_{X,Y} = 0$ .
- We have  $\rho_{X,Y} = 1$  or -1 if and only if Y = aX + b for some a, b with  $a \neq 0$ .

## **Exercise 1:** Correlation Coefficient

• Calculate  $\operatorname{Corr}(X, Y)$  for the discrete random variables X, Y with p(x, y) given by table below

			y	
	p(x,y)	1	2	$p_X(x)$
	1	0.3	0.13	0.43
x	2	0.06	0.26	0.32
	3	0.1	0.15	0.25
	$p_Y(y)$	0.46	0.54	

**2** Calculate  $\operatorname{Corr}(X, Y)$  when X, Y have the joint density function

$$f(x,y) = \begin{cases} \frac{12}{7} (x^2 + xy) & 0 \le x, y \le 1\\ 0 & \text{otherwise.} \end{cases}$$

### Covariance and Correlation Coefficient for Samples

If (X<sub>1</sub>, Y<sub>1</sub>), ..., (X<sub>n</sub>, Y<sub>n</sub>) is a sample from the bivariate distribution of (X, Y), the sample covariance, denoted by Cov(X, Y) or S<sub>X,Y</sub>, and sample correlation coefficient, denoted by Corr(X, Y) or r<sub>X,Y</sub>, are defined as

$$S_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} \left( X_i - \overline{X} \right) \left( Y_i - \overline{Y} \right)$$
$$r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y}.$$

• A computational formula for the sample covariance is

$$S_{X,Y} = \frac{1}{n-1} \left[ \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right) \left( \sum_{i=1}^{n} Y_i \right) \right].$$

The value of  $r_{X,Y}$  needs to be interpreted carefully, especially when:

- X and Y have a non-linear relationship.
- There are influential values or outliers.
- If X and Y are not normally distributed.
- You must plot the data before interpreting the value of  $r_{X,Y}$ .

# **Regression** Models

- Regression models are used whenever the primary objective of the study is to understand the nature of the **regression** function of a variable Y on another variable X.
- For example, a study of the speed, X, of a car and the stopping distance, Y.
- In regression studies Y is called the **response variable**, and X is interchangeably referred to as the **covariate**, or the independent variable, or the predictor, or the **explanatory variable**.
- Because interest lies in the conditional mean of Y given X = x. In fact we have **hierarchical modelling** consisting of

$$Y \mid X = x \sim F_{Y|X=x}(y), \ X \sim F_X(x)$$

• Where the conditional distribution of Y given X = x,  $F_{Y|X=x}(y)$ , and the marginal distribution of X,  $F_X(x)$ , may or may not specified, can depend on additional parameters.

# The Simple Linear Regression Model

• The simple linear regression model specifies that the regression function of Y on X is linear, that is,

$$\mu_{Y|X}(x) = \beta_0 + \beta_1 x,$$

- and the conditional variance of Y given X = x, denoted by  $\sigma_{\epsilon}^2$ , is the same for all values x, known as the homoscedasticity assumption.
- In this model,  $\beta_0$ ,  $\beta_1$ , and  $\sigma_{\epsilon}^2$  are unknown parameters.
- The marginal expectation of Y is given by

$$E(Y) = \beta_0 + \beta_1 \mu_X.$$

• For the simple linear regression model, where  $E(Y \mid X)$  is given by either the mean plus error form is

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

• The intrinsic error  $\epsilon$  has zero mean and is uncorrelated from the explanatory variable X, with  $Var(\epsilon) = \sigma_{\epsilon}^2$  so

$$E(\epsilon) = 0$$
,  $Cov(\epsilon, X) = 0$ , and  $Var(\epsilon) = \sigma_{\epsilon}^2$ .

If the regression function of Y on X is linear, then
The marginal variance of Y is

$$\sigma_Y^2 = \sigma_\epsilon^2 + \beta_1^2 \sigma_X^2$$

**2** The slope  $\beta_1$  is related to the covariance,  $\sigma_{X,Y}$ , and the correlation,  $\rho_{X,Y}$ , by

$$\beta_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}$$

If  $(X_1, Y_1), ..., (X_n, Y_n)$  is a sample from the bivariate distribution of (X, Y), then  $\sigma_{X,Y}$  can be estimated by  $S_{X,Y}$  so

$$\widehat{\beta}_1 = \frac{S_{X,Y}}{S_X^2}$$
, and  $\beta_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$ .

• The normal regression model specifies that the conditional distribution of Y given X = x is normal

$$Y \mid X = x \sim N(\mu_{Y|X}(x), \sigma_{\epsilon}^2).$$

• The normal simple linear regression model is also written as

$$Y = \beta_0 + \beta_1 x + \epsilon, \ \epsilon \sim N(0, \sigma_{\epsilon}^2).$$

# Example

The following is an R output for a linear regression.

```
Call:
lm(formula = dist ~ speed, data = cars)
Residuals:
Min 10 Median 30 Max
-29.069 -9.525 -2.272 9.215 43.201
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791 6.7584 -2.601 0.0123 *
speed
            3.9324 0.4155 9.464 1.49e-12 ***
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
 1
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

• The formula is given by

 $dist = -17.5791 + 3.9324 \times speed.$ 

- The residual degrees of freedom is 48 and we have 2 parameters estimates, so 50 observation.
- The small *p*-value for the coefficient estimate for speed shows that the value for estimate is significantly different from zero.
- The small *p*-value for the F-statistics shows the significance of the overall regression.
- The  $R^2$  value is 0.6511 and its signifies roughly how much of the variation in data has been explained by the regression model.

The assumptions of linear regression is that there is a linear relationship between response variable  $y_i$  and explanatory variable  $x_i$  which has the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where  $\epsilon_i$  are independent and from a normal distribution  $N(0, \sigma_{\epsilon}^2)$ . These we can check by residual plots.

### **Residual Plots**





There should be no obvious pattern in the spread of residuals and their spread constant around the dashed line. Here we have a few large positive residuals and a slight non-linear pattern in the residuals.


Residuals should fit the Q-Q plot as much as possible. Here there is some deviation in the tails.



Scale-Location

Here the blue line should be as horizontal as possible.



Point with large Cook's distance can influence the model and indicate outliers. The Cook's distance is considered high if it is greater than 0.5 and extreme if it is greater than 1.



Points with large Leverage and Residual should be checked, they may indicate outliers.



Points with large Leverage and Cook's distance should be checked.

The multiple linear regression (MLR) model specifies that the conditional expectation,

$$E(Y \mid X_1 = x_1, ..., X_k = xk) = \mu_{Y|X_1, ..., X_k}(x_1, ..., x_k),$$

of a response variable Y given the values of k predictor variables,  $X_1, ..., X_k$ , is a linear function of the predictors' values:

$$\mu_{Y|X_1,...,X_k}(x_1,...,x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

## Model Fitting with Least Square Method

- The method of least squares (LS), which is the most common method for fitting regression models.
- Let  $(X_1, Y_1), ..., (X_n, Y_n)$  denote a simple random sample from a bivariate population (X, Y).
- We would like to find estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for the equation

$$\mu_{Y|X}(x) = \beta_0 + \beta_1 x.$$

• For this we can find  $\widehat{\beta}_0$  and  $\widehat{\beta}_1$  which minimise

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

• These estimates are given by

$$\widehat{\beta}_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}, \ \widehat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}.$$

• Given the estimates in the previous pages,

$$\widehat{Y}_i = \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

### are called **fitted values**.

• The estimated intrinsic error variables

$$\widehat{\epsilon_i} = Y_i - \widehat{Y_i} = Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i,$$

### are called **residuals**.

• Because the computation of residuals requires that two parameters be estimated, we use the following estimating  $\sigma_{\epsilon}^2$ 

$$S_{\epsilon}^2 = \frac{1}{n-2} \sum_{i=1}^n \widehat{\epsilon_i}^2.$$

What we did today...



# Topic 6 Approximations and Confidence Intervals

## Lecture Contents

### l) Inti

### Module Aims and

Assessment

- Topics to be Covered
- Programming Package R

RStudio GUI

Reading List and References

### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

#### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

#### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking
- Model Fitting with Least Square Method

#### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals

#### Topic 7: Introduction to Time Series

- Times Series Data and Visualisation
- Definition and Examples
- Time Series
   Decomposition
- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

MATH1180

• Th Topi

### By the end of this session you will be able to...

- **Q** Learn approximation results relating to sample estimates.
- **②** Understand and calculate confidence intervals.

- We have seen that using a **sample** we can calculate several useful quantities.
- Intuition: the **bigger** the sample size the **better** the approximation.
- We shall review two **approximation** results:
  - The **Law of Large Numbers** is an explicit assertion that above intuition is in fact true.
  - **Central Limit Theorem** provides an approximation to the distribution of sums or averages.
- Using the above we look at properties of **point estimators**.
- In particular, the probability an estimate will be within a certain distance from the true population: **confidence interval**.

- LLN justifies the approximation of the **population** mean by the **sample** mean.
- It does not help in determining how large the sample size should be for a given quality of the approximation.
- Given an estimator θ
   of a parameter θ, we say that θ

   converges in probability to θ, which means that for any

$$P\left(\left|\widehat{\theta}-\theta\right|>\epsilon\right)\longrightarrow 0 \text{ as } n\longrightarrow\infty,$$

where n is the sample size.

- Whenever an estimator converges in probability to the quantity it is supposed to estimate, we say that the estimator is **consistent**.
- The LLN, stated below, asserts that averages possess the consistency property.

### Theorem (The Law of Large Numbers)

Let  $X_1, ..., X_n$  be independent and identically distributed and let g be a function such that  $-\infty < E[g(X_1)] < \infty$ . Then,

$$\frac{1}{n}\sum_{i=1}^{n}g(X_{i}) \text{ converges in probability to } E\left[g\left(X_{1}\right)\right].$$

### Remark 1: The Law of Large Number for Mean

Note in the above if g(x) = x, then we have

$$\frac{1}{n}\sum_{i=1}^{n}X_{i}=\overline{X} \text{ converges in probability to } E\left(X_{1}\right)=\mu,$$

i.e.,  $\overline{X}$  is a consistent estimator of the population mean  $\mu$ .

## Example

Cylinders are produced in such a way that their height is fixed at 5 centimetres (cm), but the radius of their base is uniformly distributed in the interval (9.5 cm, 10.5 cm). The volume of each of the next 100 cylinders to be produced will be measured, and the 100 volume measurements will be averaged. What will the approximate value of this average be?

- Let  $X_i$  for i = 1, ..., 100, and  $\overline{X}$  denote the volume measurements and their average, respectively.
- By the LLN,  $\overline{X}$  should be approximately equal to the expected volume of a randomly selected cylinder.
- Since the volume is given by  $X = \pi R^2 h$ , we have

$$E(X) = E(R^2) = 1572.1.$$

• Thus, the value of  $\overline{X}$  should be "close" to 1572.1.

## Distribution of Sums (Convolution)

- Given a number of random variables  $X_1, ..., X_n$ , sometimes we are interested in the distribution of their sums.
- For example, it is known that if  $X \sim \text{Poi}(\lambda_1)$  and  $Y \sim \text{Poi}(\lambda_2)$  are independent, then

$$X + Y \sim \operatorname{Poi}(\lambda_1 + \lambda_2).$$

• Similarly, if  $X \sim Bin(n_1, p)$  and  $Y \sim Bin(n_2, p)$  are independent, then

$$X + Y \sim \operatorname{Bin}(n_1 + n_2, p).$$

• In general given independent random variable  $X_1, ..., X_n$ , it may not be easy to find the distribution of their sum.

## **Bivariate Normal Distribution**

The bivariate normal distribution has the joint PDF of (X, Y)

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{1-\rho^2} \left[\frac{(x-\mu_X)^2}{2\sigma_X^2} - \frac{\rho\left(x-\mu_X\right)\left(y-\mu_Y\right)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right]\right\}$$

where  $\mu_X, \mu_Y, \sigma_X, \sigma_Y$  and  $\rho$  are given. It has the following properties.

- The marginal distribution of Y is also normal.
- If X and Y are uncorrelated then they are independent.
- If X and Y are independent normal random variables, their joint distribution is bivariate normal with parameters  $\mu_X, \mu_Y, \sigma_X, \sigma_Y$  and  $\rho = 0$ .
- Any linear combination of X and Y has a normal distribution. In particular,

$$aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\operatorname{Cov}(X, Y)).$$

### Proposition

Let  $X_1, ..., X_n$  are independent and normally distributed random variables with  $X_i \sim N(\mu_i, \sigma_i^2)$ . Let  $Y = a_1 X_1 + \cdots + a_n X_n$  be a linear combination of  $X_i$ . Then

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

where 
$$\mu_Y = a_1\mu_1 + \dots + a_n\mu_n$$
 and  $\sigma_Y^2 = a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2$ 

### Example

If  $X_1 \sim N(0,1)$ ,  $X_2 \sim N(-1,4)$  and  $X_3 \sim N(2,2)$ . Then for  $Y = X_1 + 2X_2 + 3X_3$  we have

$$Y \sim N(0 + 2 \times (-1) + 3 \times 2, 1 + 2^2 \times 4 + 3^2 \times 2) = N(4, 35).$$

## Distribution of $\overline{X}$ in the Normal Case

### Corollary

Let  $X_1, ..., X_n$  be iid  $N(\mu, \sigma^2)$ , and let  $\overline{X}$  be the sample mean. Then

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

### Exercise 1: Distribution of $\overline{X}$

It is desired to estimate the mean of a normal population whose variance is known to be  $\sigma^2 = 9$ . What sample size should be used to ensure that  $\overline{X}$  lies within 0.3 units of the population mean with probability 0.95?

## Central Limit Theorem

- A simple way to **approximate** sum or average of a large number of random variables.
- Most **important** theorem in probability and statistics.

### Theorem (The Central Limit Theorem)

Let  $X_1, ..., X_n$  be iid with mean  $\mu$  and a finite variance  $\sigma^2$ . Then for large enough  $n \ (n \ge 30$  for our purposes), we have that  $\overline{X}$ has approximately a normal distribution with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ , that is  $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ .

- The **quality** of the approximation **increases** with *n*, and also depends on the population distribution.
- For example, data from skewed populations require a larger sample size than data from, say, the uniform distribution.

## Example

The number of units serviced in a week at a certain service facility is a random variable having mean 50 and variance 16. Find an approximation to the probability that the total number of units to be serviced at the facility over the next 36 weeks is between 1728 and 1872. **Solution.** 

- Let  $X_1, ..., X_{36}$  denote the number of units that are serviced in each of the next 36 weeks, and assume they are iid.
- Set  $T = \sum_{i=1}^{36} X_i$ . Then  $E(T) = 36 \times 50 = 1800$  and  $Var(T) = 36 \times 16 = 576$ .
- Since the sample size is  $n \ge 30$ , according to the CLT the distribution of T is approximately normal with mean 1800 and variance 576.
- Thus we need P(1728 < T < 1872)

> pnorm(1872,1800,sqrt(576))-pnorm(1728,1800,sqrt(576))
[1] 0.9973002

### Estimation

- Estimation of population parameters, such as proportion, mean, variance, covariance and Pearson's correlation coefficients, is achieved by using the corresponding sample quantities.
- The Greek letter  $\theta$  will be used as a generic notation for any model or **population** parameter(s) that we are interested in estimating.
- When a **sample** is denoted in capital letters, such as  $X_1, ..., X_n$ , the  $X_i$ 's are considered random variables, that is, before their values are observed.
- The **observed** sample values, or data, are denoted in lowercase letters, that is,  $x_1, ..., x_n$ .
- A quantity used to **estimate** the true value of a parameter  $\theta$  is denoted by  $\hat{\theta}$ .
- Then  $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$  is known as the an **estimator** and  $\hat{\theta} = \hat{\theta}(x_1, ..., x_n)$  is an **estimate**.

- Estimators **approximate** population **parameters** but, in general, are different from them.
- CLT allows one to assess the **probability** that an **estimator** will be within a certain distance from the true population parameter.
- Confidence intervals have been devised to **address** the lack of **information** that is inherent in the practice of reporting only a point estimate.
- A confidence interval is an interval for which we can assert, with a given degree of confidence/certainty, that it includes the true value of the **parameter** being estimated.

### **Distribution of** $\hat{\theta}$

- By CLT if sample size is large enough, then  $\hat{\theta}$  is approximately normally distributed with mean  $\theta$ .
- By LLN estimated standard error  $S_{\widehat{\theta}}$  will be good estimate of  $\sqrt{\operatorname{Var}\left(\widehat{\theta}\right)}$ .
- Taken together, these facts imply that

$$\widehat{\theta} \sim N(\theta, S_{\widehat{\theta}}) \text{ or } Z = \frac{\widehat{\theta} - \theta}{S_{\widehat{\theta}}} \sim N(0, 1).$$

## Construction of Confidence Interval II

• Now to construct an interval which we expect to contain the true parameter, for example, 95% of the time we need

$$P(|Z| < \alpha) = P(-\alpha < Z < \alpha) = 0.95.$$

> qnorm(0.975, mean = 0, sd = 1)
[1] 1.959964 # P(Z<alpha)=0.975</pre>

• Therefore we need

$$|Z| \le 1.96,$$

i.e.,

$$-1.96 \le \frac{\widehat{\theta} - \theta}{S_{\widehat{\theta}}} \le 1.96$$

• For a particular estimate of  $\theta$  we have a 95% confidence interval

$$\widehat{\theta} - 1.96S_{\widehat{\theta}} \le \theta \le \widehat{\theta} + 1.96S_{\widehat{\theta}}.$$

## Z and T Confidence Intervals

- Confidence intervals using standard normal distribution are known as **Z** confidence intervals.
- Z intervals for mean used only if the population variance is known and either the population is normal or the sample size is at least 30.
- When sampling from normal populations, an estimator  $\hat{\theta}$  of some parameter  $\theta$  often satisfies, for all sample sizes n,

$$\frac{\widehat{\theta} - \theta}{S_{\widehat{\theta}}} \sim T_{\nu}.$$

- $T_{\nu}$  stands for the *T* distribution with  $\nu$  degrees of freedom. The degrees of freedom depend on the sample size.
- A T distribution is symmetric and its PDF tends to that of the standard normal as  $\nu$  tends to infinity.
- Then for  $(1 \alpha)100\%$  confidence interval we have the following bound on the error

$$\widehat{\theta} - t_{\nu,\alpha/2} S_{\widehat{\theta}} \le \theta \le \widehat{\theta} + t_{\nu,\alpha/2} S_{\widehat{\theta}}.$$

Kayvan Nejabati Zenouz MATH1180

### T CI for the Mean

- Let  $X_1, ..., X_n$  be a simple random sample from a population, and let  $\overline{X}$ ,  $S^2$  denote the sample mean and sample variance, respectively.
- If the population is normal (if not we need  $n \ge 20$ ), then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}.$$

• Then for  $(1 - \alpha)100\%$  confidence interval we have the following bound on the error

$$\overline{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}.$$

• In R you can use

confint(lm(x~1), level=1-alpha)

## Example

A random sample of n = 56 cotton pieces gave average percent elongation of  $\overline{X} = 8.17$  and a sample standard deviation of S = 1.42. Construct a 95% CI for  $\mu$ , the population mean percent elongation. **Solution.** 

- Because the sample is large enough, the CI can be used without the assumption of normality.
- The degrees of freedom is  $\nu = 56 1 = 55$ .
- The R command qt(0.975, 55) = 2.004 gives the exact value
- Using the approximate value of the percentile, 2.01 and the given sample information, we obtain

$$7.80 \le \mu \le 8.54.$$

What we did today...

Approximations

Confidence Intervals

Next Time

The Law of Large Numbers, The Centr Limit Theorem Estimation, T CI for the Mean

Introduction to Time Series

# Topic 7 Introduction to Time Series

## Lecture Contents

### 🚺 Intr

- Module Aims and
- Topics to be Covered
- Programming Package R
- RStudio GUI
- Reading List and References

### Topic 1: Review of Statistics and Probability

- Population and Samples
- Sampling Techniques
- Statistical Visualisation
- Populations/Samples Parameters
- Sample Spaces and Events
- Probability of Events
- Counting Techniques
- Axioms of Probability
- Conditional Probability
- Total Law of Probability and Bayes' Theorem
- Independent Events

#### Topic 2: Introduction to Random Variables

- Random Variables
- Probability Mass Function

- Cumulative Distribution Function
- Probability Density Function
- Expected Value
- Variance and Standard Deviation

#### Topic 3: Discrete and Continuous Random Variables

- The Bernoulli and Binomial Distributions
- The Geometric Distribution
- The Poisson Distribution
- The Exponential Distribution
- The Normal Distribution

#### Topic 4: Joint Distribution of Random Variables

- Joint Probability Distribution
- Conditional Probability Functions
- Conditional Expectation: Regression Function
- Independence
- Functions of Random Variables
- Variance and Covariance

#### Topic 5: Pearson's Correlation and Regression

- Pearson's Correlation Coefficient
- Regression Models
- Model Checking
- Model Fitting with Least Square Method

#### Topic 6: Approximations and Confidence Intervals

- The Law of Large Numbers
- The Central Limit Theorem
- Confidence Intervals
- Topic 7: Introduction to Time Series
- Times Series Data and Visualisation
- Definition and Examples
- Time Series Decomposition
- Lags and Differences
- Autocorrelation and Autocovariance
- Stationary Time Series
- Trend and Seasonality
- Autoregressive Moving Average
- Forecasting

### By the end of this session you will be able to...

• Learn basic properties of times series.

**2** Use software package R to analyse time series data.

## Introduction

- **Time series** analysis is concerned with experimental **data** that have been **observed** at different points in **time**.
- A few **examples** include
  - Daily **stock market**, GDP, quotations or monthly unemployment figures.
  - Population series, such as **birthrates** or school enrolments.
  - The number of **influenza cases** observed over some time period.
  - Blood pressure **measurements** traced over time useful for evaluating drugs used.
  - Functional magnetic resonance imaging of **brain-wave** time to study how the brain reacts to certain stimuli under various experimental conditions.
- We shall look at **visualisation**, **modelling**, and **forecasting** for time series.

### Times Series Data

- We usually have **correlation** introduced by the sampling of **adjacent points** in time.
- This means methods dependent on the assumption that these **adjacent observations** are **independent** and identically distributed is **not valid**.
- A basic **time series data** looks like the following.

	Q1	Q2	Q3	Q4
2010		2.00	1.76	1.45
2011	4.78	5.72	8.78	8.39
2012	10.37	5.40	5.02	6.60
2013	8.99	10.69	13.11	10.57
2014	8.69	7.94	11.62	11.20
2015	18.76	12.40	13.13	11.56
2016	14.00	13.30	15.67	14.75
2017	15.02	11.20	14.88	

Table: Quarterly Sales Data

Kayvan Nejabati Zenouz MATH1180

## Visualisation

You can plot your time series for time on the x-axis and values on the y-axis.



180

Kayvan Nejabati Zenouz MATH1180
You can create tm series in R using ts() function.

```
value <- 1:30 + rnorm(30,0,2) # Any vector of values
ts(value, frequency=365, start=c(2014, 6)) # Daily
ts(value, frequency=4, start=c(2010, 2)) # frequency 4 =>
Quarterly Data
ts(value, frequency=12, start=1990) # freq 12 => Monthly
data.
ts(value, start=c(2001), end=c(2014), frequency=1) # Yearly
Data
library(xts) # Using xts library
dates <- seq(as.Date("2016-01-01"), length=30, by="days")
xts(x = value, order.by = dates)
```

Alternatively you can use xts, or zoo library.

#### Trend, Periodicity, Noise, Forecast

- It allow us to observe the **primary patterns** in the time history, underlying **trend**.
- **Regular variation** superimposed on the trend that seems to repeat over periods, underlying **periodicity** or seasonality.
- Forecast the from current data the **future** events and determine uncertainty.
- May also be interested in analysing **several time series** at once to study **their relationship**.

#### Definition

- A time series can be defined as a collection of random variables indexed according to the order they are obtained in time.
- For each point in time t = 0, 1, 2, ..., minute, hours, day, month, etc.., we have a random variable  $X_t$ .
- In general, a collection of random variables  $\{X_t\}$  indexed by t is referred to as a **stochastic process**.
- The observed values of a stochastic process  $\{x_t\}$  are referred to as a realization of the stochastic process.
- A description of a time series, observed as a collection of n random variables at arbitrary time points  $t_1, t_2, ..., t_n$  is provided by the joint distribution function.

### White Noise

In this case we have  $E(X_t) = 0$  and  $Var(X_t) = \sigma^2$ , so

$$X_t \sim \operatorname{wn}(0, \sigma^2)$$
 or  $X_t \sim \operatorname{iid}(0, \sigma^2), \ X_t \sim \operatorname{iid} N(0, \sigma^2).$ 



184

Kayvan Nejabati Zenouz MATH1180

# Linear Trend and White Noise

In this case we have  $E(X_t) = \mu_t = \beta_0 + \beta_1 t$  and  $Var(X_t) = \sigma^2$ .



# Seasonality and White Noise

In this case we have  $E(X_t) = \mu_t = A \cos(2\pi\omega t + \phi)$  and  $\operatorname{Var}(X_t) = \sigma^2$ .



186

Kayvan Nejabati Zenouz MATH1180

### First Order Autoregressive (Random Walk with Drift)

In this case we have  $X_t = X_{t-1} + \delta + \epsilon_t$ .



## Time Series Decomposition

- A time series can be expressed as either a sum or a product of 3 components of **Seasonality**  $S_t$ , **Trend**  $T_t$  and **Error**  $\epsilon_t$ , a.k.a White Noise.
- That is we have
  - Additive time series:

$$X_t = T_t + S_t + \epsilon_t.$$

• Multiplicative time series:

$$X_t = T_t \times S_t \times \epsilon_t.$$

• A multiplicative time series can be **converted** into and additive one by taking logarithm

$$\log \left( T_t \times S_t \times \epsilon_t \right) = \log T_t + \log S_t + \log \epsilon_t.$$

#### Definition $(j^{\text{th}} \text{Lag})$

Previous values of a time series are called **lags**. The first lag of  $X_t$  is  $X_{t-1}$ . The  $j^{\text{th}}$  lag of  $X_t$  is  $X_{t-j}$ , sometimes written as  $B^j X_t$ . Lags of univariate or multivariate time series objects are conveniently computed by lag(mytsdata,j)

#### Example

• If 
$$X_t = \epsilon_t$$
, then  $X_{t-j} = \epsilon_{t-j}$ .

• If 
$$X_t = \beta_0 + \beta_1 t + \epsilon_t$$
, then  $X_{t-j} = \beta_0 + \beta_1 (t-j) + \epsilon_{t-j}$ .

#### **Definition** $(j^{\text{th}} \text{ Difference})$

The difference between  $X_t$  and  $X_{t-j}$  is denoted by  $\Delta_j X_t = X_t - X_{t-j}$ , you can iterate it k-times and get  $\Delta_j^k$ , this is computed by diff(mytsdata, lag = j, differences=k).

# Autocorrelation and Autocovariance

It is important to know if  $X_t$  is correlated with the previous values in time, i.e., if  $X_t$  and  $B^j X_t = X_{t-j}$  are related for some j.

#### **Definition** (Autocorrelation and Autocovariance)

The covariance between  $X_t$  and  $X_s$  is called the  $(s-t)^{\text{th}}$ autocovariance of the series  $X_t$ . Autocorrelation coefficient, also called the serial correlation coefficient, measures the correlation between  $X_t$  and  $X_s$ , so we have

$$(s-t)^{\text{th}} \text{Autocovariance} = \gamma_X(s,t) = \text{Cov}(X_s, X_t)$$
  
 $(s-t)^{\text{th}} \text{Autocorrelation} = \rho_X(s,t) = \frac{\text{Cov}(X_s, X_t)}{\sqrt{\text{Cov}(X_s, X_s)\text{Cov}(X_t, X_t)}}.$ 

You can use acf(mytsdata, lag = k) to see the first k autocorrelations.

Note if  $\rho_X(t,s)$  is around  $\pm 1$ , then we may hope for a linear relationship  $X_t = \beta_0 + \beta_1 X_s + \epsilon_t$ .

Kayvan Nejabati Zenouz MATH1180

### The Cross-covariance and Cross-correlation

Given two time series  $X_s$  and  $Y_t$  we may be interested they are correlated with the previous values in time, i.e., if  $X_t$  and  $B^j Y_t = Y_{t-j}$  are related for some j.

#### Definition (The Cross-covariance and Cross-correlation)

The covariance between  $X_t$  and  $Y_s$  is called the  $(s-t)^{\text{th}}$  autocovariance of between the series  $X_t$  and  $Y_t$ . Autocorrelation coefficient, also called the serial correlation coefficient, measures the correlation between  $X_t$  and  $Y_s$ , so we have

$$(s-t)^{\text{th}}\text{Autocovariance} = \gamma_{X,Y}(s,t) = \text{Cov}(X_s, Y_t)$$
$$(s-t)^{\text{th}}\text{Autocorrelation} = \rho_{X,Y}(s,t) = \frac{\text{Cov}(X_s, Y_t)}{\sqrt{\text{Cov}(X_s, X_s)\text{Cov}(Y_t, Y_t)}}$$

You can use ccf(x, y, lag.max = k) to see the first k autocorrelations.

Note if  $\rho_{X,Y}(t,s)$  is around  $\pm 1$ , then we may hope for a linear Kayvan Nejabati Zenouz MATH1180

# Stationary Time Series

#### Definition (Strictly Stationary and Stationary)

A strictly stationary time series is one for which the probabilistic behaviour of every collection of values  $\{X_{t_1}, X_{t_2}, ..., X_{t_k}\}$  is identical to that of the time shifted set  $\{X_{t_1+h}, X_{t_2+h}, ..., X_{t_k+h}\}$ . If  $X_t$  is a stationary time series, then for all t, the distribution of  $(X_t, X_{t+1}, ..., X_{t+s})$  does not depend on t.

- If a times series is **strictly stationary**, then we have  $\gamma_X(s,t) = \gamma_X(s+h,t+h)$ .
- Thus **autocovariance** function of the process **depends** only on the **time difference** between *s* and *t*, and not on the actual times.
- A stationary series is roughly horizontal, constant variance, no patterns predictable in the long-term.
- Test for stationariness by adf.test() from tseries library.

# Trend

- Assume time series Y<sub>t</sub> is being influenced by a collection of possible inputs or independent series, say, X<sub>1,t</sub>, X<sub>2,t</sub>, ..., X<sub>p,t</sub>.
- We express this relation through the linear regression model

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \dots + \beta_p X_{p,t} + \epsilon_t, \ \epsilon_t \sim \operatorname{iid}(0, \sigma^2).$$

• At a basic level you may find the **linear trend** in the time series by through a linear regression on time.

lm(mytsdata<sup>~</sup>time(mytsdata))

- The **residuals** for the linear regression above are known as the (linear) **de-trended** time series.
- You may also want to create **regression** models with **lagged** time series.

• If the times series has a **seasonal** (repeating) pattern i.e., is of the form

$$Y_t = T_t + A\cos\left(2\pi\omega t + \phi\right) + +\epsilon_t$$

where A is amplitude,  $\omega$  is frequency, and  $\phi$  phase difference.

• Then we may like to understand the pattern in data. We can work with the de-trended series i.e.,

$$Z_t = Y_t - T_t = A\cos\left(2\pi\omega t + \phi\right) + \epsilon_t$$

• This can be written as

$$Z_t = \beta_1 \cos \left(2\pi\omega t\right) + \beta_2 \sin \left(2\pi\omega t\right) + \epsilon_t.$$

• You can use the stl() function from forecast library to decompose the times series into trend and seasonal components.

- An **autoregressive** model relates a time series variable to its past values.
- An autoregressive model of order p, abbreviated AR(p), is of the form  $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \epsilon_t$ .

#### Example

First order AR(1) is given by  $X_t = \phi_1 X_{t-1} + \epsilon_t$ . We have already see a plot for Random Walk time series.

- The moving average model of order q, or MA(q) model, is defined to be  $X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$ .
- An autoregressive moving average with difference d written as ARMA(p, d, q) is a combination of AR(p) and MA(q).
- In R use auto.arima() function from forecast to fit ARIMA(p,d,q).

- Once we have a suitable model for our times series, then we can predict for future times.
- This is known as forecasting.
- In R you can use forecast() function from forecast in order to do this.

What we did today...

Time Series	
Decompsition	Multiplicative and Additive
Lags and Differences	Seasonality, Trend, Error, Forecasting
Stationary Time Series	Autocorrelation and Autocovariance
Next Time	Autoregressive Moving Average
Thanks for Your Attention!	There won't be any (revision)!
	Have a good holiday!
	197

#### Please Do Not Forget To

- Ask any **questions** now or through my contact details.
- Drop me **comments** and **feedback** relating to any aspects of the course.
- Come and see me during Student Drop-in Hours: MONDAYS 12:00-13:00 (MATHS ARCADE) AND TUESDAYS 15:00-16:00 (QM315). Alternatively, email to make an appointment.

# Thank You!