FIT2086 Lecture 1
Introduction, Models, Random Variables

Daniel F. Schmidt, with material from Geoff I. Webb

Faculty of Information Technology, Monash University

July 25, 2019

Outline

Outline

What is a “model”?

1. What is a model?
    - A mathematical description of some phenomena
2. What can we use a model for?
    - We can use it to make statements about reality
3. Where do models come from?
    - They are often learned from empirical (observational) data
4. Why is modelling important?

Data science is big business

Rank	Company	Capitalisation (US$ million)
1	Apple Inc	$749,124$
2	Alphabet	$628,610$
3	Microsoft	$528,778$
4	Amazon.com	$466,471$
5	Berkshire Hathaway	$418,880$
6	Johnson & Johnson	$357,310$
7	Facebook	$357,176$
8	Tencent	$344,879$
9	Exxon Mobil	$341,947$
10	JPMorgan Chase	$323,838$

Public Companies by Capitalisation (c. mid-2017)

Data science is big business

Rank	Company	Capitalisation (US$ million)
1	Apple Inc	$749,124$
2	Alphabet	$628,610$
3	Microsoft	$528,778$
4	Amazon.com	$466,471$
5	Berkshire Hathaway	$418,880$
6	Johnson & Johnson	$357,310$
7	Facebook	$357,176$
8	Tencent	$344,879$
9	Exxon Mobil	$341,947$
10	JPMorgan Chase	$323,838$

Public Companies by Capitalisation (c. mid-2017)

Data Science is Fun

- Data science lets you take data (numbers, measurements) and learn about the process that generated the data
- It lets you make predictions about the future using the past
  - Will Manchester United beat Real Madrid in the Champions League?
- It lets you quantify empirical evidence of phenomena
  - Do dogs really bite more frequently on the full moon?

Administrative Details

- Classes
  - 2 hour lecture, Monday, 12:00 - 14:00
  - 2 hour studio, as per Allocate+
- Outside class
  - Reading, assignments and self-learning
  - Note: you will be expected to learn R programming
- Text: Ross, S.M. (2014) Introduction to Probability and Statistics for Engineers and Scientists, 5th ed. Academic Press.

Subject Schedule & Assessment

Week	Topics	Assessment
1	Introduction, Modelling, Random Variables	
2	Expectations, Probability Distributions	
3	Sampling, Parameter Estimation and Bias	
4	Confidence Intervals	Ass. #1 Due (10%)
5	Hypothesis Testing	
6	Linear Regression	
7	Classification and Logistic Regression	
8	Model Selection and Penalized Regression	Ass. #2 Due (20%)
9	Trees and Nearest Neighbour Methods	
10	Introduction to Unsupervised Learning	
11	Simulation Based Statistical Methods	Ass. #3 Due (20%)
12	Revision	

- There is also an examination worth 50%.

Staff

- Lecturer (Clayton)
  - Dr. Daniel Schmidt (Daniel.Schmidt@monash.edu)
  - Office: 126A, Level 1, 25 Exhibition Walk
  - Consultation: Monday 10:00 – 11:00
- Head Tutor (Clayton)
  - Mr. Dang Nguyen (dan.nguyen2@monash.edu)
  - Consultation: Tuesday 12:00 – 13:00
- Tutors (Clayton)
  - Mr. Arnil Gurbaz
  - Mr. Van Nguyen
- Communication
  - Please make use of the forum as much as possible
  - Email subject must start with "FIT2086: ..."
    - Otherwise, risk email being missed
    - I will endeavour to reply to emails within two working days

Studios

- You must prepare beforehand
  - Studio material will be released before the studio is to be run
  - Based on material covered in the current week's lecture
  - Will be using R, but we will not be teaching R programming
- The basic idea behind the studios is:
  - to examine a little theory in more depth;
  - to get some hands-on experience analysing data;
  - to use computational techniques to understand concepts.
- To do well at this unit:
  - Complete all studio exercises;
  - Revise lecture material from provided readings;
  - Start on your assignments early.

What this unit is about

- Technical overview of Data Science
  - Exposure to variety of models/methods for data science
  - Some hands-on experience with data analysis
  - Gain an understanding of data and probabilistic models
- NOT learning in depth each model, method introduced
- NOT becoming an R expert
- Realistic goals for students:
  - Familiarization with basics of a few tools
  - Learning advantages/disadvantages of main techniques/models
  - Practice data analysis
  - Exposure to fundamental ideas behind data analytic tools

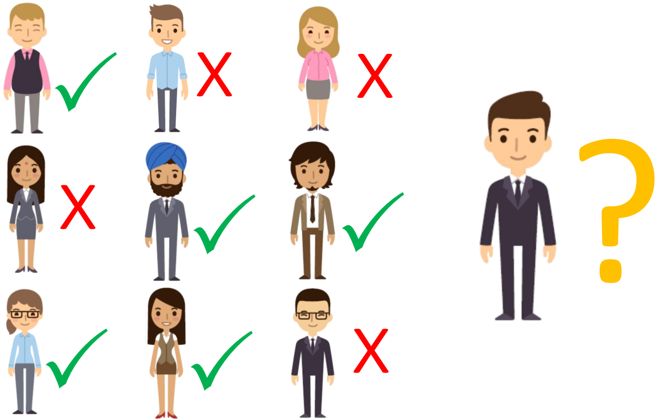
Marks and Hurdles

- Important information!
- To pass FIT2086 you must obtain:
  - 40% or more in the exam; and
  - 40% or more in the assignments; and
  - an overall unit mark of 50% or greater.
- If you get less than 40% for either exam or assignments, and the total mark is:
  - equal to or greater than 50%, a mark of 49-N will be recorded.
  - less than 50%, then the actual mark will be recorded.
- Remember: plagiarism is a serious academic offense; you can be expelled from the university.

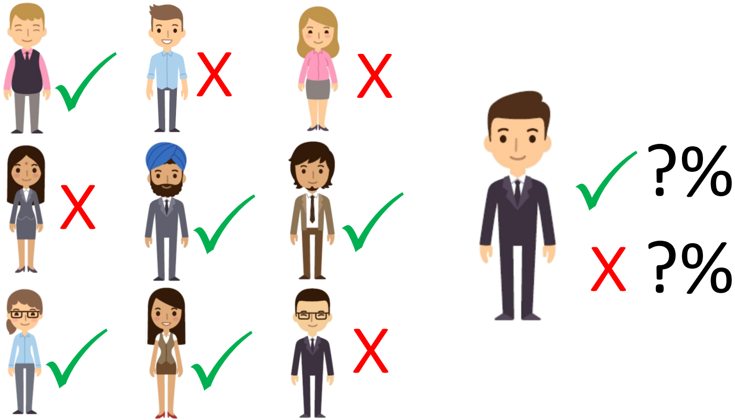
Models

- A model is an object that represents something else
  - A model airplane, a model of a building
- Data science models are mathematical or algorithmic representations
- Models are neither correct, nor incorrect: but they can be more, or less useful for different purposes
  - One model aircraft might accurately represent the relative dimensions of the wings and body
  - An alternative model might more accurately capture the aerodynamic behaviour
- Let's take a quick tour of some models used in data science...

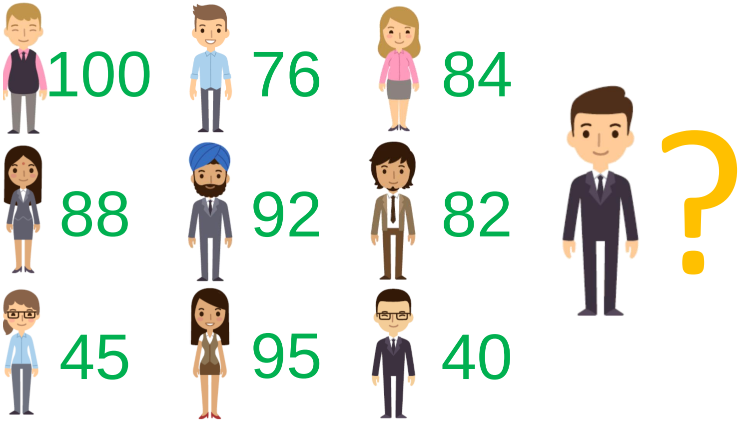
Classifiers



Probabilistic classifiers

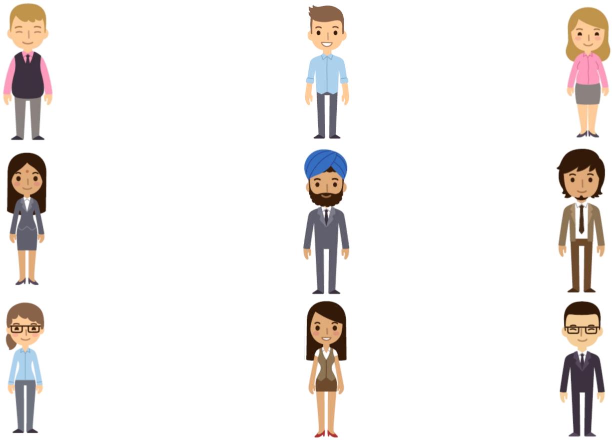


Regression

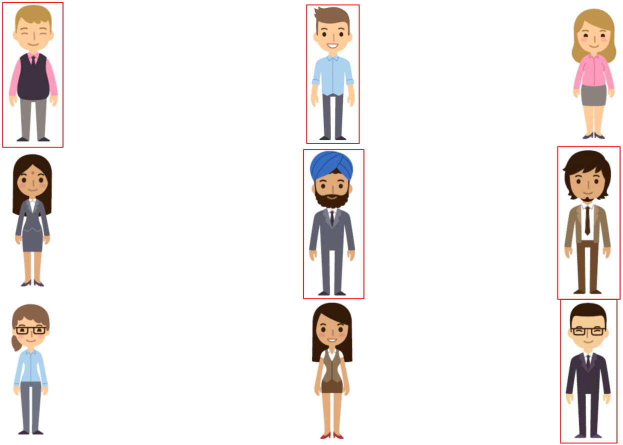


Forecasting

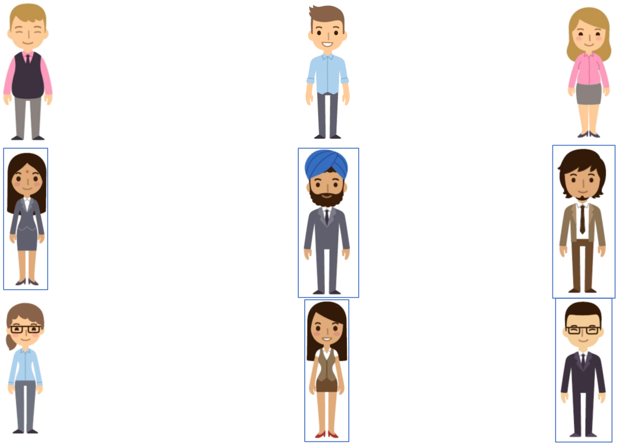
Clustering



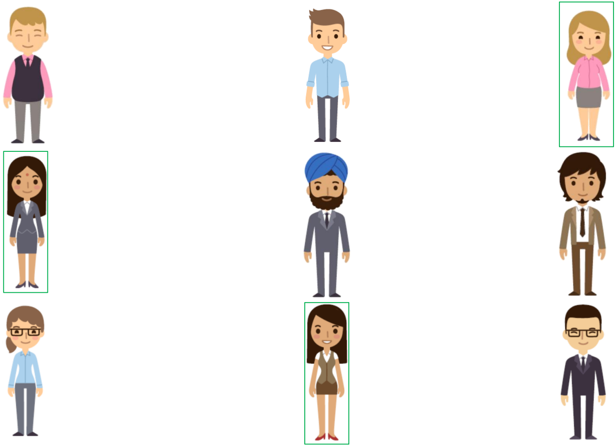
Clustering



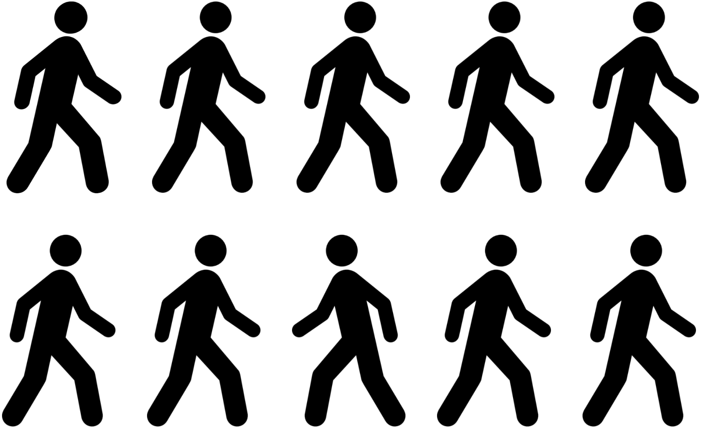
Clustering



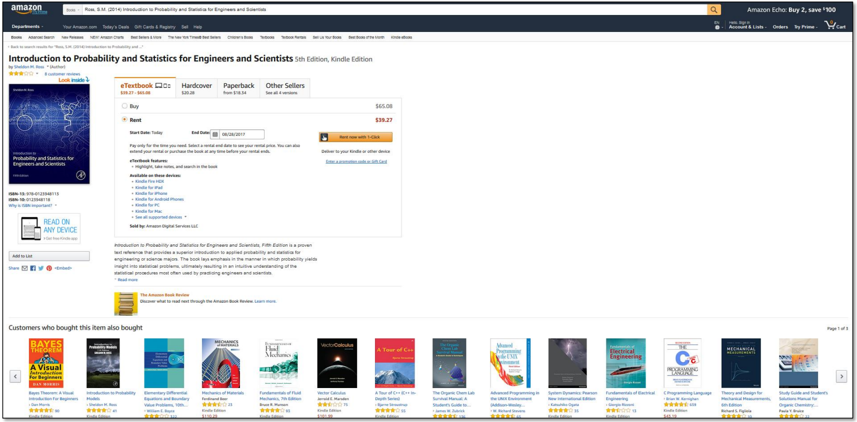
Clustering



Anomaly Detection



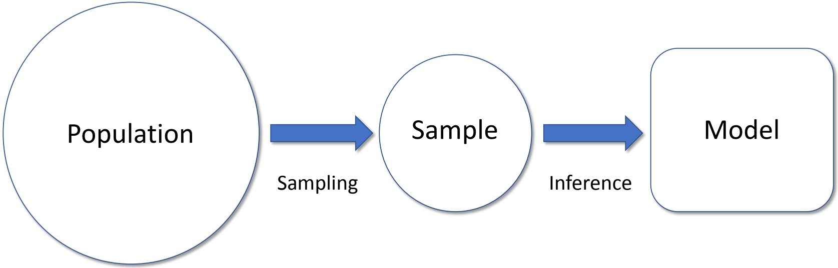
Recommendation Systems



Some Important Terms

- Population:
  - A large collection of objects/items with measurable attributes
- Sample:
  - A finite number of recordings of attributes of items from a population
- Model:
  - A mathematical or algorithmic description of the population learned/inferred from the sample

From Data to Models



Basic Types of Data

- Categorical-Nominal:
  - Discrete numbers of values, no inherent ordering
  - E.g., country of birth, sex
- Categorical-Ordinal:
  - Discrete number of states, but with an ordering
  - E.g., Education status, State of disease progression
- Numeric-Discrete:
  - Numeric, but the values are enumerable
  - e.g., Number of live births, Age (in whole years)
- Numeric-Continuous:
  - Numeric, not enumerable (i.e., real numbers)
  - E.g., Weight, Height, Distance from CBD
- Quantitative vs Qualitative:
  - Generally, categorical data is qualitative, numeric data is quantitative

Why do we Need Formal Methods for Data Science?

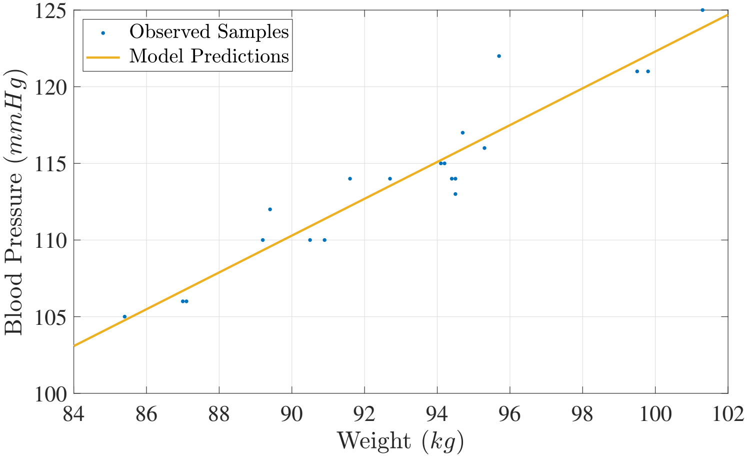
- Consider the following simple example

Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10.0	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87.0	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	56	95.7	2.09	7.0	75	99

- Task: knowing weight, can we build a model for blood pressure?

A Simple Model (1)

- We could "build" the following model



A Simple Model (2)

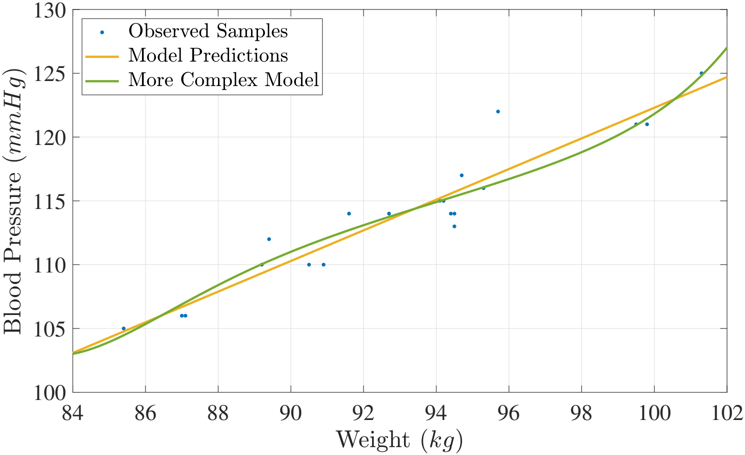
- More formally, our model is the equation:

$$\text{bp} = 1.2 \times \text{weight} + 2.2 + \text{error}$$

- This model relates a person's blood pressure to their weight
  - The relationship is linear (a straight line)
  - The coefficients were learned directly from the data
- The "error" term accounts for the discrepancy between the model predictions and the measured data points
  - We handle this error by treating it as a random quantity

A Simple Model (3)

- We could build the more complex model:



$\Longrightarrow$ fits the sample better – but is it a better model of reality?

Formal Data Science Methods

- Formal data science methods let us ...
  1. Find the coefficients of our straight line in an objective fashion
     - "Parameter estimation", learning a model
  2. Answer the question as to which of the two models we looked is the better description of the population
     - The more complex model fit the sample better, but is it warranted?
  3. Examine many variables simultaneously to find complex relationships
     - Not really possible "by hand"

Outline

Why probability and random variables? (1)

- The central quantity in data science is the data we have observed (our sample)
- We use the language of probability to describe our data
  - We treat the recorded values as realisations of random variables
- But why should we treat them as random?

Why probability and random variables? (2)

- Randomness due to experimental/measurement error
- The measurements/recordings of the observations are corrupted by some intrinsic random measurement/experimental error
- Example: measurement of voltage using commodity level voltmeter
    - Measure the voltage
    - But repeated measurements will yield slightly different results

Why probability and random variables? (3)

- Randomness due to unmeasured factors
- In this setting a measured variable could be (almost) deterministically predicted from other variable(s)
  - If these variables are not recorded, the changes in the measured variable will appear random
- Example: the temperature of water in the shower as other taps in the building are switched on and off
  - If we knew when the taps switched on, we could predict the fluctuations
  - Without this information, the changes in temperature appear random
- But even if we had this knowledge, randomness would remain due to more unmeasured factors

Why probability and random variables? (4)

- Randomness due to sampling
- A finite (but large) population of items, with well measured attributes
  - We cannot measure them all, so we select a sample of these
  - If the selection is done at random, the observations we record behave like realisations of a random process
- Example: estimating average height
  - Imagine our population of interest is all the students in this lecture
  - Select 10 students at random and measure their height
  - The particular heights recorded will vary randomly from sample to sample

Some important notation – refresher

- We will use several bits of set notation in this lecture
  - We use $\{a, b, c\}$ to denote a set with elements a, b and c
  - We use $x \in \mathcal{X}$ to denote that x is an element of the set $\mathcal{X}$
    - Example: $3 \in \{1, 2, 3, 4, 5\}$
  - We use $A \subseteq \mathcal{X}$ to denote that A is a subset of the set $\mathcal{X}$
    - Example: $\{2, 3, 4\} \subseteq \{1, 2, 3, 4, 5\}$

- Some important sets:
  - $\mathbb{Z}$ is the set of all integers;
  - $\mathbb{Z}_+$ is the set of non-negative integers;
  - $\mathbb{R}$ is the set of all real numbers;
  - $\mathbb{R}_+$ is the set of non-negative numbers.

Random Variables (1)

- A random variable (RV) is a variable that takes on a value from a set of possible values with specified probabilities
  - We can let $\mathcal{X}$ denote the possible set of values
  - For now, let's just consider cases where $\mathcal{X}$ is discrete
- We often use capital letters to denote a random variable
- Example: let X be a random variable over $\mathcal{X} = \{1, 2, 3\}$ with:

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases},$$

Random Variables (2)

- A *realisation* of a random variable is a particular value from $\mathcal{X}$ drawn at random
- Consider our example distribution over $\mathcal{X} = \{1, 2, 3\}$ with:

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/4 \end{cases},$$

- Twenty-two sample realisations are:

$$3, 3, 1, 3, 2, 1, 1, 1, 2, 3, 3, 2, 1, 3, 3, 2, 1, 2, 1, 2, 1, 1$$

- There are nine 1s, six 2s and seven 3s
  - We would expect 1s to appear more frequently the more realisations we take

Probability Distributions (1)

- We use the language of probability distributions to describe random variables
- The notation

$$\mathbb{P}(X = x), \; x \in \mathcal{X}$$

  describes the probability that the RV X takes on the value x from $\mathcal{X}$.
- We can use this notation to describe the example random variable X from the previous slides

$$\mathbb{P}(X = 1) = 1/2, \;\; \mathbb{P}(X = 2) = 1/4, \;\; \mathbb{P}(X = 3) = 1/4$$

Probability Distributions (2)

- Review of facts regarding probability distributions
- Fact 1: A probability distribution satisfies:

$$\mathbb{P}(X = x) \in [0, 1] \text{ for all } x \in \mathcal{X}$$

and

$$\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) = 1$$

Probability Distributions (3)

- Fact 2: The probability of $(X \in A_1 \text{ OR } X \in A_2)$, with $A_1, A_2 \subset \mathcal{X}$

$$\mathbb{P}(X \in A_1 \cup A_2) = \mathbb{P}(X \in A_1) + \mathbb{P}(X \in A_2) - \mathbb{P}(X \in A_1 \cap A_2),$$

  with "$\cap$" set intersection and "$\cup$" set union
- Example: If X follows the probability distribution

$$\mathbb{P}(X = 1) = 1/2, \;\; \mathbb{P}(X = 2) = 1/4, \;\; \mathbb{P}(X = 3) = 1/4$$

  then $\mathbb{P}(X \geq 2)$ is

$$\begin{aligned} \mathbb{P}(X \in \{2\} \cup \{3\}) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) \\ &= 1/4 + 1/4 \\ &= 1/2 \end{aligned}$$

Probability Distributions of Two RVs (1)

- Now let us consider the case of two RVs $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$
  - $\mathcal{X}$ and $\mathcal{Y}$ are the sets of values X and Y can take, respectively
  - $\mathcal{X} \times \mathcal{Y}$ is the set of values the pair can assume

- Example: If $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{Y} = \{1, 2\}$, then

$$\mathcal{X} \times \mathcal{Y} = \{\{1,1\}, \{2,1\}, \{3,1\}, \{1,2\}, \{2,2\}, \{3,2\}\}$$

- Example: An example distribution over $\mathcal{X} \times \mathcal{Y}$:

	$X = 1$	$X = 2$	$X = 3$
$Y = 1$	0.05	0.15	0.1
$Y = 2$	0.25	0.15	0.3

Probability Distributions of Two RVs (2)

- We can define a probability distribution over (X, Y) as before:

$$\mathbb{P}(X = x, Y = y) \in [0, 1] \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

which satisfies

$$\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y) = 1$$

- $\mathbb{P}(X = x, Y = y)$ is the *joint* probability of $X = x$ and $Y = y$
  - That is, the probability of $X = x$ AND $Y = y$

- Example: The example distribution from previous slide

$$\begin{aligned} \mathbb{P}(X = 1, Y = 1) &= 0.05 \\ \mathbb{P}(X = 1, Y = 2) &= 0.25 \\ \mathbb{P}(X = 2, Y = 1) &= 0.15 \end{aligned}$$

and so on.

The Sum Rule (1)

The Sum Rule

The sum rule is given by:

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} P(X = x, Y = y)$$

The probability $\mathbb{P}(X = x)$ is called the *marginal* probability.

- The marginal probability $\mathbb{P}(X = x)$ is the probability of seeing $X = x$ irrespective of what value Y takes on

The Sum Rule (2)

- Example:

	$X=1$	$X=2$	$X=3$
$Y=1$	0.05	0.15	0.1
$Y=2$	0.25	0.15	0.3

- Then

$$\begin{aligned}\mathbb{P}(Y=1) &= 0.05+0.15+0.1=0.3\\ \mathbb{P}(Y=2) &= 0.25+0.15+0.3=0.7\end{aligned}$$

so that the probability of seeing a $Y=2$ is significantly higher than the probability of seeing a $Y=1$, irrespective of the value of X.

Conditional Probability (1)

Conditional Probability

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

The probability $\mathbb{P}(X = x \,|\, Y = y)$ is called the probability of $X = x$, conditional on $Y = y$.

- The conditional probability $\mathbb{P}(X = x \,|\, Y = y)$ is the (joint) probability of seeing $X = x$ and $Y = y$, divided by the (marginal) probability that we have observed $Y = y$.

Conditional Probability (2)

- Example:

	$X=1$	$X=2$	$X=3$
$Y=1$	0.05	0.15	0.1
$Y=2$	0.25	0.15	0.3

- Then

$$\begin{aligned}\mathbb{P}(X=1 \,|\, Y=1) &= \mathbb{P}(X=1, Y=1)/\mathbb{P}(Y=1) \\ &= 0.05/0.3 \approx 0.1667\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X=1 \,|\, Y=2) &= \mathbb{P}(X=1, Y=2)/\mathbb{P}(Y=2) \\ &= 0.25/0.7 \approx 0.3571\end{aligned}$$

so that seeing $X=1$ is twice as likely when $Y=2$ as compared to the case that $Y=1$.

Independent Random Variables (1)

- Independent random variables are very important
- X and Y are considered independent if

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

  for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.
- This implies that

$$\mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x).$$

  $\Rightarrow$ Knowing about Y tells us nothing new about X
- An even more special class are independent and identically distributed (i.i.d.) random variables
  - $X_1 \in \mathcal{X}$, $X_2 \in \mathcal{X}$ are i.i.d. if they are independent and

$$\mathbb{P}(X_1 = x) = \mathbb{P}(X_2 = x) \text{ for all } x \in \mathcal{X}$$

Continuous Random Variables (1)

- So far we have considered only discrete random variables
- The ideas extend to the case that the values X can take on form a continuum, that is, $\mathcal{X} \subseteq \mathbb{R}$
- X now follows a probability density function (pdf) $p(x)$.
- A pdf satisfies:

$$p(x) \geq 0 \text{ for all } x \in \mathcal{X}$$

and

$$\int_{\mathcal{X}} p(x)dx = 1$$

Continuous Random Variables (2)

- The probability that X lies in an interval (a, b) is

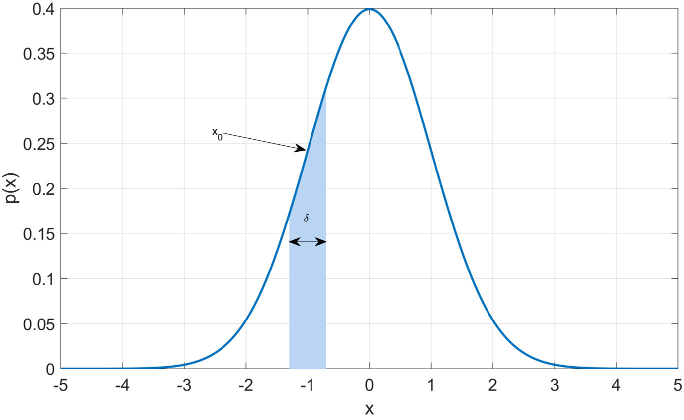
$$\mathbb{P}(a < X < b) = \int_a^b p(x)dx.$$

- More generally, the probability $X \in A$, where $A \subset \mathcal{X}$ is

$$\mathbb{P}(X \in A) = \int_A p(x)dx.$$

- This implies that $\mathbb{P}(X = x) = 0$
  $\Rightarrow$ One of the most confusing aspects of continous RVs

- Example: Probability of $(x_0 - \delta/2 < X < x_0 + \delta/2)$



Continuous Random Variables (4)

- Define the interval $A_\delta = (x_0 - \delta/2,\, x_0 + \delta/2)$ centered on x_0
- From the rules of probability we have

$$
\begin{aligned}
\mathbb{P}(x \in A_\delta) &= \int_{x_0-\delta/2}^{x_0+\delta/2} p(x)dx \\
&= \left[\int p(x)dx\right]_{x=x_0+\delta/2} - \left[\int p(x)dx\right]_{x=x_0-\delta/2}
\end{aligned}
$$

where $\int p(x)dx$ denotes the indefinite integral of $p(x)$

- It is clear that as $\delta \to 0$
  1. the interval $A_\delta \to x_0$ and
  2. $\mathbb{P}(x \in A_\delta) \to 0$

Continuous Random Variables (5)

- Consider a pdf of two continuous RVs, say X and Y
  - Use the shorthand notation $p(X = x, Y = y) \equiv p(x, y)$
- Then we have continuous analogues of the sum rule

$$p(x) = \int p(x, y) dy$$

and the conditional probability rule

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

$\Longrightarrow$ go back and compare to discrete versions

Cumulative Distribution Functions (1)

- The cumulative distribution function (cdf) of a continuous RV is:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(x')dx'$$

  that is, the probability that X is less than some value x

- Let's introduce some shorthand notation for discrete RVs:

$$\mathbb{P}(X = x) \equiv p(x)$$

- Then, if X is a discrete RV over the integers (or a subset)

$$\mathbb{P}(X \leq x) = \sum_{x' \leq x} p(x')$$

- It follows that

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$$

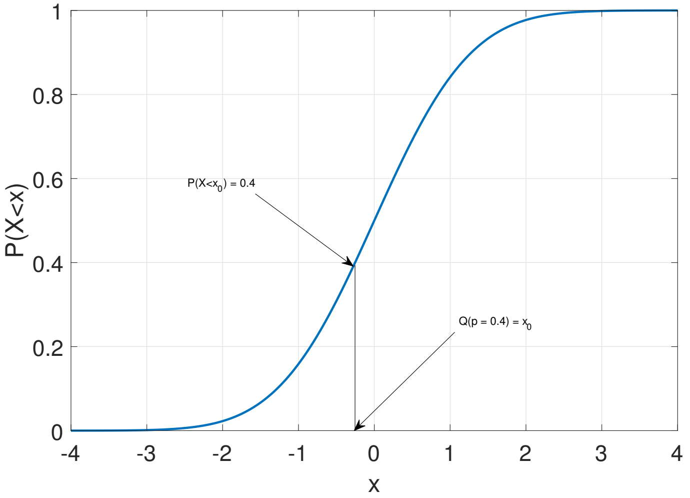
Cumulative Distribution Functions (2)

- The inverse cdf is

$$Q(p) = \{x \in \mathcal{X} : \mathbb{P}(X \leq x) = p\}$$

  which is sometimes called the quantile function.
- In words, the quantile function says: find the the value x such that the probability that $X \leq x$ is p
- For example:
  - $Q(p = 1/2)$ is the median;
  - $Q(p = 1/4)$ is the first quartile; and
  - $Q(p = 3/4)$ is the third quartile.

Cumulative Distribution Functions (3)



Reading/Terms to Revise

- Reading for this week: Chapter 4 of Ross.
- Terms you should know:
  - Random variable;
  - Conditional Probability;
  - Probability density function;
  - Cumulative distribution function