

Overview: Probability Axioms, Bayes' Theorem and Probability Distributions

In this week, we will be walking through basic probability axioms which are used in calculating probabilities. This includes Bayes' theorem, which is a well-known theorem in the Bayesian approach in statistics. We will be concluding this week's lesson with a quick look at selected probability distributions which are important in statistical methods for data science.

Besides introducing fundamental statistical concepts, this week's content also lays the foundation for the topics which we will be covering in the coming weeks. For example, knowledge on the characteristics of normal probability distribution will help you to understand the underlying assumptions of linear regression models which are one of the basic models used in predictive modelling.

1.1 Descriptive Statistics

Introduction

We begin this week with statistical concepts that are used in the first step in a data science project - exploratory data analysis (EDA). Data can come from many sources, for example survey questionnaires, transactional records, IoT (Internet of Things) sensor measurements, social media posts, Twitter text posts and YouTube videos just to name a few. During data analysis, large amount of data can be summarized and presented using descriptive statistics to be used to inform subsequent steps in the project. In the previous week, we have seen that variables can be classified as either qualitative variable or quantitative variable. Knowledge on type of the variable is important to help determine appropriate choices for visualization, data analysis, or statistical modelling.

Univariate Descriptive Statistics

Measures of Central Tendency and Dispersion

Descriptive statistics consist of two types – measures of central tendency and measures of dispersion. Measures of central tendency indicate the typical value or location of the distribution. Commonly used measures of central tendency are mean, median and mode. On the other hand, measures of variability or dispersion indicate the variation or spread of the individual values. Examples of measures of dispersion are standard deviation and mean absolute deviation. An illustration on some measures of central tendency and dispersion as well as how they complement each other in describing a data set is described in the following video by Simple Learning Pro (2015).

Mode, Median, Mean, Range and Standard Deviation (1.3)

<https://youtu.be/mk8tOD0t8M0>

Source: (Simple Learning Pro, 2015)

The choice of descriptive statistics to be used depends on the characteristics or intrinsic meaning of the variables. Outliers and extreme values can affect measures of central tendency (i.e., the mean) and dispersion (i.e., variance or standard deviation). Many introductory statistics textbooks discuss the relationship between mean, median, mode and skewness. However, we would like to put forth a note of caution here as this relationship does not always hold (Hippel, 2005).

To eliminate the influence of outliers and extreme values, we can use the trimmed mean in which a fixed number of sorted values at each end is dropped and the average of the remaining values is taken. For example, the final score in international diving events is calculated using the trimmed mean. For measures of dispersion, a robust estimate of variability is the median absolute deviation from the median given as:

Median absolute deviation = $Median(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$

where m is the median.

The measures of dispersion that we have discussed so far are computed in relation to the measures of central tendency. Another approach of discussing dispersion is by looking at statistics which are computed based on the sorted data, also known as order statistics. An example would be the range, which is simply the difference between the largest and smallest values in the data set. Obviously, the range is extremely sensitive to outliers hence is not a good measure of dispersion.

An alternative would be to use the p -th percentile, which is a value such that at least p percent of the values is less than or equal to this value and at least $(100 - p)$ percent of the values are more than or equal to this value. A closely related concept is the quantile, with quantiles indexed by fractions (hence 0.9 quantile is the same as the 90th percentile). The percentile is also used to summarize the tails of the distribution, for example the term one-percenter in popular culture refers to the wealthiest 1% (99th percentile) of people in the United States. An explanation on quantiles and percentiles is given in the following video by StatQuest with Josh Starmer (2017).

Quantiles and Percentiles, Clearly Explained!!!

<https://www.youtube.com/watch?v=IFKQLDmRK0Y>

Source: ([StatQuest with Josh Starmer](#), 2017)

For categorical or qualitative variables, the mode or percentages are common descriptive measures.

Getting Started with R

Traditionally, statistical measures such as mean and standard deviation that we discuss in this topic are calculated on paper with the assistance of desktop calculators and statistical tables. Recent advances in statistical computing have seen the emergence of statistical software such as SPSS, SAS and [R](#). These software enables complex computation, modelling and statistical inference to be performed within a relatively short time and allows the data scientist or statistician to shift the

focus from computation to the design, build and interpretation of models to analyze the data.

If you are new to R, the following video will help you get started with R.

R Programming Tutorial - Learn the Basics of Statistical Computing

<https://www.youtube.com/watch?v= V8eKsto3Ug>

Source: ([freeCodeCamp.org](https://www.freecodecamp.org), 2017)

The following R code snippet illustrates how we can compute simple descriptive statistics using the `summary()` function which produces a numerical summary of each variable in a particular data set. To request for a trimmed mean and percentiles discussed above, we can use the `mean()` and `quantile()` functions, respectively. The `hist()` function plots a histogram.

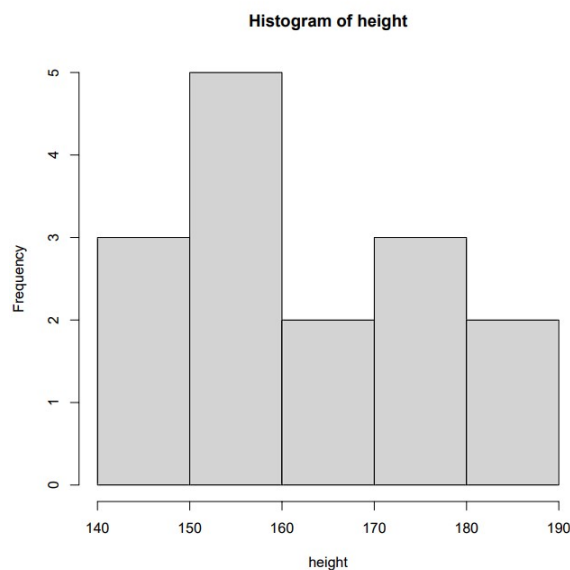
```
► Run
1 height <- c(151, 174, 141, 186, 168, 146, 179, 163, 152, 145, 160, 172, 153, 155, 141)
2 summary(height)
3 mean(height, trim=0.1)
4 quantile(height,0.1)
5 hist(height)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
141.0	151.5	160.0	161.9	173.0	186.0

```
[1] 161.6154
10%
145.4
```

Rplots.pdf

✓ Program exited with code 0



1.1.2 Multivariate Descriptive Statistics

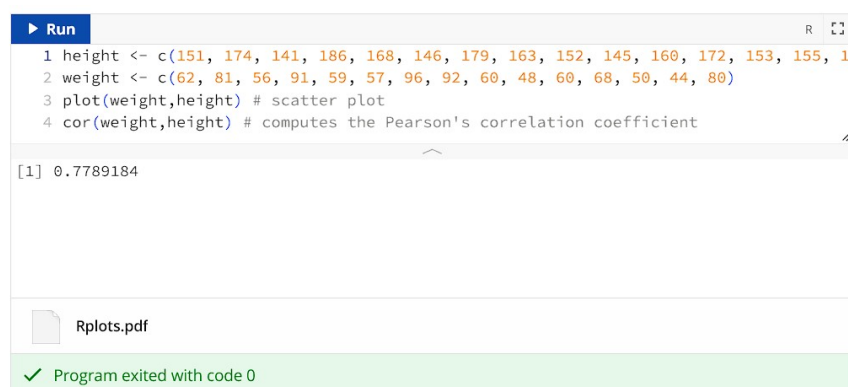
Two Quantitative Variables: Correlation

Besides measures of central tendency and dispersion, exploratory data analysis also involves examining correlation between the variables. Two quantitative variables X and Y are said to be positively correlated if high values of X correspond with high values of Y and similarly low values of X correspond with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.

Correlation is measured using correlation coefficient with a range of values between -1 (perfect negative correlation) and 1 (perfect positive correlation). The sign of the coefficient indicates the direction of the correlation (either positive or negative) and the magnitude denotes the strength of the correlation. *Pearson's correlation coefficient* between two variables X and Y is computed using the formula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

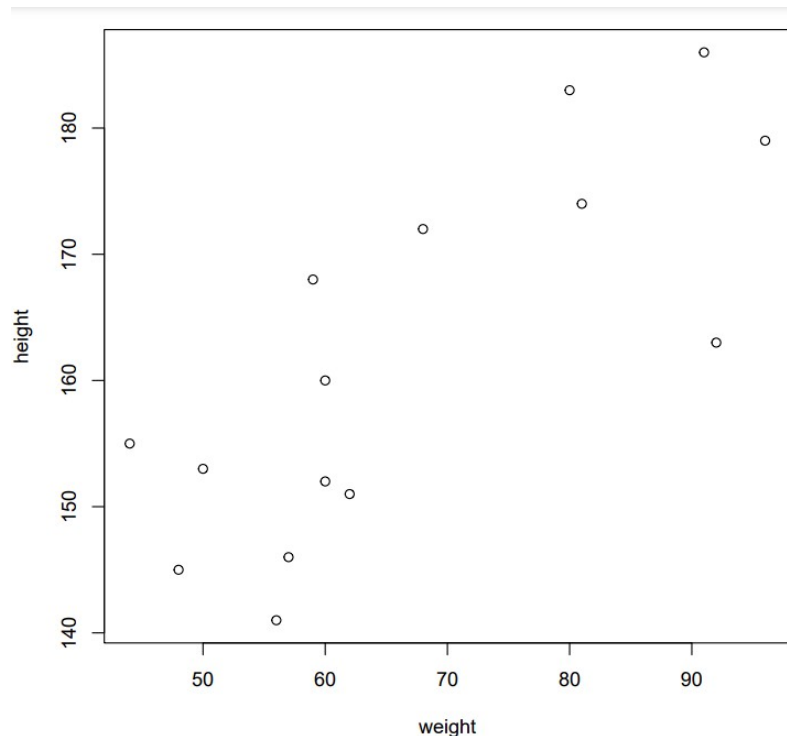
Where x_i and y_i are the observed values for X and Y, respectively, \bar{x} and \bar{y} are the respective sample means and s_x and s_y are the respective sample standard deviations. A word of caution here is that correlation coefficient may not be meaningful if the relationship between the variables is in fact non-linear. If the number of observations is relatively small, a scatter plot will be useful here to visualize the structure as well as the strength of the relationship. In R, the `plot()` function produces scatter plots for quantitative variables whereas the `cor()` function computes Pearson's correlation coefficient.



```
► Run R [3]
1 height <- c(151, 174, 141, 186, 168, 146, 179, 163, 152, 145, 160, 172, 153, 155, 141)
2 weight <- c(62, 81, 56, 91, 59, 57, 96, 92, 60, 48, 60, 68, 50, 44, 80)
3 plot(weight,height) # scatter plot
4 cor(weight,height) # computes the Pearson's correlation coefficient

[1] 0.7789184

Rplots.pdf
✓ Program exited with code 0
```



If the linear relationship is validated through the scatter plot, a Pearson's correlation coefficient of 0.7789 indicates a strong positive correlation between weight and height in the example above.

Other types of correlation coefficients such as Spearman's rho and Kendall's tau are based on the rank of the data thus, they are more robust towards outliers. However, usually Pearson's correlation coefficient and its robust alternatives are used. Rank-based coefficients are usually used for smaller data sets and specific hypothesis tests.

We would like to stress here that correlation does not imply causation although there are many times that it does. Watch the following video by One Minute Economics (2019) for an illustration on this.

Correlation Does Not Imply Causation: A One Minute Perspective on Correlation vs. Causation

<https://youtu.be/mQfacqVvOEM>

Source: (One Minute Economics, 2019)

Two Categorical Variables: Association

A contingency table can be used to summarize two qualitative or categorical variables. An $m \times n$ table is a contingency table with m rows and n columns. The columns and rows of the contingency table represent the categories of each variable, respectively. Each cell in a contingency table shows the number of individuals or observations that fall in each combination of the corresponding categories. For example, in the following video explanation on contingency table chi-square test by Khan Academy (2010), the contingency table is a 2×3 table. Since we are not touching on hypothesis tests yet at this point, you may stop viewing at 2:07 of the video.

Contingency table chi-square test | Probability and Statistics | Khan Academy

<https://youtu.be/hpWdDmgsIRE>

Source: (Khan Academy, 2010)

The *chi-square measure* or *Cramer's V* are commonly used to measure the strength of association between two qualitative variables.

Let us take a quick look on performing categorical data analysis in R, using a small data set of observations on **gender** (gender of the person; 0 = male, 1 = female) and **coffee.yes** (whether the person drinks coffee; 0 = no, 1 = yes). The **data.frame()** function is used here to create a data.frame object in R. The **table()** function is used to construct a contingency table for the categorical variables. Hypothesis test on association between the variables will be discussed in next week's topic.

```
► Run R [ ]
1 gender <- c(1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0)
2 coffee.yes <- c(0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1)
3 df <- data.frame(gender, coffee.yes)
4 table(df)

      coffee.yes
gender 0 1
  0 4 3
  1 3 5

✓ Program exited with code 0
```

Cramer's V is not included in the base R package, hence R package installation is required. R package installation can be easily done if you have RStudio installed on your local PC.

Categorical and Numeric Data

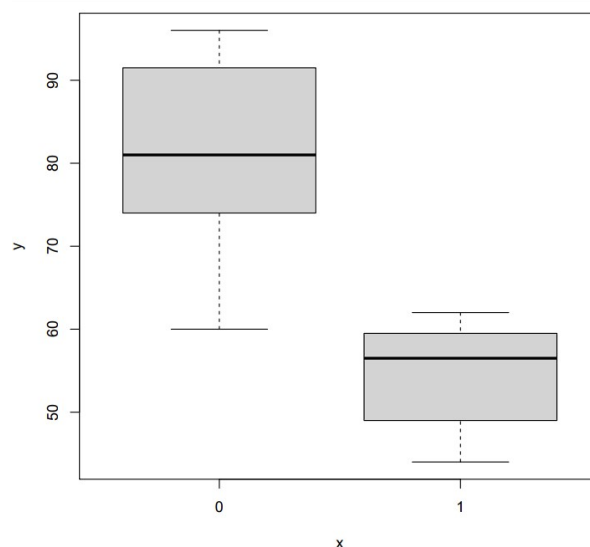
In some data science problems, we would like to describe the relationship between a categorical and a quantitative variable. Besides being used to detect outliers, boxplots can also visually compare the distributions of a quantitative variable grouped according to a categorical variable. In a boxplot, outliers are data points that are too far above or below the box boundaries. The definition of “too far” here is stated as “more than 1.5 times the interquartile range.”

The `plot()` function in R automatically produces a boxplot if the variable on the x-axis is qualitative. A code example is shown below - the `as.factor()` function converts the `gender` variable which was initially quantitative into qualitative.

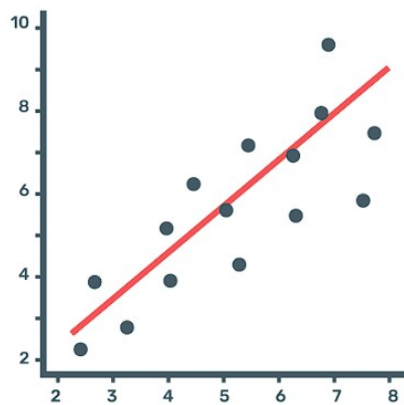
```
► Run
1 weight <- c(62, 81, 56, 91, 59, 57, 96, 92, 60, 48, 60, 68, 50, 44, 80)
2 gender <- c(1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0)
3 gender <- as.factor(gender)
4 plot(gender, weight)
```

Rplots.pdf

✓ Program exited with code 0

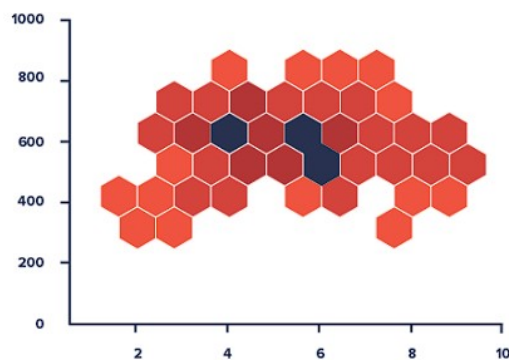


Descriptive Plots

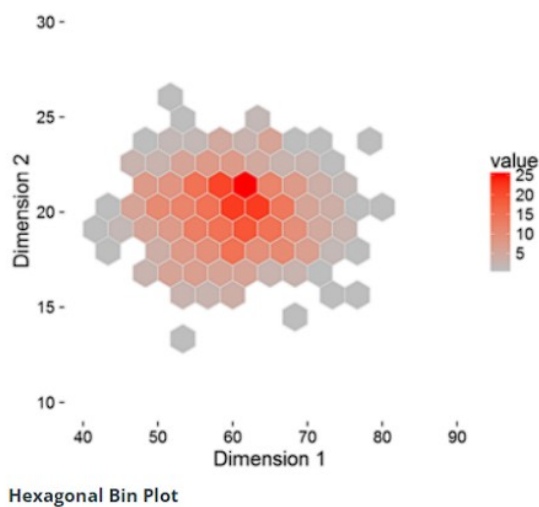


Source: Scatter plot (Adobe stock, n.d.)

For large data sets, a scatterplot will be too dense and not informative. Instead of plotting individual points, a hexagonal binning plot can be used whereby the records are grouped into hexagonal bins. The hexagons are then plotted with a colour indicating the number of records in that bin.



Source: Hexagonal binning plot (datavizproject.com, n.d.)



Source: Hexagonal bin plot (Mike Carlo, 2016)

Other alternatives are contour plot and heat maps.



Source: Heat map (Adobe stock, n.d.)

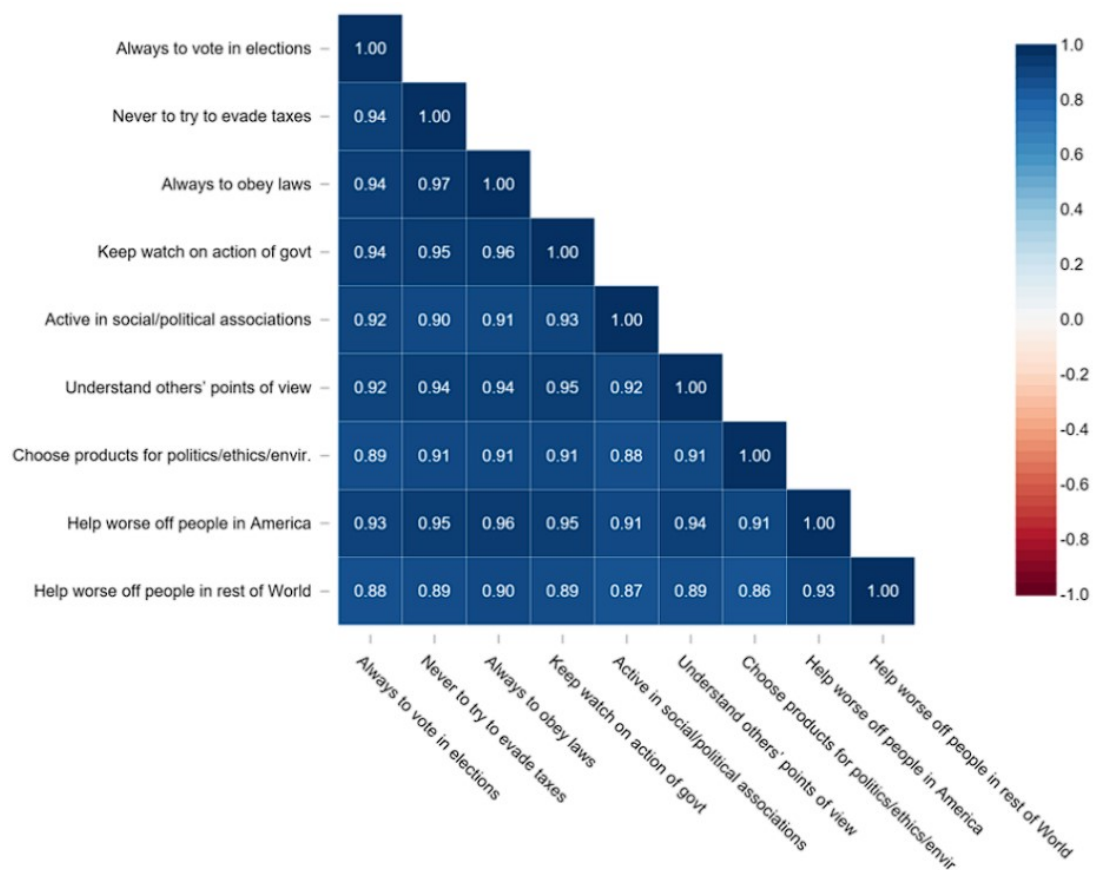
Cutomer retention by user age

		Periods out																
cohort	first_period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Apr 28, 2014	79	22%	19%	13%	19%	16%	23%	19%	20%	11%	14%	16%	10%	10%	10%	9%	6%	6%
May 5, 2014	168	23%	21%	21%	24%	24%	29%	24%	18%	22%	14%	14%	12%	13%	10%	10%	7%	
May 12, 2014	188	19%	19%	13%	21%	19%	20%	24%	21%	16%	14%	13%	10%	9%	9%	7%		
May 19, 2014	191	23%	21%	22%	22%	26%	27%	29%	26%	21%	21%	17%	15%	10%	6%			
May 26, 2014	191	21%	16%	20%	24%	27%	23%	20%	19%	15%	15%	12%	12%	6%				
Jun 2, 2014	184	24%	24%	24%	24%	21%	21%	18%	20%	16%	15%	18%	7%					
Jun 9, 2014	182	19%	16%	25%	19%	23%	28%	22%	18%	13%	10%	5%						
Jun 16, 2014	209	24%	20%	24%	22%	23%	17%	18%	15%	13%	7%							
Jun 23, 2014	217	22%	19%	19%	20%	20%	17%	19%	18%	12%								
Jun 30, 2014	221	18%	18%	24%	24%	23%	19%	20%	8%									
Jul 7, 2014	203	24%	23%	18%	16%	24%	22%	16%										
Jul 14, 2014	188	24%	18%	20%	18%	21%	10%											
Jul 21, 2014	228	19%	14%	14%	14%	7%												
Jul 28, 2014	204	14%	12%	15%	11%													
Aug 4, 2014	230	22%	17%	11%														
Aug 11, 2014	245	16%	8%															
Aug 18, 2014	252	12%																

Source: Heat map (mode.com, n.d.)

Correlation Matrix

A correlation matrix is used to present the correlation coefficients for a set of variables in a concise manner. In the plot below, the number in each box indicates the correlation coefficient between the respective pairs of variables. For example, the correlation coefficient between "Always to obey laws" and "Always to vote in elections" is 0.94.



Source: *What is a Correlation Matrix?* (Tim Bock, n.d.)

For the interested reader, the [R Graph Gallery](#) contains a collection of beautiful plots for visualizing different types of data and variables, which can be produced using R's various packages.

1.2 Introduction: Probability Axiom

Introduction

To quantify the chances of an event occurring or to predict an unknown outcome in a scenario, the concept of probability is used. A question of interest could be, for example, "What is the probability that a house is sold at a price of more than RM500,000 given pre-determined conditions of the house?" Probability is a numerical value between 0 and 1, inclusive. An event with probability value of 0 is said to be an impossible event, whereas an event with probability of 1 is certain to happen. The concept of probability is inherent in various applications from diverse areas such as stock market prediction to spam filtering.

Let us start with some basic terminologies in probability theory first.

Basic Terminologies

In probability theory, an event is a collection of outcomes from an experiment which we are interested in. In such an experiment, the set of all possible outcomes (also known as sample space) is known but we do not know the outcome in advance.

Example 1: E-mail spam filtering

In this example, spam filtering is the experiment with the two possible (binary) outcomes: (1) an e-mail is a spam e-mail, and (2) an e-mail is not a spam e-mail. This kind of scenario in which the possible outcomes are binary or categorical and the algorithm is dealing with classification problems.

How Email Spam Filters Work Based On Algorithms | Mach | NBC News

<https://youtu.be/iOxylrN4Io>

Source: ([NBC News](#), 2017)

Example 2: Measuring lifetime of an electronic component

The lifetime or time-to-failure of electronic components is of interest in reliability studies. In this case, the experiment measures the lifetime of individual components and the sample space consists of nonnegative real numbers. The possible outcomes

are on the real number line (e.g. in hours) and a reliability engineer may be interested in predicting the lifetime value using regression analysis.

Example 3: Loan default

In finance, the ability to predict loan default is of high importance for risk management. With the availability of borrower's information and economic indicators, machine learning algorithms are able to predict the probability of a borrower defaulting on his/her loan thus enabling the lender to make informed decisions. The experiment in this example refers to the provision of the loan, and there are two possible outcomes in the sample space, i.e. default or non-default.

Independent Events

An independent event is an event that is not affected by other events. For example, consider an experiment of drawing two balls from a box which contains three red balls and three blue balls. The first draw is considered as an event A, and the second draw is another event B. The possible outcomes of each event are either drawing a red ball or a blue ball. If the balls are drawn without replacement, then A and B are dependent because drawing a second ball of a certain colour is dependent on the colour of the first ball being drawn. On the other hand, if the first ball is returned to the box (with replacement) before the second ball is being drawn, then A and B are independent because the colour of the first ball being drawn will not affect the outcome of the colour of the second ball drawn.

We will revisit this kind of experiments after we have defined some probability axioms.

Mutually Exclusive Events

Two events are said to be mutually exclusive if they do not have common outcomes.

Exhaustive Events

A set of events E_1, E_2, \dots, E_k are said to be exhaustive events if one or more of them must occur. For example, an adult population is classified as either smoker or non-

smoker. An adult is randomly selected from this population, and let E_1 denotes the event that the selected adult is a smoker and E_2 denotes the event that the selected adult is a non-smoker. The events E_1 and E_2 are exhaustive because the selected adult must be either a smoker or non-smoker.

Now that we have familiarized ourselves with the basic terminologies, we are now ready to move on to defining probability axioms.

Question 1

Consider an experiment which involves tossing a die twice and tossing a coin once. The events which result from tossing the die twice are independent of the events which result from tossing a coin once.

Ans: True, The result obtained from the toss of the die will not affect the result from the toss of the coin.

Question 2

Consider the event that "it will rain tomorrow" and another event "it will rain today". These two events are independent.

Ans: False, In weather forecasting, the events on two consecutive days are usually modelled as dependent events. Such models are known as Markov chains with the conditional probability matrix given as a transition probability matrix.

Question 3

The event that the stock price of a certain company will increase the next day, is dependent on the event that the stock price of the company will increase today.

Ans: True, Stock price on consecutive days are usually modelled as dependent on each other using time series modelling.

Question 4

In a market survey, 500 participants were randomly selected and recruited to participate in a new product's test use before its official launch. The response of each of the participant is independent of each other.

Ans: True, Since the participants were randomly selected, we can assume that their responses will be independent of each other.

1.2.1 Probability Axioms

In statistics, an event is often denoted using capital letters, such as E. In this case, the probability of the event is denoted as $Pr(E)$. How are the probability values derived?

In general, there are [four perspectives to derive probability values](#) – classical (theoretical), empirical, subjective and axiomatic approaches.

There are several probability axioms or probability rules which are useful for calculating the probability of events. In the subsequent sections, let us start with two events denoted as A and B, respectively.

Complement Rule

The complement of an event A (denoted as A') refers to the collection of outcomes that does not belong to the event.

$$Pr(A') = 1 - Pr(A)$$

Example: There were 300 participants who attended a conference, of which 225 were men and 75 were women. The probability that a randomly selected participant is a male is 0.75, and the probability that a randomly selected participant is a female (complementary event of randomly selecting a male) is 0.25.

Addition Rule

The addition rule can be applied to calculate the probability that either one or both events occur. If the events are mutually exclusive, then

$$Pr(A \text{ or } B) = Pr(A) + Pr(B)$$

If the two events are not mutually exclusive, then

$$Pr(A \text{ or } B) = Pr(A) + Pr(B) - Pr(A \text{ and } B)$$

Example: There were 300 participants who attended a conference, of which 225 were men and 75 were women. One hundred and sixty five of the participants are non-vegetarians, 85 of them are vegetarians and 50 of the participants did not state their meal preference. One hundred and forty five of the participants are non-vegetarian men. The probability that a randomly selected participant is a male or non-vegetarian is $0.75 + 0.55 - 0.4833 = 0.817$.

Multiplication Rule

The multiplication rule is applied to calculate the probability that both events A and B occur. In general,

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

Or

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

Note here that $Pr(B|A)$ denotes the conditional probability of event B occurring given event A has occurred. A special case of the multiplication rule,

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

applies when the events are independent.

Example: A product must go through two stages of inspection before it can pass the quality control test. If the probability that it passes through the first stage and second stage is 0.3 and 0.2, respectively, the probability that the product will pass through the quality control test is $0.3 \times 0.2 = 0.06$. We assume that the two stages are independent of each other (which should be a reasonable assumption in this scenario.)

The addition rule and multiplication rule can be generalized to the case of more than two events.

Probability in Action: Benford's Law

Before we move on to the next topic on Bayes' theorem, let's look at an interesting application of probability through the [use of Benford's law in forensic accounting and network threat detection](#). The following video gives an explanation on this concept.

Number 1 and Benford's Law - Numberphile

<https://www.youtube.com/watch?v=XXjIR2OK1kM>

Source: ([Numberphile](#), 2013)

1.2.2 Odds and Odds Ratio

An important statistical concept for binary categorical variable (I.e., categorical variable with two possible outcomes) is the odds. Let us illustrate with an example based on customer churn. There are two possible outcomes in the categorical variable, which is either the subscriber of a service discontinues from the service or not. The odds of a customer discontinuing is the probability that the customer discontinues from the service divided by the probability that the customer continues using the service.

In general, the odds of an outcome is the proportion of times (or probability) that it happens divided by the proportion of times (or probability) that it doesn't. If the probability that the outcome happens (also known as a "success") is denoted as p , then

$$\text{Odds} = \frac{p}{1 - p}$$

A related concept, odds ratio compares the odds of one group relative to another group. Using the same example, the odds ratio of male discontinuing the service to female is the odds of a male discontinuing the service divided by the odds of a female discontinuing the service. An odds ratio is the odds of an outcome happening under one circumstance divided by the odds in another circumstance.

$$\text{Odds ratio} = \frac{\text{Odds}(\text{Group A})}{\text{Odds}(\text{Group B})}$$

Another example application of odds and odds ratio is explained in the following video by The NCCMT (2016) in the context of public health care.

NCCMT - URE - Odds Ratios

https://youtu.be/5zPSD_eN04

Source: ([The NCCMT](#), 2016)

1.3 Bayes Theorem

Introduction

Bayes' Theorem (also known as Bayes' Rule) was introduced by and named after Reverend Thomas Bayes (1701-1761) and is used to revise or update conditional probabilities with newly acquired information. Bayes' Theorem has been applied in various fields such as machine learning (remember the e-mail spam filtering algorithm?), image processing, linguistics, epidemiology, psychology and marketing (Allenby, Bakken and Rossi, 2004), just to name a few. Machine learning algorithms such as Naïve Bayes' classifier and Bayesian networks applies Bayes' Theorem in its tasks.

Before going into Bayes' theorem, we first need to understand the Rule of Total Probability. Recall the definition of mutually exclusive and exhaustive events which have been introduced in the previous topic. Let us now consider a sample space which has been partitioned into three mutually exclusive and exhaustive events, B_1 , B_2 and B_3 , and an arbitrary event A . Watch the following video by MIT OpenCourseWare (2016) for an explanation on the Rule of Total Probability.

4.2.3 Law of Probability: Video

<https://youtu.be/F3y8qpfFUs>

Source: ([MIT OpenCourseware](#), 2016)

Note that the event A consists of the mutually exclusive events $(B_1 \cap A)$, $(B_2 \cap A)$ and $(B_3 \cap A)$. From the Addition Rule, we have

$$Pr(A) = Pr(B_1 \cap A) + Pr(B_2 \cap A) + Pr(B_3 \cap A)$$

As a summary, the Rule of Total Probability is given as:

$$Pr(B) = Pr(B_1)Pr(A|B_1) + Pr(B_2)Pr(A|B_2) + Pr(B_3)Pr(A|B_3)$$

This rule can be extended to any number of mutually exclusive and exhaustive events B_1, B_2, \dots, B_k .

Bayes' Theorem

Bayes' theorem can be derived from the Rule of Total Probability by assuming that the probabilities in the right-hand side of the equation are known. We can then use this information to find the probabilities $\Pr(B_1|A)$, $\Pr(B_2|A)$ and $\Pr(B_3|A)$. Let us look at the derivation for $\Pr(B_1|A)$. From the Multiplication Rule and Rule of Total Probability, note that

$$\Pr(B_1|A) = \frac{\Pr(B_1 \cap A)}{\Pr(A)} = \frac{\Pr(B_1)\Pr(A|B_1)}{\Pr(B_1)\Pr(A|B_1) + \Pr(B_2)\Pr(A|B_2) + \Pr(B_3)\Pr(A|B_3)}$$

This formula is known as Bayes' theorem and can be generalized to any number of events B_1, B_2, \dots, B_k .

In the formula for Bayes' theorem above, $\Pr(B_1)$ is called a prior probability because it represents the probability of event B_1 without any knowledge on A . Consequently, the revised probability $\Pr(B_1|A)$ is called a posterior probability because it is calculated with additional information on event A . In some texts, $\Pr(A|B_1)$ is referred to as the likelihood (this terminology is more meaningful when A represents parameter set and B_1 are the observed data. The concept of parameters and likelihood will be visited in detail in Week 3 when we discuss parameter estimation. The probability $\Pr(A)$ is also known as evidence, an easily relatable concept when Bayes' Theorem is applied in forensic assessments.

The following video explains Bayes' theorem using a very classic example on diagnostic testing for diseases.

Conditional Probability Explained: Visual Intuition

https://youtu.be/by3_weGwnMg

Source: ([Harvard Online](#), 2020)

Watch the following video for an example on application of Bayes' Theorem for calculating the probability that an e-mail is a spam, with available prior probability and other related information.

Bayes Theorem Example

https://youtu.be/Mr2GX-K_UPA

Source: ([Selina YC Low](#), 2023)

1.3.1 Application of Bayes' Theorem

One of the application of Bayes' theorem is in [Bayesian search theory](https://www.siam.org/publications/siam-news/articles/bayesian-search-for-missing-aircraft-ships-and-people) (<https://www.siam.org/publications/siam-news/articles/bayesian-search-for-missing-aircraft-ships-and-people>). Take a look at the following [YouTube video](https://www.youtube.com/watch?v=82q3uYw6MuY) (<https://www.youtube.com/watch?v=82q3uYw6MuY>) which explains the use of this theory. This video takes around 15 minutes to complete watching.

Bayes Theorem and some of the mysteries it has solved

<https://www.youtube.com/watch?v=82q3uYw6MuY>

Source: (*Zach Star*, 2019)

Given the wide application of Bayes' theorem, which area of application are you most interested or curious about on application of Bayes' theorem? To get you started with some idea on how Bayes' theorem has been applied in real life scenarios, check out the following articles:

- Allenby, G. M., Bakken, D. G., & Rossi, P. E. (2004). The HB revolution how Bayesian methods have changed the face of marketing research. *Marketing Research*, 16(2), 20-25.
<https://research.ebsco.com/linkprocessor/plink?id=2bbfd4f8-e1d6-3606-8a6a-c9559b462fc4>
- Coyle, P. (2018, Nov 1). What is Bayesian Statistics used for? Towards Data Science. <https://towardsdatascience.com/what-is-bayesian-statistics-used-for-37b91c2c257c>
- McGrayne, S.B. (2011, May 1). Why Bayes Rules: The History of a Formula that Drives Modern Life. *Scientific American*, 304(5).
<https://www.scientificamerican.com/article/why-bayes-rules/>
- Thompson, W. C., Vuille, J., Biedermann, A., & Taroni, F. (2013). The role of prior probability in forensic assessments. *Frontiers in genetics*, 4, 220.
<https://doi.org/10.3389/fgene.2013.00220>

Time: 30 minutes

Purpose: The purpose of this task is to introduce students to applications of Bayes' Theorem in real life scenarios.

1.4 Discrete and Continuous Probability Distributions

Introduction

In Week 0, we defined discrete and continuous random variables. Do you still remember the examples of the respective type of random variables? We continue the discussion on random variables in this topic by looking at their probability distributions. As its name suggests, a probability distribution describes the distribution of the probabilities of the random variable of interest. For example, in the population of adults in Malaysia, the height of male adults could take on non-negative values, with some values being more common (consider that the average height is 164.7 cm) and others being rare (for instance, it is less likely to observe someone whose height is 200 cm). So how are all the possible values of height distributed? We will need the data of all male adults' height measurements in order to be able to know the actual distribution, but it could be reasonable to model the data using the bell-shaped normal probability distribution. We will discuss more about the normal probability distribution in later sections.

In this topic and its subtopics, we explore the characteristics of two selected discrete probability distributions and one continuous probability distribution.

Explore Activity: Distribution Explorer

Time: 30 minutes

Purpose: The purpose of this activity is to use plots created using R to visualize the distribution of selected variables.

Task: In Topic 1.1 Descriptive Statistics, we used the `hist()` function in R to plot a histogram for quantitative variable. A histogram can be used to visualize the distribution of a data set thus helping the data scientist to subsequently choose appropriate statistical methods and distributions to analyze the random variable and/or data set. Using the `AmesHouseNormal.csv` and `campaign-success.csv` data sets, use R to create a simple bar chart or histogram for two of the variables of interest. Based on the bar chart or histogram, describe the distribution (e.g. shape of

the distribution (skewness, modality), range of values, concentration of values) of the variables.

1.4.1 Discrete Probability Distributions

The probability distribution of a discrete random variable, e.g. number of defects in a batch from the production line, is known as a discrete probability distribution. There are many discrete probability distributions proposed in the literature to cater for different data characteristics. Two of the most fundamental discrete distributions are binomial distribution and Poisson distribution.

Binomial Distribution

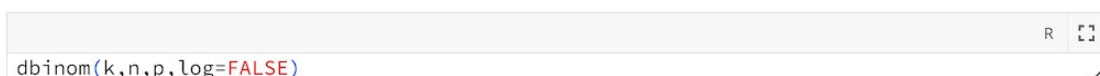
The binomial distribution is used for modelling the number of “successes” in an experiment which consists of identical and independent “trials”, also known as Bernoulli trials. Each of these trials has only two possible outcomes: success or failure. The definition of what constitutes a “success” depends on the experimenter. The classic experiment of flipping a coin is a binomial experiment, in which observing a “head” could be defined as a “success”. Other examples include:

- to buy or not to buy in a consumer decision making process
- default or non-default by a loan borrower
- defect or non-defect in a quality control process

Using the binomial distribution, if the “trials” are repeated n times, and the probability of “success” is given as p , then the probability of observing k successes is given as follows:

$$Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$


In R, the `dbinom()` function is used to calculate the probabilities.

A screenshot of an R console window. The title bar shows 'R' and a window icon. The text area contains the function call `dbinom(k,n,p,log=FALSE)` in a monospaced font. The word 'FALSE' is highlighted in red. A cursor is visible at the end of the line.

Of course, this is a purely theoretical model in which other factors that could affect the outcome were not taken into consideration. In real world, a regression model is often more useful to explain the effect of these other factors on the outcome of interest. For example, a retailer may be interested in the effect of age group in the

consumer decision process. This additional information will be able to help in targeted marketing. Similarly, the credit facility such as banks would be interested to know the probability of a loan borrower defaulting given information on the borrower such as monthly income, number of dependents and so on. Nevertheless, this theoretical model is useful to visualize the distribution and characteristics of the variable before further modelling or even anomaly detection.

Example: In investigating the number of click-sale conversion on an e-commerce website, the analyst observes a total of 200 clicks. If the probability of a click converting to a sale is 0.04, what is the probability of not observing any sales in 200 clicks?



```
1 dbinom(0,200,0.04,log=FALSE)

[1] 0.0002846077

✓ Program exited with code 0
```

Watch the following video by TED-Ed (2016) on an example application of binomial distribution in airlines industry.

Why do airlines sell too many tickets? - Nina Klietsch

<https://www.youtube.com/watch?v=ZFNstNKgEDI>

Source: (TED-Ed, 2016)

A special case of the binomial distribution is the Bernoulli distribution where only one trial is conducted ($n = 1$).

Poisson Distribution

The number of counts that occur within a pre-determined fixed interval, such as a fixed period in time, is often of interest in various phenomena. Examples include:


- In operations management a car park facility manager could be interested in the number of cars entering the car park facility between 12 pm and 1 pm on a weekday

- To optimize human resource allocation, a bank manager may want to know the number of customers requiring counter service at the bank between 10 am and 12 pm
- A network engineer would like to be able to model the number of packets arriving at a network link for system improvement

The Poisson distribution with average λ is the baseline distribution used for modelling these number of counts with the probability formula:



$$\Pr(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

where k is the number of counts and e is the exponential number. The parameter λ is the mean and also the variance of X , a phenomenon known as equidispersion. The Poisson distribution or Poisson process also plays an important role in queueing theory. In R, the `dpois` function is used to calculate the probabilities of a variable which is believed to assume the Poisson distribution.

```
R 
```


```
dpois(k, lambda, log=FALSE)
```

As an example, if the average arrival rate of customers at a bank during office hours is given as 6 per hour, then the probability of 8 customers arriving at any given hour, $P(X=8)$ can be calculated as:

```
 Run R 
```

```
1 dpois(8,6,log=FALSE)
2 ppois(8,lambda=6,lower=FALSE)
```

```
[1] 0.1032577
[1] 0.1527625
```

 Program exited with code 0

These probability values, along with other relevant information, can then be used to assist in decision-making such as deciding the optimal number of staff to be on ground duty.

Question 1

The probability that a patient recovers from a delicate heart operation is 0.9. Let X denote the number of patients who recovers in the next 100 patients having this operation. What is an appropriate probability distribution for X ?

Ans: Binomial Distribution

Explanation: For each patient, there are two possible outcomes - recover or not recover. The probability of recovering is constant from patient to patient, as assumed in this question. Therefore, X has the properties of a binomial random variable.

Question 2

Which of the following statements are true about the Poisson distribution? Select all that applies.

Ans:

- The Poisson distribution has only one parameter.
- The mean of a Poisson distribution is equal to its variance.
- The Poisson distribution is characterized by its mean.

Explanation: A Poisson random variable can take on nonnegative integer values only.

1.4.2 Continuous Probability Distributions

The probability distribution of a continuous random variable, e.g. sale price of a house, is known as continuous probability distribution. As with discrete distributions, there are a plethora of continuous distributions being proposed to account for unique characteristics of the data. The most well-known continuous probability distribution is the normal distribution.

Normal Distribution

The normal distribution (also known as Gaussian distribution after Carl Friedrich Gauss) is a bell-shaped distribution which is completely specified by its mean (μ) and standard deviation (σ). The probability density function for a normal distribution is given as

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

The normal distribution is always centered and has a peak at its mean. The normal distribution plays an important role in statistics and data science as many theorems and regression models are built on the assumption of normally distributed data or residuals.

In a normal distribution, 68% of the data lies within one standard deviation of the mean, and 95% lies within two standard deviations. Watch the following video by Simple Learning Pro (2019a) for further explanation on normal distribution.

The Normal Distribution and the 68-95-99.7 Rule (5.2)

<https://www.youtube.com/watch?v=mtbJbDwqWLE>

Source: (*Simple Learning Pro*, 2019)

Traditionally, probabilities of a normal random variable are calculated by transforming it to a standard normal variable (also known as z-score) followed by use of a statistical table.

$$Z = \frac{X - \mu}{\sigma}$$

A standard normal distribution is a normal distribution in which the mean is 0 and standard deviation is 1. In this case, the units on the x-axis can be expressed in terms of standard deviations away from the mean. Watch the following video by Simple Learning Pro (2019b) on the standard normal distribution.

Z-Scores, Standardization, and the Standard Normal Distribution (5.3)

https://www.youtube.com/watch?v=2tuBREK_mgE

Source: (*Simple Learning Pro*, 2019)

With the advent of statistical computing software, normal probabilities can be calculated using statistical software. In R, the probabilities are easily calculated using the `pnorm()` function which computes the cumulative distribution function $P(X \leq x)$.

For example, the following R code returns $P(X \leq 5000)$ for a random variable X which is normally distributed with mean 3000 and standard deviation 1000.

```
► Run R [ ]
pnorm(5000, 3000, 1000)

[1] 0.9772499

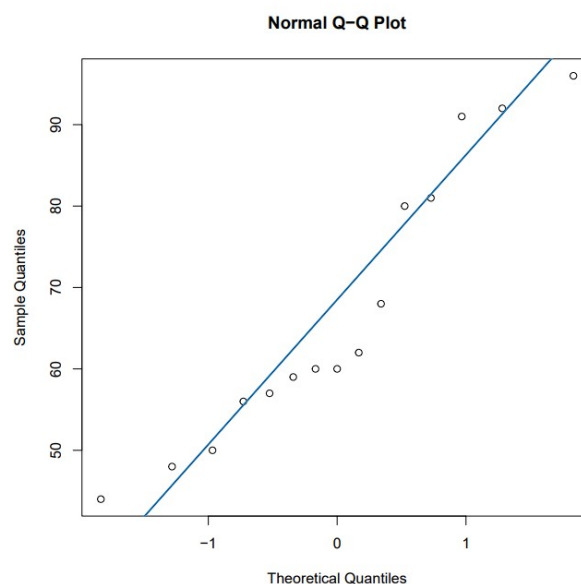
✓ Program exited with code 0
```

A quantile-quantile (QQ) plot can be used to visually determine how close a set of data points is to the normal distribution. If the points roughly fall on the diagonal line, then the sample distribution can be considered close to normal. In R, the `qqnorm()` and `qqline()` functions produce a QQ plot to examine the distribution of the variable.

```
► Run R [ ]
1 weight <- c(62, 81, 56, 91, 59, 57, 96, 92, 60, 48, 60, 68, 50, 44, 80)
2 qqnorm(weight)
3 qqline(weight, col = "steelblue", lwd = 2)

Rplots.pdf

✓ Program exited with code 0
```



QQ plots are particularly useful in linear regression analysis (to be covered in further details in Week 4) when checking for the assumption of normality for the data set.

We wrap up this week's topic with a quick preview of two important distributions frequently encountered in statistical hypothesis testing: F distribution and Student's t-distribution.

F Distribution and Student's t Distribution

The F distribution and Student's t-distribution are two important distributions frequently encountered in statistical hypothesis testing. The F distribution is closely related to another continuous distribution known as the chi-square distribution. Watch the following video by jbststatistics (2012) for an introduction to the F distribution.

An Introduction to the F Distribution

https://www.youtube.com/watch?v=G_RDxAZJ-ug&t=16s

Source: (*jbststatistics*, 2012)

The t-distribution has a bell shape (similar to the normal distribution) and is characterized by its *degrees of freedom*. It is a small sample size approximation of a normal distribution and is quite similar to the standard normal distribution for values of n greater than approximately 30. This distribution is often encountered in hypothesis tests for comparing two groups as well as tests on significance of association between predictor and response variables in linear regression analysis. Watch the following video up to minute 2:12 for an introduction to the t-distribution by 365 Data Science (2019).

Introduction to Probability: Student's T Distribution

<https://www.youtube.com/watch?v=t4hpjK1z5uY>

Source: (*365 Data Science*, 2019)

Question 1

The SAT college entrance exams are taken by thousands of students each year. The mathematics portion of the exam produces scores that are approximately normally distributed. In recent years, SAT mathematics exam scores have averaged 480 with standard deviation 100. What is the z-score for the exam score of a student who attained 550 marks?

Ans: 0.7

Explanation: The z-score is calculated as $Z = \frac{X - \mu}{\sigma}$, where μ is the population average and σ is the standard deviation.

Question 2

What is the mean and standard deviation of a standard normal distribution, respectively?

Ans: 0,1

Explanation: A standard normal distribution has a mean of 0 and standard deviation of 1.

Question 3

Which one of the following is **not true** about a normal probability distribution?

Ans: The standard deviation of the normal distribution determines its skewness.

Explanation: The standard deviation of the normal distribution determines its spread, not skewness.